pagepress

Geospatial Health 2025; volume 20:1386

The future of spatial epidemiology in the AI era: enhancing machine learning approaches with explicit spatial structure

Nima Kianfar,[1] Benn Sartorius,[2] Colleen L Lau,[2] Robert Bergquist,[3] Behzad Kiani[2]

[1]Department of Geospatial Information Systems, Faculty of Geomatics Engineering, K.N. Toosi University of Technology, Tehran, Iran; [2]UQ Centre for Clinical Research (UQCCR), Faculty of Health, Medicine, and Behavioural Sciences, The University of Queensland, Brisbane, Australia; [3]Geospatial Health, Ingerod, Brastad, Sweden

Spatial epidemiology, defined as the study of spatial patterns in disease burdens or health outcomes, aims to estimate disease risk or incidence by identifying geographical risk factors and populations at risk (Morrison *et al.*, 2024). Research in spatial epidemiology relies on both conventional approaches and Machine-Learning (ML) algorithms to explore geographic patterns of diseases and identify influential factors (Pfeiffer & Stevens, 2015). Traditional spatial techniques, including spatial autocorrelation using global Moran's I, Geary's C (Amgalan *et al.*, 2022), and Ripley's K Function (Kan *et al.*, 2022), Local Indicators of Spatial Association (LISA) (Sansuk *et al.*, 2023), hotspot analysis by Getis-Ord Gi* (Lun *et al.*, 2022), spatial lag models (Rey & Franklin, 2022), and Geographically Weighted Regression (GWR) (Kiani *et al.*, 2024) are designed to explicitly incorporate the spatial structure of data into spatial modelling, often referred to as spatially aware models (Reich *et al.*, 2021). Beyond these models, several other spatially aware approaches that have been widely applied in epidemiological studies include but are not limited to Bayesian spatial models that account for spatial uncertainty in disease mapping, such as Bayesian Hierarchical models, Conditional Autoregressive (CAR), and Besage, York, and Mollie' (BYM) models (Louzada *et al.*, 2021). Bayesian methods are statistically rigorous techniques that assume neighboring regions share similar values. Kulldorff's Spatial Scan Statistic is another traditional spatial technique that uses a moving circular window to extract significant disease clusters (Tango, 2021). Moreover, geostatistical models such as Kriging and Inverse Distance Weighting (IDW) allow for continuous spatial interpolation of health data (Nayak *et al.*, 2021).

Most spatial models are formulated to inherently embed spatial dependency and spatial non-stationarity, acknowledging that diseases are not randomly dispersed but influenced by spatial relationships, neighbourhood effects (local spatial interactions), and regional connectivity, such as spatial networks and movement (Kianfar & Mesgari, 2022). For instance, Spatial Lag Models (SLM), another widely recognized type of spatially driven technique, introduce spatial dependencies by incorporating neighbouring values as predictors. Since epidemics rarely spread randomly, the built-in spatial structure in traditional spatial analysis techniques plays a fundamental role in spatial epidemiological modelling approaches (Gaudart *et al.*, 2021). However, the high computational complexity of existing conventional methods, especially Bayesian geostatistical and related models, and others limitations in predictive accuracy due to assumptions like linear relationships or stationarity in spatial processes might reduce both their applicability and validity for large-scale implementation (Kwan, 2021).

To handle large-scale data and overcome limiting assumptions regarding the distribution of predictors, Artificial Intelligence (AI)—along with the rapid rise of ML and Deep Learning (DL) as specific AI techniques—has revolutionized spatial epidemiology by introducing new innovations in disease mapping and prediction (VoPham *et al.*, 2018). Compared to traditional spatial analysis techniques, these advanced computational AI methods generally offer enhanced predictive accuracy due to their capability of uncovering complex, non-linear relationships in epidemiological data (Wiemken & Kelley, 2020). These approaches are broadly categorized into two main types of supervised and unsupervised algorithms (Alloghani *et al.*, 2020), where the former are commonly applied for regression and classification purposes, and the latter to detect hidden patterns and identify clusters. Supervised learning approaches involve training models based on labeled data, where each input is associated with a known output. Conversely, models in unsupervised learning approaches are provided with unlabeled data.

For instance, supervised algorithms include Random Forest (RF), which is widely used in spatial epidemiology as an ensemble learning algorithm for its ability to handle high-dimensional, non-linear and complex interactions among risk factors. It reduces the

Correspondence: Behzad Kiani, UQ Centre for Clinical Research (UQCCR), Faculty of Health, Medicine, and Behavioural Sciences, The University of Queensland, Brisbane, Australia.
E-mail: B.kiani@uq.edu.au

Key words: spatial dependence, artificial intelligence, machine learning, disease mapping, spatial analysis, predictive modeling, public health.

Conflict of interest: the authors declare no potential conflict of interest, and all authors confirm accuracy.

Received: 11 March 2025.
Accepted: 14 April 2025.

Geospatial Health 2025; 20:1386
doi:10.4081/gh.2025.1386

risk of overfitting by combining multiple models through ensemble averaging, a technique that improves accuracy by averaging predictions from multiple models (bagging). Furthermore, RF models can provide feature importance rankings within datasets (Andraud *et al.*, 2021). Decision tree is another commonly used non-parametric supervised modelling approach that classifies a population into branch-like segments to construct an inverted tree with a root node, internal nodes, and leaf nodes (Song & Lu, 2015). Gradient Boosting Machines (GBM) including XGBoost, Light GBM, and CatBoost, also excel in high-accuracy epidemiological predictions by sequentially building trees and correcting errors from previous iterations (boosting) (Li, 2023). Other powerful classification methods include Support Vector Machines (SVMs), which can classify disease hotspots, predict infection risks, and identify outliers in epidemiological data (Muhammad *et al.*, 2021). Additionally, deep learning algorithms, as a subset of Artificial Neural Networks (ANNs) consisting of multi-layered neural nets (deep architectures), perform well when the underlying complex relationships among variables are poorly understood, such as predicting disease incidence where synergetic interaction effects between environmental conditions, socioeconomic factors and population dynamics make it challenging for traditional models to accurately capture patterns (Kianfar *et al.*, 2022). Unsupervised learning algorithms include K-means as a partition-based clustering algorithm that identifies disease incidence clusters and analyzes risk patterns by classifying regions into groups based on similar characteristics (Hutagalung *et al.*, 2021). Hierarchical Clustering (HC) is another unsupervised algorithm that creates a nested hierarchy of clusters, particularly applicable in analyzing health disparities at multiple scales (Uribe *et al.*, 2018).

Spatial epidemiology increasingly relies on ML-based methods to generate outputs such as disease risk maps, hotspot identification, environmental exposure assessments, and spatiotemporal prediction models. Despite their computational efficiency and strong predictive performance compared to traditional spatial techniques, ML algorithms are fundamentally "aspatial", meaning they do not explicitly account for spatial dependencies and geographical relationships within datasets (Nikparvar & Thill, 2021). While these relationships can be implicitly captured through patterns in the data, this approach may be insufficient for accurately modeling complex spatial interactions and underlying heterogeneity that influence disease distribution and the identification of influential risk factors (Rocha *et al.*, 2018). Unlike spatially explicit and geostatistical methodologies, traditional ML models rely solely on data-driven patterns, which can result in outputs that inadequately reflect local variations and spatial heterogeneity. Consequently, ML models may overlook critical spatial nuances and, while they often outperform classical spatial methods, they are not yet the most efficient approach for achieving comprehensive epidemiological insights.

Table 1 summarizes the major distinctions between the main spatial analysis techniques and traditional ML models.

Moving forward: integrating spatial dependence into ML models

Do epidemiologists need to prioritize spatial awareness over higher predictive accuracy, or can they integrate the strengths of both approaches to achieve superior epidemiological insights? The most effective way to achieve the highest level of accuracy in epidemiological studies is to equip ML frameworks with spatial dependence. These hybrid models offer a more comprehensive solution by integrating the predictive strengths of ML algorithms with spatial intelligence of traditional spatial analysis techniques. Several hybrid models have been proposed in recent years to enhance both predictive accuracy and spatial interpretability. For instance, Spatial Random Forest (SRF) has been implemented to incorporate spatial dependencies to improve model accuracy in spatially autocorrelated data (Talebi *et al.*, 2022). Geographically Weighted Random Forest (GWRF) further refines RF by allowing relationships between risk factors and outcomes to vary across space, accounting for spatial heterogeneity by adapting concepts from traditional GWR (Mollalo *et al.*, 2024). Spatial XGBoost, used in health disparity assessments and spatiotemporal predictions, is another composite modelling method that enhances spatial predictive efficiency by incorporating spatial feature engineering such as spatial lag terms and distance-based weight matrices into the ML model (Wu *et al.*, 2021). Moreover, integrating ensemble ML algorithms with Spatiotemporal Gaussian Process Regression (ST-GPR) provides us with another hybrid approach by capturing spatial dependence in data. This hybrid framework incorporates the strengths of ensemble techniques in detecting complex relationships with ST-GPRs ability to model spatial autocorrelation through covariance functions (Lv *et al.*, 2024). Another novel spatially intelligent regression method which incorporates spatial non-stationarity is Geographically Neural Network-Weighted Regression (GNNWR), which blends deep learning with GWR, allowing for spatially varying coefficients while leveraging neural networks for complex pattern recognition. In other words, it retains the interpretable framework of traditional GWR while benefiting

Table 1. Differences between spatial analysis techniques and traditional ML models.

Feature	Spatial analysis techniques	ML-model
Spatial Awareness	Explicitly incorporate spatial neighborhood and dependencies	Do not inherently model spatial relationships
Spatial Autocorrelation	Incorporate autocorrelation using spatial weights	Do not adjust for autocorrelation
Spatial Heterogeneity	Allow relationships to change across space	Assume global relationships
Data Requirements	Require well-structured spatial data with coordinates	Can work with non-spatial data efficiently
Robustness to Noise	Sensitive to spatial data quality and missing values	Can be robust with sufficient data and regularization[a]
Model Flexibility	Use predefined spatial weights and functions	Can learn complex, non-linear patterns
Interpretability	Transparent, explain spatial effects and relationships	Often a 'black box' with limited interpretability
Computational Complexity	Computationally demanding but interpretable	Require large datasets and high processing power
Scalability	Limited in handling very large datasets	Can scale effectively with big data

a Regularization is a technique used to prevent models from overfitting by adding constraints or penalties to the learning process.

from neural networks' capability of modelling non-linear dependencies. By incorporating temporal dynamics, Geographically and Temporally Neural Network-Weighted Regression (GTNNWR) also extends the GNNWR composite approach, making it one of the most enhanced GeoAI models applicable for spatiotemporal epidemiological studies (Yin *et al.*, 2024).

Notably, some potential challenges of hybrid models include computational complexities, parameter tuning difficulties, and scalability constraints, which can be overlooked to some extent as these spatially informed AI models provide significant improvements in spatial epidemiology compared to traditional non-hybrid approaches. During the COVID-19 pandemic, for instance, hybrid models demonstrated substantial improvements in hotspot identification and outbreak prediction outputs (Lucas *et al.*, 2023). This composite modeling approach allowed for both spatial awareness and higher accuracy simultaneously, leading to better resource allocation in affected areas (Du *et al.*, 2020).

In conclusion, rather than viewing traditional spatial analysis techniques and ML algorithms as competitors, the future of spatial epidemiology lies in their logical integration. By developing hybrid models that leverage both spatial intelligence and ML capabilities, researchers and policymakers can gain deeper insights into disease patterns, enhance risk prediction, and ultimately improve public health interventions.

References

Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ, 2020. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In MW Berry, A Mohamed, & BW Yap (Eds.), Supervised and Unsupervised Learning for Data Science (pp. 3–21). Springer International Publishing.

Amgalan A, Mujica-Parodi L, Skiena SS, 2022. Fast spatial autocorrelation. Knowledge Inform Systems 64:919–41.

Andraud M, Bougeard S, Chesnoiu T, Rose N, 2021. Spatiotemporal clustering and Random Forest models to identify risk factors of African swine fever outbreak in Romania in 2018–2019. Sci Rep 11:2098.

Du P, Bai X, Tan K, Xue Z, Samat A, Xia J, Li E, Su H, Liu W, 2020. Advances of four machine learning methods for spatial data handling: a review. J Geovisual Spat Anal 4:13.

Gaudart J, Landier J, Huiart L, Legendre E, Lehot L, Bendiane MK, Chiche L, Petitjean A, Mosnier E, Kirakoya-Samadoulougou F, Demongeot J, Piarroux R, Rebaudet S, 2021. Factors associated with the spatial heterogeneity of the first wave of COVID-19 in France: A nationwide geo-epidemiological study. Lancet Public Health 6:e222–31.

Hutagalung J, Ginantra NLWSR, Bhawika GW, Parwita WGS, Wanto A, Panjaitan PD, 2021. COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm. J Physics Confer Series 1783:012027.

Kan Z, Kwan M, Tang L, 2022. Ripley's K□function for Network□Constrained Flow Data. Geograph Anal 54:769–88.

Kianfar N, Mesgari MS, 2022. GIS-based spatio-temporal analysis and modeling of COVID-19 incidence rates in Europe. Spatial Spatio-Temporal Epidemiol 41:100498.

Kianfar N, Mesgari MS, Mollalo A, Kaveh M, 2022. Spatio-temporal modeling of COVID-19 prevalence and mortality using artificial neural network algorithms. Spat Spatio-Temp Epidemiol 40;100471.

Kiani B, Sartorius B, Lau CL, Bergquist R, 2024. Mastering geographically weighted regression: Key considerations for building a robust model. Geospat Health 19:1271

Kwan M-P, 2021. The stationarity bias in research on the environmental determinants of health. Health & Place 70:102609.

Li L, 2023. Application of machine learning and data mining in medicine: opportunities and considerations. In M Antonio Aceves-Fernández ed., Artificial Intelligence (Vol. 21). IntechOpen. https://doi.org/10.5772/intechopen.113286

Louzada F, Nascimento DCD, Egbon OA, 2021. Spatial statistical models: an overview under the Bayesian approach. Axioms 10:307.

Lucas B, Vahedi B, Karimzadeh M, 2023. A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. Int J Data Sci Analytics 15:247–66.

Lun X, Wang Y, Zhao C, Wu H, Zhu C, Ma D, Xu M, Wang J, Liu Q, Xu L, Meng F, 2022. Epidemiological characteristics and temporal-spatial analysis of overseas imported dengue fever cases in outbreak provinces of China, 2005–2019. Infect Dis Poverty 11:12.

Lv S, Zhu Y, Cheng L, Zhang J, Shen W, Li X, 2024. Evaluation of the prediction effectiveness for geochemical mapping using machine learning methods: A case study from northern Guangdong Province in China. Sci Total Environ 927:172223.

Mollalo A, Grekousis G, Benitez A, Florez H, Neelon B, Lenert LA, Alekseyenko A, 2024. Factors associated with Alzheimer's Disease Dementia prevalence in the United States: A county-level spatial machine learning analysis. medRxiv https://doi.org/10.1101/2024.07.16.24310529

Morrison CN, Mair CF, Bates L, Duncan DT, Branas CC, Bushover BR, Mehranbod CA, Gobaud AN, Uong S, Forrest S, Roberts L, Rundle AG, 2024. Defining spatial epidemiology: a systematic review and re-orientation. Epidemiology 35:542–55.

Muhammad LJ, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA, 2021. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN Computer Sci 2:11.

Nayak PP, Pai JB, Singla N, Somayaji KS, Kalra D, 2021. Geographic information systems in spatial epidemiology: unveiling new horizons in dental public health. J Int Soc Prev Commun Dent 11:125–31.

Nikparvar, B., & Thill, J.-C. (2021). Machine learning of spatial data. ISPRS Int J Geo-Information 10:600.

Pfeiffer DU, Stevens KB, 2015. Spatial and temporal epidemiological analysis in the Big Data era. Prev Vet Med 122:213–20.

Reich BJ, Yang S, Guan Y, Giffin AB, Miller MJ, Rappold A, 2021. A review of spatial causal inference methods for environmental and epidemiological applications. Int Statist Rev 89:605–34.

Rey SJ, Franklin R, eds., 2022. Spatial econometrics. In: Handbook of Spatial Analysis in the Social Sciences. Edward Elgar Publishing. pp. 101–22.

Rocha AD, Groen TA, Skidmore AK, Darvishzadeh R, Willemen L, 2018. Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. Remote Sensing 10:1263.

Sansuk J, Laohasiriwong W, Sornlorm K, 2023. Spatial association between socio-economic health service factors and sepsis mor-

tality in Thailand. Geospat Health 18:1215.
Song Y-Y, Lu Y, 2015. Decision tree methods: Applications for classification and prediction. Shanghai Arch Psych 27:130–5.
Talebi H, Peeters LJM, Otto A, Tolosana-Delgado R, 2022. A truly spatial random forests algorithm for geoscience data analysis and modelling. Math Geosci 54:1–22.
Tango T, 2021. Spatial scan statistics can be dangerous. Statist Methods Med Res 30:75–86.
Uribe C, Segura B, Baggio HC, Abos A, Garcia-Diaz AI, Campabadal A, Marti MJ, Valldeoriola F, Compta Y, Tolosa E, Junque C, 2018. Cortical atrophy patterns in early Parkinson's disease patients using hierarchical cluster analysis. Parkinsonism & Related Dis 50:3–9.
VoPham T, Hart JE, Laden F, Chiang Y-Y, 2018. Emerging trends in geospatial artificial intelligence (geoAI): Potential applications for environmental epidemiology. Environ Health 17:40.
Wiemken TL, Kelley RR, 2020. Machine learning in epidemiology and health outcomes research. Ann Rev Public Health 41:21–36.
Wu C, Zhou M, Liu P, Yang M, 2021. Analyzing COVID□19 using multisource data: an integrated approach of visualization, spatial regression, and machine learning. GeoHealth 5:e2021GH000439.
Yin Z, Ding J, Liu Y, Wang R, Wang Y, Chen Y, Qi J, Wu S, Du Z, 2024. GNNWR: An open-source package of spatiotemporal intelligent regression methods for modeling spatial and temporal nonstationarity. Geosci Model Develop 17:8455–68.