Using geospatial analysis and explainable machine learning to examine risk factors of out-of-hospital cardiac arrest survival

by

Jielu Zhang

(Under the Direction of Mu Lan, PhD)

Abstract

Out-of-Hospital Cardiac Arrest (OHCA) affects over 350,000 Americans annually, yet survival rates remain below 6%. Identifying causal factors and developing comprehensive interventions are critical for improving outcomes. Existing research typically falls into two categories: (1) traditional geospatial analysis, which correlates risk factors and survival outcomes but often provides limited insight into underlying mechanisms, and (2) machine learning (ML), which identifies various risk factors but often overlooks geospatial disparities. Emerging Geospatial Artificial Intelligence (GeoAI) shows that incorporating geospatial variables can significantly enhance ML performance. However, in the domain of geospatially explainable AI (GeoXAI), there is still no comprehensive framework for estimating causal effects in health geography—an essential step for devising effective regional interventions. In this dissertation, we propose a comprehensive framework designed to improve survival outcomes for OHCA patients. This framework comprises three key components, each corresponding to a chapter. First, we introduce the Overlayed Spatio-Temporal Optimization (OSTO), a spatio-temporal placement optimization method that maximizes coverage for potential OHCA patients by accounting for spatiotemporal heterogeneity, we apply this into Washington D.C. and demonstrated an improved OHCA coverage. Second, we present Spatial Counterfactual Explainable Deep Learning (SpaCE), which goes beyond AED placement optimization by exploring additional risk factors that affect survival outcomes. Using a Georgia OHCA outcome prediction task as an example, SpaCE uncovers how various risk factors—and the extent to which they correlate with OHCA survival—vary across different locations. Finally, we introduce Spatially-Aware Causal Inference (SpatialCausal) to estimate causal effects for each treatment variable across space. This method achieves an improved performance over baseline approaches in the Georgia case study. Two standing out variables—AED usage prior to EMS arrival and the identity of the OHCA witness—show a positive treatment effect on survival outcomes, indicating that expanding AED training programs and ensuring more professional individuals can significantly enhance survival chance. Through this framework, we improved OHCA survival by AED placement optimization and the causality estimation of risk factors. The flexibility of this framework demonstrates adaptability for broader geospatial health applications. The methodological advances also serve as a valuable reference for health geographers integrating GeoXAI.

Index words: Out-of-Hospital Cardiac Arrest, Automated External Defibrillator, Spatial Optimization, Geographically Explainable Artificial Intelligence, Spatially-aware Causal Inference

Using geospatial analysis and explainable machine learning to examine risk factors of out-of-hospital cardiac arrest survival

by

Jielu Zhang

B.S., Hunan Agricultural University, 2016
M.S., Nanjing Forestry University, 2020

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2025

Using geospatial analysis and explainable machine learning to examine risk factors of out-of-hospital cardiac arrest survival

by

Jielu Zhang

Major Professor:	Mu Lan
Committee:	Andrew Grundstein Gengchen Mai Donglan Zhang

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Dedication

To my beloved Grandma, who always sees the best in me and believes in my potential—even when I was just a timid young girl. During my teenage years, our winter afternoons spent basking in the sunlight at home made the world feel truly beautiful, and that warmth and wonder continue to help me remain calm and resilient through the challenges of studying and living abroad in the United States. Every joyful phone call with you renews my strength to overcome whatever obstacles come my way.

Acknowledgments

I want to express my sincere gratitude to my advisor, Dr. Mu Lan, for all her guidance and support during my PhD. She was always open to discussing ideas, listening carefully, and offering valuable suggestions. I'd also like to thank my committee members—Dr. Andrew Grundstein, Dr. Gengchen Mai, and Dr. Donglan Zhang—for their advice and support throughout this journey. I additionally thank the members in Dr. Mu Lan's group for their helpful advice: Jue Yang, Nemin Wu, Mohammed Almusidi, Rajib Hossain, and William Hinrichs. I am also grateful to the wonderful faculty and staff members at our department who have supported me throughout my doctoral journey. Specifically, I would like to thank Jennifer L. Rice, Fabiana Hayden, Sarah Baker, Leslie Geiser, Matthew Lars Calonius, for their invaluable support and assistance with all kinds of graduate and international student difficulties.

Finally, I thank my family and Zhongliang Zhou for their endless support and company during this incredible journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

An Overview of Out-of-Hospital Cardiac Arrest Disease Landscape Out-of-hospital cardiac arrest (OHCA) is the loss of functional cardiac mechanical activity associated with an absence of systemic circulation occurring outside of a hospital setting (Myat, Song, and T. Rea 2018). More than 350,000 people suffer from OHCA in the United States(US) each year(American Heart Association) and less than 6% survive (Berger 2017). According to the Georgia Department of Public Health (GDPH) (Georgia Department of Public Health), there were 55,214 recorded instances of OHCA between 2019 and 2021 in Georgia. While the survival outcome at the end of hospital care is largely unavailable, with 90.42% of the data missing, the survival status after the emergency medical services (EMS) is available. Over this three-year span, patients who were still alive at the end of EMS interventions was 54.87%, among which, those still in progress with resuscitation efforts account for 54%, those got in a return of spontaneous circulation (ROSC) accounts for 46%. To save more lives in Georgia, legislation was signed into law in 2017 establishing the Office of Cardiac Care within GDPH(Georgia Department of Public Health). Given the current low survival rates, identifying key factors affecting the survival rate of OHCA is critical to improve the survival probability in Georgia. Potential risk factors (Ha et al. 2022; Moeller et al. 2021; Harford, Darabi, et al. 2019) contributing to OHCA survival include OHCA incident-related characteristics, Social Determinants of Health (SDoH) etc. Among these various factors, OHCA incident-related characteristics —such as response time, type of emergency response, and incident location type—are expected to have more impact on survival rate based on previous research (Al-Dury et al. 2020; Kołtowski et al. 2021; Nishikimi et al. 2019).

A growing number of Explainable Artificial Intelligence (XAI) methods have been developed or applied to explore the relationship between OHCA risk factors and survival outcomes. For instance, Al-Dury et al. (2020) analyzed 45,000 OHCA cases using Random Forest algorithms and a permutation-based importance metric, determining that initial rhythm, patient age, time until CPR initiation, EMS response time, and the cardiac arrest location were key determinants ranked by descending importance. Similarly, Yoshikazu Goto, Maeda, and Yumiko Goto (2013) examined data from 390,226 adult OHCA patients in a nationwide Utstein-style Japanese database (2005–2009) and developed a decision-tree model to predict 1-month survival and favorable neurologic outcomes. Their recursive partitioning analysis of 10 prehospital predictors revealed four key factors—shockable initial rhythm, patient age, witnessed arrest, and an arrest witnessed by EMS personnel—enabling the researchers to stratify patients into four groups (good, moderately good, poor, and absolutely poor) that effectively estimated both survival and neurologic outcomes. In another study, T. D. Rea et al. (2010) analyzed prospective data on adult patients with nontraumatic OHCA from the Resuscitation Outcomes Consortium Epistry–Cardiac Arrest database. Using logistic regression, receiver operating curves, and variance measures, they investigated how effectively the Utstein elements predicted survival to hospital discharge and explained outcome differences across seven sites. Their findings underscore the need to identify additional factors that influence OHCA survival. More recently, Harford, Del Rios, et al. (2022) used 2,398 OHCA cases from the Chicago Fire Department EMS registry to develop machine learning (ML) models predicting whether hospitals would perform coronary angiography and estimating subsequent neurologic outcomes. Their developed Embedded Fully Convolutional Network (EFCN) achieved the highest performance. In a related study from the same research group, Harford, Darabi, et al. (2019) analyzed 2,639 witnessed OHCA cases from Chicago's CARES database using an EFCN model with 27 features, reaching an average class sensitivity of 0.825. Sensitivity analyses suggested layperson CPR could have led to 33 additional survivors with good neurological outcomes, and coronary angiography could have benefitted 88 more patients. Both studies highlight ML's promise for OHCA decision-making.

A key observation from our OHCA data is that cases exhibit distinct spatial patterns. However, traditional XAI methods typically yield averaged results that overlook the spatial variability in patient distribution. Integrating spatial variability into XAI models could enable more targeted, geography-specific intervention strategies.

Machine Learning in Healthcare Applications In recent years, various machine learning (ML) techniques (Mubeen et al. 2017; Neefjes et al. 2017; Alaa et al. 2019; J.-m. Kwon et al. 2019) have been applied to predict health outcomes, capitalizing on their superior performance. Within the ML domain, explainable machine learning (XAI) models have been developed to elucidate the underlying prediction mechanisms, while causal inference models have been employed to assess the effects of risk factors on health outcomes. In this work, we review current research in three key areas: prediction models for health outcomes, explainable ML methods, and causal models applied to health outcome prediction.

Regarding health outcome prediction, Killian et al. (2021) compared logistic regression, naive Bayes, support vector machines, and deep learning to predict 1-, 3-, and 5-year post-transplant hospitalizations among pediatric kidney, liver, and heart transplant recipients. Similarly, Chung et al. (2022) employed topic model cluster analysis to develop a risk prediction model for chronic limb-threatening ischemia (CLTI). Their analysis identified three distinct stages within CLTI and suggested that CLTI-free survival could serve as a robust endpoint for risk prediction. In another study, the data of 193 mild Degenerative Cervical Myelopathy patients were used to predict significant improvement in mental component summary (MCS) and physical component summary (PCS) (Khan et al. 2021). A suite of ML models, along with logistic regression, was trained on a train-test split, with the best-performing models being a generalized boosted model for MCS and an earth model for PCS. Notably, female patients with low initial MCS and symptoms such as lower limb spasticity and clumsy hands were less likely to experience significant improvement.

Despite the strong predictive performance of these models, their complexity often obscures the decision-making process. To address this, tools such as SHapley Additive exPlanations (SHAP) (Lundberg and S.-I. Lee 2017) and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) are widely used to deconstruct complex models and highlight significant features (Loh et al. 2022). For example, Kor et al. (2022) applied a gradient-boosting machine to forecast first-time acute exacerbation of chronic obstructive pulmonary disease, with SHAP revealing that chronic obstructive pulmonary disease assessment test scores and wheezing were pivotal predictors. In another study, Dindorf et al. (2020) used LIME to interpret the support vector machine model classifying post-hip surgery walking patterns, identifying key movements in the hip, knee, and ankle. In addition to XAI, causal inference models have been utilized to investigate the impact of risk factors on health outcomes. Echajei et al. (2024) integrated causal inference with ML for early diabetes diagnosis using a stacking ensemble that combined Random Forest, Extreme Gradient Boosting, and Gradient Boosting. Enhanced by advanced data preprocessing, feature engineering, and hyperparameter optimization, this approach demonstrated significant promise for effective diabetes management. S. Rao et al. (2022) developed a transformer-based model, targeted bidirectional EHR transformer, combined with doubly robust estimation to infer the causal effect of antihypertensive classes on incident cancer risk using comprehensive electronic health record data. In semi-synthetic experiments with various confounding scenarios and limited data, the model outperformed traditional benchmarks, yielding estimates consistent with randomized clinical trial findings.

Many chronic diseases, such as cardiovascular disease, exhibit distinct spatial patterns in survival outcomes and prevalence (Nazia et al. 2022; Djukpen 2012; Mena et al. 2018; Sahar et al. 2021; Son et al. 2023). Although current research shows promising results, it often fails to incorporate spatial information, overlooking its role in the relationship between risk factors and health outcomes (D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016). Integrating spatial information into explainable AI models to explore these relationships—and further incorporating such data into causal models to better estimate the causal effects of identified risk factors—remains a critical challenge.

1.2 Research Gaps and Addressed Questions

Based on existing research in OHCA health outcome prediction, risk factor identification, and recent advances in healthcare machine learning, we have identified four key research gaps:

- **Comprehensive System Gap (*Gap 1*):** No systematic approach currently exists that integrates both practical interventions (e.g., optimizing Automated External Defibrillator accessibility) and data-driven risk factor exploration to improve OHCA survival outcomes. This gap motivates the overall structure of our dissertation.

- **Spatial-Temporal AED Placement (*Gap 2*):** Prior studies addressing the maximal covering location problem for AED placement have not considered the spatially- and temporally-varying distribution of potential OHCA cases. This gap is tackled in ***Chapter 2***.
- **Spatial Information Integration (*Gap 3*):** Existing health outcome prediction models fail to incorporate spatial information which is crucial for guiding geography-specific interventions. Through the development of a counterfactual explanation based Geospatially Explainable Artificial Intelligence (GeoXAI) method, this limitation is tackled in ***Chapter 3***
- **Spatially-Aware Causal Inference Gap (*Gap 4*):** To our knowledge, deep learning–based spatial causal inference has not been explored in the field of health geography. Moreover, current causal inference models do not integrate spatial information, limiting our ability to estimate the spatially varying causal effects of risk factors on health outcomes. This challenge is tackled in ***Chapter 4***.

In order to address these gaps, our dissertation is organized into a step-by-step framework. Chapters 2, 3, and 4 specifically tackle the issues of spatio-temporal AED placement, spatial information integration, and spatially-aware causal inference, respectively. To bridge these gaps, each chapter is designed to answer designated questions.

- Spatio-temporal Placement Optimization
  - How is the spatially and temporally varying potential OHCA distribution integrated into the AED optimization model to enhance accessibility?
  - How can the cost-coverage increment curve be estimated, taking into account budget limitations in real-world AED placement scenarios?
- Spatial Counterfactual Explanation
  - How can spatial information be effectively integrated into an explainable machine learning model?
  - In pursuit of improved OHCA survival outcomes through targeted intervention, how is the counterfactual explanation concept incorporated into the developed spatially-varying explainable machine learning model to deliver a 'what if' analysis for expired OHCA cases, thereby providing direct intervention guidance?
- Spatially-aware Causal Inference
  - How can spatial information be effectively integrated into a causal inference model?
  - How can unmeasured confounders be incorporated or approximately estimated within a causal inference model?

Chapter 2

Spatiotemporal Optimization for the Placement of Automated External Defibrillators Using Mobile Phone Data

Zhang, Jielu, et al. "Spatiotemporal Optimization for the Placement of Automated External Defibrillators Using Mobile Phone Data." ISPRS international journal of geo-information 12.3 (2023)

Reprinted here with permission of the publisher.

2.1 Introduction

Out-of-hospital cardiac arrest (OHCA) is the loss of functional cardiac mechanical activity associated with an absence of systemic circulation occurring outside of a hospital setting (Myat, Song, and T. Rea 2018). More than 350,000 people suffer from OHCA in the United States (US) each year (Association 2022a), and less than 6% of them survive (Berger 2017). From collapse to treatment, every additional minute will reduce the probability of survival by up to 10% (Valenzuela et al. 1997). Public access defibrillation programs using automated external defibrillators (AEDs) are practical and have been linked to a significant increase in the OHCA survival rate (T. Aufderheide et al. 2006; Mary F Hazinski et al. 2005; Investigators 2004). However, fewer than 12% of individuals with OHCAs have an AED applied before the emergency medical services (EMSs) arrive (Myat, Song, and T. Rea 2018). Several factors responsible for this rate include an unawareness of AED locations, a lack of AED training, bystander apathy, and concerns about liability (Gundry et al. 1999; J. H. Lee et al. 2021; Merchant and Asch 2012). Another major factor is the accessibility to AEDs (Ho et al. 2014; Karlsson et al. 2019; P. Kwon et al. 2016). General guidelines proposed by the American Heart Association (AHA) (Association 2022b) suggested that AEDs should be deployed in areas at risk of cardiac arrest, which did not provide a clear plan for delimiting risk boundaries and positioning AEDs. Thus, determining the areas of potential OHCAs is the first step before optimizing the AED placement.

Many previous studies predicted the potential OHCAs using population distribution or historical OHCAs (Chan, H. Li, et al. 2013; Claesson et al. 2017; Leung et al. 2021). However, the bias of using population distribution to model real-time spatial population dynamics has long been revealed (Chan, H. Li, et al. 2013). On the other hand, due to practical factors, the historical OHCA data are not accessible in many scenarios. To our knowledge, the overwhelming majority of research on AED placement is limited to western countries with well-developed systematic programs and dedicated personnel for collecting incident locations of OHCAs, including the National Emergency Medical Services Information System in the US, the Resuscitation Outcomes Consortium in North America, and Scottish Ambulance Service in Scotland, among others. These programs are not always available in some underdeveloped countries or regions due to the global disparity and imbalance in the development of healthcare systems. Additionally, for some small areas, it is impractical to collect sufficient historical OHCA data to perform effective AED placement optimization. In Washington DC, for example, the total number of cardiac arrests was 605 in 2019, and the number of estimated OHCA cases was only 179 (CARES n.d.; Fire and Department n.d.). We overcome the drawbacks of the aforementioned methods by representing the potential distribution of OHCAs with dynamic point of interest (POI) (SafeGraph n.d.) visit data collected from mobile devices.

In contrast to historical OHCA data, which are usually sparse and uneven across nations, mobile location data can be acquired in most countries (SafeGraph n.d.) owing to the extensive reach of mobile phones. By utilizing such data, we may overcome the intrinsic hindrance in gathering OHCA cases and formulate a method that has the potential to span both developed and developing regions, bridging the existing public health disparities between them. POI visit data represent real-time population movement, effectively reflecting potential OHCA risk areas across space and time. Based on the statistics on cardiac arrest location types in the US (CARES n.d.), 60.32% OHCAs occurred in public or commercial buildings. Applying the North American Industry Classification System (NAICS) (Bureau n.d.) on POI visit data suggested that almost all the POIs are subordinate to the public and commercial buildings where most OHCAs occurred.

In addition, the majority of the studies and AHA guidelines for AED placement focus on accessibility concerning the spatial pattern of potential OHCAs, including using the incident location type to predict potential locations at risk of cardiac arrest (Becker et al. 1998; Fedoruk, Currie, and Gobet 2002), leveraging GIS techniques (Fredman et al. 2017), or optimization algorithm (Chan, H. Li, et al. 2013; Chan, Demirtas, and R. H. Kwon 2016; Sun et al. 2020) to improve the access level of AEDs and delivering AEDs with the aid of drones (Boutilier et al. 2017; Claesson et al. 2017). Although improvements have been made by integrating spatial patterns into AED deployment plans, temporal access has been largely ignored in previous research. Given that there exists a substantial difference in the occurrence and survival of OHCAs between the different times of day and different days of the week (Bagai et al. 2013; Brooks et al. 2010), a spatio-temporal analysis method is needed to better model the spatio-temporal characteristics of OHCAs.

In this study, we propose an adaptable overlayed spatio-temporal optimization (OSTO) method to optimize the AED placement using fine-grained analysis in time and space. Taking Washington DC as a case study, the results demonstrate that our proposed methods can successfully maximize the AED performance with respect to all time periods as well as improve the AED access level. We also evaluated the cost–coverage increment curve to determine the most suitable number of total deployed AEDs considering the AED coverage efficiency and financial budget. Our method presents a general framework that can be practically adapted to other facility deployment optimization planning when both spatial and temporal factors are important.

2.2 Overlayed Spatio-Temporal Optimization Method

The proposed OSTO method includes two stages (Figure 2.1). The first stage aims to maximize the AED coverage rate by hour and obtain a total of 24 solutions $\{J'_1, J'_2, J'_3, \dots, J'_{24}\}$ corresponding to 24 hours, with each solution representing a spatial distribution of optimized AEDs. For each hour, the optimized AED coverage rate $\{R_{11}, R_{22}, R_{33}, \dots, R_{2424}\}$ was calculated using the hourly solution to cover the POI visitor distribution of the hour itself. The second stage is to apply each solution into 24 sets of visitor distribution corresponding to 24 hours $\{I_1, I_2, I_3, \dots, I_{24}\}$, calculate the average performance $\{N_1, N_2, N_3, \dots, N_{24}\}$ of each solution over 24 hours, and identify the best solution out of the 24 sets. The next two subsections will illustrate details about the calculation process for these two stages.

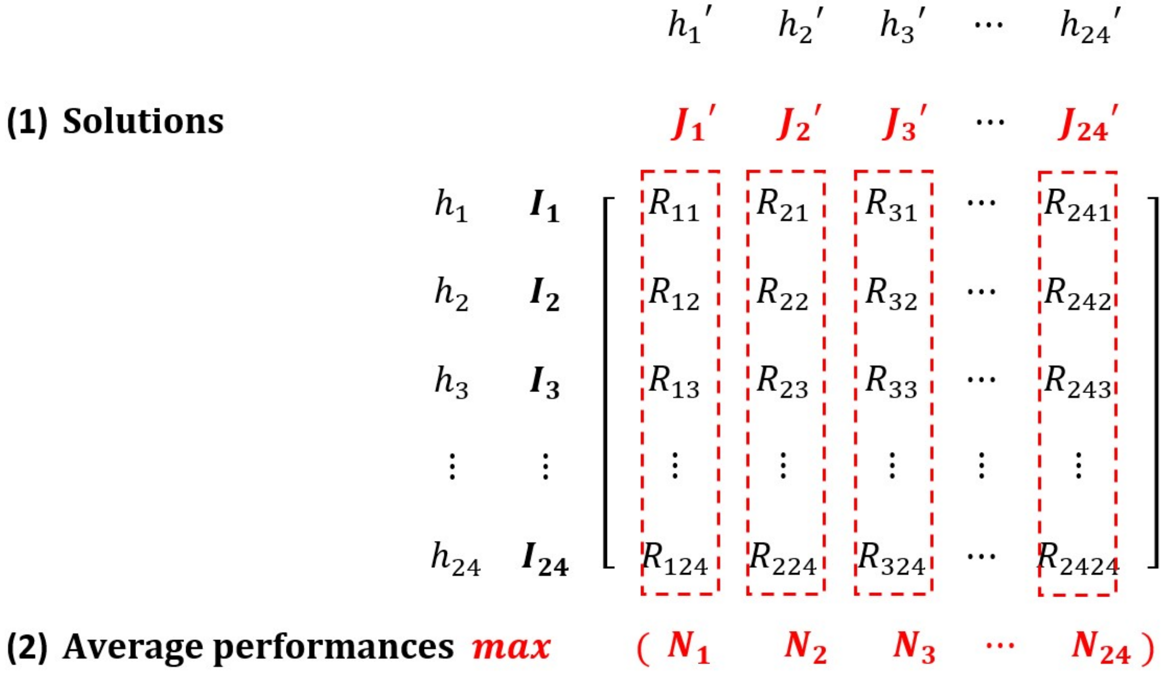

h_1' h_2' h_3' ⋯ h_{24}'
(1) Solutions
J_1' J_2' J_3' ⋯ J_{24}'
h_1 I_1 R_{11} R_{21} R_{31} ⋯ R_{241}
h_2 I_2 R_{12} R_{22} R_{32} ⋯ R_{242}
h_3 I_3 R_{13} R_{23} R_{33} ⋯ R_{243}
h_{24} I_{24} R_{124} R_{224} R_{324} ⋯ R_{2424}
(2) Average performances max (N_1 N_2 N_3 ⋯ N_{24})

Figure 2.1: A Matrix Describing the Process of the Overlayed Spatio-Temporal Optimization. Both $h_1, h_2, h_3, \ldots, h_{24}$ and $h'_1, h'_2, h'_3, \ldots, h'_{24}$ refer to 24 hours a day. $I_1, I_2, I_3, \ldots, I_{24}$ indicate 24 hourly sets of POI visitor distribution. $J'_1, J'_2, J'_3, \ldots, J'_{24}$ represents 24 hourly AED solutions computed based on the visitor distribution of the hour itself. $R_{hh'}(h, h' \in 1, 2, 3, \ldots, 24)$ means the optimized AED coverage rate applying the h'th solution on hth set of POI visitor distribution. $N_1, N_2, N_3, \ldots, N_{24}$ means the 24 average performances for each solution. For instance, N_1 is averaged over $R_{11}, R_{12}, R_{13}, \ldots R_{124}$, and $N_2, N_3, \ldots, N_{24}$ is calculated similarly.

Optimizing AED Placement by Hour This section aims to illustrate the first stage of the OSTO. The mixed-integer programming (MIP) model based on a branch-and-bound algorithm (Gurobi 2022) was deployed to solve the maximal covering location problem (MCLP), which is to locate a restrained number of facilities to maximize the population covered within a particular service distance (Church and ReVelle 1974). Our scenario of optimizing the AED placement aimed to identify a set of locations where placing AEDs would maximize the number of covered POI visitors. The objective function (1) and four constraints (2)–(5) to the MCLP for AED optimizing placement can be expressed as follows:

$$
\begin{array}{llll}
Maximize & z_h = \sum\limits_{i \,\in\, I_h} a_i y_i & for\ all\ h \in H & (2.1) \\
Subject\ to & \sum\limits_{j \,\in\, N_i} x_j y_i & for\ all\ i \in I_h & (2.2) \\
 & \sum\limits_{j \,\in\, J} x_j = k & & (2.3) \\
 & x_j = 0\ or\ 1 & for\ all\ j \in J & (2.4) \\
 & y_i = 0\ or\ 1 & for\ all\ i \in I_h & (2.5)
\end{array}
$$

where:
h = denotes each hour;
H = denotes the set of 24 hours 1, ..., h..., 24;
i = denotes each POI site;
I_h = denotes the set of POI sites for each hour;
j = denotes each AED candidate site;
J = denotes the set of AED candidate sites;
a_i = the number of visitors to be served at POI site i;
k = the number of AEDs to be located;
d_{ij} = the shortest distance from site i to site j;
S = the distance beyond which a POI site is considered uncovered;
$N_i = \{j \in J | d_{ij} S\}$, denotes the set of AED candidate sites that can cover visitors in the POI site I;
Z_h = denotes the number of covered POI visitors;

$$
x_j = \begin{cases} 1 & \text{if an } AED \text{ is allocated to site } j \\ 0 & \text{else} \end{cases}
$$

$$
y_i = \begin{cases} 1 & \text{if one or more AED candidate sites are established at sites in the set } N_i \\ 0 & \text{else} \end{cases}.
$$

The constraints can be explained as follows: for AED selection (4), if an AED is allocated to location j, the x_j is assigned 1, otherwise the x_j is assigned 0. For budget constraint (3), the total number of AEDs to be located is k. For POI sites (2) and (5), if there exists at least one established AED from which a POI site i is within the distance of S, then y_i is assigned 1; otherwise y_i is assigned 0. Embedding these constraints and the objective function into the MIP model, we conducted the model by the hour, and 24 solutions were computed. The potential AED locations provided for the MIP model to select the optimized AED locations are named AED candidate sites in this study.

Identifying the Final Solution of Optimized AEDs This section aimed to illustrate the second stage of the OSTO. The solution of optimized AED locations at each specific hour h was then applied 24 times to hourly visitor distribution, and then the average performance of each solution over 24 sets of POI distribution was calculated ($N_{J'_{h'}}$). A total of 24 $N_{J'_{h'}}$s were generated and the final solution was identified as the solution with maximal $N_{J'_{h'}}$.

Calculation process for identifying the final solution:

$$O_i = \begin{cases} 1 \quad min(d_{J'_{h'}i})\ S \\ \\ 0 \quad else \end{cases} \qquad i \in I_h;\ j' \in J'_{h'};\ h' \in\ H \tag{2.6}$$

$$R_{J'_{h'}I_h} = \frac{\sum_{i\ \in\ I_h} O_i a_i}{\sum_{i\ \in\ I_h} a_i} \qquad h \in H;\ h' \in H \tag{2.7}$$

$$N_{J'_{h'}} = Avg(R_{J_{h'}I_1},\ R_{J_{h'}I_2},\ R_{J_{h'}I_3},\ ...,\ R_{J_{h'}I_{24}}) \qquad h' \in\ H \tag{2.8}$$

$$Objective \qquad Max(N_{J'_1},\ N_{J'_2},\ N_{J'_3},\ \ldots,\ N_{J'_{24}}) \tag{2.9}$$

where:
h' = also denotes each hour, used for distinguishing itself from h;
H = denotes the set of 24 hours 1, ..., h..., 24;
i = denotes each POI site;
I_h = denotes the set of POI sites for each hour;
j' = denotes each optimized AED;
$J'_{h'}$ = denotes the set of optimized AEDs for each hour;
$Min(d_{J'_{h'}i})$ = denotes the minimum distance between each site in the set of optimized AEDs and a POI site;
S = the distance beyond which a POI site is considered uncovered;
O_i = denotes whether a POI site i is covered by a set of optimized AEDs;
a_i = the number of visitors to be served at POI site i;
$R_{J'_{h'}I_h}$ = denotes the optimized AED coverage rate for each hour h';
$N_{J'_{h'}}$ = denotes the average performance of each hour' s set of optimized AEDs.

Four steps for identifying the final solution are: (6) Calculate whether a POI site i is covered by any site of an AED solution (O_i); (7) Obtain the optimized AED coverage rate ($R_{J'_{h'}I_h}$) using a solution in a specific hour (h') to cover a POI visitor distribution in another specific hour (h); (8) Compute the averaged performance ($N_{J'_{h'}}$) of a solution in a specific hour (h'); (9) Compare the average performances between 24 hours and identify the maximal one ($Max(N_{J'_{h'}})$).

Cost-Coverage Increment Analysis The cost-coverage increment curve was evaluated by assessing the increased average performance obtained from an increasing number of AEDs. Specifically, it is calculated by dividing the increased AED average performance ($\Delta N_{J'_{h'}}$) by the increased number of AEDs to be located (Δk) (the intervention cost). Steps for calculating the average performance of a certain k number of AEDs were as follows: (1) The optimized AED locations are computed based on the visitor distribution of the hour ($\hat{h}$) with the highest performance; (2) Calculate the average performance ($N_{J'_{h'}}$) of the set of optimized AED locations on all 24 h visitor distributions; (3) The final cost-coverage increment curve is generated based on $\Delta N_{J'_{h'}}/\Delta k$.

2.3 Application in Washington DC

Data Source and Preprocessing The POI visit count data in 2019 in Washington DC originates from the SafeGraph Pattern Dataset (SafeGraph n.d.), indicating that the count of visits to each POI during each hour accumulated across the whole year of 2019. Considering the fact that the visit duration must last at least 4 min to count as valid for a given POI and that the visitor would be counted once in each hour if they stay for multiple hours (SafeGraph n.d.), visit counts and visitors are interchangeable in our study. Since we focused on deploying public-access defibrillators, we excluded residential areas that are not publicly accessible. Considering the full equipment with healthcare facilities including AEDs in hospitals, we also excluded hospital areas. After removing the POI visit counts in residentials and hospitals, 134,850,972 POI total visit counts were used for analysis. The number of visitors per hour is shown in Figure 2.2. Considering the computational efficiency, we randomly sampled $1/6$ visitor locations per day for the optimization analysis. The sample size was $N_{sample} = N_{total}/365/6 = 61,575$. The value calculated from $N_{total}/365$ means the average daily visits, and further dividing by six means that one in every six POI visit counts.

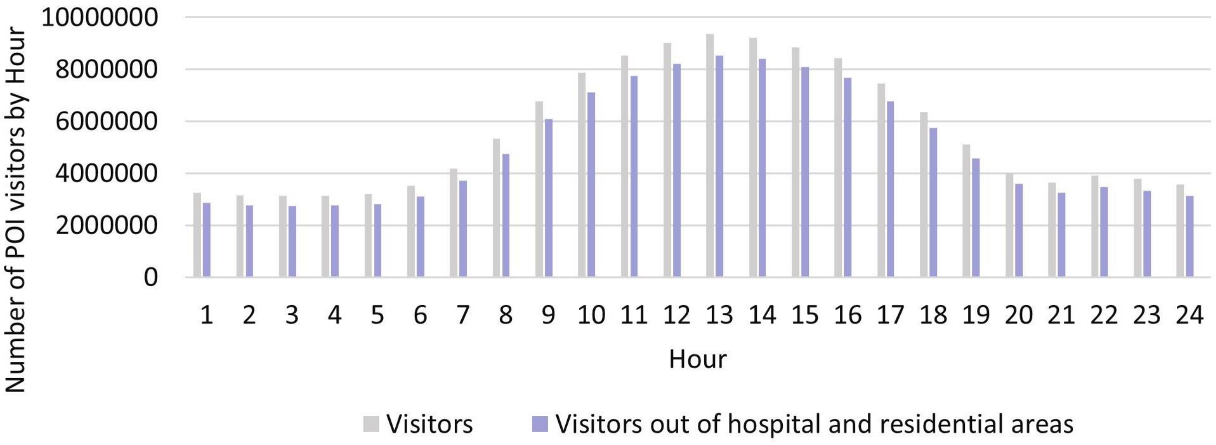


Figure 2.2: Number of POI visitors by Hour in Washington DC.

The data of 2302 AEDs in Washington DC were obtained from Open Data DC (updated until 8 December 2021) (DC n.d.). A total of 2113 AEDs were left after removing the existing AEDs in hospitals and residential areas. A total of 18,098 AED candidate sites were generated from the centroids of buildings other than hospitals and residential areas, and building data were obtained from the Open Street Map (Server n.d.).

Data on 54 hospitals were accessed from the Geographic Names Information System (GNIS) (updated until 25 August 2021) (Survey n.d.); the land use data with 137,626 polygons came from Open Data DC (updated until 18 March 2022), which includes residential data (DC n.d.).

Analysis In our model, we set k as the maximum number of locations where AEDs could be deployed. We set k as equal to 100, 200, and 300 separately to explore the optimization under different situations. The distance S beyond which a POI site is considered uncovered is set as 100 m as suggested by the AHA (T. Aufderheide et al. 2006). To simulate the mobility of visitors, we randomly spread out the visitors of each POI in each hour into a 5 min walking circle. Thus, the a_i (the number of POI visitors to be served at site i) that was used to maximize the z_h (the number of covered POI visitors) in the maximizing function ($Maximize\ z_h = \sum_i \in I_h\ a_i y_i$) was set to 1. Hence, the z_h value is only determined by y_i (whether the visitor is covered or not). The radius of the walking circle is calculated as 414.3 m based on the average comfortable gait speed (138.10 cm/s) of all the adults of different ages (Bohannon 1997). For each k value, we conducted the first stage of the OSTO, and then compared the trendlines of the AED coverage rate across 24 h between different ks. Since trends were similar between different ks, we conducted the second stage of the OSTO only with k equals to 100. The trend of the average performance ($N_{J'_{h'}}$) of each solution across 24 h was then calculated. We used the Python (Python 3.7.11) programming language to code the algebraic formulation of the model and used the Gurobi (Gurobi 2022) solver to solve the optimization problem. The whole process took 446,293.68 s of computing time with the use of a workstation with Intel® Xeon® CPU E5-1650 6 cores 3.78-GHz processor and 128 GB of RAM.

Results

Applying the OSTO in Washington DC The matrix showing the detailed result across the whole process of the OSTO in this case is illustrated in Figure 2.3. The results from the first stage of the OSTO are elaborated as follows. The trends of optimized AED coverage rates across 24 h a day for different ks are highly similar (Figure 2.4), although minor differences exist in the peak and valley hours between these three ks. In general, the peak hours for all three ks are approximately 14:00, and the valley hours are around 6:00 and 22:00, implying a general pattern of dynamic population distribution not significantly affected by the sample size k. Exploring the divergence of the optimized AED coverage rate between different hours helps us understand the regularity of how the optimization algorithm works on different visitor distributions. We analyzed the spatial clustering level of POI visitors of each hour by adopting the kernel density estimation and directional distribution analyses, and it was found that the optimized AED coverage rate is positively associated with visitor clustering level. Take three particular hours (one peak and two valleys) as an example (Figure 2.5). One standard deviation ellipse includes 68% of visitors, and two standard deviations ellipse includes 95% of visitors. Compared with 6:00, a more compacted directional ellipse and a higher kernel density in the central area at 14:00 indicates a higher clustering level of visitors, and a higher optimized coverage rate. Compared with 14:00, a more expanded directional ellipse and a lower kernel density at 22:00 indicate a lower clustering level and a lower optimized coverage rate. The optimized coverage rate can be referred to in Figure 2.4.

$$
\begin{array}{cc|ccccccccccccccc|}
 & & h_1' & h_2' & h_3' & \cdots & h_6' & \cdots & h_{14}' & \cdots & h_{19}' & \cdots & h_{21}' & h_{22}' & h_{23}' & h_{24}' \\
 & & J_1' & J_2' & J_3' & \cdots & J_6' & \cdots & J_{14}' & \cdots & J_{19}' & \cdots & J_{21}' & J_{22}' & J_{23}' & J_{24}' \\
h_1 & I_1 & 17.58 & 17.37 & 17.42 & \cdots & 17.05 & \cdots & 12.03 & \cdots & 15.10 & \cdots & 15.47 & 16.05 & 17.04 & 17.30 \\
h_2 & I_2 & 17.62 & 17.43 & 17.57 & \cdots & 17.15 & \cdots & 11.87 & \cdots & 14.89 & \cdots & 15.23 & 15.84 & 16.94 & 17.25 \\
h_3 & I_3 & 17.33 & 17.16 & 17.41 & \cdots & 17.09 & \cdots & 11.69 & \cdots & 14.61 & \cdots & 14.89 & 15.52 & 16.66 & 16.96 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
h_7 & I_7 & 16.04 & 15.95 & 16.06 & \cdots & 16.39 & \cdots & 13.78 & \cdots & 15.32 & \cdots & 15.08 & 15.32 & 16.10 & 16.11 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
h_{14} & I_{14} & 14.38 & 14.54 & 13.81 & \cdots & 15.15 & \cdots & 21.29 & \cdots & 19.75 & \cdots & 18.03 & 16.90 & 16.38 & 15.24 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
h_{19} & I_{19} & 15.33 & 15.20 & 14.72 & \cdots & 15.67 & \cdots & 17.23 & \cdots & 19.11 & \cdots & 18.42 & 17.88 & 17.32 & 16.27 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
h_{21} & I_{21} & 14.87 & 14.61 & 14.37 & \cdots & 14.75 & \cdots & 14.31 & \cdots & 16.86 & \cdots & 17.37 & 17.09 & 16.5 & 15.45 \\
h_{22} & I_{22} & 15.79 & 15.54 & 15.42 & \cdots & 15.61 & \cdots & 13.79 & \cdots & 16.61 & \cdots & 17.16 & 17.21 & 16.96 & 16.22 \\
h_{23} & I_{23} & 16.75 & 16.50 & 16.51 & \cdots & 16.51 & \cdots & 13.12 & \cdots & 16.11 & \cdots & 16.67 & 16.99 & 17.31 & 17.01 \\
h_{24} & I_{24} & 17.34 & 17.10 & 17.18 & \cdots & 16.98 & \cdots & 12.46 & \cdots & 15.55 & \cdots & 15.98 & 16.51 & 17.24 & 17.36 \\
\multicolumn{2}{c}{Max} & (15.65 & 15.61 & 15.35 & \cdots & 15.93 & \cdots & 16.17 & \cdots & \mathbf{17.25} & \cdots & 16.67 & 16.42 & 16.57 & 16.06)
\end{array}
$$

Figure 2.3: A Matrix Showing the Whole Process of the OSTO on the Case in Washington DC (k = 100). All numbers in Figure 3 are percentages. Numbers in black show the AED coverage rate when applying the solution in a specific hour (e.g., J_1') to cover a visitor distribution at all hours (e.g., h_1, h_2, ..., h_{24}). Numbers in red show 24 of the average performances $(N_{J_{h'}'})$ of each solution. The number marked in red bold is the highest average performance 17.25% (when J_{19}' is applied to all hours).

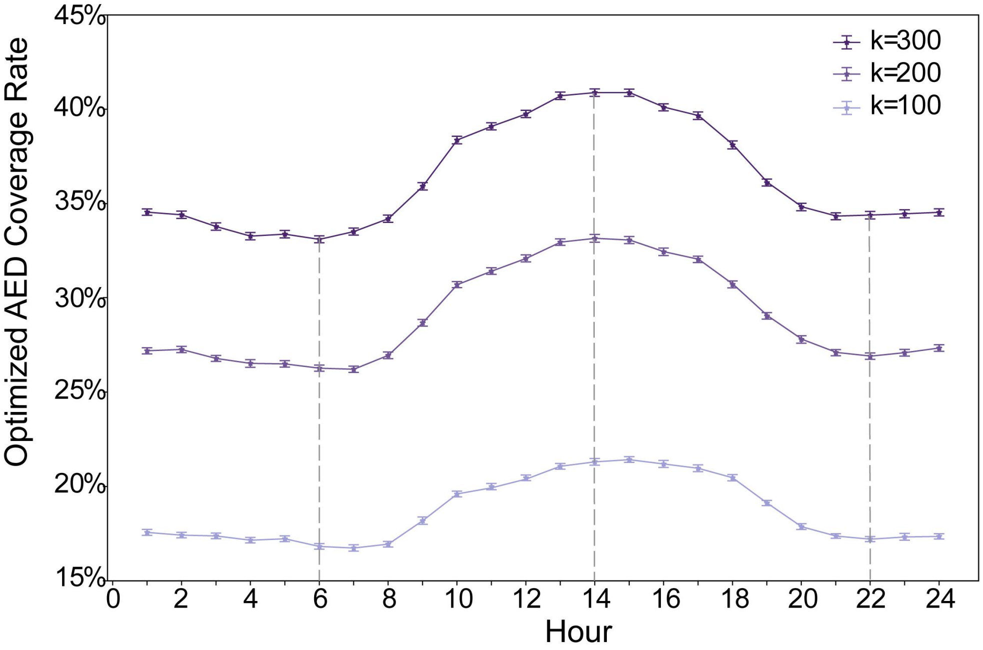


Figure 2.4: Optimized AED coverage rate. The optimized AED coverage rate ($R_{J'_h I_h}$) refers to the coverage rate using the set of optimized AEDs in a specific hour to cover the visitor distribution of the hour itself. For each hour, the error bar is generated from resampling 61,575 visitors from the total number of visitors 100 times.

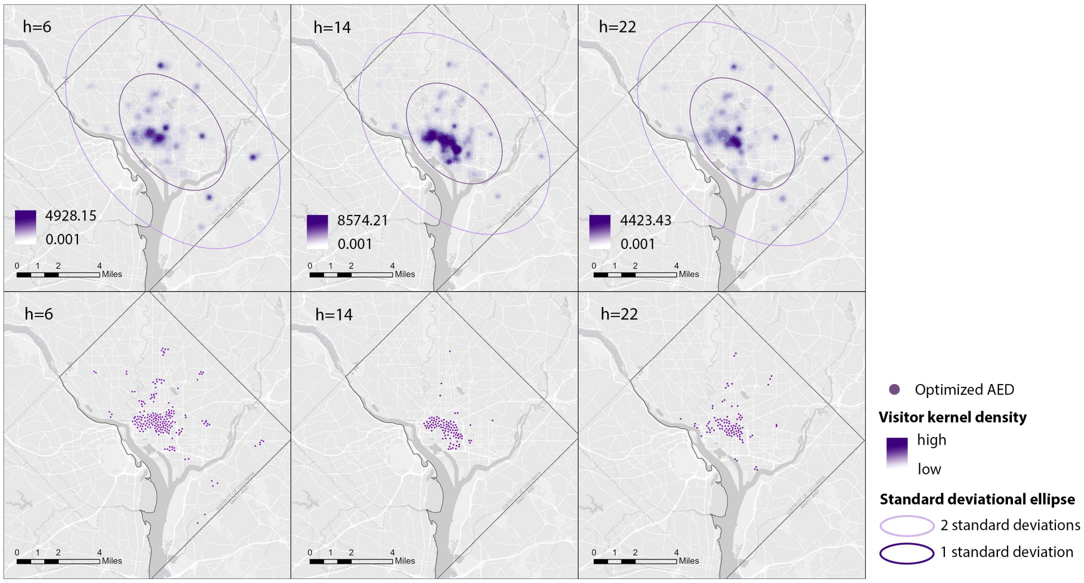


Figure 2.5: Kernel Density Maps of Visitors and Corresponding Optimized AED Locations at Three Particular Hours (k = 100).

The results from the second stage of the OSTO are also elaborated as follows. The hour of 19:00 has the highest average performance of optimized AEDs (Figures 2.3 and 2.6). The trend and the magnitude of the average performance across 24 h were compared with that of the optimized AED coverage rate with k equals 100 (Figure 2.6). The curve of the optimized AED coverage rate is almost reversed to that of average performance. We noticed exceptions from the reversed shape at the beginning and end of the curves, but in general, the reversal relationship dominates. Specific values of optimized AED coverage rates and average performances at particular hours $\{h'_6, h'_{14}, h'_{19}, \ldots, h'_{22}\}$ are shown in Figure 2.3. For the magnitude comparison, the range of the magnitude of optimized AED average performance is smaller than that of the optimized AED coverage rate. Average performance on all 24 h can help mitigate the extremely high or low optimized coverage rate caused by a particular visitor distribution.

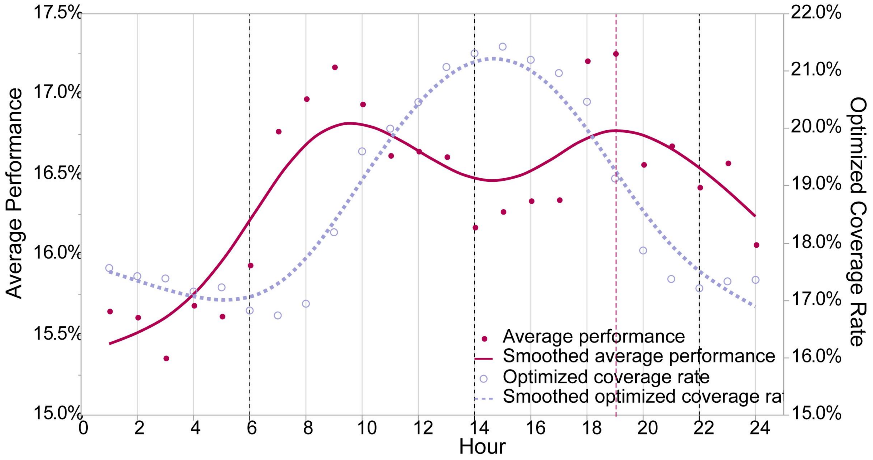


Figure 2.6: Comparison between the Average Performance of Optimized AEDs and the Optimized AED Coverage Rate across 24 h (k = 100). The average performance of each set of optimized AEDs ($N_{J'_{h'}}$) refers to the average performance applying each set of optimized AEDs to 24 sets of visitor distributions. The smoothed line is calculated based on a cubic spline with a lambda of 0.05.

Relocating Existing AEDs in Washington DC To make the results before and after optimization comparable, we calculated the existing AED coverage rate based on the average performance of existing AEDs on all visitor distributions of 24 hours. The unsatisfactory rate of 29.68% calls for an improvement in AED placement. Based on the result that the AED solution at 19:00 a higher performance than at any other time, we adopted the visitor distribution at 19:00 to optimize and relocate the existing AEDs. The AED coverage rate is improved to 73.99%. As shown in Figure 2.7, there is a great difference in AED spatial pattern before and after optimization. Before optimization, AEDs in the central area have a significantly higher density than those in peripheral areas. The range between the highest and lowest density reached 1124. After optimization, AEDs spread around, and the range between the highest and lowest densities narrowed down to 46.

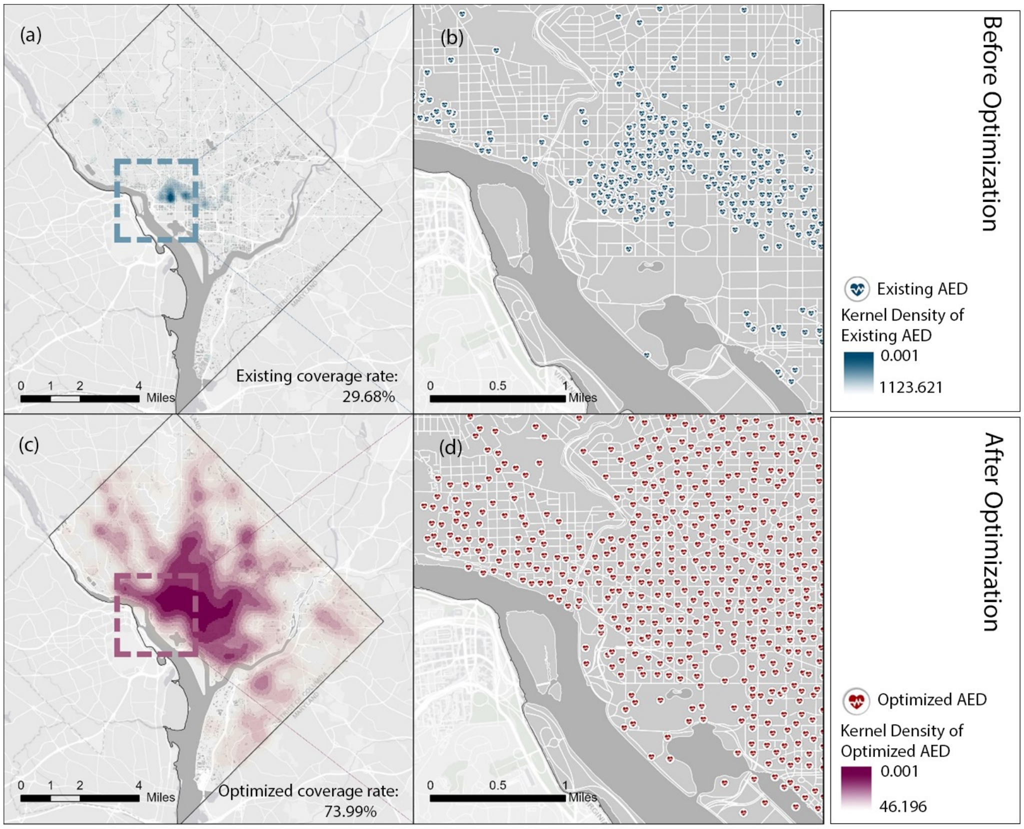


Figure 2.7: Comparison of the Distributions between Existing AEDs and Relocated AEDs in Washington DC. (b) and (d) shows the detailed locations of existing and relocated AEDs in a zoomed area. The kernel density maps (a,c) are computed based on the locations of existing and relocated AEDs.

Cost–Coverage Increment Curve In view of the similarity of the trends of optimized AED coverage rate between different ks, it is extrapolated that, regardless of the k value, the average performance of optimized AEDs will always reach the highest at 19:00. Therefore, we computed the cost–coverage increment curve based on the visitor distribution at 19:00. From the average performance curve (Figure 2.8), the average performance increases with more AEDs at first. However, the plateau is reached after 3600 AEDs. From the curve of improved average performance (cost–coverage increment curve), the gain of improved performance per additional 100 AEDs is decreasing monotonically. After the threshold of 3600, the performance will only increase marginally. The improved average performance per additional 100 AEDs (Δk) is shown as $\Delta N_{J'_{h'}}(h' = 19)$. Both smooth lines are calculated based on a cubic spline with a lambda of 0.05.

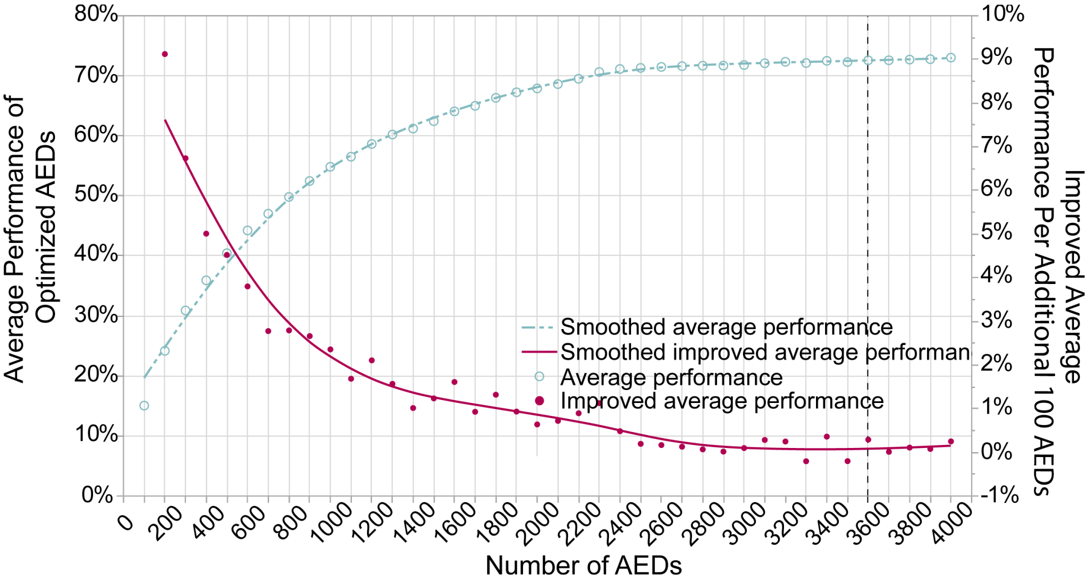


Figure 2.8: Cost—Coverage Increment Curve.

2.4 Discussion

Comparison of Current Planning and Optimized Planning of AEDs Previous spatial analytic methods, including the AHA guideline-based and population-based method, could lead to a paradoxical result with many AEDs placed in an area with a relatively low risk of OHCA. For example, the existing AED placement in Washington DC, which follows AHA guidelines, puts a disproportionate number of AEDs (1134 out of 2113) in downtown areas where the evaluated risk of potential OHCA incidence is relatively low. However, our study adopting the spatio-temporal optimization method demonstrated the improved AED coverage level. Such an increment results in shortened average time from collapse to treatment, which could positively improve the probability of survival. Compared with the existing plan in Washington DC, our proposed method deployed 298 out of 2113 AEDs downtown based on evaluating potential OHCA distribution. This deployment plan maximized the AED coverage rate across all areas at risk and balanced the supply and demand to a certain extent.

Designing an Inclusive Strategy Considering Potential OHCA Distributions across All Space—Time Ranges The optimized AED coverage rate (Figure 2.4) showed highly similar trends across 24 hours between different numbers of AED to be configured. This result implicated that the trends of the AED coverage rate were only affected by the spatial pattern of visitors at different times. This effect may result from the MIP model that always deploys AEDs proportionally concerning the distribution of the potential risk of OHCAs when maximizing the coverage level, regardless of the number of AEDs to be configured.

We also observed a reversed relationship between the trend of the optimized AED coverage rate and the trend of average performance across 24 hours (Figure 2.6). We attribute this effect to the correlation between the visitor clustering level and the optimized pattern of AED locations. For hours with more clustered visitors, the optimized AED locations on that hour will follow a high-density pattern. Although this can give a high coverage rate for that hour on the visitor distribution of the hour itself, the performance of the optimized AEDs on all hours' visitor distribution will be alleviated. A high-density set of AEDs is more challenging to couple with all kinds of visitor distributions. Selecting the solution from hours with low-density visitors could improve the overall coverage level across all spaces and times. These findings should initiate critical thinking for researchers and policymakers that the overall solution in a given geographic area needs to be designed to serve all the potential risk populations at any space and time, instead of simply maximizing a population shown in a limited space–time domain.

Finally, to make our solution of optimized AEDs more inclusive of potential OHCAs in different time periods, we applied the average performance evaluation and identified the solution with the highest performance based on the original purpose of maximizing efficiency. This efficiency maximization strategy is supported by utilitarianism to maximize the sum of benefits for all people (Rawls 2004). However, from another perspective of egalitarian theories, which encourages a policy on equalizing the relative level of accessibility between different groups (Van Wee and Geurs 2011), we might need to adjust the performance evaluation by selecting the solution with the smallest coverage gap. In further studies, careful consideration and exploration of the moral interpretation of utilitarianism and egalitarianism in adopting different policy measures should be conducted.

Limitations There are several limitations to our study. First, this study is contingent upon the reliability and accessibility of SafeGraph data. However, in real-world situations, these data may be limited by drawbacks such as the inadequate coverage of mobile device signals, the inaccurate counting of visit data, or only encompassing public or commercial buildings, which neglects 39.68% of out-of-hospital cardiac arrests (Bureau n.d.). In the future, the limitations of this study could be overcome by implementing cutting-edge communication technologies such as 5G and indoor positioning systems, as well as the procurement of more comprehensive datasets. Second, we optimistically assumed that all bystanders within proximity could locate and access AEDs and that all the AEDs could be appropriately used. However, this assumption first failed to consider the lack of guidance applications to locate AEDs. For access to AED, this assumption also ignores physical obstructions such as walls or multistory buildings that may impede access to AEDs or extend the access time, leading to an overestimation of the AED coverage level. This also ignores that using a Euclidean distance to represent an actual distance may also overestimate the rate. For the use of AEDs, this assumption assumes that everyone is trained and knows how to use AEDs correctly. Combining accessibility optimization with several practical measures, including placing AEDs in visible places, developing an app or online website for checking, and publicizing AED training, can help encourage the full use of AEDs. Including vertical space factors such as elevator speed in multistory buildings and elevation in a network analysis can improve the accuracy of predicting the AED access time, assisting in designing a more precise AED deployment plan. Thirdly, limiting the temporal analysis to monthly intervals with SafeGraph data hinders the ability to differentiate between holidays, weekdays, and weekends. A more refined time unit can enable a more detailed and adaptive plan for AED placement. Fourthly, the parameter setting in the optimization algorithm, such as the accessibility distance to AEDs and the walking circle size used to simulate visitor mobility, can impact the optimization result. Finally, the age distribution of OHCA demonstrates the peaks during infancy and after the age of 45 (Adabag et al. 2010). Taking age into the spatial and temporal optimization model may improve the accuracy of locating AEDs.

2.5 Conclusions

Using the OSTO method, our model provides an improved AED deployment solution under specific conditions as well as multiple options for a decision maker to quantify the trade-off between the budget and coverage level. Compared to the previous literature, where the instantaneous characteristics and critical time relevance of OHCA occurrences are often ignored, our results remind both researchers and policymakers that healthcare facilities' placement schemes should reasonably consider all populations across time and space. In addition, our cost-coverage increment analysis revealed that solely increasing the number of healthcare facilities may not guarantee an increasing coverage rate. The trade-off between the budget and coverage level need to be balanced in real scenarios. Furthermore, in addition to the AED deployment optimization planning shown in this study, our model is highly transferable to other scenarios for tackling different healthcare facilities' deployment tasks with certain constraints. Nevertheless, considering the current limitations illustrated in the previous section, this method can be improved, but it also takes us one step closer to applying this approach in a real scenario.

Chapter 3

SpaCE: a spatial counterfactual explainable deep learning model for predicting out-of-hospital cardiac arrest survival outcome

Zhang, Jielu, et al. "SpaCE: a spatial counterfactual explainable deep learning model for predicting out-of-hospital cardiac arrest survival outcome." International Journal of Geographical Information Science (2025): 1.32 (2025)

Reprinted here with permission of the publisher.

3.1 Introduction

The fundamental task in building predictive models in healthcare and public health is to accurately capture and quantify the relationships between risk factors and health outcomes, enabling models to guide effective interventions and policy decisions (N. D. Shah, Steyerberg, and Kent 2018). Given the complexity of this task, previous research has often focused on one of two primary goals: developing highly accurate predictive models or creating explainable methods to reveal patterns learned by these models.

In the case of predictive modeling in healthcare and public health, a key observation is that many chronic health conditions such as cardiovascular disease exhibit distinct spatial patterns (Nazia et al. 2022; Djukpen 2012; Mena et al. 2018; Sahar et al. 2021; Son et al. 2023). Consequently, integrating both health-related features and spatial effects is essential to building reliable predictive models. Traditionally, spatial statistical methods, such as Geographically Weighted Regression (GWR) (Brunsdon, S. Fotheringham, and Charlton 1998), have been instrumental in identifying spatial variations and elucidating underlying risk factors across geographic regions (Akindote et al. 2023; ŞENER and Türk 2021). In recent years, machine learning (ML) models, known for their superior predictive capabilities and ability to manage complex data patterns, have introduced a new approach to predicting health outcomes (Santangelo et al. 2023; Wiemken and Kelley 2020; Habehh and Gohel 2021). However, as ML approaches become more prevalent, some studies fail to account for spatial information, overlooking how geospatial variations influence health outcomes (D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016). Admittedly, incorporating spatial effects into traditional ML models is challenging (Mai, Janowicz, Y. Hu, et al. 2022). Treating geographical coordinates as continuous variables in a regression model often leads to model overfitting, as the model may memorize all the coordinates in the training dataset. Such a model might perform well on the training data but struggle to generalize to new, unseen data, limiting its practical utility. Thus, a pressing need exists to develop ML-based predictive models that effectively incorporate the spatial distribution of the data.

Once a predictive model is established, the next challenge lies in interpreting the patterns it has identified. Unfortunately, most ML models are highly complex, making direct interpretation difficult. Previous studies (Loh et al. 2022; Minh et al. 2022) have primarily employed post-hoc Explainable Artificial Intelligence (XAI) techniques to address this issue (Lundberg and S.-I. Lee 2017). Although these methods can generate global or local feature importance scores(D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016), they often fall short of providing actionable guidance, such as specific changes needed in variables to alter outcomes (e.g., transitioning from a high-risk to a low-risk state) or identifying optimal intervention targets. Such insights are crucial for both life-saving interventions and informed policy development.

Recognizing the limitations in current predictive models and XAI methods, we developed a novel **Spatial Counterfactual Explainable Deep Learning Method (SpaCE)**. This method comprises two core components: a Spatially Explicit Health Outcome Predictor (SEP) and a Prototype-Guided Counterfactual Explanation (PCE) algorithm, applied to predict survival outcomes for Out-of-Hospital Cardiac Arrest (OHCA). The SEP integrates spatial and health variables to improve health outcome prediction accuracy, while the PCE explains SEP predictions by identifying minimal adjustments needed to achieve an alternative outcome, generating feasible and actionable counterfactual examples. These examples offer strategic insights for intervention when an individual's predicted outcome differs from the observed result. We summarize our contribution as follows:

- **Spatially Explicit Health Outcome Predictor (SEP)**: We design SEP to integrate heterogeneous spatial and health variables effectively, enhancing predictive accuracy, demonstrated on Out-of-Hospital Cardiac Arrest (OHCA) survival outcome predictions.

- **Prototype-Guided Counterfactual Explanation (PCE)**: We develop the PCE algorithm to generate actionable counterfactual explanations by identifying minimal variable changes required to achieve alternative outcomes, thus providing practical intervention insights.

- **Insights for Public Health Policy and Intervention**: SpaCE delivers guidance at both geographic and individual levels, supporting data-driven strategies for reducing cardiac arrest mortality and offering personalized recommendations for intervention.

3.2 Related Work

ML predictive models using healthcare-related variables ML models are widely developed and applied to predict health outcomes using healthcare-related variables owing to their advantages in enhancing diagnostic speed and accuracy (Mubeen et al. 2017; Neefjes et al. 2017; Alaa et al. 2019; J.-m. Kwon et al. 2019). These ML predictive models are divided into two categories based on whether they account for spatial effects: non-spatial predictive models and spatially explicit predictive models. Non-spatial predictive ML models(D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016) include Decision Trees (DT), Random Forests (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN), among others. For instance, D. Zhang et al. (2022) utilized logistic regression, random forest, light gradient boosting, and extreme gradient boosting models to forecast in-hospital mortality among patients admitted for peripheral artery disease in the United States based on variables including patient characteristics, comorbidities, procedures, and hospital-related factors. In another study, M. Chen et al. (2017) proposed a convolutional neural network-based multimodal approach for effective prediction of chronic disease outbreaks in disease-frequent communities using structured and unstructured data from hospitals. Through experimentation on a dataset focused on regional chronic diseases such as cerebral infarction, they demonstrated that the prediction accuracy of their algorithm surpassed that of unimodal approaches for disease risk prediction using structured and unstructured data from hospitals. Choi et al. (2016) developed a recurrent neural network-based temporal model to predict all the diagnosis and medication categories for a subsequent visit using longitudinal time-stamped Electronic Health Record(EHR) data for 260K patients. These EHR variables include diagnosis codes, medication codes, or procedure codes. It is worth noting that, in general, none of these ML models applied in healthcare fields have incorporated spatial elements during their development(D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016), despite the evident spatial distribution patterns observed in diseases.

Recently, we have witnessed an increasing number of studies on spatially explicit machine learning model development (Janowicz et al. 2020; Mai, Y. Hu, et al. 2022). These models have been applied to various geospatial prediction tasks such as geographic question answering (Mai, Janowicz, Cai, et al. 2020), geospatial shape recognition (Mai, Jiang, et al. 2023; Siampou et al. 2024), POI type prediction (Yan et al. 2017; Mai, Janowicz, Yan, et al. 2020), species fine-grained recognition (Mac Aodha, Cole, and Perona 2019; Wu et al. 2024; Mai, Xuan, et al. 2023), satellite image classification (Mai, Lao, et al. 2023), trajectory generation (J. Rao et al. 2020; Klemmer et al. 2023), terrain feature detection (W. Li, Hsu, and M. Hu 2021), etc. However, most of these spatially explicit ML models do not offer model explainability except for a few. One such model is the Spatial Random Forest (Spatial RF)(Benito 2021; Wright and Ziegler 2015), which is suitable for spatial classification and spatial regression tasks. Another is the Spatial Regression Graph Convolutional Neural Network (SRGCNN) model(D. Zhu et al. 2022), which is limited to spatial regression tasks. Both of these methods show potential in healthcare outcome prediction tasks. SpatialRF facilitates the fitting of spatial regression models on spatial data using Random Forest. It achieves this by generating spatial predictors that enable the model to understand the spatial structure of the training data. This approach aims to minimize the spatial autocorrelation of model residuals and provide accurate variable importance scores. The method SRGCNN, a spatial regression graph convolutional neural network, incorporates both non-spatial and spatial effects by formalizing the spatial weights matrix W and cross-sectional data (X, y) as a fully connected graph within the GCNN framework. Experiments have demonstrated its capability to handle a broad range of geospatial data. As discussed in the paper, the authors suggest that this method could be extended into the field of public health. Considering the SRGCNN is a spatial regression model yet our task is a classification task, we leverage SpatialRF as one baseline in our study.

Explainable AI for healthcare-related outcome predictive model The XAI models used for explaining predictive models can also be divided into two categories based on whether they account for spatial effects: non-spatial XAI models and spatially explicit XAI models. Among non-spatial XAI studies for explaining predictive models in healthcare, Tseng et al. (2020) investigated the influence of intraoperative variables on acute kidney injury associated with cardiac surgery. Utilizing SHAP values (Lundberg and S.-I. Lee 2017), they identified several factors in hemodynamic variables as significant contributors to injury occurrence. Dindorf et al. (2020) applied LIME to an SVC model aimed at classifying post-hip surgery walking patterns in patients. Their analysis illuminated the pivotal role of specific movements as key determinants influencing the SVC's decision-making process. Similarly, Junfeng Peng et al. (2021) endeavored to provide a comprehensive interpretation of auxiliary diagnosis in hepatitis cases. They utilized Partial Dependence Plots(PDP) to refine the interpretation of liver disease dynamics.

To the best of our knowledge, the field of spatially explicit XAI (explainable artificial intelligence) models is still in its early stages. Currently, the only model in this category is GeoShapley (Z. Li 2024). GeoShapley is a post-hoc explanation method that treats geographic coordinates as separate feature columns in the predictive model. Its explanation approach leverages the Joint Shapley concept, treating two separate features as a combined feature for interpretation. Although GeoShapley is based on SHAP, it is not a counterfactual-based method. Although it can be used to identify feature importance, it cannot generate feasible and actionable counterfactual examples that offer strategic insights for interventions which are expected for most healthcare prediction problems.

Despite advancements in both non-spatial and spatially explicit XAI methods, these techniques primarily focus on assessing global or local feature importance(D. Zhang et al. 2022; M. Chen et al. 2017; Choi et al. 2016). In contrast, counterfactual-based explanation methods not only provide insights into feature importance but also offer targeted interventions for individual variables. For example, in the case of a patient with an "expired" status, a counterfactual approach might suggest that if the "witness status" changed from "not witnessed" to "witnessed by family member" and "AED use" from "not used" to "used," the patient might have survived. This approach offers actionable modifications for specific features. When applied within a geographic context, it can further inform region-specific intervention strategies.

Counterfactual Explanation for healthcare-related outcome predictive model Counterfactual explanations (Molnar 2020) represent a distinctive area within the XAI (Explainable Artificial Intelligence) domain. In general, counterfactual explanations can be divided into two main categories: one emphasizing causal inference (Prosperi et al. 2020; Smith and Ramamoorthy 2020; S. Morgan 2015), where counterfactuals are used to infer causal relationships, and the other focusing on counterfactuals without making causal assumptions (Dickerman and Hernán 2020; Mothilal, Sharma, and Tan 2020; Goyal et al. 2019; Poyiadzi et al. 2020). Our study aligns with the latter, non-causal approach.

For counterfactual explanations that involve causal inference, the objective is to mitigate confounding bias and estimate treatment effects from generated counterfactuals(L. Yao et al. 2021). Methods (Y. Zhu et al. 2024; L. Yao et al. 2021) in this category are often grouped by their approaches to controlling confounders, including (1) re-weighting, (2) stratification, (3) matching, (4) tree-based, and (5) representation-based. In the context of healthcare outcome prediction, several key studies illustrate these approaches. F. Li and F. Li (2019) introduced a propensity score weighting framework to estimate causal effects across multiple treatments, applying this to analyze racial disparities in medical expenditures among different racial groups using the 2009 Medical Expenditure Panel Survey (MEPS) data. Linden (2014) employed marginal mean weighting through stratification (MMWS) to measure pre and post-intervention differences in hospitalizations following a disease management program for congestive heart failure, the results underscore MMWS as a valuable alternative for evaluating healthcare interventions with observational data. Huang et al. (2023) proposed GPMatching, a novel matching method utilizing Gaussian process priors to define matching distance, which they applied to assess the effectiveness of early biological medication for Juvenile Idiopathic Arthritis using electronic medical record data. Similarly, G. Wang, J. Li, and Hopp (2016) utilized causal trees to identify patient groups with differing outcomes across healthcare providers, highlighting that patient-provider alignment based on outcome information can lead to improved expectations for patients. Despite the advancements in counterfactual causal inference for healthcare, these methods often rest on strict assumptions. Given that our work focuses on spatially explicit explanations, spatial continuity and correlation may violate the assumption that an individual's potential outcomes remain unaffected by others' treatment assignments, we leave these considerations for future work and will not explore them in the current study.

For non-causal counterfactual prediction and explanation, counterfactual examples can be generated through either model-agnostic or model-specific approaches. Model-agnostic methods do not consider the model's internal structure, instead relying solely on the model's inputs and outputs to generate counterfactual examples. Model-specific methods, on the other hand, leverage the model's internal structure to create these examples. Our study emphasizes the model-agnostic approach due to its adaptability and potential for transfer across tasks. Despite the demand, few counterfactual explanation methods in this category are tailored for healthcare-specific ML models. One general-purpose approach, DiCE (Diverse Counterfactual Explanations) (Mothilal, Sharma, and Tan 2020), can be applied broadly to any ML model, generating counterfactuals agnostic to model structure. However, DiCE does not account for the alignment of generated examples with real-world data, affecting the trustworthiness of the explanations. Developing a healthcare-specific counterfactual explanation method could substantially enhance intervention strategies. In our study, we include DiCE as a baseline for comparison.

3.3 Methodology

The SpaCE (Spatial Counterfactual Explanation) method (Figure 4.2) consists of two integral components: a Spatially-Explicit health outcome Predictor(SEP) to predict health outcomes from both health and spatial variables (longitude and latitude), and a Prototype-guided Counterfactual Explanation algorithm (PCE) to illuminate the decision-making process inherent in the SEP model.

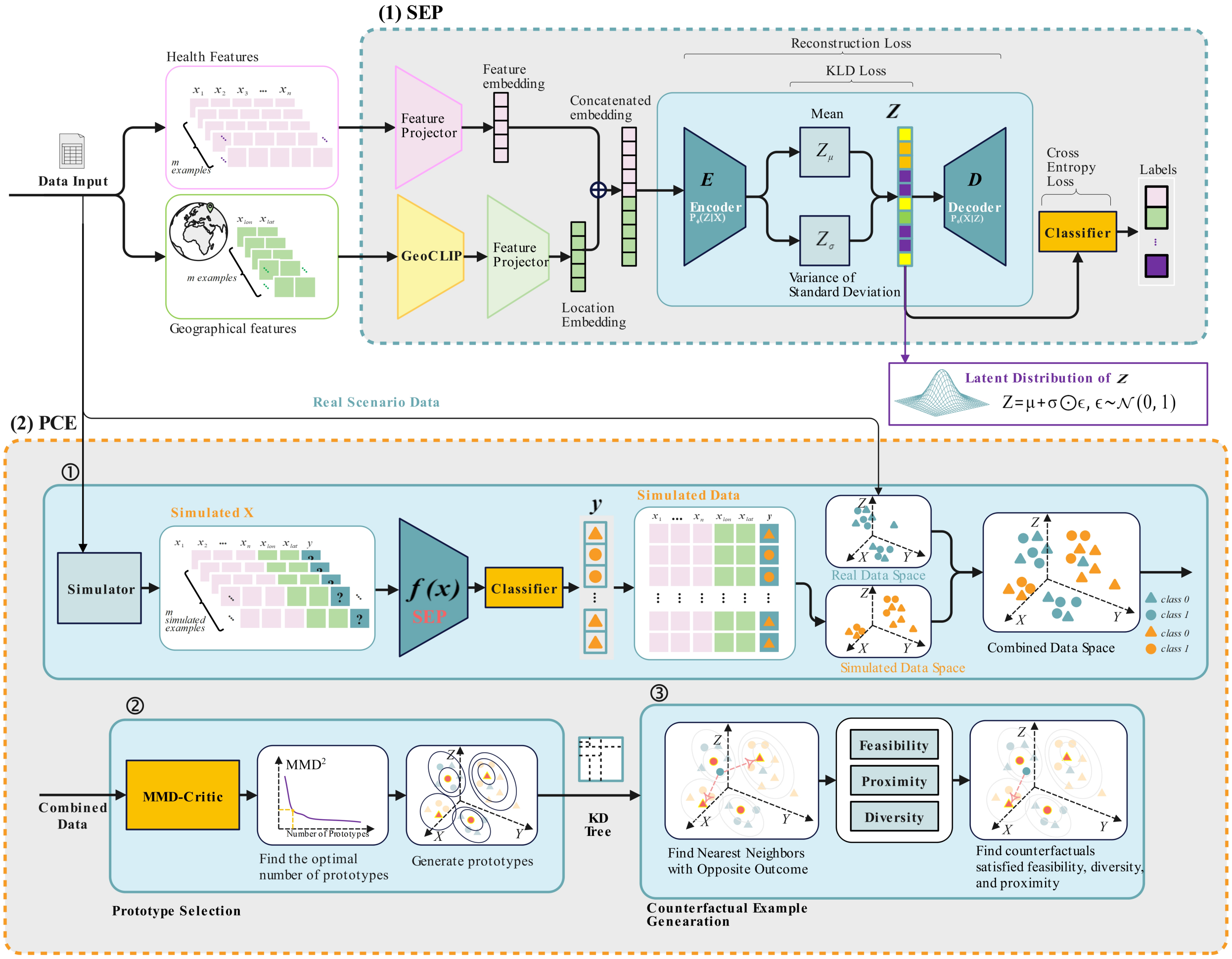


Figure 3.1: The SpaCE Framework. **Panel (1): Spatially Explicit Health Outcome Predictor (SEP)**. **Panel (2): Prototype-guided Counterfactual Explanation algorithm (PCE)**. This PCE algorithm-based pipeline includes three components: 1 **Simulated Data Generation**: generating simulated data with a trained SEP model and combining them with the actual scenario data. The triangle and square symbols symbolize binary outcomes within the dataset. 2 **Prototype Calculation**: employing the Maximum Mean Discrepancy-Critic method to aid in identifying a specific number of prototypes from the merged dataset. 3 **Counterfactual Example Generation**: generating counterfactual examples guided by prototypes while considering three criteria of feasibility, proximity, and diversity. This is the core of our method.

Spatially Explicit Health Outcome Predictor

The problem statement for health outcome prediction can be formulated as follows. Let $H = \{h_1, \ldots, h_n\}, \forall n \in [1, \ldots, N]$ denote the set of health variables, $G = \{G_{lon}, G_{lat}\}$ denote the set of longitude and latitude. The total variable set can be denoted as $X = \{(H, G)\}$. Given a dataset D that consists of M samples, for each sample m, $H_m \in R^N$, $G_m \in R^2, \forall m \in [1, \ldots, M]$. Our goal is to learn a predictive model $f : X \to y$, where y represents the outcome of the health status.

Making the outcome prediction model spatially explicit(Janowicz et al. 2020; Mai, Y. Hu, et al. 2022; W. Li, Hsu, and M. Hu 2021) is crucial given geospatial patterns observed in various diseases. A spatially explicit health outcome predictor (SEP) is developed to integrate both spatial and health variables, enhancing the performance of outcome predictions.

The architecture for SEP f is shown in Figure 4.2 Panel 1. First, we encode spatial and health variables separately. For spatial effect representation or so-called location encoding (Mac Aodha, Cole, and Perona 2019; Mai, Janowicz, Yan, et al. 2020; Mai, Janowicz, Y. Hu, et al. 2022; Mai, Lao, et al. 2023; Mai, Xuan, et al. 2023; Rußwurm et al. 2023; Wu et al. 2024), we adopt the location encoder from GeoCLIP (Vivanco Cepeda, Nayak, and M. Shah 2024). It is important to note that the GeoCLIP encoder is a point-based encoder rather than a spatial-structure-based encoder like a Graph Neural Network (GNN) (D. Zhu et al. 2022). While GNNs can quantify spatial effects, they implicitly learn spatial features during training, which requires a carefully designed loss term for effective optimization. Assessing the extent to which these models rely on spatial information for downstream tasks can be challenging due to the complexity of their learning processes. In contrast, point-based encoders like GeoCLIP can capture spatial information in a pre-trained manner, eliminating the need for training from scratch. This allows them to serve as plug-and-play components within various model architectures alongside other types of features, making them highly efficient and easily transferable to new applications. The GeoCLIP encoder, trained on a globally geotagged image dataset, generalizes well across tasks and regions. It is in an equal earth projection and applied positional encoding with random Fourier variables that transform two-dimensional coordinates ($G_m \in R^2, \forall m \in [1, \ldots, M]$) into a high-dimensional location embedding $R^D$where $D \gg 2$. For the extraction of health variables, we utilized a Multilayer Perceptron (MLP). After obtaining two embeddings for spatial and health variables, we concatenate these embeddings and feed them into a Variational Autoencoder (VAE) model to learn a fused distribution. VAE(Durk P Kingma et al. 2016) is an advanced version of autoencoder (AE). Unlike traditional AE that encodes input data as a single point, VAE treats the input as a distribution over potential values in the latent space. This probabilistic approach not only helps in disregarding outlier data but also in filtering the most significant variables, thereby improving model robustness. Moreover, the VAE is particularly adept at handling heterogeneous data. It learns to normalize different types of variables internally and weigh the importance of each variable based on its relevance to the task, thus effectively fusing two different types of variables.

The framework of the SEP model is detailed as follows:

- *Encoder Network*: The Encoder Network contains two parts. Firstly, we encode the spatial and health variables to embeddings separately. Secondly, the VAE encoder encodes the concatenated embedding to a latent variable z within a Gaussian distribution, characterized by its mean (μ) and standard deviation (σ).

- *Decoder Network*: The decoder network takes the sampled representation z as input and predicts the reconstruction of the original data. Through training, the decoder aims to generate output points that closely resemble the input.
- *Classification Network* The learned embedding z is fed into a classification layer (MLP) to predict the health outcome.

The total loss function combines reconstruction loss, Kullback-Leibler (KL) divergence, and cross-entropy loss, with the weights for each component controlled by hyperparameters α, β, and γ. The reconstruction loss ($\mathcal{L}_{MSE}$) quantifies the dissimilarity between the input data and the reconstructed data, quantified by Mean Squared Error. The KL divergence loss ($\mathcal{L}_{KL}$) measures the disparity between the inferred latent space distribution and a prior standard Gaussian distribution $\mathcal{N}(0,1)$. The cross-entropy loss ($\mathcal{L}_{CE}$) is used to measure the precision of multiclass health outcome prediction. The loss formula is given by:

$$\begin{aligned}\mathcal{L}_{SEP} =& \alpha \cdot \mathcal{L}_{MSE} + \beta \cdot \mathcal{L}_{KL} + \gamma \cdot \mathcal{L}_{CE} \\ =& \alpha \cdot \frac{1}{M} \sum_{m=1}^{M} (x_m - \hat{x}_m)^2 \\ &+ \beta \cdot \left(-\frac{1}{2} \sum_{z=1}^{Z} \left(1 + \log(\sigma_z^2) - \mu_z^2 - \sigma_z^2\right) \right) \\ &+ \gamma \cdot \left(-\frac{1}{M} \sum_{i=1}^{M} \sum_{c=1}^{C} y_{mc} \log(p_{mc}) \right)\end{aligned} \tag{3.1}$$

Where, M is the number of training samples, x_m represents the m^{th} feature of the input, and $\hat{x}_m$ represents the corresponding reconstructed output. μ_z and σ_z denote the mean and standard deviation of the latent variable z, respectively, and Z represents the dimensionality of the latent space. C is the number of classes. y_{mc} is a binary indicator of whether sample m belongs to class c (1 if true, 0 otherwise). p_{mc} is the predicted probability that sample m belongs to class c according to the model.

During training, we freeze the the pretrained location encoder module to preserve its generalized location embeddings.

Prototype-guided Counterfactual Explanation

After training the SEP model, our next aim is to understand the rationale behind its predictions. We achieve this by developing a novel XAI algorithm, named the Prototype-guided Counterfactual Explanation (PCE) algorithm, as illustrated in Panel 2 of Figure 4.2. In utilizing PCE to explain the SEP model, we initially employ the pre-trained SEP model to predict outcomes for simulated OHCA patients, integrating the learned distribution into the simulated data. Subsequently, we merge these simulated patients with actual OHCA patients to generate candidates for counterfactual examples. The PCE algorithm then performs counterfactual generation from these candidates.

To illustrate how PCE works, we will first explain its key idea and then outline the steps involved. The simplest way to generate counterfactual examples is to use real-world data with opposite outcomes. However, real-world data are often noisy and contain many outliers, making them less suitable for counterfactuals. Our pre-trained SEP model, which has learned the mapping from x variables to y outcomes, can predict and filter out noise for y outcomes. We first simulate x variable values from the real-world data and use the pre-trained SEP model to predict y outcomes, creating a "noise-free" dataset. We then combine these noise-free data with the original data to form integrated data. In PCE, we first generate prototypes from this combined dataset to further reduce the effect of outliers. Then, for each query instance, we identify the closest prototypes with the opposite outcome as counterfactuals.

Data Simulation We simulate or augment the data using a uniform sampling method from the original dataset, ensuring all simulated data points fall within the range of the original real-scenario values. The number of simulated samples matches the original data samples, based on the hypothesis that their importance is equal. There are three main reasons for using simulated data and a uniform sampling method: (1) Current real-scenario data we are working with is sparse, leveraging a pretrained prediction model to perform data augmentation helps to interpolate the data distribution, potentially smoothing the decision boundary and facilitating the identification of valid counterfactual examples. (2) We aim to address the skewness in the original data. For example, a variable such as the use of AED is highly skewed (over 90% of cases do not use AEDs). Directly using these data would generate counterfactuals that predominantly favor not to use AEDs, limiting the potential for improving survival rates through AED advocacy. (3) We aim to avoid limiting the generated counterfactual examples to specific geographical or data space. Since our collected data set only covered certain geographical areas, the use of uniform sampling ensures even geographical coverage for generality and equality.

Prototype Selection We calculate prototypes from both simulated and real-world data to serve as potential candidates for counterfactual examples. Identifying prototypes ensures that the generated counterfactuals are not outliers, providing viable intervention guidance. Here, we use the Maximum Mean Discrepancy-Critic (MMD-Critic) method (Molnar 2020) to identify prototypes. The MMD-Critic procedure consists of two phases:

- *Determining Prototype Count*: As the number of prototypes increases, the Squared MMD decreases, indicating a closer match between the prototype distribution and the overall dataset distribution. To determine the optimal prototype count, we employ elbow analysis (Humaira and Rasyidah 2020). This involves plotting the curve of Squared MMD against the increasing number of prototypes and identifying the elbow point where the rate of decrease sharply changes. We use the number at this point as the final number of prototypes.

- *Identifying the Prototypes*: Once the optimal number is determined, we set it as a parameter to the MMD-Critic model and get the results of prototype points.

The squared MMD is calculated (Eq. 5)

$$MMD^2 = \frac{1}{P^2}\sum_{i,j=1}^{P} k(p_i, p_j) - \frac{2}{PM}\sum_{i,j=1}^{P,M} k(p_i, x_j) + \frac{2}{M^2}\sum_{i,j=1}^{M} k(x_i, x_j) \tag{3.2}$$

In this equation, k is a kernel function, identifying similarities between two data points. The parameters P and M denote the counts of prototypes p and original data points x, respectively. The parameters i and j denote the random index which indicates any samples selected from prototypes p or data points x. The prototypes are derived from a mix of real data and simulated data from a pre-trained model. This blend ensures the generated counterfactuals closely mirror real-world examples.

Counterfactual Example Generation The next step involves generating counterfactual examples for each query instance. The formulation for counterfactual example generation is defined as follows. Let the set of query instances is defined as $U = \{q_1, q_2, \ldots, q_i, \ldots\}, \forall i \in [1, 2, \ldots, Q]$, where Q is the number of query instances. The set of prototype instances is defined as $S = \{p_1, p_2, \ldots, p_j, \ldots\}, \forall j \in [1, 2, \ldots, P]$, where P is the number of prototypes. The set of counterfactual examples for all query instances U is defined as $CF = \{C_{q_1}, C_{q_2}, \ldots, C_{q_i}, \ldots\}, \forall i \in [1, 2, \ldots, Q]$, where C_{q_i} represents the counterfactual example set for each query instance q_i. For each query instance q_i, the counterfactual example set is defined as $C_{q_i} = \{e_{1_i}, e_{2_i}, \ldots, e_{k_i}, \ldots\}, \forall k \in [1, 2, \ldots, E]$, where E is the number of counterfactual examples we wish to generate. The total number of counterfactual examples for the query set U is $R = QE$. Given the set of query instances U and the set of prototype instances S, the goal is to find the set of counterfactual instances CF for all query instances such that the outcomes of the counterfactual instances are opposite to the outcomes of the query instances and are realistic.

To ensure the generated counterfactuals are realistic, according to Mothilal, Sharma, and Tan (2020), we follow three principles: feasibility, proximity, and diversity. Feasibility ensures that the counterfactual examples generated are practical and applicable to individual circumstances. When identifying counterfactual examples, we consider certain variables to be immutable, such as race or gender. For each variable h in the list of immutable variables H_c, if h is a categorical variable, and the categorical variable value of h_{p_j} (one prototype example) equals that of h_{q_i} (one query instance/example), we include p_j in the set of counterfactual examples CF. If h is a continuous variable, and the continuous variable value of h_{p_j} falls within the same quantile as that of h_{q_i}, we add p_j to the set of counterfactual examples CF. Proximity refers to the notion that the generated counterfactual examples should closely resemble real data points. To achieve this, counterfactual examples are searched in a high-dimensional space facilitated by k-d Tree (k-dimensional Tree) (Ram and Sinha 2019) data structure. A k-d Tree is particularly useful for nearest neighbor searches, as it first divides the feature space into subspaces to quickly eliminate large portions of the search space. Our k-d tree is constructed from prototypes derived from both real-world scenarios and simulated data, ensuring that the generated counterfactual examples maintain proximity to the original data points. Diversity in this context refers to the flexibility to generate multiple counterfactual examples for each query instance as required. In real-world scenarios, some variables are not easy to change for certain people, such as income, due to various constraints. Therefore, providing multiple counterfactual examples allows for a broader exploration of possibilities to identify actionable interventions.

The process for finding E counterfactual examples for a specific query instance q_i is as follows. Note that query instances are limited to data points with an expired status, since we aim to observe the change from expired to survived status.

- Once we select prototypes from the combined dataset(simulated and real-scenario data), our initial step is to filter out the prototypes that belong to the opposite class of the query instance.
- For the filtered prototypes ($P^{'}$), we construct a k-d tree instance, denoted as T, using both the filtered prototypes $P^{'}$ and the query instance q_i.
- Initialize the number of neighbors, denoted as $N^{'}$, to be equal to the number of counterfactual examples, represented by E.
- Iterate while $N^{'} \leq len(P^{'})$, where $P^{'}$ is the set of filtered prototypes.
  - We query the k-d tree T to find the $N^{'}$ nearest neighbors of q_i, and denoted the preliminary generated counterfactual examples as $NN(q_i, N^{'})$.
  - We check if $NN(q_i, N^{'})$ satisfies all variable feasibility constraints and return the counterfactual examples met with constraints $NN.indices$.
  - If the number of generated counterfactual examples $NN.indices$ less than E, we increment the number of neighbors $N^{'}$ by 1.
  - We end this until the number of counterfactual examples $NN.indices$ generated equals to E or the number of neighbors $N^{'}$ equals to the number of filtered prototypes($P^{'}$).
- Return $NN(q_i, N^{'})$ as counterfactual example set C_{q_i}.

Details for how to find all counterfactual examples CF for all query instances U can be found in Algorithm 1.

As we generate counterfactual examples for each query instance, we observe that some variables have changed more times than others. To quantify the effect of different variables in the counterfactual explanation scenario, we define both individual variable importance and global variable importance.

Variable Importance

1. ***Individual Importance*** The individual variable importance is calculated for all variables with respect to a single example in the original dataset. We define four different individual importance scores based on each variable type, value, and the positive/negative impact on the outcome. The calculated individual importance is normalized by dividing by the number of counterfactual examples to ensure they sum up to 1.

Algorithm 1 Counterfactual Generation

```
Input: $query\_instance\_set(U), prototype\_set(S)$
Parameter: $num\_cfs(E), variables\_not\_vary(H_c)$
Output: The set of counterfactual examples $CF$
Initialize empty list $lst\_CF$
Initialize $num\_cfs(E)$
$prototypes\_select(P^{'}) \leftarrow S[S.class \neq query(q_i).class]$
for each $q_i$ in $U$ do
    $T \leftarrow concat[q_i, P^{'}]$
    $k\_dTree \leftarrow Build\_k\_d\_tree(T)$
    $num\_neighbors(N^{'}) \leftarrow num\_cfs(E)$
    while $N^{'} \leq len(P')$ do
        $NN \leftarrow k\_dTree.search(q_i, N^{'})$
        $NN.indices = constraints\_check(NN, H_c)$
        if $len(NN.indices) < num\_cfs(E)$ then
            $N^{'} \leftarrow N^{'} + 1$
        else
            $lst\_CF \leftarrow concat[lst\_CF, NN.indices]$
            break
        end if
    end while
    Can not find enough counterfactuals
end for
Create counterfactuals using indices from $lst\_CF$
```

- *Individual Importance for Categorical Variables*: For each categorical variable, the individual importance, denoted as I_v, is calculated as follows:

$$I_v = \frac{\sum_{e \in E} 1(v_e \neq v)}{|E|} \tag{3.3}$$

  where E represents the set of all counterfactual examples, v is a categorical variable in the query instance q, v_e is the value of v in a counterfactual example e, and $1(v_e \neq v)$ is an indicator function that returns 1 if the categorical value v_e does not match v.

- *Individual Importance for Numerical Variables*: For each numerical variable, the individual importance, denoted as I_n, is calculated as follows:

$$I_n = \frac{\sum_{e \in E} 1(|D(v) - D(v_e)| > 1)}{|E|} \tag{3.4}$$

where E is the set of all counterfactual examples, v is a numerical variable in the query instance q, v_e is the value of v in a counterfactual example e, $D(v)$ and $D(v_e)$ represent the quantile distributions of v in the query instance and counterfactual example respectively, and $1(|D(v) - D(v_e)| > 1)$ is an indicator function that returns 1 if the absolute difference in quantiles between the query instance and the counterfactual example is greater than one quantile, and 0 otherwise. $|E|$ is the total number of counterfactual examples.

- *Individual Importance Per Categorical variable Class:* For each categorical variable, the individual importance for each class is calculated based on the frequency of this class appearing in the counterfactual examples and is different from the class of query instance normalized by the total number of counterfactual examples, denoted as $I_{v,k}$:

  $$I_{v,k} = \frac{\sum_{e \in E} 1(v_e = k \wedge v \neq k)}{|E|} \tag{3.5}$$

  Let E be the set of all counterfactual examples, v the class of the categorical variable in the query instance q, v_e the class in a counterfactual example e, and k a possible class. The indicator function $1(v_e = k \wedge v \neq k)$ returns 1 if v_e equals k and differs from v, and 0 otherwise. $|E|$ denotes the total number of counterfactual examples.

- *Divergent Individual Importance for Numerical variable:* The divergent individual importance of numerical variables, denoted as I_v^+ and I_v^-, is calculated by comparing the value v in the query instance q against values v_e in counterfactual examples $e \in E$. For negative outcomes, I_v^+ measures the proportion of e where $v_e > v_q$, and for positive outcomes, where $v_e < v_q$. Conversely, I_v^- is measured under the opposite conditions. This approach assesses how variable changes impact outcomes differently. For a query instance with a negative outcome ($O_q =$ negative):

  $$I_v^+ = \frac{\sum_{e \in E} 1(v_e > v)}{|E|}, \quad I_v^- = \frac{\sum_{e \in E} 1(v_e < v)}{|E|} \tag{3.6}$$

  For a query instance with a positive outcome ($O_q =$ positive):

  $$I_v^+ = \frac{\sum_{e \in E} 1(v_e < v)}{|E|}, \quad I_v^- = \frac{\sum_{e \in E} 1(v_e > v)}{|E|} \tag{3.7}$$

  Here, 1 is an indicator function that returns 1 if the condition is true and 0 otherwise, and $|E|$ is the total number of counterfactual examples. The roles of I_v^+ and I_v^- reverse depending on whether the outcome O_q is positive or negative.

2. ***Global Importance (Georgia State-level in case study)***

   - Global importance, denoted as G_v is determined by aggregating individual importances for each variable across all examples(query instances) in the original dataset. This aggregation involves summing individual importances across all query instances and then averaging these sums over all instances.

Coefficient The coefficient for an ordinal variable is determined by graphically representing a smoothed linear relationship between the ordinal values of the variable (x-axis) and their corresponding global importance values (y-axis). The slope of this line is used as the coefficient for the ordinal variable. Conversely, the coefficient for a continuous variable is derived by estimating the net effect of the positive and negative individual importances.

Geographical Area-level Variable Importance and Coefficient In addition to investigating the impact of spatial variables on health outcomes, this study examines the spatial variation in the relationship between health variables and outcomes. To achieve this, we segment the study area according to specific levels of administrative boundaries. For each distinct zone, we evaluate both the global importance and the coefficient for every variable. In the case study, the geographical-area level is set at the county level.

Evaluation Metric

In our study, we utilize two sets of evaluation metrics. To assess the performance of SEP, we employ standard metrics, including Precision, Recall, F1 score, Area under the Receiver Operating Characteristic Curve(AUCROC), and Area Under the Precision-Recall curve(AUCPR), to compare it against other baseline models. For evaluating the quality of the generated counterfactual examples, we adopte proximity-based metrics, as recommended by Lucic et al. (2022), to quantify the deviation between the original dataset and the counterfactual examples generated. Our counterfactual explanation model is compared with a recently published model DiCE as the baseline (Mothilal, Sharma, and Tan 2020).

Here, we formulate three proximity-based metrics. The first metric, mean distance (d_{mean}), represents the average distance of the generated counterfactual examples from the original input. It is computed by first measuring the distance between a single query instance and each of its corresponding counterfactual examples, then calculating the average of these distances for all generated counterfactuals, and finally averaging these averages across all query instances. The second metric, mean relative distance ($d_{R_{mean}}$) is computed by calculating the ratio of the distance for each query instance from our method to the distance from the baseline method, and then averaging these ratios across all query instances in the dataset. A $d_{R_{mean}}$ value less than 1 indicates that our counterfactual methods generate examples that are, on average, closer to the original input than those produced by the baseline. Furthermore, we also quantified how often the distance from a query instance to its counterfactual examples generated by our method was smaller than the distance using the baseline method as ($\%_{closer}$). A $\%_{closer}$ greater than 0.5 indicates our model is better than the baseline.

3.4 Experiment Setup

Data Source

This study utilizes Georgia cardiac arrest incident data (Figure 4.3) recorded from 2019 to 2021. The data is accessed from the Georgia Department of Public Health, reported by the Emergency Medical Service (EMS), and collected by The National Emergency Medical Services Information System. The data collection process is approved by the university's Institutional Review Board. A total of 57,223 individual cardiac arrest patients and 24 variables are reported in this dataset(Table 4.1). The variable Patient Outcome at End of EMS Event is used as the targeted outcome variable. After data processing, a total of 5,385 cardiac arrest patients with geographic coordinates are filtered for analysis.

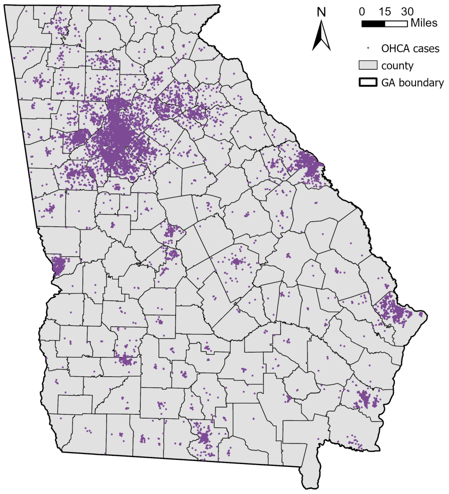


Figure 3.2: Out-of-Hospital Cardiac Arrest Distribution in Georgia from 2019 to 2021

Table 3.1: Variable Description of Georgia Out-of-Hospital Cardiac Arrest Dataset

Category	Variable
Incident Details	Incident Date, Scene Latitude, Scene Longitude, Incident Location Type
Patient Information	Patient Gender, Patient Race, Patient Age
Response Details	Response Beginning Vehicle Odometer, Response On Scene Vehicle Odometer, Incident Call Date Time, Incident Unit Arrived On Scene Date Time
Initial Assessment	Initial Patient Acuity
Cardiac Arrest Details	Cardiac Arrest Etiology, Cardiac Arrest Indications Resuscitation Attempted By EMS, Cardiac Arrest Witnessed By Whom
CPR and AED Details	CPR Provided or Not Prior to EMS Arrival, Who Provided CPR Prior to EMS Arrival, AED Used or Not Prior to EMS Arrival, Who Used AED Prior to EMS Arrival, Types of CPR Provided List
Patient Outcome	Patient Outcome at End of EMS Event, Medical Device Type of Shock, Outcome Emergency Department Disposition Description, Outcome Hospital Disposition Description

*AED: Automated External Defibrillator, CPR: Cardiopulmonary Resuscitation

Data Preprocessing

Data processing steps include: (1) The location type of cardiac arrest incidents is classified using the Tenth Revision, Clinical Modification (ICD-10-CM) provided by the CDC (Centers for Disease Control and Prevention). (2) Racial categorization is refined according to the standards of the Office of Management and Budget. (3) Outliers are identified and removed. (4) Data integrity is maintained by excluding variables with more than 35% missing values. Then we remove examples that have missing values. (5) For data analysis, nominal variables are processed using one-hot encoding, while ordinal variables utilize ordinal encoding. Additionally, all numerical variables are standardized.

Fifteen variables are finally used, including latitude, longitude, etiology, gender, race, age, initial patient acuity, who witnessed the cardiac arrest, CPR provided prior to EMS arrival, who provided CPR prior to EMS, AED use prior to EMS arrival, types of CPR provided, duration from call to EMS arrival, location type, and patient outcome at the end of the EMS event. The patient outcome at the end of the EMS event is the targeted outcome variable.

SpaCE Model Training and Counterfactual Explanation

Spatially Explicit Health Outcome Predictor Training After data pre-processing, our analysis includes 5,385 patients with OHCA, each with OHCA coordinates. The survival outcomes for these patients are classified into three categories: expired, ongoing resuscitation, and survived. We train our SEP model using 80% of the data and test it on the remaining 20%. It is important to mention that we use three labels when training the SEP model, but only two labels (expired and survived) are used when conducting the counterfactual explanation. We exclude ongoing resuscitation from our analysis as it represents an intermediate state, which complicates the generation of valid counterfactuals. Instead, we focus on generating examples where individuals transition directly from expired to survived.

Prototype-Guided Counterfactual Explanation Estimation After obtaining a well-trained SEP model, we utilized it to generate pseudo data (5,385 examples) and combined it with real-scenario data (5,385 examples) to create a dataset with 10,770 examples. We filter the data and obtain 8,248 final examples that have either expired or survived labels. Using the PCE algorithm, we identify in total of 400 prototypes as counterfactual candidates. The query instances (4,167 examples) are patients who have expired within the combined dataset. We generate five counterfactual examples from prototypes for each query instance. We chose five as it ensures computational efficiency and provides sufficient diversity in options to effectively guide interventions for changing the survival status of a cardiac arrest patient. Based on the generated counterfactual examples, we calculate the global importance and global coefficient for each variable. Additionally, we create the dependence plot with the x-axis representing the variable class and the y-axis showing the variable class importance. We estimate the global importance and coefficients at the county level to assess the spatial variation in the effects and direction of each variable on health outcomes.

3.5 Result

Comparison between SEP and traditional machine learning models

To illustrate the effectiveness of our developed predictive model, SEP, in capturing spatial effects, we compared its performance with the state-of-the-art spatially explicit model, SpatialRF, in predicting OHCA survival outcomes. SpatialRF (Benito 2021; Wright and Ziegler 2015) is a recently developed machine learning approach that incorporates spatial information by calculating a distance matrix based on spatial coordinates and training data, aiming to minimize spatial autocorrelation in model residuals. Our SEP model achieved an AUC-ROC score of 0.682, outperforming SpatialRF's score of 0.619 (Table 3.2). To further emphasize the importance of spatial variables in outcome prediction, we also evaluated our model against traditional machine learning models that do not include spatial variables. These models included Random Forest (RF), Gradient Boosted Decision Trees (GBDT), LightGBM, and Support Vector Classifier (SVC). Our SEP model consistently demonstrated superior performance across various metrics, as shown in Table 3.2. Notably, it achieved the highest AUC-ROC score of 0.682, surpassing the closest traditional model, GBDT, which scored 0.622.

Comparison between PCE and DiCE

We compare our method, PCE, with the gradient-based DiCE method, which is a recently published counterfactual explanation framework(Mothilal, Sharma, and Tan 2020). This comparative analysis is conducted using three distinct evaluation metrics stated in Section **??** across three different types of distance (Euclidean, Manhattan, and Cosine Similarity). As indicated in Table 3.3, the counterfactual examples generated by our method, PCE, are demonstrably closer to actual OHCA scenarios than those produced by DiCE. Specifically, the mean distance d_{mean} calculated for all three distance types using our method consistently yields smaller values compared to those from DiCE. Furthermore, the ratio $d_{R_{mean}}$, representing the distance from our method relative to that from DiCE, consistently stays below 1 for all distance types, indicating a consistent outperformance by our method. Additionally, the percentage $d_{R_{mean}}$, exceeds 50% for both the Euclidean and Manhattan distances, suggesting that our method surpasses DiCE in these metrics.

Figure 3.3 shows histograms for all three types of distance whose x-axes represent the calculated mean distance between each query instance and its corresponding counterfactual examples, and y-axes represent the count of query instances within certain distance bins. A closer inspection of this distribution histogram reveals that the distance between the counterfactual examples generated from PCE and the original real-scenario dataset is substantially smaller compared with those from DiCE.

Table 3.2: Evaluation result for OHCA survival outcome prediction across multiple models

Metrics	**Models**						
	RandomForest	GBDT	LightGBM	XGBoost	SVC	SpatialRF	**SEP**
Precision	0.384	0.415	0.417	0.400	0.405	-	**0.440**
Recall	0.501	0.583	0.579	0.570	0.576	-	**0.625**
F1	0.435	0.485	0.485	0.470	0.476	-	**0.517**
AUCROC	0.571	0.622	0.616	0.599	0.609	0.619	**0.682**
AUCPR	0.386	0.444	0.447	0.418	0.419	-	**0.469**

* The **bold** numbers indicate the highest performance per metric, the numbers underlined represent the second-highest performance. SEP represents the Spatially Explicit Health Outcome Prediction model. The hyphen (-) means the metric calculation in SpatialRF is not provided.

Table 3.3: Evaluation Comparison between PCE and DiCE

Metric	**Method**	**Euclidean**	**Manhattan**	**Cosine**
d_{mean}	DiCE	2.48	7.72	0.26
	PCE	**1.24**	**2.80**	**0.17**
$d_{R_{mean}}$	PCE/DiCE	0.54	0.42	0.79
$\%_{closer}$	PCE<DiCE	0.62	0.62	0.42

*The **bold** number shows the best performance. The smaller the distance metric, the better.

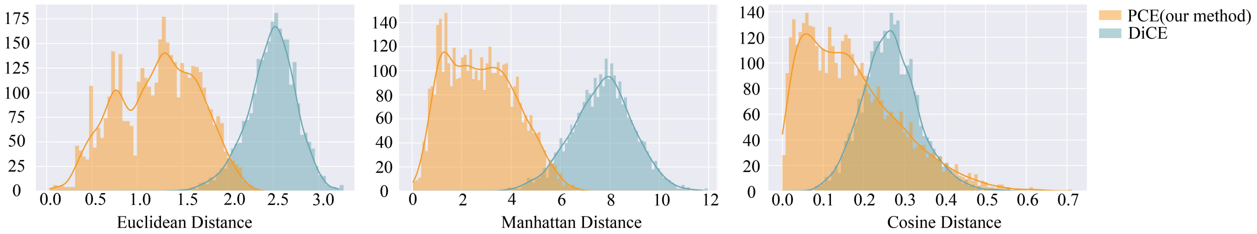


Figure 3.3: Distance distribution map between PCE and DiCE

Results from SpaCE on OHCA Survival Outcome Prediction

Latent Embedding Visualization To investigate the potential relationship between the learned latent embedding Z and geographical locations, we employ Uniform Manifold Approximation and Projection (UMAP)(McInnes, Healy, and Melville 2018) to project the high-dimensional embedding into a two-dimension space, subsequently comparing it with the geographical distribution of the data points. Given the substantial number of data points, we adopt a sampling approach. Specifically, we sample 10 data points from each of the top 10 counties with the highest number of points. Then, we plot the learned latent embeddings of these 100 points colored by counties and compare this embedding space with real scenario spatial space. Figure 3.4 shows the visualization results. From panel(a), we observe two distinct groups. The larger group on the left shows a clear correlation where data points close in the embedding space are also proximal in spatial space, and vice versa. The smaller group on the right displays some discrepancies, likely due to the influence of other health-related variables on the embedding space. Overall, there is general alignment between the embedding and spatial coordinates, indicating effective integration of spatial information into the latent space.

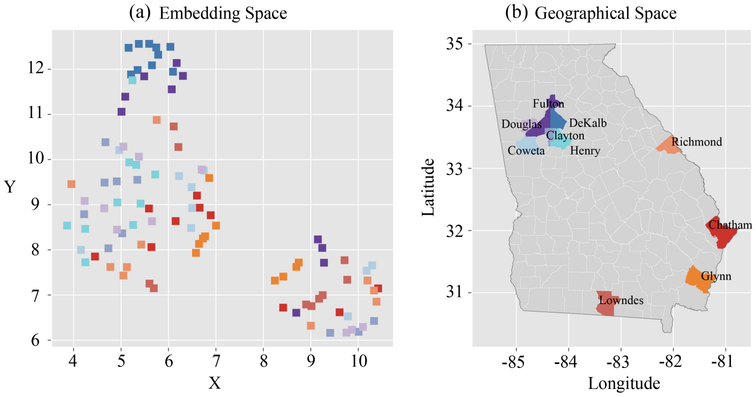


Figure 3.4: Comparison between learned embedding space and real scenario spatial space.

Generated Counterfactual Examples with Feasibility and Diversity To assess the distinction between feasible and unfeasible counterfactual examples, we conducted a feasibility scenario simulation, the results are displayed in Table 3.4. We maintained certain demographic variables, such as gender and race, as constants; instead, we focused on modifying other factors. In scenarios deemed unfeasible, the counterfactual examples erroneously altered immutable characteristics, changing race from 6 (white) to 5 (other race) and 3 (Hispanic or Latin), and gender from 1 (male) to 0 (female). Conversely, the feasible counterfactual examples retained these variables and concentrated on altering other variables. To show the diversity of our method, we can select any reasonable number of generated counterfactual examples to provide various options for OHCA patients.

Individual Variable Importance and Coefficient To illustrate an individual-level result, we randomly selected an OHCA patient from our dataset, generated five counterfactual examples (Table 3.5), and calculated the importance of each variable (Figure 3.5(a)) and coefficients (Figure 3.5(b)) for this patient. Considering only ordinal or continuous variables can be used to calculate coefficients, we have excluded the nominal variables, location type, and etiology in Figure 3.5(b). Analysis of the importance table revealed that variables such as geographic location (longitude and latitude), EMS response time(duration), and the person administering CPR(who used CPR) significantly influenced the shift from a non-survival to a survival outcome. From the coefficients depicted in Figure 3.5(b), we observed that duration is highly negatively correlated with survival outcomes, indicating that shorter EMS response times are crucial. Specifically shown in Table 3.5, four out of the five counterfactual examples suggested reducing EMS

Table 3.4: Counterfactual examples generated from unfeasible and feasible scenario

Variable	**query instance**	**unfeasible counterfactuals**		**feasible counterfactuals**	
		cf1	cf2	cf1	cf2
latitude	33.40	34.32	33.66	33.33	30.85
longitude	-84.60	-85.09	-84.41	-82.54	-83.34
Types of CPR	1.0	-	3.0	2.0	2.0
Etiology	6.0	3.0	-	-	-
Gender	**1.0**	**0.0**	**0.0**	-	-
Race	**6.0**	**5.0**	**3.0**	-	-
Location type	5.0	3.0	-	-	2.0
Patient initial acuity	2.0	0.0	-	-	-
Witnessed by whom	0.0	-	1.0	2.0	-
CPR used or not	1.0	-	-	-	-
Who use CPR	3.0	-	1.0	1.0	-
AED used or not	0.0	-	-	-	-
Age	0.0	3.0	1.0	2.0	1.0
EMS response time (Duration)	3.0	-	0.0	-	2.0
Survival outcome	0.0	1.0	1.0	1.0	1.0

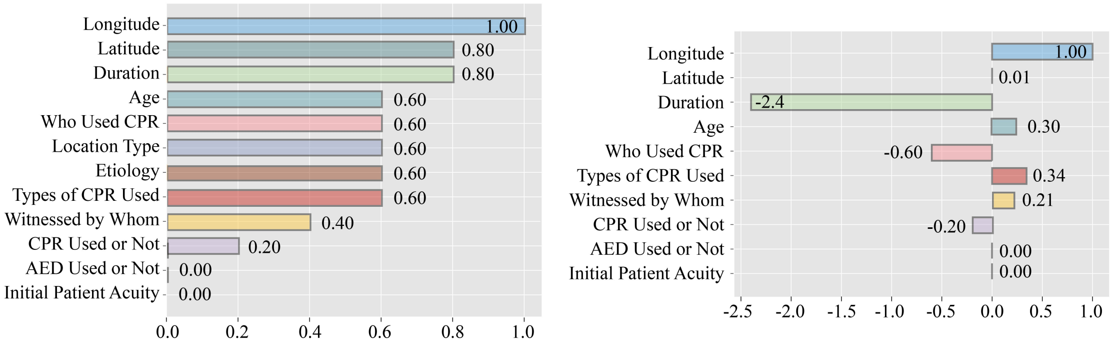
*The variables underlined means these variables should be kept constant and unchanged. Hyphen (−) means the values in generated counterfactual examples are unchanged compared with the value in the query instance. Survival outcome (0: expired, 1: survived). cfs: counterfactual examples

response time from category 3 (more than 13 minutes) to category 2 (less than 13 minutes), underscoring the importance of timely EMS intervention. Furthermore, most counterfactual examples recommended changing the CPR provider(who use CPR) from a professional (category 3) to a family member (category 1) or a community access responder (category 2), likely because these responders can reach OHCA patients faster than EMS personnel in such scenarios. In this instance, the variables "AED usage" and "initial patient acuity" have zero importance since our model tends to maximize the individual's survival chance by keeping these two variables constant when generating counterfactual examples, highlighting our method's ability to develop personalized intervention strategies tailored to individual cases.

Table 3.5: Generated counterfactual examples for a single OHCA patient

Variable	**query instance**	**Counterfactuals (cf)**				
		cf1	cf2	cf3	cf4	cf5
latitude	33.40	33.33	30.85	33.98	30.77	33.47
longitude	-84.60	-82.54	-83.34	-83.71	-83.76	-84.18
Types of CPR	1.0	2.0	2.0	-	-	4.0
Etiology	6.0	-	-	2.0	5.0	2.0
Gender	1.0	-	-	-	-	-
Race	6.0	-	-	-	-	-
Location type	5.0	-	2.0	4.0	7.0	-
Patient initial acuity	2.0	-	-	-	-	-
Witnessed by whom	0.0	2.0	-	-	3.0	-
CPR used or not	1.0	-	-	-	0.0	-
Who use CPR	3.0	1.0	-	-	2.0	1.0
AED used or not	0.0	-	-	-	-	-
Age	0.0	2.0	1.0	1.0	-	-
EMS response time (Duration)	3.0	-	2.0	2.0	2.0	2.0
Survival outcome	0.0	1.0	1.0	1.0	1.0	1.0

*Hyphen (−) means the values in generated counterfactual examples are unchanged compared with the value in the query instance. Survival outcome (0: expired, 1: survived). cfs: counterfactual examples



(a) Individual Importance of Survival

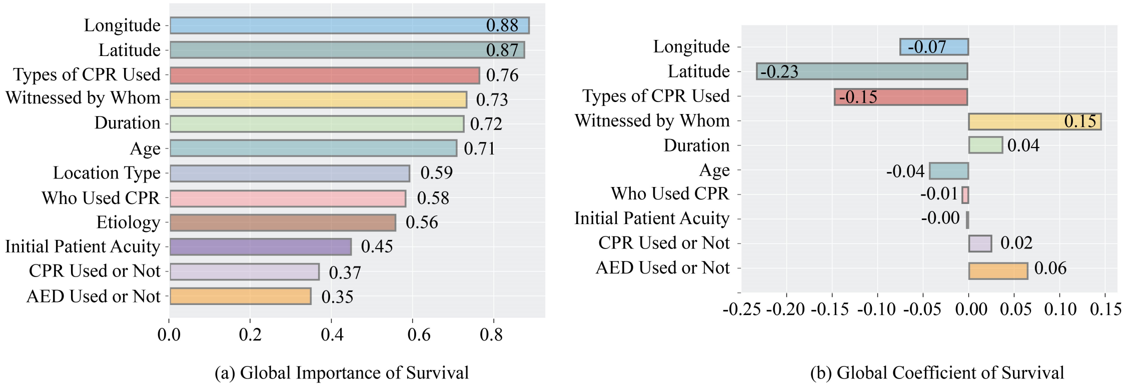
(b) Individual Coefficient of Survival

Figure 3.5: Individual importance and coefficient for a single OHCA patient in predicting survival outcome. Considering only ordinal or continuous variables can be used to calculate coefficients, we have excluded the nominal variables, location type, and etiology in panel (b).

Global(State-level) Variable Importance and Coefficient The global variable importance (Figure 3.6(a)) and coefficient (Figure 3.6(b)) are calculated for all examples in Georgia state. We also plot a dependence plot (Figure 3.7) based on the calculated global importance for all variables. This plot shows how the survival probability (importance of survival) changes when the variable class value changes. Given that the query instances used in PCE pertain to expired OHCA patients, and our objective is to generate counterfactual examples resulting in survived status, the variable importance and dependence plot serve as estimations of importance to survival. As we can see from Figure 3.6(a) panel, the geographical variable of longitude and latitude exhibit high significance in improving survival probability. Apart from these two spatial variables, the top six health variables include types of CPR used, OHCA witnessed by whom, the duration, age, location type of OHCA incident, and who use CPR. We also look into the efficient and dependence plot of these six variables, as shown in Figure 3.6(b) and Figure 3.7.

- Variables with positive coefficient: For the variable OHCA witnessed by whom, patients witnessed by professionals or family members are more likely to survive than those not witnessed or witnessed by bystanders. This aligns with our common sense. Regarding emergency response durations, there is a positive correlation between survival rates and the promptness of the response. However, a closer inspection reveals a significant decline in survival rates when the response time exceeds 13 minutes. This suggests that factors other than EMS response time may play a more critical role in determining survival outcomes for instances that are less than 13 minutes. Additionally, incidents occurring in sports areas and service areas.

- Variables with negative coefficient: There is a negative correlation between both latitude and longitude and survival outcomes, indicating that survival chances increase the further south and west the location. The impact of location on survival is complex and interacts with socioeconomic factors, social determinants of health, and factors related to OHCA incidents. An exploration of the geospatial distribution of the original data revealed that the northeastern area utilized CPR less frequently than the southwestern area, which may contribute to the negative correlation. Additionally, the choice of CPR types employed is negatively correlated with survival rates. The variable types of CPR reflect the frequency and variety of devices and techniques used during resuscitation efforts. For instance, in a case where three types of CPR are utilized on a patient, this might include compression with an external plunger-type device, ventilation with a bag valve mask, and ventilation with a pocket mask. These findings suggest that employing simpler, yet effective CPR techniques may enhance survival outcomes. Age also plays a crucial role, as age increases, survival chances decrease. Moreover, patients who receive CPR from family members or professionals before the arrival of EMS have a higher survival chance compared to those assisted by bystanders or community access responders.



(a) Global Importance of Survival

(b) Global Coefficient of Survival

Figure 3.6: Global(State-level) importances and coefficients for all examples in predicting OHCA survival outcome. Considering only ordinal or continuous variables can be used to calculate coefficients, we have excluded the nominal variables, location type, and etiology in panel (b).

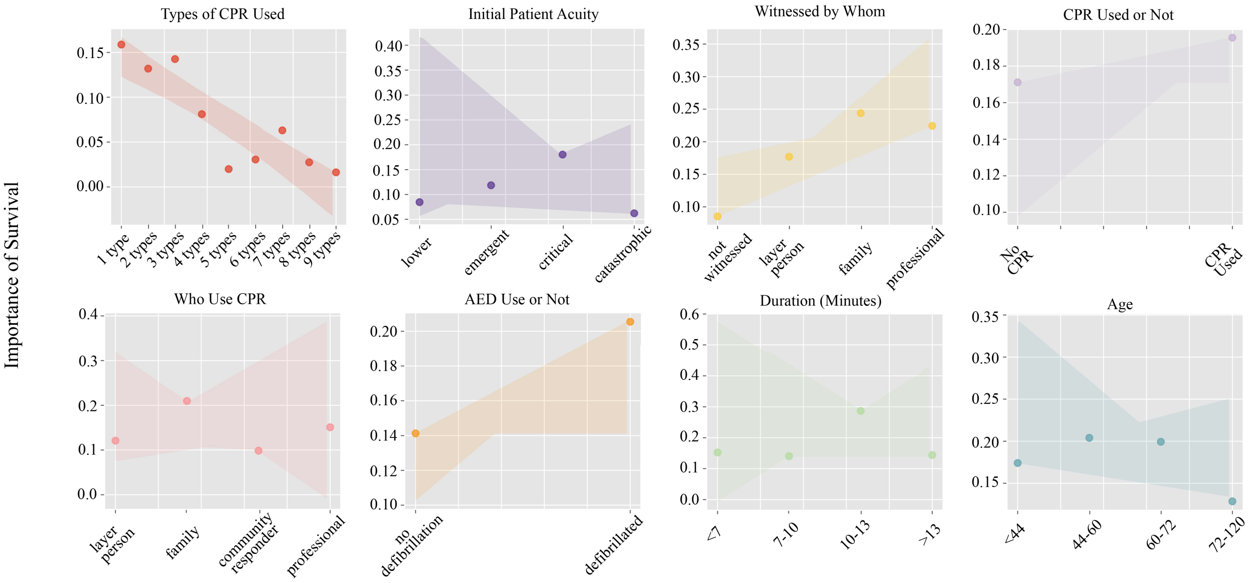


Figure 3.7: Global(state-level) dependence plot in predicting OHCA survival outcome. The y-axis represents variable class importance in enhancing survival status. The shaded region represents changing trend. For nominal variables without a specific order, such as etiology and location type, we do not display the trend here.

County-level Explanation To assess the spatial variation in the impact of health-related variables on survival outcomes across different counties in Georgia, we conduct a county-level analysis and visualize the results through maps of variable importance (Figure 3.8) and coefficient (Figure 3.9). The variable importance and coefficient are broken into 5 classes in the maps using quantile method. We organize our discussion of results according to the spatial orientation, focusing on the northern, southern, and entire state perspectives. The boundary for north and south Georgia in the map is perceptually draw based on the fall line.

In North Georgia, the variable importance map(Figure 3.8) generally indicates low significance for most variables in the Atlanta metro area. However, this area is critical due to its high incidence of OHCA. Consequently, for the Atlanta metro area, we have shifted our focus to other variables that show greater importance: the duration from EMS dispatch to arrival and patient age. As illustrated in Figure 3.9, these two variables exhibit a consistent negative correlation with survival outcomes in the Atlanta area. The coefficient map for EMS response time indicates that patients with shorter response times have a higher survival probability in the Atlanta metro area. Similarly, the coefficient map for age reveals that younger patients generally have a better chance of survival.

In South Georgia, the importance map (Figure 3.8) indicates that most variables hold greater significance compared to those in North Georgia. From the coefficient map (Figure 3.9), it is evident that certain variables display a strong correlation (indicated by deep color intensity) with survival outcomes. These variables include initial patient acuity, witnessed by whom, who used CPR, and age. For initial patient acuity, a significant negative correlation is observed, suggesting that higher acuity levels correspond to lower survival chances, aligning with intuitive expectations. The variable witnessed by whom displays a positive correlation with survival outcomes, indicating that OHCA patients witnessed by healthcare professionals are more likely to survive than those witnessed by bystanders. Further analysis combining the witnessed by whom coefficient map with the AED use map pinpoints counties such as Telfair, Pulaski, Coffee, and Appling where both witnessing and AED use are positively correlated with survival outcomes. This insight supports the advocacy for targeted AED training initiatives in these specific areas. Additionally, when the witnessed by whom map is combined with the CPR use or not map, a positive correlation between CPR use and witnessed by whom with survival outcomes is also observed in southern counties, including Atkinson, Coffee, Telfair, and Irwin. This suggests that promoting CPR training initiatives in these counties could increase the availability of trained responders and thus improve survival rates. For the variable concerning who performed CPR, almost all counties show a negative correlation with survival rates, indicating that patients assisted by professionals have lower survival chances compared to those helped by family members, community responders, or bystanders. This may be due to the latter groups' ability to reach patients more promptly than professionals.

Across the entire state of Georgia, the age coefficient map reveals a significant negative correlation with survival outcomes.

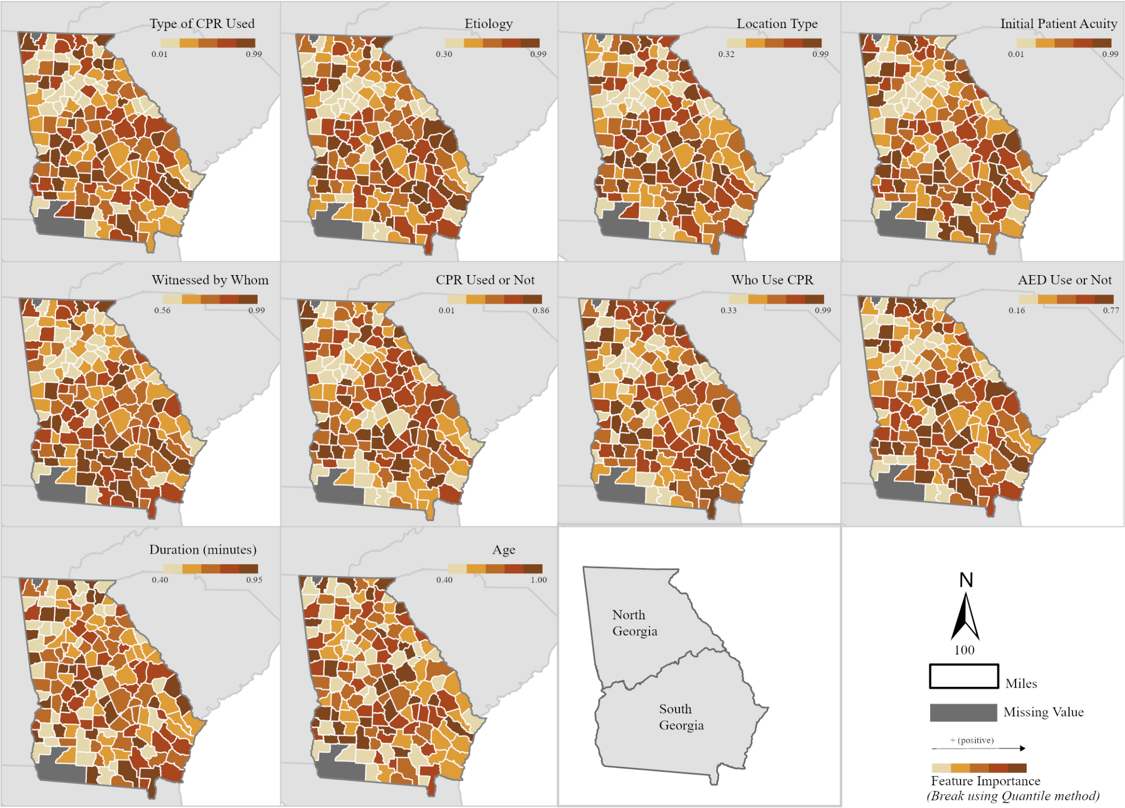


Figure 3.8: Variable Importance at County level across Georgia

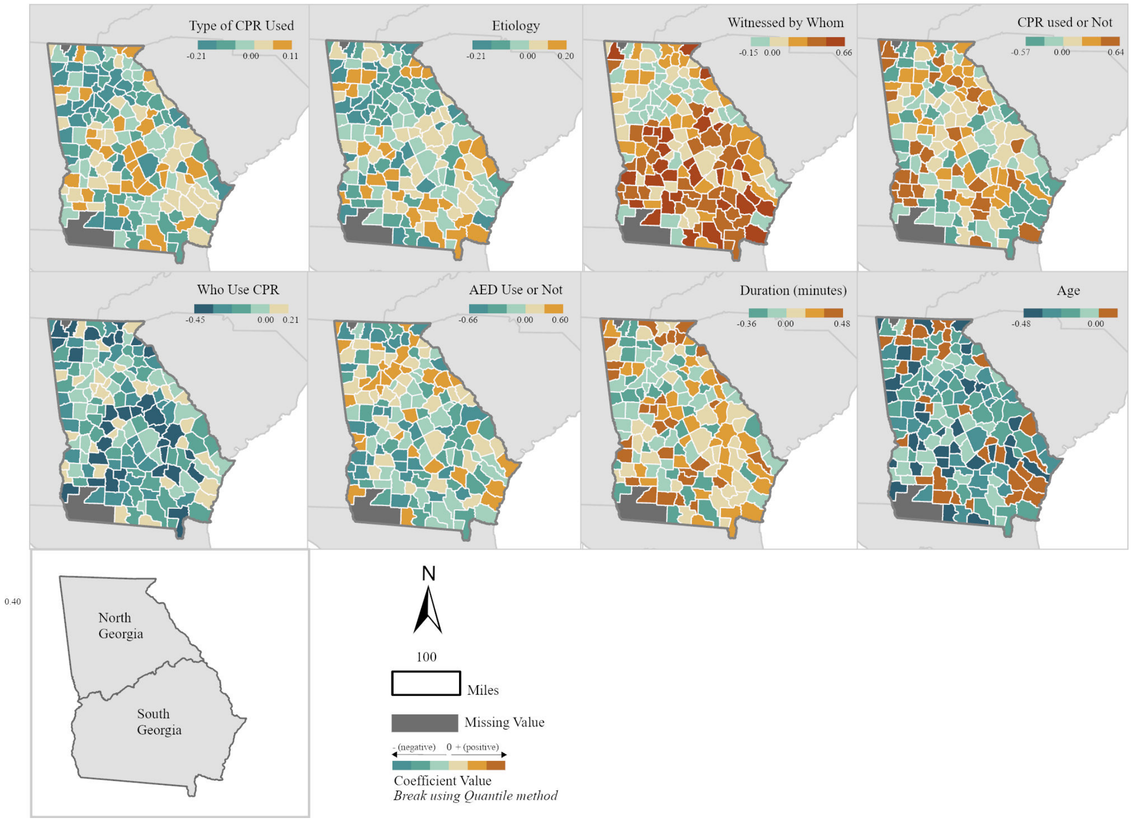


Figure 3.9: Variable Coefficient at County level across Georgia

3.6 Discussion

Advantages of our predictive model over previous models Our research on OHCA survival outcome prediction demonstrates that the SEP model we developed outperforms the state-of-the-art model, SpatialRF. The SEP model, which integrates spatial and health variables, achieved a 10.2% improvement in OHCA survival prediction accuracy compared to SpatialRF, indicating a stronger capability in capturing spatial effects. While SpatialRF can simulate spatial structure using a distance matrix and account for spatial autocorrelation, it may fall short in capturing more complex spatial relationships. In contrast, our model's location encoders, which leverage Fourier transformations, effectively capture subtle spatial patterns that traditional models might miss. Additionally, it is noted that SpatialRF, depending on its implementation, can be computationally demanding due to spatial adjustments and distance matrix calculations, particularly when working with large datasets.

Moreover, our SEP model showed a 9.7% improvement in OHCA survival outcome prediction compared to traditional machine learning models such as RF, GBDT, LightGBM, and SVC, which do not consider spatial variables. This indicates that spatial variables play a critical role in survival outcome prediction, and our SEP model successfully integrates spatial effect. Second, the resulting map (Figure 3.4) showed a high similarity between cardiac arrest cases compared at the county level in a two-dimensional embedding space plot and a geographical space plot. This further confirms that our model effectively learned and incorporated spatial features.

The validity and advantages in our explainability model By comparing the PCE with the DiCE to explain the prediction of the SEP model, we found that the counterfactual examples generated by PCE are more closely aligned with the real scenario data than DiCE. This indicates that our model has greater potential to provide practical suggestions. Further analysis of the PCE explanation results on the SEP model in OHCA survival outcome prediction, at both individual and geographical levels, supports this conclusion. At the geographical level, the explanation results for variable importance and coefficient maps indicate that EMS response time needs improvement in North Georgia, particularly around the Atlanta area. Additionally, these maps highlight that promoting AED and CPR training initiatives in South Georgia would be beneficial for improving OHCA survival rates. At the individual level, the generated counterfactual examples in the OHCA survival outcome prediction task enable practical intervention suggestions for specific individuals. These suggestions are reasonable and practical, as they maintain certain self-attribute variables, such as race and gender, unchanged, and make the generated counterfactual examples as close to the real scenario examples as possible. Overall, these results offer targeted plans for improving OHCA survival rates at both geographical and individual levels, demonstrating the effectiveness of our explanations.

While SHAP, LIME, and GeoShapley methods can rank feature importance at both global and local levels, they do not provide actionable guidance on how to adjust specific features to change the outcome for an individual case. In contrast, counterfactual-based methods not only rank feature importance but also suggest specific feature modifications that could lead to a different outcome, aligning directly with the goal of improving health outcomes through targeted interventions. Our counterfactual-based XAI method achieves this by generating counterfactual examples. For instance, for a patient with an "expired" status, our method might indicate that changing the "witness status" from "not witnessed" to "witnessed by family member" and changing "AED use" from "not used" to "used" could have led to survival. This counterfactual-based method provides clear, actionable modifications for specific features. With the successful development and application of our counterfactual explanation model, PCE, in predicting OHCA outcomes, our method holds the potential for adaptation to other geohealth datasets, enabling the development of multi-level intervention strategies accordingly. Furthermore, with adjustments to align with principles of feasibility, diversity, and proximity, our method can be applied to other domains where data exhibit spatial patterns and support counterfactual outcomes.

Health Policy Recommendation Our analysis provides valuable insights into the factors influencing OHCA survival outcomes in Georgia, revealing significant regional differences that necessitate tailored health policy interventions. In North Georgia, particularly within the Atlanta metro area, the variable importance map highlights that the duration from EMS dispatch to arrival and patient age are critical factors. The consistent negative correlation between shorter EMS response times and higher survival probabilities underscores the urgency of improving EMS efficiency. This finding suggests that investing in advanced dispatch systems, better traffic management for emergency vehicles, and strategic placement of EMS stations could substantially enhance survival rates in this densely populated area. Additionally, since younger patients tend to have better survival outcomes, health policies could also focus on preventive measures and awareness programs targeting older adults to mitigate their higher risks. In South Georgia, the analysis indicates that multiple variables, including initial patient acuity, who witnessed the event, and who performed CPR, play crucial roles. The significant negative correlation of higher acuity levels with survival suggests the need for early intervention strategies and better acute care management. The positive impact of being witnessed by healthcare professionals and the use of AEDs implies that increasing the presence and readiness of medical personnel, as well as public access to AEDs, could improve outcomes. Training community members in AED use and CPR, especially in identified high-impact counties like Telfair, Pulaski, Coffee, and Appling, could empower bystanders to act swiftly and effectively, potentially bridging the gap before professional help arrives. Additionally, allocating more resources to the older population is essential to improve survival chances across the state.

Limitations and Future Works Although our method has shown improved performance and explainability, we recognize that there is still room for further exploration. For example, despite being a complete dataset spanning the years from 2019 to 2021, the data is not able to cover all geographical areas and data space, this gives limitations to the quality of the generated counterfactual examples. Additionally, the input variables for our model are limited to those related to OHCA incidents, while individual demographic variables are omitted due to privacy policies. However, these features could potentially impact health outcomes. Evaluating our model on datasets that include such information would provide a more complete picture, helping to identify significant features of health outcomes.

3.7 Conclusion

In this study, we propose a Spatial Counterfactual Explanation (SpaCE) method for health outcome prediction and explanation, and demonstrate its effectiveness in the task of OHCA survival outcome prediction. Overall, this study presents three key contributions. First, we develop a spatially explicit health outcome predictor (SEP) model that effectively captures the nuances of both spatial and health variable distributions. Second, we introduce a tailored Prototype-guided Counterfactual Explanation (PCE) algorithm designed to elucidate the decision-making process of this model. This provides valuable insights for intervention strategies and policymaking by examining counterfactual examples. Finally, we apply the model to predict and analyze the Out-of-Hospital Cardiac Arrest (OHCA) dataset in Georgia, United States. Our model outperforms other state-of-the-art ML models in prediction accuracy and generates the most closely resembling counterfactual examples to the query case. Through analysis of generated counterfactual examples, our method effectively identifies risk factors, quantifies their impact on health outcomes at both individual and county levels and provides tailored intervention strategies. This illustrative case study highlights the versatility of our method, demonstrating its potential applicability to diverse individual, public, and population health contexts.

Chapter 4

SpatialCausal: A Spatially-Aware Causal Inference Framework for Out-of-Hospital Cardiac Arrest Survival Outcome Prediction

Zhang, Jielu, et al. "SpatialCausal: A Spatially-Aware Causal Inference Framework for Out-of-Hospital Cardiac Arrest Survival Outcome Prediction." To be submitted to a peer-reviewed journal International Journal of Geographical Information Science

4.1 Introduction

Out-of-Hospital Cardiac Arrest (OHCA)(B. McNally et al. 2011) is defined as the loss of functional cardiac mechanical activity in association with an absence of systemic circulation occurring outside a hospital setting. With more than 350,000 people suffering from OHCA in the United States (US) each year, only lower than 6% of them survived(Smida et al. 2022). Identifying key factors affecting the survival rate of OHCA is critical to improving the probability of survival. Previous studies on OHCA survival outcome prediction leverage statistics(Sayegh et al. 1999; T. D. Rea et al. 2010) to identify risk factors. In recent years, explainable machine learning (XAI) methods(Harford, Del Rios, et al. 2022; Harford, Darabi, et al. 2019; Al-Dury et al. 2020), known for their superior predictive and interpretable capabilities to manage complex data patterns, have been introduced to OHCA survival prediction task. Among these studies, most did not consider spatial information when developing XAI models. Only a few have incorporated spatial information to estimate spatially varying effects for OHCA outcome prediction—for example, SpaCE(J. Zhang et al. 2025). However, in this process, the spatially varying effects of risk factors learned by SpaCE represent correlation rather than causality. Causality(L. Yao et al. 2021; Pearl 2009; Stephen L Morgan and Winship 2014), also referred to as a causal effect, implies that the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Causal inference is the process of drawing a conclusion about a causal connection based on the conditions under which an effect occurs. Uncovering the causal mechanisms and estimating the causal effect underlying geographical health disparities is critical for developing effective, targeted intervention guidance across geospatial areas.

Various causal effect estimation methods for observational data have emerged with the advancement of machine learning (ML)(Luo, Jian Peng, and J. Ma 2020; Jiao et al. 2024; L. Yao et al. 2021). One well-known causal inference framework is the potential outcome framework, also known as the Rubin Causal Model (RCM)(Imbens and Rubin 2010), which leverages potential outcomes to estimate causal effects—representing the possible results that could occur under different treatment assignments. When applied to health outcome prediction, methods within this framework can be categorized into reweighting(S. Shi et al. 2024), stratification(H. Jin and Rubin 2008), matching(Zubizarreta et al. 2023), tree-based(G. Wang, J. Li, and Hopp 2016), and representation-based approaches(L. Yao et al. 2021). These methods assist in estimating individual treatment effects (ITE)(Wager 2024) and average treatment effects (ATE)(Wager 2024) by leveraging observational health data. However, most of these causal effect estimation methods typically assume that spatial factors do not influence treatment assignment or outcome variation(H. Jin and Rubin 2008; Zubizarreta et al. 2023; G. Wang, J. Li, and Hopp 2016). In real-world scenarios, spatial information and geographical clustering of exposures and outcomes are common. Ignoring such spatial information can lead to biased estimates of causal effects.

With the development of Geospatial Artificial Intelligence (GeoAI), Geospatially Explainable Artificial Intelligence (GeoXAI) has also seen notable advances with its sophisticated techniques for uncovering underlying spatial mechanisms. Recent innovations such as GeoShapley (Z. Li 2024) and SpaCE (J. Zhang et al. 2025) are GeoXAI models developed specifically to estimate spatially-varying effects for both spatial and non-spatial variables. GeoShapley applies SHAP(SHapley Additive exPlanations)-based methods to spatial regression tasks, while SpaCE—a Spatial Counterfactual Explainable Deep Learning model—captures nonlinear spatial-health relationships using a Variational Autoencoder (VAE). While these methods effectively model spatial patterns, they fall short in addressing the crucial task of causal effect estimation. This limitation poses challenges for generating targeted intervention strategies, which are essential for advancing health geography and guiding decisions across geospatial contexts.

In this study, we developed a spatially-aware causal inference framework, named as SpatialCausal, that explicitly incorporates spatial confounders into the causal inference model while leveraging Variational Autoencoder(VAE) based architecture incorporating implicit confounders to estimate spatially varying causal effects. By integrating spatial information, our method facilitates causal effect estimation both at the individual level and across broader geographical areas. By implicitly learning unmeasured confounders leverage VAE-based architecture, we are able to estimate causal effect more reasonably. We applied this study to OHCA data in Georgia, which is the same data as in Chapter 3, aiming to explore whether the causal effect of an intervention varies across different locations and to what extent these variations occur. We futher compared our methods with other traditional causal effect methods to demonstrate the effectiveness of our developed framework. The results provide a more nuanced intervention guidance about how to improve OHCA health outcome.

4.2 Related Work

Geospatially Explainable AI Models for Health Outcome Prediction Several emerging models have been proposed to enhance spatial explainability in predictive tasks. One such model is Spatial Random Forest (Spatial RF) (Benito 2021; Wright and Ziegler 2015), which is well-suited for spatial classification and regression tasks. Another is the Spatial Regression Graph Convolutional Neural Network (SRGCNN) (D. Zhu et al. 2022), specifically designed for spatial regression. Both methods show promise in healthcare outcome prediction. Spatial RF enables spatial regression modeling by generating spatial predictors that help capture the spatial structure of training data. This approach aims to reduce spatial autocorrelation in model residuals while providing accurate variable importance scores. In contrast, SRGCNN formalizes the spatial weights matrix W and cross-sectional data (X, y) as a fully connected graph within a graph convolutional neural network (GCNN) framework, incorporating both spatial and non-spatial effects. Experimental results highlight its effectiveness in handling a wide range of geospatial data, and the authors suggest its potential applicability in public health research. Additionally, GeoShapley (Z. Li 2024) is a post-hoc explanation method that treats geographic coordinates as distinct feature columns in predictive models. It employs the Joint Shapley concept, considering two separate features as a combined feature for interpretation. Meanwhile, SpaCE(Spatial Counterfactual Explainable Deep Learning)(J. Zhang et al. 2025) enhances the explainability of spatially explicit predictive models by leveraging counterfactual explanations. While these methods can identify global and local feature importance, they do not provide causal effect estimates for treatment variables, which are essential for guiding real-world interventions.

Causal Inference Model for Healthoutcome Prediction Machine learning-based causal inference methods have emerged as powerful tools in health outcome prediction, providing robust mechanisms to estimate causal effects from observational data. Unlike traditional statistical approaches, machine learning models can handle complex, high-dimensional datasets, enabling better confounding control and heterogeneity analysis. In the potential outcome framework based causal inference, the objective is to mitigate confounding bias and estimate treatment effects from counterfactual outcome(L. Yao et al. 2021). Methods (Y. Zhu et al. 2024; L. Yao et al. 2021) in this category are often grouped by their approaches to controlling confounders, including reweighting(S. Shi et al. 2024), stratification(H. Jin and Rubin 2008), matching(Zubizarreta et al. 2023), tree-based(G. Wang, J. Li, and Hopp 2016), and representation-based(L. Yao et al. 2021). These methods help in estimating individual treatment effects (ITE) and average treatment effects (ATE) by leveraging observational data. In the context of healthcare outcome prediction, several key studies illustrate these approaches. As for reweighting method, F. Li and F. Li (2019) introduced a propensity score weighting framework to estimate causal effects across multiple treatments, applying this to analyze racial disparities in medical expenditures among different racial groups using the 2009 Medical Expenditure Panel Survey (MEPS) data. In terms of stratification, Linden (2014) employed marginal mean weighting through stratification (MMWS) to measure pre and post-intervention differences in hospitalizations following a disease management program for congestive heart failure, the results underscore MMWS as a valuable alternative for evaluating healthcare interventions with observational data. As for matching-based method, Huang et al. (2023) proposed GPMatching, a novel matching method utilizing Gaussian process priors to define matching distance, which they applied to assess the effectiveness of early biological medication for Juvenile Idiopathic Arthritis using electronic medical record data. As for causal tree-based causal inference, G. Wang, J. Li, and Hopp (2016) utilized causal trees to identify patient groups with differing outcomes across healthcare providers, highlighting that patient-provider alignment based on outcome information can lead to improved expectations for patients. Despite advancements in causal inference for healthcare, these methods often overlook spatial information, which is crucial for estimating spatially varying causal effects and guiding geographically targeted interventions.

Spatially-aware Causal Inference Model for Health Outcome Prediction To the best of our knowledge, the field of deep-learning based Spatially-aware Causal Inference model is still in its early stage. Most causal inference methods are from statistics field. For example, Arpino and Mattei (2016) explicitly model interference as a function of units' characteristics and apply this approach to assess a policy in Tuscany (Italy) targeting small handicraft firms. Their findings reveal that the policy's benefits decrease for treated firms experiencing high interference, and that ignoring interference results in a slight underestimation of the average causal effect. Papadogeorgou et al. (2022) extends the classic potential outcomes framework to spatio-temporal point processes by modeling the treatment point process as a stochastic intervention. The authors define causal estimands based on the expected number of outcome events within a specified region under different stochastic treatment assignment strategies, an application estimating the effects of American airstrikes in Iraq suggests that increasing daily airstrikes may lead to more insurgent attacks overall and potentially displace attacks from Baghdad to locations up to 400 km away. In the few of deep learning based method, Tec, Scott, and Zigler (2023) propose weather2vec, a method to learn scalar or vector representations of non-local information (e.g., weather), which can then be used for confounding adjustment in causal inference. Through simulations and two air pollution case studies, they show that this approach helps account for known regional confounders. In summary, few studies have investigated spatially-aware causal inference leveraging machine learning methods, particularly within the area of health outcome prediction.

4.3 Methodology

There are two key challenges in developing spatially-aware causal inference models. The first key challenge is determining how to effectively integrate spatial confounder into the analytical framework. Standard causal inference techniques, such as propensity score reweighting and covariate adjustment, typically ignore spatial information in treatment assignment or outcome variations. However, in real-world scenarios, spatial information and geographical clustering of certain data attribute are nontrivial. Ignoring these spatial information can lead to biased effect estimates. Another significant challenge stems from the reliance on observed confounders for causal effect estimation. Traditional methods assume that all relevant confounders are measured, but in practice, unobserved or hidden confounders—such as socio-economic status, healthcare accessibility, and environmental factors—can introduce substantial bias. These unmeasured confounders can simultaneously influence both the treatment (e.g., access to healthcare interventions) and the outcome (e.g., disease progression), making it difficult to isolate the true causal effect. Without robust strategies to account for these hidden confounders, causal effect estimation may be unreliable. In summary, the dual challenges of incorporating spatial information and addressing unmeasured confounders highlight the need for a robust spatially-aware causal inference models. To address the first challenge, we propose leveraging spatial representation learning method which has been demonstrated effectively to learn spatial confounders (Mai, Janowicz, Yan, et al. 2020; Mai, Cundy, et al. 2022; Mai, Xuan, et al. 2023). These learned confounder vectors are then incorporated to balance the distribution of covariates between treatment groups during downstream causal effect analysis. To address the second challenge, we adopt an alternative approach tailored to a surrogate-rich setting: estimating a latent-variable model that simultaneously discovers hidden confounders and infers their influence on both treatments and outcomes. These hidden confounders are learned from not only the health feature network but also the spatial feature network. Specifically, we employ maximum-likelihood approximation methods using variational autoencoders (VAEs)(Diederik P Kingma, Welling, et al. 2013; Durk P Kingma et al. 2016), which maximize the Evidence Lower Bound (ELBO) as a tractable surrogate for the true data likelihood. Accordingly, we developed a spatially-aware causal inference framework, SpatialCausal, that explicitly addresses spatial confounder integration while incorporating implicit confounders to estimate spatially-varying causal effects. By integrating spatial confounders, our method facilitates causal effect estimation both at the individual level and across broader geographical areas. This proposed framework aims to explore whether the causal effect of an intervention varies across different locations and to what extent these variations occur, providing a more nuanced interventions guidance about how to improve health outcomes in diverse geographical contexts.

Preliminary

Assumption 1 (No Unobserved Confounding) For most methods, we need to assume that the variables we have measured are enough to account for any existing connection between the treatment and the outcome that isn't actually causal. This assumption(Imbens and Rubin 2010; Cunningham 2021) ensures that, once we consider these measured variables, the treatment is assigned in a way that is effectively random. This idea is known as conditional exchangeability, conditional ignorability, or causal sufficiency. Mathematically, this is written as:

$$Y(t) \perp\!\!\!\perp T \mid X, \quad \forall t \in T$$

So the potential outcomes are independent of the particular value of the treatment.

Assumption 2 (Positivity Assumption) The probability of receiving each treatment level is strictly greater than zero for all possible values of the covariates(Imbens and Rubin 2010; Cunningham 2021). Mathematically, this is expressed as:

$$P(T = t \mid X = x) > 0, \quad \forall t \in T, \quad \text{for all } X = x \text{ where } P(X = x) > 0.$$

This means that for any given covariate values $X = x$, there must be a positive probability of receiving any treatment a.

Assumption 3 (Consistency) This means that for any collection of potential outcomes, say $Y(t)$ for all $t \in T$, if the actual treatment received is $T = t$(Imbens and Rubin 2010; Cunningham 2021), then:

$$Y(T) = Y(t) = Y.$$

This establishes the connection between the potential outcome and the observed (factual) outcome.

Adjustment for Spatial Confounder

To accurately estimate treatment effects while incorporating spatial information, we introduce a spatial representation model that encodes spatial coordinates (latitude, longitude) into high-dimensional vectors. We then integrate the spatial representation at location s with health-related covariates X_s. By adjusting for both spatial information and health-related covariates (Figure 4.1), our approach enables more precise estimation of individual treatment effects (ITE) and average treatment effects (ATE).

Definition 1 (Spatial Confounder) Let X_s be the observed covariates for location s, and let L_s represent the represented spatial information. Define the **location encoding function** $\phi(X_{loc})$ such that:

$$L_s = \phi(X_{loc}), \quad \forall s \in S$$

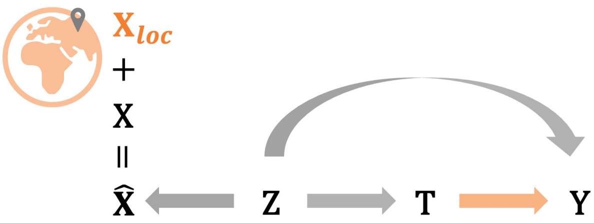


Figure 4.1: Causal Graph

Proposition 1 (The Spatial Confounder as a Covariate Adjustment Mechanism) If spatial confounder is not explicitly controlled, standard methods may introduce bias:

$$E[Y_s(T) \mid X_s] \neq E[Y_s(T) \mid X_s, L_s], \quad \forall s \in S$$

which leads to inaccurate treatment effect estimates. The spatial representation acts as a term that ensures:

$$Y_s \perp\!\!\!\perp T_s \mid (X_s, L_s), \quad \forall s \in S$$

This guarantees that treatment assignment is conditionally independent of potential outcomes once we adjust for spatial confounder.

Proposition 2 Identifiability of the Individual Treatment Effect Given the spatial representation L_s, we can estimate the ITE at location s as:

$$\text{ITE}_s = E[Y_s \mid X_s, L_s, T_s = 1] - E[Y_s \mid X_s, L_s, T_s = 0]$$

Proposition 3 Identifiability of the Average Treatment Effect The ATE across all locations is given by:

$$E\left[E[Y_s \mid X_s, L_s, T_s = 1] - E[Y_s \mid X_s, L_s, T_s = 0]\right]$$

Given that some unmeasured confounders may be missing from our dataset, potentially violating the *no unobserved confounding* assumption, and considering the strong capability of maximum-likelihood approximation model that integrates diverse modalities, including geospatial and tabular data, we leverage a *Variational Autoencoder-based Causal Model*. In this model, we first concatenate the health-related covariates X_s with spatial confounder L_s as $\hat{X}$, and we assume that the joint distribution $p(Z, \hat{X}, t, y)$ of the latent confounders Z and the observed confounders $\hat{X}$ can be approximately recovered solely from the observations $(\hat{X}, t, y)$. We will prove that the ITE can be estimated if the distribution $p(Z, \hat{X}, t, y)$ is known.

Theorem If we recover the joint distribution $p(Z, \hat{X}, t, y)$, where $\hat{X}$ is a vector that concatenates both local covariates X and spatial confounders L, then we can recover ITE under the causal model.

We prove that $p(y \mid \hat{X}, do(t = 1))$ is identifiable under the theorem's premise, the case for $t = 0$ follows identically, and the expectation defining ITE is directly recovered from the probability function. ATE is identified if ITE is identified. This means that we can identify the causal effect of treatment t on outcome y can be estimated using conditional probabilities from the observed data. Using the properties of do-calculus, we rewrite $p(y \mid \hat{X}, do(t = 1))$ as:

$$p(y \mid \hat{X}, do(t = 1)) = \int_Z p(y \mid \hat{X}, t = 1, Z) p(Z \mid \hat{X}) dZ.$$

Since $\hat{X}$ now includes both non-spatial and spatial features ($\hat{X} = [X, L]$), and the overall potential confounder distribution Z is derived from these features, all relevant confounders are effectively accounted for. Given that $p(Z, \hat{X}, t, y)$ is fully observed under the theorem's assumption, all terms in the final equation are identifiable, thus proving the theorem.

4.4 Architecture

The *SpatialCausal* architecture (see Figure 4.2) consists of two main components: (1) a Spatial Confounder Representation Model, which captures relative geospatial relationship between individuals, and (2) a maximum-likelihood approximation-based causal inference model, which adjusts for spatial confounders to estimate spatially varying causal effects. For the causal inference model, we draw inspiration from the deep learning-based latent-variable model (Louizos et al. 2017), which embeds a T-learner(Nagai et al. 2024), a classical meta learner, within an encoder–decoder architecture. In our implementation, we employ a Variational Autoencoder (VAE) trained alongside the spatial confounder representation model. Given the VAE's proven ability to learn latent distributions and handle diverse modalities, it not only captures unmeasured confounders but also models the complex distribution of concatenated geo-embeddings and health feature embeddings, enhancing the robustness of spatially aware causal effect estimation.

Spatial Confounder Representation Model

Previous studies have demonstrated that representing and integrating spatial geometry (e.g., points, polylines, and polygons) is beneficial for various geospatial prediction tasks, such as geographic question answering (Mai et al. 2020a), geospatial shape recognition (Mai et al. 2023a; Siampou et al. 2024), POI type prediction (Yan et al. 2017; Mai et al. 2020b), species fine-grained recognition (Mac Aodha et al. 2019; Mai et al. 2023c; Wu et al. 2024), satellite image classification (Mai et al. 2023b), trajectory generation (Rao et al. 2020; Klemmer et al. 2023), and terrain feature detection (Li et al. 2021), among others. Integrating spatial information into causal effect models helps account for potential unmeasured confounders and enables the estimation of spatially varying effects, providing a deeper understanding of geospatial variations across different features(A. S. Fotheringham and Z. Li 2023).

For spatial effect representation (Mac Aodha, Cole, and Perona 2019; Mai, Janowicz, Yan, et al. 2020; Mai, Janowicz, Y. Hu, et al. 2022; Mai, Lao, et al. 2023; Mai, Xuan, et al. 2023), we adopt the location encoder from GeoCLIP (Vivanco Cepeda, Nayak, and M. Shah 2024). It is important to note that the GeoCLIP encoder is a point-based encoder rather than a spatial-structure-based encoder like a Graph Neural Network (GNN) (D. Zhu et al. 2022). While GNNs can quantify spatial effects, they implicitly learn spatial features during training, which requires a carefully designed loss term for effective optimization. Assessing the extent to which these models rely on spatial information for downstream tasks can be challenging due to the complexity of their learning processes. In contrast, point-based encoders like GeoCLIP can capture spatial information in a pre-trained manner, eliminating the need for training from scratch. This allows them to serve as plug-and-play components within various model architectures alongside other types of features, making them highly efficient and easily transferable to new applications. The GeoCLIP encoder is in an equal earth projection and applied positional encoding with Random Fourier variables that transform two-dimensional coordinates ($G_m \in R^2, \forall m \in [1, \ldots, M]$) into a high-dimensional location embedding $R^D$where $D \gg 2$. For the extraction of health variables, we utilized a Multilayer Perception (MLP). After obtaining two embeddings for spatial and health variables, we concatenate these embeddings and feed them into a VAE-based causal inference model to learn a fused distribution of data and estimate the spatially varying causal effect (Figure 4.2).

Spatially-Aware Causal Inference Model

The Variational Autoencoder (VAE) (Durk P Kingma et al. 2016) is an advanced extension of the traditional Autoencoder (AE). Unlike conventional AEs, which encode input data as a single point in the latent space, VAEs model the input as a probability distribution over potential values. This probabilistic approach not only mitigates the influence of outliers but also enhances feature selection, thereby improving model robustness. Additionally, VAEs are particularly well-suited for handling heterogeneous data, as they internally normalize different variable types and assign weights based on their relevance to the task. In our study, we define the input $\hat{X}$ as a set of non-spatial variables (X) and spatial variables (X_{loc}). The VAE effectively integrates these heterogeneous components, enabling a more comprehensive feature representation. Furthermore, when combined with a causal inference model, the VAE demonstrates significant potential for capturing unmeasured confounders by leveraging the learned probabilistic distribution. This integration facilitates a more accurate estimation of spatial treatment effects.

Therefore, we developed a spatially-aware VAE-based causal inference model to infer the complex non-linear relationships between $\hat{X}$ and (Z, t, y), where the input is defined as $\hat{X} = \{X, L\}$, L is a spatial representation of X_{loc}. Our approach aims to approximate the joint distribution $p(Z, X, L, t, y)$ and effectively captures the intricate dependencies among these variables. This spatially-aware causal inference model contains two parts, an inference model and generative model, which can be named as encoder and decoder in a traditional VAE model.

(1) Inference model ***Variational posterior distribution*** Since the complex, non-linear transformations in neural networks make exact inference intractable, we employ variational inference (VI) along with

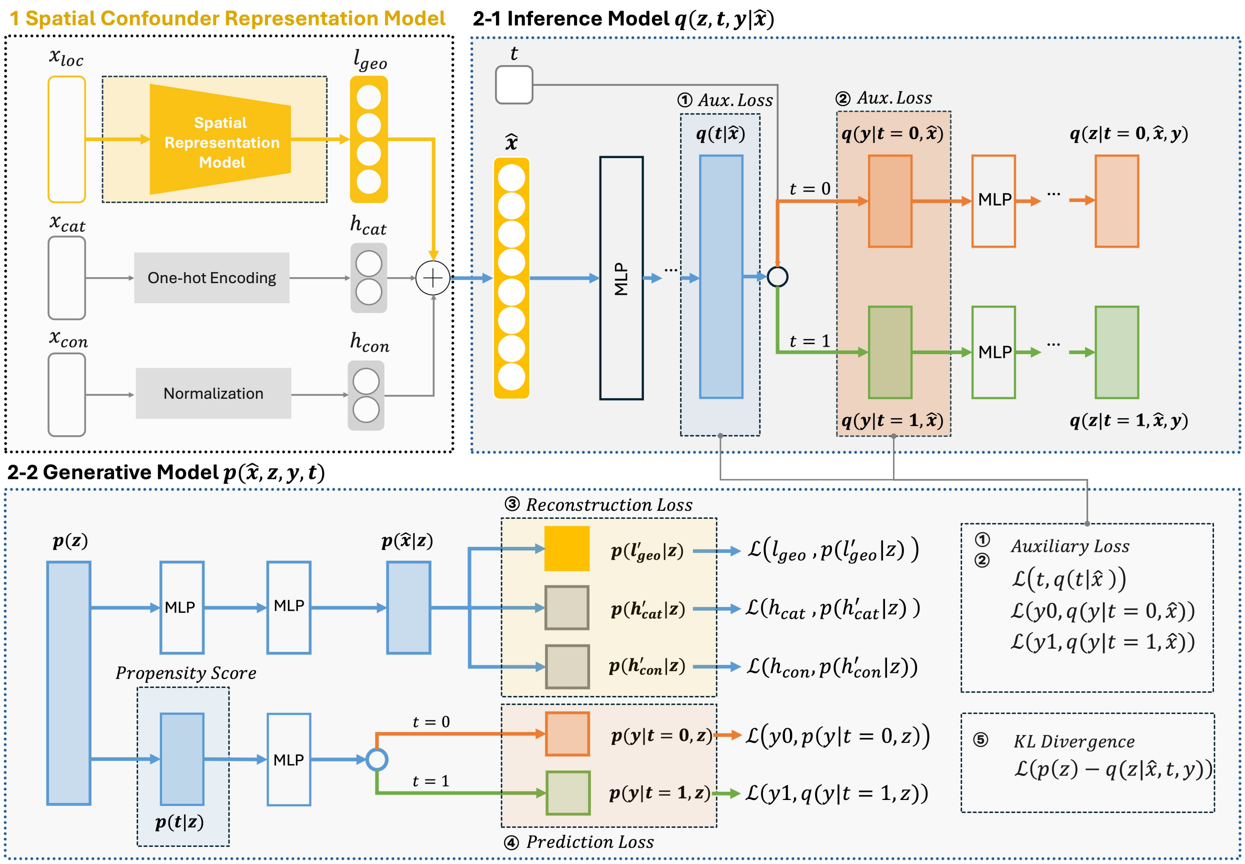

1 Spatial Confounder Representation Model
2-1 Inference Model $q(z,t,y|\hat{x})$
x_{loc}
l_{geo}
Spatial Representation Model
x_{cat}
h_{cat}
One-hot Encoding
x_{con}
h_{con}
Normalization
t
$\hat{x}$
MLP
① $Aux. Loss$
$q(t|\hat{x})$
② $Aux. Loss$
$q(y|t=0,\hat{x})$
$t=0$
$t=1$
MLP
MLP
$q(z|t=0,\hat{x},y)$
$q(y|t=1,\hat{x})$
$q(z|t=1,\hat{x},y)$
2-2 Generative Model $p(\hat{x},z,y,t)$
③ $Reconstruction\ Loss$
$p(z)$
MLP
MLP
$p(\hat{x}|z)$
$p(l'_{geo}|z)$
$\mathcal{L}(l_{geo}, p(l'_{geo}|z))$
$p(h'_{cat}|z)$
$\mathcal{L}(h_{cat}, p(h'_{cat}|z))$
$p(h'_{con}|z)$
$\mathcal{L}(h_{con}, p(h'_{con}|z))$
$Propensity\ Score$
$p(t|z)$
MLP
$t=0$
$t=1$
$p(y|t=0,z)$
$\mathcal{L}(y0, p(y|t=0,z))$
$p(y|t=1,z)$
$\mathcal{L}(y1, q(y|t=1,z))$
④ $Prediction\ Loss$
① ② $Auxiliary\ Loss$
$\mathcal{L}(t, q(t|\hat{x}))$
$\mathcal{L}(y0, q(y|t=0,\hat{x}))$
$\mathcal{L}(y1, q(y|t=1,\hat{x}))$
⑤ $KL\ Divergence$
$\mathcal{L}(p(z) - q(z|\hat{x},t,y))$

Figure 4.2: The *SpatialCausal* Framework. **Panel (1): a Spatial Confounder Representation Model** that captures the spatial characteristics of the neighborhood. **Panel (2): a Variational Autoencoder-based Spatial Causal Inference Model**, this panel is composed of two parts as follows. *Panel (2-1): Inference model in Spatial Causal Effect Estimation. Panel (2-2): Generative model in Spatial Causal Effect Estimation.*

inference networks. As is well known, variational inference approximates the true posterior distribution $p(Z|\hat{X})$ with a structured, parametric distribution $q_\phi(Z|\hat{X})$, where ϕ represents learnable parameters. To leverage this within a neural network framework, we developed a spatially-aware inference model that output the parameters of a fixed-form posterior approximation over the latent variables Z, such as a Gaussian distribution, given the observed variables $\hat{X} = \{X, L\}$. This approach enables efficient and scalable inference. The variational posterior distribution over latent variables $\mathbf{z}_i$, conditioned on observed variables

x_i, spatial confounder l_i, treatment assignment t_i, and outcome y_i, is defined as:

$$q(z_i|x_i, l_i, t_i, y_i) = \prod_{j=1}^{D_z} \mathcal{N}(\mu_j = \bar{\mu}_{i,j}, \sigma_j^2 = \bar{\sigma}_{i,j}^2) \tag{4.1}$$

$$\bar{\mu}_i = t_i \mu_{t=1,i} + (1 - t_i)\mu_{t=0,i} \qquad \bar{\sigma}_i^2 = t_i \sigma_{t=1,i}^2 + (1 - t_i)\sigma_{t=0,i}^2 \tag{4.2}$$

$$\mu_{t=0,i}, \sigma_{t=0,i}^2 = g_2 \circ g_1(x_i, l_i, y_i) \qquad \mu_{t=1,i}, \sigma_{t=1,i}^2 = g_3 \circ g_1(x_i, l_i, y_i) \tag{4.3}$$

where:

- x_i represents the observed non-spatial features for instance i.
- l_i denotes the spatial confounder.
- t_i is a binary treatment assignment (e.g., $t_i = 0$ or $t_i = 1$).
- y_i is the observed outcome.
- z_i represents the latent variable.
- $\mathcal{N}(\mu, \sigma^2)$: a Gaussian (Normal) distribution with mean μ and variance σ^2.
- g_1: shared neural networks, mapping observed variables $(\mathbf{x}_i, \mathbf{l}_i, y_i)$ to shared distribution parameters.
- g_2, g_3: separate neural networks, mapping separately for $t_i = 0$ and $t_i = 1$ to different distribution parameters.

Auxiliary distributions For out-of-sample predictions, it is necessary to know the prediction of treatment assignment t and its corresponding outcome y before inferring the distribution over Z. To address this, we introduce two auxiliary distributions that facilitate the prediction of t_i and y_i for new samples. The treatment assignment t_i is modeled as a Bernoulli-distributed random variable with probability parameterized by a neural network g_4 and activated by the sigmoid function σ:

$$q(t_i|\mathbf{x}_i, \mathbf{l}_i) = \text{Bern}(\pi = \sigma(g_4(\mathbf{x}_i, \mathbf{l}_i))) \tag{4.4}$$

The conditional outcome distribution $q(y_i|x_i, t_i)$ is modeled as either a Gaussian, Bernoulli distribution, or a Categorical distribution, depending on whether the outcome is continuous, binary, or categorical.

- *Continuous case* ($y \in R$):

$$q(y_i|\mathbf{x}_i, \mathbf{l}_i, t_i) = \mathcal{N}(\mu = \bar{\mu}_i, \sigma^2 = \bar{v}) \qquad \bar{\mu}_i = t_i(g_6 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) + (1 - t_i)(g_7 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) \tag{4.5}$$

- *Binary case* ($y \in \{0, 1\}$):

$$q(y_i|\mathbf{x}_i, \mathbf{l}_i, t_i) = \text{Bern}(\pi = \bar{\pi}_i) \qquad \bar{\pi}_i = t_i(g_6 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) + (1 - t_i)(g_7 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) \tag{4.6}$$

- *Categorical case* ($y \in \{1, \ldots, K\}$):

$$q(y_i|\mathbf{x}_i, \mathbf{l}_i, t_i) = \text{Cat}(\pi = \bar{\pi}_i) \qquad \bar{\pi}_i = t_i(g_6 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) + (1 - t_i)(g_7 \circ g_5(\mathbf{x}_i, \mathbf{l}_i)) \tag{4.7}$$

where:

- g_5: shared neural networks, mapping observed variables $(\mathbf{x}_i, \mathbf{l}_i)$ to shared distribution parameters.
- g_6, g_7: separate neural networks, mapping separately for $t_i = 0$ and $t_i = 1$ to different distribution parameters.
- $\mathcal{N}(\mu, \sigma^2)$ means a Gaussian (Normal) distribution with mean μ and variance σ^2. Bern is Bernoulli distribution, Cat is Categorical distribution.

(2) Generative model In the generative model, the goal is to utilize the learned latent variables Z to reconstruct non-spatial features X, spatial confounders L, the treatment variable t, and the outcome y. To predict y across different treatment groups, we adopt an architecture inspired by TARnet (Shalit, Johansson, and Sontag 2017). However, instead of conditioning on observed features, we condition on the latent variables Z to make predictions. The probability distributions governing $p(z_i)$, $p(x_i, l_{s_i}|z_i)$, $p(t_i|z_i)$, and $p(y_i|t_i, z_i)$ are defined as follows:

$$p(z_i) = \prod_{j=1}^{D_z} \mathcal{N}(z_{ij}|0, 1) \tag{4.8}$$

where $p(z_i)$ represents the prior distribution of the latent variable z_i, which is assumed to be factorized as a standard Gaussian distribution over each dimension. D_z denotes the dimensionality of z_i, and each component z_{ij} follows an independent standard normal distribution with mean 0 and variance 1.

$$p(x_i, l_i|z_i) = \left(\prod_{j=1}^{D_x} p(x_{ij}|z_i)\right) \cdot \left(\prod_{k=1}^{D_l} p(l_{ik}|z_i)\right) \tag{4.9}$$

where $p(x_i, l_i|z_i)$ represents the joint distribution of the observed variable x_i and the spatial confounder l_i conditioned on the latent variable z_i. The distribution of x_i is assumed to factorize over its D_x dimensions, with each component x_{ij} being conditionally independent given z_i. Similarly, the distribution of l_i factorizes over its D_l dimensions, where each component l_{ik} is also conditionally independent given z_i. Furthermore, the probability distribution $p(x_i|z_i)$ is defined based on the type of the observed variable $p(x_i)$, while $p(l_i|z_i)$ is specified under the assumption that $p(l_i)$ follows a continuous distribution.

- *Continuous case* ($x \in R$):

$$p(x_i|z_i) = \mathcal{N}(\mu = \bar{\mu}_i, \sigma^2 = \bar{v}) \tag{4.10}$$

- *Binary case* ($x \in \{0, 1\}$):

$$p(x_i|z_i) = \text{Bern}(\sigma(f(z_i))) \tag{4.11}$$

- *Categorical case* ($x \in \{1, \ldots, K\}$):

$$p(x_i|z_i) = \text{Cat}(\sigma(f(z_i))) \tag{4.12}$$

- *Spatial confounder case* ($x \in \{1, \ldots, K\}$):

$$p(l_i|z_i) = \mathcal{N}(\mu = \bar{\mu}_i, \sigma^2 = \bar{v}) \tag{4.13}$$

where $\mathcal{N}(\mu, \sigma^2)$ means a Gaussian distribution with mean μ and variance σ^2. Bern is Bernoulli distribution, Cat is Categorical distribution. $\sigma(\cdot)$ is the sigmoid(binary case) or softmax function (categorical case). $f(\cdot)$ is a parameterized neural network.

$$p(t_i|z_i) = \text{Bern}(\sigma(f_1(z_i))) \tag{4.14}$$

where $p(t_i|z_i)$ represents the probability distribution of the observed variable t_i given the latent variable z_i.where $f_1(\cdot)$ is neural network parameterized by its own parameters.

The conditional outcome distribution $p(y_i|t_i, z_i)$ is modeled as either a Gaussian, Bernoulli distribution, or a Categorical distribution, depending on whether the outcome is continuous, binary, or categorical. For a continuous outcome, we parameterize the probability distribution as a Gaussian with its mean given by a TARnet architecture, i.e., a treatment-specific function, and its variance fixed to $\hat{v}$. For a binary outcome, we use a Bernoulli distribution similarly parameterized by a TARnet. For a categorical outcome, we use a Categorical distribution.

- *Continuous case* ($x \in R$):

$$p(y_i|t_i, z_i) = \mathcal{N}(\mu = \hat{\mu}_i, \sigma^2 = \hat{v}) \qquad \hat{\mu}_i = t_i f_2(z_i) + (1 - t_i) f_3(z_i) \tag{4.15}$$

- *Binary case* ($x \in \{0, 1\}$):

$$p(y_i|t_i, z_i) = \text{Bern}(\pi = \hat{\pi}_i) \qquad \hat{\pi}_i = \sigma(t_i f_2(z_i) + (1 - t_i) f_3(z_i)) \tag{4.16}$$

- *Categorical case* ($x \in \{1, \ldots, K\}$):

$$p(y_i|t_i, z_i) = \text{Cat}(\pi = \hat{\pi}_i) \qquad \hat{\pi}_i = \sigma(t_i f_2(z_i) + (1 - t_i) f_3(z_i)) \tag{4.17}$$

where $f_2(\cdot)$ and $f_3(\cdot)$ are neural networks parameterized by its own parameters, $\hat{\pi}_i$is a result of going through a neural network $f_k(\cdot)$ activated by σ, a sigmoid (binary case) or softmax function (categorical case)

(3) Spatially-Aware Causal Inference Model Loss The following given loss function is based on variational inference, where we approximate the intractable posterior $p(z_i)$ with a learned distribution $q(z_i|x_i, l_i, t_i, y_i)$. This function represents the variational lower bound, Evidence Lower Bound(ELBO) for the graphical model. It serves as the single objective for both inference and generative models, meaning it jointly optimizes the generative model which aims to reconstruct observed data and the inference model which approximates the latent variable posterior.

$$\mathcal{L} = \sum_{i=1}^{N} E_{q(z_i|x_i,l_i,t_i,y_i)} \left[\log p(x_i, l_i, t_i|z_i) + w_{y_i} \log p(y_i|t_i, z_i) + \log p(z_i) - \log q(z_i|x_i, l_i, t_i, y_i)\right] \quad (4.18)$$

where the expectation $E_{q(\cdot)}$ ensures that we optimize the objective over sampled latent variables, the first two terms of $\log p(\cdot|z_i)$ represents the data likelihood given the latent variable z_i, w_{y_i} is the class weight for the true label y_i, the last two term is Kullback-Leibler Divergence term ensuring the posterior distribution is close to the true posterior.

The function to compute class weights for an imbalanced dataset is given by:

$$w_{y_i} = \frac{N}{U \cdot n_{y_i}} \quad (4.19)$$

where w_{y_i} is the weight for class y_i, N is the total number of samples, U is the number of unique classes, n_{y_i} is the number of samples belonging to class y_i.

Moreover, to estimate the parameters of the auxiliary distributions we add two extra terms in the variational lower bound:

$$\mathcal{L}_{\text{SpatialCausal}} = \mathcal{L} + \sum_{i=1}^{N} \left(\log q(t_i|x_i, l_i) + w_{y_i} \log q(y_i|x_i, l_i, t_i)\right), \quad (4.20)$$

4.5 Experiment Setup

Data Source This study utilizes Georgia cardiac arrest incident data (Figure 4.3) recorded from 2019 to 2021. The data is accessed from the Georgia Department of Public Health, reported by the Emergency Medical Service (EMS), and collected by The National Emergency Medical Services Information System. The data collection process is approved by the university's Institutional Review Board. A total of 57,223 individual cardiac arrest patients and 24 variables are reported in this dataset(Table 4.1). The variable Patient Outcome at End of EMS Event is used as the targeted outcome variable. After data processing, a total of 5,385 cardiac arrest patients with geographic coordinates are filtered for analysis.

Data Preprocessing Data processing steps include: (1) The location type of cardiac arrest incidents is classified using the Tenth Revision, Clinical Modification (ICD-10-CM) provided by the CDC (Centers for Disease Control and Prevention). (2) Racial categorization is refined according to the standards of the Office of Management and Budget. (3) Outliers are identified and removed. (4) Data integrity is maintained by excluding variables with more than 35% missing values. Then we remove examples that have missing values. (5) For data analysis, nominal variables are processed using one-hot encoding, while ordinal variables utilize ordinal encoding. Additionally, all numerical variables are standardized.

Fifteen variables are finally used, including latitude, longitude, etiology, gender, race, age, initial patient acuity, who witnessed the cardiac arrest, CPR provided prior to EMS arrival, who provided CPR prior to EMS, AED use prior to EMS arrival, types of CPR provided, duration from call to EMS arrival, location type, and patient outcome

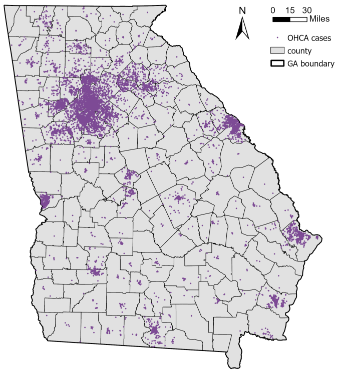


Figure 4.3: Out-of-Hospital Cardiac Arrest Distribution in Georgia from 2019 to 2021

Table 4.1: Variable Description of Georgia Out-of-Hospital Cardiac Arrest Dataset

Category	Variable
Incident Details	Incident Date, Scene Latitude, Scene Longitude, Incident Location Type
Patient Information	Patient Gender, Patient Race, Patient Age
Response Details	Response Beginning Vehicle Odometer, Response On Scene Vehicle Odometer, Incident Call Date Time, Incident Unit Arrived On Scene Date Time
Initial Assessment	Initial Patient Acuity
Cardiac Arrest Details	Cardiac Arrest Etiology, Cardiac Arrest Indications Resuscitation Attempted By EMS, Cardiac Arrest Witnessed By Whom
CPR and AED Details	CPR Provided or Not Prior to EMS Arrival, Who Provided CPR Prior to EMS Arrival, AED Used or Not Prior to EMS Arrival, Who Used AED Prior to EMS Arrival, Types of CPR Provided List
Patient Outcome	Patient Outcome at End of EMS Event, Medical Device Type of Shock, Outcome Emergency Department Disposition Description, Outcome Hospital Disposition Description

at the end of the EMS event. The patient outcome at the end of the EMS event is the targeted outcome variable with binary class expired or survived.

***SpatialCausal* Model Training** Given the imbalance in the y labels, we incorporated class weights into the loss function to enhance prediction robustness. Additionally, we applied early stopping, a regularization technique that helps prevent overfitting during training. To further stabilize training, we pre-trained $q(z)$, the inferred latent distribution, ensuring that its output aligns with a standard normal distribution before the main training phase. This step mitigates potential gradient explosions that may arise from poor initialization of $q(z)$'s parameters. Specifically, we optimized $q(z)$ for 50 iterations using Kullback–Leibler (KL) divergence minimization to promote stability. In summary, this pre-training process refines $q(z)$ to produce a distribution close to standard normal, preventing gradient instabilities in early training steps.

4.6 Result

In this section, we first examine the results from the SpatialCausal model. We leverage predicted probabilities under different treatment conditions to compute both the Average Treatment Effect (ATE) for the entire dataset and the Individual Treatment Effect (ITE) for each subject. Furthermore, we aggregate these individual treatment effects at the county level to derive a spatially varying treatment effect, which can be utilized for geographic healthcare guidance. In addition to presenting the SpatialCausal results, we compare our method's performance with several classical baseline approaches to highlight its effectiveness, and we conduct ablation studies to determine the optimal configuration of SpatialCausal.

Average Treatment Effect

The Average Treatment Effect (ATE) is determined by averaging the individual treatment effects observed across the dataset. For each treatment variable, we develop a dedicated causal model that accounts for all spatial and non-spatial confounders. This model estimates the individual treatment effects for every sample, and by aggregating these effects, we obtain the average treatment effect, which serves as the ATE for that specific treatment variable. Figure 4.4 clearly visualizes the ranking of both positive and negative ATE values across all treatment variables. The analysis shows that receiving AED treatment before EMS arrival has the highest positive effect (0.30), followed by witnessed status (0.14). In essence, patients who received AED treatment prior to EMS arrival had a significantly higher likelihood of survival compared to those who did not, and those witnessed by professionals had a greater survival probability than those observed by laypersons or family members.

Beyond the two positive treatment effects, several variables exhibit negative impacts on survival outcomes. The most pronounced negative effect is associated with patient initial acuity, indicating that patients in a more critical state at first assessment have a lower chance of survival than those with less severe conditions. The second strongest negative effect stems from the location type variable where the OHCA occurs—cardiac arrests in residential areas are linked to lower survival probabilities compared to those occurring outside such areas. Additionally, the duration from EMS response to arrival shows a significant negative effect, underscoring that longer response times drastically reduce survival chances. Further negative treatment effects are observed for the etiology of the OHCA and the number of CPR types applied during the incident. Specifically, patients whose cardiac arrest is cardiac-related have lower survival rates than those whose arrests result from causes like drowning or drug overdose, and those receiving more than two types of CPR tend to have lower survival probabilities compared to those receiving just one type. Other factors with comparatively lower negative effects include age, the use of CPR before EMS arrival, and the category of the individual performing CPR (whether a layperson, family member, community responder, or professional). The analysis confirms that elderly patients generally face lower survival probabilities than younger ones. Interestingly, even though administering CPR before EMS arrival shows only a modest negative effect (-0.05), it is associated with a lower survival rate compared to not administering CPR at all. Moreover, CPR provided by professionals before EMS arrival correlates with a lower survival probability than when performed by laypersons or family members—a counterintuitive result. In addition to the magnitude of the treatment effects from Figure 4.4, Table 4.4 details the performance metrics of the causal models, including the AUCROC and AUCPR scores, offering a comprehensive view of the model's efficacy. Figure 4.5 visualizes the AUCROC scores for all spatially-aware causal inference models.

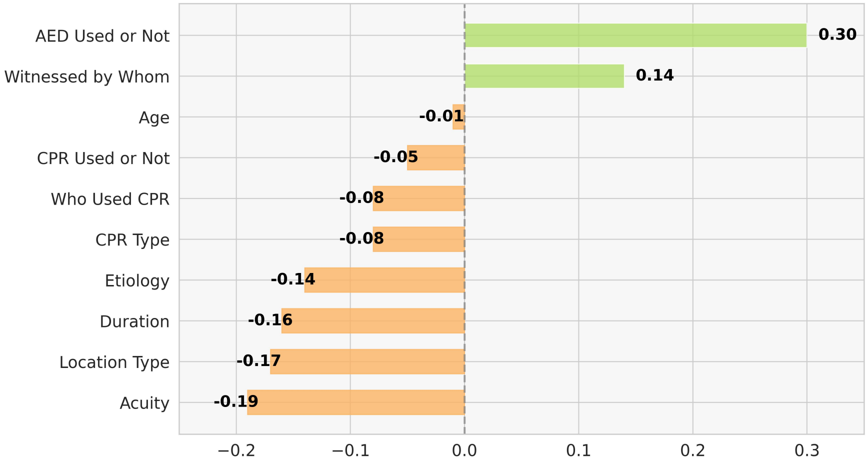


Figure 4.4: Out-of-Hospital Cardiac Arrest Average Treatment Effect for all treatment variables

Individual Treatment Effect

Individual Treatment Effect (ITE) measures the treatment effect for a specific individual and is calculated as the difference between the predicted outcome probability when the treatment is applied and the predicted outcome probability when the treatment is not applied. As shown in Table 4.2, two example cases illustrate the original and counterfactual scenarios when the treatment value changes. *Case 1:* The ground truth represents a scenario where the treatment is 0 (AED not used) and the survival outcome is 0 (expired). The counterfactual scenario assumes the treatment is 1 (AED used), with a predicted survival outcome of 1 (survived). *Case 2:* The ground truth corresponds to a scenario where the treatment is 1 (AED used) and the survival outcome is 1 (survived). The counterfactual scenario assumes the treatment is 0 (AED not used), with a predicted survival outcome of 0 (expired). The ITE calculation for Case 1 is obtained by subtracting the true outcome ($y = 0$) from the predicted survival probability when the treatment is 1. For Case 2, the ITE is calculated by subtracting the predicted survival probability when the treatment is 0 from the true outcome ($y = 1$). For other treatment variables, the ITE is calculated using the same approach.

Spatially-varying treatment effect

We estimate the spatially-varying treatment effect (Figure 4.6) by aggregating the spatially-aware individual treatment effects at the county level. Specifically, we average individual treatment effects for each county and visualize the results on a spatially-varying treatment effect map. From an overall perspective, these maps reveal that most counties exhibit a positive causal effect on survival outcomes for two key treatment variables: AED use before EMS arrival

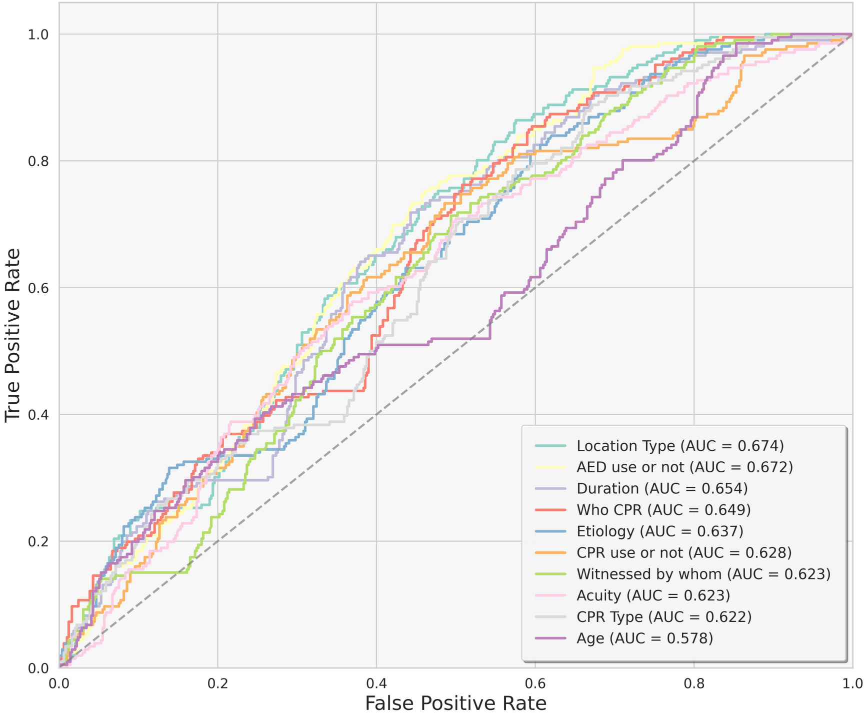


Figure 4.5: The AUCROC Performance Comparison between Various Treatment-based Causal Inference Models

and OHCA witnessed by whom (e.g., layperson, family member, or professional). In contrast, for other treatment variables, the majority of counties display a negative treatment effect. However, it is important to acknowledge the data limitations and the highly geographically clustered distribution of individual OHCA patients, this leads to missing data in certain counties and gaps in the estimated treatment effects.

The survival rate map is presented in the lower-right corner of Figure 4.6. Counties with the lowest survival rates (i.e., less than 0.1) are highlighted in yellow. Notably, the three easternmost yellow-colored counties—Emanuel, Screven, and Bulloch—fall within this lowest survival rate category. To guide potential intervention strategies for improving outcomes in these counties, we refer to the corresponding causal effect map. Actionable variables identified for intervention include AED usage, EMS dispatch-to-arrival time, witness identity, CPR type, and the individual administering CPR. Across all three counties, AED use exhibits a positive causal effect on survival, suggesting that initiatives such as AED training programs and AED location accessibility improvements may enhance survival outcomes. Conversely, EMS response time demonstrates a negative causal effect, indicating that optimizing the emergency transportation system to reduce response time could be beneficial. Similarly, the type of CPR administered also shows a negative effect, highlighting the importance of promoting more efficient and effective CPR methods. In Emanuel County specifically, survival outcomes may be improved by increasing the number of EMS professionals or by implementing professional emergency response training programs to raise the likelihood of pa-

Table 4.2: Counterfactual examples for treatment variable being using AED or not

	Case 1		Case 2	
Treatment **(AED use or not)**	0 (no)	1 (yes)	0 (no)	1 (yes)
gender	male	male	male	male
cpr use or not	yes	yes	yes	yes
etiology	other	other	other	other
location type	establishment	establishment	residence	residence
race	White	White	African American	African American
CPR type	2 types	2 types	1 types	1 type
acuity	emergent	emergent	critical	critical
witness	not witnessed	not witnessed	professional	professional
who use cpr	professional	professional	family	family
age	48	48	59	59
duration (seconds)	436	436	679	679
Latitude	32.01485	32.01485	33.64338	33.64338
Longitude	-81.11279	-81.11279	-84.0096	-84.0096
groudtruth y	0	Not applicable	Not applicable	1
ITE	0.614805		0.498918	

[*] For Case 1, the ground truth corresponds to a scenario where the treatment is 0 (AED not used) and the survival outcome is 0 (expired). For Case 2, the ground truth represents a scenario where the treatment is 1 (AED used) and the survival outcome is 1 (survived). All other columns, except for these two, represent counterfactual examples.

tients being witnessed by trained personnel. In Bulloch County, enhancing CPR training programs could increase the probability of CPR being administered by trained individuals, thereby contributing to improved survival rates.

Baselines

In this work, we propose SpatialCausal, a new method for causal inference that we compare against several established baselines, including S-learner(Salditt, Eckes, and Nestler 2024), T-learner(Künzel et al. 2019; Salditt, Eckes, and Nestler 2024), X-learner(Künzel et al. 2019; Salditt, Eckes, and Nestler 2024), R-learner(Nie and Wager 2021), and Dragonnet(C. Shi, Blei, and Veitch 2019). The first four are known as meta-learners(Künzel et al. 2019)—they combine multiple machine learning algorithms to estimate ATE across different subgroups. The S-learner(Salditt, Eckes, and Nestler 2024) incorporates the treatment variable as an additional feature within a single model, whereas the T-learner(Künzel et al. 2019; Salditt, Eckes, and Nestler 2024) fits separate models for treatment and control groups. The X-learner is particularly useful in scenarios where treatment groups are imbalanced(Künzel et al. 2019; Salditt, Eckes, and Nestler 2024), as it leverages information from both groups to iteratively refine individual treatment-effect estimates. The R-learner(Nie and Wager 2021) first generates out-of-fold predictions for both outcomes and treatment assignments, then regresses outcome residuals on treatment residuals to isolate the causal effect. Finally, Dragonnet(C. Shi, Blei, and Veitch 2019) employs a deep learning architecture that integrates representation learning with causal inference techniques, aiming to produce more accurate and less biased estimates of treatment

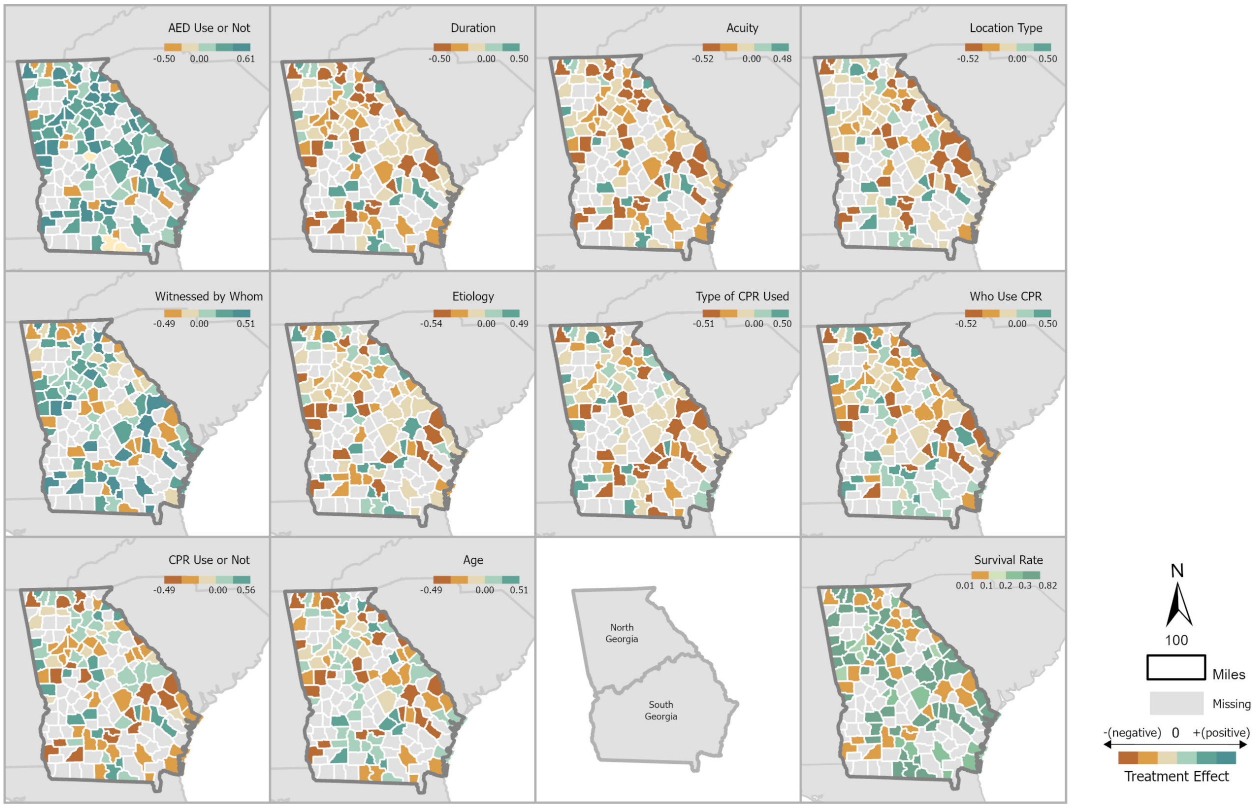


Figure 4.6: Spatially-Varying Causal Effect Across County Level

effects from observational data. By comparing SpatialCausal with these diverse baselines, we aim to showcase its effectiveness and robustness in capturing treatment heterogeneity.

Our baseline experiments employ meta-learners that integrate various machine learning models, including Gradient Boosting Decision Trees (GBDT), eXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Multi-Layer Perceptron (MLP). Along with Dragonnet, all these models use coordinates as two separate columns during training. We maintain the same train-test split as SpatialCausal and use grid search for hyperparameter tuning when testing these models. Note that different treatment variables yield distinct models; the corresponding results are presented in Table 4.3. Our model, SpatialCausal, achieves the highest overall average performance with an AUCROC score of 0.636, while the second-best score, 0.623, is from the R-learner with integrated GBDT. Moreover, as illustrated in the Figure 4.7, our model demonstrates more stable performance across 10 different treatment variable models compared to the others, highlighting its superior robustness.

Ablation studies

Our ablation studies consist of three components. First, we compare the performance of the plain causal model and the spatially-aware causal model, demonstrating the benefits of incorporating spatial information into our approach.

Table 4.3: Meta-learners and Dragonnet baselines compared with SpatialCausal

Models		**Treatment Variables**										Overall Avg. Perf.	
		AED use or not	**location type**	**who use cpr**	**duration**	**etiology**	**CPR use or not**	**witness**	**acuity**	**CPR type**	**age**		
S-Learner	*GBDT*	0.544	0.520	0.433	0.556	0.447	0.459	**0.694**	0.460	0.440	0.611	0.517	0.5031
	XGB	0.562	0.457	0.521	0.378	0.497	0.368	0.663	0.500	0.283	0.567	0.480	
	RF	0.169	0.543	0.538	0.697	0.650	0.536	0.586	0.832	0.397	0.511	0.546	
	MLP	0.649	0.385	0.363	0.378	0.361	0.612	0.622	0.349	0.349	0.635	0.471	
T-Learner	*GBDT*	0.246	0.643	0.541	0.687	0.566	0.901	0.322	**0.970**	0.499	0.543	0.592	0.5705
	XGB	0.584	0.709	0.556	0.676	0.292	**0.955**	0.277	0.713	0.516	0.515	0.579	
	RF	0.244	0.603	0.580	0.707	0.652	0.685	0.508	0.774	0.457	0.531	0.574	
	MLP	0.461	0.638	0.415	0.583	0.618	0.668	0.534	0.634	0.388	0.432	0.537	
X-learner	*GBDT*	0.556	0.503	0.474	0.501	0.459	0.414	0.615	0.468	0.427	0.485	0.490	0.4946
	XGB	0.597	0.417	0.503	0.512	0.445	0.465	0.619	0.487	0.434	0.517	0.500	
	RF	0.502	0.533	0.447	0.473	0.459	0.466	0.523	0.533	0.442	0.517	0.489	
	MLP	0.396	0.559	0.497	0.520	0.436	0.464	0.555	0.515	0.408	**0.643**	0.499	
R-learner	*GBDT*	0.094	0.810	0.558	0.746	**0.695**	0.829	0.576	0.828	0.591	0.502	0.623	0.5922
	XGB	0.496	0.562	0.467	0.671	0.499	0.641	0.583	0.628	0.464	0.516	0.553	
	RF	0.172	0.741	0.511	0.558	0.626	0.745	0.598	0.879	0.560	0.516	0.591	
	MLP	0.107	**0.820**	0.522	**0.757**	0.685	0.733	0.585	0.726	0.580	0.513	0.603	
Dragonnet		0.651	0.623	0.564	0.607	0.562	0.578	0.586	0.590	0.575	0.651	0.598	
SpatialCausal		**0.672**	0.674	**0.649**	0.654	0.637	0.628	0.623	0.623	**0.622**	0.578	**0.636**	

*Numbers are AUCROC score. Different treatment variables yield distinct models.

Second, we examine how the dimensionality of the learned hidden state affects the results to ensure optimal model configuration. Third, we assess the effect of freezing versus not freezing the spatial information confounding model on the outcomes.

Spatial information VS Non-spatial information The plain model refers to the causal model that does not incorporate spatial information. In contrast, the spatially-aware model integrates spatial information into the causal inference training process, capturing potential spatial effects. For both causal models, we treat each variable as a binary treatment separately, while all other variables—including spatial and non-spatial variables—are considered confounders. Since race and gender are not actionable treatment variables in causal inference, they are excluded from the treatment variables. Consequently, the number of causal models trained is equal to the number of treatment variables. As shown in Table 4.4, we trained a total of ten causal inference models, each evaluated using AUC-ROC(Area Under the Receiver Operating Characteristic Curve) and AUC-PR(Area Under the Precision-Recall Curve). We then average these performance metrics across the ten models for both the plain and spatially-aware causal models to obtain the final aggregated performance. Our results indicate that, across individual treatment models, our spatially-aware causal model consistently outperforms the plain model. Additionally, when considering the final average performance, the spatially-aware causal model demonstrates superior results across both evaluation metrics. Specifically, the AUC-ROC of the spatially-aware model improves by 15.04% compared to the plain model, highlighting the effectiveness of incorporating spatial information into causal inference.

Different dimensions for learned hidden state Z The learned hidden state Z captures both health and spatial variable information. We hypothesize that its dimensionality influences model performance. To determine the optimal configuration, we conducted experiments with dimensions set to 10, 20, 30, 40, and 50. As shown in Table 4.5, the highest AUCROC performance 0.636 is achieved when the dimension is 20. Notably, six out of ten variable models attained peak performance at this setting, and the results were more stable across all variable models compared with other dimension settings, underscoring the robustness of this configuration.

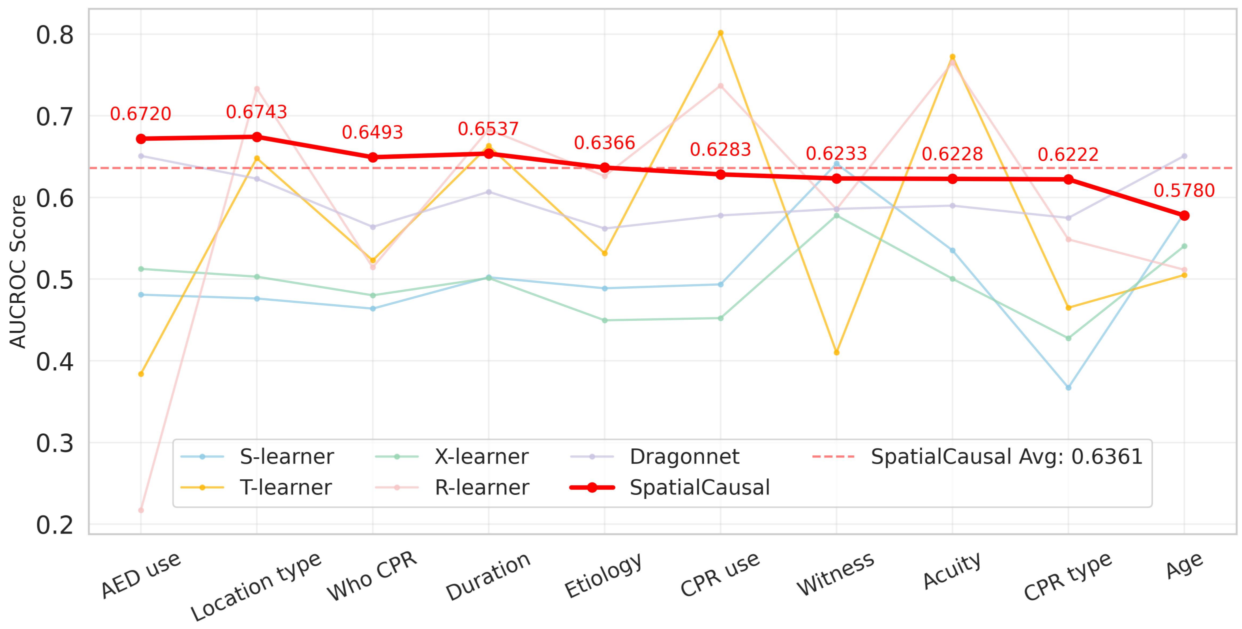


Figure 4.7: Comparison between Causal Inference Models in Stability and Robustness

Table 4.4: Comparison between plain and spatially-aware causal inference model

Treatment Variable	**Plain Causal Inference**			**Spatially-Aware Causal Inference**		
	AUC ROC	**AUC PR**	**ATE**	**AUC ROC**	**AUC PR**	**ATE**
AED use or not	0.5643	0.3214	0.2195	0.6719	0.3997	0.3048
location type	0.5574	0.3033	-0.1834	0.6743	0.4199	-0.1651
who use cpr	0.5332	0.2812	-0.0556	0.6492	0.4229	-0.0759
duration	0.5034	0.2878	-0.1833	0.6537	0.4091	-0.1612
etiology	0.5494	0.3187	-0.1298	0.6366	0.4162	-0.1363
CPR use or not	0.5270	0.3008	-0.1466	0.6282	0.3838	-0.0516
witness	0.5095	0.3037	0.18873	0.6232	0.3892	0.1425
acuity	0.5669	0.29817	-0.222	0.6228	0.3662	-0.1943
CPR type	0.5667	0.3350	-0.0969	0.6222	0.3899	-0.0770
age	0.5006	0.2860	0.01224	0.5780	0.3726	-0.0143
Average	0.5378	0.3036		**0.6360**	**0.3970**	

Freeze spatial information confounding Model VS Not freeze The spatial information confounding model utilizes a GeoCLIP location encoder, which employs an equal Earth projection combined with positional encoding using random Fourier features. This process transforms two-dimensional coordinates ($G_m \in R^2, \forall m \in [1, \ldots, M]$) into a high-dimensional location embedding $R^D$where $D \gg 2$. We evaluate two scenarios: one in which the GeoCLIP location encoder is trained jointly with the causal inference model, and another in which it

Table 4.5: Impact of Learned Hidden State Dimensionality on SpatialCausal Performance

Z dim	Treatment Variables										Avg. Perf.
	AED use or not	location type	who use cpr	duration	etiology	CPR use or not	witness	acuity	CPR type	age	
10	0.692	0.620	0.614	0.630	0.549	0.643	0.544	0.603	0.605	0.517	0.602
20	0.672	**0.674**	**0.649**	**0.654**	0.637	0.628	**0.623**	**0.623**	0.622	**0.578**	**0.636**
30	**0.704**	0.612	0.636	0.541	0.638	**0.648**	0.588	0.554	0.612	0.519	0.605
40	0.692	0.667	0.635	0.508	**0.640**	0.640	0.515	0.603	**0.631**	0.510	0.604
50	0.695	0.660	0.531	0.519	0.567	0.647	0.555	0.562	0.630	0.494	0.586

*Numbers are AUCROC score. Different treatment variables yield distinct models.

is kept frozen. As shown in Table 4.6, the encoder trained with the causal inference model demonstrates superior performance compared to the frozen scenario.

Table 4.6: Comparison between Freezing Versus Non-Freezing in the Spatial Confounder Representation Model

Status of Spa. Info. Con. Model	AED use or not	location type	who use cpr	duration	etiology	CPR use or not	witness	acuity	CPR type	age	Avg. Perf.
Freeze	0.612	0.567	0.626	0.609	0.549	0.549	0.546	0.470	0.606	0.487	0.562
Train	0.672	0.674	0.649	0.654	0.637	0.628	0.623	0.623	0.622	0.578	**0.636**

*Numbers are AUCROC score. Different treatment variables yield distinct models.

4.7 Discussion

Health Policy Recommendation The findings indicate that the use of automated external defibrillators (AEDs) before the arrival of emergency medical services (EMS) has a positive impact on survival rates in most counties across Georgia. To maximize the benefits of AED availability, policymakers should prioritize the strategic placement of AEDs in public locations and integrate AED accessibility into broader community health initiatives. Special attention should be given to areas where data suggests a significant survival benefit, ensuring that AEDs are readily available in high-risk locations.

However, in some counties where AED use does not yield a strong positive effect, alternative strategies should be considered. Evidence suggests that in certain regions, cardiopulmonary resuscitation (CPR) has a greater impact on survival outcomes than AED use. This underscores the critical role of trained professionals in administering CPR effectively. In the Atlanta region, as indicated by the witnessed by whom map, patients who experience out-of-hospital cardiac arrest (OHCA) and are witnessed by professionals have a higher likelihood of survival compared to those witnessed by laypersons. Similarly, in southern counties, as shown on the who performed CPR map, the availability of professionals trained in CPR is a key factor in improving survival rates. To enhance survival outcomes in these regions, targeted recruitment and training efforts should be implemented to expand the pool of healthcare workers and first responders equipped with CPR expertise. Strengthening emergency medical training programs in these areas could significantly improve the overall effectiveness of pre-hospital interventions.

Additionally, the type of CPR used map suggests that simplifying CPR procedures across most counties could further improve survival rates. Raising awareness about appropriate CPR technique selection through school curricula, public health campaigns, and community-based outreach programs can enhance the likelihood that bystanders and responders apply the most effective CPR methods when needed. Standardizing CPR protocols and emphasizing hands-only CPR training for laypersons could further increase bystander intervention rates and improve patient outcomes.

Advantages of our model Our proposed **SpatialCausal** model introduces key advancements in spatially-aware causal inference for health geography. First, it addresses the challenge of unmeasured confounders, a common limitation in traditional causal inference methods. Real-world health data often contain hidden biases from factors such as socio-economic status, healthcare accessibility, and environmental influences. To mitigate these biases, our model estimates latent confounders using variational autoencoders (VAEs) and a maximum-likelihood approximation approach, improving the reliability of causal effect estimation. Second, *SpatialCausal* enables spatially varying causal effect estimation by explicitly incorporating spatial dependencies into the causal inference process. By leveraging spatial representation learning, our approach captures spatial information and integrates it into the analysis. Unlike conventional methods that assume uniform treatment effects across locations, our model accounts for geographical variations, providing a more nuanced understanding of how treatment effects differ across regions. Additionally, our model provides actionable insights for public health policy and intervention by allowing causal effect estimation at both individual and geographic levels. This enables policymakers and healthcare professionals to identify areas where interventions—such as improving AED accessibility or reducing EMS response time—would be most effective in improving survival outcomes. By capturing spatial variations in treatment effects, our approach supports the development of targeted public health strategies aimed at reducing disparities in health outcomes. Finally, *SpatialCausal* advances the field of spatial causal inference by integrating spatial dependencies and latent confounders within a potential outcome framework, addressing limitations in existing spatial models that focus primarily on pattern analysis without causal inference. This methodological contribution provides a foundation for future studies on spatially varying causal effects, offering a more robust and interpretable approach for understanding geographical health disparities and guiding effective interventions.

Limitations and Future Works Although our method has shown improved performance and effectiveness in estimating causal effect, we recognize that there is still room for further exploration. For example, despite being a complete dataset spanning the years from 2019 to 2021, the data is not able to cover all geographical areas and data space, this gives limitations to the quality of the causal effect for certain counties. Additionally, the input variables for our model are limited to those related to OHCA incidents, while individual demographic variables are omitted due to privacy policies. However, these features could potentially impact health outcomes. Evaluating our model on datasets that include such information would provide a more complete picture, helping to identify significant features of health outcomes.

4.8 Conclusion

This study introduces *SpatialCausal*, a spatially-aware causal inference framework designed to address the dual challenges of incorporating spatial information and accounting for unmeasured confounders in health outcome analysis. By leveraging spatial representation learning, our method effectively integrates geographic information into causal effect estimation, while variational autoencoders enable the identification of latent confounders, reducing bias in causal inference. By providing more accurate causal estimates, *SpatialCausal* enables data-driven decision-making in public health, offering tailored intervention strategies for diverse geographic regions. As one of the first studies to apply a potential outcomes framework to spatial causal inference, this work establishes a foundation for future research on understanding and addressing health disparities through spatially-aware causal analysis.

Chapter 5

Conclusion

In this dissertation, we developed a comprehensive framework aimed at improving OHCA survival rates in Georgia through the integration of practical AED placement optimization and data-driven spatial analysis using machine learning. Given the currently low AED coverage rate 29.7% in Georgia, our framework addresses this critical gap by enhancing AED accessibility with the development of an *OSTO* method that accounts for the spatial-temporal distribution of potential OHCA incidents. Furthermore, we introduced a spatial counterfactual deep learning approach *SpaCE* designed to identify key risk factors and generate 'what if' scenarios for individuals at risk, thus providing actionable insights for intervention. Finally, to further uncover the causal mechanism behind the complex relationship, we developed *Spatial-Causal*, a spatially-aware causal inference model, and estimate the spatially-varying causal effect between the identified risk factors and health outcomes at both individual level and county level. Together, these components constitute a comprehensive system that not only delineates the magnitude and direction of risk factors on health outcomes but also offers a reference model for future research. The conclusions of each chapter are as follows, which explicitly highlights the contributions and implications of our work.

In chapter 2, we addressed certain limitations of current AHA guidelines for AED deployment, which often lack the detail necessary for dynamic, spatiotemporal scenarios. By framing AED placement as a location optimization problem under budget and resource constraints, we introduced the Overlayed Spatio-Temporal Optimization (OSTO) method. This innovative approach accounts for the heterogeneous distribution of potential OHCA incidents over time and space, providing a more tailored and effective strategy for AED placement. We validated the OSTO method using anonymized mobile device location data from Washington, DC, and the results demonstrated a significant improvement in AED coverage through systematic, optimization-based planning. Our cost-coverage increment analysis further revealed that merely increasing the number of AEDs does not guarantee proportional gains in coverage; rather, an optimal balance between budget limitations and coverage must be achieved. Moreover, the flexibility and adaptability of the OSTO framework make it highly transferable to other healthcare facility deployment tasks facing similar constraints.

In chapter 3, we introduced and validated the Spatial Counterfactual Explainable Deep Learning (SpaCE) model, an approach that integrates geospatial and health data to predict health outcomes and generate interpretable, actionable insights. By combining a spatially explicit health outcome predictor (SEP) with a tailored Prototype-guided Counterfactual Explanation (PCE) algorithm, SpaCE effectively uncovers the complex interplay between risk factors, geospatial patterns, and disease outcomes. Our evaluation on OHCA survival outcomes in Georgia demonstrated that the SpaCE model not only outperforms state-of-the-art baseline models—with a 10.2% improvement reflected in a 0.682 AUCROC score—but also reveals the significant influence of geospatial context on risk factors. The counterfactual examples generated by the model offer valuable insights, enabling us to quantify the impact of each variable at both individual and county levels and to propose targeted intervention strategies. Overall, the SpaCE framework enhances both predictive accuracy and model explainability, providing a versatile tool for health outcome prediction and intervention planning. Its ability to seamlessly integrate spatial information with health variables marks an advancement in health geography research, with promising applications across diverse health application contexts.

In chapter 4, we aim to uncover the mechanism between the risk factors we identified in chapter 3 and OHCA health outcome, this chapter partially addressed the critical challenge of estimating spatially varying causal effects by introducing the SpatialCausal framework—a spatially-aware causal inference method that integrates geographic information while effectively adjusting for both measured and unmeasured confounders. By leveraging spatial representation learning and variational autoencoder-based causal model, SpatialCausal overcomes the dual challenges of incorporating spatial information and identifying latent confounders, thus reducing bias in causal effect estimation. Our application of SpatialCausal to out-of-hospital cardiac arrest survival outcomes not only achieved an improved performance over baseline approaches but also underscored the significant impact of geospatial context on health outcomes. This enhanced accuracy in causal estimation enables more targeted, data-driven public health interventions tailored to the needs of diverse geographic regions. As one of the first studies to employ a potential outcomes framework for spatial causal inference, this work lays a robust foundation for future research aimed at understanding and addressing health disparities through spatially-aware causal analysis. The versatility and adaptability of SpatialCausal make it a valuable tool for both researchers and policymakers seeking to refine intervention strategies and improve overall health outcomes in various geospatial contexts.

Comparing the risk factors identified in Chapters 3 and 4 reveals that while the magnitudes of the risk factors differ at the whole dataset level, their directional influences on OHCA outcomes remain almost consistent across Chapters 3 and 4. In the SpaCE model results from Chapter 3, the variables witnessed by whom, types of CPR used, and AED usage show the top three absolute effects on OHCA health outcomes. But in the chapter 4 SpatialCausal analysis ranks AED usage as the most influential factor, followed by duration and acuity, with witnessed by whom ranking fifth rather than third. This discrepancy is understandable considering the model architecture design, as the SpaCE model primarily explores correlations rather than direct causal relationships. Additionally, during the causal analysis, treatments are treated as binary variables, and continuous variables are reclassified accordingly, which can contribute to these minor differences. Ultimately, our goal is to eliminate spurious correlations—considered confounders in the causal model—and to emphasize the true causal relationships. Therefore, the final conclusions are based on the causal results presented in Chapter 4.

Overall, these chapters demonstrate a comprehensive approach to addressing the multifaceted challenges in AED deployment, health outcome prediction, and causal inference. By integrating spatial information into every stage—from optimization to explanation to causal estimation—our work provides a versatile and robust framework that enhances our understanding of how to reposition AEDs based on the dynamic spatial and temporal distribution of at-risk populations, as well as how identified risk factors influence health outcomes at both individual and geographical levels. Furthermore, this adaptable framework holds promise for application in other similar geospatial health research areas.