Environmental and Ecological Statistics
https://doi.org/10.1007/s10651-026-00754-1

Nonparametric proportional hazards model with differential regularization applied to spatial survival data

Lorenzo Tedesco[1] · **Francesco Finazzi**[1]

Received: 3 March 2026 / Revised: 8 June 2026 / Accepted: 20 July 2026

Abstract
The proportional hazards (PH) model is one of the most widely used models in survival analysis, typically assuming a log-linear relationship between covariates and the hazard function. However, in the context of spatial survival data, where the time-to-event variable is associated with a spatial location within a given domain, this assumption is often unrealistic in capturing spatial effects. Thus, this paper proposes modeling the location effect through a nonparametric function of spatial location. The function is approximated using finite element methods on a triangulated mesh to accommodate irregular domains. Estimation is carried out within the classical partial likelihood framework, with smoothness of the spatial effect enforced through differential penalization. Using sieve methods, we establish the consistency and asymptotic normality of the parametric component. Simulations and two empirical applications demonstrate superior performance compared to existing approaches.

Keywords Differential regularization · Nonparametric statistics · Proportional hazards models · Spatial statistics · Survival analysis

1 Introduction

Since its introduction by Cox (1972), the proportional hazards (PH) model has become a fundamental tool in survival analysis. Central to this framework is the hazard function, which quantifies the instantaneous risk of an event occurring at a particular time, given that the subject has not yet experienced the event up to that point.

Handling Editor: Luiz Duczmal.

✉ Lorenzo Tedesco
lorenzo.tedesco@unibg.it

[1] Department of Economics, University of Bergamo, Via dei Caniana 2, 24127 Bergamo, Italy

Published online: 07 August 2026

Within the PH framework, the hazard is expressed as the product of two parts: a baseline hazard, which is the same for all individuals, and a term that incorporates the effects of covariates. The standard version of the model assumes that the covariates have a log-linear effect, meaning the logarithm of the hazard is a linear combination of the covariates. This setup provides a convenient interpretation of the regression coefficients as hazard ratios: each coefficient reflects the multiplicative change in risk associated with a one-unit increase in the corresponding covariate. These coefficients can be estimated efficiently using partial likelihood, which focuses on the ordering of event times rather than their exact distribution. This approach has the important advantage of profiling out the baseline hazard, allowing it to remain unspecified while still yielding consistent and efficient estimates of the covariate effects. As a result, inference can be conducted on hazard ratios without making parametric assumptions about the underlying hazard function.

The assumption of log-linearity, however, can be too restrictive in practice. A more flexible alternative is the nonparametric proportional hazards model, which allows some covariate effects to be modeled in a nonparametric way rather than forcing a purely linear form. In this formulation, the hazard depends on two sets of covariates: one that enters the model linearly, with a finite set of regression coefficients, and another that enters through an unspecified smooth function.

Estimation approaches for the PH model that incorporate nonparametric covariate effects through smooth functions fall into three main categories. The first class consists of local likelihood methods (Tibshirani and Hastie 1987; Fan et al. 1997; Chen et al. 2010), which rely on kernel smoothing and the careful selection of bandwidth parameters. The second widely used class consists of spline-based approaches (O'Sullivan 1988; Hastie and Tibshirani 1990; Kooperberg et al. 1995), in which the smooth function is approximated using a basis expansion. This formulation enables flexible modeling while preserving parsimony and interpretability. The third class consists of deep learning extensions of survival models. See, for instance, Nagpal et al. (2021); Katzman et al. (2018) and Wiegrebe et al. (2024) for a complete review. These methods have demonstrated strong predictive performance in high-dimensional and nonlinear settings. However, they may also involve substantial computational complexity, reduced interpretability, and limited availability of classical asymptotic theory for the resulting estimators.

In this paper, we focus on the case where the nonparametric smooth function reflects a spatial effect in a bidimensional domain. In such contexts, standard nonparametric approaches fail to adequately represent spatial variation in the hazard. In fact, both local likelihood methods and spline-based approaches are primarily designed for univariate smoothing, and their direct extensions to two-dimensional spatial domains often suffer from instability, boundary bias, or prohibitive computational cost. Deep learning extensions, while highly flexible, typically do not explicitly incorporate the geometry of the spatial domain, which may limit their ability to represent spatial variation over irregular regions. Consequently, these methods may fail to capture complex spatial structures in the hazard function, highlighting the need for a spatially aware nonparametric extensions of the PH model.

For a review of spatial PH models, we refer the reader to Hanson and Zhou (2014). We restrict our discussion to the main and more recent contributions.

Early contributions to spatial PH modeling focused on extending the standard framework to accommodate spatial dependence between observations through joint distributional assumptions. An example is Li and Lin (2006), where the authors propose a semiparametric normal transformation model for spatial survival data. In this framework, observations marginally follow a PH model, while their joint distribution is defined by transforming the data into approximately normal variables and assuming a multivariate normal distribution for the transformed outcomes. A key limitation of this approach is that, if the transformation is misspecified, the assumption of multivariate normality may be violated, potentially leading to biased inference. Notably, this model was originally motivated by the claim that "direct nonparametric maximum likelihood estimation in such models is practically prohibited due to the high-dimensional, intractable integration in the likelihood function and the infinite-dimensional nuisance baseline hazard parameter." This highlights the importance of basing the estimation on the partial likelihood, which, as in the standard PH model, treats the infinite-dimensional baseline hazard as a nuisance parameter, rather than attempting estimation from the full likelihood. Following this principle, the proposed model assumes independence among observations, incorporates a nonparametric spatial effect into the hazard function to capture spatial effects, and employs the partial likelihood for estimation, thereby showing that a nonparametric estimation can be carried out in a straightforward manner.

Recent contributions to spatial survival modeling have developed along several directions. Within the frequentist framework, Xue et al. (2020) introduce a geographically weighted Cox regression that incorporates location through pairwise spatial kernels. Thus, this can be seen as a spatially aware local likelihood method. A similar work is Hu and Huffer (2020) adapted for the Kaplan–Meier and Nelson–Aalen estimators.

Another spatial PH alternative is the composite likelihood approach of Paik and Ying (2012), which assumes the Farlie–Gumbel–Morgenstern distribution and models the dependence parameter as a function of geographic and demographic pairwise distances. Apart from the restrictive dependence structure, composite likelihood methods—while computationally convenient—can be less efficient than full likelihood approaches, as they ignore higher-order dependencies.

Bayesian formulations have also been proposed, such as Banerjee et al. (2003); Taylor and Rowlingson (2017), which partition the domain into clusters and assign random effects. A more recent version based on a geographically weighted Chinese restaurant process prior is proposed in Geng and Hu (2022), while Zhou and Hanson (2018) provide a unified framework for arbitrarily censored spatial survival data implemented in `spBayesSurv` and Zhang et al. (2023) proposed a method that can be adapted also for nonlinear covariate effects. In Hennerfeind et al. (2006), instead, a nonparametric spatial effect is modeled using splines, where spatial dependence is incorporated through the choice of priors on the spline parameters. These methods, however, rely on a parametric specification of the baseline hazard, thereby overlooking a key advantage of leaving the baseline hazard as a nuisance parameter.

Finally, a limitation common to all the cited spatial PH models is the assumption that spatial dependence is driven solely by Euclidean distance, thereby neglecting the geometry of the domain. This simplification can lead to biased results in settings with irregular boundaries, nonconvex shapes, or internal holes.

In this work, we introduce a spatial PH model that remains within the classical partial likelihood framework. We model a continuous, nonparametric spatial effect over a two-dimensional domain, enabling a smooth representation of spatial variation in the hazard. The method is implemented via finite element methods (FEM) on a mesh constructed from a triangulation of the domain of interest, ensuring accurate representation even for highly irregular domains. Spatial variation in survival data can be modeled through geostatistical approaches, which rely on continuous coordinates (e.g., latitude and longitude), or through lattice-based approaches, which model dependence among discrete spatial units. Our method can accommodate both, and while focusing on geostatistical data, we also describe how it naturally extends to areal data, see Sect. A of the Appendix.

The smoothness of the spatial effect is enforced through a differential penalization term, which yields a concave maximization problem for a high-dimensional, differentiable objective function. This formulation enables efficient estimation using derivative-based optimization methods, facilitated by the fact that the derivative is available in closed form and simple to evaluate. A related penalization strategy for spatial regression is discussed in Sangalli et al. (2013), although it cannot be readily extended to survival data. Another distinction is that our estimation procedure is developed within the framework of sieve theory, which enables the derivation of robust asymptotic properties. Thus, our approach could also serve as a further step toward achieving asymptotic results in spatial regression with differential regularization.

The paper is organized as follows. In Sect. 2, we specify the model. In Sect. 3, we develop the estimation procedure based on penalized likelihood and the sieve method with finite element approximations. Section 4 establishes the asymptotic properties of the proposed estimator. Section 5 presents simulation results illustrating finite-sample performance, while Sect. 6 provides empirical applications. Section 7 contains our concluding remarks. Relevant proofs are in the Appendix.

2 Model

Let T be a nonnegative time-to-event random variable representing the occurrence of the event of interest. We consider a compact spatial domain with nonempty interior $\Omega \subset \mathbb{R}^2$ with a regular boundary $\partial\Omega \in C^2(\mathbb{R}^2)$. Let $\boldsymbol{X} \in \mathbb{R}^b$ denote a b-dimensional vector of covariates, and let $\boldsymbol{P}$ be a vector taking values in Ω, representing the spatial location at which an observation is made. We assume that, conditional on the values $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{P} = \boldsymbol{p}$, the hazard function of T follows the PH model

$$\lambda(t \mid \boldsymbol{x}, \boldsymbol{p}) = \lambda_0(t) \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}_0 + h_0(\boldsymbol{p})\right), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function shared across all individuals; $\boldsymbol{\beta}_0 \in \mathbb{R}^b$ is a vector of regression coefficients that quantifies the log-linear effects of the covariates $\boldsymbol{X}$ on the hazard; and $h_0 : \Omega \to \mathbb{R}$ is an unknown smooth function that captures residual spatial variation in the hazard associated with location $\boldsymbol{P}$, that is not explained by the observed covariates.

As in the standard interpretation of the proportional hazards model, the coefficient β_{0j} represents the log-hazard ratio associated with a one-unit increase in the j-th covariate, holding all other covariates and the spatial effect constant. The spatial term $h_0(\boldsymbol{p})$ serves to identify regions of elevated or reduced hazard relative to the baseline $\lambda_0(t)$. Since h_0 is identifiable only up to an additive constant, we impose the centering condition

$$\int_{\Omega} h_0(\boldsymbol{p})\, \boldsymbol{dp} = 0, \tag{2}$$

which guarantees uniqueness and will prove to be computationally convenient.

Because of right-censoring, it is not possible to directly observe T. Instead, we observe

$$Y = \min(T, C), \quad \delta = I(T \leq C),$$

where C denotes a censoring variable and $I(\cdot)$ is the indicator function. We consider the following assumption.

Assumption 1 (i) Conditional on $\boldsymbol{X}$ and $\boldsymbol{P}$, the censoring variable C is independent of T. (ii) There exists $\tau > 0$ such that $\mathbb{P}(T \geq \tau) > 0$ and $\mathbb{P}(C \geq \tau) > 0$, ensuring that the support of Y is nondegenerate. (iii) The true coefficient vector $\boldsymbol{\beta}_0$ is an interior point of a compact set $\mathcal{B} \subset \mathbb{R}^b$. (iv) The matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$ is positive definite. (v) The spatial effect h_0 is smooth and satisfies (2).

Assumption 1 (i) is standard in survival analysis and ensures unbiased estimation under right-censoring. Assumption 1 (ii) guarantees nondegenerate support of the observed data, ensuring sufficient follow-up for reliable inference. Assumption 1 (iii) imposes compactness of the parameter space, a technical condition that facilitates consistency of the estimator. Assumption 1 (iv) requires positive definiteness of $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]$, preventing collinearity and ensuring identifiability of regression effects. Assumption 1 (v) enforces smoothness and centering of the spatial effect, which provide identifiability and regularity for asymptotic analysis.

Theorem 1 *Suppose Assumption 1 holds. Then, the model in (1) is uniquely identified.*

The proof is reported in Appendix B.

3 Estimation via penalized likelihood and sieve method

In order to characterize the regression structure and capture spatial heterogeneity, we have developed an estimation procedure based on penalized likelihood. Specifically, we assume that the observed data arise from an independent and identically distributed sample of $(Y, \delta, \boldsymbol{X}, \boldsymbol{P})$ of size n, namely $\{(Y_i, \delta_i, \boldsymbol{x}_i, \mathbf{p}_i)\}_{i=1}^n$. Our objective is to estimate the regression coefficients $\boldsymbol{\beta}_0$ and the spatial effect h_0 by maximizing a penalized log-partial-likelihood functional Q_n defined on a suitable space $\mathcal{B} \times \mathcal{H}$ as

$$
\begin{aligned}
Q_n(\boldsymbol{\beta}, h) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \Big(\boldsymbol{x}_i^\top \boldsymbol{\beta} + h(\boldsymbol{p}_i) - \log \frac{1}{n} \sum_{j=1}^{n} I(Y_j \geq Y_i) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta} \\
+ h(\boldsymbol{p}_j)) \Big) - \frac{\lambda}{2} \int_{\Omega} (\Delta h(\boldsymbol{p}))^2 \, \mathrm{d}\boldsymbol{p}.
\end{aligned}
\tag{3}
$$

Here, λ is a positive smoothing parameter (potentially dependent on n), and Δ is the Laplacian operator, $\Delta h = \frac{\partial^2 h}{\partial p_1^2} + \frac{\partial^2 h}{\partial p_2^2}$, which measures the local curvature of the effect h. The penalty term controls the smoothness of the estimated function h, with a larger λ enforcing a smoother function.

The functional space $\mathcal{H}$ must be chosen carefully. Since the penalty term $\int_\Omega (\Delta h)^2$ must be well defined, we require that $\mathcal{H} \subset H^2(\Omega)$, the Sobolev space of functions in $L^2(\Omega)$ with all distributional derivatives up to order 2 also in $L^2(\Omega)$. By the Sobolev embedding theorem, $H^2(\Omega) \subset C^0(\Omega)$, ensuring that any $h \in H^2(\Omega)$ is continuous and can be evaluated pointwise at any location. To facilitate the estimation procedure, we consider homogeneous Neumann boundary conditions, meaning the flux across the boundary is zero: $\nabla h \cdot \mathbf{n} = 0$ on $\partial\Omega$, where $\mathbf{n}$ is the outward-pointing normal unit vector. Combining these requirements with the identifiability constraint, we consider the following functional parameter space:

$$
\mathcal{H} = \left\{ h \in H^2_{\mathbf{n}}(\Omega) \;\middle|\; \int_\Omega h \, dx = 0, \|h\|_{H^2(\Omega)} < M_{\mathcal{H}} \right\},
$$

where $H^2_{\mathbf{n}}(\Omega) = \left\{ h \in H^2(\Omega) \;\middle|\; \nabla h \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \right\}$, the norm $\| \cdot \|^2_{H^2(\Omega)}$ is defined as

$$
\|h\|^2_{H^2(\Omega)} = \sum_{|\alpha| \leq 2} \int_\Omega |D^\alpha h(\boldsymbol{p})|^2 \, d\boldsymbol{p},
$$

with $\alpha = (\alpha_1, \alpha_2)$ being a multi-index, $|\alpha| = \alpha_1 + \alpha_2$, and $D^\alpha h = \partial^{|\alpha|} h / \partial p_1^{\alpha_1} \partial p_2^{\alpha_2}$ denotes the corresponding distributional derivative, and $M_{\mathcal{H}}$ is a constant large enough. The bound $\|h\|_{H^2(\Omega)} < M_{\mathcal{H}}$ is a consequence of the smoothness of h_0 and the fact the Ω is compact.

Maximizing the functional in (3) is an infinite-dimensional optimization problem. To make it computationally tractable, we replace the infinite-dimensional space $\mathcal{H}$ with a sequence of finite-dimensional subspaces $\mathcal{H}_n$ that become dense in $\mathcal{H}$ as $n \to \infty$. This is the core idea of the method of sieves (Chen 2007). We use the finite element method to construct these subspaces.

3.1 The finite element method

The FEM is a versatile numerical technique for approximating differential operators. Its core idea is to construct a finite-dimensional approximation space, typically spanned by locally supported basis functions. These basis functions are defined with respect to a subdivision of the computational domain Ω into simpler subdomains (Quarteroni and Quarteroni 2009).

We consider a triangulation $\mathcal{T}_\eta$ of Ω, in which the domain is partitioned into nonoverlapping triangles such that any two adjacent triangles share either a complete edge or a single vertex. The parameter η denotes the mesh size, typically taken as the diameter of the largest triangle in the mesh. In this setting, the triangulation also serves to approximate the boundary $\partial\Omega$ by a polygon (or, more generally, a union of polygonal segments), ensuring that even curved boundaries are represented in a piecewise linear fashion.

3.1.1 Illustration with C^0 linear elements

From this triangulation, one often constructs a finite-dimensional function space spanned by locally supported basis functions. In the case of continuous, piecewise linear finite elements (C^0), each basis function ψ_k is associated with a specific vertex $\boldsymbol{\xi}_k$ of the triangulation. By definition, ψ_k takes the value 1 at its corresponding vertex $\boldsymbol{\xi}_k$ and 0 at all other vertices, i.e.,

$$\psi_k(\boldsymbol{\xi}_j) = \delta_{kj},$$

where δ_{kj} is the Kronecker delta. Figure 1 illustrates two such basis functions, ψ_i and ψ_j, associated with neighboring nodes of a planar mesh. Any function h in this C^0 finite element space can be expressed as

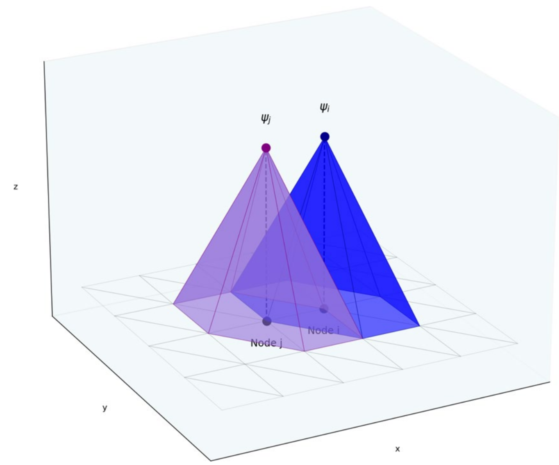


Fig. 1 Example of two linear finite element basis functions, ψ_i and ψ_j, associated with neighboring nodes on a planar mesh. Each function is locally supported, taking the value one at its corresponding vertex and zero at all other vertices

$$h(\boldsymbol{p}) = \sum_{k=1}^{K} c_k \, \psi_k(\boldsymbol{p}) = \mathbf{c}^\top \boldsymbol{\psi}(\boldsymbol{p}),$$

where K is the total number of vertices, and the coefficients c_k correspond to nodal values, $c_k = h(\boldsymbol{\xi}_k)$.

3.1.1.1 C^1 space used in our estimator While the C^0 FEM construction presented above is helpful for intuition, our penalization involves the Laplacian squared term $\int_\Omega (\Delta h)^2$, which is only well defined for $h \in H^2(\Omega)$. Therefore, for estimation we employ an H^2-conforming, C^1 finite element space (e.g., the Argyris element, see Brenner and Scott, 2008). For each triangulation $\mathcal{T}_\eta$ we define

$$V_\eta = \left\{ v \in C^1(\overline{\Omega}) \;\middle|\; v|_T \in \mathbb{P}_5(T) \text{ for every } T \in \mathcal{T}_\eta \right\},$$

where $\mathbb{P}_5(T)$ denotes the space of bivariate polynomials on the triangle T of total degree ≤ 5. The associated degrees of freedom (DOFs) enforce C^1 continuity across elements (values, first and second derivatives at vertices, and edge-midpoint normal derivatives), ensuring $V_\eta \subset H^2(\Omega)$.

Let $\{\psi_k\}_{k=1}^{K(\eta)}$ be the global C^1 basis associated with these DOFs, defined by Kronecker interpolation with respect to the DOF functionals. Any $h \in V_\eta$ can be written as

$$h(\boldsymbol{p}) \;=\; \sum_{k=1}^{K(\eta)} c_k \, \psi_k(\boldsymbol{p}) \;=\; \mathbf{c}^\top \boldsymbol{\psi}(\boldsymbol{p}),$$

where now $\mathbf{c}$ collects DOF values (not just nodal values).

3.2 Sieve space construction

To apply sieve theory, we define a sequence of triangulations and corresponding function spaces. Let $\{\eta(n)\}_{n\in\mathbb{N}} \subset (0, 1]$ be a decreasing sequence converging to zero. For each $\eta = \eta(n)$, we have a conforming triangulation $\mathcal{T}_\eta$ of the domain Ω satisfying

$$\max_{T\in\mathcal{T}_\eta} \operatorname{diam}(T) \leq \eta \; \operatorname{diam}(\Omega).$$

This family of triangulations is said to be *quasi-uniform*, meaning there exists a constant $\rho > 0$ such that for all η,

$$\min_{T\in\mathcal{T}_\eta} \operatorname{diam}(B_T) \geq \rho \, \eta \; \operatorname{diam}(\Omega),$$

where B_T is the largest ball contained in triangle T. This condition prevents triangles from becoming arbitrarily thin, ensuring good approximation properties. The discrete space corresponding to $H^2_{\mathbf{n}}(\Omega)$ is

$$H_n = V_{\eta(n)} \cap H^2_{\mathbf{n}}(\Omega).$$

Finally, the sieve space for our problem is the finite-dimensional approximation of $\mathcal{H}$:

$$\mathcal{H}_n = \left\{ h \in H_n \;\middle|\; \int_\Omega h\, d\boldsymbol{p} = 0,\ \|h\|_{H^2(\Omega)} < M_{\mathcal{H}} \right\}.$$

The quality of this approximation is guaranteed by standard C^1 FEM interpolation theory. For any $h \in \mathcal{H}$, we can define a projection $\mathcal{J}^n : \mathcal{H} \to \mathcal{H}_n$ as

$$\mathcal{J}^n h = \mathcal{I}^n h - \frac{1}{|\Omega|} \int_\Omega \mathcal{I}^n h\, d\boldsymbol{p},$$

where $\mathcal{I}^n$ is a C^1-conforming interpolation operator (e.g., the Argyris interpolant). Then (see, e.g., Brenner and Scott, 2008),

$$\|h - \mathcal{J}^n h\|_\infty \le C\, \eta(n)\, \|h\|_{H^2(\Omega)}, \tag{4}$$

for a constant C independent of h and $\eta(n)$. This shows that the approximation error vanishes at a rate $\mathcal{O}(\eta(n))$ in the L^∞-norm, confirming that $\mathcal{H}_n$ is a suitable sieve space.

The estimation problem is now restricted to the sieve space $\Theta_n = \mathcal{B} \times \mathcal{H}_n$. The estimator $\hat{\theta}_n = (\hat{\boldsymbol{\beta}}_n, \hat{h}_n)$ of $\theta_0 = (\boldsymbol{\beta}_0, h_0)$ is found by maximizing the penalized sample likelihood $Q_n(\theta)$ over the sieve space Θ_n:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta_n} Q_n(\theta). \tag{5}$$

3.3 Numerical implementation

Let $\boldsymbol{\psi}_n(\cdot) = (\psi_1(\cdot), \ldots, \psi_{K(n)}(\cdot))^\top$ be the vector of $K(n)$ nodal basis functions for the conforming C^1 finite element space V_n. Any $h \in \mathcal{H}_n$ can be written as

$$h(\boldsymbol{p}) = \boldsymbol{\psi}_n(\boldsymbol{p})^\top \mathbf{c}, \quad \mathbf{c} \in \mathbb{R}^{K(n)}.$$

Let the mass and stiffness matrices be

$$(\mathbf{R}_{0,n})_{ij} = \int_\Omega \psi_i \psi_j\, d\boldsymbol{p}, \qquad (\mathbf{R}_{1,n})_{ij} = \int_\Omega \nabla\psi_i \cdot \nabla\psi_j\, d\boldsymbol{p}.$$

The zero-mean constraint $\int_\Omega h\, d\boldsymbol{p} = 0$ can therefore be written as $\mathbf{1}^\top \mathbf{R}_{0,n}\mathbf{c} = 0$. Let $\mathbf{r}_0 = \mathbf{R}_{0,n}^\top \mathbf{1}$ and let $\mathbf{Z}_n \in \mathbb{R}^{K(n)\times(K(n)-1)}$ be an orthonormal basis of $\ker(\mathbf{r}_0^\top)$; then any feasible $\mathbf{c}$ can be written as $\mathbf{c} = \mathbf{Z}_n \mathbf{h}$, with $\mathbf{h} \in \mathbb{R}^{K(n)-1}$.

Introduce $g \in V_n$ as the L^2–projection of the (weak) Laplacian onto V_n, defined by

$$\int_{\Omega} g\, v\, d\boldsymbol{p} \;=\; \int_{\Omega} (\Delta h)\, v\, d\boldsymbol{p} \;=\; -\int_{\Omega} \nabla h \cdot \nabla v\, d\boldsymbol{p} \quad \forall v \in V_n,$$

where the boundary term vanishes by the homogeneous Neumann condition. Note that g is in general only an approximation of Δh, with equality holding when $\Delta h \in V_n$. Writing $g(\boldsymbol{p}) = \boldsymbol{\psi}_n(\boldsymbol{p})^\top \mathbf{d}$ yields the linear relation

$$\mathbf{R}_{0,n}\, \mathbf{d} \;=\; -\mathbf{R}_{1,n}\, \mathbf{c} \quad \Longrightarrow \quad \mathbf{d} \;=\; -\mathbf{R}_{0,n}^{-1}\mathbf{R}_{1,n}\, \mathbf{c}.$$

Hence the Laplacian squared penalty becomes

$$\int_{\Omega} (\Delta h)^2\, d\boldsymbol{p} \;\approx\; \|g\|^2_{L^2(\Omega)} = \mathbf{d}^\top \mathbf{R}_{0,n}\mathbf{d} = \mathbf{c}^\top \mathbf{R}_{1,n}\mathbf{R}_{0,n}^{-1}\mathbf{R}_{1,n}\mathbf{c}.$$

After the reparameterization $\mathbf{c} = \mathbf{Z}_n\mathbf{h}$, the estimator in (5) is obtained by maximizing

$$\hat{Q}(\boldsymbol{\beta}, \boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i\Big(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\psi}_n(\boldsymbol{P}_i)^\top \mathbf{Z}_n\mathbf{h} - \log S_n^{(0)}(\boldsymbol{\beta}, \mathbf{h}, Y_i)\Big) \;-\; \frac{\lambda}{2}\,\mathbf{h}^\top \mathbf{A}_n\mathbf{h}, \tag{6}$$

where

$$S_n^{(0)}(\boldsymbol{\beta}, \mathbf{h}, t) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq t)\, \exp\Big(\boldsymbol{x}_j^\top \boldsymbol{\beta} + \boldsymbol{\psi}_n(\boldsymbol{p}_j)^\top \mathbf{Z}_n\mathbf{h}\Big)$$

and

$$\mathbf{A}_n \;=\; \mathbf{Z}_n^\top\, \mathbf{R}_{1,n}\mathbf{R}_{0,n}^{-1}\mathbf{R}_{1,n}\, \mathbf{Z}_n.$$

The objective is differentiable and concave in $(\boldsymbol{\beta}, \mathbf{h})$, so derivative-based maximization applies. In fact we propose to use BFGS quasi-Newton algorithm (Nocedal and Wright 2006) with analytic gradients. For that, define

$$\eta_j = \boldsymbol{x}_j^\top \boldsymbol{\beta} + \boldsymbol{\psi}_n(\boldsymbol{p}_j)^\top \mathbf{Z}_n\mathbf{h}, \qquad \mathbf{r}_j = \mathbf{Z}_n^\top \boldsymbol{\psi}_n(\boldsymbol{p}_j).$$

For $t \geq 0$, set

$$S_{n,X}^{(1)}(\boldsymbol{\beta}, \mathbf{h}, t) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq t)\, e^{\eta_j}\, \boldsymbol{x}_j, \qquad S_{n,h}^{(1)}(\boldsymbol{\beta}, \mathbf{h}, t) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \geq t)\, e^{\eta_j}\, \mathbf{r}_j.$$

Then the gradient of the objective in (6) is

$$\nabla_{\boldsymbol{\beta}}\hat{Q}(\boldsymbol{\beta}, h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i\left(\boldsymbol{x}_i - \frac{S_{n,X}^{(1)}(\boldsymbol{\beta}, \mathbf{h}, Y_i)}{S_n^{(0)}(\boldsymbol{\beta}, \mathbf{h}, Y_i)}\right),$$

$$\nabla_{\mathbf{h}}\hat{Q}(\boldsymbol{\beta}, h) = \frac{1}{n}\sum_{i=1}^{n} \delta_i\left(\mathbf{r}_i - \frac{S_{n,h}^{(1)}(\boldsymbol{\beta}, \mathbf{h}, Y_i)}{S_n^{(0)}(\boldsymbol{\beta}, \mathbf{h}, Y_i)}\right) \;-\; \lambda\, \mathbf{A}_n\, \mathbf{h}.$$

The proposed estimator can therefore be written as

$$\hat{\theta}_n = \left(\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}, \;\; \hat{h}_n(\cdot) = \boldsymbol{\psi}_n(\cdot)^\top \mathbf{Z}_n \hat{\mathbf{h}}\right), \tag{7}$$

where $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{h}}) = \underset{(\boldsymbol{\beta},\mathbf{h})}{\arg\max}\, \hat{Q}(\boldsymbol{\beta}, \mathbf{h})$.

3.4 Selecting the tuning parameter

For the selection of the tuning parameter λ, we adopt a data-driven cross-validation strategy, as in Cygu et al. (2021). In this approach, the optimal value of λ is chosen to either minimize the cross-validated partial likelihood deviance (CV-PLD).

Using cross-validation, the training data are partitioned into K folds. For each $k = 1, \dots, K$, the model is fitted on the retained data (all but fold k) to obtain penalized estimates $(\hat{\boldsymbol{\beta}}_{-k}(\lambda), \hat{f}_{-k}(\lambda))$, which are then evaluated on the held-out fold. The CV-PLD is defined as

$$\text{CV-PLD}(\lambda) = -2 \sum_{k=1}^{K} \left\{ \ell_{\text{partial}}\big(\hat{\boldsymbol{\beta}}_{-k}(\lambda), \hat{h}_{-k}(\lambda); \mathcal{D}\big) - \ell_{\text{partial}}\big(\hat{\boldsymbol{\beta}}_{-k}(\lambda), \hat{h}_{-k}(\lambda); \mathcal{D}_{-k}\big) \right\}, \tag{8}$$

where $\ell_{\text{partial}}(\cdot;\mathcal{D})$ denotes the log-partial likelihood evaluated on the full dataset $\mathcal{D}$, that is

$$\ell_{\text{partial}}(\boldsymbol{\beta}, h; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \left(\boldsymbol{x}_i^\top \boldsymbol{\beta} + h(\boldsymbol{P}_i) - \log \frac{1}{n} \sum_{j=1}^{n} I(Y_j \geq Y_i) \exp\{\boldsymbol{x}_j^\top \boldsymbol{\beta} + h(\boldsymbol{P}_j)\} \right),$$

and $\ell_{\text{partial}}(\cdot;\mathcal{D}_{-k})$ is the same expression computed on the retained (training) data $\mathcal{D}_{-k}$. Subtracting the two ensures that only the contribution of the held-out fold is isolated while preserving the correct risk sets, which typically improves upon the simpler strategy of evaluating the partial likelihood solely on the held-out fold. The optimal value of λ is then the minimizer of (8).

Alternatively, the CV-C-index (Harrell et al. 1996) criterion aggregates the out-of-fold risk scores and computes Harrell's concordance index for the PH model (as implemented, for example, in the `survival` package; see Dai and Breheny, 2019). In this case, the optimal λ maximizes the cross-validated C-index.

4 Asymptotic analysis

We now turn to the asymptotic properties of the sieve estimator $\hat{\theta}_n$. Our approach follows the general framework of sieve M-estimation developed in Chen (2007). Let $d(\cdot,\cdot)$ denote a metric on the parameter space Θ, defined by

$$d(\theta, \tilde{\theta}) = \|\beta - \tilde{\beta}\| + \|h - \tilde{h}\|_\infty,$$

where $\theta = (\beta, h)$ and $\tilde{\theta} = (\tilde{\beta}, \tilde{h})$ are two generic elements of $\Theta = \mathcal{B} \times \mathcal{H}$, $\|\beta - \tilde{\beta}\|$ denotes the Euclidean norm of the difference between the finite-dimensional parameters $\beta, \tilde{\beta} \in \mathcal{B} \subset \mathbb{R}^b$, and $\|h - \tilde{h}\|_\infty = \sup_{\boldsymbol{p}\in\Omega} |h(\boldsymbol{p}) - \tilde{h}(\boldsymbol{p})|$ denotes the supremum norm of the difference between the functions $h, \tilde{h} \in \mathcal{H}$. This metric allows us to measure closeness in both the finite-dimensional and infinite-dimensional components of $\theta = (\beta, h)$.

As is customary in sieve methods, let $K(n)$ denote the dimension of the approximating space $\mathcal{H}_n$. Before presenting the main theorems, we introduce regularity conditions that ensure the sieve estimator is well behaved and facilitate the asymptotic derivations.

Assumption 2 (i) The support $\mathcal{X}$ of the covariates $\boldsymbol{X}$ is bounded, i.e., $\sup_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\| \le M_{\mathcal{X}}$. (ii) The triangulation $\mathcal{T}_\eta$, indexed by η, is assumed to be quasi-uniform.

The first step is to establish consistency of the sieve estimator. The following theorem shows that, under mild growth restrictions on $K(n)$ and the penalty parameter, the estimator converges in probability to the true parameter at a controlled rate.

Theorem 2 *Suppose that Assumptions 1, 2 hold, and that*

$$K(n) \log K(n) = o(n) \quad \text{and} \quad \lambda_n = o(1). \tag{9}$$

Define

$$\varepsilon_n \;=\; \max\left\{\, \delta_n,\; \eta(n),\; \sqrt{\lambda_n}\,\right\}, \qquad \delta_n \asymp \sqrt{\frac{K(n)\log K(n)}{n}}.$$

Then

$$d(\hat{\theta}_n, \theta_0) \;=\; O_P(\varepsilon_n).$$

The proof is reported in Appendix C. We now show that, with a suitable choice of sieve dimension, the parametric component $\hat{\boldsymbol{\beta}}_n$ is asymptotically normal.

Theorem 3 *Suppose assumptions of Theorem 2 hold, and* $\varepsilon_n^2 = o(n^{-1/2})$, $\lambda_n \epsilon_n = o(n^{-1/2})$ *and* $\eta(n)\epsilon_n = o(n^{-1/2})$. *Then, it holds*

$$\sqrt{n}\,\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \;\Rightarrow\; \mathcal{N}\!\left(0,\, \Sigma_\beta\right),$$

for a suitable covariance matrix Σ_β *specified in the proof.*

The proof is reported in Appendix D. Together, Theorems 2 and 3 establish that the sieve estimator is both consistent and asymptotically normal in its finite-dimensional component, providing the basis for inference on $\boldsymbol{\beta}_0$.

We now relate the mesh size $\eta(n)$ to the sieve dimension $K(n)$. For a quasi-uniform triangulation with mesh parameter $\eta(n)$, the number of triangles satisfies $\#\mathcal{T}_{\eta(n)} \sim C\,\eta(n)^{-2}$. Each triangle carries a fixed number of local degrees of freedom (21 for the Argyris element). Moreover, the global C^1-continuity only modifies the scaling by a constant factor. Hence, the dimension of the FEM space grows like $K(n) = \dim V_{\eta(n)} \asymp \eta(n)^{-2}$, which is equivalent to $\eta(n) \asymp K(n)^{-1/2}$. With this relation, it is easy to check that the rate conditions of Theorem 3 are satisfied by taking $K(n) = n^{\alpha}$ and $\lambda_n = n^{-\gamma}$ for $\frac{1}{8} < \alpha < \frac{1}{2}$ and $\gamma > \frac{1}{2}$.

5 Simulation results

To evaluate the performance of the proposed model, we conducted simulations on a horseshoe-shaped domain Ω as described in Ramsay (2002) and Wood et al. (2008). The true spatial effect is displayed in Fig. 2. Note that $\int_{\Omega} h_0(\boldsymbol{p})d\boldsymbol{p} \approx 0$. We observed independent replicates of $(Y, \delta, \boldsymbol{X}, \boldsymbol{P}) = (\min(T, C), I(T \leq C), \boldsymbol{X}, \boldsymbol{P})$, where $\boldsymbol{X} = (X_1, X_2)$ has $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \text{Bernoulli}(0.5)$, and $\boldsymbol{P}$ is drawn uniformly from a grid over the domain with spacing 0.025. Conditional on $(\boldsymbol{X}, \boldsymbol{P})$, the event time T follows an exponential distribution with rate parameter $\exp(-4\log(10) + \boldsymbol{X}^{\top}\boldsymbol{\beta}_0 + h_0(\boldsymbol{P}))$, where $\boldsymbol{\beta}_0 = (0.25, -1)^{\top}$, matching the model in (1). The censoring time C is exponentially distributed with rate varying in $\{0.27, 2.00\}$, corresponding to approximately 15% and 30% censoring, respectively. The analysis is conducted using sample sizes of $n = 500, 1000$, and 2000.

We construct the mesh using the package of Sangalli (2021) and refine the domain so that the maximum size η of the mesh triangles is consistent with the asymptotic analysis, setting $\eta = 0.1n^{-0.45}$. We employ first-order FEM basis functions to reduce computational complexity. An illustration for the case $n = 500$ is provided in Fig. 3. Blue dots represent uncensored observations, red dots indicate censored observations, and the gray background depicts the constructed mesh.

We performed $N = 250$ Monte Carlo replications. The penalized partial likelihood in (3) was maximized using the quasi-Newton algorithm described in Sect. 3.3. At each replication, the regularization parameter λ was selected via the CV-PLD procedure (Sect. 3.4) over the grid $\Lambda_n = \{\lambda_j = |\Omega|\, n^{-0.55} e^{\ell_j},\ \ell_j = \log(0.05) + \frac{j-1}{9}(\log(50) - \log(0.05)),\ j = 1, \ldots, 10\}$,

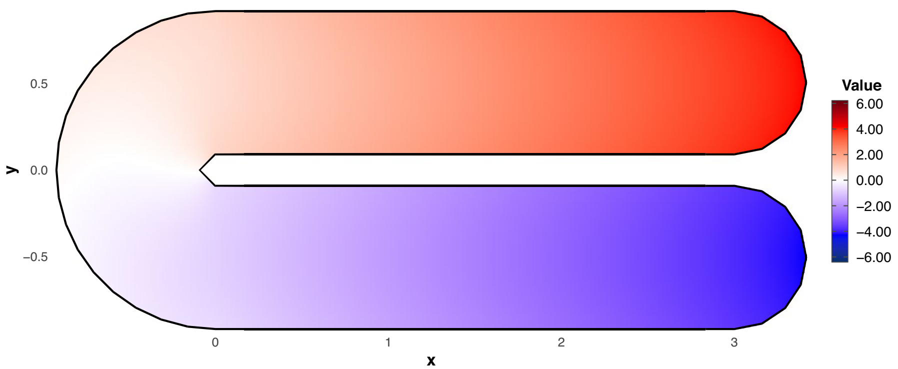


Fig. 2 True spatial effect h_0 on the horseshoe-shaped domain Ω, used as ground truth throughout the simulation study

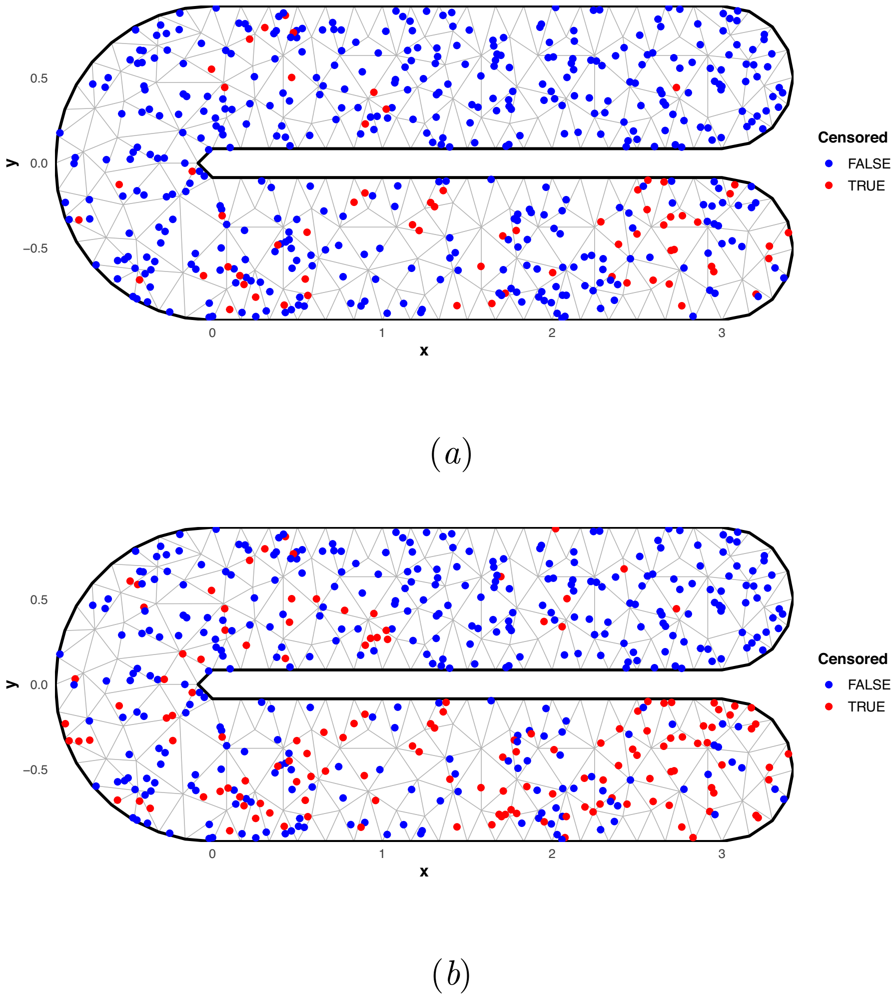


Fig. 3 Sampling design and finite element mesh for one representative replicate at $n = 500$, under (*a*) 15% and (*b*) 30% censoring. Blue dots are uncensored observations, red dots are censored

where $|\Omega|$ indicates the area of Ω and normalizes the Laplacian penalty so that the regularization strength is independent of the size of the spatial domain. The optimal value $\hat{\lambda} \in \Lambda_n$ was determined using five-fold cross-validation.

Figure 4 illustrates the effect of the smoothing parameter λ on the estimated spatial effect, showing $\hat{h}$ from a single Monte Carlo replication at $n = 500$ under 15% censoring for six values of λ ranging from strong under-smoothing to strong over-smoothing.

Figure 5 reports the average estimation error over the $N = 250$ Monte Carlo replications, defined as $N^{-1} \sum_{i=1}^{N} \{\hat{h}^i(s) - h_0(s)\}$, where $\hat{h}^i$ denotes the estimated spatial field from the ith Monte Carlo replication. The bias appears to be small and is

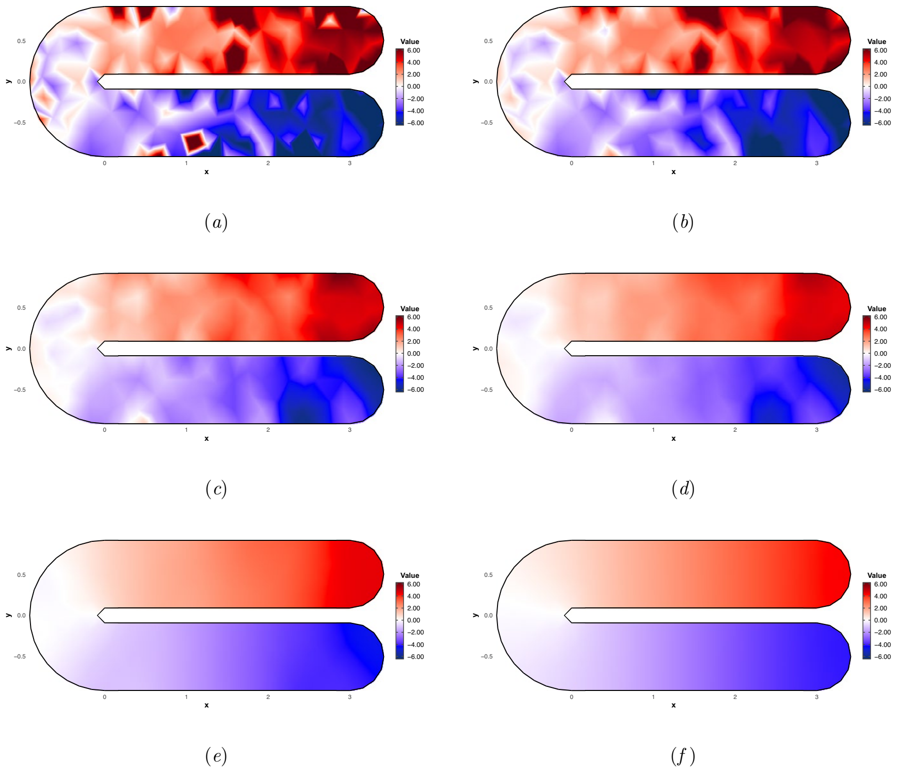


Fig. 4 Estimated spatial effect $\hat{h}$ from a single Monte Carlo replication at $n = 500$ under 15% censoring, for six values of the smoothing parameter λ: **a** $\lambda = 3.05 \times 10^{-6}$, **b** $\lambda = 6.57 \times 10^{-5}$, **c** $\lambda = 1.42 \times 10^{-3}$, **d** $\lambda = 6.57 \times 10^{-3}$, **e** $\lambda = 1.42 \times 10^{-1}$, **f** $\lambda = 3.05$

mainly concentrated near the boundary, where the Neumann conditions constrain the behavior of the derivative of the spatial field. This is consistent with the well-known behavior of nonparametric estimators on bounded domains and does not compromise the overall quality of the estimation. Moreover, the bias decreases as the sample size increases and the censoring rate decreases. It is more pronounced along the lower branch of the horseshoe-shaped domain, where the censoring mechanism induces a larger number of censored observations.

The proposed spatial PH model is compared with the standard PH model and the Generalized Additive Model (GAM) PH. In the latter, spatial variation is captured by a two-dimensional thin-plate regression spline, $s(x, y)$, with basis dimension $k = 100$. All simulations were carried out in R version 4.5.2 on an Intel Xeon Platinum 8460Y+ workstation with 1000 GB of RAM. The standard PH and GAM PH models are fitted with the `survival` v3.8.6 (Therneau 2015) and `mgcv` v1.9.4 (Wood 2018) packages, respectively. For computational efficiency, the proposed method is implemented in `C++` and interfaced to R via the `Rcpp` package v1.1.1 (Eddelbuettel and François 2011).

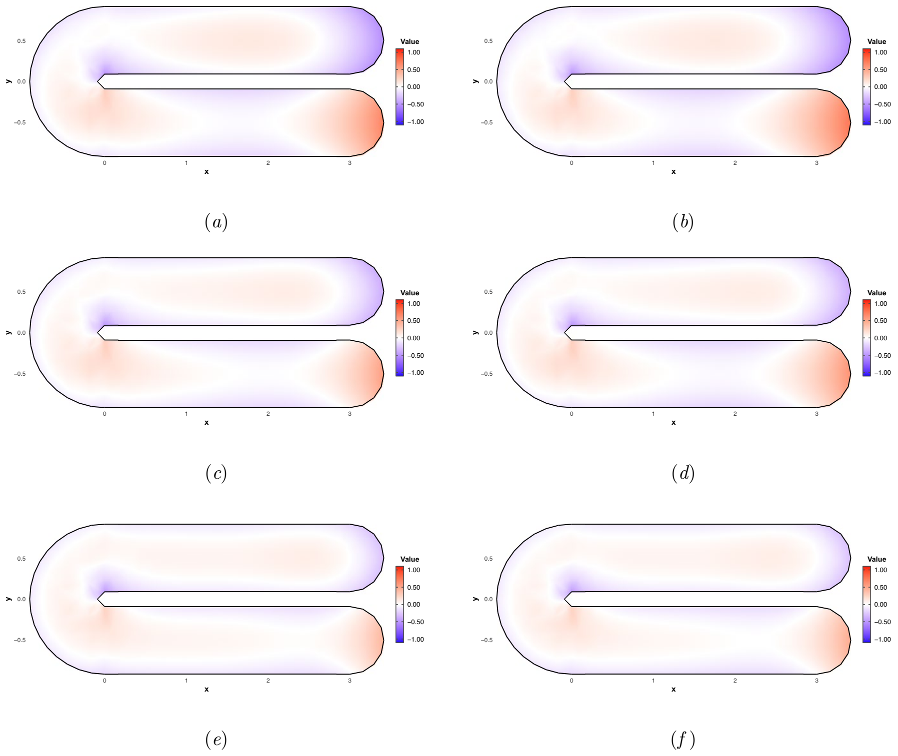


Fig. 5 Average pointwise estimation error $\hat{h} - h_0$ over $N = 250$ Monte Carlo replications. Panels **a**, **c**, **e** correspond to 15% censoring at $n = 500$, 1000, 2000; panels **b**, **d**, **f** to 30% censoring at the same sample sizes

Table 1 reports the estimation performance. Reported metrics include coefficient bias, defined as $\textbf{\textit{bias}}\,\beta_j = N^{-1} \sum_{i=1}^{N} \hat{\beta}_j^{\,i} - \beta_{0,j}$ and the root mean squared error (RMSE) for regression parameters, defined as $\textbf{\textit{RMSE}} = \sqrt{N^{-1} \sum_{i=1}^{N} \|\hat{\boldsymbol{\beta}}^{\,i} - \boldsymbol{\beta}_0\|^2}$, where $\hat{\boldsymbol{\beta}}^{\,i} = (\beta_1^i, \beta_2^i)$ is the estimate from the ith Monte Carlo replication, L^2 norms of bias for the spatial effect estimates, and 95% coverage probabilities for the components of $\boldsymbol{\beta}$ constructed based on the Hessian and normal approximation. Specifically, for each replication i and coefficient β_j, we build the Wald interval $\hat{\beta}_j^{\,i} \pm 1.96\,\widehat{\boldsymbol{se}}(\hat{\beta}_j^{\,i})$, with $\widehat{\boldsymbol{se}}(\hat{\beta}_j^{\,i})$ taken from the diagonal of the inverse Hessian of the penalized log-partial likelihood, and report the proportion of replications in which it contains $\beta_{0,j}$. The L^2 error of the spatial effect on a region $R \subseteq \Omega$ is computed as

$$\|\hat{h} - h_0\|_{L^2(R)} = \sqrt{\frac{1}{|R|} \sum_{k:\, c_k \in R} \left(\hat{h}(c_k) - h_0(c_k)\right)^2 a_k},$$

where $\{c_k\}$ and $\{a_k\}$ are the centroids and areas of the triangles of a fine reference mesh of Ω, and $|R| = \sum_{k:\, c_k \in R} a_k$. This area-weighted Riemann sum approximates

Table 1 Monte Carlo estimation performance over $N = 250$ replications on the horseshoe domain, by sample size (n) and censoring percentage (Cens.)

n	Cens (%)	β estimation			Spatial effect h_0			Coverage		Time(s)
		Bias β_1	Bias β_2	RMSE $\boldsymbol{\beta}$	L_2 RMSE	L_2 Int	L_2 Bdry	CP95 β_1	CP95 β_2	
Spatial PH										
500	15	– 0.003	0.017	0.105	0.488	0.179	0.252	0.952	0.944	3.18
	30	– 0.003	0.021	0.118	0.526	0.193	0.269	0.948	0.932	3.22
1000	15	– 0.002	0.019	0.077	0.360	0.131	0.198	0.960	0.936	4.51
	30	– 0.002	0.023	0.084	0.393	0.143	0.212	0.960	0.948	3.82
2000	15	– 0.001	0.001	0.055	0.285	0.103	0.164	0.932	0.952	5.90
	30	– 0.001	0.001	0.055	0.304	0.110	0.173	0.944	0.944	6.40
GAM PH										
500	15	– 0.049	0.201	0.232	1.820	–	–	–	–	0.28
	30	– 0.098	0.395	0.422	2.772	–	–	–	–	0.26
1000	15	– 0.049	0.203	0.221	1.645	–	–	–	–	0.47
	30	– 0.097	0.394	0.412	2.631	–	–	–	–	0.45
2000	15	– 0.050	0.199	0.212	1.564	–	–	–	–	0.88
	30	– 0.096	0.384	0.400	2.545	–	–	–	–	0.87
Standard PH										
500	15	– 0.153	0.614	0.641	–	–	–	–	–	0.01
	30	– 0.151	0.610	0.638	–	–	–	–	–	0.01
1000	15	– 0.153	0.600	0.624	–	–	–	–	–	0.01
	30	– 0.152	0.593	0.618	–	–	–	–	–	0.01
2000	15	– 0.153	0.611	0.632	–	–	–	–	–	0.01
	30	– 0.152	0.606	0.628	–	–	–	–	–	0.01

For each method we report the bias of the regression coefficients β_1, β_2, the joint RMSE, the area-weighted L^2 error of the spatial effect estimate on the full domain (L^2 RMSE), restricted to the interior (L^2 Int.) and to the boundary strip (L^2 Bdry.), the 95% coverage probabilities for β_1 and β_2 based on the Hessian and normal approximation, and the mean wall-clock fitting time per replicate (s). Entries marked "–" do not apply

$|R|^{-1} \int_R (\hat{h} - h_0)^2$. We report this quantity over three regions: the full domain Ω, the interior Ω_{int}, and a boundary strip $\Omega_{bnd} = \Omega \setminus \Omega_{int}$. A triangle is classified as belonging to Ω_{bnd} whenever any of its vertices lies on $\partial\Omega$, and to Ω_{int} otherwise.

Overall, the simulation results emphasize four main insights. First, as the sample size increases and the censoring rate decreases, performance improves markedly: both regression coefficient RMSE and L^2 error of the spatial effect shrink, confirming consistency. Second, the spatial PH estimator consistently outperforms the GAM approach in terms of spatial effect recovery, with substantially smaller L^2 error across all scenarios. Moreover, the split between L^2 on the interior of the domain and close to the boundary further shows, as expected, that the error is systematically larger near $\partial\Omega$, while still decreasing with n. Third, both the GAM approach and the standard PH model exhibit persistent bias in $\boldsymbol{\beta}$ regardless of sample size or censoring, underscoring model misspecification, whereas the proposed estimator is essentially unbiased and its 95% Wald intervals attain coverage close to the nominal level across all scenarios, providing empirical support for the asymptotic normality result. Fourth, the computational cost of the proposed method remains moderate, with mean fitting times of only a few seconds per replicate even for $n = 2000$, thanks to the `C++` implementation. Taken together, these findings highlight the robustness of the proposed spatial PH method and its clear advantages in capturing both covariate effects and spatial structure.

6 Empirical application

In this section, we present two empirical studies that illustrate the practical use of the proposed methodology. The first examines emergency medical service (EMS) response times for ambulance units in the city of San Francisco, while the second focuses on crowdsourced seismic data collected during an earthquake in the Campi Flegrei area, Italy. Together, these applications highlight the flexibility of the model across very different domains: urban public safety and real-time earthquake monitoring. Both cases demonstrate the model's ability to handle complex spatial domains—ranging from the intricate coastline of a peninsula to the volcanic topography of a caldera—while adjusting for high-dimensional temporal and environmental covariates.

6.1 San Francisco EMS ambulance response time

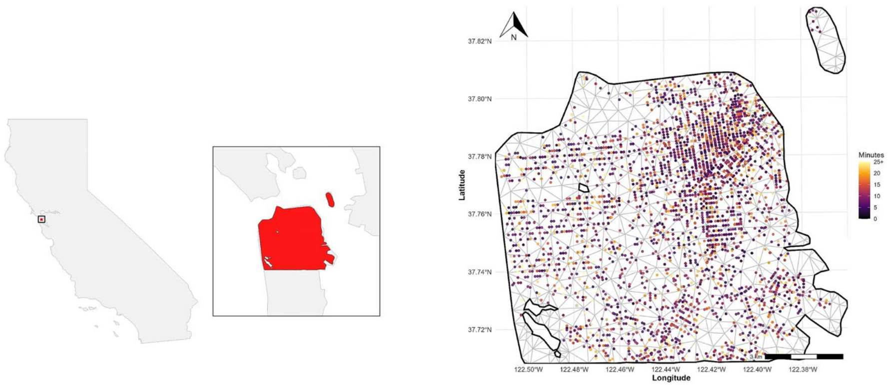
The proposed methodology is illustrated using San Francisco Division of Emergency Medical Services (EMS) dispatch data of February 2025, comprising 7890 different interventions. Each intervention records a sequence of chronological milestones, including the time the call was received, the unit dispatch time, and the start of the unit's response. The event of interest is the duration until the first unit's arrival on scene, which is subject to right-censoring originating from two distinct sources. First, operational censoring occurs when a unit is diverted to a higher-priority incident or canceled before reaching the scene. In such instances, the arrival timestamp is missing, and the censoring time is defined as the duration from the initial call to the last recorded operational activity. Second, administrative censoring was applied to limit the influence of extreme outliers by capping all response times at 60 min. For reference, the 99th percentile of response time is approximately 37 min. Any intervention exceeding 60 min was therefore treated as right-censored at that threshold.

The model is adjusted for several environmental and temporal covariates that might influence medical unit mobility. Weather data were sourced from the Open-Meteo archive, which is relative to the average coordinates of the observed locations (37.77° N, 122.42° W). Weather data include temperature (`temp`) measured in degrees Celsius, hourly precipitation (`precip`) in millimeters, and wind speed (`wind`) in kilometers per hour. Daily periodicity is captured via harmonic terms s_k, c_k for $k = 1, 2$, representing 24- and 12-h cycles derived from the decimal hour of the call. We also include a binary indicator for high-priority emergency calls (`is_emergency`), and a holiday indicator (`is_holiday`) based on the New York Stock Exchange (NYSE) calendar. The latter serves as a proxy for urban traffic dynamics, as NYSE holidays align with the closure of San Francisco's financial district and school systems, significantly reducing road congestion.

In order to accommodate San Francisco's complex coastal geometry and internal water bodies, we define the spatial domain of interest as the city's geographic boundaries where major water features, such as Lake Merced, are excluded, while Treasure Island is retained. The resulting domain is shown in Fig. 6a, displayed in the context of the state of California. This domain is then discretized using a two-dimensional finite element mesh generated via constrained Delaunay refinement. The resulting triangulation consists of 468 nodes. Figure 6b displays the resulting mesh along with the spatial distribution of EMS incidents, where point colors indicate observed response times in minutes.

We apply the proposed method, selecting the smoothing parameter λ as in the simulation study. The estimated regression coefficients $\hat{\beta}$ are reported in Table 2. The reported numerical values correspond to the estimated coefficients, while the relative p-values are computed using standard errors obtained from the inverse Hessian of the penalized log-likelihood.

We observe that the indicator variable `is_emergency` is the most dominant predictor, with an estimated coefficient of 1.913 ($p < 0.001$). This implies a substantially higher arrival hazard for emergency-status calls. The coefficient for `precip` is 0.035 and is marginally significant ($p = 0.066$), suggesting a modest increase in the hazard under precipitation. Furthermore, the 12-h harmonic component s_2 is statistically significant ($p = 0.019$), indicating a semi-diurnal rhythm in EMS efficiency, possibly associated with peak demand periods or shift changes. The estimated effects of `wind`, `temp`, and `is_holiday` are instead not statistically different from zero at the 10% significance level. For comparison,



(a) (b)

Fig. 6 Spatial domain and data for San Francisco in February 2025: **a** location of the San Francisco study area within California; **b** finite element mesh and EMS response-time observations, with points indicating incident locations colored according to observed time

Table 2 Estimated covariate effects from the proposed spatial PH model and the standard PH model for San Francisco EMS ambulance response times (February 2025)

Covariate	Spatial PH			Standard PH		
	Estimate	Std. error	p-value	Estimate	Std. error	p-value
is_emergency	1.913	0.028	< 0.001	1.886	0.028	< 0.001
temp	– 0.004	0.005	0.469	– 0.007	0.005	0.227
precip	0.035	0.019	0.066	0.038	0.019	0.045
wind	– 0.002	0.002	0.348	– 0.001	0.002	0.505
is_holiday	0.032	0.026	0.216	0.032	0.026	0.221
s_1 (24 h)	– 0.004	0.018	0.815	– 0.009	0.018	0.605
c_1 (24 h)	– 0.015	0.021	0.487	– 0.011	0.021	0.603
s_2 (12 h)	– 0.040	0.017	0.019	– 0.044	0.017	0.011
c_2 (12 h)	0.007	0.017	0.684	0.017	0.017	0.326

Estimates are log-hazard ratios; s_k and c_k denote sine and cosine terms capturing 24-h ($k = 1$) and 12-h ($k = 2$) temporal cycles

Table 2 also reports estimates obtained from a standard Cox proportional hazards model that omits the spatial component. The two specifications yield broadly consistent point estimates and similar standard errors, indicating that the covariate effects are robust to the inclusion of the spatial field. Some differences nonetheless emerge: the spatial model attributes a slightly larger effect to `is_emergency` (1.913 vs. 1.886), suggesting that once geographic heterogeneity is properly accounted for, the operational priority gap between emergency and non-emergency calls becomes more pronounced. Conversely, under the standard Cox model the effect of `precip` crosses the conventional 5% threshold ($p = 0.045$) and the magnitude of the 12-h harmonic s_2 is slightly inflated. This pattern is consistent with the well-known phenomenon whereby neglecting unobserved spatial heterogeneity can bias covariate effects and distort their significance, as part of the residual geographic variation is absorbed into the covariates rather than the spatial term.

Beyond these global trends, the model uncovers geographic disparities. Figure 7 illustrates the estimated spatial function $\hat{h}(\boldsymbol{p})$, where lower values indicate weaker response efficiency. These variations persist after adjusting for covariates, indicating that structural characteristics of the urban environment influence expected arrival times. Higher arrival hazards (faster expected responses) are concentrated in central and north-eastern areas, whereas south-western and south-eastern neighborhoods exhibit lower hazards. Treasure Island displays a clear north–south gradient, with relatively higher hazards in the north and lower hazards in the south. These disparities likely reflect differences in road-network configuration, station proximity, topographical constraints, and localized access conditions. The model thus provides a representation of EMS risk, offering insights into the geographic and temporal distribution of EMS arrival performance across the urban landscape.

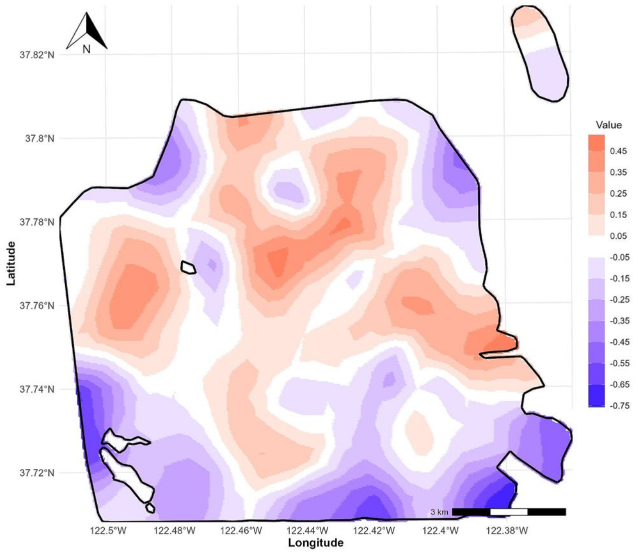


Fig. 7 Estimated spatial effect $\hat{h}(\boldsymbol{p})$, where higher values correspond to faster expected emergency unit arrivals

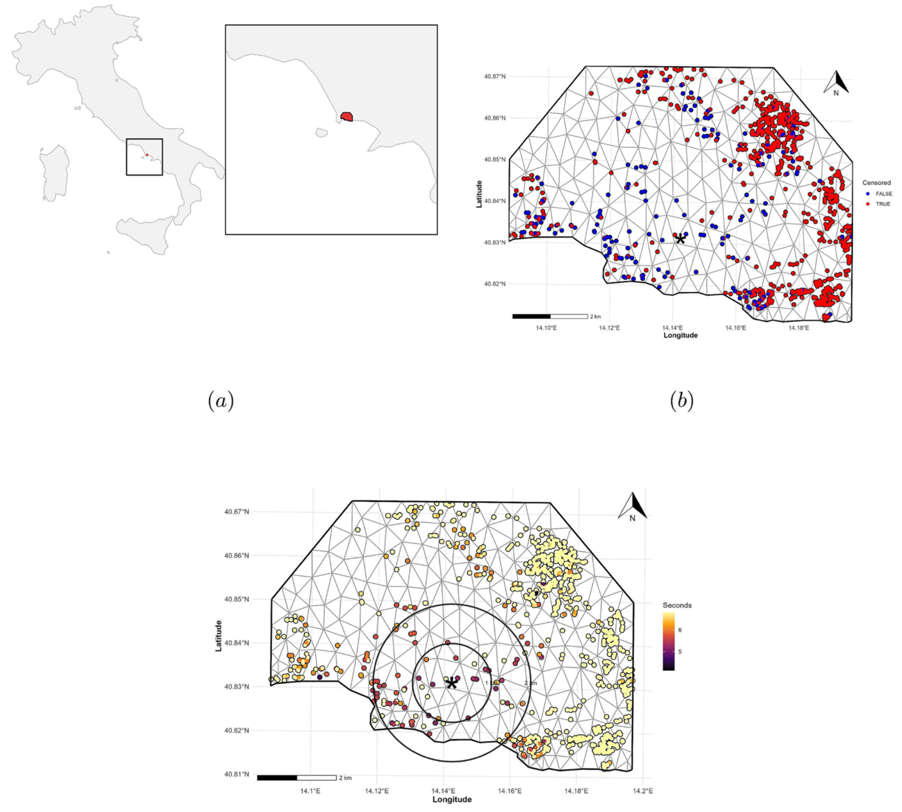
6.2 Crowdsourced seismic data analysis

Earthquake Network (EQN) is a citizen science initiative for real-time earthquake monitoring using crowdsourced smartphones (Finazzi 2016). When a smartphone detects shaking, it sends a signal (called trigger) to a central server. Based on all the triggers received, the server decides whether an earthquake is occurring. If an earthquake is detected, an alert containing a warning about the incoming seismic wave is sent to the population. Earthquake detection also represents a censoring event, meaning any subsequent triggers are ignored.

Here, we examine the EQN triggering times collected during the magnitude 3.1 earthquake occurred on 18 February 2025 at 02:22:19 UTC. The epicenter was located in the Campi Flegrei area (Italy), as shown in Fig. 8a at a depth of 2 km.

We claim that triggering times carry information about site amplification, i.e., the amplification of seismic waves due to local geology. For a given distance from the epicenter, we expect smartphones to detect seismic waves slightly earlier where amplification is high, and slightly later where amplification is low (de-amplification). Additionally, we expect the number of censored smartphones to be higher in de-amplification areas because the smartphone may not detect the seismic wave at all.

For each smartphone i we observe Y_i, which may be equal to the triggering time or to the time of the censoring event (i.e., the earthquake detection by EQN). The only model covariate is the hypocentral distance. Letting $(\text{lat}_i, \text{lon}_i)$ denote the smartphone coordinates, we compute the central angle $\Delta\sigma_i$ using the haversine formula and define the 3D depth-adjusted distance



(a) (b)

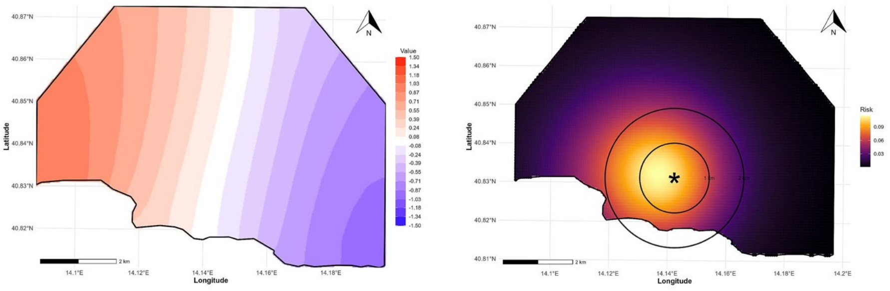
(c)

Fig. 8 Spatial domain and data for Campi Flegrei earthquake data (Feb 18, 2025): **a** location of the study area within Italy, **b** triangular mesh with censored (red) and uncensored (blue) observations, and **c** observed smartphone triggering times with distance rings from the epicenter (star) colored according to observed time

$$\text{dist}_i = \sqrt{(\text{depth})^2 + \left(2R\sin(\Delta\sigma_i/2)\right)^2},$$

where $R = 6371$ km is the Earth's radius and depth $= 2$ km. To ensure a fair proportion of uncensored values within the study area, we restrict the data to smartphones located within 5 km from the epicenter. This resulted in a sample of 976 triggering times, approximately 80% of which are right-censored.

As before, we construct a two-dimensional finite element mesh that conforms to the irregular boundary of the coastline of Campi Flegrei and the distance limit from the epicenter. The mesh is obtained by refining the boundary geometry and consists of 229 vertices connected through triangular elements. The resulting mesh, along with smartphone locations colored by censoring status, is shown in Fig. 8b, while the observed triggering times are illustrated in Fig. 8c.



(a) (b)

Fig. 9 Model results for the Campi Flegrei earthquake: **a** estimated spatial effect $\hat{h}(\boldsymbol{p})$, where higher values indicate a greater hazard (faster triggering); **b** estimated relative hazard $\exp(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}} + \hat{h}(\boldsymbol{p}))$ at 12:00 PM, where warmer colors denote higher values

The estimated hypocentral distance effect is equal to $-$ 1.122, with standard error 0.102 and p-value < 0.001. This confirms that, as expected, smartphones located further from the epicenter have a lower hazard of triggering, that is, they tend to trigger later or not at all. For comparison, fitting a standard Cox proportional hazards model that omits the spatial component yields a substantially larger effect in magnitude ($-$ 1.304, standard error 0.089, $p < 0.001$). The discrepancy is consistent with the interpretation that, when spatial heterogeneity is not accounted for, part of the site amplification signal is absorbed into the distance coefficient, inflating its apparent effect. Once the spatial field captures the local geological variation, the distance effect retains its expected sign and high significance but takes a more moderate value, reflecting the purely geometric attenuation of seismic waves with hypocentral distance. On the other hand, Fig. 9a displays the estimated spatial effect after accounting for the hypocentral distance. The model uncovers a clear east–west gradient across the Campi Flegrei area, which is consistent with the findings of Finazzi et al. (2025) who obtained the amplification map for the Campi Flegrei area from the analysis of the smartphone shaking intensity rather than triggering time. Finally, Fig. 9b illustrates the estimated spatial variation of the relative hazard for the specific earthquake. The highest hazard values are obviously concentrated around the epicenter. However, due to the spatial effect considered in the model (which describes site amplification), the gradient of the hazard is not radial with respect to the epicenter.

7 Conclusions

We introduced a nonparametric proportional hazards model that incorporates a smooth spatial effect within the classical partial likelihood framework. By combining a Laplacian-based differential penalty with finite element sieve approximations, our approach can handle complex and irregular spatial domains while maintaining a concave optimization problem and producing interpretable regression coefficients.

From a theoretical perspective, we proved that the model is identifiable, that the sieve estimator is consistent, and that the parametric component is asymptotically normal. These results ensure that standard inferential procedures for the regression coefficients remain valid despite the presence of an infinite-dimensional spatial effect. Empirically, simulation studies demonstrate that the proposed method accurately reconstructs latent spatial variation, reducing bias and mean squared error relative to standard proportional hazards models and yielding high-resolution, interpretable spatial risk surfaces. Two empirical applications further illustrate how explicitly modeling spatial heterogeneity can reveal meaningful structures in the data-generating process that would otherwise remain unobserved.

The proposed methodology, developed for spatial survival analysis, naturally extends to other spatial data structures such as areal and network data, and more broadly to settings requiring nonparametric effects within a proportional hazards framework. The finite element formulation provides a flexible and principled way to represent smooth effects over complex domains, making the approach applicable to a wide range of problems beyond those considered here.

There are several avenues for future research. One important area for further research concerns the goodness-of-fit of the proportional hazards assumption in the presence of a nonparametric spatial effect. Classical diagnostic tools, such as Schoenfeld residuals, are derived under the assumption of purely parametric covariate effects and therefore require careful adaptation in this setting. The development of residual- or score-based diagnostics that explicitly account for the estimated spatial component would enable practitioners to evaluate time-varying effects and identify deviations from proportionality following model fitting. Other avenues to explore include formally quantifying the uncertainty of the spatial effect, selecting the penalty structure based on the data, and extending spatio-temporal survival models to include latent effects that evolve over time.

Appendix

The Appendix is organized as follows. Section explains how to adapt the model in case of areal data. Section B provides the proof of Theorem 1 on model identifiability. Section C contains the proof of Theorem 2 establishing consistency of the estimator. Section D gives the proof of Theorem 3. Finally, Sect. E collects auxiliary lemmas on functional space compactness, continuity, differentiability, and uniqueness results used throughout the proofs.

A. Areal data

The proposed methodology has so far been developed for geostatistical survival data, where observations are associated with point-referenced spatial locations $\boldsymbol{p}_1, \dots, \boldsymbol{p}_n \in \Omega$. In many applications, however, survival outcomes are available only for areal units, corresponding to disjoint spatial subdomains of Ω. Let $\{D_1, \dots, D_n\}$ denote such a collection of regions. For instance, in epidemiological or environmental studies, data may be aggregated at the level of administrative districts, census tracts, or neighborhoods. In this setting, the spatial effect cannot be evaluated at a point, but must instead be represented through an areal functional of the latent effect. A natural choice is the mean of h over region D_i, leading to the model

$$\lambda(t \mid \boldsymbol{x}, D_i) \;=\; \lambda_0(t) \exp\left(\boldsymbol{x}^\top \boldsymbol{\beta}_0 \;+\; \frac{1}{|D_i|} \int_{D_i} h(\boldsymbol{p})\, d\boldsymbol{p} \right),$$

where $|D_i|$ denotes the area of region D_i. Estimation of $(\boldsymbol{\beta}, h)$ proceeds by maximizing a penalized criterion analogous to the one introduced in (3), but with integrals over areal units in place of pointwise evaluations. The finite element framework naturally accommodates this extension: nodal evaluations of h are replaced by weighted averages over the mesh elements contained in each D_i. In practice, this requires the construction of an incidence matrix that maps finite element basis functions to the regions $\{D_1, \dots, D_n\}$, together with numerical quadrature to approximate $\int_{D_i} h(\boldsymbol{p})\, d\boldsymbol{p}$ for each i. From a theoretical perspective, the asymptotic results of Sect. 4 can be extended to this setting by noting that areal averages correspond to additional linear operators acting on the underlying effect h. A rigorous treatment requires technical modifications, but the main arguments parallel those in the point-referenced case once the weights $|D_i|^{-1}$ are incorporated.

B. Proof of Theorem 1

Proof First, we show that, from the observed right-censored data, the conditional hazard $\lambda(t \mid \boldsymbol{x}, \boldsymbol{p})$ is itself identified on a nontrivial time interval. For that, fix $(\boldsymbol{x}, \boldsymbol{p})$ in the support of $(\boldsymbol{X}, \boldsymbol{P})$. From the joint law of $(Y, \delta, \boldsymbol{X}, \boldsymbol{P})$ and compute, for $t \geq 0$,

$$\lambda(t \mid \boldsymbol{x}, \boldsymbol{p}) = \lim_{\Delta \downarrow 0} \frac{\mathbb{P}\Big(t \leq Y < t + \Delta,\ \delta = 1 \;\Big|\; Y \geq t,\ \boldsymbol{X} = \boldsymbol{x},\ \boldsymbol{P} = \boldsymbol{p} \Big)}{\Delta}.$$

Assumption 1 (i) ensures that this limit equals the failure hazard of $T | \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{P} = \boldsymbol{p}$ without censoring confounding. Assumption 1 (ii) guarantees $\mathbb{P}(Y \geq t \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{P} = \boldsymbol{p}) > 0$ for all $t \in [0, \tau)$, so the above ratio is well defined and $\lambda(t \mid \boldsymbol{x}, \boldsymbol{p})$ is identified on $[0, \tau)$.

Now, we show that if two parameter triplets generate the same $\lambda(t \mid \boldsymbol{x}, \boldsymbol{p})$, then the triplets must coincide. Suppose two triplets $(\lambda_0, \beta_0, h_0)$ and $(\tilde{\lambda}_0, \tilde{\beta}, \tilde{h})$, with $h_0, \tilde{h} \in \mathcal{H} = \{h : \Omega \to \mathbb{R} \text{ smooth}, \int_\Omega h = 0\}$, generate the same conditional hazard on $[0, \tau)$ for $(\boldsymbol{x}, \boldsymbol{p})$ in the support:

$$\lambda_0(t) \exp\big(\boldsymbol{x}^\top \boldsymbol{\beta}_0 + h_0(\boldsymbol{p})\big) = \tilde{\lambda}_0(t) \exp\big(\boldsymbol{x}^\top \tilde{\boldsymbol{\beta}} + \tilde{h}(\boldsymbol{p})\big), \qquad \text{for a.e. } (x, p) \text{ and all } t \in [0, \tau).$$

Taking logs and defining

$$\boldsymbol{\beta}^\star = \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \qquad r(\boldsymbol{p}) = \tilde{h}(\boldsymbol{p}) - h_0(\boldsymbol{p}), \qquad a(t) = \log\tilde{\lambda}_0(t) - \log\lambda_0(t),$$

we obtain

$$\boldsymbol{x}^\top\boldsymbol{\beta}^\star + r(\boldsymbol{p}) = a(t) \qquad \text{for a.e. } (\boldsymbol{x},\boldsymbol{p}) \text{ and all } t \in [0,\tau). \tag{10}$$

Equality of hazards implies $\boldsymbol{x}^\top\boldsymbol{\beta}_0 = \boldsymbol{x}^\top\tilde{\boldsymbol{\beta}} + c$ for a.e. $(\boldsymbol{x},\boldsymbol{p})$ where c is a constant, and, without loss of generality, we can assume 0 is in the support of X and so $c = 0$. Hence $\boldsymbol{x}^\top\boldsymbol{\beta}^\star = 0$ for $\boldsymbol{X}$-a.e. $\boldsymbol{x}$. Therefore

$$\mathbb{E}\big[(\boldsymbol{X}^\top\boldsymbol{\beta}^\star)^2\big] = (\boldsymbol{\beta}^{\star\top}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top]\boldsymbol{\beta}^\star = 0.$$

Assumption 1 (iv) forces $\boldsymbol{\beta}^\star = 0$; hence $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ and (10) reduces to $r(\boldsymbol{p}) = a(t)$ for a.e. $\boldsymbol{p}$ and all $t \in [0,\tau)$. Since the left-hand side is free of t, $a(t)$ is constant on $[0,\tau)$; write $a(t) \equiv c$. Thus

$$r(\boldsymbol{p}) = c \qquad \text{for } \boldsymbol{P}\text{-a.e. } \boldsymbol{p} \in \Omega. \tag{11}$$

Assumption 1 (v) implies $r(\boldsymbol{p}) = c$ for Lebesgue-a.e. $p \in \Omega$. Since $h_0, \tilde{h} \in \mathcal{H}$,

$$0 = \int_\Omega r(\boldsymbol{p})\,dp = \int_\Omega c\,d\boldsymbol{p} = c\,|\Omega|,$$

so $c = 0$ and $r = 0$ a.e. on Ω.

The function r is smooth; a smooth function that is zero a.e. must be identically zero on Ω. Hence $\tilde{h} \equiv h_0$ on Ω. Finally, with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ and $\tilde{h} = h_0$, the hazard equality implies $\tilde{\lambda}_0(t) = \lambda_0(t)$ for all $t \in [0,\tau)$. □

C. Proof of Theorem 2

Proof Let $\theta = (\boldsymbol{\beta}, h)$ and define the population objective

$$Q(\theta) = \mathbb{E}\Big[\delta\Big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}) - \log s^{(0)}(\boldsymbol{\beta}, h, Y)\Big)\Big],$$

where

$$s^{(0)}(\boldsymbol{\beta}, h, t) = \mathbb{E}\big[I(Y \geq t)\exp\big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P})\big)\big].$$

The result follows from Theorems 3.1–3.2 of Chen (2007) for sieve M–estimation, upon verifying, identification and local curvature (Condition 3.1), sieve approximation properties (Condition 3.2), continuity and compactness (Conditions 3.3–3.4), uniform convergence over the sieves (Condition 3.5) and stochastic regularity (Conditions 3.6–3.8), that we state and verify below. Regarding the rate ϵ_n, consider that the stochastic term is

$$\delta_n \asymp \sqrt{\frac{K(n)\log K(n)}{n}}$$

by our entropy bound, and the sieve approximation error satisfies

$$\|\theta_0 - \pi_n\theta_0\| = \|h_0 - \mathcal{J}^n h_0\|_\infty = O\big(\eta(n)\big)$$

by standard FEM theory (here $\pi_n\theta_0 = (\boldsymbol{\beta}_0, \mathcal{J}^n h_0)$).

Because $\mathcal{H}_n \subset \mathcal{H}$ and $\|h\|_{H^2(\Omega)} \le M_{\mathcal{H}}$ on $\mathcal{H}$, the Laplacian penalty

$$J(h) = \int_\Omega (\Delta h)^2\, d\boldsymbol{p}$$

is uniformly bounded on $\mathcal{H}_n$: $\sup_{h\in\mathcal{H}_n} J(h) \le C < \infty$. Let $Q^{pen}(\theta) = Q(\theta) - \frac{\lambda_n}{2}J(h)$. Then the penalty perturbs the criterion uniformly by

$$\sup_{\theta\in\Theta_n}\big|\, Q^{pen}(\theta) - Q(\theta)\,\big| \le C\,\lambda_n. \tag{12}$$

We claim that Q has local quadratic curvature at θ_0: there exist constants $c > 0$ and $r > 0$ such that

$$Q(\theta_0) - Q(\theta) \;\ge\; c\,d(\theta,\theta_0)^2 \quad \text{whenever } d(\theta,\theta_0) \le r.$$

Let $\tilde{\theta}_n$ be any maximizer of Q^{pen} over Θ_n. Since $Q^{pen}(\tilde{\theta}_n) \ge Q^{pen}(\theta_0)$, we have

$$Q(\theta_0) - Q(\tilde{\theta}_n) \;\le\; \big|Q(\theta_0) - Q^{pen}(\theta_0)\big| + \big|Q^{pen}(\tilde{\theta}_n) - Q(\tilde{\theta}_n)\big| \;\le\; 2C\,\lambda_n.$$

Therefore, we obtain

$$c\,d(\tilde{\theta}_n,\theta_0)^2 \;\le\; 2C\,\lambda_n \quad\Longrightarrow\quad d(\tilde{\theta}_n,\theta_0) \;=\; O\big(\sqrt{\lambda_n}\big).$$

Thus, the penalty perturbs the population maximizer by at most order $\sqrt{\lambda_n}$ and contributes an additive $\sqrt{\lambda_n}$ term to the overall convergence rate. These three contributions yield $\varepsilon_n = \max\left\{\, \delta_n,\ \eta(n),\ \sqrt{\lambda_n}\,\right\}$.

We now state and verify Conditions 3.1–3.8 of Chen (2007). $\square$

Condition 3.1 *(i) $Q(\theta_0) > -\infty$, and if $Q(\theta_0) = +\infty$ then $Q(\theta) < +\infty$ for all $\theta \in \Theta_k \setminus \{\theta_0\}$ for all $k \ge 1$; (ii) there are a nonincreasing positive function $\delta(\cdot)$ and a positive function $g(\cdot)$ such that for all $\varepsilon > 0$ and for all $k \ge 1$,*

$$Q(\theta_0) - \sup_{\{\theta\in\Theta_k : d(\theta,\theta_0)\ge\varepsilon\}} Q(\theta) \ge \delta(k)g(\varepsilon) > 0.$$

Proof of Condition 3.1(i). The value $Q(\theta_0)$ is finite because $\mathbb{E}\big|\boldsymbol{X}^\top\boldsymbol{\beta}_0 + h_0(\boldsymbol{P})\big| < \infty$ and $\mathbb{E}|\log s^{(0)}(\boldsymbol{\beta}_0, h_0, Y)| < \infty$ by the same moment bounds used in Lemma 3. $\square$

Proof of Condition 3.1(ii). Let $\theta = (\boldsymbol{\beta}, h)$ be an arbitrary point in the parameter space Θ, and let $\theta_0 = (\boldsymbol{\beta}_0, h_0)$ be the true parameter value. We define the perturbation $\xi = \theta - \theta_0 = (\boldsymbol{\beta} - \boldsymbol{\beta}_0, h - h_0) = (\boldsymbol{u}, g)$.

From Lemma 4, the functional $Q(\theta)$ is twice Fréchet differentiable. Therefore, we can write a second-order Taylor expansion of $Q(\theta)$ around θ_0:

$$Q(\theta) = Q(\theta_0) + DQ(\theta_0)[\xi] + \frac{1}{2}D^2Q(\theta_0)[\xi,\xi] + o(d(\theta,\theta_0)^2),$$

where $DQ(\theta_0)[\xi]$ is the first Fréchet derivative in the direction ξ, and $D^2Q(\theta_0)[\xi,\xi]$ is the second Fréchet derivative.

As established by Lemma 5, θ_0 is the unique maximizer of the population objective function $Q(\theta)$. A necessary condition for θ_0 to be an extremum is that the first derivative of Q at θ_0 is zero for any direction ξ. That is:

$$DQ(\theta_0)[\xi] = 0.$$

With the first-order term being zero, the Taylor expansion simplifies to:

$$Q(\theta) - Q(\theta_0) = \frac{1}{2}D^2Q(\theta_0)[\xi,\xi] + o(d(\theta,\theta_0)^2).$$

Now, from Lemma 4, we have:

$$D^2Q(\theta_0)[\xi,\xi] = \mathbb{E}\left[-\delta\,\frac{s^{(2)}(\theta_0,Y)[\xi,\xi]\,s^{(0)}(\theta_0,Y) - \left(s^{(1)}(\theta_0,Y)[\xi]\right)^2}{\left(s^{(0)}(\theta_0,Y)\right)^2}\right].$$

The fraction inside the expectation can be recognized as the conditional variance of the term $A(\boldsymbol{X},\boldsymbol{P}) = \boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P})$ for an individual in the risk set at time Y. Thus, we denote this by $\mathrm{Var}_Y(A)$. We can write the second derivative more compactly as:

$$D^2Q(\theta_0)[\xi,\xi] = -\mathbb{E}[\delta\cdot\mathrm{Var}_Y(A)]. \tag{13}$$

Since the event indicator δ is nonnegative and variance is always nonnegative, we have $D^2Q(\theta_0)[\xi,\xi] \le 0$. For the second derivative to be strictly negative definite, we must show that $D^2Q(\theta_0)[\xi,\xi] = 0$ if and only if $\xi = \mathbf{0}$.

$D^2Q(\theta_0)[\xi,\xi] = 0$ implies that $\mathrm{Var}_Y(\boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P})) = 0$ for almost all observed event times Y. This means that for almost every event time, the quantity $\boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P})$ must be constant for all individuals in the corresponding risk set. Given the assumptions on the distributions of $\boldsymbol{X}$ and $\boldsymbol{P}$, this can only hold if $\boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P})$ is constant almost surely.

Following the same line of reasoning as in the proof of Theorem 1, we obtain that $D^2Q(\theta_0)[\xi,\xi] = 0$ if and only if $\xi = (\boldsymbol{u}, g) = \mathbf{0}$. Therefore, the Hessian operator is negative definite at θ_0.

We have shown that the quadratic form $I(\xi) = -D^2Q(\theta_0)[\xi,\xi]$ is positive definite. Furthermore, the functional $I(\xi)$ is continuous with respect to ξ in the d-metric. The parameter space $\Theta = \mathcal{B}\times\mathcal{H}$ is compact under this metric. Consequently, the function $f(\xi) = I(\xi)/d(\mathbf{0},\xi)^2$ is continuous and positive on the compact unit sphere $\{\xi \mid d(\mathbf{0},\xi) = 1\}$. It must, therefore, attain a minimum value, which we denote as $2c > 0$.

This implies that for any $\xi \ne \mathbf{0}$:

$$I(\xi) \geq 2c \cdot d(\mathbf{0}, \xi)^2 \quad \Longrightarrow \quad D^2 Q(\theta_0)[\xi, \xi] \leq -2c \cdot d(\theta, \theta_0)^2.$$

Substituting this quadratic bound back into our simplified Taylor expansion, we get:

$$Q(\theta) - Q(\theta_0) \leq \frac{1}{2}(-2c \cdot d(\theta, \theta_0)^2) + o(d(\theta, \theta_0)^2) = -c \cdot d(\theta, \theta_0)^2 + o(d(\theta, \theta_0)^2).$$

This inequality implies that for any $\varepsilon > 0$, there exists a local neighborhood around θ_0 within which $Q(\theta_0) - Q(\theta) \geq c' d(\theta, \theta_0)^2$ for some $c' > 0$.

Because θ_0 is the unique global maximizer of Q on a compact set, the separation $Q(\theta_0) - \sup_{d(\theta,\theta_0)\geq\varepsilon} Q(\theta)$ is guaranteed to be positive for any $\varepsilon > 0$. The local quadratic nature of this separation, which we have just proven, is the crucial property that drives the convergence rates of the estimator. Therefore, we can state the existence of a positive function $g(\varepsilon) = c\varepsilon^2$ that provides the required lower bound.

Finally, since this bound was established on the full parameter space Θ, it holds automatically for any subset, including the sieve spaces Θ_k. Thus, we can set the nonincreasing function $\delta(k) \equiv 1$, which completes the proof. □

Let the sieve parameter spaces be

$$\Theta_n = \mathcal{B} \times \mathcal{H}_n, \qquad n = 1, 2, \dots,$$

where $\mathcal{B} \subset \mathbb{R}^b$ is the fixed compact set from Assumption 1 and $\mathcal{H}_n \subset \mathcal{H}$ are the finite element subspaces constructed in Sect. 3.2. With this notation, we now consider the following condition.

Condition 3.2 (i) $\Theta_n \subset \Theta_{n+1} \subset \Theta$ *for every* $n \geq 1$; (*ii*) *for the true parameter* $\theta_0 = (\boldsymbol{\beta}_0, h_0) \in \Theta$ *there exists a sequence* $\pi_n\theta_0 \in \Theta_n$ *such that* $d(\theta_0, \pi_n\theta_0) \to 0$ *as* $n \to \infty$.

Proof of Condition 3.2(i). The coefficient space $\mathcal{B}$ is independent of n, hence $\mathcal{B} \subset \mathcal{B}$ trivially. For the functional part, our triangulations satisfy the condition $\mathcal{T}_{\eta(n+1)}$ a refinement of $\mathcal{T}_{\eta(n)}$ with $\eta(n+1) < \eta(n)$; consequently every piecewise polynomial function in $\mathcal{H}_n$ is also contained in $\mathcal{H}_{k+1}$. Lemma 1 implies that $\Theta_n \subset \Theta_{n+1}$. □

Proof of Condition 3.2(ii). Define the projection operator $\mathcal{J}^n : \mathcal{H} \to \mathcal{H}_n$ as in Sect. 3.2 and set

$$\pi_n\theta_0 = \left(\boldsymbol{\beta}_0, \mathcal{J}^n h_0\right) \in \Theta_n.$$

Because $\boldsymbol{\beta}_0$ is copied verbatim, only the functional component needs an error bound. By the FEM interpolation estimate (4):

$$\|h_0 - \mathcal{J}^n h_0\|_\infty \leq C\,\eta(n)\,\|h_0\|_{H^2(\Omega)} \quad \longrightarrow 0 \quad \text{as } n \to \infty,$$

since $\eta(n) \downarrow 0$ and $h_0 \in H^2(\Omega)$. Therefore we have

$$d\left(\theta_0, \pi_n\theta_0\right) = \|h_0 - \mathcal{J}^n h_0\|_\infty \longrightarrow 0. \tag{14}$$

□

Condition 3.3 (*i*) *For every fixed* $n \geq 1$ *the mapping* $Q(\theta)$ *is upper semicontinuous on* Θ_n *under* d; (*ii*) $\left|Q(\theta_0) - Q(\pi_n\theta_0)\right| = o\big(\delta\big(k(n)\big)\big)$ *for any sequence* $k(n) \uparrow \infty$.

Proof of condition 3.3(i). Lemma 3 established that Q is (everywhere) continuous on the full space Θ. Continuity clearly implies both upper and lower semicontinuity, so the restriction of Q to any subset – in particular to each finite-dimensional, closed set Θ_n – is automatically upper semicontinuous. □

Proof of condition 3.3(ii). We have shown in (14) that $d\big(\theta_0, \pi_k\theta_0\big) \to 0$ as $n \to \infty$. Because Q is continuous (Lemma 3),

$$\lim_{n\to\infty} Q\big(\pi_n\theta_0\big) = Q(\theta_0), \quad \text{i.e.} \quad \left|Q(\theta_0) - Q(\pi_n\theta_0)\right| = o(1),$$

which is the assertion, as we set $\delta(k(n)) = 1$. □

Condition 3.4 *For every* $n \geq 1$, *the sieve space* $\Theta_n = \mathcal{B} \times \mathcal{H}_n$ *is compact under the metric* $d(\cdot, \cdot)$.

Proof of Condition 3.4. By assumption $\mathcal{B} \subset \mathbb{R}^b$ is compact. The fact that $\mathcal{H}_n$ is compact under the $\|\cdot\|_\infty$ is shown in Lemma 2. The Cartesian product of two compact sets is compact, and the metric d is simply the sum of the metrics on the two factors. Thus Θ_n is compact under d. □

Condition 3.5 (*i*) *for every fixed* $k \geq 1$, $\sup_{\theta\in\Theta_k}\left|\hat{Q}_n(\theta) - Q(\theta)\right| \xrightarrow{p} 0$; (*ii*) *along any sequence* $k = k(n) \to \infty$ *satisfying* $\sup_{\theta\in\Theta_{k(n)}} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = o_p(\delta(k))$.

Proof of Condition 3.5(i). For $Z = (Y, \delta, \boldsymbol{X}, \boldsymbol{P})$ define

$$m(Z;\theta) = \delta\big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}) - \log s^{(0)}(\boldsymbol{\beta}, h, Y)\big), \qquad \theta = (\boldsymbol{\beta}, h) \in \Theta.$$

Let $P_n f = n^{-1}\sum_{i=1}^n f(Z_i)$ and $Pf = \mathbb{E}[f(Z)]$. The (unpenalized) sample criterion can be written as

$$\widetilde{Q}_n(\theta) = P_n m_n(\cdot;\theta), \qquad m_n(Z;\theta) = \delta\big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}) - \log S_n^{(0)}(\boldsymbol{\beta}, h, Y)\big),$$

so that the penalized empirical objective is

$$\hat{Q}_n(\theta) = \widetilde{Q}_n(\theta) - \tfrac{\lambda_n}{2}\int_\Omega (\Delta h)^2\, d\boldsymbol{p},$$

and the population counterpart is $Q(\theta) = Pm(\cdot;\theta)$. Add and subtract $P_n m(\cdot;\theta)$:

$$\hat{Q}_n(\theta) - Q(\theta) = (P_n - P)\, m(\cdot;\theta) + P_n\big[m_n(\cdot;\theta) - m(\cdot;\theta)\big] - \tfrac{\lambda_n}{2}\int_\Omega (\Delta h)^2\, d\boldsymbol{p}. \quad (15)$$

Denote the remainder by

$$R_n(\theta) = P_n\big[m_n(\cdot;\theta) - m(\cdot;\theta)\big].$$

By Lemma 3 there exist constants $0 < c_0 \leq C_0 < \infty$ (independent of n and k) such that, for all θ and $t \in [0, \tau]$,

$$c_0 \, \mathbb{P}(Y \geq t) \leq s^{(0)}(\theta, t) \leq C_0 \, \mathbb{P}(Y \geq t).$$

Hence $s^{(0)}$ is uniformly bounded away from zero on $[0, \tau]$. Assumption 2 further implies $\|\boldsymbol{X}\| \leq M_{\mathcal{X}}$ a.s. and $\|h\|_\infty \leq M_{\mathcal{H}}$ on every sieve $\mathcal{H}_k$, so $|m(Z;\theta)| \leq F(Z)$ for some integrable, nonrandom envelope F that does not depend on k.

Each sieve Θ_k is compact and finite-dimensional (Lemma 2), and $m(\cdot;\theta)$ is Lipschitz in θ on Θ_k with envelope F. Hence the class $\mathcal{F}_k = \{m(\cdot;\theta) : \theta \in \Theta_k\}$ is Glivenko–Cantelli:

$$\sup_{\theta \in \Theta_k} \left|(P_n - P)m(\cdot;\theta)\right| \xrightarrow{p} 0. \tag{16}$$

From the definition,

$$R_n(\theta) = P_n\Big[\delta\Big(\log s^{(0)}(\theta, Y) - \log S_n^{(0)}(\theta, Y)\Big)\Big].$$

By the mean value theorem and the lower bound $s^{(0)}(\theta, t) \geq c_0 \, \mathbb{P}(Y \geq t)$,

$$\left| \log S_n^{(0)}(\theta, t) - \log s^{(0)}(\theta, t)\right| \leq \frac{1}{c_0 \, \mathbb{P}(Y \geq t)} \left|S_n^{(0)}(\theta, t) - s^{(0)}(\theta, t)\right|.$$

The class

$$\mathcal{G}_k = \Big\{(z, t) \mapsto I(Y \geq t) \exp(\boldsymbol{X}^\top \boldsymbol{\beta} + h(\boldsymbol{P})) \; : \; \theta \in \Theta_k, \; t \in [0, \tau]\Big\}$$

is also Glivenko–Cantelli for fixed k (finite-dimensional Lipschitz parametrization and bounded envelope). Therefore,

$$\sup_{\theta \in \Theta_k, \, t \in [0, \tau]} \left|S_n^{(0)}(\theta, t) - s^{(0)}(\theta, t)\right| \xrightarrow{p} 0,$$

which implies

$$\sup_{\theta \in \Theta_k} |R_n(\theta)| \xrightarrow{p} 0. \tag{17}$$

Since $\sup_{h \in \mathcal{H}_n} \int_\Omega (\Delta h)^2 \, d\boldsymbol{p} \leq M_{\mathcal{H}}^2$,

$$0 \leq \tfrac{\lambda_n}{2} \sup_{h \in \mathcal{H}_n} \int_\Omega (\Delta h)^2 \, d\boldsymbol{p} \leq \tfrac{\lambda_n}{2} M_{\mathcal{H}}^2 = o_p(1). \tag{18}$$

From (15), together with (16), (17), and (18), we obtain

$$\sup_{\theta \in \Theta_k} \left|\hat{Q}_n(\theta) - Q(\theta)\right| \xrightarrow{p} 0,$$

which establishes Condition 3.5(i). □

Proof of condition 3.5(ii). Let $K(n) = \dim(\mathcal{H}_n)$ and $K_\Theta(n) = b + K(n)$. Empirical process entropy bounds for Lipschitz parameterizations yield

$$\sup_{\theta\in\Theta_n} \left|(P_n - P)m(\cdot;\theta)\right| = O_p\left(\sqrt{\frac{K_\Theta(n)\log K_\Theta(n)}{n}}\right).$$

Under the assumption of growth rule stated in Theorem 2, this is $o_p(1)$. The same argument applied to the class $\mathcal{G}_n$ gives

$$\sup_{\theta\in\Theta_n,\, t\geq 0} \left|S_n^{(0)}(\theta,t) - s^{(0)}(\theta,t)\right| = o_p(1),$$

hence $\sup_{\theta\in\Theta_n} |R_n(\theta)| = o_p(1)$ by the log–Lipschitz bound above. Using the fact that $\delta(\cdot) \equiv 1$ in our framework, we obtain

$$\sup_{\theta\in\Theta_n} \left|\hat{Q}_n(\theta) - Q(\theta)\right| = o_p(1).$$

□

Condition 3.6 *The sample* $\{Z_t\}_{t=1}^n = \{(Y_t, \delta_t, \boldsymbol{X}_t, \boldsymbol{P}_t)\}_{t=1}^n$ *is i.i.d. or m–dependent.*

Proof of Condition 3.6. By assumption the observations are i.i.d. □

Condition 3.7 *There exists a constant* $C_1 > 0$ (*independent of* n) *such that for all sufficiently small* $\varepsilon > 0$ *and for every sieve* Θ_n,

$$\sup_{\{\theta\in\Theta_n :\, d(\theta,\theta_0)\leq\varepsilon\}} \boldsymbol{Var}\big(\ell(\theta,Z) - \ell(\theta_0,Z)\big) \leq C_1\,\varepsilon^2.$$

Proof of Condition 3.7. Fix $\theta = (\boldsymbol{\beta}, h)$ and write $\eta = (\boldsymbol{u}, g) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0, h - h_0)$. Decompose

$$\ell(\theta,Z) - \ell(\theta_0,Z) = \delta\Big(\boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P}) - \big[\log s^{(0)}(\theta,Y) - \log s^{(0)}(\theta_0,Y)\big]\Big).$$

By the mean value theorem, for some $t \in (0,1)$,

$$\log s^{(0)}(\theta,Y) - \log s^{(0)}(\theta_0,Y) = \frac{s^{(1)}(\theta_t,Y)[\eta]}{s^{(0)}(\theta_t,Y)}, \qquad \theta_t = \theta_0 + t\eta,$$

with

$$s^{(1)}(\theta,Y)[\eta] = \mathbb{E}\big[I(\tilde{Y}\geq Y)\exp(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}))\big(\boldsymbol{X}^\top\boldsymbol{u} + g(\boldsymbol{P})\big)\big].$$

By Lemma 3 there exists $c_0 > 0$ such that $s^{(0)}(\theta_t,Y) \geq c_0$ uniformly in t, θ, Y. Using Assumption 2, we have $\|\boldsymbol{X}\| \leq M_{\mathcal{X}}$ and $\|h\|_\infty \leq M_{\mathcal{H}}$ on Θ_n, and therefore

$$\left|\log s^{(0)}(\theta,Y) - \log s^{(0)}(\theta_0,Y)\right| \leq C\left(\|\boldsymbol{u}\| + \|g\|_\infty\right) = C\,d(\theta,\theta_0)$$

for some constant C that does not depend on n or on the sieve. Therefore,

$$\left|\ell(\theta,Z)-\ell(\theta_0,Z)\right| \le \delta\Big(\|\boldsymbol{X}\|\,\|\boldsymbol{u}\|+\|g\|_\infty\Big)+\delta\,C\,d(\theta,\theta_0)\le C'\,d(\theta,\theta_0),$$

using $\|\boldsymbol{X}\|\le M_{\mathcal{X}}$ a.s. Hence

$$\boldsymbol{Var}\big(\ell(\theta,Z)-\ell(\theta_0,Z)\big)\le \mathbb{E}\Big[\big(\ell(\theta,Z)-\ell(\theta_0,Z)\big)^2\Big]\le (C')^2\,d(\theta,\theta_0)^2\le C_1\,\varepsilon^2,$$

uniformly over $d(\theta,\theta_0)\le\varepsilon$ and $\theta\in\Theta_n$. □

Condition 3.8 *For any $\delta>0$ there exists $s\in(0,2)$ and a random variable $U(Z)$ with $\mathbb{E}\big[(U(Z))^\gamma\big]\le C_2$ for some $\gamma\ge 2$ such that*

$$\sup_{\{\theta\in\Theta_n:\,d(\theta,\theta_0)\le\delta\}}\left|\ell(\theta,Z)-\ell(\theta_0,Z)\right|\le\delta^s\,U(Z).$$

In fact we can take $s=1$ and $U(Z)\equiv C$ (a finite constant).

Proof of Condition 3.8. The Lipschitz estimate obtained in Condition 3.7 yields, for all θ in a δ–ball around θ_0,

$$\left|\ell(\theta,Z)-\ell(\theta_0,Z)\right|\le C'\,d(\theta,\theta_0)\le C'\,\delta.$$

Thus the inequality holds with $s=1$ and $U(Z)\equiv C'$. Since C' is deterministic, $\mathbb{E}\big[(U(Z))^\gamma\big]=C'^\gamma<\infty$ for any $\gamma\ge 2$, which verifies Condition 3.8. □

D. Proof of Theorem 3

Proof We apply Theorem 4.3 of Chen (2007) verifying Conditions 4.1–4.5 for the functional $f(\theta)=\boldsymbol{a}^\top\boldsymbol{\beta}$ and then use Cramér–Wold to obtain the joint limit for $\boldsymbol{\beta}$; here $\boldsymbol{a}\in\mathbb{R}^b$ is arbitrary.

By Theorem 4.3 of Chen (2007) and our rate assumption $\|\hat{\theta}_n-\theta_0\|^2=o_p(n^{-1/2})$ (implied by $\varepsilon_n^2=o(n^{-1/2})$ and $\|\hat{\theta}_n-\theta_0\|=O_p(\varepsilon_n)$), it holds

$$\sqrt{n}\,\big(\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}_n-\boldsymbol{a}^\top\boldsymbol{\beta}_0\big)\ \Rightarrow\ \mathcal{N}\big(0,\ \sigma^2_{v^*_a}\big).$$

Cramér–Wold Theorem then yields $\sqrt{n}\,(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\Rightarrow\mathcal{N}(0,\Sigma_\beta)$, for a suitable $b\times b$ covariance matrix Σ_β.

Note that, since the penalty is orthogonal to $\boldsymbol{\beta}$ scores and its magnitude is $O(\lambda_n)=o(n^{-1/2})$, it does not affect the first-order limit for $\hat{\boldsymbol{\beta}}_n$.

We now prove Conditions 4.1–4.5 of Chen (2007). Write $\theta=(\boldsymbol{\beta},h)$, $\theta_0=(\boldsymbol{\beta}_0,h_0)$, and define the population unpenalized partial log-likelihood contribution

$$\ell(\theta,Z)=\delta\Big(\boldsymbol{X}^\top\boldsymbol{\beta}+h(\boldsymbol{P})-\log s^{(0)}(\boldsymbol{\beta},h,Y)\Big),\qquad s^{(0)}(\boldsymbol{\beta},h,t)=\mathbb{E}\big[I(Y\ge t)\,e^{\boldsymbol{X}^\top\boldsymbol{\beta}+h(\boldsymbol{P})}\big].$$

Let $\hat{Q}_n(\theta)$ denote the actual penalized sample criterion of (3), where the population term $s^{(0)}$ is replaced by the empirical $S_n^{(0)}$ and a penalty $\frac{\lambda_n}{2}\int_\Omega(\Delta h)^2$ is subtracted.

Let $V = \{\xi = (\boldsymbol{u}, g) : \boldsymbol{u} \in \mathbb{R}^b,\ g \in \mathcal{H}\}$ and define the Fisher-type bilinear form at θ_0

$$\langle \xi_1, \xi_2 \rangle = -\mathbb{E}\big[D^2\ell(\theta_0, Z)[\xi_1, \xi_2]\big] = \mathbb{E}\Big[\delta\ \boldsymbol{Cov}_Y\Big(A_{\xi_1}, A_{\xi_2}\Big)\Big],$$
$$A_\xi(\boldsymbol{X}, \boldsymbol{P}) = \boldsymbol{X}^\top \boldsymbol{u} + g(\boldsymbol{P}),$$

where $\boldsymbol{Cov}_Y(\cdot,\cdot)$ denotes the covariance under the risk-set weights at time Y (see the proof of Theorem 2). This is an inner product on the completion $\overline{V}$ by the strict curvature in Lemma 5. For fixed $\boldsymbol{a} \in \mathbb{R}^b$, the Riesz representer $v_{\boldsymbol{a}}^* = (\boldsymbol{u}_{\boldsymbol{a}}^*, g_{\boldsymbol{a}}^*) \in \overline{V}$ is defined by

$$\langle \xi, v_{\boldsymbol{a}}^* \rangle = \frac{\partial}{\partial\theta}(\boldsymbol{a}^\top \boldsymbol{\beta}_0)[\xi] = \boldsymbol{a}^\top \boldsymbol{u} \quad \text{for all } \xi = (\boldsymbol{u}, g) \in V.$$

Let $\pi_n v_{\boldsymbol{a}}^*$ be the V_n-projection (as in Sect. 3.2).

Decompose the gap between the sample penalized criterion and the population sample average:

$$\hat{Q}_n(\theta) = P_n\ell(\theta, \cdot) + P_n\big(\ell_n - \ell\big)(\theta) - \frac{\lambda_n}{2}\int_\Omega (\Delta h)^2,$$

where ℓ_n is obtained from ℓ by replacing $s^{(0)}$ with $S_n^{(0)}$.

By the entropy bound used in the proof of Theorem 2 and the growth rule $K(n)\log K(n) = o(n)$, the class

$$\mathcal{G}_n = \Big\{(z, t) \mapsto I(Y \ge t)e^{\boldsymbol{X}^\top \boldsymbol{\beta} + h(\boldsymbol{P})} : \theta \in \Theta_n,\ t \ge 0\Big\}$$

is Glivenko–Cantelli and Donsker uniformly in a Θ-neighborhood of θ_0. Hence

$$\sup_{\theta\in\Theta_n,\ t\ge 0} \big|S_n^{(0)}(\theta, t) - s^{(0)}(\theta, t)\big| = O_p\Big(\sqrt{\tfrac{K(n)\log K(n)}{n}}\Big) = O_p(\delta_n) \tag{19}$$

and therefore

$$\sup_{\theta\in\Theta_n} \big|P_n(\ell_n - \ell)(\theta)\big| = O_p(\delta_n). \tag{20}$$

Moreover the penalty does not depend on Z, so it vanishes inside empirical process terms; it only perturbs the objective level by $O(\lambda_n)$ and its directional derivative in any direction $\xi = (\boldsymbol{u}, g)$ equals $-\lambda_n \int_\Omega (\Delta h)(\Delta g)$, which is zero whenever $\boldsymbol{u} \ne 0$ and $g = 0$. Thus, the penalty does not change the first-order behavior of $\boldsymbol{\beta}$. Its only influence comes indirectly through the nuisance part g, and this enters the β–score as a product of the penalty weight and the nuisance estimation error, i.e., $O_p(\lambda_n \|\hat{h} - h_0\|_\infty) = o_p(n^{-1/2})$ by assumption as $\|\hat{h} - h_0\| = O_p(\varepsilon_n)$. Hence the effect vanishes under $\sqrt{n}$–scaling. □

Condition 4.1 *(i) There is* $\omega > 0$ *such that* $|f(\theta) - f(\theta_o) - \frac{\partial f(\theta_o)}{\partial \theta}[\theta - \theta_o]| = O(\|\theta - \theta_o\|^{\omega})$ *uniformly in* $\theta \in \Theta_n$ *with* $\|\theta - \theta_o\| = o(1)$; *(ii)* $\left\| \frac{\partial f(\theta_o)}{\partial \theta} \right\| < \infty$; *(iii) there is* $\pi_n v^* \in \Theta_n$ *such that* $\|\pi_n v^* - v^*\| \times \|\hat{\theta}_n - \theta_o\| = o_P(n^{-1/2})$.

Proof of Condition 4.1. Condition 4.1 (*i*) and (*ii*) hold with $\omega = 2$ because $f(\theta) = \boldsymbol{a}^\top \boldsymbol{\beta}$ is linear and $\|\partial f(\theta_0)/\partial\theta\| < \infty$. For 4.1 (*iii*), by the FEM approximation bound (4), $\|\pi_n v_{\boldsymbol{a}}^* - v_{\boldsymbol{a}}^*\| = O(\eta(n))$, and $\|\hat{\theta}_n - \theta_0\| = O_p(\varepsilon_n)$ by Theorem 2. Hence

$$\|\pi_n v_{\boldsymbol{a}}^* - v_{\boldsymbol{a}}^*\| \cdot \|\hat{\theta}_n - \theta_0\| = O_p(\eta(n)\varepsilon_n) = o_p(n^{-1/2}),$$

verifying 4.1 (*iii*). □

Condition 4.2 *Let* $\mu_n(g(Z)) = n^{-1} \sum_{i=1}^n g(z_i) - E[g(Z_i)]$ *denote the empirical process indexed by a function* g. *Then,*

$$\sup_{\{\theta \in \Theta_n : \|\theta - \theta_o\| \le \delta_n\}} \mu_n\left(l(\theta, Z) - l(\theta \pm \varepsilon_n \pi_n v^*, Z) - \frac{\partial l(\theta_o, Z)}{\partial \theta}[\pm \varepsilon_n \pi_n v^*] \right) = O_P(\varepsilon_n^2).$$

Proof of Condition 4.2. The map $\theta \mapsto \ell(\theta, Z)$ is twice Fréchet differentiable in a neighborhood of θ_0 (Lemma 4); thus pathwise differentiability holds. Using the Donsker property and (19)–(20), we obtain

$$\sup_{\{\theta \in \Theta_n :\ \|\theta - \theta_0\| \le \delta_n\}} \left| \mu_n\left(\dot{\ell}(\theta, Z)[\pi_n v_{\boldsymbol{a}}^*] - \dot{\ell}(\theta_0, Z)[\pi_n v_{\boldsymbol{a}}^*] \right) \right| = o_p(n^{-1/2}),$$

which is Condition 4.2′ and hence implies 4.2; see Chen (2007). □

Condition 4.3 $K(\theta_o, \hat{\theta}_n) - K(\theta_o, \hat{\theta}_n \pm \varepsilon_n \pi_n v^*) = \pm \varepsilon_n (\hat{\theta}_n - \theta_o, \pi_n v^*) + o(n^{-1})$, *where* $K(\theta_0, \theta) \equiv \mathbb{E}[l(\theta_0, Z_i) - l(\theta, Z_i)]$.

Proof of Condition 4.3. Taylor-expanding the population criterion $Q(\theta) = P\ell(\theta, \cdot)$ along the line $\theta_0 \pm \varepsilon_n \pi_n v_{\boldsymbol{a}}^*$ and using that $-D^2 Q(\theta_0)[\xi_1, \xi_2] = \langle \xi_1, \xi_2 \rangle$ with linear remainder $o(\varepsilon_n^2)$, we get

$$\begin{aligned} \mathbb{E}\left[\dot{\ell}(\hat{\theta}_n, \cdot)[\pi_n v_{\boldsymbol{a}}^*]\right] &= \langle \hat{\theta}_n - \theta_0, \pi_n v_{\boldsymbol{a}}^* \rangle + o(n^{-1/2}) \\ &= \langle \hat{\theta}_n - \theta_0, v_{\boldsymbol{a}}^* \rangle + o(n^{-1/2}) = \boldsymbol{a}^\top(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o(n^{-1/2}), \end{aligned}$$

which is Condition 4.3′; hence 4.3 holds. The penalty contributes nothing to these equalities in the $\boldsymbol{\beta}$ direction, and its contribution along g is of order $o(\lambda_n \varepsilon_n) = o(n^{-1/2})$. □

Condition 4.4 [.4.4] (*i*) $\mu_n\left(\frac{\partial l(\theta_o, Z)}{\partial \theta}[\pi_n v^* - v^*] \right) = o_P(n^{-1/2})$; (*ii*) $E\left\{ \frac{\partial l(\theta_o, Z)}{\partial \theta}[\pi_n v^*] \right\} = o(n^{-1/2})$.

Proof of Condition 4.4. Because $\|\pi_n v_{\boldsymbol{a}}^* - v_{\boldsymbol{a}}^*\| = O(\eta(n))$ and the score map $\xi \mapsto \dot{\ell}(\theta_0, Z)[\xi]$ is square-integrable and continuous in ξ under $\langle\cdot,\cdot\rangle$, we have

$$\mu_n\Big(\dot{\ell}(\theta_0,\cdot)[\pi_n v_{\boldsymbol{a}}^* - v_{\boldsymbol{a}}^*]\Big) = o_p(n^{-1/2}),$$
$$\mathbb{E}\Big[\dot{\ell}(\theta_0, Z)[\pi_n v_{\boldsymbol{a}}^*]\Big] = \mathbb{E}\Big[\dot{\ell}(\theta_0, Z)[v_{\boldsymbol{a}}^*]\Big] + o(n^{-1/2}) = o(n^{-1/2}),$$

since $\mathbb{E}[\dot{\ell}(\theta_0, Z)[v_{\boldsymbol{a}}^*]] = 0$ by the definition of the representer and the score identity. □

Condition 4.5 $n^{1/2}\,\mu_n\Big(\frac{\partial l(\theta_o,Z)}{\partial\theta}[v^*]\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{v^*}), \quad$ with $\sigma^2_{v^*} > 0.$

Proof of Condition 4.5 The i.i.d. assumption and finite variance imply the Central Limit Theorem for the score in direction $v_{\boldsymbol{a}}^*$:

$$\sqrt{n}\,\mu_n\Big(\dot{\ell}(\theta_0,\cdot)[v_{\boldsymbol{a}}^*]\Big) \to_d \mathcal{N}\big(0, \sigma^2_{v_{\boldsymbol{a}}^*}\big), \qquad \sigma^2_{v_{\boldsymbol{a}}^*} = \boldsymbol{Var}\Big(\dot{\ell}(\theta_0, Z)[v_{\boldsymbol{a}}^*]\Big) > 0,$$

which is Condition 4.5. Note that the penalty does not enter the score, and the $S_n^{(0)}$ instead of $s^{(0)}$ replacement alters the criterion by $O_p(\delta_n)$ uniformly by (20); combined with the fact that $\delta_n\epsilon_n = o(n^{-1/2})$, its contribution to the linearization is negligible for the $\sqrt{n}$-asymptotics. □

E. Lemmas

Lemma 1 *Let $\{\mathcal{T}_{\eta(n)}\}_{n\geq 1}$ be a sequence of conforming, shape-regular triangulations of a bounded C^2 domain $\Omega \subset \mathbb{R}^2$ such that the refinement is nested, i.e., every $\mathrm{T} \in \mathcal{T}_{\eta(n)}$ is a union of triangles from $\mathcal{T}_{n+1}$. Let V_n be the Argyris C^1 finite element space on $\mathcal{T}_{\eta(n)}$, i.e.,*

$$V_n = \Big\{\ v \in C^1(\overline{\Omega}) \ :\ \ v|_{\mathrm{T}} \in \mathbb{P}_5(\mathrm{T}) \text{ for all } \mathrm{T} \in \mathcal{T}_{\eta(n)}\ \Big\}.$$

Define $H_n = V_n \cap H^2_{\mathbf{n}}(\Omega)$ and, for a fixed $M_{\mathcal{H}} > 0$,

$$\mathcal{H}_n = \Big\{\ h \in H_n\ :\ \int_\Omega h = 0,\ \|h\|_{H^2(\Omega)} \leq M_{\mathcal{H}}\ \Big\}.$$

Then the spaces are nested:

$$V_n \subset V_{n+1}, \qquad H_n \subset H_{n+1}, \qquad \mathcal{H}_n \subset \mathcal{H}_{n+1} \quad \text{for all } n \geq 1.$$

Proof Fix $v \in V_n$. For each triangle $\mathrm{T} \in \mathcal{T}_{\eta(n)}$, $v|_{\mathrm{T}} \in \mathbb{P}_5(\mathrm{T})$ is a polynomial. By nested refinement, T is partitioned into $\{\mathrm{T}'\}_{\mathrm{T}'\subset\mathrm{T},\,\mathrm{T}'\in\mathcal{T}_{n+1}}$, and for each $\mathrm{T}' \subset \mathrm{T}$ the restriction $v|_{\mathrm{T}'}$ is still a polynomial of total degree ≤ 5, i.e., $v|_{\mathrm{T}'} \in \mathbb{P}_5(\mathrm{T}')$. Hence v is piecewise $\mathbb{P}_5$ on $\mathcal{T}_{\eta(n+1)}$. Global C^1-continuity on $\mathcal{T}_{n+1}$ follows because across any fine edge lying strictly inside the triangle T, both traces of v come from the same polynomial $v|_T$ and thus coincide smoothly, while across fine edges that coincide with a coarse edge, C^1 continuity holds since $v \in C^1(\overline{\Omega})$ already. Therefore $v \in V_{n+1}$, so $V_n \subset V_{n+1}$.

Since $H_n = V_n \cap H^2_{\mathbf{n}}(\Omega)$ and $H_{n+1} = V_{n+1} \cap H^2_{\mathbf{n}}(\Omega)$, the inclusion $V_n \subset V_{n+1}$ implies $H_n \subset H_{n+1}$, as the homogeneous Neumann boundary condition does not depend on the mesh.

Finally, if $h \in \mathcal{H}_n$, then $h \in H_n \subset H_{n+1}$. The mean-zero constraint $\int_\Omega h = 0$ and the uniform bound $\|h\|_{H^2(\Omega)} \le M_{\mathcal{H}}$ are independent of n, so $h \in \mathcal{H}_{n+1}$. Thus the nestedness $\mathcal{H}_n \subset \mathcal{H}_{n+1}$ holds as well. □

Lemma 2 *Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain and fix $C > 0$. Define*

$$\mathcal{B}_C = \left\{ u \in H^2(\Omega) \,\middle|\, \|u\|_{H^2(\Omega)} \le C, \int_\Omega u\,dx = 0 \right\}.$$

Then $\mathcal{B}_C$ is compact in $L^\infty(\Omega)$.

Proof By the Sobolev–Morrey embedding on bounded Lipschitz domains in $d = 2$,

$$H^2(\Omega) \hookrightarrow C^{0,\alpha}(\overline{\Omega}) \quad \text{for any } 0 < \alpha < 1,$$

continuously: there exists $K = K(\Omega, \alpha)$ such that

$$\|u\|_{C^{0,\alpha}(\overline{\Omega})} \le K\|u\|_{H^2(\Omega)} \quad \forall\, u \in H^2(\Omega).$$

Hence for $u \in \mathcal{B}_C$ we have $\|u\|_{C^{0,\alpha}} \le KC$. Thus $\mathcal{B}_C$ is uniformly bounded and Hölder–equicontinuous in $C^{0,\alpha}(\overline{\Omega})$.

Since $\overline{\Omega}$ is compact, the embedding $C^{0,\alpha}(\overline{\Omega}) \hookrightarrow C^0(\overline{\Omega})$ is compact (Arzelà–Ascoli). Therefore $\mathcal{B}_C$ is relatively compact in $C^0(\overline{\Omega})$, and hence in $L^\infty(\Omega)$.

It remains to show $\mathcal{B}_C$ is closed in L^∞. Let $u_k \in \mathcal{B}_C$ and $u_k \to u$ in $L^\infty(\Omega)$. The sequence (u_k) is bounded in the Hilbert space $H^2(\Omega)$, so (after extracting a subsequence, not relabeled) $u_k \rightharpoonup v$ weakly in $H^2(\Omega)$ for some $v \in H^2(\Omega)$. Since $u_k \to u$ in L^∞, in particular $u_k \to u$ in L^2; passing to the limit in L^2 shows $v = u$ in L^2, hence $u \in H^2(\Omega)$. By weak lower semicontinuity,

$$\|u\|_{H^2} \le \liminf_{k\to\infty} \|u_k\|_{H^2} \le C.$$

Moreover,

$$\int_\Omega u = \lim_{k\to\infty} \int_\Omega u_k = 0$$

since $u_k \to u$ in L^1 (bounded domain and L^∞–convergence). Thus $u \in \mathcal{B}_C$. Hence $\mathcal{B}_C$ is closed in L^∞. Relative compactness with closedness in L^∞ implies that $\mathcal{B}_C$ is compact in $L^\infty(\Omega)$. □

Lemma 3 *Equip* $\Theta = \mathcal{B} \times \mathcal{H}$ *with the metric* $d\big((\boldsymbol{\beta}, h), (\tilde{\boldsymbol{\beta}}, \tilde{h})\big) = \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\| + \|h - \tilde{h}\|_\infty$. *Under Assumptions 1–2, the mapping*

$$Q(\theta) = \mathbb{E}\Big[\delta\big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}) - \log s^{(0)}(\boldsymbol{\beta}, h, Y)\big)\Big], \qquad \theta = (\boldsymbol{\beta}, h) \in \Theta,$$

is continuous on Θ.

Proof Write $s^{(0)}(\boldsymbol{\beta}, h, y) = \mathbb{E}\big[I(Y \geq y)\exp\{\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P})\}\big]$. By Assumption 2(i) there is $M_\mathcal{X} < \infty$ with $\|\boldsymbol{X}\| \leq M_\mathcal{X}$ a.s., and by the definition of $\mathcal{B}$ we have $\|\boldsymbol{\beta}\| \leq M_\mathcal{B}$. Moreover, by Sobolev embedding in $d = 2$ applied to $\mathcal{H}$ (bounded in H^2), there exists $M_\infty < \infty$ such that $\|h\|_\infty \leq M_\infty$ for all $h \in \mathcal{H}$. Set

$$C_* = M_\mathcal{B} M_\mathcal{X} + M_\infty.$$

Then for all $(\boldsymbol{\beta}, h) \in \Theta$ and all $y \geq 0$,

$$e^{-C_*}\,\mathbb{P}(Y \geq y) \;\leq\; s^{(0)}(\boldsymbol{\beta}, h, y) \;\leq\; e^{C_*}\,\mathbb{P}(Y \geq y), \tag{21}$$

since $e^{\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P})} \in [e^{-C_*}, e^{C_*}]$ a.s. In particular, $s^{(0)}(\boldsymbol{\beta}, h, Y) > 0$ a.s.

Fix two parameters $\theta_1 = (\boldsymbol{\beta}_1, h_1)$ and $\theta_2 = (\boldsymbol{\beta}_2, h_2)$. By the mean–value theorem for the exponential and the bound above,

$$\begin{aligned}
\big|s^{(0)}(\theta_1, y) - s^{(0)}(\theta_2, y)\big| &= \Big|\mathbb{E}\big[I(Y \geq y)\big(e^{\eta_1} - e^{\eta_2}\big)\big]\Big| \\
&\leq \mathbb{E}\Big[I(Y \geq y)\, e^{\max\{\eta_1, \eta_2\}}\Big|(\boldsymbol{X}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)) + (h_1 - h_2)(\boldsymbol{P})\Big|\Big] \\
&\leq e^{C_*}\,\mathbb{P}(Y \geq y)\,\big(M_\mathcal{X}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|h_1 - h_2\|_\infty\big),
\end{aligned}$$

where $\eta_j = \boldsymbol{X}^\top\boldsymbol{\beta}_j + h_j(\boldsymbol{P})$. Using (21) and the mean–value theorem for log,

$$\begin{aligned}
\big|\log s^{(0)}(\theta_1, y) - \log s^{(0)}(\theta_2, y)\big| \;&\leq\; \frac{\big|s^{(0)}(\theta_1, y) - s^{(0)}(\theta_2, y)\big|}{\min\{s^{(0)}(\theta_1, y), s^{(0)}(\theta_2, y)\}} \\
&\leq\; e^{2C_*}\,\big(M_\mathcal{X}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|h_1 - h_2\|_\infty\big),
\end{aligned}$$

where the factor $\mathbb{P}(Y \geq y)$ cancels, so the bound is uniform in y.

Define, for $z = (Y, \delta, \boldsymbol{X}, \boldsymbol{P})$ and $\theta = (\boldsymbol{\beta}, h)$,

$$\zeta(z;\theta) = \delta\Big(\boldsymbol{X}^\top\boldsymbol{\beta} + h(\boldsymbol{P}) - \log s^{(0)}(\boldsymbol{\beta}, h, Y)\Big).$$

Then for any $\theta_1, \theta_2 \in \Theta$,

$$\begin{aligned}
\big|\zeta(z;\theta_1) - \zeta(z;\theta_2)\big| &\leq \delta\Big(\big|\boldsymbol{X}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\big| + \big|(h_1 - h_2)(\boldsymbol{P})\big| + \big|\log s^{(0)}(\theta_1, Y) - \log s^{(0)}(\theta_2, Y)\big|\Big) \\
&\leq \delta\, L_*\Big(\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| + \|h_1 - h_2\|_\infty\Big),
\end{aligned}$$

with $L_* = (1 + e^{2C_*})\max\{M_\mathcal{X}, 1\}$, which is deterministic and finite. Taking expectations and using $\delta \leq 1$,

$$|Q(\theta_1) - Q(\theta_2)| \le \mathbb{E}\big[\,|\zeta(z;\theta_1) - \zeta(z;\theta_2)|\,\big] \le L_*\, d(\theta_1,\theta_2).$$

Thus Q is globally Lipschitz on (Θ, d). □

Lemma 4 *Consider the metric space (Θ, d) of Lemma 3 and retain Assumptions 1–2. Let $\theta = (\boldsymbol{\beta}, h) \in \Theta$, and directions $\eta_1 = (\boldsymbol{u}_1, g_1)$, $\eta_2 = (\boldsymbol{u}_2, g_2)$ in the linear space $\mathbb{R}^b \times \{g : \|g\|_\infty < \infty, \int_\Omega g = 0\}$ such that $\theta + t\eta_r \in \Theta$ for all sufficiently small t (e.g., θ is in the relative interior of Θ). For $k = 0, 1, 2$ define*

$$s^{(k)}(\theta, y)[\eta_1, \dots, \eta_k] = \mathbb{E}\Big[\mathbf{1}\{Y' \ge y\}\ \exp\big(\boldsymbol{X}'^{\top}\boldsymbol{\beta} + h(\boldsymbol{P}')\big)\ \prod_{r=1}^{k}\big(\boldsymbol{X}'^{\top}\boldsymbol{u}_r + g_r(\boldsymbol{P}')\big)\Big],$$

where $(Y', \boldsymbol{X}', \boldsymbol{P}')$ is an i.i.d. copy of $(Y, \boldsymbol{X}, \boldsymbol{P})$ independent of $(Y, \boldsymbol{X}, \boldsymbol{P})$. Then $Q : \Theta \to \mathbb{R}$ is twice Fréchet differentiable on the relative interior of Θ, with

$$DQ(\theta)[\eta_1] = \mathbb{E}\Big[\,\delta\Big(\boldsymbol{X}^{\top}\boldsymbol{u}_1 + g_1(\boldsymbol{P}) - \frac{s^{(1)}(\theta, Y)[\eta_1]}{s^{(0)}(\theta, Y)}\Big)\Big],$$

$$D^2Q(\theta)[\eta_1, \eta_2] = \mathbb{E}\Big[-\delta\ \frac{s^{(2)}(\theta, Y)[\eta_1, \eta_2]\, s^{(0)}(\theta, Y) - s^{(1)}(\theta, Y)[\eta_1]\, s^{(1)}(\theta, Y)[\eta_2]}{\big(s^{(0)}(\theta, Y)\big)^2}\Big].$$

Assumption 2 (i) gives $\|\boldsymbol{X}\| \le M_{\mathcal{X}}$ a.s., $\mathcal{B}$ is compact so $\|\boldsymbol{\beta}\| \le M_{\mathcal{B}}$, and the Sobolev–Morrey embedding yields a uniform bound $\|h\|_\infty \le C_H M_{\mathcal{H}}$ for all $h \in \mathcal{H}$. Let $C_* = M_{\mathcal{B}} M_{\mathcal{X}} + C_H M_{\mathcal{H}}$. Then for any $\theta = (\boldsymbol{\beta}, h)$ and $y \ge 0$,

$$e^{-C_*}\,\mathbb{P}(Y' \ge y)\ \le\ s^{(0)}(\theta, y)\ \le\ e^{C_*}\,\mathbb{P}(Y' \ge y). \tag{22}$$

Moreover, for $\eta = (\boldsymbol{u}, g)$ write $L(\eta) = M_{\mathcal{X}}\|\boldsymbol{u}\| + \|g\|_\infty$. By the mean–value theorem and boundedness of the exponential,

$$|s^{(1)}(\theta, y)[\eta]| \le e^{C_*}\,\mathbb{P}(Y' \ge y)\, L(\eta),$$
$$|s^{(2)}(\theta, y)[\eta_1, \eta_2]| \le e^{C_*}\,\mathbb{P}(Y' \ge y)\, L(\eta_1) L(\eta_2).$$

Combining with (22) yields the uniform-in-y ratio bounds

$$\left|\frac{s^{(1)}(\theta, y)[\eta]}{s^{(0)}(\theta, y)}\right| \le e^{2C_*} L(\eta), \qquad \left|\frac{s^{(2)}(\theta, y)[\eta_1, \eta_2]}{s^{(0)}(\theta, y)}\right| \le e^{2C_*} L(\eta_1) L(\eta_2), \tag{23}$$

and likewise $\big|s^{(1)}(\theta, y)[\eta_1]\, s^{(1)}(\theta, y)[\eta_2] / (s^{(0)}(\theta, y))^2\big| \le e^{4C_*} L(\eta_1) L(\eta_2)$. Now, for $t \mapsto \theta + t\eta_1$ define

$$\Psi_t(Y, \delta, \boldsymbol{X}, \boldsymbol{P}) = \boldsymbol{X}^{\top}\boldsymbol{u}_1 + g_1(\boldsymbol{P}) - \frac{s^{(1)}(\theta + t\eta_1, Y)[\eta_1]}{s^{(0)}(\theta + t\eta_1, Y)}.$$

By dominated convergence, we have $\Psi_t \to \Psi_0$ a.s., and the uniform ratio bound (23) gives $|\Psi_t| \le C(1 + L(\eta_1))$ for a deterministic C. Thus

$$\frac{Q(\theta + t\eta_1) - Q(\theta)}{t} = \mathbb{E}\big[\delta\,\Psi_t\big] \;\longrightarrow\; \mathbb{E}\big[\delta\,\Psi_0\big],$$

which is the stated formula for $DQ(\theta)[\eta_1]$.

Similarly, consider $t \mapsto DQ(\theta + t\eta_2)[\eta_1]$. By the chain rule for log, Taylor's formula, and the definitions of $s^{(k)}$,

$$\frac{d}{dt}\Big|_{t=0}\left(\frac{s^{(1)}(\theta + t\eta_2, y)[\eta_1]}{s^{(0)}(\theta + t\eta_2, y)}\right) = \frac{s^{(2)}(\theta, y)[\eta_1, \eta_2]\, s^{(0)}(\theta, y) - s^{(1)}(\theta, y)[\eta_1]\, s^{(1)}(\theta, y)[\eta_2]}{(s^{(0)}(\theta, y))^2}.$$

The interchange of differentiation and expectation that defines $s^{(k)}$ is justified by dominated convergence using the same e^{C_*} and $L(\cdot)$ bounds, and the ratio is bounded uniformly in y by (23). Therefore

$$\frac{DQ(\theta + t\eta_2)[\eta_1] - DQ(\theta)[\eta_1]}{t} \longrightarrow -\mathbb{E}\left[\delta\, \frac{s^{(2)}(\theta, Y)[\eta_1, \eta_2]\, s^{(0)}(\theta, Y) - s^{(1)}(\theta, Y)[\eta_1]\, s^{(1)}(\theta, Y)[\eta_2]}{(s^{(0)}(\theta, Y))^2}\right].$$

Again, the integrand is dominated by a deterministic constant times $L(\eta_1)L(\eta_2)$, so the limit passes through the expectation. Bilinearity and continuity in (η_1, η_2) are immediate from the bounds, which also give the $o(\|\eta_2\|)$ remainder uniformly in η_1. Thus Q is C^2 Fréchet on the interior of Θ with the stated DQ and D^2Q.

Lemma 5 *$Q(\theta)$ attains its unique global maximum in (Θ, d) at $\theta_0 = (\beta_0, h_0)$.*

Proof By Lemma 3, Q is continuous on (Θ, d) with $d\big((\boldsymbol{\beta}, h), (\tilde{\boldsymbol{\beta}}, \tilde{h})\big) = \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\| + \|h - \tilde{h}\|_\infty$. By compactness of $\mathcal{B}$ (Assumption 1 (iii)) and the compact embedding $H^2(\Omega) \hookrightarrow C^0(\Omega)$ in $d = 2$, the set $\mathcal{H} = \{h \in H^2_{\mathbf{n}}(\Omega) : \int_\Omega h = 0,\ \|h\|_{H^2} \le M_{\mathcal{H}}\}$ is compact in the $\|\cdot\|_\infty$–topology (see Lemma 2). Hence $\Theta = \mathcal{B} \times \mathcal{H}$ is compact under d, and by Weierstrass Q attains a global maximum on Θ. Uniqueness follows from strict concavity of Q, see (13). The fact that the maximum is θ_0 is implied by Theorem 1. □

Author contributions Lorenzo Tedesco and Francesco Finazzi have contributed equally to this work. Both authors were equally involved in the conception and design of the study, the development of the methodology, the analysis and interpretation of the results, and the drafting and revision of the manuscript. Both authors have read and approved the final version of the manuscript.

Funding Open access funding provided by Università degli Studi di Bergamo within the CRUI-CARE Agreement.

Data Availability The data and code that support the findings of this study are openly available at https://github.com/tedescolor/SPCOX.

References

Cox DR (1972) Regression models and life tables (with discussion). J R Statist Soc B 34:187–220
Tibshirani R, Hastie T (1987) Local likelihood estimation. J Am Stat Assoc 82(398):559–567
Fan J, Gijbels I, King M (1997) Local likelihood and local partial likelihood in hazard regression. Ann Stat 25(4):1661–1690
Chen K, Guo S, Sun L, Wang J-L (2010) Global partial likelihood for nonparametric proportional hazards models. J Am Stat Assoc 105(490):750–760
O'Sullivan F (1988) Nonparametric estimation of relative risk using splines and cross-validation. SIAM J Sci Stat Comput 9(3):531–542
Hastie T, Tibshirani R (1990) Exploring the nature of covariate effects in the proportional hazards model. Biometrics 46:1005–1016
Kooperberg C, Stone CJ, Truong YK (1995) Hazard regression. J Am Stat Assoc 90(429):78–94
Nagpal C, Yadlowsky S, Rostamzadeh N, Heller K (2021) Deep Cox mixtures for survival regression. Machine learning for healthcare conference. PMLR, pp 674–708
Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y (2018) Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Med Res Methodol 18(1):24
Wiegrebe S, Kopper P, Sonabend R, Bischl B, Bender A (2024) Deep learning for survival analysis: a review. Artif Intell Rev 57(3):65
Hanson T, Zhou H (2014) Spatial survival model. Wiley StatsRef Statistics Reference Online, pp 1–8
Li Y, Lin X (2006) Semiparametric normal transformation models for spatially correlated survival data. J Am Stat Assoc 101(474):591–603
Xue Y, Schifano ED, Hu G (2020) Geographically weighted Cox regression for prostate cancer survival data in Louisiana. Geogr Anal 52(4):570–587
Hu G, Huffer F (2020) Modified Kaplan-Meier estimator and Nelson-Aalen estimator with geographical weighting for survival data. Geogr Anal 52(1):28–48
Paik J, Ying Z (2012) A composite likelihood approach for spatially correlated survival data. Comput Stat Data Anal 56(1):209–216
Banerjee S, Wall MM, Carlin BP (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. Biostatistics 4(1):123–142
Taylor BM, Rowlingson BS (2017) spatsurv: An R package for Bayesian inference with spatial survival models. J Stat Softw 77:1–32
Geng L, Hu G (2022) Bayesian spatial homogeneity pursuit for survival data with an application to the seer respiratory cancer data. Biometrics 78(2):536–547
Zhou H, Hanson T (2018) A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. J Am Stat Assoc 113(522):571–581
Zhang Z, Stringer A, Brown P, Stafford J (2023) Bayesian inference for Cox proportional hazard models with partial likelihoods, nonlinear covariate effects and correlated observations. Stat Methods Med Res 32(1):165–180
Hennerfeind A, Brezger A, & Fahrmeir L (2006) Geoadditive survival models. J Am Stat Assoc 101(475):1065–1075
Sangalli LM, Ramsay JO, Ramsay TO (2013) Spatial spline regression models. J R Stat Soc Ser B Stat Methodol 75(4):681–703
Chen X (2007) Large sample sieve estimation of semi-nonparametric models. Handb Econ 6:5549–5632

Quarteroni A, Quarteroni S (2009) Numerical models for differential problems, vol 2. Springer, Milan
Brenner SC, Scott LR (2008) The mathematical theory of finite element methods. Springer, New York
Nocedal J, Wright SJ (2006) Numerical optimization. Springer, New York
Cygu S, Dushoff J, Bolker BM (2021) pcoxtime: Penalized Cox proportional hazard model for time-dependent covariates. arXiv:2102.02297 arXiv preprint
Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15(4):361–387
Dai B, Breheny P (2019) Cross validation approaches for penalized Cox regression. arXiv:1905.10432 arXiv preprint
Ramsay T (2002) Spline smoothing over difficult regions. J R Stat Soc Ser B Stat Methodol 64(2):307–319
Wood SN, Bravington MV, Hedley SL (2008) Soap film smoothing. J R Stat Soc Ser B Stat Methodol 70(5):931–955
Sangalli LM (2021) Spatial regression with partial differential equation regularisation. Int Stat Rev 89(3):505–531
Therneau T et al. (2015) A package for survival analysis in R. R Package Version 2(7):2014
Wood S (2018) Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL. R Package Version 1:8–23
Eddelbuettel D, François R (2011) Rcpp: Seamless R and C++ integration. J Stat Softw 40:1–18
Finazzi F (2016) The earthquake network project: toward a crowdsourced smartphone-based earthquake early warning system. Bull Seismol Soc Am 106(3):1088–1099
Finazzi F, Cotton F, Bossu R (2025) Citizens' smartphones unravel earthquake shaking in urban areas. Nat Commun 16(1):9527