Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data

Zheng Dong[1], Shixiang Zhu[1], Yao Xie[1], Jorge Mateu[2], and Francisco J. Rodríguez-Cortés[3]

[1]*H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology*
[2]*Department of Mathematics, Universitat Jaume I, Castelló de la Plana, Valencia, 12071, Spain*
[3]*Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia*

Abstract

Most COVID-19 studies commonly report figures of the overall infection at a state- or county-level, reporting the aggregated number of cases in a particular region at one time. This aggregation tends to miss out on fine details of the propagation patterns of the virus. This paper is motivated by analyzing a high-resolution COVID-19 dataset in Cali, Colombia, that provides every confirmed case's exact location and time information, offering vital insights for the spatio-temporal interaction between individuals concerning the disease spread in a metropolis. We develop a non-stationary spatio-temporal point process, assuming that previously infected cases trigger newly confirmed ones, and introduce a neural network-based kernel to capture the spatially varying triggering effect. The neural network-based kernel is carefully crafted to enhance expressiveness while maintaining results interpretability. We also incorporate some exogenous influences imposed by city landmarks. The numerical results on real data demonstrate good predictive performances of our method compared to the state-of-the-art as well as its interpretable findings.

1 Introduction

The outbreak of coronavirus disease 2019 (COVID-19) since 2020 has swept the world and is still developing. It causes a dramatic loss of human lives (Chriscaden, 2020) and presents an unprecedented challenge to public health, food systems, and the world systems. Tracking the dynamics of COVID-19 enables the human being to take target protecting measures to curb the pandemic's spread and design health surveillance systems. However, limited and biased information about local COVID-19 cases makes it extremely difficult to control strategies against the pandemic effectively.

There is a large amount of *aggregated* data consistently collected and publicly available, which contains rich information about COVID-19 cases. For instance, Johns Hopkins Center for Systems

Science and Engineering (JHU CSSE) establishes an interactive COVID-19 dashboard to track the global coronavirus development (John Hopkins University, 2020) which reports the daily confirmed cases and deaths worldwide up to the state level. New York Times (2020) also tracks daily the county-level counts of confirmed cases and deaths in the United States. Such data help the scientific researcher model the disease transmission on an aggregated level and play a pivotal role in tracking the propagation patterns of the virus and helping policymakers act effectively to revitalize economic and social development.

However, such aggregated data lack precise information about individual cases and present a significant challenge in modeling the spatio-temporal dynamics of human-to-human disease transmission when capturing the fine spatial heterogeneity of case distribution in a small region. Aggregated data cannot reflect the actual situation, which will lead the administrative officials to make biased decisions. For example, it is reported in Bizzarri et al. (2020) that the unreliable preliminary data, as well as inaccurate models, significantly affected the political decisions of the Italian administration. Another example in Guenther et al. (2020) documents a superspreader event in Germany in a meat processing plant; modeling such an event requires accounting for the precise plant location information, and aggregated data may miss such crucial local information.

In this paper, we consider an unprecedented high-resolution dataset for individual cases of COVID-19 in Cali, Colombia, the second-largest city in the country. This data records individual confirmed cases during six months, from March 15 to September 30 of 2020, with time and location information of each case. To take full advantage of the fine-grained dataset, we develop a non-stationary spatio-temporal point process model, assuming that previously infected events trigger the newly confirmed cases. We assume the triggering effect is non-stationary (Hendry and Pretis, 2016) since the virus is likely to spread more slowly in sparsely populated rural than densely populated areas. This fact entails stationary point processes non-applicable: the stationary kernel is "shift-invariant" and only depends on the temporal and location differences between events. Moreover, we consider the exogenous promotion of densely populated city landmarks in the model since the COVID-19 virus proves to spread quickly through respiratory droplets (Jasper et al., 2020), and aerosol transmission in crowded and inadequately ventilated spaces (Leclerc et al., 2020). We represent parameters of the non-stationary kernel by neural networks to enhance model flexibility while maintaining the interpretability of results. The model is estimated by solving a maximum likelihood problem via a computationally efficient strategy to tackle the intractable numerical integration in the log-likelihood function. We conduct an extensive real-data study, which reveals the unique transmission dynamics of COVID-19 and confirms that a few landmarks in the city play an essential role in spreading the virus. The model and results will help policymakers monitor coronavirus dynamics and provide a template for tracking real-time data for future epidemics and implementing health surveillance systems. Since similar high-resolution datasets will not be so rare in the future, the need for such an approach is not limited to the situation of Cali.

The paper is organized as follows. The rest of this section discusses some relevant literature on COVID-19 modeling and spatio-temporal point processes. We then introduce our motivating (and unique) dataset in Section 2. In Section 3, we review some fundamentals about point processes and propose our framework with a non-stationary spatio-temporal kernel, and illustrate our fine-crafted parameterization scheme with a simple neural network. Section 4 presents the computational strategies for model estimation with an approximation to the likelihood. Section 5 interprets the results from real data and compares them with several benchmark models. Lastly, Section 6 concludes the paper.

Related work Compartmental models are widely developed to describe the overall COVID-19 infection in a region. The simplest SIR compartmental model (Harko et al., 2014) assigns the population into three compartments with labels $\boldsymbol{S}$ (susceptible), $\boldsymbol{I}$ (infectious), and $\boldsymbol{R}$ (recovered), respectively. Deterministic differential equations fit the transition rates between each kind of compartment. Advanced compartmental models are further designed by reframing the basic one with different compartments (Lin et al., 2020; Nande et al., 2020). SEIRD in Korolev (2021) and forced SEIRD models in Loli Piccolomini and Zama (2020) are adopted in various epidemic scenarios by introducing compartments of exposed and deceased populations into the system. Other extensions, such as splitting the infected population according to infection severity (Nande et al., 2020) and introducing unreported infected population (Lin et al., 2020) are also considered. Compartmental models assume a stable population of the inspected region, thus perform well when applied to large regions such as a country or state. However, they usually do not consider detailed spatial information such as population migration across regions.

Much work has been done on predicting the number of COVID-19 cases and deaths. Kraemer (2020) and Woody et al. (2020) adopted Generalized linear models to predict the number of daily cases and deaths during the first-wave COVID-19 in China and the United States, respectively. Autoregressive models are also widely used to forecast confirmed cases at a state-level (Mamode Khan et al., 2020; Triaccaa and Triacca, 2021; Agosto and Giudici, 2020). There are also several studies (Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems, 2021; Institute for Health Metrics and Evaluation, 2020) adopted by the Centers for Disease Control and Prevention (CDC) for COVID-19 case forecast in the United States. Our approach differs from these methods in two ways: (a) Our model provides finer-grained predictions based on the unique data, and (b) we focus on capturing the spatio-temporal correlation between confirmed cases and emphasize the interpretability of the proposed model.

Spatio-temporal analysis of the COVID-19 plays a pivotal role in understanding the dynamics of the spread of COVID-19. Angulo et al. (2013) introduces a spatio-temporal BME-SIR model integrating the disease representation at different locations to generate disease predictions. Bai et al. (2020) divides the regional-level COVID-19 time series data in the United States into several periods and develops a piecewise stationary SIR model coupled with spatio-temporal dependence. In addition, a vector autoregressive model developed by Zhu et al. (2021a) considers local spatio-temporal correlations, mobility, and demographic factors, aiming to estimate COVID-19 cases and deaths at a county level in the United States. In Chiang et al. (2020), a multivariate Hawkes process is adopted to model the occurrence of confirmed cases across the U.S. counties by incorporating social and health covariates. However, most of these methods use spatially or temporally aggregated data, which hinders us from understanding the spread of COVID-19 at an individual level.

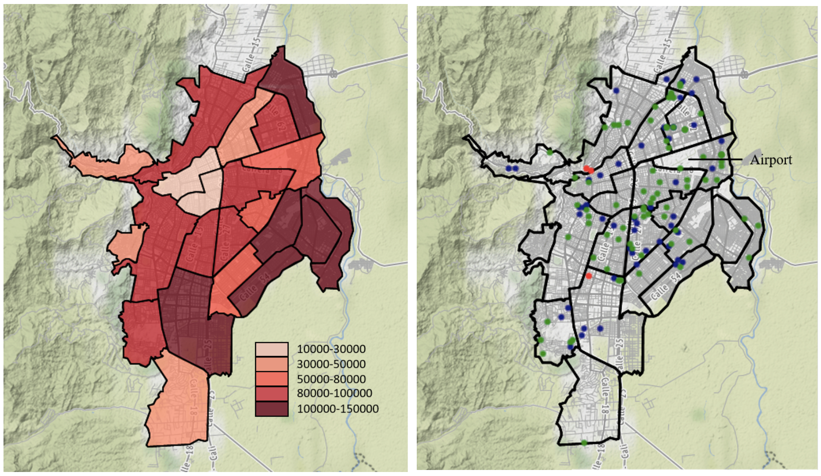
A few studies attempt to model the dynamics of COVID-19 using point processes. Gajardo and Müller (2021) proposes a point process regression framework of COVID-19 cases and deaths conditioned on mobility and economic covariates. Giudici et al. (2021) focuses on country-level case prediction in 27 European countries by augmenting spatio-temporal point process model with mobility network covariates. Li et al. (2021) introduces a generative and intensity-free point process model based on an imitation learning framework to track the spread of COVID-19 and forecast county-level cases in the United States. Compared to the previous methods, our approach is more flexible by considering non-stationarity in the spatial correlation, which is highly interpretable and expressive in representing the spread of the disease.

We also note that some related studies use similar techniques developed in this paper. First, similar to our proposed framework, Du et al. (2016); Mei and Eisner (2016); Zhang et al. (2020) model discrete events using neural-network-based point process models. However, most of these works aim to enhance the representative power by taking advantage of the recurrent neural structure (Hochreiter and Schmidhuber, 1997) or the attention mechanism (Vaswani et al., 2017) to represent the historical information, which lacks interpretability and is unable of capturing long-term effects. Second, a wide array of research focuses on characterizing the triggering effects between events using a fine-crafted kernel function. For example, original works Ogata (1988, 1998) introduce a parametric kernel in Epidemic Type Aftershock Sequence (ETAS) to capture the triggering effects between earthquakes. Recent works (Zhu et al., 2021b,c) extend the ETAS model by introducing neural networks and allowing for non-stationarity in representing the spatial correlation. The main difference of our method is that we consider both inter- and intra-influences between spatial kernel feature functions, which significantly improve the expressiveness of the model. A few works incorporate the exogenous effects into point process models by adding terms in the conditional intensity function (Zhu et al., 2021d; Rizoiu et al., 2017; Farajtabar et al., 2017).

2 Data description and preliminary analysis

The COVID-19 dataset provided by the Municipal Public Health Secretary of Cali[1] documents the individual-level confirmed COVID-19 cases, collected from Cali, one of the major cities in Colombia, the capital of the Valle del Cauca department and the most populated city in southwest Colombia, with 2,227,642 residents according to the 2018 census. As shown in Fig. 1(a), more than half of the population concentrates in neighborhoods of low socioeconomic strata located mainly in the east, northeast, and west. Almost a tenth of the population under the line of poverty agglomerates in the city's eastern neighborhoods. The population with higher socioeconomic

[1] https://www.cali.gov.co/salud/



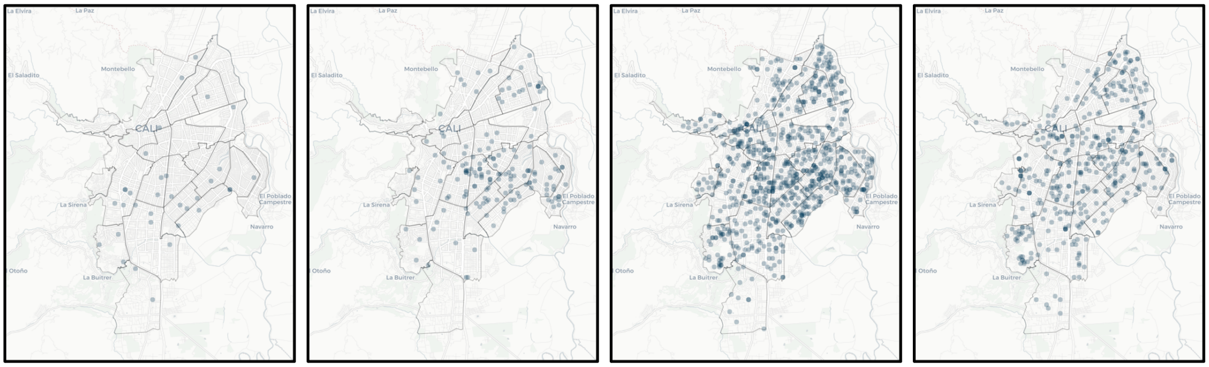
(a) Population distribution

(b) Landmarks

Figure 1: (a) Population distribution in Cali. Each polygon bounded by black lines represents a comuna (a municipality-level subdivision in Cali); there are 22 comunas in the city of Cali. (b) Landmarks in Cali. Each dot represents the landmark's location, and its color indicates the type of the landmark, where the red dot is a town hall, the blue dot is a church, and the green dot is a school.

strata distributes in the other city areas, concentrating the wealthiest population in the city's south. The city spans 560.3 square kilometers (216.3 square miles) with 120.9 square kilometers (46.7 square miles) of the urban area, making it the second-largest and the third most populated city in the country. As the only major Colombian city with access to the Pacific coast, Cali is the leading industrial and economic center in the country's south, with one of Colombia's fastest-growing economies. Cali's international airport is located in the northeast part of the city, and it is Colombia's third-largest airport in terms of passengers (Wikipedia, 2021).

The dataset records 38,611 cases from March 15 to September 30 of 2020, including 28 weeks. Specifically, a COVID-19 case was recorded once confirmed, with the diagnosed date of the patient and the geographical location (measured in longitude and latitude) of their residence. The testing procedures were carried out across the entire urban area, with similar testing rates in each comuna. Unlike other commonly-seen COVID-19 datasets that only report the aggregated number of cases or deaths at a state or county level, this dataset records the exact location and time information of each single confirmed case. In practice, we observe periodic weekly oscillations in daily reported cases and deaths, which may be caused by testing bias (higher testing rates on certain days of the week). To reduce such bias, we aggregate the number of cases and deaths of each county by weeks. Fig. 2 presents the spatial distribution of confirmed cases at four particular weeks in Cali. We note that the first confirmed case of COVID-19 in Colombia appeared on March 6, 2020. On March 12, the country soon declared a state of emergency. On March 15, Cali reported the first positive person. Then the authorities announced the mandatory isolation for the entire



(a) March 29, 2020 (b) May 17, 2020 (c) July 12, 2020 (d) August 23, 2020

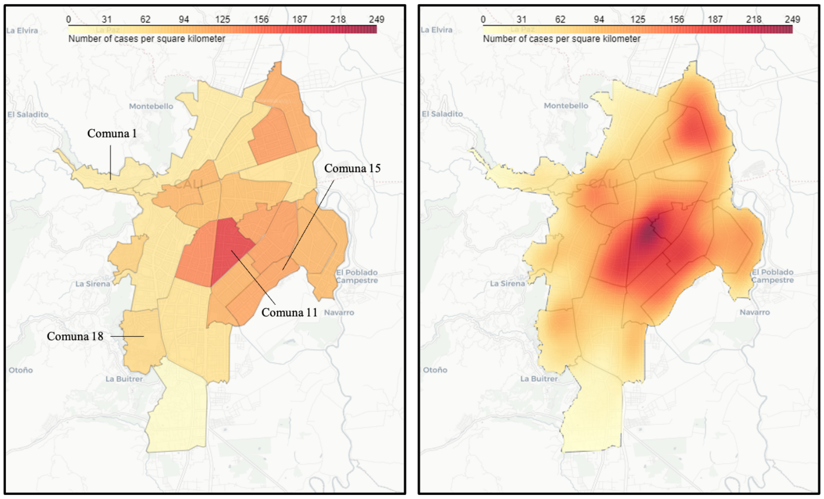
Figure 2: Snapshots of confirmed COVID-19 cases at four particular weeks. Each dot represents the location of a confirmed case. Note that darker dots indicate multiple dots being overlapped.

city for just eight days (Presidency of the Republic of Colombia., 2020). The first case reported in the city based on people who went to health services occurred in high socioeconomic strata. However, the disease quickly spread and concentrated in the most vulnerable areas with low socioeconomic strata. After early efforts of the government to contain the pandemic, inevitably, the virus spread throughout the city, affecting a large part of the population. The above public health decisions are known not significantly to affect the dynamics of the virus spreading. Thus, we do not consider the impact of these decisions in our model for simplicity.

Besides COVID-19 events, we also collect the location of three kinds of landmarks in Cali, including churches, schools, and town halls, from the Administrative Department of Municipal Planning[2], as these locations play an important role in understanding the wide and rapid spreading of the virus. According to James et al. (2021), there is a high COVID-19 positive rate among attendees to events at places, such as churches. As a clear note in this line, among 92 attendees at a rural Arkansas church during March 6–11, 35 (38%) developed laboratory-confirmed COVID-19, and three persons died (James et al., 2021). The landmark dataset has three town halls, 49 small and large churches, and 77 schools. Fig. 1(b) shows the exact locations of these collected landmarks.

Our preliminary study suggests that the confirmed cases are unevenly distributed across the city and correlated in time and space. In Fig. 3, we show the spatial distribution of all the confirmed cases. As we can see, most of the reported cases concentrate in the city's center, particularly in Comuna 11. More cases are reported in the eastern Cali than in the western Cali, which presents a heterogeneous spatial profile of the COVID-19 cases in Cali. The first three panels in Fig. 4 show the partial autocorrelation functions (PACF) (Brockwell and Davis, 1991) of daily confirmed cases for three comunas in Cali. Short lags (less than one week) appear to be highly relevant to the current confirmed cases at each comuna, highlighting a significant temporal

[2] `https://www.cali.gov.co/planeacion/`



(a) Case density by comuna (b) Case density estimated by KDE

Figure 3: Spatial distribution of the confirmed cases at two spatial resolutions. The color depth indicates the number of confirmed cases in one square kilometer; (a) shows the case density per comuna. (b) represents the spatially continuous case density estimated by KDE.

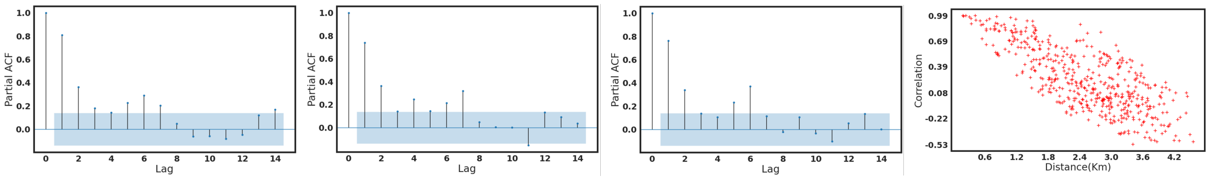
dependence. The last panel of Fig. 4 shows the spatial correlation versus the distance between different locations in Cali. Specifically, we investigate the time series of cases occurrence rate (estimated by KDE) at 1,000 arbitrary locations. As we can see, a strong spatial correlation is observed in the vicinity of an arbitrary location, while the correlation between two locations weakens with their distance.

3 Methodology

This section presents our non-stationary spatio-temporal point process model for COVID-19 cases. In the following, we first revisit some essential background of spatio-temporal point processes. Then we propose a novel point process model with a non-stationary kernel function, which captures complex triggering effects between events in time and space. Lastly, we characterize the influence of city landmarks as an exogenous promotion.

3.1 Background: Spatio-temporal point processes

Spatio-temporal point processes (STPPs) is a popular model for discrete events data that occur in space and time González et al. (2016); Reinhart (2017). Denote the observation space as $\mathcal{X} = [0, T] \times \mathcal{S} \subseteq \mathbb{R}^+ \times \mathbb{R}^2$, where T is the time horizon and $\mathcal{S}$ represents the space of geographic coordinate system (GCS). Each confirmed case is a *discrete event* defined by a data tuple $x := (t, s)$, where $t \in [0, T]$ is the time when the individual was diagnosed with COVID-19 and $s \in \mathcal{S}$ represents the location of residence of confirmed case. Let $\mathcal{H}_t := \{x_i = (t_i, s_i) | t_i < t\}$ denote the



(a) PACF 1 (b) PACF 2 (c) PACF 3 (d) Spatial correlation

Figure 4: (a), (b), and (c): PACFs of the time series of confirmed cases in three comunas. The x-axis is the time lag in days. The shaded area represents non-significant PACFs. (d) Correlation coefficients between the series of confirmed cases density at five arbitrarily chosen locations and those at other 199 locations in their neighborhoods against the separating distances measured in kilometers.

events' history before time t. Let $\mathbb{N}$ be a counting measure on $[0, T] \times \mathcal{S}$ corresponding to $\mathcal{H}_T$, i.e. for any $S \subset [0, T] \times \mathcal{S}, \mathbb{N}(S) = |S \cap \mathcal{H}_T|$, the number of occurred events in the set S. For any function $f : [0, T] \times \mathcal{S} \to \mathbb{R}$, the integral w.r.t. the counting measure is defined as

$$\int_S f(\tau, r) d\mathbb{N}(\tau, r) = \sum_{(t_i, s_i) \in S \cap \mathcal{H}_T} f(t_i, s_i).$$

Given the observed history $\mathcal{H}_t$, the probability structure of the point process is characterized by the conditional intensity function $\lambda(t, s)$ (for notational simplicity, we omit the dependence on $\mathcal{H}_t$), which is defined as:

$$\lambda(t, s) dt \cdot |B(s, ds)| = \mathbb{E}[d\mathbb{N}(t, s) | \mathcal{H}_t]. \tag{1}$$

Here $B(s, ds)$ is a ball centered at s in the space $\mathcal{S}$ with radius ds, and $|B(s, ds)|$ is the Lebesgue measure.

Hawkes processes (Hawkes, 1971) is a type of self-exciting point process that captures the triggering effects between events. Assuming that influences from past events are linearly additive towards the current event, the conditional intensity for a Hawkes point process takes the form of

$$\lambda(t, s) = \lambda_0 + \int_0^t \int_{\mathcal{S}} k(t, \tau, s, u) d\mathbb{N}(\tau, u), \tag{2}$$

where $\lambda_0 > 0$ denotes the background intensity, and $k(t, t', s, s')$ is a triggering kernel function that captures the influence of past events on the likelihood of event occurrence at the current time; in this work, we do not assume the kernel function to be positive or shift-invariant (to capture the non-stationary process as we will define later on).

The parameters can be estimated by maximum likelihood estimation (MLE). Given the observed point pattern $\boldsymbol{x}$, we can write the log-likelihood as

$$\ell(\boldsymbol{x}) = \sum_{i=1}^{\mathbb{N}([0,T] \times \mathcal{S})} \log \lambda(t_i, s_i) - \int_0^T \int_{\mathcal{S}} \lambda(\tau, u) du d\tau, \tag{3}$$

where $\mathbb{N}([0,T] \times \mathcal{S})$ is the number of observed events (see the derivation of the log-likelihood in Appendix A).

The Epidemic Type Aftershock-Sequences (ETAS) model is one of the most common spatio-temporal point processes (Ogata, 1988, 1998), which has been widely adopted in modeling typical spatio-temporal datasets such as earthquakes. ETAS model uses a Gaussian diffusion kernel

$$k(t,t',s,s') = \frac{Ce^{-\beta(t-t')}}{2\pi\sqrt{|\Sigma|}(t-t')} \cdot \exp\left\{-\frac{(s-s'-\mu)^\top \Sigma^{-1}(s-s'-\mu)}{2(t-t')}\right\},$$

where $\Sigma \equiv \text{diag}(\sigma_x^2, \sigma_y^2)$ is a two-dimensional diagonal matrix representing the covariance of the spatial correlation, β is the decaying rate, μ is the mean shift, and C is a constant. However, the diffusion kernel is stationary and only depends on the spatio-temporal distance between two events. In addition, the kernel assumes the spatial correlation is isotropic and unable to capture complex spatial dependence.

3.2 A non-stationary triggering Gaussian kernel

We introduce a non-stationary triggering kernel, which can vary continuously over space and plays a vital role in modeling the heterogeneous spatial correlation across different regions. For model simplicity and computational efficiency, we adopt the commonly used assumption that the triggering effect of a past event is separable in space and time:

$$k(t,t',s,s') = \nu(t,t') \cdot \upsilon(s,s'),$$

where $\nu(t,t')$ is a kernel that captures the dependence between time t and t', and $\upsilon(s,s')$ is a spatial kernel that captures the non-stationary correlation between location s and s'.

A stationary temporal kernel As the virus spreads and affects a significant portion of the population in a short period, we can assume temporal virus transmission is through a shift-invariant kernel with exponential decay:

$$\nu(t,t') = Ce^{-\frac{1}{2\sigma_0^2}(t-t')^2}, \quad t > t'.$$

Here $C > 0$ is a parameter that controls the magnitude of the kernel, $\sigma_0 > 0$ is a parameter that controls the decaying rate of the event's temporal influences, and we assume $t > t'$ to capture the fact that historical event at time t' has an impact on the current time t but not vice versa.

A non-stationary spatial kernel The complex nature of the spatial spread of COVID-19 requires a non-homogenenous and non-stationary spatial kernel function in the point process. Given two arbitrary locations $s, s' \in \mathcal{S}$, we define the spatial kernel $\upsilon(s,s')$ as a inner product between two feature mappings ϕ_s and $\phi_{s'}$, i.e,

$$\upsilon(s,s') = \langle \phi_s, \phi_{s'} \rangle, \quad s, s' \in \mathcal{S},$$

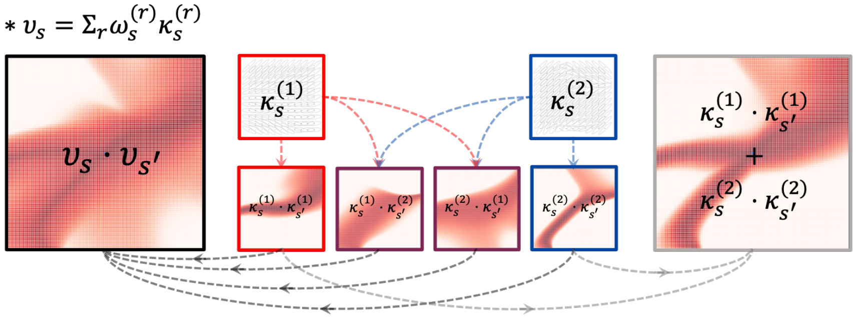

$* \upsilon_s = \Sigma_r \omega_s^{(r)} \kappa_s^{(r)}$
$\upsilon_s \cdot \upsilon_{s'}$
$\kappa_s^{(1)}$
$\kappa_s^{(2)}$
$\kappa_s^{(1)} \cdot \kappa_{s'}^{(1)}$
$\kappa_s^{(1)} \cdot \kappa_{s'}^{(2)}$
$\kappa_s^{(2)} \cdot \kappa_{s'}^{(1)}$
$\kappa_s^{(2)} \cdot \kappa_{s'}^{(2)}$
$\kappa_s^{(1)} \cdot \kappa_{s'}^{(1)}$
$+$
$\kappa_s^{(2)} \cdot \kappa_{s'}^{(2)}$

Figure 5: An example of the non-stationary spatial kernel with two feature functions evaluating at location s (the center of the box), i.e., $\upsilon(s,s') = \langle \phi_s, \phi_{s'} \rangle$, $\forall s' \in \mathcal{S}$, where $\phi_s = \kappa_s^{(1)} + \kappa_s^{(2)}$. The purple boxes indicate the cross-correlated terms ($\kappa_s^{(1)} \cdot \kappa_{s'}^{(2)}$ and $\kappa_s^{(2)} \cdot \kappa_{s'}^{(1)}$); the red and blue boxes indicate the self-correlated terms ($\kappa_s^{(1)} \cdot \kappa_{s'}^{(1)}$ and $\kappa_s^{(2)} \cdot \kappa_{s'}^{(2)}$).

where the inner product for functions $\langle f, g \rangle := \int_{\mathbb{R}^2} f(u)g(u)du$. We represent the feature mapping ϕ_s as a weighted sum of a set of R independent kernel-induced feature functions $\{\kappa_s^{(r)} := \kappa^{(r)}(s, \cdot)\}_{r=1}^R$:

$$\phi_s = \sum_{r=1}^{R} w_s^{(r)} \kappa_s^{(r)},$$

where $\kappa^{(r)} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ is a general kernel and $w_s^{(r)}$ is the corresponding weight of that feature function at location s. The location-dependent weight satisfies $\sum_{r=1}^R w_s^{(r)} = 1$ at any arbitrary location s. Hence the spatial kernel can be rewritten as

$$\upsilon(s,s') = \sum_{1 \le r_1, r_2 \le R} w_s^{(r_1)} w_{s'}^{(r_2)} \left\langle \kappa_s^{(r_1)}, \kappa_{s'}^{(r_2)} \right\rangle .$$

The rationale of this design is two-fold: (a) Using a linear combination of the product of feature functions enhances the representative power of the spatial kernel. Note that when $r_1 = r_2$, the kernel captures self-correlation (self-similarity of feature functions) and otherwise captures the cross-correlation (similarity between two feature functions). (b) The spatial kernel can also be highly interpretable if $\kappa^{(r)}$ takes a specific parametric form; following the idea in Higdon et al. (1998); Zhu et al. (2021c), we choose κ_s to be a Gaussian function centered at s with covariance matrix Σ_s, since the spatial correlation between two events decays as their distance increases in general. The spatial kernel is specified to be:

$$\upsilon(s,s') = \sum_{1 \le r_1, r_2 \le R} \frac{w_s^{(r_1)} w_{s'}^{(r_2)}}{2\pi |\Sigma_s^{(r_1)} + \Sigma_{s'}^{(r_2)}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (s - s')^\top (\Sigma_s^{(r_1)} + \Sigma_{s'}^{(r_2)})^{-1} (s - s') \right\}. \quad (4)$$

See detailed derivation of (4) in Appendix B. Fig. 5 gives an example of the spatial kernel with two feature functions.

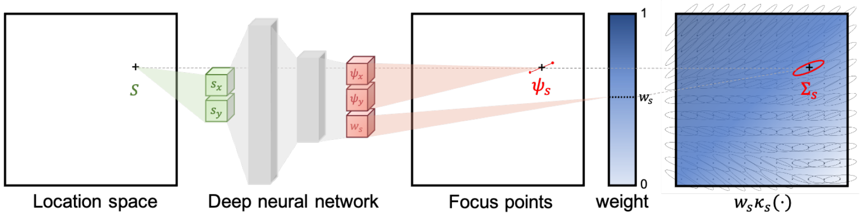


Figure 6: An illustration of a deep neural network that maps an arbitrary location s to a spatial kernel, consisting of a feature function κ_s (represented through focus points) and weight w_s.

Now we specify the kernel-induced feature function κ_s. According to Higdon et al. (1998), there exists a one-to-one mapping between a bivariate normal distribution specified by Σ_s and its one standard deviation ellipse. Note that κ_s is centered at s, so the ellipse's center is fixed at s. Thus we can specify the ellipse by a pair of focus points and the fixed area A. The focus points are denoted by $\boldsymbol{\psi}_s = (\boldsymbol{\psi}_x(s), \boldsymbol{\psi}_y(s))$ and $-\boldsymbol{\psi}_s = (-\boldsymbol{\psi}_x(s), -\boldsymbol{\psi}_y(s))$, where $\boldsymbol{\psi}_s \in \Psi \subset \mathbb{R}^2$. Hence, given $\boldsymbol{\psi}_s$ and A, the corresponding Σ_s can be written as

$$\Sigma_s = \tau_z^2 \begin{pmatrix} Q + \frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\alpha & \frac{\|\boldsymbol{\psi}_s\|^2}{2}\sin 2\alpha \\ \frac{\|\boldsymbol{\psi}_s\|^2}{2}\sin 2\alpha & Q - \frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\alpha \end{pmatrix},$$

where $Q = \sqrt{4A^2 + \|\boldsymbol{\psi}_s\|^4\pi^2}/2\pi, \alpha = \tan^{-1}(\boldsymbol{\psi}_y(s)/\boldsymbol{\psi}_x(s))$, $\tau_z > 0$ is a scaling parameter that controls the overall level of the covariance (see the derivation in Appendix C). We consider A as a hyper-parameter.

Neural network-based kernel representation We develop a neural network-based representation for the kernel-induced feature function similar to the idea in Zhu et al. (2021c,e). A key feature of our non-stationary spatial kernel is that for any location $s \in \mathcal{S}$, we can estimate a mapping that obtain the focus point $\boldsymbol{\psi}_s$ and the corresponding location-dependent weight w_s. To this end, we represent the mapping $\varphi : \mathcal{S} \to \Psi \times [0, 1]$ from the location to the space of focus points Ψ and the weights $[0, 1]$ using a fully-connected multi-layer neural network. The input of the neural network is the two-dimensional location vector s, and the output is the concatenation of the corresponding focus point $\boldsymbol{\psi}_s$ and its weight w_s. Here, each hidden layer is equipped with a softplus activating function $f(x) = \log(1 + e^x)$ (see the detailed specification of the neural network in Section 5). Neural networks allow a flexible representation of the covariance and the corresponding kernel-induced feature function due to their well-known universal approximation power. In our implementation, we adopt the same network architecture for all R kernel-induced feature functions, as illustrated in Fig. 6.

3.3 Exogenous promotion of city landmarks

To consider the influence of city landmarks, we consider each landmark as a constant exogenous promotion to the virus spread at their locations. To achieve this, we adopt an idea similar to

Zhu et al. (2021d) and introduce an additional term to the conditional intensity function $\lambda(t, s)$ (2):

$$\lambda(t, s) = \lambda_0 + \sum_{l=1}^{L} \gamma_l g(s|s_l, \Sigma_l) + \sum_{t'<t} k(t, t', s, s') \ . \tag{5}$$

The second and third terms represent the exogenous promotion at location s and the endogenous excitation at location s and time t, respectively. We use L to denote the number of landmarks, and γ_l indicates the significance of landmark l. We assume that the exogenous effect induced by landmarks decays with distance to them. Hence, the influence of landmark l located at s_l is modeled by a Gaussian function $g(s|s_l, \Sigma_l)$ centered at location $s_l \in \mathcal{S}$ with covariance Σ_l. Here we define $\Sigma_l := \sigma_l^2 \mathbf{I}$, where $\mathbf{I}$ is an identity matrix.

4 Efficient computation of the log-likelihood function

The log-likelihood of the spatio-temporal point process defined in (3) is often intractable due to the double integral term. Numerical integral can also be expensive: if the number of randomly sampled points in a three-dimensional space is K, and the total number of events is N, the computational complexity is $\mathcal{O}(KN)$ $(K \gg N)$ using commonly-used numerical integration techniques. In our case, we can write the integral term as

$$\int_0^T \int_{\mathcal{S}} \lambda(\tau, u) du d\tau = \lambda_0 |\mathcal{S}| T + \int_0^T \sum_{l=1}^{L} \gamma_l \underbrace{\int_{\mathcal{S}} g(u|s_l, \Sigma_l) du}_{(i)} \, d\tau + \int_0^T \sum_{t_i<\tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau \cdot \underbrace{\int_{\mathcal{S}} \upsilon(u, s_i) du}_{(ii)}, \tag{6}$$

where $|\mathcal{S}|$ is the Euclidean area of the city, and evaluating $(i), (ii)$ are difficult in general because (a) Both (i) and (ii) require the integral over the geographical space of Cali $\mathcal{S}$, which has an irregular shape; (b) In (ii), $\upsilon(u, s_i)$ is location-dependent and parameterized by a neural network.

We circumvent these two difficulties by simplifying without significantly impacting the model's accuracy: (a) We expand the integration region $\mathcal{S}$ to the entire geographical space $\mathbb{R}^2$ and account for the boundary effect error by ϵ_1. Note that the kernel g or κ_s are Gaussian concentrated around s and most events are located in the interior of $\mathcal{S}$ when choosing sufficiently large $\mathcal{S}$. As suggested by Ogata (1998), such boundary effect is usually negligible due to the decreased activity in the region's edges. (b) We assume the distance between two focus points ($\boldsymbol{\psi}_s$ and $-\boldsymbol{\psi}_s$) at an arbitrary location s is bounded by a threshold $2c$ (which can be obtained by rescaling the output of neural networks); a large distance between focus points leads to an overstretched ellipse, which is unrealistic in practice. Therefore, when performing numerical integration, we approximate the kernel-induced feature function κ_s by a standard Gaussian function denoted by κ_s^0, which corresponds to a standard deviation ellipse centered at s with area A. The resulted relative error of the integral approximation is denoted by ϵ_2. In short, these two assumptions reduce the double integral (6) to an analytical form that can be evaluated directly without numerical integration.

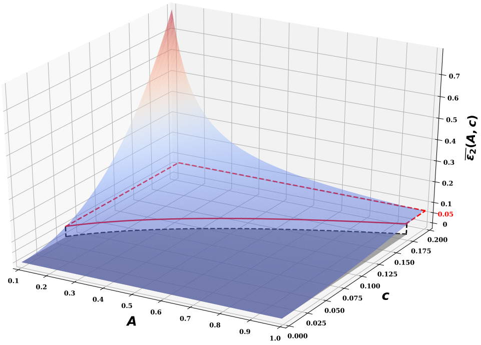


Figure 7: Surface plot for the upper bound of the relative error $\overline{\epsilon_2}$ with regards to hyper-parameters A and c. The horizontal coordinates represent the value of A and c, respectively, and the vertical coordinate represents the value of $\overline{\epsilon_2}(A, c)$. The red solid line is a surface contour valued at 0.05. The grey shaded area in the horizontal plane represents the set of (A, c) that satisfies $\overline{\epsilon_2}(A, c) < 0.05$. We can observe that the higher the value of A and the the lower the value of C, the smaller the upper bound of the relative error.

Proposition 1 (Approximation of the integral in the likelihood function)**.** *Assume the area of the corresponding ellipse of κ_s is A and the distance between its focus points is restricted to be smaller than $2c$, then the integral in (6) can be approximated by*

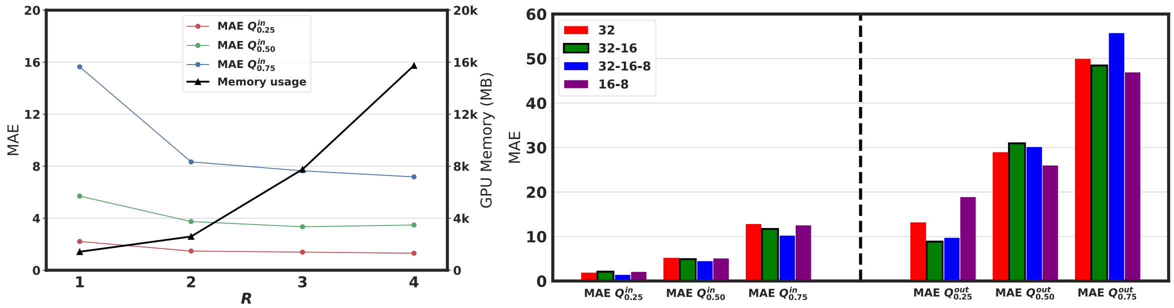
$$\int_0^T \int_{\mathcal{S}} \lambda(\tau, r) dr d\tau = (1+\epsilon_2)\left[\lambda_0|\mathcal{S}|T + T\sum_{l=1}^{L}\gamma_l + \sqrt{2\pi}C\sigma_0 \sum_{i=1}^{\mathbb{N}([0,T]\times\mathcal{S})}\left\{h\left(\frac{T-t_i}{\sigma_0}\right) - \frac{1}{2}\right\}\right] - \epsilon_1, \tag{7}$$

where the function h is the cumulative density function of the standard normal distribution and ϵ_1, ϵ_2 are the boundary effect error and the relative error of the integral approximation, respectively. Ignoring the boundary effect error ϵ_1, the relative error $\epsilon_2 \in (-1, +\infty)$ can be bounded by:

$$|\epsilon_2| < \max\left\{U - 1, 1 - \frac{1}{U}\right\},$$

where $U = (\sqrt{4A^2 + c^4\pi^2} + c^2\pi)/2A$. (see the proof in Appendix D)

Remark. *Proposition 1 leads to a computationally efficient calculation of the integral with complexity $\mathcal{O}(N)$. We denote the upper bound of the relative error as $\overline{\epsilon_2}$ and its dependence on hyper-parameters A and c is illustrated in Fig. 7. In general, a larger c results in a more expressive spatial kernel but requires a larger A to control the approximation error. In practice, we select $c = 0.1$ and $A = 0.35$ to limit the relative error ϵ_2 under 0.05 and ensure a certain level of expressiveness for the spatial kernel.*



(a) Performance of different Rs

(b) Performance of different Neural network architectures

Figure 8: Performance of the proposed model with different numbers of components R in the feature mapping ϕ_s or different neural network architectures: (a) MAE and GPU memory usage of the in-sample estimations with $R = 1, 2, 3, 4$. The red, green, and blue lines represent three different quartiles of MAE for the in-sample estimations, respectively, and the black line represents the increase of GPU memory usage when R grows. (b) MAE of the in-sample and out-of-sample estimations with four different neural network architectures. The color code and the corresponding number series represent different neural network structures; for example, "32-16" indicates a two-hidden-layer neural network, and there are 32 and 16 nodes for each layer. The left three groups show the MAEs of in-sample estimation. The right three groups show the MAEs of out-of-sample estimation. In the following experimental results, we adopt the architecture 32-16.

5 COVID-19 data case study in Cali

In this section, we present the numerical results for studying the real COVID-19 data in Cali, which is described in Section 2. We first investigate the model's explanatory power by evaluating the in-sample performance and visualize the estimated kernel-induced feature functions and their corresponding spatial kernel. We also study the exogenous effects of the city landmarks. Finally, we compare the out-of-sample predictive performance of the proposed method with four baseline approaches. In this section, $\{\text{MAE}Q_q^{in}, \text{MAE}Q_q^{out}\}$ denote to the lower q-quantile of the mean absolute error (MAE) (Willmott and Matsuura, 2005) for the in-sample and out-of-sample estimation, respectively.

Our experimental settings are as follows. We consider a mixture kernel with $R = 3$ components, which achieves the balance between the predictive performance and the computational efficiency according to the results shown in Fig. 8(a). Fig. 8(b) compares the out-of-sample performance for four network architectures; we choose a network architecture that achieves good performance for our data: a two-hidden-layer neural network with 32 and 16 nodes in each hidden layer for each kernel-induced feature function. We select the hyper-parameters $A = 0.35$ and $c = 0.1$ based on actual needs, and estimate model's parameters $\{\lambda_0, C, \sigma_0, \tau_z, \{\gamma_l\}_{l=1}^L, \{\sigma_l\}_{l=1}^L, \{\varphi^{(r)}\}_{r=1}^R\}$ by solving the maximum likelihood problem via gradient descent. We train the model with the entire training set in each epoch. The initial learning rate is 1 and will decay to 0.1 of its last

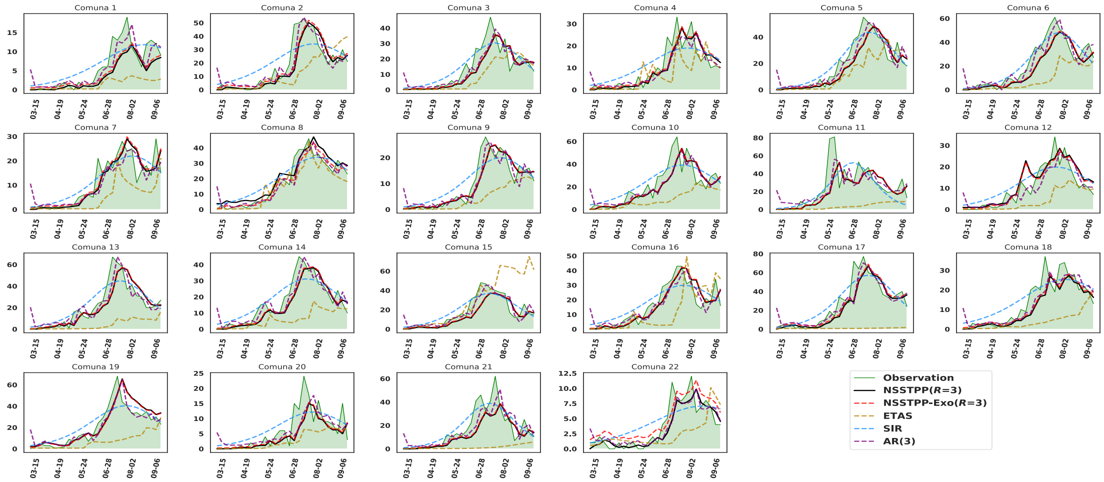
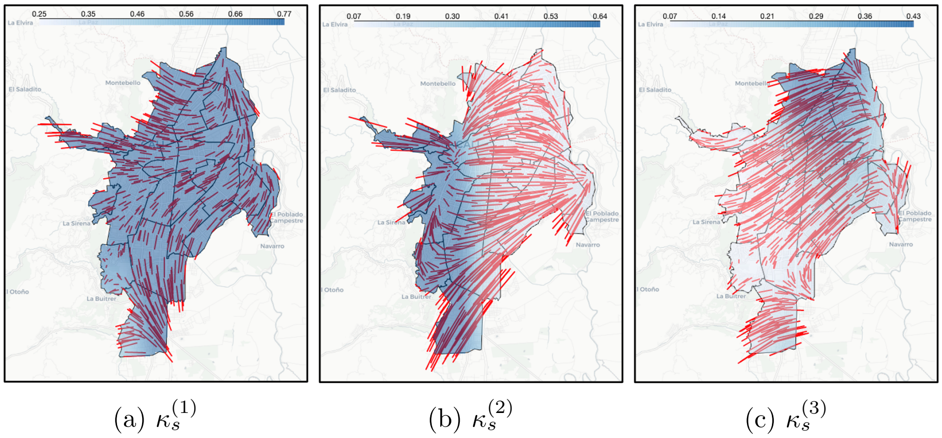


Figure 9: Comparison of the proposed model with baseline models. The green lines and shaded areas represent the ground truth. The black and red lines indicate the in-sample estimation of our non-stationary point process model. The yellow, blue, and purple lines represent the in-sample estimation of the ETAS model, SIR model, and AR(3) model, respectively.

Table 1: Performance of in-sample estimation. The numbers in the brackets are one standard deviation.

Models	Log-likelihood($\times 10^4$)	MAE $Q^{\text{in}}_{0.25}$	MAE $Q^{\text{in}}_{0.5}$	MAE $Q^{\text{in}}_{0.75}$
Random	/	5.000	11.000	18.000
SIR	/	1.862	3.759	7.391
AR(3)	/	1.307	2.880	**6.496**
ETAS	4.868 (0.0058)	1.486	4.737	14.895
NSSTPP−Exo (R=1)	8.671 (0.0772)	0.834	3.145	7.922
NSSTPP−Exo (R=2)	9.138 (0.0886)	0.806	2.728	7.119
NSSTPP−Exo (R=3)	9.190 (0.0906)	0.853	**2.613**	7.000
NSSTPP (R=3)	**9.331** (0.0937)	**0.797**	2.620	6.757

value when there is no likelihood increment for 10 epochs. The algorithm stops when the likelihood oscillation is less than 1 for 30 epochs. We use Adam optimizer (Kingma and Ba, 2017) for all experiments. In the following, we refer to the proposed framework as a Non-Stationary Spatio-Temporal Point Process (NSSTPP).



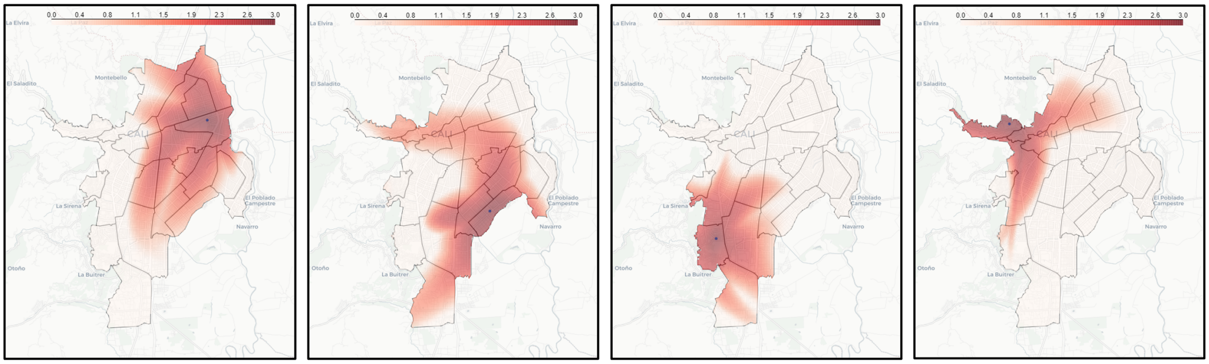
(a) $\kappa_s^{(1)}$ (b) $\kappa_s^{(2)}$ (c) $\kappa_s^{(3)}$

Figure 10: Visualization of three learned kernel-induced feature functions over Cali. Each panel shows one κ_s over space. The red line segments are edges that connect two focus points of location s. The blue shaded area shows the intensity of weight $w_s^{(r)}$ of each $\kappa_s^{(r)}$ over space. Darker colors mean larger weights.

5.1 Model interpretation

To evaluate our model's goodness-of-fit, we compare the in-sample estimations of different models on the one-week-ahead number of cases, which is performed as follows. We first fit the model using the entire 28 weeks of data. The in-sample estimation can then be obtained by feeding the same data into the fitted model and finding an empirical expectation of the conditional intensity at a given week according to the equation (5). We compare our model with five baselines that are commonly adopted in modeling infectious epidemics: (a) Homogeneous Poisson process (as a sanity check); (b) Susceptible-Infectious-Recovered (SIR) model; (c) Autoregressive (AR) time series model; (d) Epidemic-type aftershock sequence (ETAS) model; (e) Our model without exogenous effects (NSSTPP−Exo). See Appendix E for a detailed review of the baseline methods and their hyper-parameter choices. Fig. 9 shows the estimated number of cases by different models in each comuna of Cali. More results are summarized in Table 1, where we adopt two commonly-used metrics for performance evaluation, including log-likelihood and MAE. The results show that our method outperforms other baseline approaches in both log-likelihood and MAE. Besides, we observe a significant performance gain compared to the ETAS model, which emphasizes the importance of the non-stationarity of the spatial kernel in capturing complex spatio-temporal pattern.

We study the in-sample explanatory power of our model and interpret the estimation results on the data in Cali. First, we visualize three learned spatial kernel-induced feature functions, which reveal the underlying spatio-temporal transmission dynamics of COVID-19 in Cali, as shown in Fig. 10. Recall that at any location s, $\kappa_s^{(r)}$ is a Gaussian kernel with a spatially varying covariance matrix represented by two focus points of its one standard deviation ellipse. Therefore,



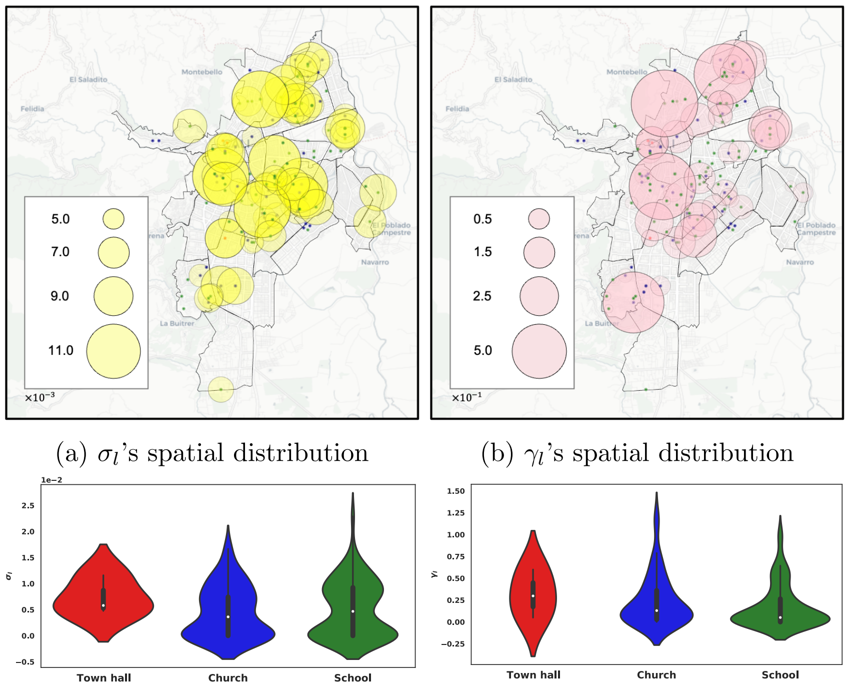
(a) Airport (b) Center of Comuna 15 (c) Center of Comuna 18 (d) Center of Comuna 1

Figure 11: Evaluation of the spatial kernel $\upsilon(s, \cdot)$ with s fixed at four typical locations over space. These panels intuitively show the spatial influence of the regional hubs located in different parts of the city. The dots represent the fixed location. The color depth indicates the intensity of the kernel value. Darker color represents a higher kernel value.

we connect two focus points of each sampled covariance matrix over space using a red line segment for visualization. The angle and length of each red line can be interpreted as the direction and strength of influence at the particular location. The color depth of the background represents the value of the corresponding weight $w_s^{(r)}$ at location s of each $\kappa_s^{(r)}$, indicating the significance of $\kappa_s^{(r)}$ at that location. These results suggest that the virus is spreading rapidly across the region following the diagonal direction from Southwestern to the city's Northeastern. We can also observe a more subtle but complicated spreading pattern near the border of the city.

Fig. 11 visualize the estimated spatial kernel $\upsilon(s, \cdot)$ given one of its input s, which can be treated as the influence of the location s. Here we present four examples, including the airport, the center of Comuna 1, Comuna 15, and Comuna 18. Each example demonstrates that each location radiates the influence to its surrounding region in a different manner. The results show that the airport significantly influences the other city region as most of the northern area has relatively high kernel values. As the most populated community in Cali, Comuna 15 also casts its influences on the city's Southeastern side. In addition, the impact of the location in Comuna 1 extends narrowly to two different directions, which correspond to two major routes in Cali. We note that these examples also emphasize the significance of the non-stationarity of the proposed method.

We also visually and quantitatively examine the exogenous effect of the city landmarks, as shown in Fig. 12. Recall that the exogenous effect of each landmark is assumed to be an isotropic bivariate normal distribution, where γ_l and σ_l can be interpreted as the *intensity* and the *sphere of influence* of the exogenous effects of landmarks l, respectively. We visualize the learned σ_l and γ_l on the map of Cali in Fig. 12(a),(b). We also report the distributions of these two learned parameters for different categories of landmarks. As we can observe, the exogenous effects of the landmarks located in the center of the city (the most severely affected areas) tend to have



(a) σ_l's spatial distribution

(b) γ_l's spatial distribution

(c) σ_l's distribution for different categories

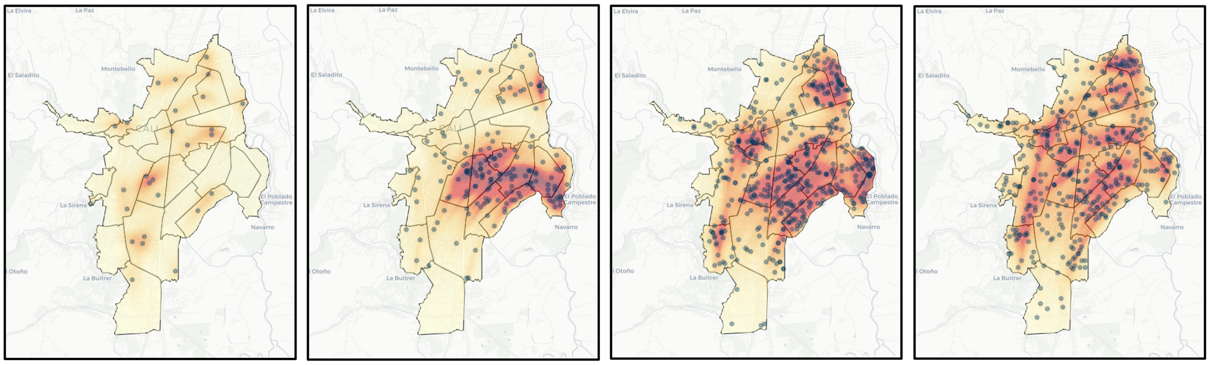
(d) γ_l's distribution for different categories

Figure 12: Estimated exogenous effects of landmarks in Cali. (a) and (b) visualize the learned $\{\sigma_l\}_{l=1}^{L}$ and $\{\gamma_l\}_{l=1}^{L}$ on the map of Cali, respectively. (c) and (d) show the distributions of learned parameters for different categories of landmark.

smaller intensities (γ_l) but larger sphere of influences (σ_l) to their neighborhood. This result also indicates that town halls may have a more significant influence than other landmarks. We see a real explanation here: The landmarks located at the center receive more people during the day and they act as super spreaders of the virus which is indicated by larger spheres of influence.

5.2 Predictive performance

Now we assess the model's predictive power by performing the one-week-ahead out-of-sample prediction of the number of cases. The out-of-sample prediction withholds the future data after a certain time point, trains the model based on the previous data, and then uses the estimated model to forecast the data in the next week. Fig. 13 shows the predicted conditional intensity at four particular weeks, which represent four different stages of the pandemic: (a) the early stage, (b) the week before the first outbreak, (c) the week before the second outbreak, and (d) the week in the stabilized plateau of the pandemic development. As we can observe, our method can capture the spatial occurrences of these cases, detect regions with sparsely distributed cases by showing a lower intensity, and show a higher intensity in other regions with densely distributed cases. We



(a) March 22, 2020 (b) May 17, 2020 (c) June 28, 2020 (d) August 30, 2020

Figure 13: Predicted conditional intensity at four different weeks. The black dot represents an actual case reported in that week. The color depth indicates the conditional intensity at the corresponding location. A darker color means a higher risk for citizens to be infected.

Table 2: Out-of-sample estimation performance.

Models	MAE $Q^{\text{out}}_{0.25}$	MAE $Q^{\text{out}}_{0.5}$	MAE $Q^{\text{out}}_{0.75}$
Random	5.190	8.660	14.900
SIR	2.253	5.713	8.554
AR(3)	2.219	**3.776**	8.915
ETAS	4.413	8.234	14.153
NSSTPP−Exo (R=1)	**1.732**	6.051	8.779
NSSTPP−Exo (R=2)	1.962	5.151	8.575
NSSTPP−Exo (R=3)	1.762	5.190	8.342
NSSTPP (R=3)	2.051	4.702	**7.450**

then examine the one-week-ahead prediction of the number of cases. Table 2 summarizes the quantitative results between the proposed method and the baselines. The result confirms that our model significantly outperforms other baseline methods. It is worth noting that SIR and the linear AR model only provide predicted aggregated number in each comuna, which is less challenging than our finer-grained prediction.

6 Conclusions

Based on an unprecedented fine-grained COVID-19 dataset in Cali, Colombia, we propose a spatio-temporal point process framework equipped with a non-stationary kernel to model the epidemic transmission at an individual level. The kernel is composed of a set of kernel-induced feature functions. Each feature function is represented by a neural network aiming to enhance the model flexibility while being interpretable. We also develop an efficient log-likelihood estimation by approximating the double integral using an analytical expression. Our numerical study in Cali has shown that the proposed approach achieves promising predicting performance and the learned model is highly interpretable.

We believe our methodology combines in a natural while novel way theory of point processes with artificial intelligence methods (such as neural networks), providing a unified framework for dealing with highly non-stationary spatio-temporal point patterns. The method is more general than the focused application and can be used, extended, and adapted to several natural phenomena represented by locations in space and time.

There are many ways the proposed method can be extended, and one possibility could be considering non-Gaussian kernels and alternative neural network methods. In any case, the data should always guide these ad-hoc adaptations.

The global results finding in this work for the city of Cali show an increased risk of contracting COVID-19 in the center, northeast, and northeast of the city, which are located the communes with more unsatisfied basic needs. On the other hand, in the south of the city, the risk of contagion is lower, and it is an area where people with greater purchasing power live. Considering the locations of the landmarks of the city into the model as transmission sources is undoubtedly an indispensable tool for predicting the spread of the virus. These outputs are very close to the reality experienced in that region of Colombia during the pandemic. The current official figures for Cali city show significant progress in the fight against the COVID-19, although a new peak is feared in the near future.

Acknowledgements

The authors are grateful to the Municipal Public Health Secretary of Cali, Valle del Cauca, Colombia for providing the COVID-19 data used in this paper. The work of Dong, Zhu, and Xie is partially supported by National Science Foundation CCF-1650913, CMMI-2015787, DMS-1938106, DMS-1830210. Francisco J. Rodríguez-Cortés has been partially supported by Universidad Nacional de Colombia, HERMES projects, Grant/Award Number: 51279.

References

Kimberly Chriscaden. Impact of covid-19 on people's livelihoods, their health and our food systems. https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems, 2020.

John Hopkins University. Coronavirus resource center, 2020. https://coronavirus.jhu.edu.

New York Times. Coronavirus data in the united states, 2020. https://www.nytimes.com/article/coronavirus-county-data-us.html.

Mariano Bizzarri, Mario Di Traglia, Alessandro Giuliani, Annarita Vestri, Valeria Fedeli, and Alberto Prestininzi. New statistical ri index allow to better track the dynamics of covid-19 outbreak in italy. *Scientific Reports*, 10, 2020. URL `https://doi.org/10.1038/s41598-020-79039-x`.

Thomas Guenther, Manja Czech-Sioli, Daniela Indenbirken, Alexis Robitailles, Peter Tenhaken, Martin Exner, Matthias Ottinger, Nicole Fischer, Adam Grundhoff, and Melanie Brinkmann. Investigation of a superspreading event preceding the largest meat processing plant-related sars-coronavirus 2 outbreak in germany. *SSRN Journal*, 2020. doi: 10.2139/ssrn.3654517.

David F Hendry and Felix Pretis. All change! the implications of non-stationarity for empirical modelling, forecasting and policy. *SSRN Journal*, 2016.

Fuk-Woo Chan Jasper, Yuan Shuofeng, Kok Kin-Hang, Kai-Wang To Kelvin, Chu Hin, Yang Jin, and et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, 2020. doi: https://doi.org/10.1016/S0140-6736(20) 30154-9.

Quentin J. Leclerc, Naomi M. Fuller, Lisa E. Knight, Sebastian Funk, and Gwenan M. Knight. What settings have been linked to sars-cov-2 transmission clusters? *Wellcome Open Research*, 2020.

Tiberiu Harko, Francisco S. N. Lobo, and M. K. Mak. Exact analytical solutions of the susceptible-infected-recovered (sir) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 236, 2014. URL `https://doi.org/10.1016/j.amc.2014.03.030`.

Qianying Lin, Shi Zhao, Daozhou Gao, Yijun Lou, Shu Yang, Salihu S. Musa, Maggie H. Wang, Yongli Cai, Weiming Wang, Lin Yang, and Daihai He. A conceptual model for the coronavirus disease 2019 (covid-19) outbreak in wuhan, china with individual reaction and governmental action. *International Journal of Infectious Diseases*, 2020. URL `https://doi.org/10.1016/j.ijid.2020.02.058`.

Anjalika Nande, Ben Adlam, Justin Sheen, Michael Z. Levy, , and Alison L. Hill. Dynamics of covid-19 under social distancing measures are driven by transmission network structure. *medRxiv*, 2020. URL `https://doi.org/10.1101/2020.06.04.20121673`.

Ivan Korolev. Identification and estimation of the seird epidemic model for covid-19. *Journal of econometrics*, 220(1), 2021. URL `https://doi.org/10.1016/j.jeconom.2020.07.038`.

Elena Loli Piccolomini and Fabiana Zama. Monitoring italian covid-19 spread by a forced seird model. *PloS one*, 15(8), 2020. URL `https://doi.org/10.1371/journal.pone.0237417`.

Moritz U G et al. Kraemer. The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 2020. doi: 10.1126/science.abb4218.

Spencer Woody, Mauricio Garcia Tec, Maytal Dahan, Kelly Gaither, Michael Lachmann, Spencer Fox, Lauren Ancel Meyers, and James G Scott. Projections for first-wave covid-19 deaths across the us using social-distancing measures derived from mobile phones. *medRxiv*, 2020.

Naushad Mamode Khan, Ashwinee Devi Soobhug, and Maleika Heenaye-Mamode Khan. Studying the trend of the novel coronavirus series in mauritius and its implications. *PloS one*, 15(7), 2020. URL `https://doi.org/10.1371/journal.pone.0235730`.

Marco Triaccaa and Umberto Triacca. Forecasting the number of confirmed new cases of covid-19 in italy for the period from 19 may to 2 june 2020. *Infectious Disease Modelling*, 2021. URL `https://doi.org/10.1016/j.idm.2021.01.003`.

Arianna Agosto and Paolo Giudici. A poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks*, 8(3):77, 2020. doi: 10.3390/risks8030077. URL `https://doi.org/10.3390/risks8030077`.

Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems. COVID-19 modeling, 2021. URL `https://covid19.gleamproject.org/`.

Institute for Health Metrics and Evaluation. Modeling COVID-19 scenarios for the united states. *Nature Medicine*, 27(1):94–105, October 2020. doi: 10.1038/s41591-020-1132-9.

Jose Angulo, Hwa-Lung Yu, Andrea Langousis, Alexander Kolovos, Jinfeng Wang, Ana Esther Madrid, and George Christakos. Spatiotemporal infectious disease modeling: a bme-sir approach. *PloS one*, 8 (9), 2013. URL `https://doi.org/10.1371/journal.pone.0072168`.

Yue Bai, Abolfazl Safikhani, and George Michailidis. Non-stationary spatio-temporal modeling of COVID-19 progression in the u.s. *medRXiv preprint*, September 2020. doi: 10.1101/2020.09.14.20194548. URL `https://doi.org/10.1101/2020.09.14.20194548`.

Shixiang Zhu, Alexander Bukharin, Liyan Xie, Mauricio Santillana, Shihao Yang, and Yao Xie. High-resolution spatio-temporal model for county-level covid-19 activity in the us. *ACM Transactions on Management Information Systems (TMIS)*, 12(4):1–20, 2021a.

Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *medRxiv*, 2020. doi: 10.1101/2020.06.06.20124149. URL `https://www.medrxiv.org/content/early/2020/12/20/2020.06.06.20124149`.

Álvaro Gajardo and Hans-Georg Müller. Point process models for COVID-19 cases and deaths. *Journal of Applied Statistics*, pages 1–16, March 2021. doi: 10.1080/02664763.2021.1907839. URL `https://doi.org/10.1080/02664763.2021.1907839`.

Paolo Giudici, Paolo Pagnottoni, and Alessandro Spelta. Network self-exciting point processes to measure health impacts of COVID-19. *SSRN Electronic Journal*, 2021. doi: 10.2139/ssrn.3892998. URL `https://doi.org/10.2139/ssrn.3892998`.

Shuang Li, Lu Wang, Xinyun Chen, Yixiang Fang, and Yan Song. Understanding the spread of COVID-19 epidemic: A spatio-temporal point process view. *CoRR*, abs/2106.13097, 2021. URL `https://arxiv.org/abs/2106.13097`.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2016. doi: 10.1145/2939672.2939875. URL `https://doi.org/10.1145/2939672.2939875`.

Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *CoRR*, abs/1612.09328, 2016. URL `http://arxiv.org/abs/1612.09328`.

Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive Hawkes process. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11183–11193. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/zhang20q.html.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988. ISSN 01621459. URL http://www.jstor.org/stable/2288914.

Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998. URL https://doi.org/10.1023/A:1003403601725.

Shixiang Zhu, Shuang Li, Zhigang Peng, and Yao Xie. Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 2021b.

Shixiang Zhu, Alexander Bukharin, Liyan Xie, Shihao Yang, Pinar Keskinocak, and Yao Xie. Early detection of covid-19 hotspots using spatio-temporal data. *arXiv preprint arXiv:2106.00072*, 2021c.

Shixiang Zhu, Ruyi Ding, Minghe Zhang, Pascal Van Hentenryck, and Yao Xie. Spatio-temporal point processes with attention for traffic congestion event modeling. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12, 2021d. doi: 10.1109/TITS.2021.3068139.

Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017. doi: 10.1145/3038912.3052650. URL http://dx.doi.org/10.1145/3038912.3052650.

Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention, 2017.

Wikipedia. Cali. https://en.wikipedia.org/wiki/Cali, 2021.

Presidency of the Republic of Colombia. Decrees during the COVID-19 pandemic, 2020. https://coronaviruscolombia.gov.co/Covid19/decretos.html.

Allison James, Lesli Eagle, Cassandra Phillips, D. Stephen Hedges, Cathie Bodenhamer, Robin Brown, J. Gary Wheeler, and Hannah Kirking. High covid-19 attack rate among attendees at events at a church — arkansas, march 2020, 2021. https://www.cdc.gov/mmwr/volumes/69/wr/mm6920e2.htm.

Peter J. Brockwell and Richard A. Davis. Estimation of the mean and the autocovariance function. In *Springer Series in Statistics*, pages 218–237. Springer New York, 1991. doi: 10.1007/978-1-4419-0320-4_7. URL https://doi.org/10.1007/978-1-4419-0320-4_7.

Jonatan A. González, Francisco J. Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: A review. *Spatial Statistics*, 18:505–544, 2016.

Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 2017. doi: 10.1214/17-STS629.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1): 83–90, 1971. ISSN 00063444. URL `http://www.jstor.org/stable/2334319`.

D. Higdon, Jenise L. Swall, and J. Kern. Non-stationary spatial modeling, 1998.

Shixiang Zhu, Haoyun Wang, Xiuyuan Cheng, and Yao Xie. Neural spectral marked point processes. *arXiv preprint arXiv:2106.10773*, 2021e.

CJ Willmott and K Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82, 2005. doi: 10.3354/cr030079. URL `https://doi.org/10.3354/cr030079`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

A Derivation of the point process log-likelihood

Assume that we have total number of $\mathbb{N}([0,T] \times \mathcal{S})$ observations in $\boldsymbol{x}$. For any given $t \in [0,T]$, we assume that n events happened before t and denote the occurrence time of the latest event as t_n. Let $\Omega = [t, t+dt) \times B(s, ds)$ where $s \in \mathcal{S}$. Let $F(t) = \mathbb{P}(\boldsymbol{x}_{n+1}, t_{n+1} < t | \mathcal{H}_{t_n} \cup \boldsymbol{x}_n)$ be the conditional cumulative probability function, and $\mathcal{H}_{t_n} \cup \boldsymbol{x}_n$ represents the history events happened up to time t_n and at t_n. Let $f(t,s) \triangleq f(t, s|\mathcal{H}_{t_n} \cup \boldsymbol{x}_n)$ be the corresponding conditional probability density function of new event happening in Ω. As defined in (1), $\lambda(t,s)$ can be expressed as

$$\begin{aligned}\lambda(t,s) &= \mathbb{P}\{\boldsymbol{x}_{n+1} \in \Omega | \mathcal{H}_t\} = \mathbb{P}\{\boldsymbol{x}_{n+1} \in \Omega | \mathcal{H}_{t_n} \cup \boldsymbol{x}_n \cup \{t_{n+1} \geq t\}\} \\ &= \frac{\mathbb{P}\{\boldsymbol{x}_{n+1} \in \Omega, t_{n+1} \geq t | \mathcal{H}_{t_n} \cup \boldsymbol{x}_n\}}{\mathbb{P}\{t_{n+1} \geq t | \mathcal{H}_{t_n} \cup \boldsymbol{x}_n\}} \\ &= \frac{f(t,s)}{1-F(t)}\end{aligned}$$

We multiply the differential of time and space $dtds$ on both side of the equation, and integral over s

$$dt \cdot \int_{\mathcal{S}} \lambda(t,u)du = \frac{dt \cdot \int_{\mathcal{S}} f(t,u)du}{1-F(t)} = \frac{dF(t)}{1-F(t)} = -d\log(1-F(t)).$$

Hence, integrating over t on (t_n, t) leads to $F(t) = 1 - \exp(-\int_{t_n}^t \int_{\mathcal{S}} \lambda(\tau,u)dud\tau)$ because $F(t_n) = 0$. Then we have

$$f(t,s) = \lambda(t,s) \cdot \exp\left(-\int_{t_n}^t \int_{\mathcal{S}} \lambda(\tau,u)dud\tau\right) ,$$

The joint p.d.f. for a realization is then, by the chain rule, $f(x_1, ..., x_{\mathbb{N}([0,T]\times\mathcal{S})}) = \prod_{i=1}^{\mathbb{N}([0,T]\times\mathcal{S})} f(t_i, s_i)$. Then the log-likelihood of an observed sequence $\boldsymbol{x}$ can be written as

$$l(\boldsymbol{x}) = \sum_{i=1}^{\mathbb{N}([0,T]\times\mathcal{S})} \log\lambda(t_i, s_i) - \int_0^T \int_{\mathcal{S}} \lambda(\tau,u)dud\tau.$$

B Derivation of the non-stationary spatial kernel

In this section we prove the formulation of the function $v(s,s')$ between two bivariate normal kernels as appears in (4). Let two independent bivariate Gaussian variables $X_s, X_{s'}$ be centered at locations s, s' with $\Sigma_s, \Sigma_{s'}$ parameterized by

$$\Sigma_s = \begin{pmatrix} a^2 & \rho ab \\ \rho ab & b^2 \end{pmatrix}, \ \Sigma_{s'} = \begin{pmatrix} {a'}^2 & \rho' a' b' \\ \rho' a' b' & {b'}^2 \end{pmatrix},$$

By common knowledge, the probability density function f_Z of the sum Z of two independent random variables X, Y, i.e $Z = X + Y$, is the convolution of the probability density functions f_X and f_Y, i.e.

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x)dx,$$

In our case, let us denote the probability density function of $X_s, X_{s'}$ as $\kappa_s(\cdot), \kappa_{s'}(\cdot)$. Then, we have

$$f_{X_s+X_{s'}}(x) = \int_{\mathbb{R}^2} \kappa_s(u)\kappa_{s'}(x-u)du,$$

We also have the following equalities due to the property of Gaussianity

$$\kappa_s(2s-u) = \kappa_s(u), \kappa_{s'}(2s'-u) = \kappa_{s'}(u).$$

Writing $x = 2s'$, we therefore have

$$f_{X_s+X_{s'}}(2s') = \int_{\mathbb{R}^2} \kappa_s(u)\kappa_{s'}(u)du = v(s,s'),$$

Now it is easy to write that $X_s + X_{s'} \sim \mathcal{N}(s+s', \Sigma_s + \Sigma_{s'})$, and thus

$$\begin{aligned} v(s,s') = f_{X_s+X_{s'}}(2s') &= \frac{1}{2\pi|\Sigma_s+\Sigma_{s'}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(s'-s)^\top(\Sigma_s+\Sigma_{s'})^{-1}(s'-s)\right\} \\ &= \frac{1}{q_1}\exp\left\{-\frac{1}{q_2}(s-s')^\top W(s-s')\right\}, \end{aligned}$$

where

$$\begin{aligned} W &= \begin{pmatrix} b^2+{b'}^2 & -(\rho ab+\rho' a'b') \\ -(\rho ab+\rho' a'b') & a^2+{a'}^2 \end{pmatrix}, \\ q_1 &= 2\pi|\Sigma_s+\Sigma_{s'}|^{\frac{1}{2}} \\ &= 2\pi\sqrt{-(2\rho\rho' aa'bb' + a^2((\rho^2-1)b^2-{b'}^2) + {a'}^2(({\rho'}^2-1){b'}^2-b^2))}, \\ q_2 &= -2(2\rho\rho' aa'bb' + a^2((\rho^2-1)b^2-{b'}^2) + {a'}^2(({\rho'}^2-1){b'}^2-b^2)). \qquad \square \end{aligned}$$

C Derivation of the covariance function

Assume an ellipse centered at the origin with area A and two focus points of the ellipse in $\mathbb{R}^2$ which are $(\boldsymbol{\psi}_x, \boldsymbol{\psi}_y), (-\boldsymbol{\psi}_x, -\boldsymbol{\psi}_y)$, where $\boldsymbol{\psi}_x, \boldsymbol{\psi}_y \in \mathbb{R}$. In what follows, we use the same notation of Σ as before. For the ellipse parameters, we denote the semi-major and semi-minor axis of the ellipse as σ_1, σ_2. According to the ellipse formula, we have

$$\left\{ \begin{array}{rl} \pi\sigma_1\sigma_2 & = A \\ \sigma_1^2 - \sigma_2^2 & = \boldsymbol{\psi}_x^2 + \boldsymbol{\psi}_y^2 = \|\boldsymbol{\psi}\|^2 \end{array} \right. .$$

We can also compute

$$\sigma_1 = \left(\frac{\sqrt{4A^2+\|\boldsymbol{\psi}\|^4\pi^2}}{2\pi} + \frac{\|\boldsymbol{\psi}\|^2}{2}\right)^{\frac{1}{2}}, \quad \sigma_2 = \left(\frac{\sqrt{4A^2+\|\boldsymbol{\psi}\|^4\pi^2}}{2\pi} - \frac{\|\boldsymbol{\psi}\|^2}{2}\right)^{\frac{1}{2}}. \tag{8}$$

As the rotation angle α of the ellipse is $\alpha = \tan^{-1}(\boldsymbol{\psi}_y/\boldsymbol{\psi}_x)$, we can write the bivariate normal random variable X as follows

$$X = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} Z,$$

where $Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ with covariance matrix $\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$.

Now we introduce the kernel scale parameter τ_z, and write down the covariance matrix of X as

$$\Sigma = \tau_z^2 \begin{pmatrix} \sigma_1^2\cos^2\alpha + \sigma_2^2\sin^2\alpha & (\sigma_1^2-\sigma_2^2)\cos\alpha\sin\alpha \\ (\sigma_1^2-\sigma_2^2)\cos\alpha\sin\alpha & \sigma_1^2\sin^2\alpha + \sigma_2^2\cos^2\alpha \end{pmatrix}.$$

Plugging (8) into this equation we get

$$\Sigma_s = \tau_z^2 \begin{pmatrix} Q + \frac{\|\boldsymbol{\psi}\|^2}{2}\cos 2\alpha & \frac{\|\boldsymbol{\psi}\|^2}{2}\sin 2\alpha \\ \frac{\|\boldsymbol{\psi}\|^2}{2}\sin 2\alpha & Q - \frac{\|\boldsymbol{\psi}\|^2}{2}\cos 2\alpha \end{pmatrix}, \tag{9}$$

where $Q = \sqrt{4A^2 + \|\boldsymbol{\psi}\|^4\pi^2}/2\pi, \alpha = \tan^{-1}(\boldsymbol{\psi}_y/\boldsymbol{\psi}_x)$.

D Proof of Section 4

For convenience, we denote our approximation, $\lambda_0|\mathcal{S}|T+T\sum_{l=1}^{L}\gamma_l+\sqrt{2\pi}C\sigma_0\sum_{i=1}^{\mathbb{N}([0,T]\times\mathcal{S})}\left\{h\left(\frac{T-t_i}{\sigma_0}\right)-\frac{1}{2}\right\}$, of the intractable double integral as I_{Approx}. Error ϵ_1 satisfies $\int_0^T\int_{\mathcal{S}}\lambda(\tau,r)drd\tau+\epsilon_1=\int_0^T\int_{\mathbb{R}^2}\lambda(\tau,r)drd\tau$ according to the first assumption. Based on the second assumption, we approximate the location-dependent kernel induced feature functions κ_s with κ_s^0 to solve the intractable integration in (6). We first derive the upper bound $\eta_{\text{bound}}(A,c)$ of the approximation error $\left\langle\kappa_s^{(r_1)},\kappa_{s'}^{(r_2)}\right\rangle-\left\langle\kappa_s^0,\kappa_{s'}^{(r_2)}\right\rangle$. Given any location s and history event s', for the convenience of computation while without loss of generality, we can locate the origin of the coordinate system at s and align x-axis and y-axis with the semi-major and semi-minor axis of the one standard ellipse of κ_s. Thus according to Section B, the second inner-product $\left\langle\kappa_s^0,\kappa_{s'}^{(r_2)}\right\rangle$ equals to the probability density function $f_{X_0+X_{s'}^{(r_2)}}(2(s'-s))$ of the summation of two independent random variable X_0 and $X_{s'}^{(r_2)}$, where $X_0\sim\mathcal{N}(\mathbf{0},\Sigma_0),\Sigma_0=\frac{\tau_z^2A}{\pi}\mathbf{I}$ and $X_{s'}^{(r_2)}\sim\mathcal{N}(s'-s,\Sigma_{s'}^{(r_2)})$, $\Sigma_{s'}^{(r_2)}$ is the covariance matrix of $\kappa_{s'}^{(r_2)}$ in the preset coordinate system. For the first term we re-write the inner-product in polar coordinate system and have

$$\begin{aligned}
\left\langle\kappa_s^{(r_1)},\kappa_{s'}^{(r_2)}\right\rangle &= \int_{\mathbb{R}^2}\kappa_s^{(r_1)}(u)\kappa_{s'}^{(r_2)}(u)du \\
&= \int_0^{2\pi}\int_0^{+\infty}\kappa_s^{(r_1)}(r,\theta)\kappa_{s'}^{(r_2)}(r,\theta)rdrd\theta \\
&\overset{\text{(i)}}{=} \int_0^{2\pi}\int_0^{+\infty}\frac{r}{2\pi\tau_z^2\sqrt{Q^2-\frac{\|\boldsymbol{\psi}_s\|^4}{4}}}\exp\left\{-\frac{r^2}{2\tau_z^2}\left(\frac{\cos^2\theta}{Q+\frac{\|\boldsymbol{\psi}_s\|^2}{2}}+\frac{\sin^2\theta}{Q-\frac{\|\boldsymbol{\psi}_s\|^2}{2}}\right)\right\}\cdot\kappa_{s'}^{(r_2)}(r,\theta)drd\theta \\
&= \int_0^{2\pi}\int_0^{+\infty}\frac{r}{2\tau_z^2A}\exp\left\{-\frac{r^2\pi^2}{2\tau_z^2A^2}\left(Q-\frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\theta\right)\right\}\cdot\kappa_{s'}^{(r_2)}(r,\theta)drd\theta \\
&\overset{\text{(ii)}}{\leq} \int_0^{2\pi}\int_0^{+\infty}\frac{r}{2\tau_z^2A}\exp\left\{-\frac{r^2\pi}{\tau_z^2(\sqrt{4A^2+c^4\pi^2}+c^2\pi)}\right\}\cdot\kappa_{s'}^{(r_2)}(r,\theta)drd\theta, \quad \forall r_1,r_2.
\end{aligned} \tag{10}$$

Here Q is defined in Append C. We plug in the analytical form of $\kappa_s^{(r_1)}(r,\theta)$ at (i). The inequality at (ii) holds because:

$$\begin{aligned}
Q-\frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\theta \geq Q-\frac{\|\boldsymbol{\psi}_s\|^2}{2} &= \frac{\sqrt{4A^2+\|\boldsymbol{\psi}_s\|^4\pi^2}-\|\boldsymbol{\psi}_s\|^2\pi}{2\pi} \\
&= \frac{2A^2}{\pi(\sqrt{4A^2+\|\boldsymbol{\psi}_s\|^4\pi^2}+\|\boldsymbol{\psi}_s\|^2\pi)} \\
&\geq \frac{2A^2}{\pi(\sqrt{4A^2+c^4\pi^2}+c^2\pi)} \quad \text{(because } \|\boldsymbol{\psi}_s\|\leq c\text{)}\ .
\end{aligned}$$

For the final formula on the right of the inequality operator, we find that

$$\begin{aligned}
&\frac{1}{2\tau_z^2 A}\exp\left\{-\frac{r^2\pi}{\tau_z^2(\sqrt{4A^2+c^4\pi^2}+c^2\pi)}\right\} \\
=&\frac{\sqrt{4A^2+c^4\pi^2}+c^2\pi}{2A}\cdot\frac{1}{2\pi\frac{\tau_z^2}{2}\left(\sqrt{4A^2/\pi^2+c^4}+c^2\right)}\exp\left\{-\frac{r^2}{2\frac{\tau_z^2}{2}\left(\sqrt{4A^2/\pi^2+c^4}+c^2\right)}\right\}.
\end{aligned}$$

It takes the form of the multiplication of a constant and a probability density function of a Gaussian distribution with zero mean and covariance matrix Σ_1, where $\Sigma_1 = \tau_z^2\frac{\sqrt{4A^2/\pi^2+c^4}+c^2}{2}\mathbf{I}$. We assume a random variable X_1 conforms to the corresponding distribution $X_1 \sim \mathcal{N}(\mathbf{0}, \Sigma_1)$ and denote the constant $\frac{\sqrt{4A^2+c^4\pi^2}+c^2\pi}{2A}$ as U (the same one in the proposition 1). Combining the result of (10) we can write the upper bound of the approximation error as

$$\left\langle \kappa_s^{(r_1)}, \kappa_{s'}^{(r_2)}\right\rangle - \left\langle \kappa_s^0, \kappa_{s'}^{(r_2)}\right\rangle \le U \cdot f_{X_1+X_{s'}^{(r_2)}}(2(s'-s)) - f_{X_0+X_{s'}^{(r_2)}}(2(s'-s)) \ . \tag{11}$$

We introduce a new notation $\upsilon_0(s,s')$ to denote the spatial kernel with κ_s replaced with κ_s^0, that is $\upsilon_0(s,s') = \sum_{(r_1,r_2)\in[R]\times[R]} w_s^{(r_1)} w_{s_i}^{(r_2)} \left\langle \kappa_s^0, \kappa_{s_i}^{(r_2)}\right\rangle$, $[R] = \{1,2,...,R\}$. Based on above results, for any given history event s_i we have

$$\begin{aligned}
&\int_{\mathbb{R}^2} \upsilon(u,s_i)du - \int_{\mathbb{R}^2} \upsilon_0(u,s_i)du \\
=&\int_{\mathbb{R}^2} \sum_{(r_1,r_2)\in[R]\times[R]} w_u^{(r_1)} w_{s_i}^{(r_2)} \left(\left\langle \kappa_u, \kappa_{s_i}^{(r_2)}\right\rangle - \left\langle \kappa_u^0, \kappa_{s_i}^{(r_2)}\right\rangle\right) du \\
\le&\int_{\mathbb{R}^2} \sum_{(r_1,r_2)\in[R]\times[R]} w_u^{(r_1)} w_{s_i}^{(r_2)} \left(U\cdot f_{X_1+X_{s_i}^{(r_2)}}(2(s_i-u)) - f_{X_0+X_{s'}^{(r_2)}}(2(s'-u))\right) du \quad \text{(Plug in (11))} \\
=&\int_{\mathbb{R}^2} \sum_{r_2=1}^{R} w_{s_i}^{(r_2)} \left(U\cdot f_{X_1+X_{s'}^{(r_2)}}(2(s'-u)) - f_{X_0+X_{s'}^{(r_2)}}(2(s'-u))\right) du \quad \text{(sum over } r_1\text{)} \\
=&\sum_{r_2=1}^{R} w_{s_i}^{(r_2)} \left(U\int_{\mathbb{R}^2} f_{X_1+X_{s'}^{(r_2)}}(2(s'-u))du - \int_{\mathbb{R}^2} f_{X_0+X_{s'}^{(r_2)}}(2(s'-u))du\right) \\
=&\sum_{r_2=1}^{R} w_{s_i}^{(r_2)} (U-1) \,\text{(Integration of probability density function over } \mathbb{R}^2 \text{ equals to 1)} \\
=&\,U-1 \ .
\end{aligned}$$

Also notice that

$$\begin{aligned}
\int_{\mathbb{R}^2} \upsilon_0(u,s_i)du &= \int_{\mathbb{R}^2} \sum_{(r_1,r_2)\in[R]\times[R]} w_u^{(r_1)} w_{s_i}^{(r_2)} \left\langle \kappa_s^0, \kappa_{s_i}^{(r_2)}\right\rangle du \\
&= \int_{\mathbb{R}^2} \sum_{r_2=1}^{R} w_{s_i}^{(r_2)} \left\langle \kappa_s^0, \kappa_{s_i}^{(r_2)}\right\rangle du \quad \left(\text{because} \sum_{r_1=1}^{R} w_u^{(r_1)} = 1\right) \\
&= \sum_{r_2=1}^{R} w_{s_i}^{(r_2)} \int_{\mathbb{R}^2} \left\langle \kappa_s^0, \kappa_{s_i}^{(r_2)}\right\rangle du = \sum_{r_2=1}^{R} w_{s_i}^{(r_2)} = 1 \ .
\end{aligned}$$

Thus the upper bound of ϵ_2 can be controlled by

$$\begin{aligned}
\epsilon_2 &= \left(\int_0^T \int_{\mathbb{R}^2} \lambda(\tau, r) dr d\tau - I_{\text{Approx}}\right) \Big/ I_{\text{Approx}} \\
&= \left\{\lambda_0 |\mathcal{S}| T + T \sum_{l=1}^{L} \gamma_l + \int_0^T \sum_{t_i < \tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau \cdot \int_{\mathbb{R}^2} \upsilon(u, s_i) du - I_{\text{Approx}}\right\} \Big/ I_{\text{Approx}} \\
&= \left\{\int_0^T \sum_{t_i < \tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau \left(\int_{\mathbb{R}^2} \upsilon(u, s_i) du - 1\right)\right\} \Big/ I_{\text{Approx}} \\
&= \left\{\int_0^T \sum_{t_i < \tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau \int_{\mathbb{R}^2} \left(\upsilon(u, s_i) - \upsilon_0(u, s_i)\right) du\right\} \Big/ I_{\text{Approx}} \\
&\leq \left\{(U - 1) * \int_0^T \sum_{t_i < \tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau\right\} \Big/ I_{\text{Approx}} \ ,
\end{aligned}$$

because $I_{\text{Approx}} > \int_0^T \sum_{t_i<\tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau = \sqrt{2\pi} C \sigma_0 \sum_{i=1}^{\mathbb{N}([0,T]\times\mathcal{S})} \left\{h\left(\frac{T - t_i}{\sigma_0}\right) - \frac{1}{2}\right\}$, we have

$$\epsilon_2 < U - 1 \ . \tag{12}$$

On the other hand,

$$\begin{aligned}
\left\langle \kappa_s^{(r_1)}, \kappa_{s'}^{(r_2)} \right\rangle &= \int_0^{2\pi} \int_0^{+\infty} \frac{r}{2\tau_z^2 A} \exp\left\{-\frac{r^2\pi^2}{2\tau_z^2 A^2}\left(Q - \frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\theta\right)\right\} \cdot \kappa_{s'}^{(r_2)}(r, \theta) dr d\theta \\
&\overset{\text{(iii)}}{\geq} \int_0^{2\pi} \int_0^{+\infty} \frac{r}{2\tau_z^2 A} \exp\left\{-\frac{r^2\pi(\sqrt{4A^2 + c^4\pi^2} + c^2\pi)}{4\tau_z^2 A^2}\right\} \cdot \kappa_{s'}^{(r_2)}(r, \theta) dr d\theta \\
&= \frac{1}{U} f_{X_2 + X_{s'}^{(r_2)}}(2(s' - s)) - f_{X_0 + X_{s'}^{(r_2)}}(2(s' - s)) \ ,
\end{aligned} \tag{13}$$

where $X_2 \sim \mathcal{N}(\mathbf{0}, \Sigma_2), \Sigma_2 = \frac{\tau_z^2 A}{\pi U}\mathbf{I}$. The establishment of (iii) can be deduced by

$$\begin{aligned}
Q - \frac{\|\boldsymbol{\psi}_s\|^2}{2}\cos 2\theta \leq Q + \frac{\|\boldsymbol{\psi}_s\|^2}{2} \\
&= \frac{\sqrt{4A^2 + \|\boldsymbol{\psi}_s\|^4\pi^2} + \|\boldsymbol{\psi}_s\|^2\pi}{2\pi} \\
&\leq \frac{\sqrt{4A^2 + c^4\pi^2} + c^2\pi}{2\pi} \ .
\end{aligned}$$

We can use the lower bound (13) to similarly obtain

$$\int_{\mathbb{R}^2} \upsilon(u, s_i) du - \int_{\mathbb{R}^2} \upsilon_0(u, s_i) du \geq \frac{1}{U} - 1 \ .$$

Thus we have

$$\epsilon_2 \geq \left\{\left(\frac{1}{U} - 1\right) * \int_0^T \sum_{t_i < \tau} C e^{-\frac{1}{2\sigma_0^2}(\tau - t_i)^2} d\tau\right\} \Big/ I_{\text{Approx}} \geq \frac{1}{U} - 1 \ . \ (\text{because } \frac{1}{U} - 1 < 0)$$

Combining the result (12) we have

$$|\epsilon_2| < \max\left\{U - 1, 1 - \frac{1}{U}\right\} . \qquad \square$$

E Detailed description of the baseline methods

In this section, we provide a detailed description of three baseline models used in Section 5 and their hyper-parameter selection.

Homogeneous Poisson process assumes events occur at a constant intensity λ over space. The parameter indicates the expected number of events occurring in a unit interval or region. We estimate λ using the average number of confirmed cases over time and space and randomly sample events in the spatio-temporal space. The results of homogeneous Poisson process act as a sanity check.

Susceptible-Infectious-Recovered (SIR) model is one of the most fundamental compartmental models that aim to model infectious disease spread. It splits the whole population into three compartments of susceptible ($\boldsymbol{S}$), infected ($\boldsymbol{I}$) and recovered ($\boldsymbol{R}$) individuals. SIR makes each compartment a function of t since the population of each compartment may vary over time, and three ordinary differential equations about these three functions can describe the integral SIR system. Parameters β_{SIR} and γ_{SIR} represent the emerging rate of new infections and the recovery rate of patients, respectively. Both parameters are fitted according to the real data based on least squares. We fit a SIR model for each comuna, choosing the initial infected population $\boldsymbol{I}(0)$ to be the number of cases at the week of the first case.

Linear prediction is another popular method used to do forecasting tasks. We choose an autoregressive (AR) time series model to predict the number of infected cases. It specifies that the current output value depends linearly on its history and a white noise term. A parameter p in AR model represents the number of most recent lags that the current output depends on, which can be determined by choosing the appropriate number of significant lags of PACF about the data series. We choose $p = 3$ for the AR model according to PACF plots of confirmed case series in each comuna in Fig. 14.

ETAS is a benchmark model in modeling specific spatio-temporal data, as we mentioned in Section 3. We replace the spatio-temporal kernel in (2) with a Gaussian diffusion kernel. We estimate model parameters by applying stochastic gradient descent with regard to model likelihood.

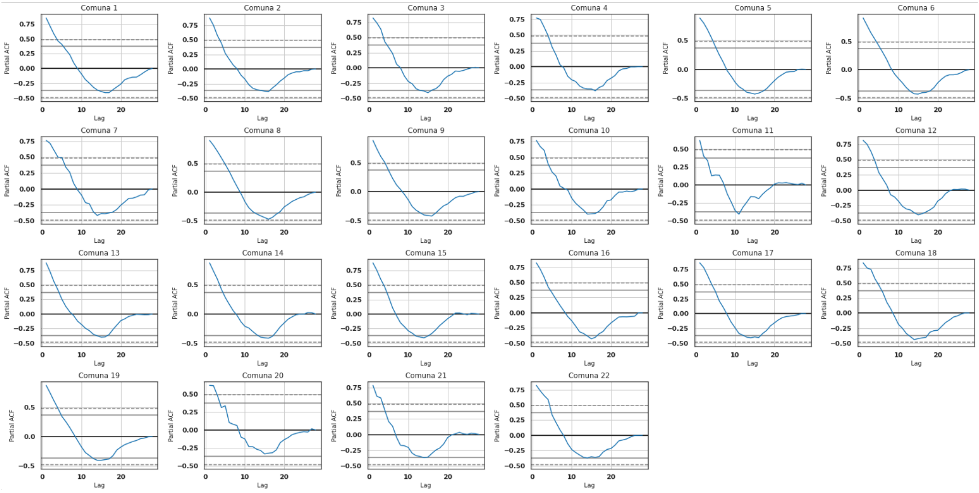


Figure 14: PACF plot for each comuna. The x-axis is the lag number of the time series itself to the current output and the y-axis is the value of PACF at the corresponding lag. For example, the PACF of a time series $\boldsymbol{X}$ at lag 2 refers to the partial autocorrelation between $\boldsymbol{X}_t$ and $\boldsymbol{X}_{t-2}$. The dash line represents the lower bound of significant PACF value.