Statistical Learning for Spatial Data: Theory and Practice

PhD Defense, Emilia Siviero
December 2nd, 2024
Télécom Paris

Why Spatial Data?

Natural phenomena:

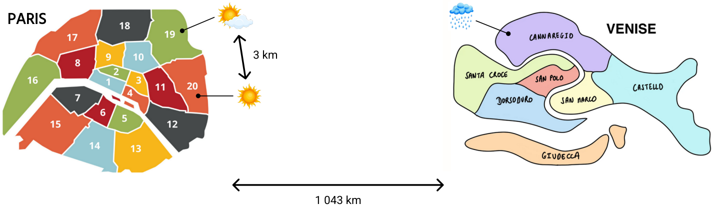
- weather
- earthquakes
- rivers
- environment
- vegetation

Humanmade:

- urban planning
- public services
- crimes
- agriculture

Spatial Data Characteristics

1. Spatial dependence structure



2. One single realization

Natural event

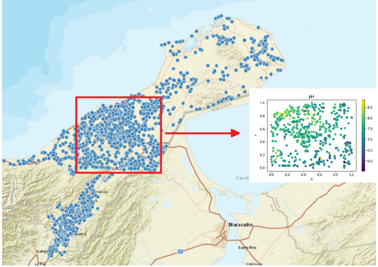
Deterioration of the environment

General Spatial Model

$$\mathbf{X} = \left\{ \mathbf{X}_s,\ s \in \mathcal{S} \right\}$$

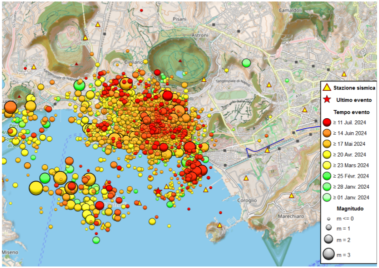
Geostatistical Data:

- **Observations:** fixed, irregularly or regularly sampled
- **Goal:** modeling, prediction
- **Example:** hydrogeology (pH value)



Point Patterns:

- **Observations:** locations (and number n) are random
- **Goal:** capturing a pattern in data
- **Example:** seismology



Geostatistics: History and Applications

1950s: Danie G. Krige's study in mineral deposit (Krige, 1951)

1960s: Georges Matheron lays foundations of Geostatistics theory (Matheron, 1962)

Covariance function describes the dependence structure of data

Unknown in practice $\Longrightarrow$ need to be **estimated**

Meteorology: weather patterns, climate trends, and atmospheric phenomena (Goovaerts, 2000)

Environment: changes due to human influence or other natural forces (Webster and Oliver, 2007; Cressie, 1993b)

Healthcare: spatial patterns of disease incidence and mortality (Oliver et al., 1998)

Point Processes: History and Applications

1970s: Temporal Hawkes Processes (HP) (Hawkes, 1971)

1980s: Introduction of HP to earthquake modeling (Ogata, 1988)

1990s: extension to spatio-temporal data: earthquakes exhibit both spatial and temporal clustering (Musmeci and Vere-Jones, 1992)

Seismology: mainshock-aftershock pattern (clustering and triggering) (Ogata, 1988; Daley and Vere-Jones, 2003)

Criminology: 'near-repeat victimization' pattern (Mohler et al., 2011; D'Angelo et al., 2022)

Epidemiology: spread of infectious diseases, patterns in disease occurrence (Meyer and Held, 2014; Kresin et al., 2022)

Research Questions

How to learn from spatial data that presents a **dependence structure**?
How does the dependence structure of the observed phenomenon affect the **performance** of the algorithms?

GEOSTATISTICS

1. How **accurate** is the empirical covariance estimator?

2. What is the **non-asymptotic** performance of the Kriging predictor?

POINT PROCESS

3. How to overcome the **numerical and modeling challenges** when learning from a spatio-temporal Hawkes process?

4. How to accurately model **real-world situations?**

A Statistical Learning View of Simple Kriging

Machine Learning:

- **Assets:** statistical learning theory for independent data, non-parametric theory
- **Limits:** very few theoretical guarantees for spatial data

Spatial Analysis:

- **Assets:** take advantage of spatial structure (modelled by covariance function)
- **Limits:** very few non-parametric theories

- **Limits:** lack of non-asymptotic results for spatial data

Challenge 1: Provide statistical guarantees for **prediction**, under the form of **non-asymptotic** bounds, for **non-parametric** methods in the context of spatial data.

State-of-the-Art

Methods	Parametric	Zimmerman 1989; Zimmerman and Cressie 1992	✗ requires selection of a model and estimation of unknown parameters
	Non-Parametric	Hall and Patil 1994; Elogne et al. 2008	✓ more flexible methods for massive spatial datasets
Results	Asymptotic	Stein 1999	✗ mainly asymptotic results
	Non-Asymptotic with independent copies	Qiao et al. 2018	✗ concentration bounds for independent copies of the spatial process

Notations

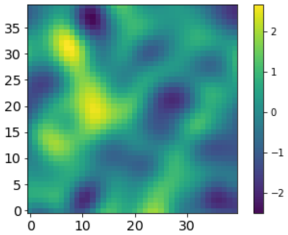
- $\mathcal{S} \subseteq \mathbb{R}^2$: **spatial domain**
- $C(s,t)$: **covariance function** $\qquad C(s,t) = \mathrm{Cov}(\mathbf{X}_s, \mathbf{X}_t)$
- $\mathbf{X}(\mathbf{s}_d)$: **observations of X** at locations $\mathbf{s}_d = (s_i)_{1\leq i\leq d}$. $\qquad \mathbf{X}(\mathbf{s}_d) = (\mathbf{X}_{s_i})_{1\leq i\leq d}$
- $\mathbf{c}_d(s)$: **covariance vector** $\qquad \mathbf{c}_d(s) = \left(\mathrm{Cov}(\mathbf{X}_s, \mathbf{X}_{s_i})\right)_{1\leq i\leq d}$
- $\Sigma(\mathbf{s}_d)$: **covariance matrix** $\qquad \Sigma(\mathbf{s}_d) = \mathrm{Var}(\mathbf{X}(\mathbf{s}_d))$
- $\mathbf{X}'$: **one single realization of X** at locations $\sigma_n = (\sigma_i)_{1\leq i\leq n}$. $\qquad \mathbf{X}'(\sigma_n) = \left(\mathbf{X}'_{\sigma_i}\right)_{1\leq i\leq n}$

Simple Kriging Problem

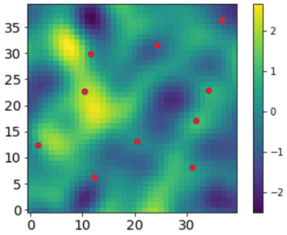
Simple Kriging: **predict** the value of **X** at some unobserved location s, based on d sampled observations $(\mathbf{X}_{s_i})_{i\leq d}$, assuming a **linear combination** of the observations: $f_\lambda(s) = \langle \lambda(s), \mathbf{X}(\mathbf{s}_d)\rangle$, such that λ minimizes the variance.

$$f_{\lambda^*}(s) = \mathbf{X}(\mathbf{s}_d)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$$

$\hookrightarrow$ the weights λ^* depend on covariance function and s



$\{X_s, s \in S\}$ random field



$(X_{s_i})_{i\leq d}$ sampled observations

Plug-in Predictive Rule

Theoretical:

$f_{\lambda^*}(s) = \mathbf{X}(\mathbf{s}_d)^\top \Sigma(\mathbf{s}_d)^{-1}\mathbf{c}_d(s)$

⟶ covariance function known

⟶ predictor is **BLUP** (Best Linear Unbiased Predictor)

covariance estimation ⟶ **plug-in rule**

Empirical:

$f_{\widehat{\lambda}}(s) = \mathbf{X}(\mathbf{s}_d)^\top \widehat{\Sigma}(\mathbf{s}_d)^{-1}\widehat{\mathbf{c}}_d(s)$

⟶ covariance function unknown and need to be **estimated**

⚠ no guarantees of optimality

⟹ **Motivation:** Need to establish **rate bounds** that assess the **generalization capacity** of the resulting predictive map

Statistical Learning Guarantees

Accuracy of predictor: measured by the integrated Mean Squared Error (IMSE) over the spatial domain $\mathcal{S}$:

$$\begin{aligned} L_{\mathcal{S}}(f_\lambda) &= \mathbb{E}_X\left[\int_{s\in\mathcal{S}} (f_\lambda(s) - \mathbf{X}_s)^2 \, ds\right] \\ &= \int_{s\in\mathcal{S}} \left(Var(\mathbf{X}_s) + \ \lambda(s)^\top \Sigma(\mathbf{s}_d)\lambda(s) - 2\,\mathbf{c}_d(s)^\top \lambda(s)\right) ds \end{aligned}$$

Statistical guarantees of predictor: The global excess risk quantifies the gap between the optimal theoretical predictor and the empirical predictor errors:

$$L_{\mathcal{S}}(f_{\widehat{\lambda}}) - L_{\mathcal{S}}\left(f_{\lambda^*}\right) = \mathbb{E}_X\left[\int_{s\in\mathcal{S}} \left(f_{\widehat{\lambda}}(s) - f_{\lambda^*}(s)\right)^2 ds\right]$$

Limitation: **unique realisation for learning** with only n observations

Assumption 1: **X** second order **stationary** with **isotropic** covariance (Cressie, 1993): constant mean $\mu \in \mathbb{R}$, and invariant covariance C (depends only on distance h):

$$\exists c, C(s,t) = c(\|t-s\|) = c(h)$$

$\longrightarrow \surd$ **Solution:** **X** is sufficiently **homogeneous** inside the spatial domain (its characteristics are identical from one point to another)

Assumptions (suite)

- ***Assumption 2:*** **X Gaussian** random field with zero mean and positive definite covariance function

  $\longrightarrow \surd$ **strict stationarity**, **all laws** are known, Bochner's theorem (Stein, 1999; Hall and Patil, 1994)

- ***Assumption 3:*** **In-fill** asymptotic: number of observations **within** spatial domain $\mathcal{S}$ increases (denser and denser grid) and **regular** grid (Cressie, 1993)

  $\longrightarrow \surd$ **accurate** and **unbiased** estimator

Estimation of the Dependence Structure

How **accurate** is the empirical covariance estimator?

How data points are related to each other, based on their spatial proximity?

$\longrightarrow$ **covariance** and **semi-variogram** functions describe how the spatial correlation between data points changes with distance

Covariance:

$$c(h) = \mathbb{E}\left[(\mathbf{X}_{s+h} - \mu)(\mathbf{X}_s - \mu)\right]$$

Semi-variogram:

$$\gamma(h) = \frac{1}{2}\mathbb{E}\left[(\mathbf{X}_{s+h} - \mathbf{X}_s)^2\right]$$

Property: For all $h \in \mathbb{R}$, $\gamma(h) = c(0) - c(h)$.

Non-parametric Estimation of the Dependence Structure

Empirical semi-variogram (Matheron, 1962):

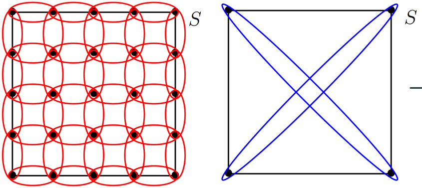
$$\widehat{\gamma}(h) = \frac{1}{2n_h} \sum_{(s_i,s_j)\in N(h)} \left(\mathbf{X}_{s_i} - \mathbf{X}_{s_j}\right)^2,$$

where $N(h)$ is the set of pairs of sites at a distance h (set of neighbors) and n_h its cardinality.

Advantages:

- **flexible** approach to the massive character of spatial datasets
- does not require knowledge of mean
- **unbiased** estimator (for regular grids)
- under Gaussianity assumption, $\widehat{\gamma}(h)$ is sum of **independent** χ^2 variables

Additional Assumptions



$\longrightarrow \surd$ **inferior bound** on n_h

- ***Assumption 4:*** $\exists\, \theta, \forall h \geq \theta,\ c(h) = 0$ **(border hypothesis)**

- ***Assumption 5:*** c is of class $\mathcal{C}^1$ and its gradient is bounded by Q **(regularity/smoothness hypothesis)**

  $\longrightarrow \surd$ for the **estimation error** at **all lags**

Corollary (Siviero et al., 2023)

Suppose that Assumptions 1–5 are satisfied. Then, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$\sup_{h \geq 0} \left| \widehat{c}(h) - c(h) \right| \leq C_3 \sqrt{\log(4n/\delta)/n} + Q/(\sqrt{n} - 1),$$

as soon as $n \geq C_3' \log(4n/\delta)$, where C_3 and C_3' are positive constants depending on θ and on the bounds of the eigenvalues of the covariance matrix solely.

Sketch of Proof

- **Distribution of $\widehat{\gamma}(h)$:** Under the Gaussian assumption **(Hyp 2)**,

$$\widehat{\gamma}(h) \sim \frac{1}{n_h} \overbrace{\sum_{i=1}^{n_h}}^{\textbf{(Hyp 3 and 4)}} \ell_i(h)\chi_i^2,$$

  where $\ell_i(h)$'s are the eigenvalues of $L(n,h)\Sigma(\sigma_n)$.

- **Poisson tail bounds:** Thanks to recent results in (Bercu et al., 2015) and (Wang and Ma, 2020):

$$\mathbb{P}\left(|\widehat{\gamma}(h) - \gamma(h)| \geq t\right) \leq e^{-C_1 nt} + e^{-C_1' nt^2},$$

  where C_1 and C_1' are positive constants depending on θ **(Hyp 4)**.

- **Estimation for all lags:** Thanks to a piece-wise constant estimator and the regularity assumption **(Hyp 5)**.

Statistical Learning Guarantees for the Kriging Method

What are the **non-asymptotic guarantees** for the Kriging predictor?

ERM: $f_{\widehat{\lambda}}$ is the empirical risk minimizer of

$$\widehat{L}_{\mathcal{S}}(f_\lambda) = \int_{s\in\mathcal{S}} \Big(\widehat{c}(0) + \ \lambda(s)^\top \widehat{\Sigma}(\mathbf{s}_d)\lambda(s) - 2\, \widehat{\mathbf{c}}_d(s)^\top \lambda(s)\Big)\, ds$$

GOAL: define **non-asymptotic** bound of global excess risk:

$$L_{\mathcal{S}}(f_{\widehat{\lambda}}) - L_{\mathcal{S}}(f_{\lambda^*}) = \int_{s\in\mathcal{S}} \Big(\widehat{\lambda}(s) - \lambda^*(s)\Big)^\top \Sigma(\mathbf{s}_d)\widehat{\lambda}(s) + \lambda^*(s)^\top \Sigma(\mathbf{s}_d)\Big(\widehat{\lambda}(s) - \lambda^*(s)\Big) - 2\, \mathbf{c}_d(s)^\top \Big(\widehat{\lambda}(s) - \lambda^*(s)\Big)\, ds$$

Non-asymptotic Bound for the Global Excess Risk

Theorem (Siviero et al., 2023)

Suppose that Assumptions 1–5 are satisfied. Then, for any $\delta \in (0,1)$, we have with probability at least $1-\delta$:

$$L_{\mathcal{S}}(f_{\widehat{\Lambda}_d}) - L_{\mathcal{S}}(f_{\lambda^*}) \leq C_6\, d^2 \sqrt{\log(4n/\delta)/n} + C_6'\, d^2\, Q/(\sqrt{n}-1),$$

as soon as $n \geq C_6'' \log(4n/\delta)$, where C_6, C_6' and C_6'' are positive constants depending on θ and on the bounds of the eigenvalues of the covariance matrix solely.

Sketch of Proof

$$\sup_{s\in\mathcal{S}} ||\widehat{\lambda}(s) - \lambda^*(s)|| \leq \underbrace{|||\Sigma(\boldsymbol{s}_d)^{-1}|||}_{N_1} \underbrace{\sup_{s\in\mathcal{S}} ||\widehat{\boldsymbol{c}}_d(s) - \boldsymbol{c}_d(s)||}_{N_2} + \underbrace{|||\widehat{\Sigma}(\boldsymbol{s}_d)^{-1} - \Sigma(\boldsymbol{s}_d)^{-1}|||}_{N_3} \underbrace{\sup_{s\in\mathcal{S}} ||\widehat{\boldsymbol{c}}_d(s)||}_{N_4},$$

where

- N_1: From bounds on eigenvalues of $\Sigma(\boldsymbol{s}_d)$
- N_2: From Corollary
- N_3: Non-asymptotic bound on the accuracy of the precision matrix estimation
- N_4: From Corollary and Assumption 5

Illustrative Experiments

Averaged MSE on 100 realisations, with two covariance models, for varying θ

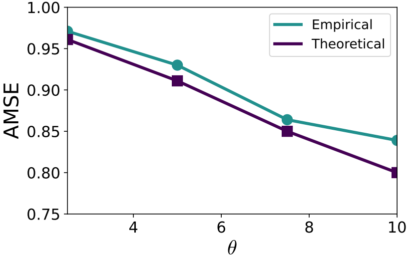


Figure 1: *Truncated power law* (TPL)

✓ satisfies all the assumptions

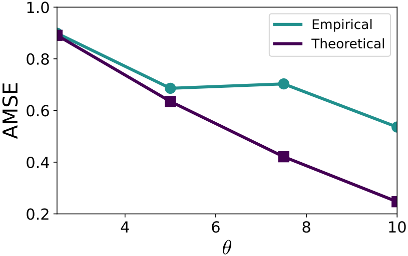


Figure 2: *Gaussian*

× not **Hyp 4**, but vanishes quickly

$\longrightarrow$ ✓ experiments **corroborate** our theoretical results: the error depends on θ (role of technical assumptions is verified)

$\longrightarrow$ ✓ results for Gaussian covariance encourages to **relax Hyp 4**

Contributions

1. **Flexible covariance estimation:**

   We develop **tail bounds** for the **non-parametric** covariance estimator.

2. **Statistical guarantees for Kriging:**

   We develop a novel theoretical framework offering **guarantees** for empirical Kriging rules in the form of **non-asymptotic bounds**.

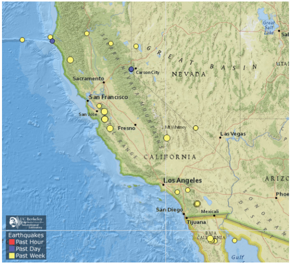
3. **Our numerical experiments** on simulated and real meteorological data corroborate our theoretical results.
   GitHub: `github.com/EmiliaSiv/Simple-Kriging-Code`

Flexible Parametric Inference for Space-Time Hawkes Processes

Motivation

Seismology

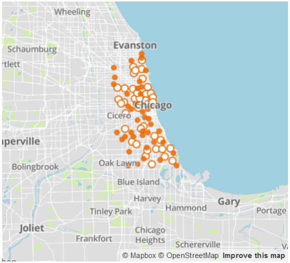
- Mainshock-aftershock clustering behavior



- (Vere-Jones, 1970; Ogata, 1988)

Criminology

- Near-repeat victimization pattern



- (Mohler, 2014; Zhu and Xie, 2022)

Challenge 2: Develop a **new, efficient, and accurate** method to predict from spatio-temporal data, such that it is **flexible** in modeling **real-world situations**.

Hawkes Processes

- **Point process:** collection of events, randomly distributed over time or space

  ↪ behavior characterized by **conditional intensity function**.

- **Hawkes (or self-exciting) process:** each event increases the likelihood of future events in its neighborhood

  ↪ **intensity** depends on time, location, and history of the process.

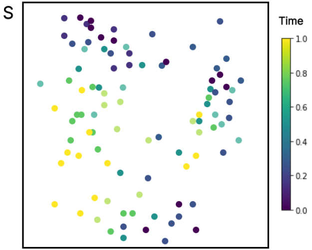
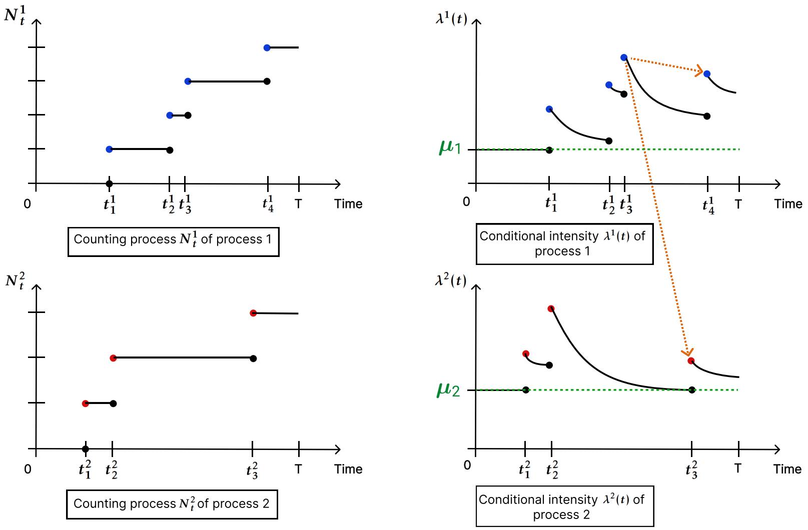


Figure: Univariate spatio-temporal point process

Conditional Intensity Function



Counting process N_t^1 of process 1

Conditional intensity $\lambda^1(t)$ of process 1

Counting process N_t^2 of process 2

Conditional intensity $\lambda^2(t)$ of process 2

Notations

Given $D \geq 1$ type of events, for each $i \in \{1, \cdots, D\}$, the intensity function of the i-th process of a multivariate spatio-temporal Hawkes process (MSTHP):

$$\lambda_i(x, y, t|\mathcal{H}_t) = \underbrace{\mu_i}_{\text{baseline}} + \sum_{j=1}^{D} \underbrace{\sum_{u_n^j \in \mathcal{H}_t^j}}_{\text{based on history}} \underbrace{\alpha_{ij}}_{\text{excitation scaling}} \underbrace{g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j)}_{\text{triggering kernel}}$$

based on history · excitation scaling · baseline · triggering kernel

- $T \in \mathbb{R}_+$: stopping time, $\mathcal{S} \subset \mathbb{R}^2$: compact set of the space domain
- $\boldsymbol{\mu}_i$: parameter controlling spontaneous event apparition rate with $\mu_i > 0$
- $\mathcal{H}_t^i$: collection of past events $u_n^i = (x_n^i, y_n^i, t_n^i),\ (x_n^i, y_n^i) \in \mathcal{S},\ t_n^i \in [0, t]$
- $\boldsymbol{\alpha}_{ij}$: describes excitation behavior between events with $0 \leq \boldsymbol{\alpha}_{ij} < 1$
- $g_{ij} : \mathcal{S} \times [0, T] \mapsto \mathbb{R}_+$: spatio-temporal *kernel* (excitation function): influence of past events onto future events

$\longrightarrow$ parameters: $\boldsymbol{\theta} = \{\boldsymbol{\mu}_i, \boldsymbol{\alpha}_{ij}, \boldsymbol{\eta}_{ij}\}_{ij}$

State-of-the-Art

Two assumptions are commonly made:

- **Space-time separability** (Mohler, 2014; Ilhan and Kozat, 2020)

  $\surd$ brings **simplicity**

  $\times$ **not realistic** (space-time interactions)

- **Constrained kernel models** (Chen et al., 2021)

  $\surd$ **computational efficiency**

  $\times$ **restrictive** for some real-world situations

$\Longrightarrow$ **Motivation:** Need to develop a parametric method allowing **(1) any kind of kernels** and **(2) space-time interactions**

Flexible Inference Approach for MSTHP

How to **accurately** model **real-world situations** using MSTHP?

Inference Approach: Key Components

Inspired by recent work of (Staerman et al., 2023), the inference approach relies on **three key components**:

1. *Discretization:* define a three dimensional regular grid $\mathcal{G} = \mathcal{G}_{\mathcal{S}} \times \mathcal{G}_T$ with $\Delta_{\mathcal{S}}, \Delta_T > 0$ the stepsizes of the spatial and temporal grids, project the observed events on these grids and define $\widetilde{\mathcal{H}}^i_T$.

   $\hookrightarrow \surd$ we can rewrite the **intensity in a discretized manner** $\tilde{\lambda}$.

2. *Finite-support Kernels:* consider the spatio-temporal kernels to be of finite lengths $W_{\mathcal{S}}$ and $W_{\mathcal{T}}$, and define $L_T = \lfloor W_T/\Delta_T \rfloor + 1$, $L_{\mathcal{S}} = \lfloor 2W_{\mathcal{S}}/\Delta_{\mathcal{S}} \rfloor + 1$ the number of points on the discretized temporal and spatial support.

   $\hookrightarrow \surd$ **reduce computational burden**.

Discretized Loss

$$
\begin{aligned}
\mathcal{L}_{\mathcal{G}}(\boldsymbol{\theta}, \widetilde{\mathcal{H}}_T) &= \sum_{i=1}^{D} \left(\Delta_{\mathcal{S}}^2 \Delta_T \sum_{v_x, v_y=0}^{G_{\mathcal{S}}} \sum_{v_t=0}^{G_T} \left(\tilde{\lambda}_i[v_x, v_y, v_t] \right)^2 - 2 \sum_{\tilde{u}_n^i \in \widetilde{\mathcal{H}}_T^i} \tilde{\lambda}_i \left[\frac{\tilde{x}_n^i}{\Delta_{\mathcal{S}}}, \frac{\tilde{y}_n^i}{\Delta_{\mathcal{S}}}, \frac{\tilde{t}_n^i}{\Delta_T} \right] \right) \\
&= (T + \Delta_T)(2S_{\mathcal{X}} + \Delta_{\mathcal{X}})(2S_{\mathcal{Y}} + \Delta_{\mathcal{Y}}) \sum_{i=1}^{D} \boldsymbol{\mu}_i^2 \\
&+ 2\Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{i=1}^{D} \boldsymbol{\mu}_i \sum_{j=1}^{D} \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ij}\, g_{ij}^{\Delta}[\tau] \overbrace{\Phi_j(\tau; G)}^{P_1} \\
&+ \Delta_{\mathcal{X}}\Delta_{\mathcal{Y}}\Delta_T \sum_{i,j,k=1}^{D} \sum_{\tau_x, \tau_x'=1}^{L_{\mathcal{X}}} \sum_{\tau_y, \tau_y'=1}^{L_{\mathcal{Y}}} \sum_{\tau_t, \tau_t'=1}^{L_T} \boldsymbol{\alpha}_{ij}\, \boldsymbol{\alpha}_{ik}\, g_{ij}^{\Delta}[\tau]\, g_{ik}^{\Delta}[\tau'] \overbrace{\Psi_{j,k}(\tau, \tau'; G)}^{P_2} \\
&- 2 \sum_{i=1}^{D} \left(N_T^i \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{\tau_x=1}^{L_{\mathcal{X}}} \sum_{\tau_y=1}^{L_{\mathcal{Y}}} \sum_{\tau_t=1}^{L_T} \boldsymbol{\alpha}_{ij}\, g_{ij}^{\Delta}[\tau] \underbrace{\Phi_j(\tau; \widetilde{\mathcal{H}}_T^i)}_{P_3} \right),
\end{aligned}
$$

Inference Approach: ERM and Precomputations

3. *Precomputations:* in the loss, terms appear that do not depend on the set of parameters θ:
    - P_1: $\Phi_j(\tau; G)$
    - P_2: $\Phi_j(\tau; \widetilde{\mathcal{H}}_T^i)$
    - P_3: $\Psi_{j,k}(\tau, \tau'; G)$

    $\hookrightarrow \surd$ these terms can be **precomputed** at initialization and used at each step of the optimization procedure.

$\Longrightarrow$ **Gradient-based optimization:** approach efficiently computes exact gradients for each parameter

Applications to Real Data: Seismic Activity in California

Table 1: Negative Log Likelihood (NLL) values on test sets of various extracted earthquake datasets (NCEDC, nce, 2014) with several triggering (separable and non-separable) kernels. The best NLL is in **bold** and the second best is underlined.

Setting	1987 - 1989	2003 - 2014	1967 - 2003
TG + TG	2.77	1.76	0.72
TG + EXP	3.25	2.14	0.65
TG + KUM	2.98	2.66	0.57
POW + TG	2.11	**1.04**	**0.18**
POW + EXP	**1.72**	1.57	0.20
POW + KUM	2.06	1.50	0.29
NS1 [1]	3.77	2.68	0.88
NS2 [2]	3.77	2.67	0.87

[1] Function from the class of non-separable functions in (Cressie and Huang, 1999)
[2] Spatio-temporal function from the (Gneiting, 2002) class

Applications to Real Data: Burglary in Chicago

Table 2: NLL values on test sets of various extracted burglary datasets of the Chicago Crime Dataset with several triggering (separable and non-separable) kernels.

Setting	2008	2002 - 2004	2002 - 2006
TG + TG	-0.24	0.26	0.51
TG + EXP	-0.24	0.38	0.60
TG + KUM	-0.23	0.35	0.54
POW + TG	0.54	1.04	1.10
POW + EXP	1.27	1.03	1.08
POW + KUM	0.83	0.86	0.91
NS1	-0.37	-0.43	-0.28
NS2	**-0.95**	**-0.49**	**-0.31**

Comparison with State-of-the-Art Methods

Table 3: Spatial and Temporal NLL values on test sets of various extracted real-world datasets.

Dataset	Earthquake		COVID-19		Citybike	
Models	Spatial	Temporal	Spatial	Temporal	Spatial	Temporal
NSTPP [1]	0.886	-0.623	1.9	-2.25	2.38	-1.09
DeepSTPP [2]	4.92	-0.174	0.361	-1.09	**-4.94**	-1.13
DSTPP [3]	0.413	-1.1	0.35	-2.66	0.529	-2.43
Our approach	**-0.501**	**-10.021**	**-0.887**	**-6.336**	0.083	**-4.275**

[1] NSTPP from (Chen et al., 2021)
[2] DeppSTPP from (Zhou et al., 2022)
[3] DSTPP from(Yuan et al., 2023)

Contributions

1. We develop an **efficient and flexible** method for estimating STHP model parameters allowing **(1) any** (separable) parametric kernel, and **(2)** space-time **non-separable** kernels

   $\longrightarrow \surd$ our method **enhances precision and adaptibility** when dealing with complex dependencies in real data.

2. **Our numerical experiments** on simulated and real data show **flexibility**, **adaptability** to phenomenon's characteristics, and **accuracy** compared to SOTA.

   $\longrightarrow$ GitHub: `github.com/EmiliaSiv/Flexible-Parametric-Inference-for-Space-Time-Hawkes-Processes`

Perspectives

Perspectives: Geostatistics

- Gradually **relax** some hypotheses:
  - Assumption 2 (*stationary hypothesis*): locally stationary processes
  - Assumption 4 (*border hypothesis*) **less restrictive**: $c(h) \searrow 0$
  - Assumption 5 (*regularity hypothesis*): other smoothing techniques

  $\longrightarrow \surd$ **extend** our main results to a **more general framework**

- **Irregular grid**: define different sets of neighbors, difficulties when controlling the spectrum of the covariance matrix

  $\longrightarrow \surd$ cover **more** real-world situations

Perspectives: Hawkes Processes

- Non-constant baseline and irregular discretization grid

  $\longrightarrow \surd$ **improve accuracy**, based on additional information on phenomenon's characteristics

- **Marked** spatio-temporal Hawkes processes:

  $$\lambda_i(x, y, t, M|\mathcal{H}_t) = \boldsymbol{\mu}_i + \sum_{j=1}^{D} \sum_{u_n^j \in \mathcal{H}_t^j} \boldsymbol{\alpha}_{ij}\, g_{ij}(x - x_n^j, y - y_n^j, t - t_n^j, M - M_n^j),$$

  where M_n^j is the mark of the event

  $\longrightarrow \surd$ additional important features (magnitude of an earthquake, type of crime, etc)

- **Non-separability** in marked processes

  $\longrightarrow \surd$ accounting for **space-time and marks interactions**

- **Python library** with MIND team (Inria)

Publications and Presentations

Work in progress: Hydrogeology and Spatial Analysis, with Juan Guzmán

Publications:

- E. Siviero, E. Chautru, & S. Clémençon (2023). A Statistical Learning View of Simple Kriging. TEST, 33(1), 271-296.
- E. Siviero, G. Staerman, S. Clémençon, & T. Moreau (2024). Flexible Parametric Inference for Space-Time Hawkes Processes. ArXiv preprint arXiv:2406.06849 (Submitted).

Presentations:

- CAp 2022 (poster), COMPSTAT 2022 (oral)
- MIND team Seminar 2023 (oral), COMPSTAT 2024 (oral)

GitHub:

- `github.com/EmiliaSiv/Simple-Kriging-Code`
- `github.com/EmiliaSiv/Flexible-Parametric-Inference-for-Space-Time-Hawkes-Processes`

Thanks for your attention !

References

(2014). Northern California Earthquake Data Center, NCEDC Dataset. *UC Berkeley Seismological Laboratory*.

Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32.

Bercu, B., Delyon, B., and Rio, E. (2015). *Concentration Inequalities for Sums and Martingales*. Springer, Cham.

Brockwell, P. and Davis, R. (1987). *Time Series: Theory and Methods*. Springer, New York, NY.

Chen, R. T. Q., Amos, B., and Nickel, M. (2021). Neural spatio-temporal point processes.

Cressie, N. (1993). *Statistics for Spatial Data*, pages 1–26. John Wiley and Sons, Ltd, New York, NY.

Cressie, N. and Hawkins, D. (1980). Robust estimation of the variogram. *Mathematical Geology*, 12:115–125.

Cressie, N. and Huang, H.-C. (1999). Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions. *Journal of the American Statistical Association*, 94(448):1330–1340.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer.

Genton, M. (1998). Highly robust variogram estimation. *Mathematical Geosciences*, 30(2):213–221.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600.

Gratton, Y. (2002). Le krigeage: la méthode optimale d'interpolation spatiale. *Les articles de l'Institut d'Analyse Géographique*, 1(4).

Hall, P. and Patil, P. (1994). Properties of Nonparametric Estimators of Autocovariance for Stationary Random Fields. *Probability Theory and Related Fields*, 99(3):399–424.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 33(3):438–443.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503.

Ilhan, F. and Kozat, S. S. (2020). Modeling of spatio-temporal Hawkes processes with randomized kernels. *IEEE Transactions on Signal Processing*, 68:4946–4958.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. *arXiv preprint arXiv:1807.02582*.

Krige, D. G. (1951). A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Matheron, G. (1962). *Traité de Géostatistique Appliquée. Tome 1*. Number 14 in Mémoires du BRGM. Tecnip, Paris.

Mingoti, S. A. and Rosa, G. (2008). A note on robust and non-robust variogram estimators. *Rem: Revista Escola de Minas*, 61:87 – 95.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

Musmeci, F. and Vere-Jones, D. (1992). A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, 44:1–11.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50:379–402.

Pardo-Igúzquiza, E. and Olea, R. A. (2012). Varboot: A spatial bootstrap program for semivariogram uncertainty assessment. *Computers & Geosciences*, 41:188–198.

Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a poisson intensity on the real line. *Electronic journal of statistics*, 4:172–238.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1–41.

Shchur, O., Gao, N., Biloš, M., and Günnemann, S. (2020). Fast and flexible temporal point processes with triangular maps. *Advances in neural information processing systems*, 33:73–84.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. (2021). Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*.

Staerman, G., Allain, C., Gramfort, A., and Moreau, T. (2023). Fadin: Fast discretized inference for Hawkes processes with general parametric kernels. In *International Conference on Machine Learning*, pages 32575–32597. PMLR.

Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. Springer-Verlag, New York. Some theory for Kriging.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(1):1–45.

Wang, L. and Ma, T. (2020). Tail Bounds for Sum of Gamma Variables and Related Inferences. *Communications in Statistics - Theory and Methods*, 0(0):1–10.

Wang, W., Tuo, R., and Jeff Wu, C. (2020). On prediction properties of Kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930.

Wedin, P.-Å. (1973). Perturbation Theory for Pseudo-Inverses. *BIT Numerical Mathematics*, 13(2):217–232.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results*, volume 131. Springer.

Yuan, Y., Ding, J., Shao, C., Jin, D., and Li, Y. (2023). Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3173–3184, New York, NY, USA. Association for Computing Machinery.

Zhou, Z., Yang, X., Rossi, R., Zhao, H., and Yu, R. (2022). Neural point process for learning spatiotemporal event dynamics. In Firoozi, R., Mehr, N., Yel, E., Antonova, R., Bohg, J., Schwager, M., and Kochenderfer, M., editors, *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 777–789. PMLR.

Zhu, S., Li, S., Peng, Z., and Xie, Y. (2021). Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5391–5402.

Zhu, S. and Xie, Y. (2022). Spatiotemporal-textual point processes for crime linkage detection. *The Annals of Applied Statistics*, 16(2):1151–1170.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380.