2nd year master's internship (with the intention of continuing to a PhD)

Estimation of spatio-temporal models by Graph Neural Networks

Research units :

- Geostatistics team, Centre for Geosciences, Mines Paris, PSL
- MIA Paris Saclay, AgroParisTech/INRAE, Université Paris-Saclay

Supervisors : Lucia Clarotto (AgroParisTech), Mike Pereira and Thomas Romary (Mines Paris)
Duration : 6 months (starting spring 2025) + possible PhD afterwords
Profile : The candidate should be a 2nd year master or last year engineer student in Statistics/Machine Learning, with courses on deep learning or spatial statistics. Scientific programming skills in Python are required.
Salary : ∼ 750€/month for the internship
Location : Centre for Geosciences, Mines Paris PSL (Fontainebleau) and/or MIA Paris Saclay, Campus Agro Paris-Saclay (Palaiseau)

The candidate will have an office in both locations, and benefit from the work environment of both laboratories, with many PhD students & postdocs working on statistical modeling and machine learning for geosciences and life sciences.

Context

In geostatistics, Gaussian processes are commonly used to model spatial and spatio-temporal data, as they allow simple prediction of the variable of interest at unmeasured unmeasured sites, while quantifying the prediction uncertainty. In this context, data are considered to be derived from a particular realization of a Gaussian random field whose covariance function must be estimated from the data. A classic approach to inference consists in choosing a parameterized family of covariance functions, then selecting the parameters that maximize the likelihood associated with the data. In practice, this approach often represents a bottleneck, as the evaluation of the likelihood function alone can quickly become become costly from a computational point of view when the quantity of data becomes large, particularly in the spatio-temporal domain. It is therefore desirable to have methods for deducing the parameters of a covariance model without using the likelihood function. Recently, several methods using neural networks (notably CNN and GNN) have been have been proposed to address this problem, using the so-called "amortized inference" [7]. Two main approaches can be mentioned.

The first aims to train a neural network capable of identifying the parameters of a covariance function, from a realization of a Gaussian random field with this covariance [2, 3, 6, 4]. The second aims rather to obtain an approximation of the likelihood as a function of parameters and observations, thus providing an easy-to-calculate proxy [5]. The neural likelihood surface can then be maximized for a fixed set of observations in order to obtain an estimator of the model parameters associated with these observations.

Goals

A first-year master's internship in 2024 focused on coding an architecture similar to that proposed in [4] under Pytorch, based on complex graph-based convolutional networks (GNNs). The second-year master's internship aims to train and validate the architecture proposed during the previous internship (on simulated datasets) and to develop an architecture adapted to the spatio-temporal context. One of the challenges of this generalization is to define a new spatio-temporal neighborhood structure in the GNN and estimate a larger number of parameters, which has never been implemented before. The next step will be to compare the proposed strategy with state-of-the-art spatio-temporal inference methods on simulated datasets. Next, the approach will be adapted to models derived from Stochastic Partial Differential Equations [1]. An application to a real data set of solar radiation measurements will conclude the internship.

This internship may be continued as part of a PhD, financed by the Geolearning chair, offering continuity and an opportunity for further research. Topics that could be addressed in the thesis are the generalization of the inference method for non-stationary models, or the extension of inference to the Bayesian framework, using variational approximation methods.

How to candidate ?

To apply or if you have any questions, please contact Lucia Clarotto (lucia.clarotto@agroparistech.fr), Mike Pereira (mike.pereira@minesparis.psl.eu) and Thomas Romary (thomas.romary@minesparis.psl.eu). Please send us your CV and cover letter.

Références

[1] Lucia Clarotto, Denis Allard, Thomas Romary, and Nicolas Desassis. The spde approach for spatio-temporal datasets with advection and diffusion. *Spatial Statistics*, 62 :100847, 2024.

[2] Florian Gerber and Douglas Nychka. Fast covariance parameter estimation of spatial gaussian process models using neural networks. *Stat*, 10(1) :e382, 2021.

[3] Amanda Lenzi, Julie Bessac, Johann Rudi, and Michael L Stein. Neural networks for parameter estimation in intractable models. *Computational Statistics & Data Analysis*, 185 :107762, 2023.

[4] Matthew Sainsbury-Dale, Jordan Richards, Andrew Zammit-Mangion, and Raphaël Huser. Neural bayes estimators for irregular spatial data using graph neural networks. *arXiv preprint arXiv :2310.02600*, 2023.

[5] Julia Walchessen, Amanda Lenzi, and Mikael Kuusela. Neural likelihood surfaces for spatial processes with computationally intensive or intractable likelihoods. *arXiv preprint arXiv :2305.04634*, 2023.

[6] Christopher K Wikle and Andrew Zammit-Mangion. Statistical deep learning for spatial and spatio-temporal data. *arXiv preprint arXiv :2206.02218*, 2022.

[7] Andrew Zammit-Mangion, Matthew Sainsbury-Dale, and Raphaël Huser. Neural methods for amortized inference. *arXiv preprint arXiv :2404.12484*, 2024.