Boletín de Estadística e Investigación Operativa
Vol. 35, No. 3, Noviembre 2019, pp. 1-15

Estadística

Modelling Preferential Sampling in time

Andreia Monteiro
University of Minho & CIDMA
✉ andreiaforte50@gmail.com

Raquel Menezes
University of Minho & CBMA
✉ rmenezes@math.uminho.pt

Maria Eduarda Silva
Faculty of Economics
University of Porto & CIDMA
✉ mesilva@fep.up.pt

Abstract

Preferential sampling in time occurs when there is stochastic dependence between the process being modeled and the times of the observations. Examples occur in fisheries if the data are observed when the resource is available, in sensoring when sensors keep only some records in order to save memory and in clinical studies, when a worse clinical condition leads to more frequent observations of the patient. In all such situations the observation times are informative on the underlying process. To make inference in time series observed under Preferential Sampling we propose, in this work, a numerical method based on a Laplace approach to optimize the likelihood and to approximate the underlying process we adopt a technique based on stochastic partial differential equation. Numerical studies with simulated and real data sets are performed to illustrate the benefits of the proposed approach.

Keywords: Preferential sampling, time series, continuous time autoregressive process, SPDE, Laplace.
AMS Subject classifications: 62M10.

1. Introduction

Real time series sometimes exhibit various types of "irregularities": missing observations, observations collected not regularly over time for practical reasons, observation times driven by the series itself, or outlying observations. However, the vast majority of methods of time series analysis are designed for regular time series only. There are few methods available in the literature for the analysis of irregularly spaced series. Some authors, such as [8], [9], [1] and [3] have suggested an embedding into continuous diffusion processes, with the aim of using the well established tools for univariate autoregressive moving average (ARMA) processes. A particular case of irregularly spaced time series is that in which the sampling procedure over time depends also on the observed values. Typically, in medical studies, a patient is observed most frequently when he presents a worse clinical condition. Otherwise, if the patient shows a better medical condition, the follow-up time process will not occur so often. As such, observed measurements can bring information about diseases condition, but also the frequency at which these measurements are made can provide information about the health status of the patients. In such situations, there is stochastic dependence between the process being modeled and the times of the observations.

In this context, following [5] in the context of spatial statistics, [13] introduce the concept of Preferential Sampling in the temporal dimension as a formal definition for the dependence between the process generating the times of the observations and the data values. The authors propose a model-based approach to make inference and prediction also able to deal with irregularly spaced time series, under Preferential Sampling or not. They consider a Monte Carlo approach for maximum likelihood estimation of the model. The suggested framework considers the observed time points as the realization of a time point process stochastically dependent on an underlying latent process (e.g. an individual health indicator, when subjected to regular monitoring).

The convergence of the estimation algorithm proposed by [13] is very slow and the running time becomes burdensome for longer time series and a large number of Monte Carlo samples. Besides these, the algorithm is sensitive to starting values and the large variability between likelihood values in each Monte Carlo iteration makes the likelihood difficult to optimize. In view of the aforementioned issues, in this work we suggest an alternative numerical method that uses the Laplace approximation for the marginal likelihood and we adopt a technique based on stochastic partial differential equation (SPDE) to approximate the underlying process.

The above mentioned numerical techniques based on the Laplace approximation and SPDE have become usual when dealing with complex models and large data sets, [6] and [4]. These changes will hopefully result in a large increase in the stability of our parameter estimates, particularly in comparison

with previous method based on Monte Carlo approximation.

The paper is organized as follows. In Section 2, based on [13], we describe the model for preferential sampling in time dimension. In Section 3 we present the methodological details of the proposed numerical method based on a Laplace approach to optimize the likelihood and the approximation of the underlying process using SPDE. In Section 4, using numerical studies, we document the performance of suggested approach. We further compare the estimates with those obtained under traditional approach for irregularly spaced data and with those obtained under INLA Bayesian approach. In Section 5 we show the application of the previously described methodology to a real data set related to monitoring the level of a biomedical marker, after a cancer patient undergoes a bone marrow transplant. Section 6 is devoted to make some concluding remarks.

2. A model for Preferential Sampling in time

Consider an unobserved stochastic process in time $S(t)$, represented by a Continuous Time Autoregressive model of order 1, CAR(1), that satisfies the differential equation

$$dS(t) + \alpha_0 S(t)dt = dW(t)$$

where, α_0 is the autoregressive coefficient and $W(t)$ is a Wiener process with variance parameter σ_w^2.

$S(\cdot)$ is a stationary Gaussian process with $E[S(t)] = 0$. Now admit that $S(t)$ is observed at times t_i, $i = 1, \cdots, n$, yielding a data set (t_i, y_i), where the corresponding $Y_i = Y(t_i)$ is the noisy version of $S(t_i)$. Following [13], a model for the data takes the form:

$$Y(t_i) = \mu + S(t_i) + N(0, \tau^2) \tag{2.1}$$

Admitting that the sampling times are stochastic the joint distribution of S, $T = (t_1, \ldots, t_n)$ and $Y = (Y_1, \ldots, Y_n)$, $[S, T, Y]$[1] must be specified. Considering the stochastic dependence between S and T, the model to deal with Preferential Sampling is defined through $[S, T, Y]$ written as:

$$[S][T\,|S][Y\,|T, S\,] \tag{2.2}$$

where, conditional on S and T, Y is a set of mutually independent Gaussian variates with τ^2 being the measurement error variance and conditional on S, T is an inhomogeneous Poisson process with intensity

$$\lambda\,(t) = \exp\{a + \beta S\,(t)\} \tag{2.3}$$

[1] $[\cdot]$ means "the distribution of"

where β is the parameter that controls the degree of preferentiality, for example, when $\beta > 0$ the sample times are concentrated, predominantly, near the maximum of the observed values and when $\beta = 0$ it corresponds to the situation of an homogeneous, non-preferential, sampling.

Although the construction of this model is driven by a Preferential Sampling context, it may be applied to model any type of irregularly spaced time series.

One of its advantages is to make predictions at unobserved time points.

The predicted value of $S(\cdot)$ at an unsampled time point $t_{n_i} < t_0 < t_{n_j}$, $S(t_0|T)$, is given by $S(t_0|T) = \mathrm{E}\big[S(t_0)|Y(T)\big]$. Considering that the process CAR(1) is Markovian, [2, p.358] shows that the conditional mean of $S(t_0)$ given $Y(T)$ is

$$\begin{aligned} S(t_0|T) &= \mathrm{E}\big[S(t_0)|Y(T)\big] \\ &= \exp\left(-\alpha_0(t_0 - t_{n_i})\right) Y(T) + \mu\left(1 - \exp\left(-\alpha_0(t_0 - t_{n_i})\right)\right) \end{aligned} \tag{2.4}$$

The variance of the prediction is

$$\sigma^2(t_0) = Var\big[S(t_0)|Y(T)\big] = \frac{\sigma_w^2}{2\alpha_0}\left(1 - \exp\left(-2\alpha_0(t_0 - t_{n_i})\right)\right) \tag{2.5}$$

3. Laplace approach - Methodological details

To obtain the parameters of the model we use maximum likelihood estimation. For the shared latent process model, the likelihood function for data T and Y can be expressed as

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [T, Y, S]dS = \int_S [S][T, Y|S]dS = \int_S [S][T|S][Y|T, S]dS \tag{3.1}$$

where $\boldsymbol{\theta} = (\mu, \sigma_w, \alpha_0, \tau, \beta)$ represents the set comprising the model parameters.

Previously, in [13], a partition of S into $S = \{S_0, S_1\}$ was considered, where S_0 denotes the values of S at each of n times $t_i \in T$, and S_1 are the values of S at the remaining $(N - n)$. The integral in (3.1) has been rewritten as

$$\begin{aligned} L(\theta) &= \int_S [S_1|S_0][S_0][T|S][Y|S_0]\frac{[S|Y]}{[S_1|S_0][S_0|Y]}dS \\ &= \int_S [T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]dS \\ &= E_{S|Y}\left[[T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right] \end{aligned} \tag{3.2}$$

Taking into account that the conditional expectation in (3.2) can be approximated by Monte Carlo, Maximum Likelihood Estimates (MCMLE's) are obtained by maximizing the Monte Carlo likelihood

$$L_{MC}(\boldsymbol{\theta}) = m^{-1} \sum_{j=1}^{m} [T|S_j] \frac{[Y|S_{0j}]}{[S_{0j}|Y]} [S_{0j}] \tag{3.3}$$

where S_j are assumed as realizations of the distribution of S conditional on Y. S_{0j} denotes the values of S_j restricted to the n observed time points and m, is the total number of Monte Carlo replicates.

An alternative method (henceforth LAP) to the Monte Carlo simulation proposed by [13] is to utilize Automatic Differentiation of a Laplace Approximation to the marginal likelihood, to evaluate directly equation (3.1).

If we assume that the likelihood function $L(\boldsymbol{\theta})$ can be written as

$$L(\boldsymbol{\theta}) = \int_S exp(-f(S, \boldsymbol{\theta}))dS \tag{3.4}$$

where $f(S, \boldsymbol{\theta})$ denotes the negative joint log-likelihood of the data, $\boldsymbol{\theta}$ is the vector of parameters (fixed effects) and S the random effects. The Laplace approximation for $L(\boldsymbol{\theta})$ is

$$L^*(\boldsymbol{\theta}) = (2\pi)^{N/2} det(H(\boldsymbol{\theta}))^{-1/2} exp(-f(\widehat{S}(\boldsymbol{\theta}), \boldsymbol{\theta}))$$

where

$$\widehat{S}(\boldsymbol{\theta}) = arg_S min f(S, \boldsymbol{\theta}) \tag{3.5}$$

and $H(\boldsymbol{\theta})$ is the Hessian of f with respect to S evaluated at $\widehat{S}(\boldsymbol{\theta})$,

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial S^2} f(S, \boldsymbol{\theta})|_{S=\widehat{S}(\boldsymbol{\theta})}$$

The estimate of $\boldsymbol{\theta}$ minimizes the negative of the logarithm of the Laplace approximation,

$$-\log L^*(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi) + \frac{1}{2}\log det(H(\boldsymbol{\theta})) + f(\widehat{S}(\boldsymbol{\theta}), \boldsymbol{\theta}) \tag{3.6}$$

This objective function and its derivatives acquired by using automatic differentiation, are required to apply standard nonlinear optimization algorithms (e.g., nlmimb) to optimize the objective function and obtain the estimate for $\boldsymbol{\theta}$. To speed up significantly the optimization of the likelihood function we use programming language C++. Using the R package TMB, short for Template Model Builder, [10], we define the joint log-likelihood of the data and (i.e. conditional on) the random effects as a C++ template function. The Laplace

approximation of the marginal likelihood is then evaluated and maximized, where the random effects are automatically integrated out. This approximation, and its derivatives, are obtained using automatic differentiation (up to order three) of the joint likelihood. In the case of the Preferential Sampling model, we have to define the joint negative log-likelihood as

$$f(S, \boldsymbol{\theta}) = -log([S][T|S][Y|S, T])$$

and integrate out the latent field S to evaluate approximately (3.1).

Uncertainty of the estimate $\widehat{\boldsymbol{\theta}}$ or of any differentiable function of the estimate $\zeta(\widehat{\boldsymbol{\theta}})$ that the user specifies, is obtained by the δ-method:

$$Var(\zeta(\widehat{\boldsymbol{\theta}})) = -\left\{ \frac{\partial \zeta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \left[\frac{\partial^2 (logL^*(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]^{-1} \frac{\partial \zeta(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{3.7}$$

These uncertainty calculations also require derivatives of (3.6). However, derivatives are not straight-forward to obtain using automatic differentiation in this context.

To increase computational efficiency, we approximate $[S]$ in (3.1) using a technique based on stochastic partial differential equations (SPDE). This allows to create a temporal mesh and corresponding components of the sparse precision matrix of a Gaussian Markov Random Field (GMRF) in time-dimension. Following [12], we use the representation of a Gaussian process with Matérn covariance structure as the solution of the following SPDE,

$$\left(\phi^{-2} - \Delta\right)^{\alpha/2} (\omega S(t)) = \epsilon(t),\ t \in \mathbb{R}^+, \tag{3.8}$$

where $\epsilon(t)$ is Gaussian white noise, Δ is the Laplacian and ϕ is the range parameter of the Matérn covariance function $\gamma(u)$ in its standard parametrization,

$$\gamma(u) = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} (u/\phi)^{\nu} K_{\nu} (u/\phi) : u \geq 0$$

where K_ν is the modified Bessel function of second kind and order $\nu > 0$ and σ^2 is the marginal variance. The integer value of ν determines the mean square differentiability of the underlying process, which matters for predictions made using such a model. However, ν is usually fixed since it is poorly identified in typically applications. The remaining parameters in (3.8) are $\alpha = \nu + 1/2$, from this we can identify the exponential covariance with $\nu = 1/2$, and ω that controls the variance,

$$\omega^2 = \frac{\Gamma(1/2)}{\Gamma(1)(4\pi)^{1/2}\phi^{-1}\sigma^2}$$

We approximate the process S by $\widetilde{S}$, where

$$\widetilde{S}(t) = \sum_{k=1}^{n} \psi_k(t) W_k,\ t \in \mathbb{R}^+$$

where $\psi_k(\cdot)$ are piecewise linear basis functions at a set of time knots and $W = W_1, ..., W_n$ is a zero-mean multivariate Gaussian variate with covariance matrix Q^{-1}. The construction is done by projecting the SPDE onto the basis representation in what is essentially a Finite Element method. For $\alpha = 1$ the required form of Q is

$$Q = \omega^2(\phi^{-2}C + G_2)$$

where C and G_2 are sparse matrices whose explicit expressions can be found in [12].

4. Numerical Studies

We now intend to proceed with the assessment of the LAP method, comparing the results of its parameter estimates with those of the traditional Kalman filter approach to irregularly spaced data (cts package [17]) and with those obtained from INLA, proposed by [14] and available in the R-INLA software package.

To simulate a time series under Preferential Sampling we use the thinning algorithm, [11]. We first generate a realization of S from model (2.1) with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$, discretized in 800 equally spaced time points. These values correspond to $Var[S(\cdot)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ and $\phi = \frac{1}{\alpha_0} = 5$, being the latter related to the lag beyond which there is no correlation for practical purposes. To generate Y from model (2.1), we consider $\mu = 0$ and $\tau = 0.1$, conducting three separate sampling procedures over the realization of S:

- Preferential Sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (2.3) and $\beta = 2$, which corresponds to the situation when the sampling times are concentrated, predominantly, near the maximum of the observed values;
- irregular sampling: we obtain $n = 70$ sampling times T from (2.3) and with $\beta = 0$, illustrating the situation without Preferential Sampling;
- Preferential Sampling: conditional on the values of S, we obtain $n = 70$ sampling times T following an inhomogeneous Poisson process with intensity function defined in (2.3) and $\beta = -2$, which corresponds to the situation when the sampling times are concentrated, predominantly, near the minima of the observed values.

The parameters μ, σ, ϕ, τ and β are the target of estimation. We compare the parameter estimates, obtained from a total of 500 independent samples, for three alternative methods:

- LAP, implemented through C++, via package TMB;
- LAP, implemented via package INLA;
- Kalman filter approach, implemented via package cts.

INLA relies on Laplace approximation methods to numerically approximate posterior distributions. This method performs Gaussian approximations of the parameters by inferring their mode. Although posterior distributions do not necessarily have to be Gaussian, INLA relies on the fact that for most real problems and data sets, the conditional posterior of the latent field looks "almost" Gaussian, [14]. This is clearly assisted by the, non-negligible, impact of the Gaussian priors on the posteriors.

In our study, the prior distributions will be the default non informative and for the SPDE model, for σ and ϕ, we consider the Penalized Complexity prior, PC-prior, as derived in [7].

4.1. Results of parameter estimation

The results of the mean and standard errors of each parameter, obtained from a total of 500 independent samples are summarized in Table 1. In this numerical study we consider as initial values ($\boldsymbol{\theta_0}$) the "true" values. In Section 4.2 a study will be conducted to evaluate the sensitivity to initial values in the estimation of the parameters.

In Figures 3, 4 and 5 (see the appendix) we have the corresponding boxplots for the preferential ($\beta = 2$), non-preferential ($\beta = 0$) and preferential ($\beta = -2$) simulated data sets, respectively, with true parameter values marked as red line. (PS corresponds to LAP method proposed in Section 3).

By analysing Table 1 and Figures 3, 4 and 5, we conclude that under Preferential Sampling, LAP, via TMB offers more accurate estimates than LAP via INLA, except in the case of σ. Comparing with the traditional Kalman filter, LAP showed considerable success mainly for μ, σ and ϕ. The parameter β seems to be underestimated using LAP and R-INLA in the case of $\beta = 2$ and overestimated for $\beta = -2$.

4.2. Sensitivity Analysis

To investigate the sensitivity of the estimation procedures to initial values, we estimate the parameters considering initial values $\boldsymbol{\theta_0}$: (i) the "true" values (ii) the parameters estimated by traditional Kalman filter approach. An estimate for the initial value of β, given a sample data set Y, can be obtained as follows. Suppose that $Y = \{(t_i, y_i) : i = 1, \ldots, n\}$, where y_i denotes the measured value and t_i is

		PS Data $\beta = 2$		
	True	LAP	INLA	CTS
$\widehat{\mu}$	**0**	**0.167** (0.483)	**0.600** (0.423)	**1.929** (0.480)
$\widehat{\sigma}$	**1.581**	**1.471** (0.355)	**1.550** (0.333)	**0.906** (0.157)
$\widehat{\phi}$	**5**	**5.873** (2.531)	**7.486** (3.097)	**2.339** (1.462)
$\widehat{\tau}$	**0.1**	**0.166** (0.099)	**0.151** (0.352)	**0.176** (0.090)
$\widehat{\beta}$	**2 ; 0**	**1.359** (0.258)	**1.076** (0.204)	

		Not PS Data $\beta = 0$		
	True	LAP	INLA	CTS
$\widehat{\mu}$	**0**	**-0.010** (0.386)	**-0.013** (0.381)	**0.003** (0.362)
$\widehat{\sigma}$	**1.581**	**1.496** (0.201)	**1.580** (0.194)	**1.529** (0.207)
$\widehat{\phi}$	**5**	**5.061** (1.605)	**5.536** (1.533)	**5.065** (1.672)
$\widehat{\tau}$	**0.1**	**0.211** (0.144)	**0.045** (0.110)	**0.233** (0.135)
$\widehat{\beta}$	**0**	**-0.005** (0.098)	**-0.004** (0.077)	

		PS Data $\beta = -2$		
	True	LAP	INLA	CTS
$\widehat{\mu}$	**0**	**-0.174** (0.384)	**-1.768** (1.880)	**-1.919** (0.480)
$\widehat{\sigma}$	**1.581**	**1.426**(0.279)	**1.699**(0.455)	**0.913** (0.153)
$\widehat{\phi}$	**5**	**5.223** (1.695)	**6.443**(1.748)	**2.317**(1.228)
$\widehat{\tau}$	**0.1**	**0.155**(0.101)	**0.080** (0.113)	**0.170**(0.094)
$\widehat{\beta}$	**-2**	**-1.344** (0.241)	**-0.530**(0.786)	

Table 1: Maximum likelihood estimates, under LAP (implemented via TMB package), LAP (implemented via INLA package) and by Kalman filter approach (implemented via cts package), mean (standard errors) obtained from a total of 500 independent samples.

the corresponding time of the observation. A preliminary β_0 can be obtained through a simple algorithm such as: first, use a kernel-type intensity estimator of the locations to derive $\widehat{\lambda}(t)$; and, then, choose β_0 such that $\log \widehat{\lambda}(t) \simeq const + \beta_0 Y(t)$.

		Preferential Data set		Not Preferential Data set	
	True	True $\boldsymbol{\theta_0}$	$\boldsymbol{\theta_0}$ from CTS	True $\boldsymbol{\theta_0}$	$\boldsymbol{\theta_0}$ from CTS
$\widehat{\mu}$	**0**	**0.247** (0.417)	**0.247** (0.417)	**-0.030** (0.388)	**-0.030** (0.388)
$\widehat{\sigma}$	**1.581**	**1.412** (0.255)	**1.413** (0.255)	**1.500** (0.207)	**1.500** (0.207)
$\widehat{\phi}$	**5**	**5.244** (1.699)	**5.242** (1.700)	**5.167** (1.739)	**5.167** (1.739)
$\widehat{\tau}$	**0.1**	**0.164** (0.104)	**0.163** (0.106)	**0.204** (0.138)	**0.203** (0.138)
$\widehat{\beta}$	**1.5;0**	**1.175** (0.159)	**1.175** (0.159)	**0.002** (0.100)	**0.002** (0.100)

Table 2: MLE's under LAP, mean (standard errors) obtained from a total of 250 independent samples, considering as initial values ($\boldsymbol{\theta_0}$) the "true" values and other considering the parameters estimated by traditional Kalman filter.

The results of the mean and standard errors of each parameter, obtained from a total of 250 independent samples are summarized in Table 2.

The proposed method seems to be quite robust to initial values of θ in both scenarios, under preferential and not preferential sample data.

5. Application to real data

We now consider the problem of monitoring the level of one biomedical marker, platelet (PLT), after a cancer patient undergoes a bone marrow transplant. The data, composed by 54 measurements over 91 days of log(PLT) shown in Figure 1, is studied by [15] as missing data problem. These data are made available in package astsa [16] with the name of "blood".

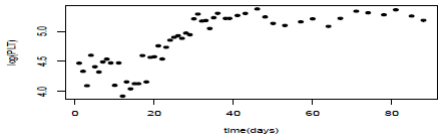

Figure 1: Measurements of biomedical marker platelet, in the logarithm scale, log(PLT).

The biomedical marker PLT was also studied by [13]. We now intend to relate the results of the two approaches (Monte Carlo and Laplace), both targeting preferential sampling issues. The estimated parameters, using LAP method together with estimated standard errors and using Monte Carlo method [13], are summarized in Table 3.

Parameter	LAP	Monte Carlo
$\widehat{\mu}$	4.99 (0.290)	4.97
$\log(\widehat{\omega})$	2.55 (0.20)	-
$\widehat{\sigma}$	0.33	0.52
$\log(\widehat{\phi})$	3.56 (0.71)	-
$\widehat{\phi}$	35.12	54.85
$\log(\widehat{\tau})$	- 2.086 (0.13)	-
$\widehat{\tau}$	0.12	0.14
$\widehat{\beta}$	-0.94 (0.32)	-1.51

Table 3: Maximum likelihood estimates under LAP and Monte Carlo method.

Comparing the above parameter estimates, we conclude that the estimated value for β, using LAP method also has negative sign but a bit lower

than Monte Carlo method. Anyway, the corresponding confidence interval for $\widehat{\beta}$ is $(-1.57; -0.30)$, confirming that β estimated from Monte Carlo and LAP approaches are in accordance. The estimates for the mean parameter, considering the two approaches, present equivalent results.

Predictions of the biomarker within the period of the observations are obtained plugging the estimated parameters in equations (2.4) and (2.5) and we obtain the predictions of the biomarker within the period of observations. Figure 2, top panel, shows the 95% prediction intervals for (log of) the biomarker, obtained from MCMLE's in (3.3), while the middle panel represents the 95% prediction intervals obtained from the MLE's from the Kalman filter approach and bottom panel represents the 95% prediction intervals obtained from the MLE's from LAP approach suggested in Section 3. In this situation the predictions obtained from LAP present lower variance than the predictions obtained from Monte Carlo approach, revealing greater precision.

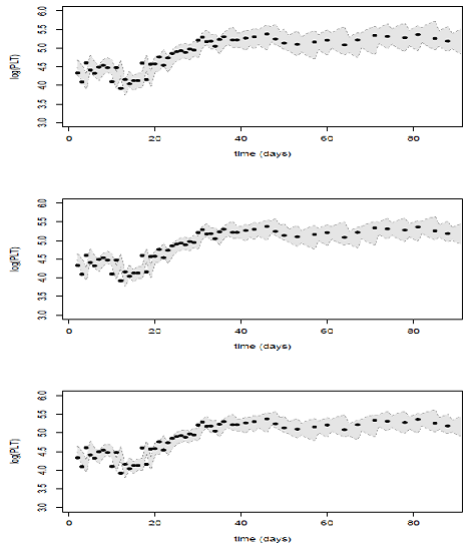

Figure 2: Prediction 95% confidence intervals using predictions acquired from MCMLE's in (3.3) (top), MLE's from the Kalman filter approach (middle) and LAP approach described in Section 3 (bottom).

6. Concluding Remarks and Future Work

We present in this work an alternative, based on a Laplace approximation, to the Monte Carlo Simulation proposed by [13]. This alternative, increases the stability of our parameter estimates and presents quite satisfactory results for estimation. We emphasize that the results for estimated parameters, in

both approaches, are quite satisfactory when compared with the traditional one that uses Kalman filter to deal with irregularly spaced time series. However, LAP method, is much more computationally efficient and runs faster, while Monte Carlo maximum likelihood estimates takes approximately 20 minutes to estimate parameters in a single simulation, LAP method takes approximately 21 seconds. Although INLA is slightly faster (16 seconds), LAP presents more accurate results and provides user high levels of flexibility, due to the direct specification of the joint likelihood.

In this work we assumed that the variable of interest is sampled in time according to a sampling design that depends on the values of the underlying process, ignoring the past of the observation processes. However, this kind of assumption of a memoryless process for the observations process having an evolution without aftereffects might be unrealistic for some real contexts, where the dependence on the past is crucial. We intend, for future investigation, to consider that the sampling design may depend on entire past history of the process, meaning all the times of the observations as well as the values of these observations.

Acknowledgments

The authors acknowledge Center for Research & Development in Mathematics and Applications of Aveiro University within project UID/MAT/04106/2019 and the project PTDC/MAT-STA/28243/2017.

References

[1] Belcher, J., Hampton, J. and Wilson, G.T. (1994). Parameterization of continuous time autoregressive models for irregularly sampled time series data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 141-155.

[2] Brockwell, P.J. and Davis, R.A. (2002). *Introduction to time series and forecasting*, Springer.

[3] Brockwell, P.J. (2009). Lévy–driven continuous–time arma processes. *Handbook of financial time series*, 457-480.

[4] Diggle, P. and Giorgi, E. (2017). Preferential sampling of exposure levels. In: *Handbook of Environmental and Ecological Statistics*, chap. 21, CRC Press.

[5] Diggle, P.J., Menezes, R. and Su, T.l. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **59**(2), 191-232.

[6] Dinsdale, D. and Salibian-Barrera, M. (2018). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*.

[7] Fuglstad, G.A., Simpson, D., Lindgren, F. and Rue, H. (2018). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 1-8.

[8] Jones, R.H. (1981). Fitting a continuous time autoregression to discrete data. In: *Applied time series analysis II*, 651-682, Elsevier.

[9] Jones, R.H. (1985). Time series analysis with unequally spaced data. *Handbook of statistics*, **5**, 157-177.

[10] Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. and Bell, B.M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**(5), 1-21. Doi: 10.18637/jss.v070.i05

[11] Lewis, P.A.W. and Shedler, G.S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, **26**(3), 403-413.

[12] Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **73**(4), 423-498.

[13] Monteiro, A., Menezes, R. and Silva, M.E. (2018). Modelling irregularly spaced time series under preferentail sampling. *Revstat Statistical Journal* (accepted).

[14] Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society*, Series B (statistical methodology) **71**(2), 319-392.

[15] Shumway, R.H. and Stoffer, D.S. (2017). *Time Series Analysis and Its Applications: With R Examples*, 4th ed. edn, Springer Texts in Statistics, Springer New York.

[16] Stoffer, D. (2017). astsa: Applied Statistical Time Series Analysis. `http://CRAN.R-project.org/package=astsa`

[17] Wang, Z. (2013). cts: An R package for continuous time autoregressive models via kalman filter. *Journal of Statistical Software*, **53**(5), 1-19. `http://www.jstatsoft.org/v53/i05/`

About the authors

Andreia Alves Forte de Oliveira Monteiro has a Doctoral degree in Applied Mathematics, under the joint doctoral programme with the Universities of Minho, Aveiro and Porto. She is Invited Assistant Professor in the University of Minho, Instituto Politécnico do Cávado e Ave and Escola Superior de Tecnologia e Gestão de Águeda and member of the Center for Research & Development in Mathematics and Applications (CIDMA). Her research interest area is spatial and temporal data modelling.

Raquel Menezes is Assistant Professor at the Department of Mathematics of Minho University, and collaborator of the Research Centre of Statistics and its Applications of Lisbon University (CEAUL). Her main research interest is Spatial Statistics, primarily Geostatistics, also working on topics as non-parametric curves estimation and analysis of Spatio-Temporal data. She has supervised 5 PhD students and more than 15 master students in Statistics. She has been acting as principal investigator of projects funded by the Portuguese National Funding Agency for Science, Research and Technology, and being member of other funded projects. In last 10 years, she published more than 22 papers in international peer-reviewed journals.

Maria Eduarda Silva is Associate Professor with Habilitation at School of Economics and Management and member of the Center for Research & Development in Mathematics and Applications (CIDMA). Her main research interest is Time Series analysis and its applications but is also working on analysis of Spatio-Temporal data. She has supervised 10 PhD students and more than 20 master students in Statistics and Applied Mathematics. She has been acting as investigator of projects funded by the Portuguese National Funding Agency for Science, Research and Technology. In last 10 years, she published more than 30 peer-reviewed.

Appendix

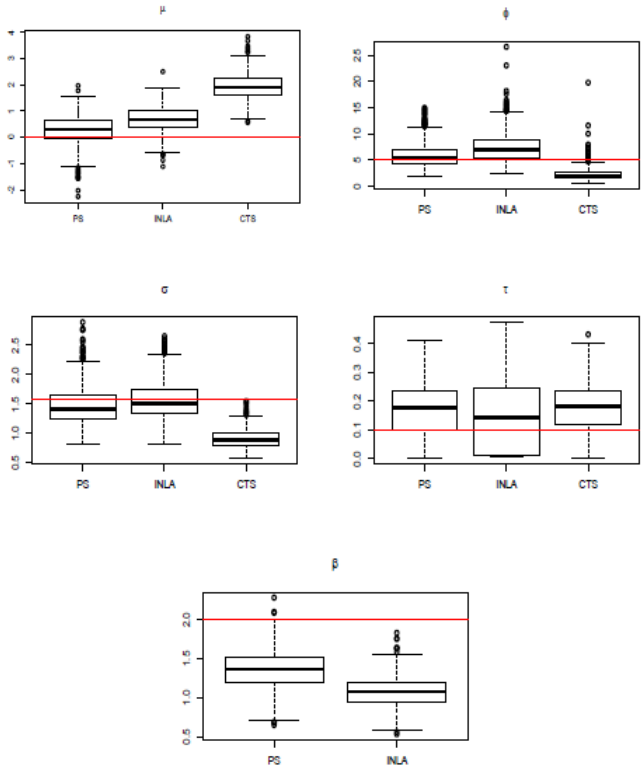

Figure 3: Boxplots for models parameters estimated over 500 preferentially sample simulated data sets, $\boldsymbol{\beta} = \mathbf{2}$, with true parameter values marked as red line.

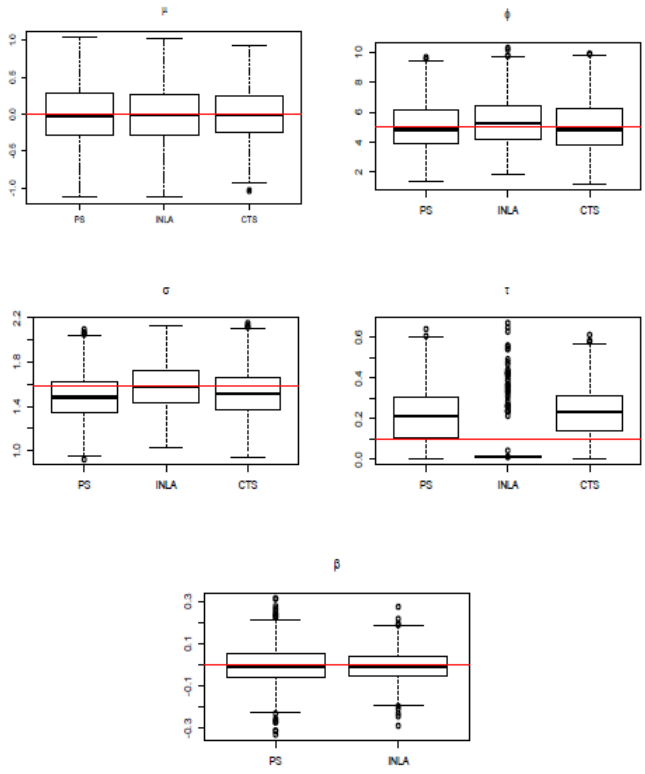

Figure 4: Boxplots for models parameters estimated over 500 non-preferentially sample simulated data sets, $\boldsymbol{\beta} = \mathbf{0}$, with true parameter values marked as red line.

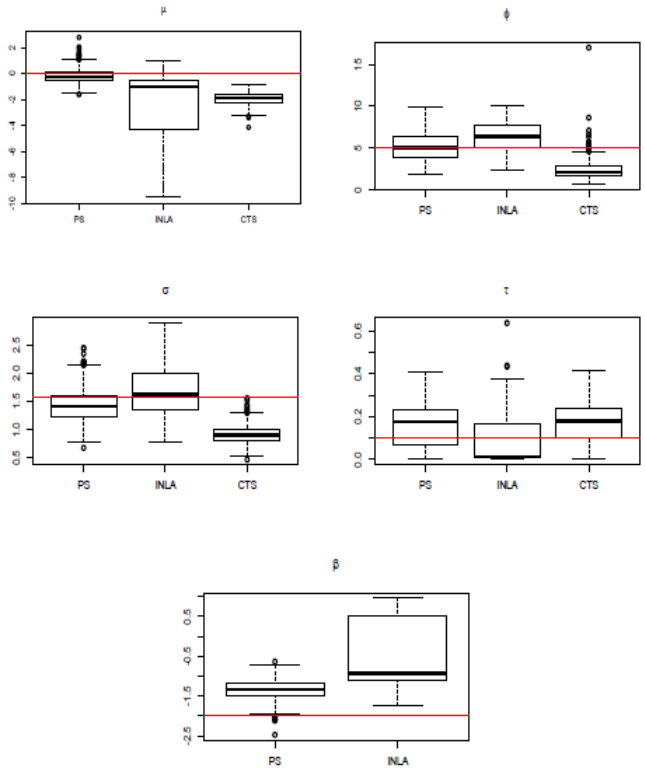

Figure 5: Boxplots for models parameters estimated over 500 preferentially sample simulated data sets, $\boldsymbol{\beta} = \mathbf{-2}$, with true parameter values marked as red line.