REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 661–672

MODELLING IRREGULARLY SPACED TIME SERIES UNDER PREFERENTIAL SAMPLING

Authors: Andreia Monteiro
– University of Minho & CIDMA, Portugal
andreiaforte50@gmail.com

Raquel Menezes
– University of Minho, Portugal
rmenezes@math.uminho.pt

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

Received: April 2018 Revised: September 2018 Accepted: October 2018

Abstract:

- Irregularly spaced time series are commonly encountered in the analysis of time series. A particular case is that in which the collection procedure over time depends also on the observed values. In such situations, there is stochastic dependence between the process being modeled and the times at which the observations are made. Ignoring this dependence can lead to biased estimates and misleading inferences. In this paper, we introduce the concept of preferential sampling in the temporal dimension and we propose a model to make inference and prediction. The methodology is illustrated using artificial data as well a real data set.

Key-Words:

- *preferential sampling; time series; continuous time autoregressive process; SPDE.*

AMS Subject Classification:

- 62M10, 62M20.

1. INTRODUCTION

Analysis of experimental data that have been observed at different points in time leads to specific problems in statistical modeling and inference. In traditional time series the main emphasis is on the case when a continuous variable is measured at discrete equispaced time points, [22]. There is an extensive body of literature on analyzing equally spaced time series data, see for example [3] and [6]. However, unevenly spaced (also called unequally or irregularly spaced) time series data naturally occurs in many scientific domains. Natural disasters such as earthquakes, floods, or volcanic eruptions typically occur at irregular time intervals. In observational astronomy, for example, measurements of properties such as the spectra of celestial objects are taken at irregularly spaced times determined by seasonal, weather conditions, and availability of observation time slots. In clinical trials (or more generally, longitudinal studies), a patient's state of health may be observed only at irregular time intervals, and different patients are usually observed at different points in time.

It must be noted that sometimes equally spaced time series are treated as irregularly spaced time series, namely time series with missing observations and multivariate data sets that consist of time series with different frequencies, even if the observations of each time series are reported at regular intervals. One of the first to treat evenly sampled gene expression time series with missing values as unevenly sampled data is [19].

There are few methods available in the literature for the analysis of irregularly spaced series. Some authors, such us [10], [12], [2] and [5] have suggested an embedding into continuous diffusion processes, with the aim of using the well established tools for univariate autoregressive moving average (ARMA) processes.

Observations with irregularly spaced sampling times are much harder to work with, partly because the established and efficient algorithms developed for equally spaced sampling times are no longer applicable [15]. A common approach to perform parametric estimation is to construct a log-likelihood function in terms of the unknown parameter [4]. When the sampling times are considered deterministic, the traditional approach is to build the classical Gaussian log-likelihood function. However, because the inversion of the covariance matrix has to be performed, numerical evaluation of this Gaussian log-likelihood function is in general very expensive [14]. One way to overcome this computational effort is to regulate the sampling scheme, using some form of interpolation, and consider it as being equally spaced. Under the assumption of equally spaced sampling times, the Gaussian log-likelihood function can be approximated, at least for a sufficiently large sample, by the Whittle log-likelihood function [24]. This approach has been successfully applied to irregularity caused by missing values, [16]. While, it may be reasonable to use this methodology, to deal with the minor irregularities in sampling times caused by missing values, the interpolation procedure will typically change the dynamic of the underlying process, leading to biased estimates for the parameters [9]. Moreover, there is little understanding of which particular interpolation method is the most appropriate on a given data set. Alternatively, a convenient continuous time domain dynamic model may be assumed for the underlying continuous time stationary process such as the Continuous time ARMA (CARMA) model. The application of Kalman recursion techniques to the parametric estimation of CARMA processes is reviewed in [22]. Additionally, [13] estimate the parameters of an irregularly sampled CARMA process using a Bayesian framework.

A particular case of irregularly spaced data is that in which the collection procedure along time depends also, for practical constraints, on the observed values. For example, a certain health indicator for an individual may be measured at different time points and with different frequencies depending on his health state. In a completely different setting, the times of occurrence of transactions in the financial markets depend largely on the value of the underlying asset. In environmental monitoring applications, or in the context of smart cities if it is decided to monitor more frequently when a value considered critical to human health is exceeded. Therefore, additional information on the phenomena under study is obtained from the frequency or time occurrence of the observations. In such situations, there is stochastic dependence between the process being modeled and times of the observations, which may be coined as temporal preferential sampling following [8] in the context of spatial statistics.

In this work, we propose a model-based approach to analyze a time series observed under preferential sampling. 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 or the underlying asset). This latent process is assumed as Gaussian without loss of generality.

The paper is organized as follows. Section 2 describes our proposed model for preferential sampling in time dimension, namely to make inference and prediction. In Section 3 we describe the Monte Carlo Maximum Likelihood Estimation. In section 4 we conduct a numerical illustration, in an artificial data set, to analyze the quality of the proposed model. We then 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 5 is devoted to make some concluding remarks.

2. A MODEL FOR PREFERENTIAL SAMPLING

In time series, data are obtained by sampling a phenomenon $S(t) : t > 0$ at a discrete set of times t_i, $i = 1, ..., n$. Admiting the possibility that the sampling design may be stochastic, $T = (t_1, ..., t_n)$ denotes a stochastic process of observation times. In many situations, $S(t)$ cannot be measured without error, hence, if $Y(t_i)$ denotes the measured value at time t_i, a model for the data takes the form:

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

where μ is a constant mean effect and $S(\cdot)$ is a stationary Gaussian process with $E[S(t)] = 0$. An equivalent formulation is that conditional on $S(\cdot)$, the $Y(t_i)$ are mutually independent, normally distributed with mean $\mu + S(t_i)$ and common variance τ^2.

We consider $S(\cdot)$ as a continuous time autoregressive process 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, $S(\cdot)$ is asymptotically stationary if an only if $\alpha_0 > 0$ and $W(t)$ is a Brownian motion with variance parameter σ_w^2. For notation simplification let us denote $Y_i = Y(t_i)$. Then $Y = (Y_1, ..., Y_n)$ is multivariate Gaussian with mean $\mu\mathbf{1}$ and covariance matrix $\Sigma_Y = \frac{\sigma_w^2}{2\alpha_0} R_y(\alpha_0) + \tau^2 I_n$, where $\mathbf{1}$ is a n-length vector of ones, I_n is the $n \times n$ identity matrix and

$R_y(\alpha_0)$ has elements $r_{ij} = \rho(|t_i - t_j|; \alpha_0)$ defined by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \exp(-\alpha_0 |h|) \tag{2.2}$$

being $\gamma(\cdot)$ the covariance function.

Admitting that the sampling times are stochastic, a complete model needs to specify the joint distribution of S, T and Y. Considering the stochastic dependence between S and T, the model to deal with preferential sampling is defined through $[S, T, Y]$ written as:

$$[S][T|S][Y|S(T)] \tag{2.3}$$

where $[\cdot]$ means "the distribution of", $S = \{S(t) : t > 0\}$, $T = (t_1, ..., t_n)$ and $S(T)$ represents $\{S(t_1), ..., S(t_n)\}$.

We define a specific class of models through the additional assumptions: conditional on S, T is an inhomogeneous Poisson process with intensity $\lambda(t) = \exp\{a + \beta S(t)\}$ and unconditionally T is a log-Gaussian Cox process. The log-Gaussian Cox process is a flexible class of point pattern models that allows conditioning the sampling times to the variable of interest. β is the parameter that controls the degree of preferentiality, for example, $\beta = 2$ corresponds to a situation when the sampling times are concentrated, predominantly, near the maximum of the observed values and $\beta = 0$ corresponds to the situation of an homogeneous, non-preferential, sampling. Conditional on S and T, Y is a set of mutually independent Gaussian variates with τ^2 being the measurement error variance.

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 $S(t_0|T) = \mathrm{E}\big[S(t_0)|Y(T)$. Considering that the process CAR(1) is Markovian, [6, 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. MONTE CARLO MAXIMUM LIKELIHOOD ESTIMATION

We consider a discretization of the S process with N points and a partition of S into $S = \{S_0, S_1\}$, 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 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 all the model parameters.
An algebraic simplification of $[Y|T,S]$ is $[Y|S_0]$ so, we can rewrite the integral as

$$L(\boldsymbol{\theta}) = \int_S [S][T|S][Y|S_0]\frac{[S|Y]}{[S|Y]}dS. \tag{3.2}$$

Considering that $[S] = [S_1, S_0] = [S_1|S_0][S_0]$ and replacing the term $[S|Y]$ in the denominator of expression (3.2) by $[S|Y] = [S_0, S_1|Y] = [S_1|S_0, Y][S_0|Y] = [S_1|S_0][S_0|Y]$, equation (3.2) becomes

$$\begin{aligned} L(\theta) &= \int_S [S_1|S_0][S_0][T|S][Y|S_0]\frac{[S|Y]}{[S_1|S_0][S_0|Y]}dS \\ &= \int_S [T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0][S|Y]dS \\ &= E_{S|Y}\left[[T|S]\frac{[Y|S_0]}{[S_0|Y]}[S_0]\right]. \end{aligned} \tag{3.3}$$

Taking into account that the above conditional expectation can be approximated by Monte Carlo, MLE's are obtained by maximizing the Monte Carlo likelihood

$$L_{MC}(\boldsymbol{\theta}) = m^{-1}\sum_{j=1}^{m}[T|S_j]\frac{[Y|S_{0j}]}{[S_{0j}|Y]}[S_{0j}] \tag{3.4}$$

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. We may notice that j takes a value from 1 to m, the total number of Monte Carlo replicates. With this purpose, we use a technique known as conditioning by kriging [18] and we use the following construction. The new sample $S_j = U + \Sigma_S A^T\left(A\Sigma_S A^T + \tau^2 I_n\right)^{-1}(V - AU)$ where A is the $n \times N$ matrix whose ith row consists of $N-1$ 0s and a single 1 to identify the position of t_i within $T = (t_1, ..., t_n)$; $U = \Sigma_S^{1/2}u \sim MVN(0, \Sigma_S)$ with $u \sim N(0,1)$ and $\Sigma_S^{1/2}$ is obtained from the Cholesky decomposition and $V \sim MVN(y, \Sigma_Y)$. Then S_j has the required multivariate Gaussian distribution of S given $Y = y$. In practice, we use antithetic pairs of realizations to reduce Monte Carlo variance [8].

$T|S_j$ in (3.4) is an inhomogeneous Poisson process with intensity

$$\lambda(t) = \exp\{a + \beta S_j(t)\}. \tag{3.5}$$

For computational reasons, we work with logarithm and thus,

$$\log([T|S_j]) = \sum_{i=1}^{n}(a + \beta S_j(t_i)) - n\log\left(\int_0^T \exp(a + \beta S_j(t))dt\right). \tag{3.6}$$

As the S_j replicate is not known in $[0,T]$ domain, we can not calculate the integral presented in expression (3.6), so, we approximate the integral using the composed trapezium formula for unequally spaced data.

$[S_{0j}]$ in (3.4) is multivariate Gaussian with mean 0 and covariance matrix $\Sigma_{S_{0j}} = \frac{\sigma_w^2}{2\alpha_0}R_{S_{0j}}(\alpha_0)$, where $R_{S_{0j}}(\alpha_0)$ is the $n \times n$ correlation matrix with elements $r_{ij} = \rho\left(|t_i - t_j|; \alpha_0\right)$ defined by (2.2).

$[S_{0j}|Y]$ in (3.4) is multivariate Gaussian with mean $\mu_{S_{0j}|Y} = \Sigma_{S_{0j}}\Sigma_Y^{-1}(y - \mu\mathbf{1})$ and covariance matrix $\Sigma_{S_{0j}|Y} = \Sigma_{S_{0j}} - \Sigma_{S_{0j}}\Sigma_Y^{-1}\Sigma_{S_{0j}}^t$. For more details about conditional distribution see for e.g. [1].

Obtained the Maximum Likelihood Estimates (MLE's), we can plug them into (2.4) and (2.5), treating them as known. We are in position of doing the so-called plug-in predictions.

4. NUMERICAL ILLUSTRATION

In this section we document the performance of the model with time series simulated under preferential and non preferential (irregular and regular sampling) scenarios. The simulation allows control the degree of preferentiality. In addition, we apply our modeling procedure to a time series related to the biomedical marker level of platelet after a cancer patient undergoes a bone marrow transplant. Taken together, these examples suggest that our model is effective at detecting potential preferential sampling situations, estimating an adequate model and obtaining predictions for the process. We compare the results from our model with the traditional Kalman filter approach to irregularly spaced data (cts package [23]). We begin by describing the procedure to simulate a time series under preferential sampling.

4.1. Artificial data

To generate a time series under preferential sampling we first generate a realization of S from model (2.1) with $\alpha_0 = 0.2$ and $\sigma_w^2 = 1$, discretized in 400 equally spaced time points. These values correspond to $Var[S(.)] = \sigma^2 = \frac{\sigma_w^2}{2\alpha_0} = (1.581)^2$ 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 (3.5) and $\beta = 2$;
- irregular sampling: we obtain $n = 70$ sampling times T from (3.5) and with $\beta = 0$, illustrating the situation without preferential sampling;
- regular sampling: we obtain $n = 70$ sampling times with equidistant observations.

To illustrate the results of these sampling schemes, we represent in Figure 1 a realization of the process S (gray line) and the three resulting data sets. We have 70 sampling times (black points), considering $\beta = 2$ in the process intensity function, in which the preferential nature of the sampling process results in sample times falling predominantly near the maxima. For 70 sampling times (white points), we consider $\beta = 0$, the situation without preferential sampling and with irregularly sampling points. For the remaining 70 points (star points), we have the situation of regular spaced sampling times.

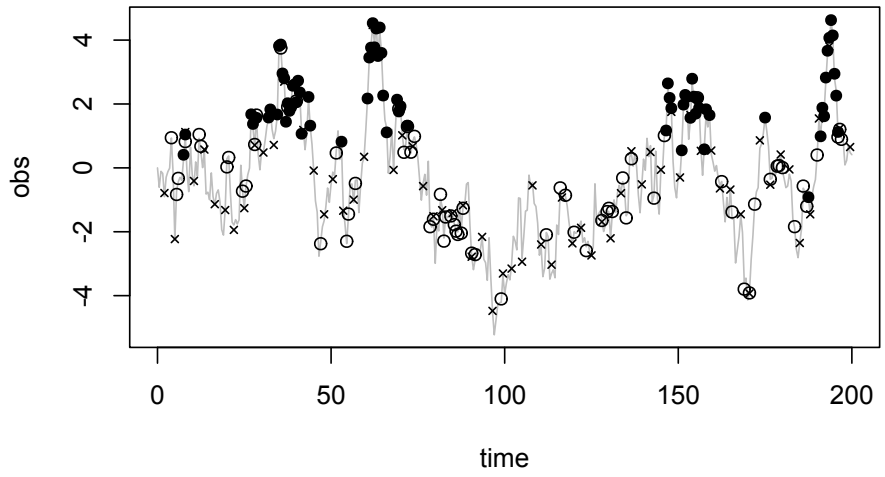

Figure 1: Sample times with preferential sampling nature (black points), without preferential sampling and irregularly spaced time points (white points), regular spaced time points (star points) and underlying process S (gray line).

The parameters μ, σ, ϕ, τ and β are the target of estimation. The estimates are obtained under (3.4), henceforward denoted by MCMLE's and from the Kalman filter, denoted by MLE's. For the maximization of our Monte Carlo log-likelihood function we considered a total of grid points $N = 400$ and a total number of MÇ replicates $m = 1000$. Mean and standard errors for the estimates obtained from 250 independent simulated samples are summarized in Table 1.

Table 1: Maximum likelihood estimates, under PS model (MCMLE's) and by cts package (MLE's), mean (standard errors) obtained from a total of 250 independent samples.

		PS Data set ($\beta = 2$)		Irregularly Sampling ($\beta = 0$)		Regular sampling	
	True	PS model	CTS	PS Model	CTS	PS Model	CTS
$\widehat{\mu}$	**0**	**0.13** (0.18)	**0.38** (0.31)	**0.04** (0.12)	**0.26** (0.34)	**0.02** (0.22)	**0.71** (0.62)
$\widehat{\sigma}$	**1.58**	**1.53** (0.21)	**0.99** (0.18)	**1.64** (0.11)	**1.52** (0.21)	**1.60** (0.13)	**1.45** (0.24)
$\widehat{\phi}$	**5**	**5.71** (1.01)	**3.17** (2.55)	**5.20** (0.48)	**5.52** (1.96)	**5.12** (0.89)	**6.78** (2.93)
$\widehat{\tau}$	**0.1**	**0.12** (0.04)	**0.27** (0.13)	**0.11** (0.01)	**0.30** (0.18)	**0.11** (0.02)	**0.55** (0.28)
$\widehat{\beta}$	**2 or 0**	**1.76** (0.39)		**0.00** (0.07)		**0.00** (0.02)	

Analysing Table 1 we conclude that the model for Temporal Preferential Sampling presents estimates for the parameters less biased, even when the preferability degree is null, with regular and irregularly sampling.

To analyse the impact of ignoring preferential sampling on the quality of predictions, we conducted a second simulation study. We simulated 250 realizations of S and for each we constructed a preferential sampling data set. Then, the proposed MCMLE's and the MLE's from the Kalman filter approach were obtained and plugged-in equation (2.4) to predict $S(t)$ at 50 equally spaced time points. These together with the corresponding standard errors, in (2.5), allowed us to calculate prediction 95% confidence intervals and estimate their coverage.

Figure 2 represents one simulation of $S(t)$ (black line), the corresponding preferential sampling data (black points) and the predictions acquired from MCMLE's (white points) and MLE's (gray points). MLE's which do not take into account the preferential character of the data lead to predictions with larger bias (overestimation of the observations) and smaller variance than that of MCMLE's. In fact, in the overall simulation results confidence intervals from MCMLE's present an estimated coverage of 88% while the MLE's provide an estimated coverage of just 73%. Thus, the proposed model leads to estimates that are less biased but with larger variance, reflecting the uncertainty associated with the observations.

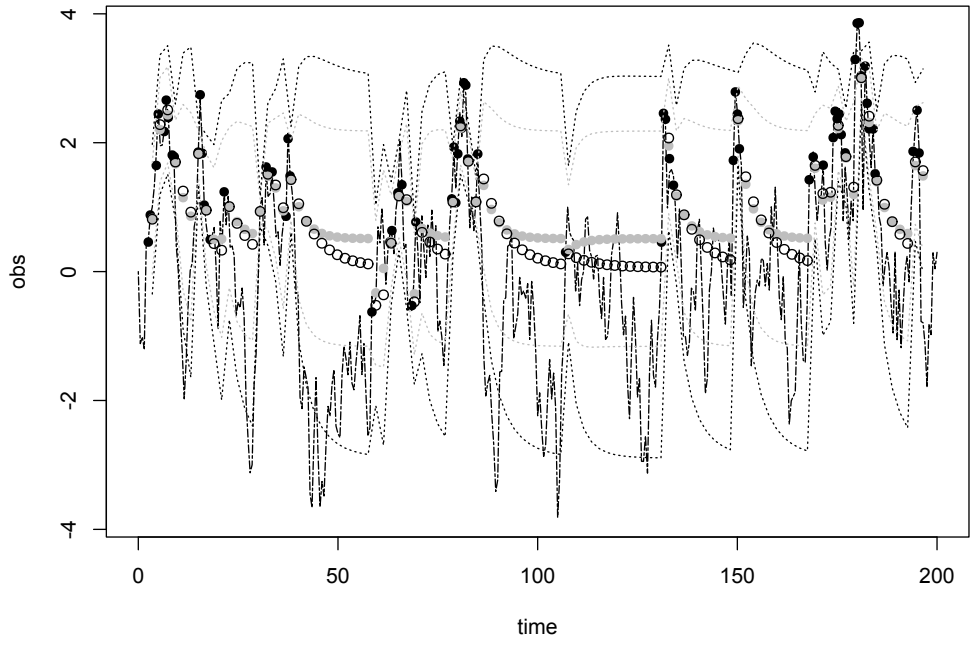

Figure 2: Predictions acquired from MCMLE's (white points) and MLE's (gray points), dashed line are confidence bands, black points are the preferential sampling data and black line is the underlying process S.

Further studies with β taking non-integer and negative values (sampling times are concentrated, predominantly, near the minima of the observed values) lead to similar conclusions.

4.2. Biomedical marker

We consider the problem of monitoring the level of a biomedical marker, platelet, after a cancer patient undergoes a bone marrow transplant. The data in Figure 3, studied in [20] as missing data problem, are 91 measurements made different days on variable log(platelet) [PLT]. In the first 35 days the data were observed daily and then irregularly, once the indicator began to show better results. According to [11], "Platelet count at about 100 days post transplant has previously been shown to be a good indicator of subsequent long term survival". This data is available in the package astsa [21] with the name of "blood".

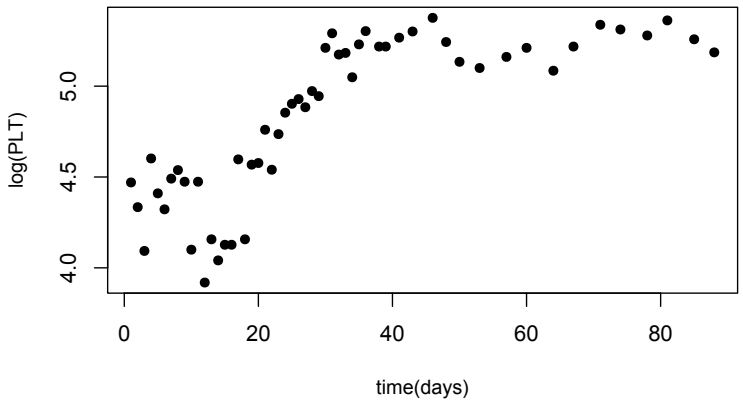

Figure 3: Measurements of the log(platelet) [PLT].

The MCMLE's for model parameters are: $\hat{\mu} = 1.99$, $\hat{\phi} = 66.18$, $\hat{\sigma} = 0.72$, $\hat{\tau} = 0.11$ and $\hat{\beta} = -2.01$. The estimated value for β with its negative sign indicates that the data was, in fact, observed under a preferential framework whereby the patient is observed more frequently when the biomarker shows lower values. Predictions of the biomarker within the period of observations are obtained plugging-in the estimated parameters in equations (2.4) and (2.5). Figure 4 top panel shows the 95% prediction intervals for (log of) the biomarker while the bottom panel represents the 95% prediction intervals obtained from the MLE's from the Kalman filter approach, with $\hat{\mu} = 1.57$, $\hat{\phi} = 53.94$, $\hat{\sigma} = 0.42$ and $\hat{\tau} = 0.13$. As expected in view of the simulation results, the predictions obtained from MCMLE present larger variance reflecting the uncertainty associated with the preferential data under analysis.

This kind of study is important, for example, to analyse when a new measurement of the patient's health indicator should be taken.

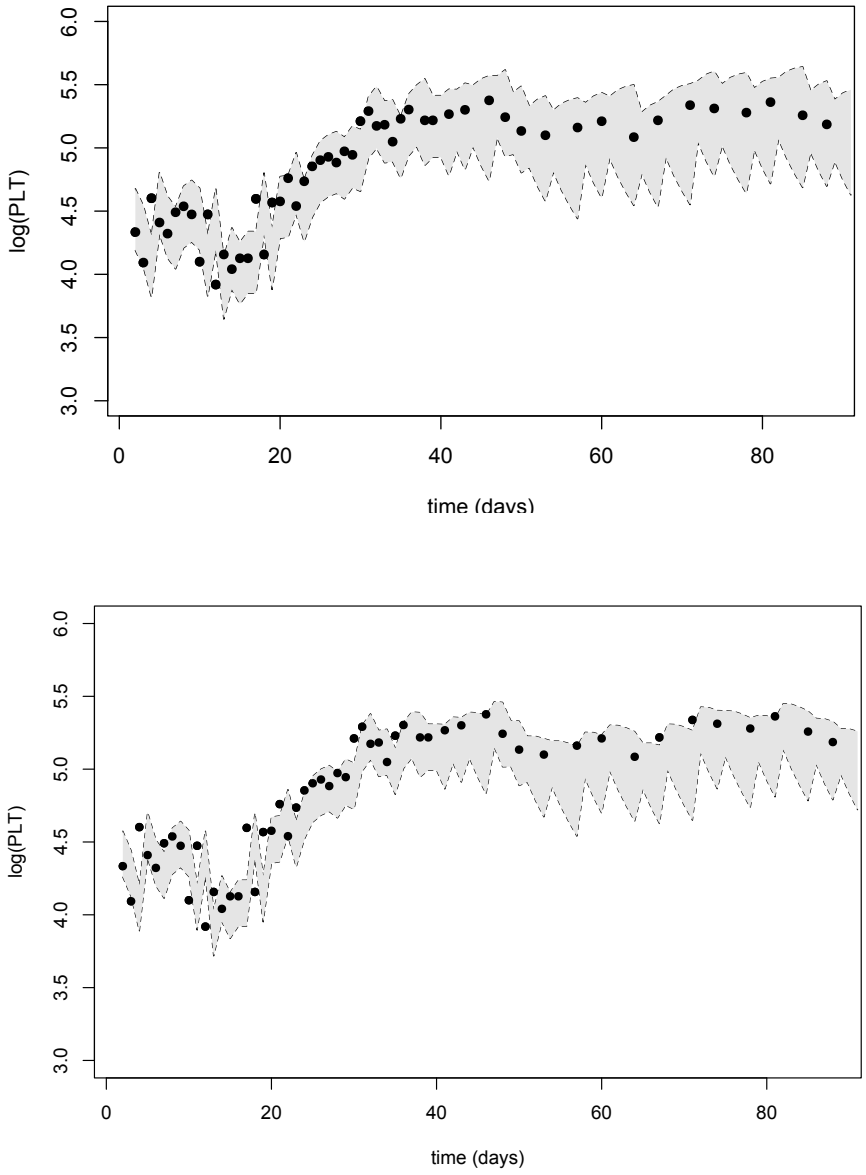

Figure 4: Prediction 95% confidence intervals using predictions acquired from MCMLE's (top) and MLE's (bottom).

5. CONCLUDING REMARKS AND FUTURE WORK

We propose, in this work, a methodology to deal with irregularly spaced time series but also a methodology that takes into account the frequency or time occurrence of the observations. The proposed model not only provides good estimates for model parameters but also reveals quite satisfactory results for prediction. A key aspect of this methodology is that it provides a tool, for example in the context of clinical trials, supporting a better knowledge of the underlying stochastic process, goal of study.

In their work, [7] affirm that the use of a single parameter in (3.5) to capture both the strength of the preferentiality and the amount of non-uniformity in sampling locations is somewhat inflexible. Alternatively, a more flexible and computational more efficient class of models, based on the proposal of [17], is discussed. These authors suggest an extension to the model proposed by [8], by adding a second Gaussian process and use of stochastic partial differential equation models. For future investigation we intend to adapt those suggestions to the time dimension.

ACKNOWLEDGMENTS

The authors acknowledge Foundation FCT (Fundação para a Ciência e Tecnologia) for funding through Individual Scholarship PhD PD/BD/ 105743/2014, Centre of Mathematics of Minho University and Center for Research & Development in Mathematics and Applications of Aveiro University within project UID/MAT/04106/2019.

REFERENCES

[1] Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

[2] 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)*, **56**, 141–155.

[3] Box, G.E.; Jenkins, G.M.; Reinsel, G.C. and Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.

[4] Brockwell, P.J. (2001). Continuous-time ARMA processes, *Handbook of Statistics*, **19**, 249–276.

[5] Brockwell, P.J. (2009). Levy-driven continuous-time ARMA processes, *Handbook of Financial Time Series*, 457–480.

[6] Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting* (second ed.), Springer.

[7] Diggle, P.J. and Giorgi, E. (2016). Preferential sampling of exposure levels, *Hanbook of Environmental and Ecological Statisics*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

[8] 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**, 191–232.

[9] Erdogan, E.; Ma, S.; Beygelzimer, A. and Rish, I. (2005). *Statistical models for unequally spaced time series*. In: "Proceedings of the 2005 SIAM International Conference on Data Mining", 626–630.

[10] Jones, R.H. (1981). Fitting a continuous time autoregression to discrete data, *Applied Time Series Analysis II*, Elsevier, 651–682.

[11] Jones, R.H. (1984). *Fitting multivariate models to unequally spaced data*. In "Time Series Analysis of Irregularly Observed Data" (E. Parzen, Ed.), Lecture Notes in Statistics, vol. 25, Springer, New York, pp. 158–188.

[12] Jones, R.H. (1985). Time series analysis with unequally spaced data, *Handbook of Statistics*, **5**, 157–177.

[13] Kelly, B.C.; Becker, A.C.; Sobolewska, M.; Siemiginowska, A. and Uttley, P. (2014). Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets, *The Astrophysical Journal*, **788**, 1–33.

[14] Lange, K. (2010). *Numerical Analysis for Statisticians*, Springer Science & Business Media.

[15] Li, Z. (2014). *Methods for irregularly sampled continuous time processes*, Ph.D. Thesis, UCL (University College of London).

[16] Little, R.J. and Rubin, D.B. (2014). *Statistical Analysis with Missing Data*, John Wiley & Sons.

[17] Pati, D.; Reich, B.J. and Dunson, D.B. (2011). Bayesian geostatistical modelling with informative sampling locations, *Biometrika*, **98**(1), 35–48.

[18] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, CRC press.

[19] Ruf, T. (1999). The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series, *Biological Rhythm Research*, **30**(2), 178–201.

[20] Shumway, R.H. and Stoffer, D.S. (2017). *Time Series Analysis and Its Applications: With R Examples*, Edition 4, Springer, New York.

[21] Stoffer, D.S. (2017). astsa: Applied Statistical Time Series Analysis, *R Package Version*, **1**, https://CRAN.R-project.org/package=astsa.

[22] Tomasson, H. (2015). Some computational aspects of Gaussian CARMA modelling, *Statistics and Computing*, **25**(2), 375–387.

[23] 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/.

[24] Whittle, P. (1961). Gaussian estimation in stationary time series, *Bull. Internat. Statist. Inst.*, **39**, 105–129.