Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

# Monitoring Short Term Changes of Malaria Incidence in Uganda with Gaussian Processes

Ricardo Andrade-Pacheco[1], Martin Mubangizi[2], John Quinn[2,3], and Neil Lawrence[1]

[1] University of Sheffield, Department of Computer Science, UK
[2] Makerere University, College of Computing and Information Science, Uganda
[3] UN Global Pulse, Pulse Lab Kampala, Uganda

**Abstract.** A method to monitor communicable diseases based on health records is proposed. The method is applied to health facility records of malaria incidence in Uganda. This disease represents a threat for approximately 3.3 billion people around the globe. We use Gaussian processes with vector-valued kernels to analyze time series components individually. This method allows not only removing the effect of specific components, but studying the components of interest with more detail. The short term variations of an infection are divided into four cyclical phases. Under this novel approach, the evolution of a disease incidence can be easily analyzed and compared between different districts. The graphical tool provided can help quick response planning and resources allocation.

**Keywords:** Gaussian processes, malaria, kernel functions, time series.

## 1 Introduction

More than a century after discovering its transmission mechanism, malaria has been successfully eradicated from different regions of world [15]. However, it is still endemic in 100 countries and represents a threat for 3.3 billion people approximately [20]. In Uganda, malaria is among the leading causes of morbidity and mortality [19]. Different types of interventions can be carried on to prevent and treat malaria [20]. Their success depend on how well the disease can be anticipated and how fast the population reacts to it. In this regard, mathematical modelling can be a strong ally for decision-making and health services planning. Spatiotemporal modelling for mapping and prediction of infection dynamics is a challenging problem. First of all, because of the costs and difficulties of gathering data. Second, because of the challenges of developing a sound theoretical model that agrees with the data observed.

The Health Management Information System (HMIS) operated by the Uganda Ministry of Health provides weekly records of the number of patients treated for malaria in different hospitals across the country. Unfortunately, the number of reporting hospitals is not consistent across time. This variation is prone to create artificial trends in the observed data. Hence, the underreporting effect has to be estimated to be removed.

A common approach for time series analysis is to decompose the observed variation into specific patterns such as *trends*, *cyclic effects* or *irregular fluctuations* [4, 3, 7]. Gaussian process (GP) models are a natural approach for analyzing

functions that represent time series. GPs provide a robust framework for non-parametric probabilistic modelling [18]. The use of covariance kernels enable to analyse non-linear patterns by embedding an inference problem into an abstract space with a *convenient structure*[14]. By combining different covariance kernels (via additions, multiplications or convolutions) into a single one, a GP is able to describe more complex functions. Each of the individual kernels contributes by encoding a specific set of properties or pattern of the resulting function [5].

We propose a monitoring system for communicable diseases based on Gaussian processes. This methodology is able to isolate the relevant components of the time series and study the short term variations of the disease. The output of this system is a graphical tool that discretizes the disease progress into four phases of simple interpretation.

## 2 Background

Say we are interested in learning the functional relation, between inputs and output, based on a set of observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. GP models introduce an additional *latent variable* $f_{\mathbf{x}}$, whose covariance kernel $K$ is a function of the input values. Usually, $y_i$ is considered a distorted version of the latent variable.

To deal with multiple outputs, GP models resort to generalizations of kernel functions to the vector-valued case [1]. In time series literature, vector-valued functions are commonly treated in the family of VAR models [12], while in geostatistics literature *co-Kriging* generalizations are used [8, 11]. These approaches are equivalent. Let $h_{\mathbf{x}} = (f_{\mathbf{x}}^1, \ldots, f_{\mathbf{x}}^d)^\top$ be a vector-valued GP, its corresponding covariance matrix is given by

$$\left[\operatorname{cov}(h_{\mathbf{x}}, h_{\mathbf{z}})_{ij}\right] = \left[\operatorname{cov}(f_{\mathbf{x}}^i, f_{\mathbf{z}}^j)\right]. \tag{1}$$

The diagonal elements of the correlation matrix $\left[\operatorname{cov}(h_{\mathbf{x}}, h_{\mathbf{z}})_{ii}\right]$ are just the covariance functions of the real-valued GP elements. The non-diagonal elements represent the *cross-covariance functions* between components [9, 10, 2].

## 3 Method Used

Suppose we have data generated from the combination of two independent signals (see Figure 1a). Usually, not only we are not able to observe the signals separately, but the combined signal they yield is corrupted by noise in the data collected (see Figure 1b). For the sake of this example, suppose that the two signals of the example represent a long term trend (the smooth signal) and a seasonal component (the sinusoidal signal). For an observer, the oscillations of the seasonal component masks the behaviour of the long term trend. At some point, however, the observer might want to know whether the trend is increasing or decreasing. Similarly, there might be interest in studying only the seasonal

component isolated from the trend. For example, in economics and finance, business recession and expansion periods are determined by studying the cyclic component of a set of indicators [16]. The cyclic component tells if an indicator is above or below the trend, and its differences tell if it is increasing or decreasing.

We propose a similar approach for monitoring disease incidence time series, but in our case, we will use a non-parametric approach. To extract the original signals, the observed data can be modelled using a GP with a combination of kernels, say exponentiated quadratics, one having a shorter lengthscale than the other. Figures 1c and 1d shows a model of the combined and independent signals. We also use a vector-valued GP to model directly the derivative of the time series, rather than using simple differences of the observed trend. As a result, we are able to provide uncertainty estimates about the speed of the changes around the trend. Our approach is based on modelling linear functionals of an underlying GP [13]. If $h_{\mathbf{x}} = (f_{\mathbf{x}}, \partial f_{\mathbf{x}}/\partial x_i)^\top$, its corresponding kernel is defined as

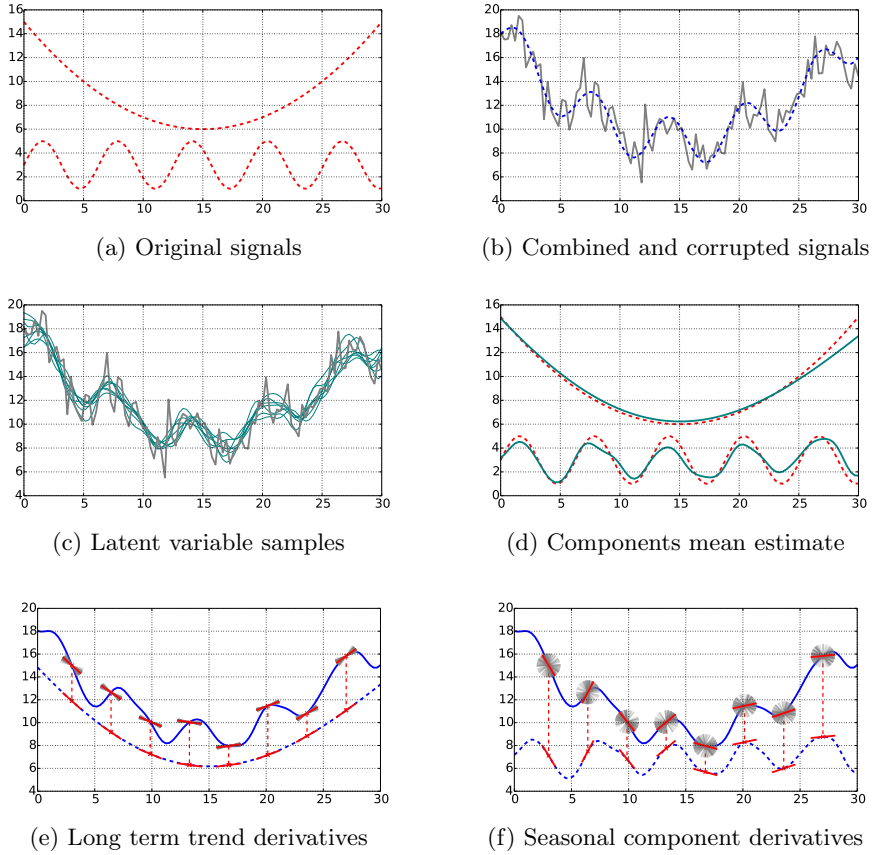
$$\Gamma(\mathbf{x}_i, \mathbf{x}_j) = \begin{bmatrix} K(\mathbf{x}_i, \mathbf{x}_j) & \frac{\partial}{\partial x_j} K(\mathbf{x}_i, \mathbf{x}_j) \\ \frac{\partial}{\partial x_i} K(\mathbf{x}_i, \mathbf{x}_j) & \frac{\partial^2}{\partial x_i x_j} K(\mathbf{x}_i, \mathbf{x}_j) \end{bmatrix}. \tag{2}$$

In most multi-output problems, observations of the different outputs are needed to learn their relation. Here, the relation between $f_{\mathbf{x}}$ and its derivative is known beforehand through the derivative of $K$. Thus $\partial f_{\mathbf{x}}/\partial x_i$ can be learnt by relying entirely on $f_{\mathbf{x}}$. For the signals described above, Figures 1e and 1f show the corresponding derivatives computed using a kernel of the form of (2). The derivatives of the long term trend are computed with high confidence, while the derivatives of the seasonal component have more uncertainty. The last is due to the magnitude of the seasonal component relative to the noise magnitude.

## 4 Uganda Case

In this exposition we focus on Kabarole district, but provide a snapshot of the monitoring system for all the country. Our base assumption about the infection process of malaria is that it evolves with some degree of smoothness across time. Smooth functions can be represented by a kernel such that the closer the observations in the input space, the more similar values of the output. The Matérn kernel family satisfies this condition, as it defines dependence through the distance between points with some exponential decay [18]. Different members of this family encode different degrees of smoothness, being the limit case the exponentiated quadratic kernel or RBF, which is infinitely differentiable. To illustrate our method we will use an RBF kernel. Results with (rougher) Matérn kernels do not differ much when used instead.

Despite malaria is a disease influenced by environmental factors like temperature or water availability, we could not observe a seasonal effect in HMIS data [6]. If that was the case, the model could be improved incorporating a periodic kernel in the covariance structure. Yet, the model fit can be improved if a second RBF kernel is added. In this case, one kernel has a short lengthscale and



(a) Original signals

(b) Combined and corrupted signals

(c) Latent variable samples

(d) Components mean estimate

(e) Long term trend derivatives

(f) Seasonal component derivatives

Fig. 1: Series decomposition. Panel (a) shows two independent signals. Panel (b) shows the combination of both signals (dashed line) and a distorted signal after adding some noise (solid line). Panel (c) shows latent variable samples representing the combined signal (thin lines). Panel (d) compares the mean estimate of each component (solid line) with the original signals (dashed line). Panels (e) and (f) show the components derivatives. Tangent lines to the individual components are shown in red. The solid blue lines represent the mean estimate of the composed signal. The gray lines are random realizations of process derivative. For comparison, the estimates of the individual signals (dashed lines) are shown below the composed signal.

therefore represents short term variations, while the other represents long term changes.

An important factor to consider about HMIS data is that the number of health facilities is highly variable. See Figure 2a. This variation is prone to create artificial trends in the incidence of malaria reported. Such trends can be removed by incorporating a linear kernel that describes the relation between reporting facilities and malaria cases. Unlike the RBF kernels mentioned above,

which take time as input, the linear kernel takes the number of health facilities as input. In Table 1, we present a comparison of the model predictive performance, when using different kernels, based on the leave-one-out predictive probabilities [17]. The best predictive performance is achieved when considering short and long term changes and a correction for misreporting facilities.

Table 1: Comparison of LOO-CV log predictive probabilities, when using different kernels. The subindex $\ell$ refers to the lengthscale of the kernel (measured in years).

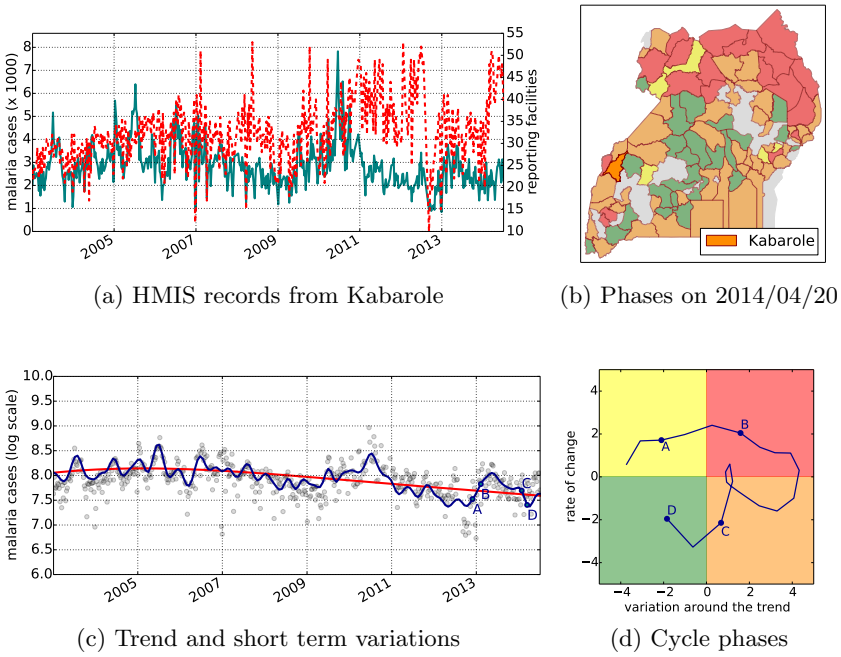
| Kernel | LOO-CV (log) |
|---|---|
| $RBF_{\ell=0.64}$ | -40.54 |
| $RBF_{\ell=0.14} + RBF_{\ell=10}$ | -16.26 |
| $RBF_{\ell=0.12} + RBF_{\ell=10} + Linear$ | 41.21 |

Figure 2c shows the trend and short term component of the number of malaria cases. Variations of a disease incidence around its trend represent short term changes in the population health. Outbreak detection and control of non-endemic diseases take place in this time frame. For some endemic diseases, this variation can be associated to seasonal factors [6]. Quick response actions, such as distribution of medicine and allocation of patients to health centres, have to take place in this time regime to be effective. The short term variations can be classified in four phases as shown in Figure 2d (values are standardized). The upper left quadrant represents an incidence below the trend, but increasing; the upper right quadrant represents an incidence above the trend and expanding; the bottom right quadrant represents an incidence above the trend, but decreasing; and the bottom left quadrant represents an incidence below the trend and decreasing.

This tracking system of short term variations is independent of the order of the disease counts, and can be used to monitor the infection progress in different districts. It is easy to identify districts where the disease is being controlled or where the infection is progressing at an unusual rate. Figure 2b shows the monitoring system on the whole country. Those districts where the variation coefficient of both the process and its derivative are less than 1 (meaning a weak signal vs noise) were left in gray color.

## 5 Final Remarks

We have proposed a disease monitor based on vector-valued Gaussian processes. Our approach is able to account for uncertainty in both the level of each component and the direction of change. The simplicity for doing inference with this




(a) HMIS records from Kabarole

(b) Phases on 2014/04/20

(c) Trend and short term variations

(d) Cycle phases

Fig. 2: Malaria incidence tracker in Uganda. Panel (a) compares the number of malaria cases (solid line) and the number of reporting health facilities (dashed line). Panel (b) shows the disease phase in each district. The colors are assigned according to the quadrants in panel (d). Panel (c) shows the long term trend (dashed line) and the short term variations (solid line). Gray bullets represent the observed records. Panel (d) shows a tracking system of the short term variations. The bullets A-D in panels (c) and (d) correspond to the same time points.

model is not compromised by the use of a vector-valued approach. The model can be benefited if spatial information is available and encoded in the kernel function. Further research is needed to explore the benefits of this model in practice. We expect that an analysis from this perspective can add situational awareness and contribute to interventions planning and resources allocation when facing infectious diseases.

## Acknowledgements

Ricardo Andrade-Pacheco is supported by CONACYT and SEP scholarships.

## References

1. M. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

2. L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.
3. M. Baxter and R. G. King. Measuring business cycles: approximate band-pass filters for economic time series. *Review of economics and statistics*, 81(4):575–593, 1999.
4. W. P. Cleveland and G. C. Tiao. Decomposition of seasonal time series: A model for the census X-11 program. *Journal of the American statistical Association*, 71(355):581–587, 1976.
5. N. Durrande, J. Hensman, M. Rattray, and N. D. Lawrence. Gaussian process models for periodicity detection. *arXiv preprint arXiv:1303.7090*, 2013.
6. S. I. Hay, R. W. Snow, and D. J. Rogers. From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. *Parasitology Today*, 14(8):306–313, 1998.
7. A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.
8. G. Matheron. Pour une analyse krigeante de donnés régionalisées. Technical report, École des Mines de Paris, Fontainebleau, France, 1982.
9. C. A. Micchelli and M. Pontil. Kernels for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2004.
10. C. A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.
11. D. E. Myers. Matrix formulation of co-Kriging. *Journal of the International Association for Mathematical Geology*, 14(3):249–257, 1982.
12. H. Quenouille. *The analysis of multiple time-series*. Griffin's statistical monographs & courses. Griffin, 1957.
13. S. Särkkä. Linear operators and stochastic partial differential equations in Gaussian process regression. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 151–158. Springer, 2011.
14. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.
15. P. I. Trigg and A. V. Kondrachine. Commentary: malaria control in the 1990s. *Bulletin of the World Health Organization*, 76(1):11, 1998.
16. F. van Ruth, B. Schouten, and R. Wekker. The statistics Netherlands business cycle tracer. Methodological aspects; concept, cycle computation and indicator selection. Technical report, Statistics Netherlands, 2005.
17. A. Vehtari, V. Tolvanen, T. Mononen, and O. Winther. Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *arXiv preprint arXiv:1412.7461*, 2014.
18. C. K. I. Williams and C. E. Rasmussen. *Gaussian processes for Machine Learning*. MIT Press, 2006.
19. World Health Organization. World health statistics 2015. Technical report, WHO Press, Geneva, 2015.
20. World Health Organization and others. World malaria report 2014. Technical report, WHO Press, Geneva, 2014.