Malaria surveillance with multiple data sources using Gaussian process models

Martin Mubangizi*, Ricardo Andrade-Pacheco†, Michael Smith*, John A. Quinn*‡ and Neil Lawrence†
*Makerere University, Kampala, Uganda
{mmubangizi,msmith,jquinn}@cit.ac.ug
†University of Sheffield, UK
{acq11ra,N.Lawrence}@sheffield.ac.uk
‡UN Global Pulse, Kampala, Uganda

***Abstract*—A statistical framework for monitoring the health of a population should ideally be able to combine data from a wide variety of sources, such as remote sensing, telecoms, and official health records, in a principled manner. Gaussian process regression is commonly used to visualise disease incidence by interpolating values across a map; in this article, we show how it can be extended to deal with many different types of information by introducing a flexible covariance structure across data sources. Combining many data sources in a single model provides a number of practical advantages, such as the ability to to automatically determine the importance of each data source through likelihood optimisation, and to deal with missing values. We show the basic idea with an application of malaria density modeling across Uganda using administrative records and remote sensing vegetation index data, and then go on to describe further extensions such as the incorporation of human mobility data extracted from mobile phone call detail records (CDRs).**

I. Introduction

Malaria remains endemic across much of the world, in spite of mitigation measures by both governments and international agencies. Health department intervention is now principally response-driven; at those times and locations with the greatest malaria infection rates the provision of treatment needs to be able to match the number of cases without stock-outs or staff-shortages. Hence planning stock and staff deployment depends on accurate and timely information regarding the distribution of malaria cases. In Uganda, the Ministry of Health receives weekly counts of reported malaria cases from all districts. However, this data is compromised by cases of non-reporting at both the district and health center levels [1], the cases reported are often based on unverified diagnoses, and there are various other sources of measurement error.

In order to resolve ambiguity about how the disease burden is distributed, models can be constructed which relate infection levels across time and space, or incorporate covariates which provide extra information. These covariates may be environmental (rainfall levels, temperature, vegetation strength) or social (population density, migration/movement patterns, demographics), for example. In this regard, NDVI index, which is widely used to estimate vegetation density [2], turns out to be good proxy for rainfall [3] and has proved useful in identifying suitable habitats for mosquito breeding [4].

Any attempt to use remote sensing data, such as NDVI, for carrying out inference on administrative records, will face the problem of trying to mix two data sources with differing space and time resolutions. For example, while HMIS data is reported weekly and aggregated at a district level[f], NDVI is provided a much higher resolution in a grid and is reported every 5 days.

Gaussian process regression is commonly used in epidemiology to interpolate disease counts across space. In this paper, we explain how it can be extended to a coregionalised form in order to incorporate information from covariates. By specifying a covariance structure relating a number of inputs and outputs, it is possible to combine several different types of data in a single, principled framework. We illustrate this model using weekly counts of malaria incidence by district in Uganda, and show that for certain regions, the incorporation of environmental remote sensing data can significantly improve the estimates of the infection rate compared to baseline models. We then describe how social data can be incorporated, in particular information about movements of the population derived from mobile phone call detail records.

This paper is organised as follows. Section II discusses some of the related work; Section III presents the data used and introduces the model framework. Application of the model to environmental covariates is discussed in IV, and Section V discusses use of mobility data. We conclude, with suggestions for future work, in Section VI.

II. Related Work

The need to use data from multiple sources to enhance disease modeling has been an active research area [5], [6]. [7] cites challenges that this research has been faced with. This also led to search for new data sources that may provide signals of changes in disease rates, including absenteeism [8], sales of over-the-counter health products [9], emergency call centers [10], and automatic malaria diagnosis results [11]. Examples of research that has focused on using multiple data sources, such as [6], acknowledge the need for data from multiple sources in biosurveillance. BioPHusion [5], for instance, is a framework that can use real time data from several sources for awareness and timely response.

It is widely understood that determining the geographical distribution of a disease is vital in its control [12] and in estimating the cost of that control [13]. To this end, considerable effort has gone into producing risk maps of diseases at different

[f]HMIS data might be available at smaller aggregation levels, however the information available to the authors had a district aggregation.

 ISBN: 978-0-7972-1533-7

spatial scales—by country [14], continent [15] and at global [12], [16] scale. These example risk-maps are over a long time-scale however, looking at seasonal averages. The methods we propose offer predictions of disease counts at a weekly time-scale, allowing more detailed and precise estimates for operational use. One feature these studies have in common is the use of remote sensing for disease prediction; we use this same idea but at a much shorter time-scale.

The use of mobile phone CDRs for modelling the effects of human mobility on the distribution of malaria infection is also gaining traction in the literature (for example see [17] for a review, but also see [18] for issues around this data).

Gaussian process regression in epidemiology

Gaussian Process Regression, or Kriging models, were first introduced in the 1960s for geostatistics, and since then the method however has had application across many disciplines. In brief, the method works by estimating the correlation structure of the data (over time, space or other dimension of interest) then using these estimates of correlation the values of the output can be estimated from training data. This basic regression can be extended to combine multiple output variables, by estimating their coregionalised correlations. Further, the uncertainty and absence of data can be incorporated, allowing our confidence in each data point to be taken into account. Finally the output also includes confidence intervals, giving important information about the reliability of each of the model's estimates. The use of Kriging in public health datasets is commonly used to interpolate disease incidence across space. For example, Kleinschmidt et al. [19] applied this method to malaria mapping.

III. Data and Methods

A. Data

a) Uganda Health Management Information System (HMIS): HMIS, hosted at the Ministry of Health in Uganda, manages countrywide reported cases of diseases of public health importance including malaria. HMIS receives weekly counts of reported malaria cases from health centers aggregated at district level. For each week, the number of cases in each district is reported, with an associated statistic on the proportion of health centres which were included. Often the number reporting is low, causing degradation in the quality of the data, to the point where, unprocessed, the data is of little use [1].

b) Population estimates: This data was obtained from World Pop[e], which provides estimates of the number of people living in 100m square grid cells across the entire country.

c) Normalised Differenced Vegetation Index (NDVI): NDVI gives a measure of how vigorous the vegetation is across space. In this study we use vegetation index data from eMODIS obtained from the Famine Early Warning System of the United States Geological Survey (USGS FEWS)[d]. This data was obtained at a spatial resolution of 250m, every five days. To reduce the computational complexity and make the remote sensing data more representative, a population-density-weighted average of the remote sensing data was calculated for each district.

[e]http://www.worldpop.org.uk/data

[d]http://earlywarning.usgs.gov/fews/

B. Methods

A Gaussian process (GP) regression is a machine learning algorithm for relating an output $\mathbf{y}$ (e.g. disease incidence) with a set of inputs $\mathbf{X}$ (e.g. longitude and latitude). The core assumption of this mathematical model is that there is an unobserved or latent variable $\mathbf{f}$ that depends on $\mathbf{X}$, but for which we only have access through its distorted version $\mathbf{y}$. This unobserved variable is a Gaussian process with some mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ which depend on the inputs [20]. The distortion is given by independent random noise at each observation.

It is possible to extend GP regression to deal with many outputs, rather than just one [21]. Broadly speaking, this approach consist of defining a multiple output kernel functions able to incorporate information from different outputs and use it to model the correlation between them. Here we are interested in showing through an application how these kind of models can be used for integrating different sources of information for malaria modelling.

Assume we have d sets of outputs and inputs $\{\mathbf{y}_1, \mathbf{X}_1\}, \ldots, \{\mathbf{y}_d, \mathbf{X}_d\}$, where all $\mathbf{X}_j$ belong to the same domain. The number of observations in each set can be different and the domains of the outputs do not have to be the same. A first approach for learning all these tasks could be to model each with a separate GP. However if we know that the outputs might be correlated we could also try to model them together. This way, information from one domain can constrain the values in another.

The mathematical formulation for such coregionalised models is broadly the same as for standard, single-output GP regression. We use the same pairing of outputs and inputs from the d original sets,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_p \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_p \end{pmatrix}.$$

The covariance matrix is defined in a block structure, where each block contains the weighted cross-correlation. Thus, given a kernel matrix $\mathbf{K}$ and a matrix of weights $\mathbf{B}$, the multiple output kernel $\mathbf{M}_K$ is defined as

$$\begin{aligned} \mathbf{M}_K =& \mathbf{B} \otimes \mathbf{K}(\mathbf{X}, \mathbf{X}) \\ =& \begin{pmatrix} B_{1,1} \cdot \mathbf{K}(\mathbf{X}_1, \mathbf{X}_1) & \ldots & B_{1,d} \cdot \mathbf{K}(\mathbf{X}_1, \mathbf{X}_d) \\ \vdots & \ddots & \vdots \\ B_{d,1} \cdot \mathbf{K}(\mathbf{X}_d, \mathbf{X}_1) & \ldots & B_{d,d} \cdot \mathbf{K}(\mathbf{X}_d, \mathbf{X}_d) \end{pmatrix}. \end{aligned} \tag{1}$$

Complex covariance structures can be defined by constructing $\mathbf{K}$ from other kernels [22] or by using a linear combination of multiple output kernels, thus defining

$$\mathbf{M}_K = \sum_{r=1}^{R} \mathbf{B}_r \otimes \mathbf{K}_r(\mathbf{X}, \mathbf{X}). \tag{2}$$

 ISBN: 978-0-7972-1533-7

IV. Application

A. Temporal Modeling

As mentioned earlier, malaria models can be improved by considering covariates such as NDVI index. Here we show an example for modelling both variables across time. For this task, vector autoregressive models or a general linear model might be considered as a first option for studying the relation between this two variables. However these models require the input and output variables to be sampled at regular and equal time and space intervals. This is usually resolved by the interpolation of one of the variables. One of the advantages of coregionalised GP regression is this step is not required, and the uncertainty in an interpolated value is already incorporated into the result.

For each district, we trained single GP regression model for malaria incidence and a joint model with NDVI. Then we predicted malaria incidence 180 days ahead. In the first model, the prediction only depended on past observation of the same variable. In the second model the prediction was aided by the training observations of NDVI which overlaped the period of malaria prediction.

Figure 1 shows a comparison of a single GP regression model of malaria incidence with a joint model with NDVI information. It can be clearly seen that HMIS and NDVI are strongly correlated (values are standardized in the figure), and that the joint model performs better at predicting values from the test set.

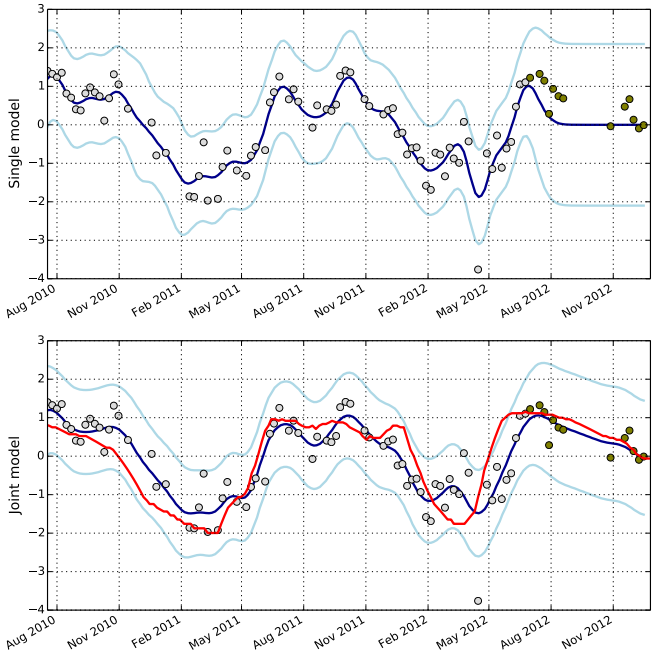

Fig. 1. Malaria incidence in Napak. The image above shows a the predictions using an independent model. The image below shows the predictions using a joint model. Training and test points are shown in gray and green circles. Predictive mean and confidence intervals are shown in solid blue lines. NDVI data is shown in red.

The similarity between malaria incidence and NDVI does not generalise across all districts. To identify those districts where there seems to be a stronger relation between these two variables we used the the quantity

$$\beta = \frac{B_{1,2}}{\sqrt{B_{1,1}B_{2,2}}}, \tag{3}$$

where $B_{i,j}$ are the entries of $\mathbf{B}$ (the coregionalisation matrix). Despite the similarity in the equation (3) with the definition of correlatin between two variables, it is worth highlighting that we are not giving $\beta_{1,2}$ the same interpretation.

We found that the mean squared errors (MSE) of the coregionalised model tend to be smaller than the ones from the single model, in districts with larger values of β. Figure 2 shows the ratio of MSE between the models for districts with $\beta > 0$.

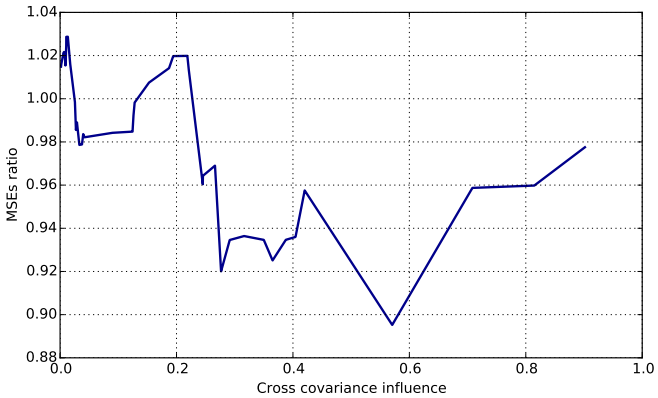

Fig. 2. MSEs vs cross covariance influence. The line shows the total MSEs of all districts with a value equal or larger than β (cross covariance influence).

Intuitively, the joint model should be as good as the single model, as in the worst scenario, where no correlation is found, $\mathbf{B}$ would be the identity and therefore we would be assuming independence. There are however a few reasons why this intuition is not totally right and, as shown in figure 1, where we can expect the joint model to have a poor performance.

First of all, model that uses a kernel like the one in (1), but where $\mathbf{B}$ is the identity, is not entirely independent. Although correlation across outputs is zero, by learning the parameters of $\mathbf{K}$ with information of both outputs we are forcing the model to share information. If both variables are different, models where the kernel parameters are learnt separately can be better. By using a kernel defined as in 2 we can create a covariance structure where a covariance structure of the joint model, leads to actual independent individual covariance structures, where the kernel parameters are not shared.

Another case where the joint model can perform poorly is when outputs are not correlated, but still there are spurious correlations. In such situation, we would only be learning and sharing noise across outputs. This can be originated by the fact that two variables behave similar in for some period or because one of the variables has scarse observations that almost all learning depends on the observations of the other variable. In our experiments, we found that in district Kole $\beta \approx 1$, while there seems to be no relation between this two variables.

A computational problem that may arise due the number of elements in $\mathbf{B}$ increases quadratically with respect to the

 ISBN: 978-0-7972-1533-7

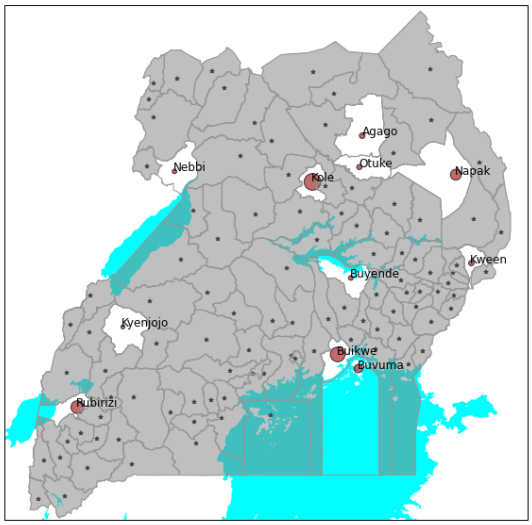

Fig. 3. Map of Ugandan districts, those with grey background have $\beta \leq 0.3$, while those with white background have $\beta > 0.3$. Points in each district have been drawn at the weighted centroid that used population as a weight.

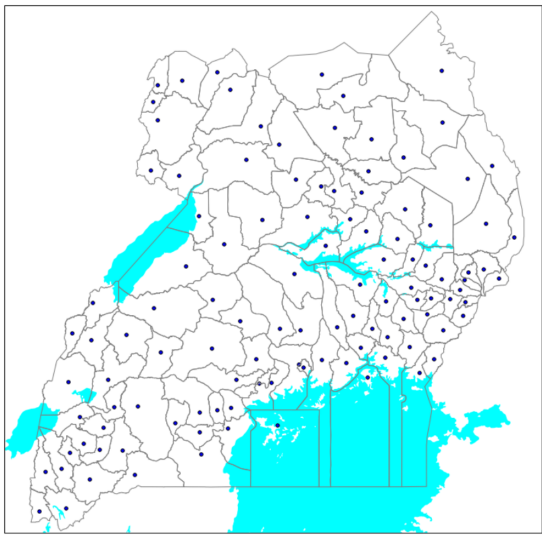

Fig. 4. Calculated positions of population-weighted district centroids.

number of outputs. Because GPs are an inference method that relies on gradient optimization, when number of parameters is large, the contribution of each one to the objective function can become negligible for some initializations.

B. *Spatial Modeling*

The GP coregionalised model can also help in combining HMIS data with satellite environment covariates (such as topology, NDVI, land surface temperature, land cover and land Use data) to produce a continuous surface of malaria disease risk. Here HMIS and the covariates can be treated as outputs of the model. The inputs of HMIS we can associate with reporting locations if known, but if not available then a population-weighted centroid (such as in Figure 4) can be calculated and used. The inputs of the environmental covariates will be the locations at which their values will be sampled, in this case also population can be used if its distribution is known. Since the model can exploit correlation across different outputs in space, the HMIS values will be smoothed out to generate a risk surface.

V. Telecoms-derived human mobility data

Early models of epidemiology considered disease dispersion to depend on geographical proximity of places, or on simple gravity models of human movement, although movement patterns can be complex and significantly affect the distribution of infectious diseases [23]. Population mobility can be obtained from CDRs for example by simply counting, for each time frame, how many people moved between each pair of cell towers on a telecoms network. Thus when a single user makes or receives a call routed through cell tower i, then later makes or receives a call routed through cell tower j, we increment the count of $i \rightarrow j$ movements. This results in a transition matrix T_{CDR}, whose entries denote the fraction of people moving from one location to another.

This information, originally recorded by each tower, can be aggregated at different areas (e.g. districts), to show the average movement between them. This gives an idea of the proximity between regions that can be more informative than the actual distance between regions for analyzing infectious diseases.

Since malaria is not transmitted directly between humans, but need a mosquito as intermediary, the measure of proximity we need goes beyond human mobility. For each individual that travels from one region to another, we also need to incorporate information about infection rate in the region of origin and probability of infection the destiny region. As proxies of both quantities we can use the parasite rate and reproductive number from [12]. Finally, we should weight the movements across regions by the population en each one. Thus we can think of a transition matrix, whose elements are defined as

$$T(i,j) = T_{CDR}(i,j) \cdot P(i) \cdot RC(j) \cdot PR(i) \ . \qquad (4)$$

Due to the lack of up to date census information in Uganda, we use values calculated from the population estimation of 2010 worldpop[g].

With this transiton matrix, we can modify the distances between regions given by their centroids to include mobility information. Figure 5 shows how we shift the centroid of each district towards those that have more access to it, based on daily movements. We can use the coordinates of this new space as the inputs in a GP and then apply the methods we have discussed so far.

VI. Conclusion

We have presented coregionalised Gaussian process regression as a method that can support both spatial and temporal modeling of disease incidence using a range of different types of data. Whereas standard GP regression, or Kriging, is a well-established method in epidemiology in general and malaria surveillance in particular, this model provides the ability to

[g]http://www.worldpop.org.uk/data

 ISBN: 978-0-7972-1533-7

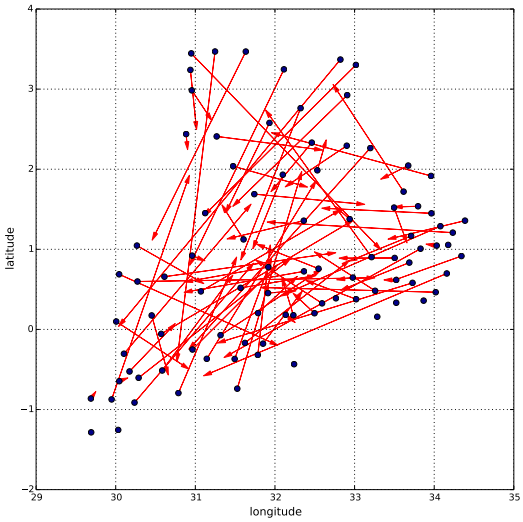

Fig. 5. Space generated by telecom data. The points correspond to the district centroids. The arrows show how each centroid is offset towards those that are closer to it in terms of the telecom data.

augment the basic regression with other data types that might be informative. In turn, while the coregionalised method has been used successfully in a number of other domains where fusion of data of different types is necessary, it has not previously been proposed in epidemiology. Using data from Uganda, we have illustrated the operation of this type of model and the types of inference it can support with remote sensing and telecoms data. We are currently collecting a more extensive dataset in order to evaluate the predictive power of these models against alternative models.

Acknowledgment

Author MM was partly funded by Google. Author RAP was funded by CONACYT and SEP scholarships. We are grateful to Orange Uganda for providing mobility data.

References

[1] World Health Organization, "Assessment of health facility data quality. Data quality report card Uganda, 2010-2011," WHO Press, Geneva, Tech. Rep., 2011.

[2] A. R. Huete, H. Q. Liu, K. Batchily, and W. J. D. A. van Leeuwen, "A comparison of vegetation indices over a global set of TM images for EOS-MODIS," *Remote sensing of environment*, vol. 59, no. 3, pp. 440–451, 1997.

[3] A. C. A. Clements, H. L. Reid, G. C. Kelly, and S. I. Hay, "Further shrinking the malaria map: how can geospatial science help to achieve malaria elimination?" *The Lancet infectious diseases*, vol. 13, no. 8, pp. 709–718, 2013.

[4] S. Hay, J. Omumbo, M. Craig, and R. Snow, "Earth observation, geographic information systems and¡ i¿ plasmodium falciparum¡/i¿ malaria in sub-saharan africa," *Advances in Parasitology*, vol. 47, pp. 173–215, 2000.

[5] H. Rolka, J. OConnor, and D. Walker, "Public health information fusion for situation awareness," in *Biosurveillance and Biosecurity*, ser. Lecture Notes in Computer Science, D. Zeng, H. Chen, H. Rolka, and B. Lober, Eds. Springer Berlin Heidelberg, 2008, vol. 5354, pp. 1–9. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89746-0_1

[6] A. W. Moore, B. Anderson, K. Das, and W.-K. Wong, "Combining multiple signals for biosurveillance," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 15, pp. 321–331.

[7] D. D. Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House, "Four key challenges in infectious disease modelling using data from multiple sources," *Epidemics*, no. 0, pp. –, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S175543651400053X

[8] L. Lenert, J. Johnson, D. Kirsh, and R. M. Aryel, "Absenteeism," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 24, pp. 361–368.

[9] W. R. Hogan and M. M. Wagner, "Sales of over-the-counter healthcare products," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 22, pp. 321–331.

[10] R. M. Aryel and M. M. Wagner, "Emergency call centers," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 25, pp. 369–374.

[11] M. Mubangizi, C. Ikae, A. Spiliopoulou, and J. A. Quinn, "Coupling spatiotemporal disease modeling with diagnosis," *Twenty-Sixth AAAI Conference*, 2012.

[12] P. Gething, A. Patil, D. Smith, C. Guerra, I. Elyazar, G. Johnston, A. Tatem, and S. Hay, "A new world malaria map: Plasmodium falciparum endemicity in 2010," *Malaria Journal*, vol. 10, no. 1, p. 378, 2011. [Online]. Available: http://www.malariajournal.com/content/10/1/378

[13] J. Omumbo, A. Noor, I. Fall, and R. Snow, "How well are malaria maps used to design and finance malaria control in africa?" *PLoS ONE*, vol. 8, no. 1, 2013.

[14] A.-S. Stensgaard, P. Vounatsou, A. W. Onapa, P. E. Simonsen, E. M. Pedersen, C. Rahbek, and T. K. Kristensen, "Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in uganda: predictors of risk and geographical patterns of co-endemicity," *Malaria journal*, vol. 10, p. 298, 2011. [Online]. Available: http://europepmc.org/articles/PMC3216645

[15] H. G. M. Zour, S. Wanji, M. Noma, U. V. Amazigo, P. J. Diggle, A. H. Tekle, and J. H. F. Remme, "The geographic distribution of loa loa in africa: Results of large-scale implementation of the rapid assessment procedure for loiasis (raploa)," *PLoS Negl Trop Dis*, vol. 5, no. 6, p. e1210, 06 2011.

[16] S. I. Hay, C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. F. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow, "A world malaria map: Plasmodium falciparum endemicity in 2007," *PLoS Med*, vol. 6, no. 3, p. e1000048, 03 2009.

[17] D. K. Pindolia, A. J. Garcia, A. Wesolowski, D. L. Smith, C. O. Buckee, A. M. Noor, R. W. Snow, and A. J. Tatem, "Human movement data for malaria control and elimination strategic planning," *Malar J*, vol. 11, no. 1, p. 205, 2012.

[18] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of The Royal Society Interface*, vol. 10, no. 81, p. 20120986, 2013.

[19] Kleinschmidt, I and Bagayoko, M and Clarke, GPY and Craig, M and Le Sueur, D, "A spatial statistical approach to malaria mapping," *International Journal of Epidemiology*, vol. 29, no. 2, pp. 355–361, 2000.

[20] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA, 2006.

[21] M. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," *Foundations and Trends in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.

[22] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2004, vol. 3.

[23] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. Buckee, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, 10 2012.