On the Challenges of Assimilating Data

Ricardo Andrade-Pacheco

University of Oxford

27 August 2015
Outline

- The problem we are working on
- The challenges we have found in our way
  - The ones we can live with
  - The ones we have to live with
- A prototype solution
About Malaria

Why is it important?

- Endemic in 100 countries.
- A threat for 3.3 billion people approximately.
- Among the leading causes of morbidity and mortality in Uganda.
About Malaria

Why is it important?

- Endemic in 100 countries.
- A threat for 3.3 billion people approximately.
- Among the leading causes of morbidity and mortality in Uganda.

Prevention

- Insecticide-treated mosquito nets
- Indoor residual spraying
- Chemoprevention (pregnant women, infants, seasonal)
About Malaria

Why is it important?

- Endemic in 100 countries.
- A threat for 3.3 billion people approximately.
- Among the leading causes of morbidity and mortality in Uganda.

Prevention

- Insecticide-treated mosquito nets
- Indoor residual spraying
- Chemoprevention (pregnant women, infants, seasonal)

Response

- Diagnostic testing
- Treatment
About Malaria

Why is it important?

- Endemic in 100 countries.
- A threat for 3.3 billion people approximately.
- Among the leading causes of morbidity and mortality in Uganda.

Prevention

- Insecticide-treated mosquito nets
- Indoor residual spraying
- Chemoprevention (pregnant women, infants, seasonal)

Response

- Diagnostic testing
- Treatment

The success of control and elimination policies depend on how well the disease can be anticipated and how fast the population reacts to it.
Data provided:

- Health facilities records across the whole country
- Number of individuals treated for malaria
- Weekly data aggregated by district
Initial Challenges

Change in districts boundaries definition.

Figure: Uganda 2003

Figure: Uganda 2015
Initial Challenges

Noise and errors.

Figure: Apac district (split into Apac, Oyam and Kole).
Initial Challenges

Variation in the number of reporting facilities.

Figure: Arua district (split into Arua, Koboko and Maracha).
Initial Challenges

Variation in the number of reporting facilities.

Figure: Arua district (split into Arua, Koboko and Maracha).
Initial Model

\[ \log y_i = f_{x_i} + \epsilon_i, \]

where

- \((f_{x_i}) \sim \mathcal{GP};\)
- \(\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)\) is a noise term with homogeneous variance.
Kernel Selection

Assumptions:

- The infection of malaria evolves with some degree of smoothness across time.

- The number of health facilities reporting has an effect on the incidence of malaria observed.
Kernel Selection

Figure: Kween district
A Model with Atypical Observations

\[ \log y_i = f_{x_i} + \epsilon_i + \zeta_i, \]

where

- \((f_{x_i}) \sim \mathcal{GP}\);
- \(\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)\) is a noise term with homogeneous variance;
- \(\zeta_i\) represents sources of variation not explained by the previous terms (e.g., reporting errors).
We expect reporting errors to occur only in a few observations, and these being characterized by

$$\epsilon_i + \zeta_i \gg \epsilon_i.$$  

If we assume that

$$z_i = (\log y_i, r_i)^\top - (\log y_{i-1}, r_{i-1})^\top \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma}),$$

for some $\hat{\mu}$ and $\hat{\Sigma}$, as the reporting errors are sparse, any point that contains a term $\zeta_i \neq 0$ will be unlikely under $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$. 

Outlier Detection

unlikely: any point outside the (rotated) ellipse $A$ centered on $\mu$ and with semi-axis defined by $3 \times \Sigma_{11}$ and $3 \times \Sigma_{22}$.

Figure: Kalungu district
If the points outside $A$ are atypical, an homogeneous model like

$$\log y_i = f_{x_i} + \epsilon_i,$$

should be outperformed by a model like

$$\log y_i = f_{x_i} + \epsilon_i \mathbb{I}_{\{z_i \in A\}} + \delta_i \mathbb{I}_{\{z_i \notin A\}},$$

where $\delta_i \sim \mathcal{N}(0, \sigma^2_{\delta_i})$ has heterogeneous variance across observations, such that $\sigma^2_{\delta_i} > \sigma^2_{\epsilon}$. 
Outlier Detection

Figure: Heteroscedastic example
Outlier Detection

Figure: Nwoya district
Harmonization

Figure: Iganga district (split into Iganga, Namutumba and Luuka).
Harmonization

Figure: Iganga district (split into Iganga, Namutumba and Luuka).

What to do?
A trivial model that explains the data of a district $A$ just as a constant mean plus noise, i.e.,

$$y_{Ai} = \gamma_A + \epsilon_{Ai},$$

can be defined as a GP with a kernel of the form

$$K(x_i, x_j) = \gamma_A^2.$$
Harmonization

Say district $A$ is split into $A'$ and $B$.

If there are no biases in the measurements of any of the districts, then it should be satisfied that

$$\gamma_A = \gamma_{A'} + \gamma_B.$$
Harmonization

Say district $A$ is split into $A'$ and $B$.

If there are no biases in the measurements of any of the districts, then it should be satisfied that

$$\gamma_A = \gamma_{A'} + \gamma_B.$$  

Let $y_i = (y_{Ai}, y_{A'i}, y_{Bi})^\top$, the corresponding kernel that satisfies having a *nested-mean* is defined as

$$\Gamma(x_i, x_j) = [\gamma_{j(i)} \gamma_{j(j)}],$$

where $j(\cdot)$ is an index associated to any of the districts.
Harmonization

We can achieve a less noisy and consistent time series.

**Figure**: Signals estimated using a composed kernel
Harmonization

We can achieve a less noisy and consistent time series.

Figure: Signals estimated using a composed kernel

... but there is a smallprint.
Important considerations

- Total number of health facilities is not clear
- Unknown coverage of HMIS data against total population
- Patients treated for malaria are not usually diagnosed with a test
Important considerations

- Total number of health facilities is not clear
- Unknown coverage of HMIS data against total population
- Patients treated for malaria are not usually diagnosed with a test

... so, we do not have any means to assess the accuracy of our estimates.
How to use what we have gained so far?
How to use what we have gained so far?

How to contribute given the current circumstances?
Say we have an observed output $y$, we can model it as

$$y = f_x + \epsilon,$$

where $f_x \sim \mathcal{GP}(\mathcal{M}, K)$. 

> Back to the Basic Model

> Say we have an observed output $y$, we can model it as

$$y = f_x + \epsilon,$$

where $f_x \sim \mathcal{GP}(\mathcal{M}, K)$. 

> By combining different covariance kernels, a GP is able to describe complex functions.

> Each of the individual kernels contributes by encoding a specific set of properties or pattern of the resulting function.
Back to the Basic Model

Say we have an observed output $y$, we can model it as

$$y = f_x + \epsilon,$$

where $f_x \sim \mathcal{GP}(\mathcal{M}, K)$.

$K$ defines the dependance structure

- By combining different covariance kernels, a GP is able to describe complex functions.
- Each of the individual kernels contributes by encoding a specific set of properties or pattern of the resulting function.
Signal decomposition with Gaussian processes

Figure: Two independent signals
Signal decomposition with Gaussian processes

Figure: Combined signal with noise
Signal decomposition with Gaussian processes

Figure: Signal estimation with a composed kernel
Signal decomposition with Gaussian processes
Signal decomposition with Gaussian processes
If \( f_x \sim \mathcal{GP}(0, K) \), then

\[
\begin{bmatrix}
  f_x \\
  \frac{\partial f_x}{\partial x}
\end{bmatrix} \sim \mathcal{GP}(0, \Gamma),
\]

where

\[
\Gamma = \begin{bmatrix}
  K & \frac{\partial}{\partial x} K \\
  \frac{\partial}{\partial x} K & \frac{\partial^2}{\partial x^2} K
\end{bmatrix}.
\]
Large-scale signal derivative

Figure: Signals estimated using a composed kernel
Short-scale signal derivative

Figure: Signals estimated using a composed kernel
Figure: Short and long-scale variations in HMIS data
Warning system

Figure: Classification of variations around the long-term trend
Figure: Short variations in HMIS data
Uganda’s monitor

http://ric70x7.github.io/research.html
Next steps

- Further research is needed to explore the benefits of this model in practice.

- Future plans to implement it with other diseases.
Final remark

- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.
Final remark

- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.

- We have passed the stage where we were able to provide just a mean and variance estimate, and now are able to provide density functions estimates.
Final remark

- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.

- We have passed the stage where we were able to provide just a mean and variance estimate, and now are able to provide density functions estimates.

- But these are complex outputs that need to be synthesized for different users.
Final remark

- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.

- We have passed the stage where we were able to provide just a mean and variance estimate, and now are able to provide density functions estimates.

- But these are complex outputs that need to be synthesized for different users.

- The monitoring system proposed tries to achieve that goal.
Collaborators

Martin Mubangizi
College of Computing and Information Science
Makerere University
Uganda

John Quinn
UN Global Pulse
Pulse Lab Kampala
Uganda

Neil Lawrence
Department of Computer Science
University of Sheffield
UK


