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

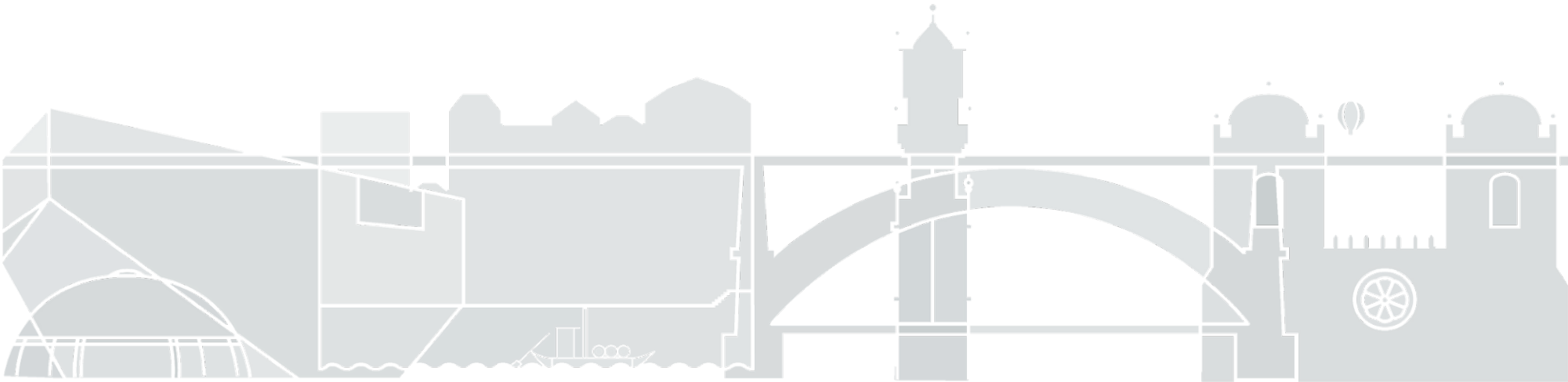
Ricardo Andrade-Pacheco, Martin Mubangizi,
John Quinn, Neil Lawrence

ECML/PKDD 2015, Workshop on
Advanced Analytics and Learning on Temporal Data

11 September 2015

Outline

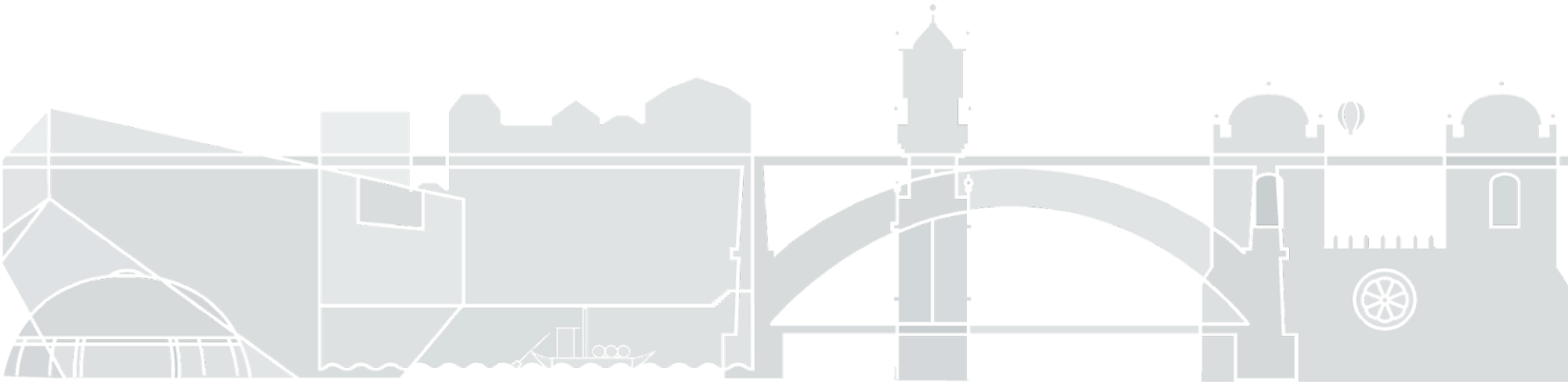
- The problem we are working on
- The challenges we have found in our way
- The solution proposed



About Malaria

Why we care?

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



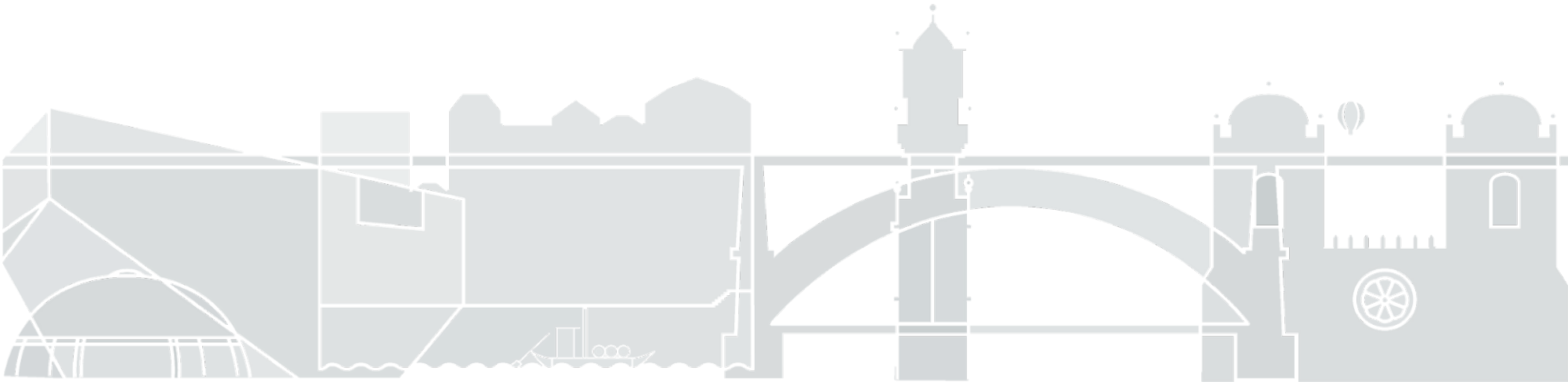
About Malaria

Why we care?

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

Can be prevented

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



About Malaria

Why we care?

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

Can be prevented

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

Reaction mechanisms available

- Diagnostic testing
- Treatment

About Malaria

Why we care?

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

Can be prevented

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

Reaction mechanisms available

- Diagnostic testing
- Treatment

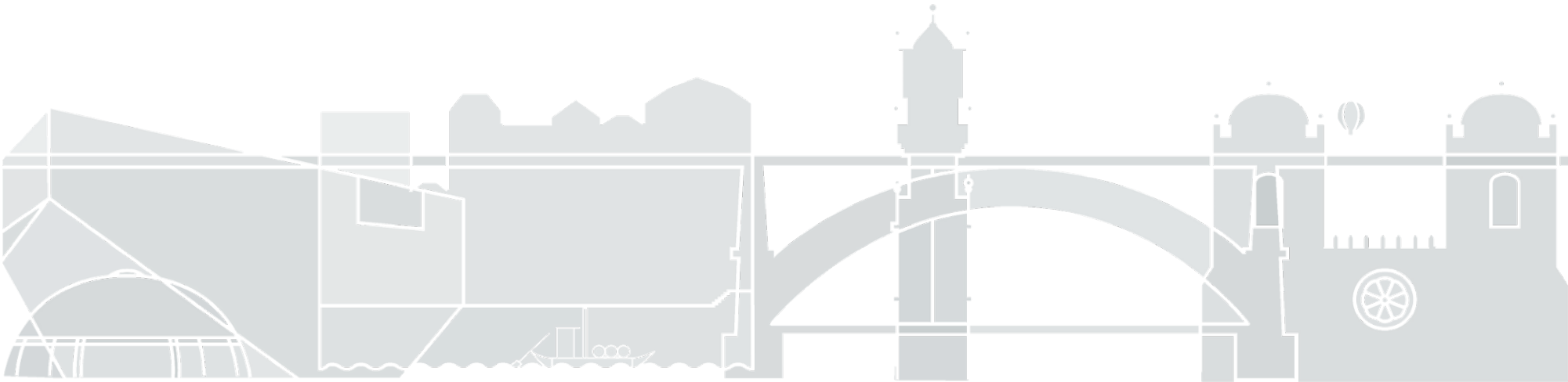
We need information to anticipate the disease and respond.

Health Management Information System (HMIS)

The Ministry of Health shared some of its information with us.

Data provided:

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



Challenges

Noise and errors

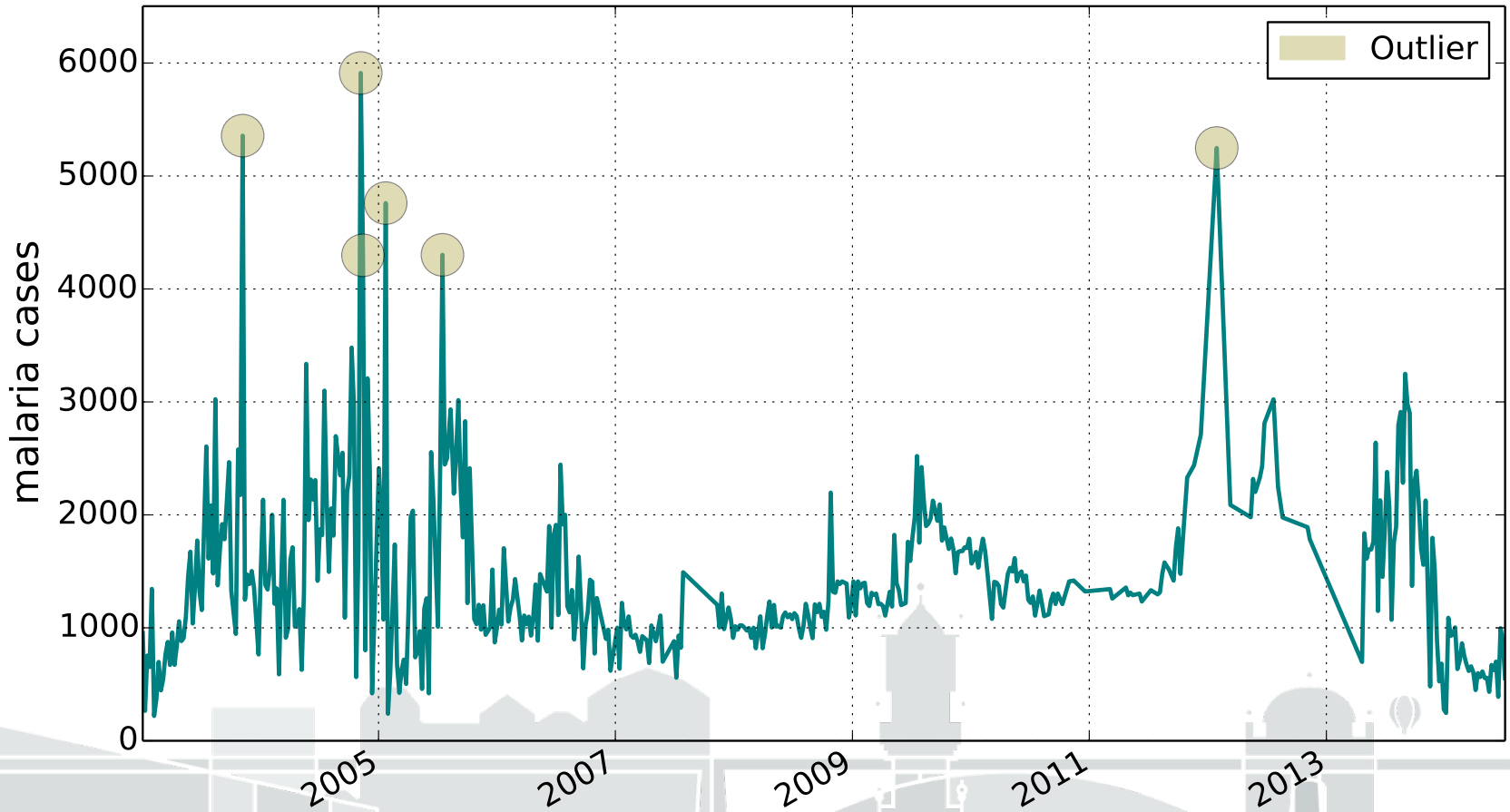


Figure : Kotido district

Challenges

Variation in the number of reporting facilities.

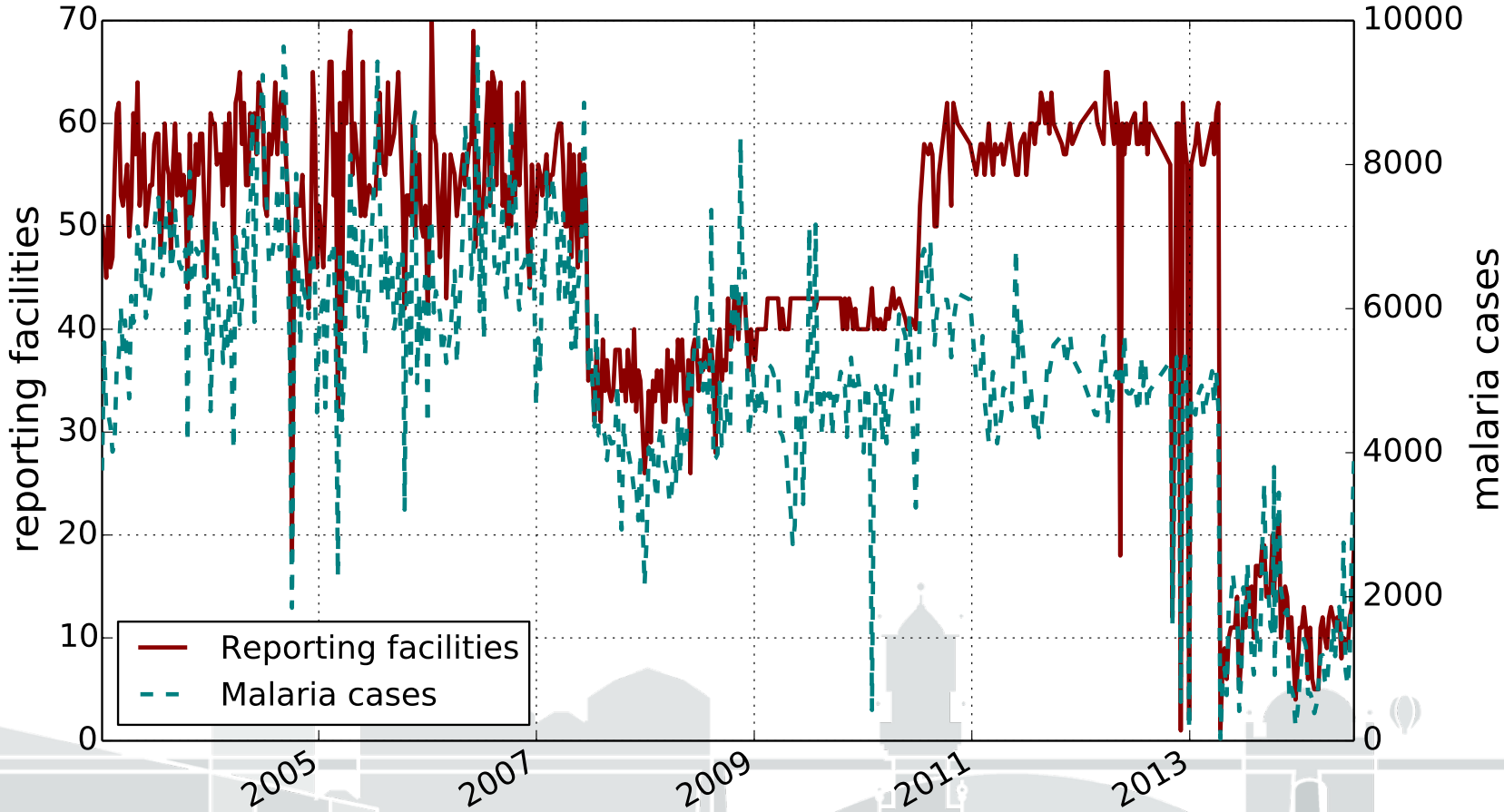


Figure : Arua district

Challenges

Change in districts boundaries definition.

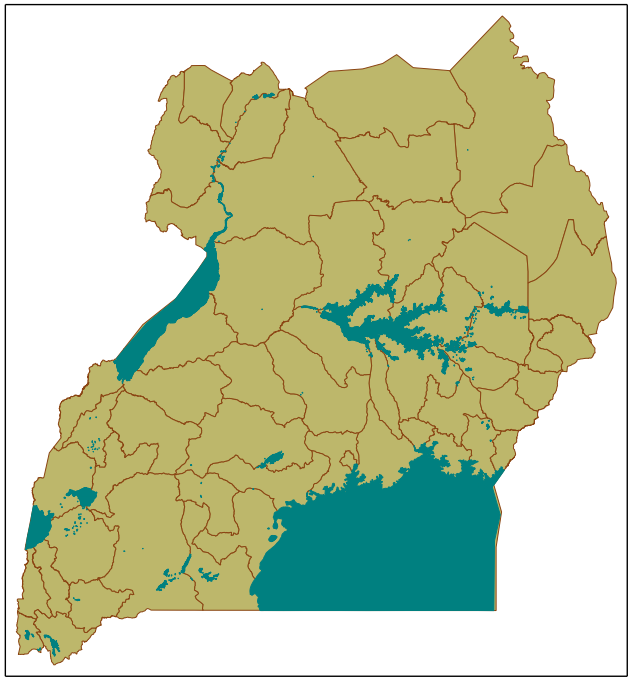


Figure : 2003 (56 districts)

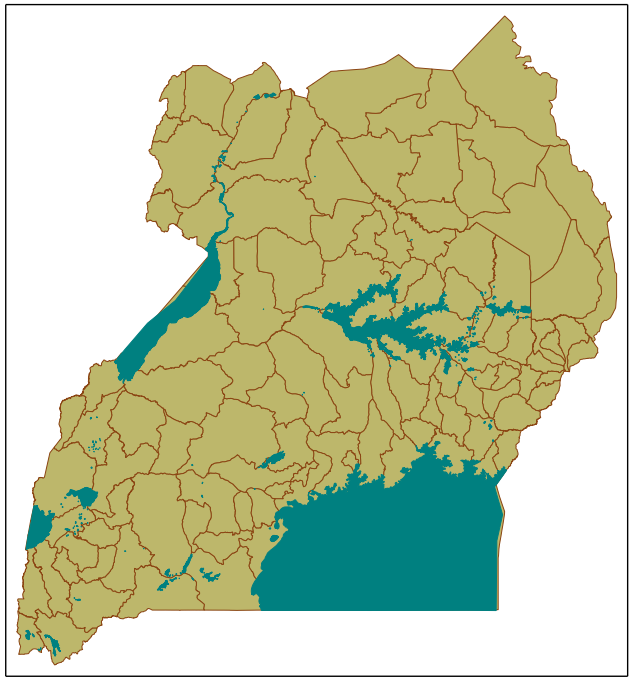
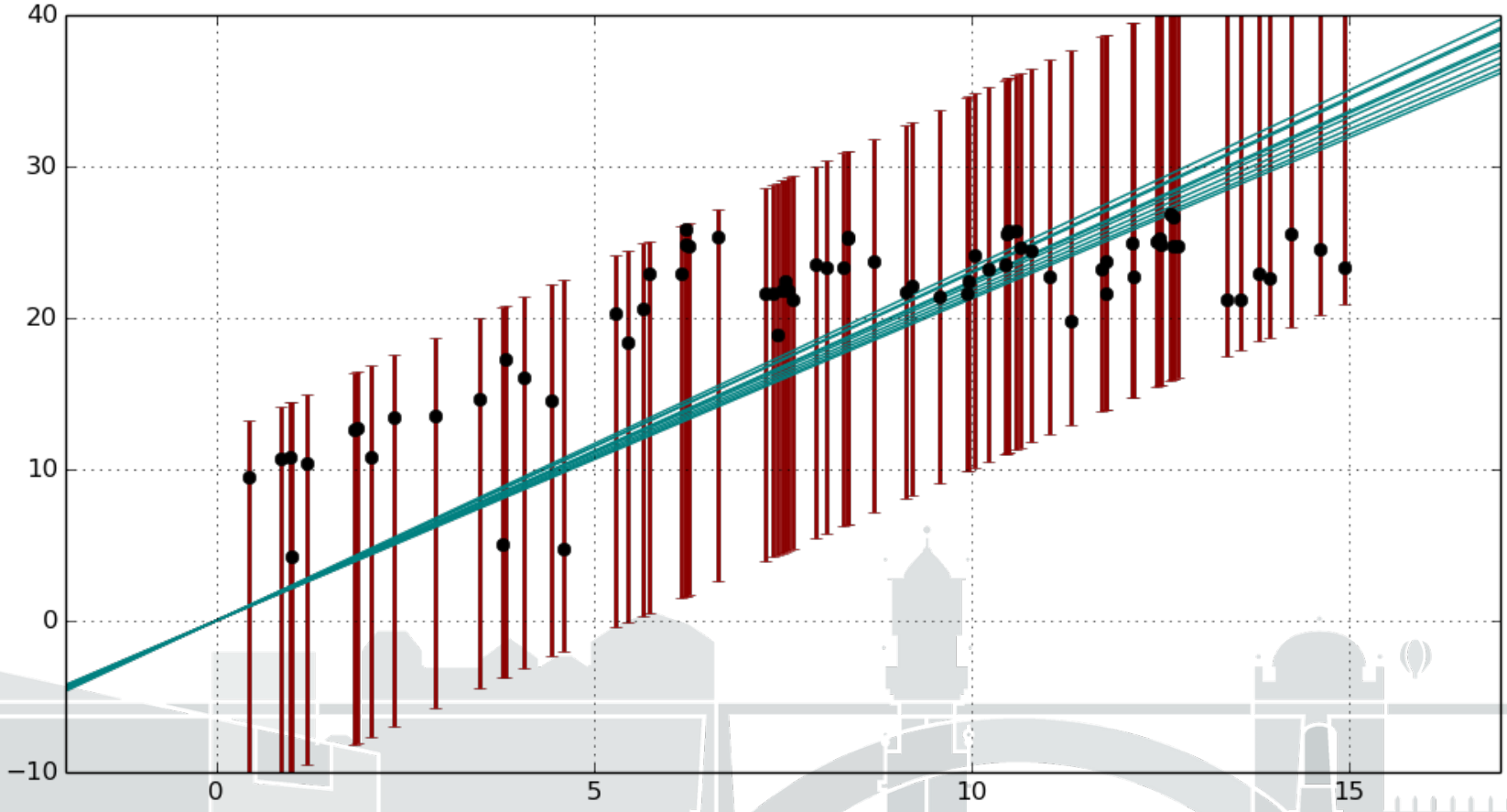


Figure : 2015 (112 districts)

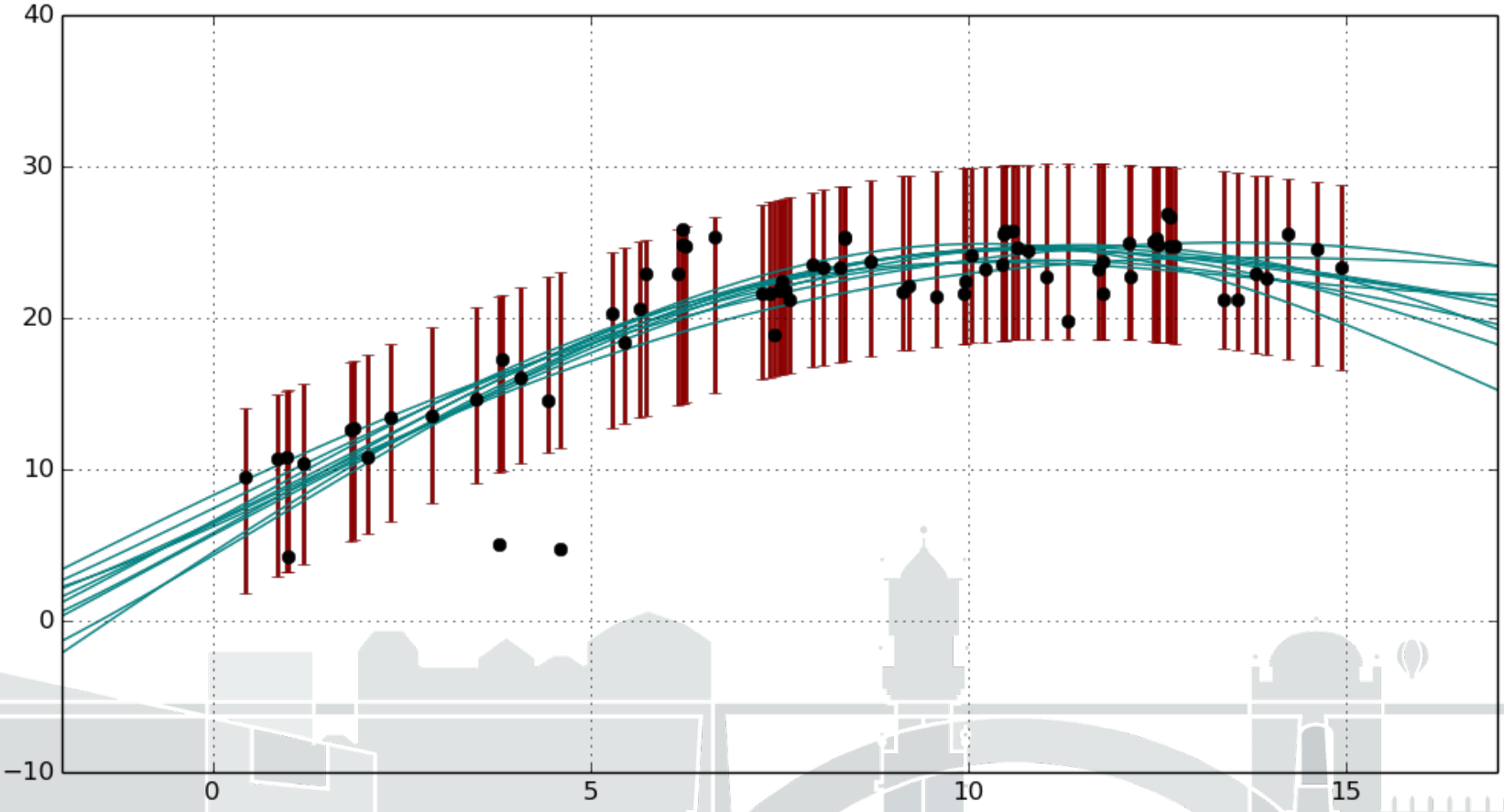
Gaussian processes

A wide range of functions can be learnt through the use of different covariance functions and noise assumptions.



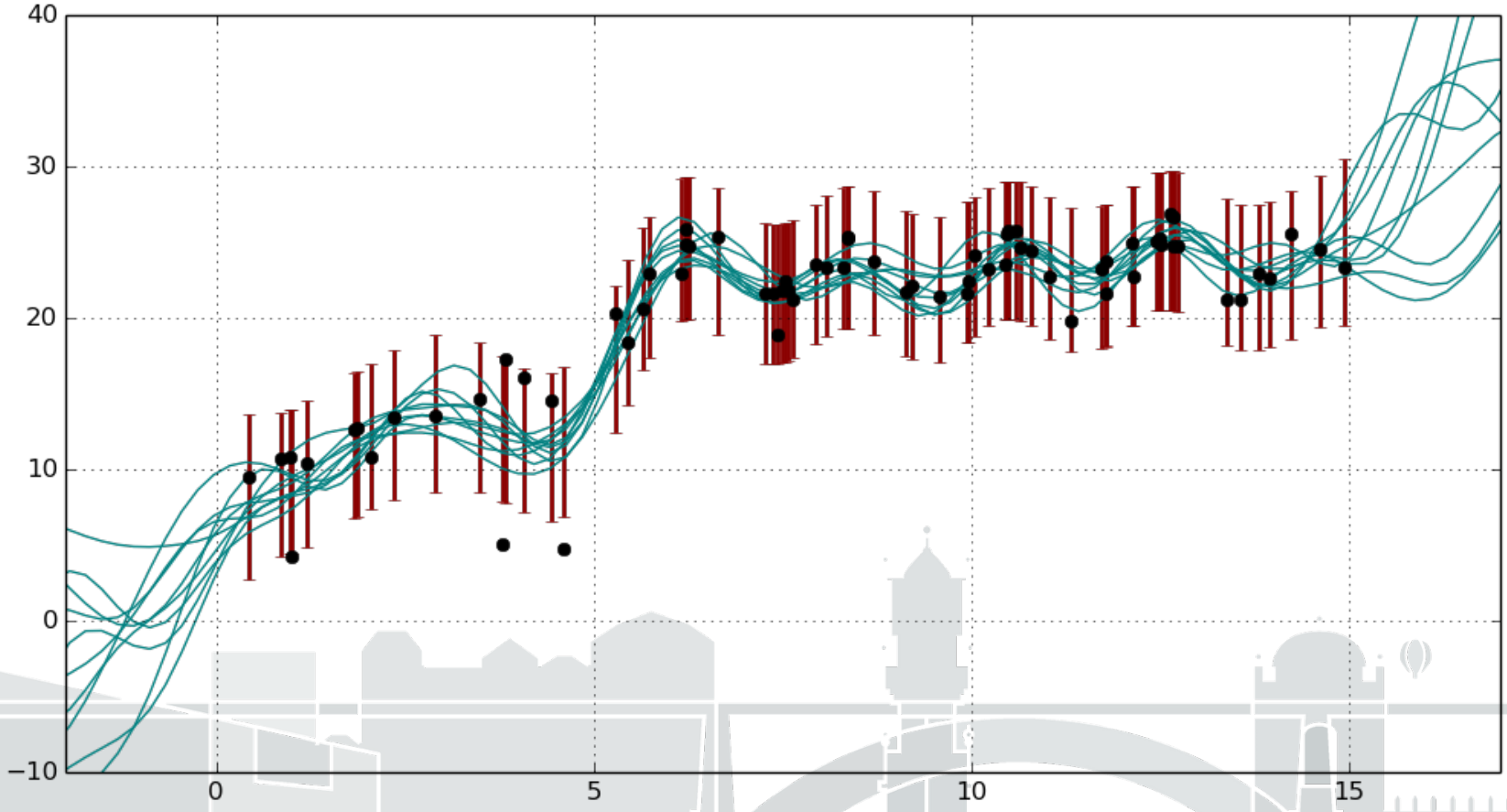
Gaussian processes

A wide range of functions can be learnt through the use of different covariance functions and noise assumptions.



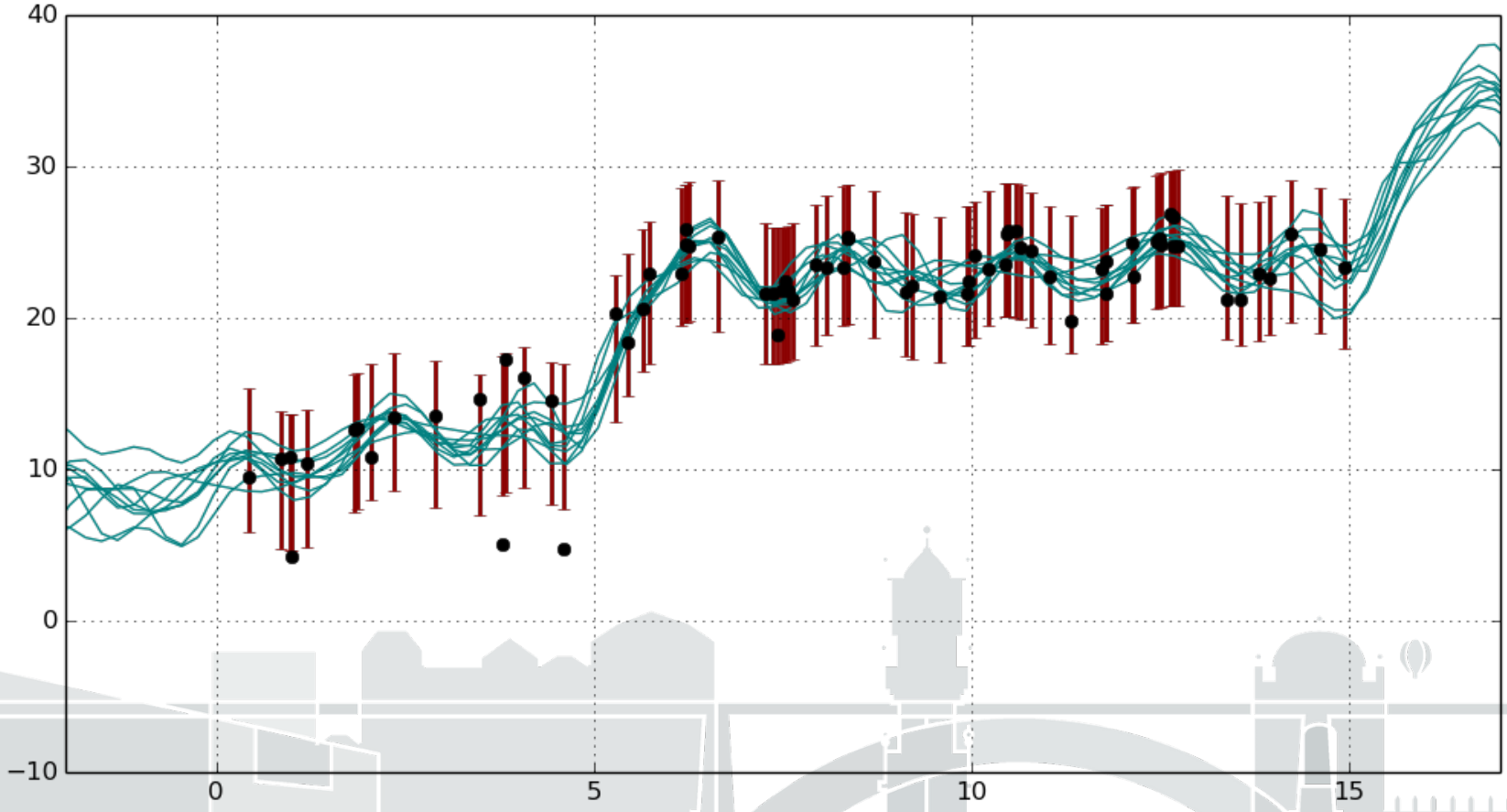
Gaussian processes

A wide range of functions can be learnt through the use of different covariance functions and noise assumptions.



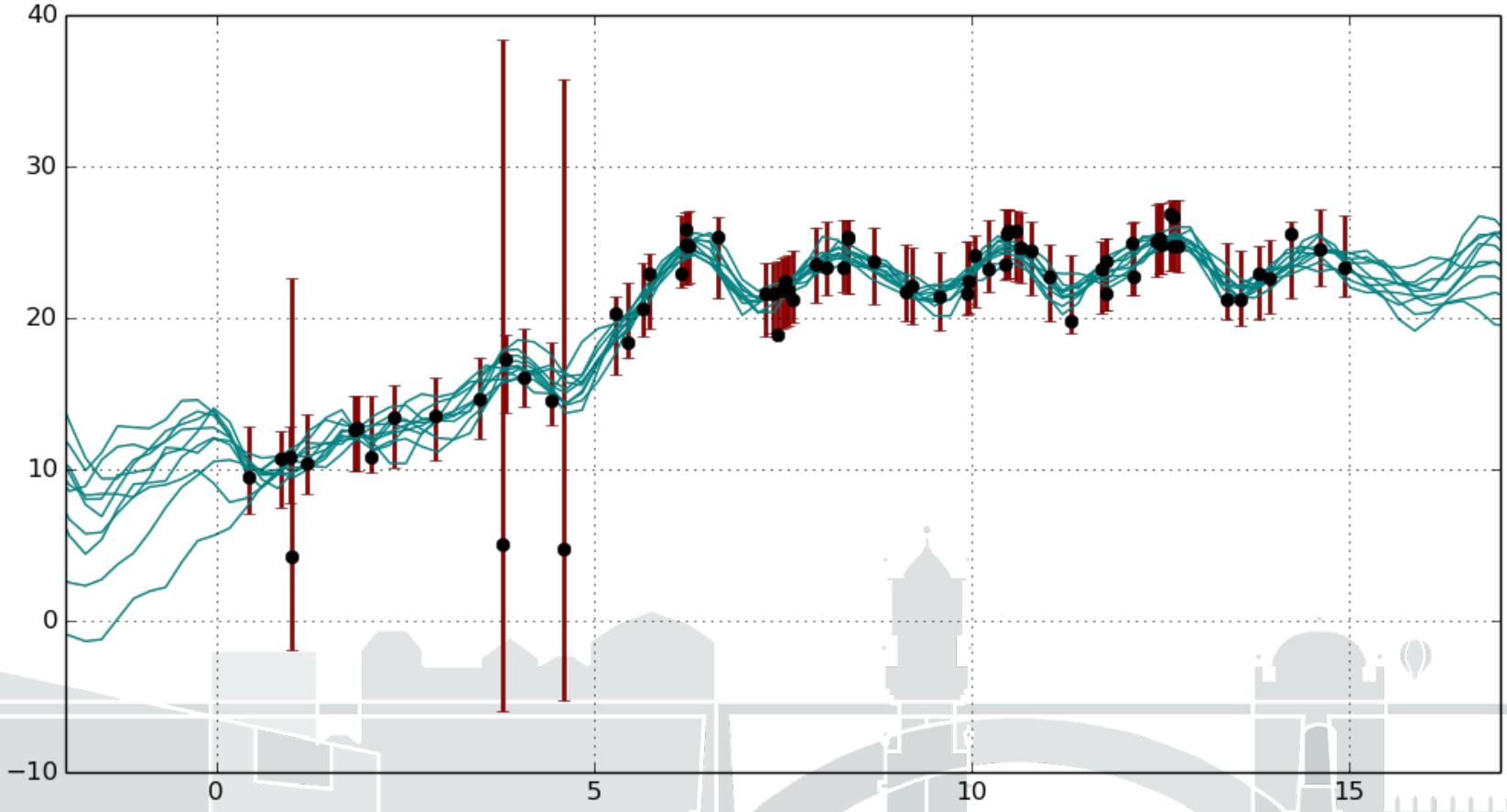
Gaussian processes

A wide range of functions can be learnt through the use of different covariance functions and noise assumptions.



Gaussian processes

A wide range of functions can be learnt through the use of different covariance functions and noise assumptions.



Gaussian processes

We can achieve a less noisy and consistent time series.

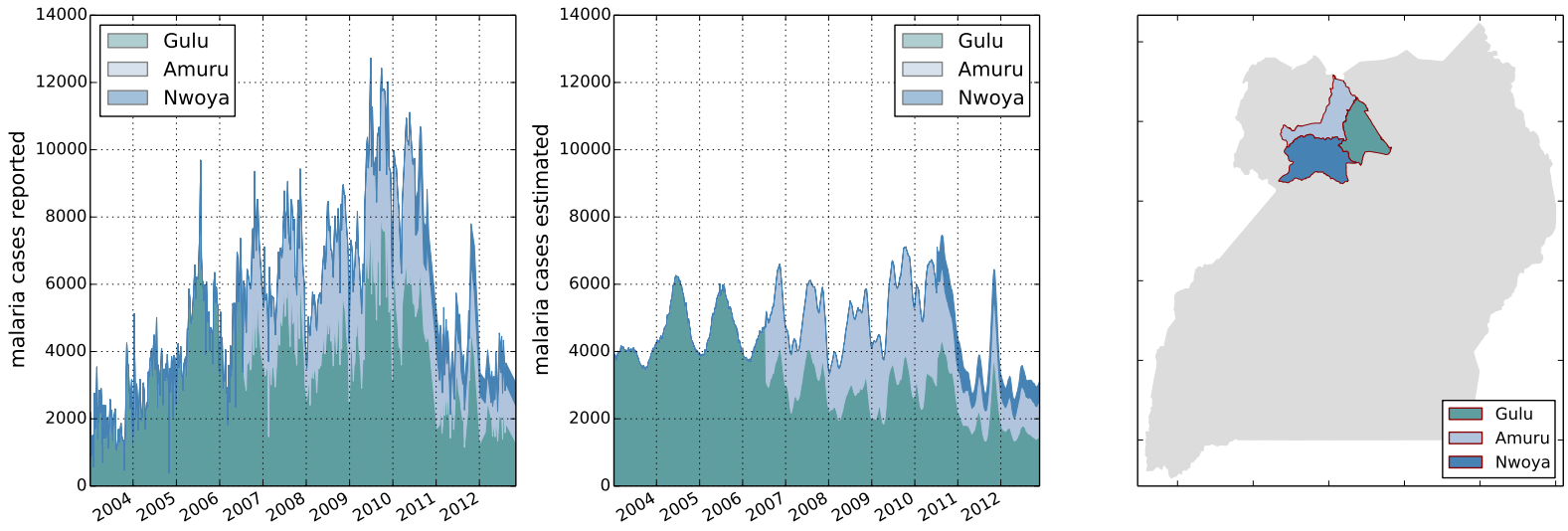


Figure : Signals estimated using a composed kernel

Gaussian processes

We can achieve a less noisy and consistent time series.

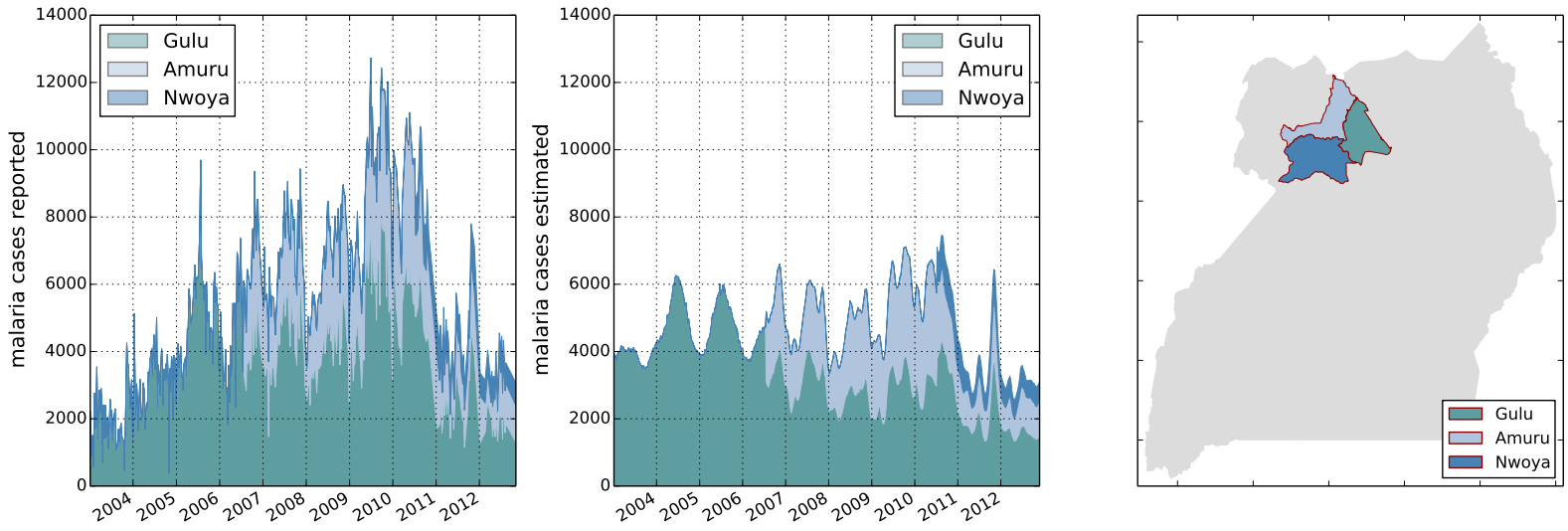
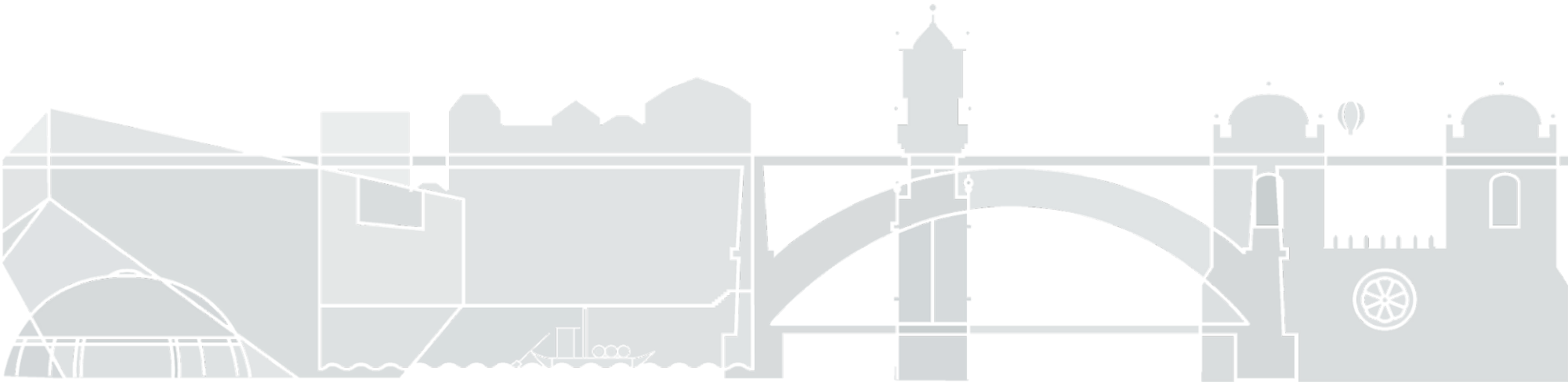


Figure : Signals estimated using a composed kernel

... but there is a smallprint.

Important considerations

- Unknown coverage of HMIS data against total population
- Total number of health facilities is not clear
- Being treated for malaria $\neq$ having malaria



Important considerations

- Unknown coverage of HMIS data against total population
- Total number of health facilities is not clear
- Being treated for malaria $\neq$ having malaria

... so, we do not have means to assess the accuracy of our estimates.

- How to contribute given the current circumstances?

Gaussian process basics

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

$$y = f_{\mathbf{x}} + \epsilon$$

where $f_{\mathbf{x}} \sim \mathcal{GP}(\mathcal{M}, K)$.

The kernel function K defines the dependence structure of the process.

Signal decomposition with Gaussian processes

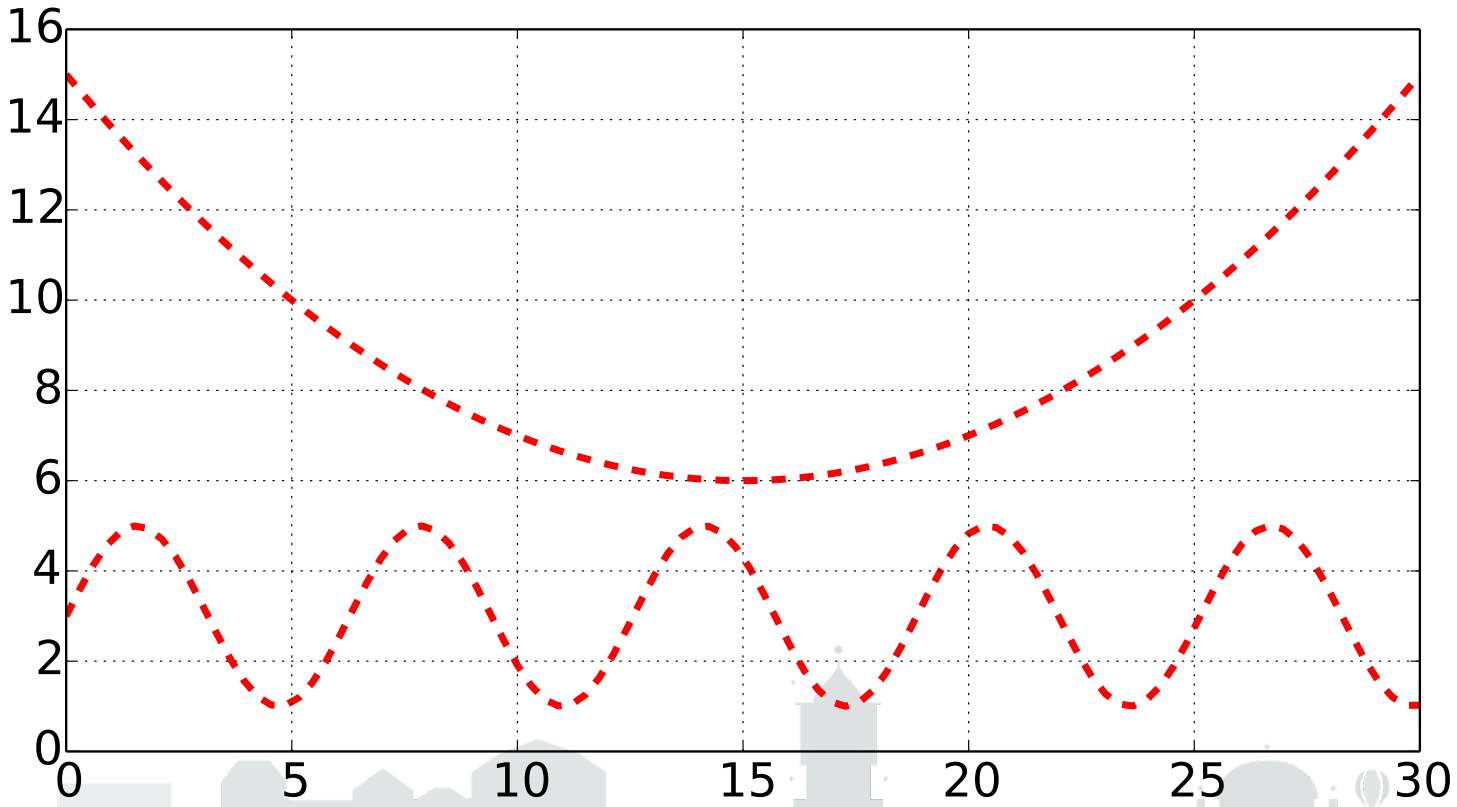


Figure : Two independent signals

Signal decomposition with Gaussian processes

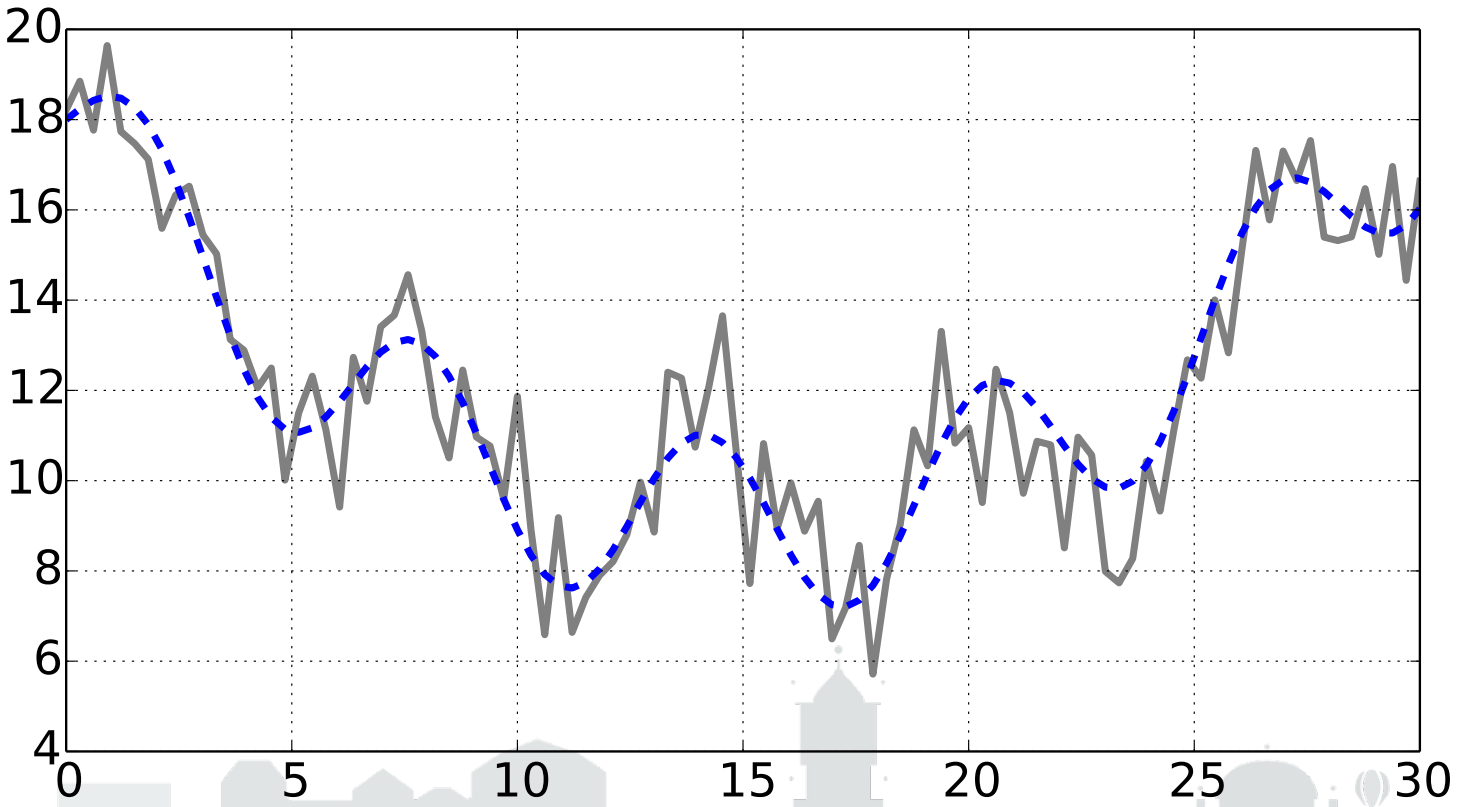


Figure : Combined signal with noise

Signal decomposition with Gaussian processes

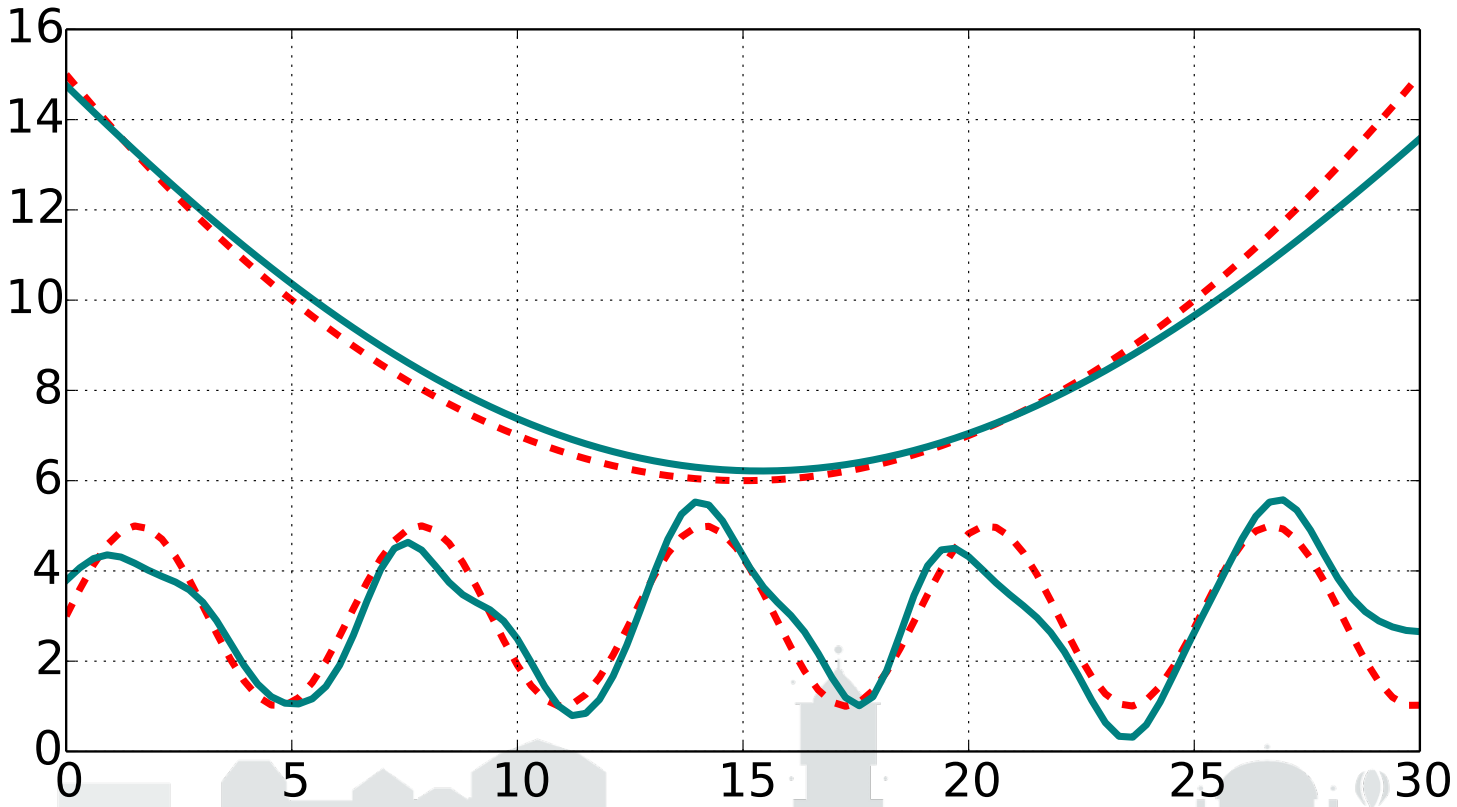
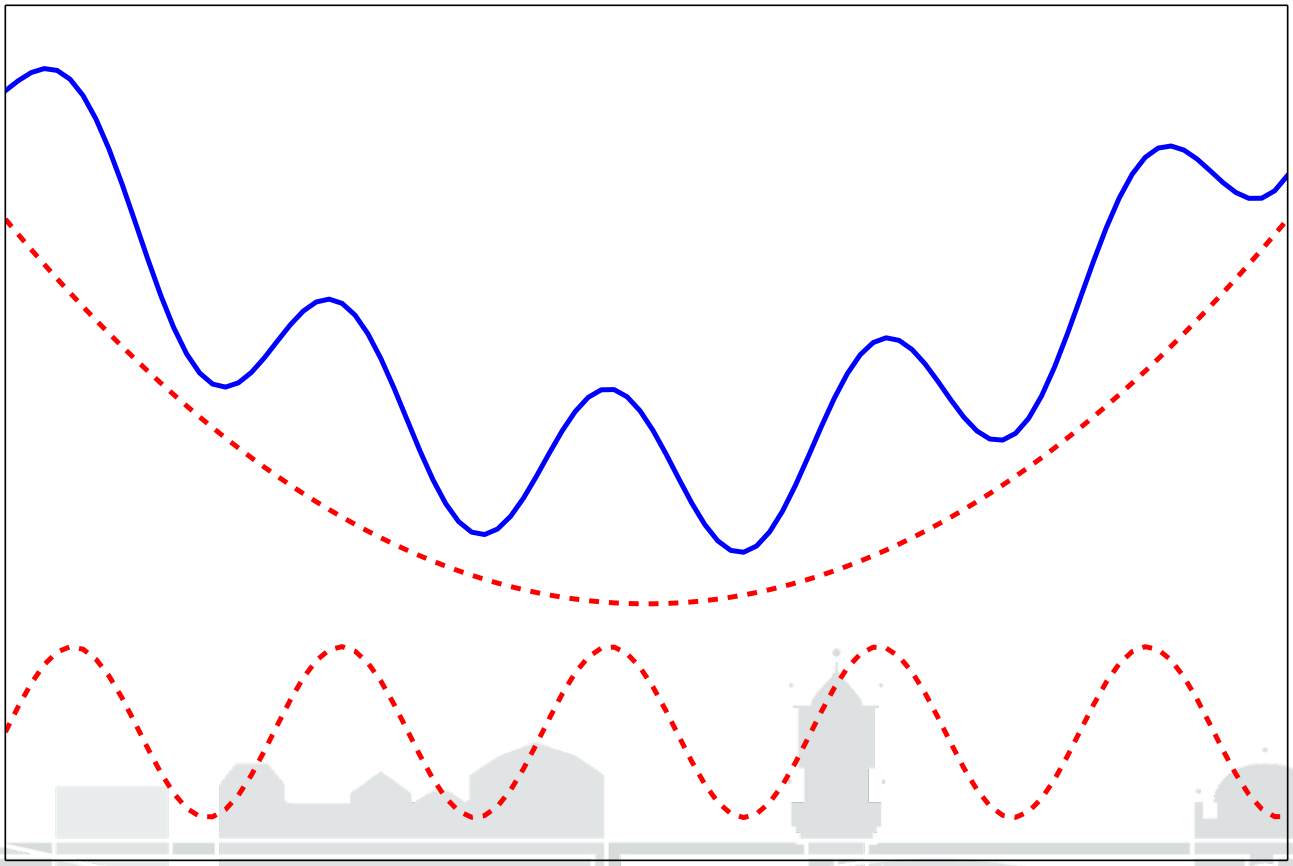
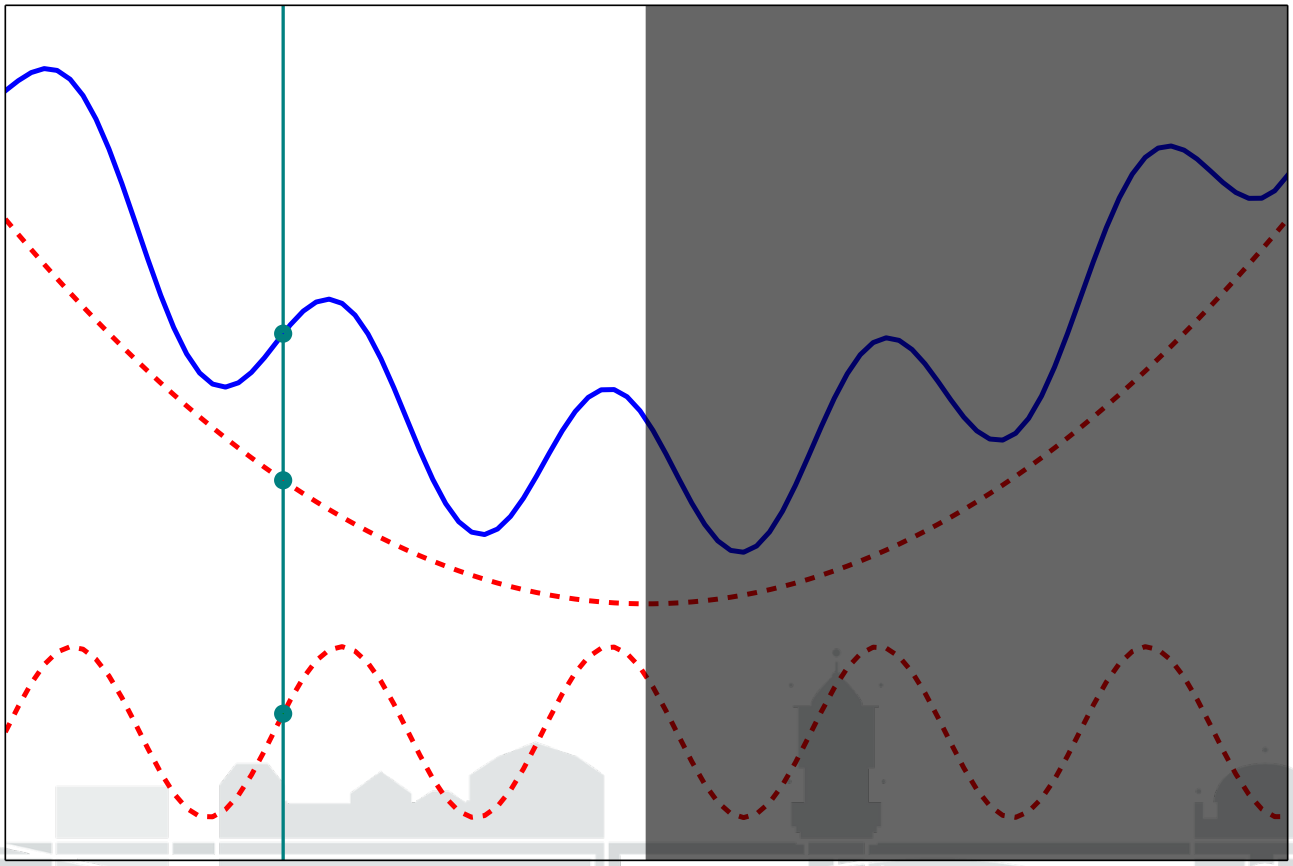


Figure : Signal estimation with a composed kernel

Signal decomposition with Gaussian processes



Signal decomposition with Gaussian processes



Simple computation of the derivative of each signal

If $f_{\mathbf{x}} \sim \mathcal{GP}(0, K)$, then

$$\begin{bmatrix} f_{\mathbf{x}} \\ \frac{\partial f_{\mathbf{x}}}{\partial \mathbf{x}} \end{bmatrix} \sim \mathcal{GP}(\mathbf{0}, \Gamma)$$

where

$$\Gamma = \begin{bmatrix} K & \frac{\partial}{\partial \mathbf{x}} K \\ \frac{\partial}{\partial \mathbf{x}} K & \frac{\partial^2}{\partial \mathbf{x}^2} K \end{bmatrix}$$

Large-scale signal derivative

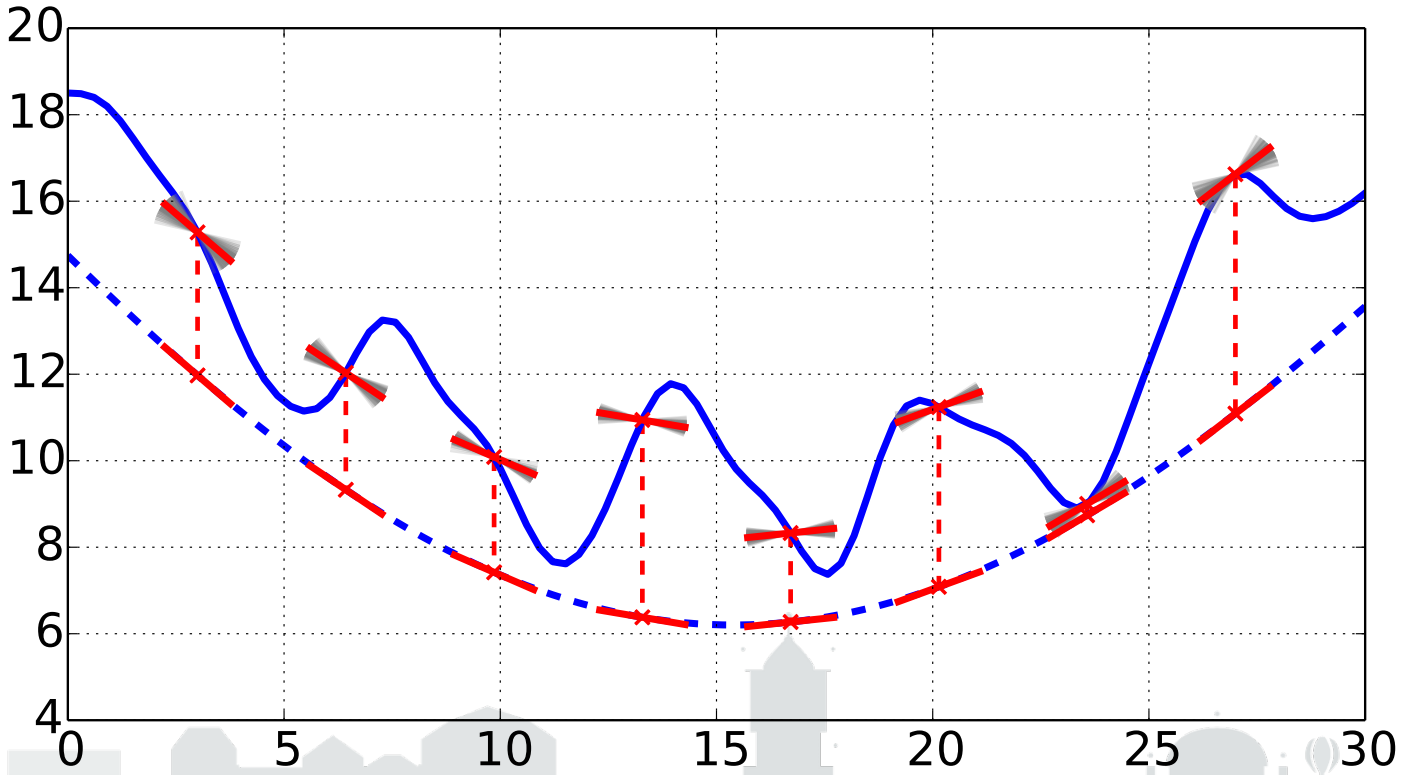


Figure : Signals estimated using a composed kernel

Short-scale signal derivative

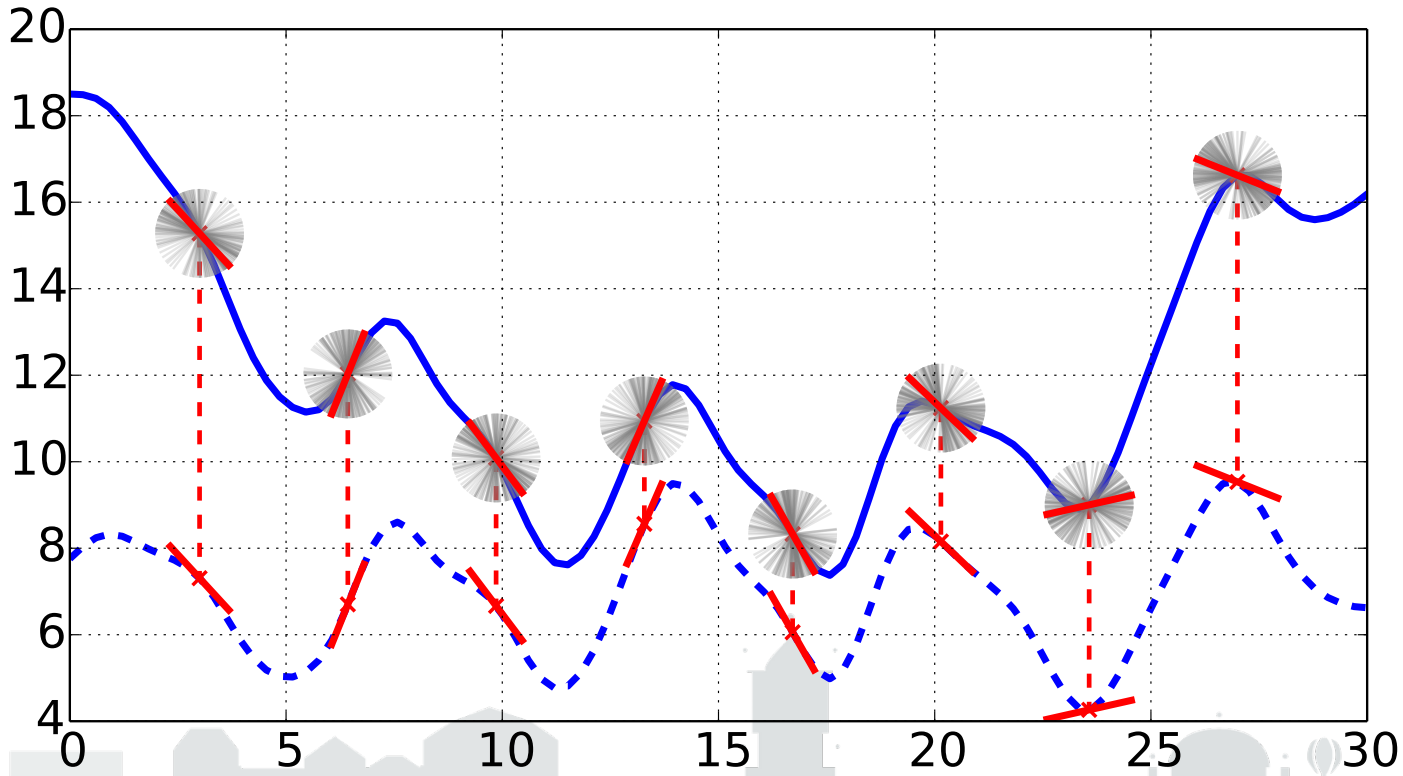


Figure : Signals estimated using a composed kernel

Back to our problem

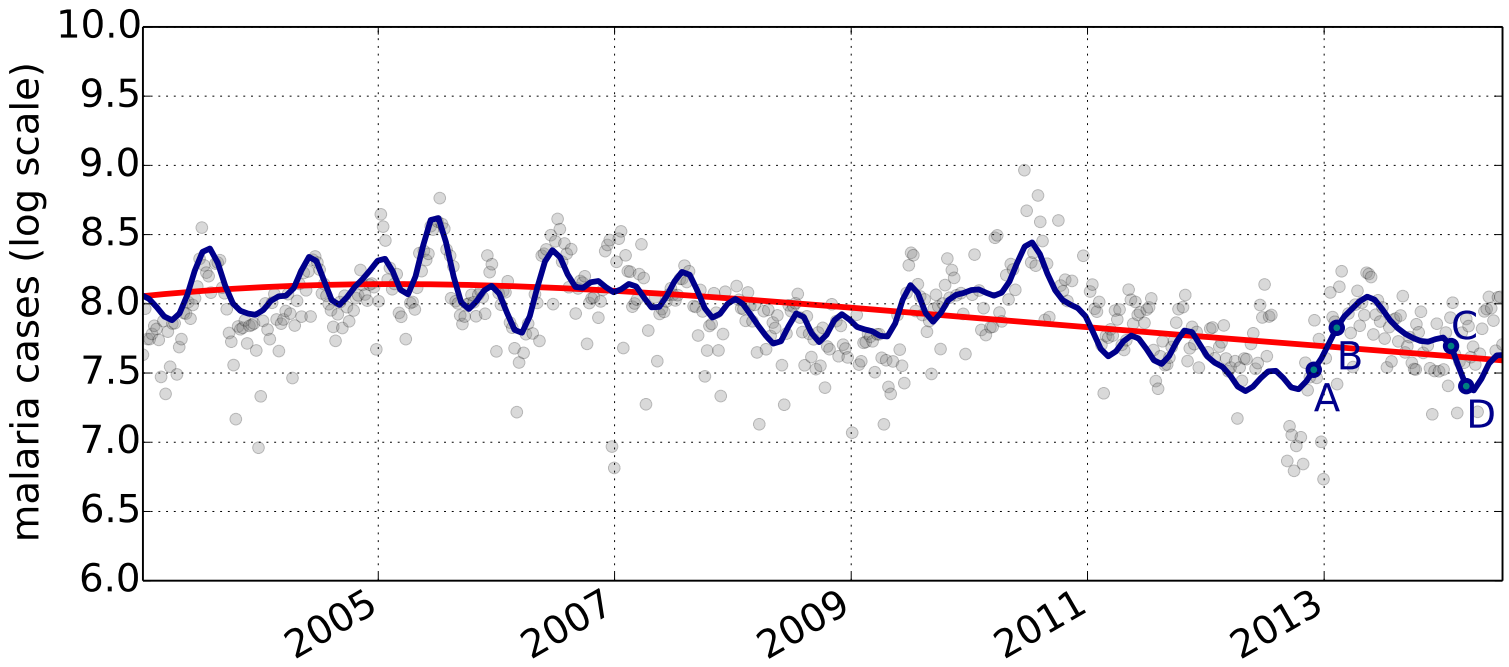


Figure : Short and long-scale variations in HMIS data

Warning system

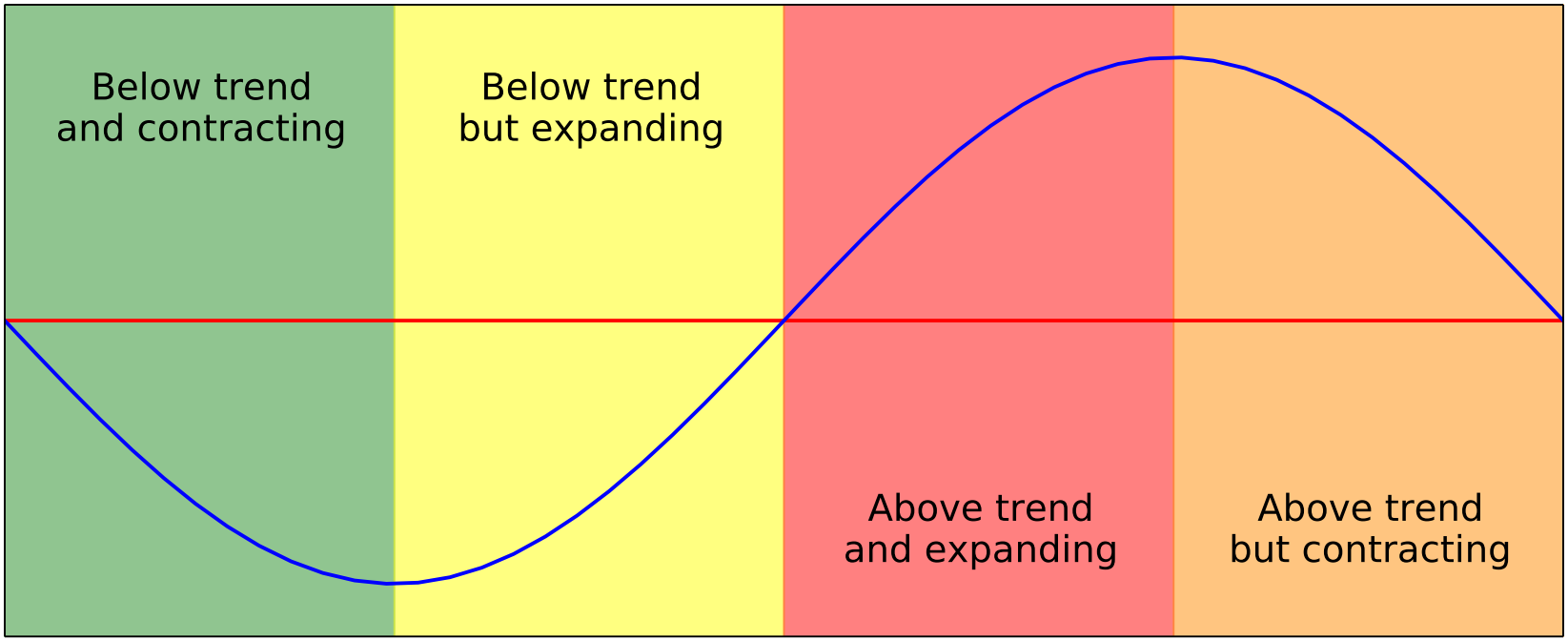
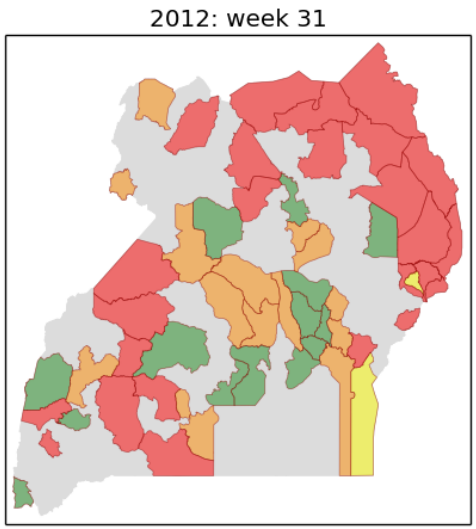
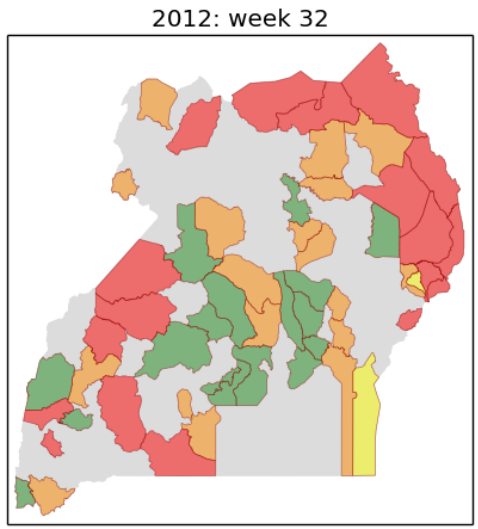


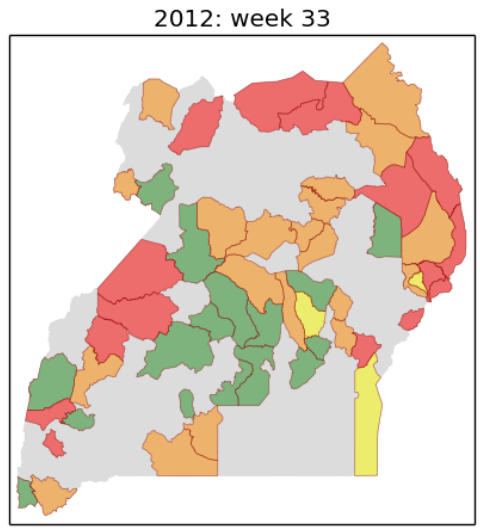
Figure : Classification of variations around the long-term trend

Uganda's monitor

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

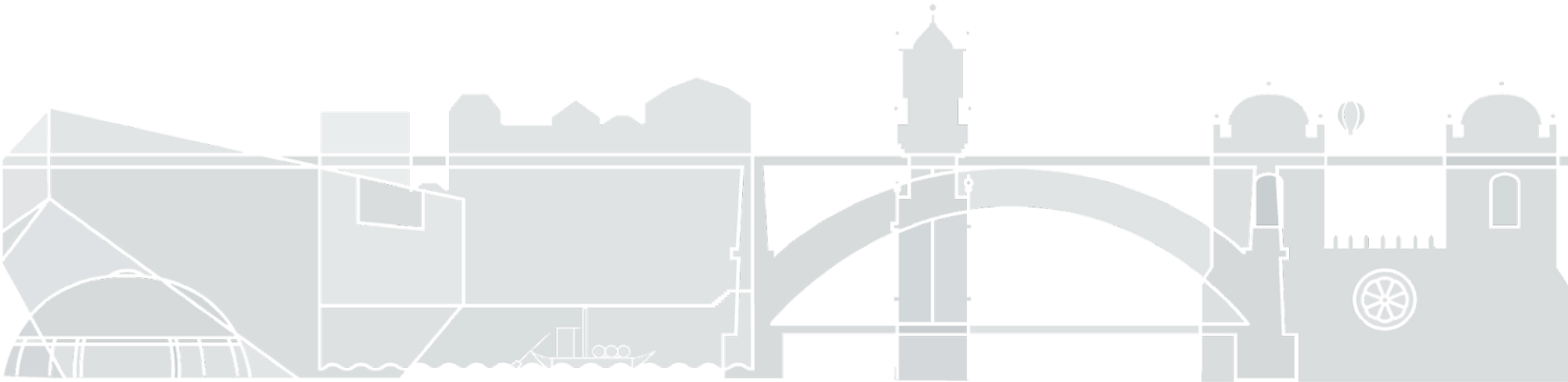






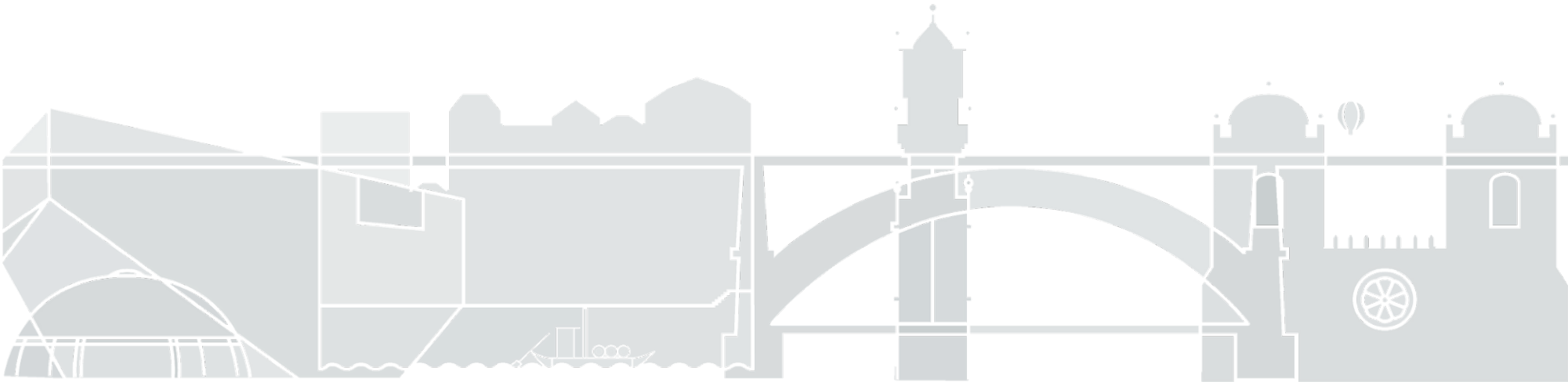
Next steps

- Further research is needed to explore the benefits of this model in practice.
- Future plans to implement it with other diseases.



Final remark

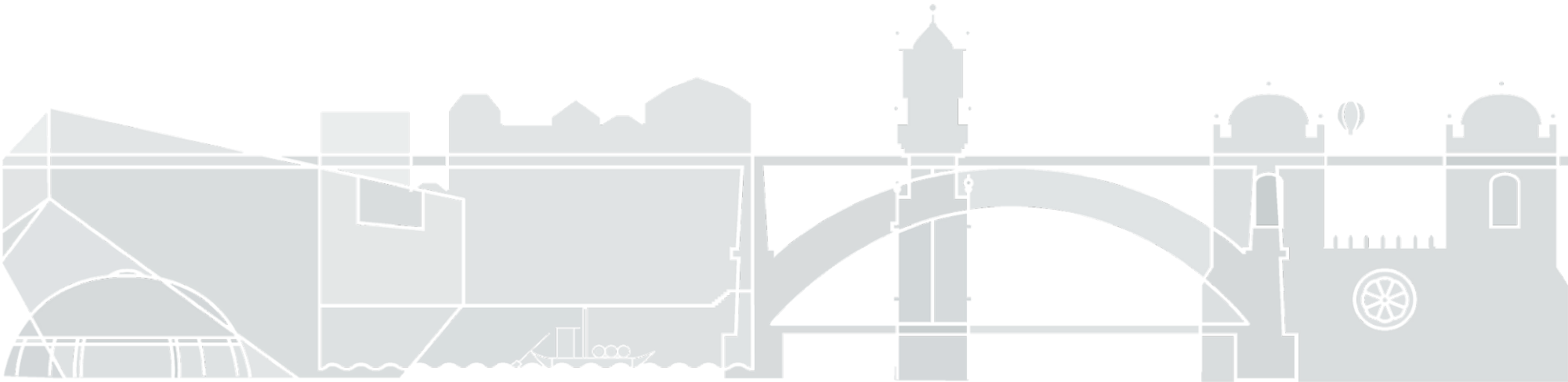
- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.



Final remark

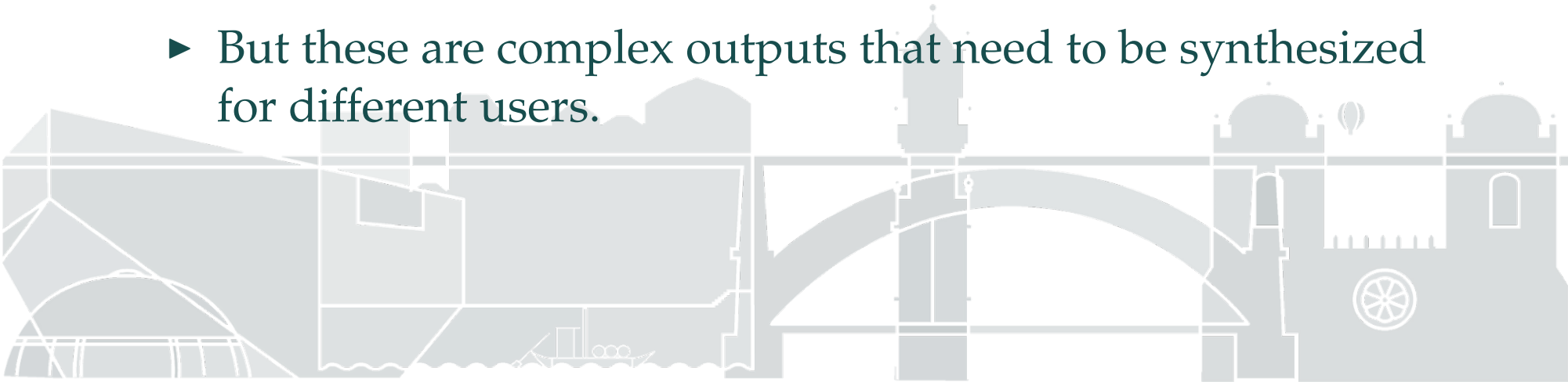
- Perhaps one of the most important challenges of statistics will always be to communicate with domain-oriented sciences and planners from different sectors.

- 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.

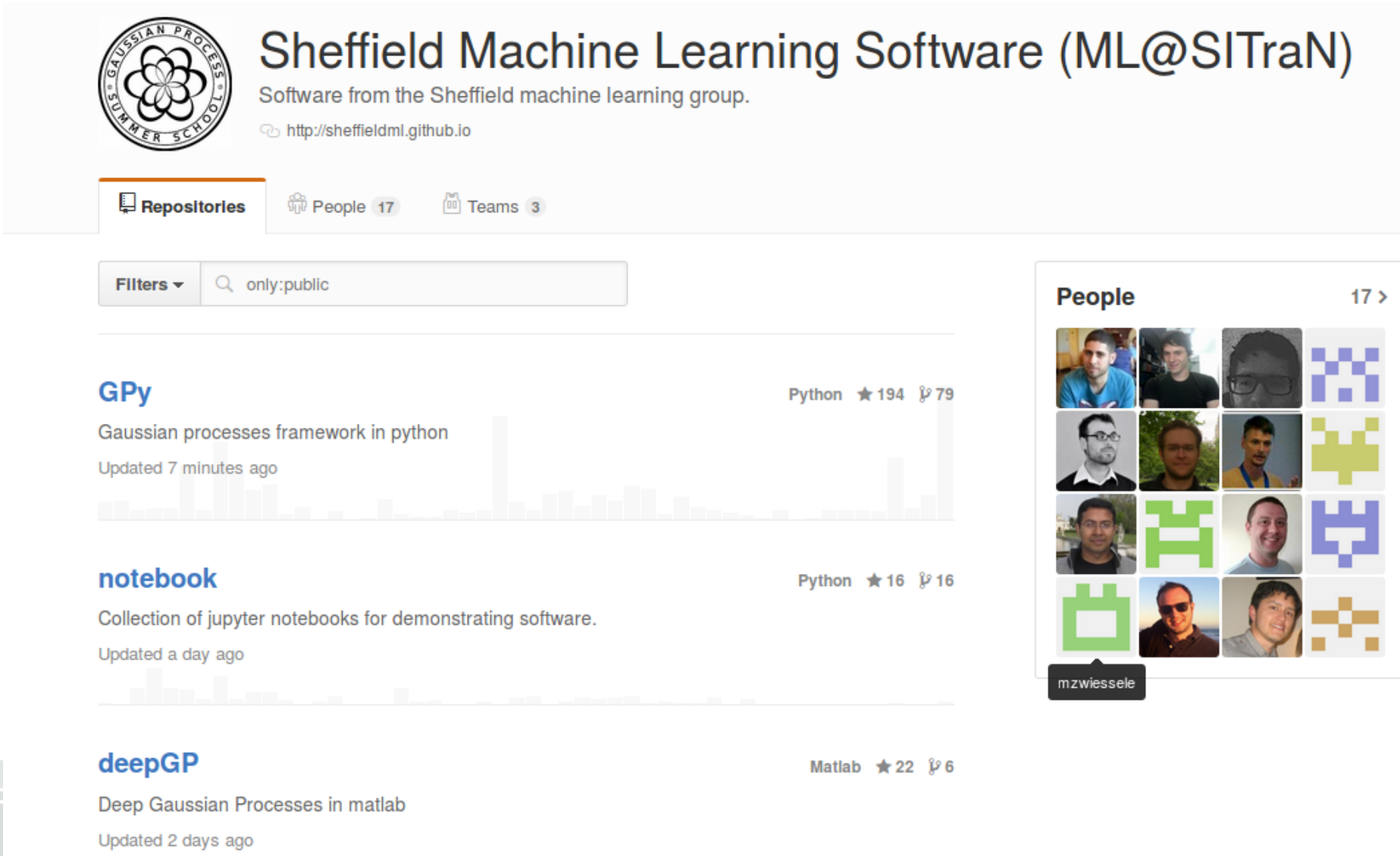


Final remark

- 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.



All our models were implemented using GPy



- https://github.com/SheffieldML/GPy

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

References I

[1] Luca Baldassarre, Lorenzo Rosasco, Annalisa Barla, and Alessandro Verri. Multi-output learning via spectral filtering. *Machine Learning*, 87(3):259–301, 2012.

[2] Marianne Baxter and Robert G King. Measuring business cycles: approximate band-pass filters for economic time series. *Review of economics and statistics*, 81(4):575–593, 1999.

[3] William P Cleveland and George 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.

[4] Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D Lawrence. Gaussian process models for periodicity detection. *arXiv preprint arXiv:1303.7090*, 2013.

[5] Simon I. Hay, Robert W. Snow, and David J. Rogers. From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. *Parasitology Today*, 14(8):306–313, 1998.

[6] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[7] Georges Matheron. Pour une analyse krigeante de donnés régionalisées. Technical report, École des Mines de Paris, Fontainebleau, France, 1982.

[8] Charles A. Micchelli and Massimiliano Pontil. Kernels for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2004.

References II

[9] Charles A. Micchelli and Massimiliano Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

[10] Donald E. Myers. Matrix formulation of co-Kriging. *Journal of the International Association for Mathematical Geology*, 14(3):249–257, 1982.

[11] H. Quenouille. *The analysis of multiple time-series*. Griffin's statistical monographs & courses. Griffin, 1957.

[12] Simo 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.

[13] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.

[14] Floris van Ruth, Barry Schouten, and Roberto Wekker. The statistics Netherlands business cycle tracer. Methodological aspects; concept, cycle computation and indicator selection. Technical report, Statistics Netherlands, 2005.

[15] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian processes for Machine Learning*. MIT Press, 2006.

[16] World Health Organization. World health statistics 2015. Technical report, WHO Press, Geneva, 2015.

[17] World Health Organization and others. World malaria report 2014. Technical report, WHO Press, Geneva, 2014.

[18] Akira Moiseevich Yaglom. *Correlation theory of stationary and related random functions I: Basic results*. Springer-Verlang, 1986.