Mapping the Risk of International Infectious Disease Spread (MRIIDS)

A project funded through USAID’s “Combating Zika and Future Threats: A Grand Challenge for Development” program  
Milestone 6: Increased complexity of the simple model to include data from Milestone 5

27 November, 2017

# Milestone Description

Increased complexity of the simple model to include data from Milestone 5. Automated testing and validating procedures implemented where possible. Explore procedures for multiple models comparison and accounting for uncertainty.

# General Approach

In Milestone 4, we presented a simple transmission model that made use of historical case counts (data stream 1), information about the transmissibility of the pathogen (data stream 2) and geographical characterization (data stream 3) to predict future risk. In this approach, the geographical characterization was not fully integrated in the model: first transmissibility was estimated from case count data, then future incidence was predicted, and finally, the predicted incidence was distributed geographically according to the spatial distribution and population density of each geographical unit.

To achieve the goals outlined in Milestone 6, we built on the model presented in Milestone 4 (ML4) to integrate geographical information into our inference procedure. Having achieved this, other characteristics defined at the geographical scale of reference such as the health care capacity of a location, can be added to the model.

In essence, by integrating the spatial information into the inference and prediction phases, we have developed a complex model that has the potential to account for the multiple data streams initially described.

Therefore, our new complex model has the ability to:

* estimate model parameters: including parameters linked to geographical spread (or potentially health care capacity,
* predict the regional/international spread of Ebola, relying on those parameters’ estimates.

We have also established the procedure for the validation process using historical data and are exploring various possibilities for multi-model comparison.

# Presentation of the model

The number of cases at a location at time is given by the equation

where is the reproduction number at location at time and is the probability of moving from location to location . The quantity is the reproduction number at time at location . is affected by a number of other factors e.g., the intrinsic transmissibility of a pathogen, the health care capacity at location etc. Its dependence on these factors is formalized as

where is an index/score quantifying the health care capacity at , denotes a function, is the basic reproduction number (data stream 2) and is time..

The probability of moving between locations is derived from the relative flow of populations between locations. This latter quantity is estimated using a population flow model such as gravity model. Under gravity model, the flow of individuals from area to area , , is proportional to the product of the populations of the two areas, and and inversely proportional to the distance between them , all quantities are raised to some power.

In practice, and are assumed to be . The exponent modulates the effects distance on the flow of populations. A large value of indicates that the distances traveled by populations tend to be short.

The relative risk of spread at a location from a location is thus the population flow into location from location .

The probability of movement from location to location is given by

where is the probability of staying at location . As the above equation indicates, by varying , we can capture the dynamics of population flow across spatial units. For instance, if is large, then the flow out of location would be small. Thus, if this parameter is geographically heterogeneous, we obtain imbalanced flow of population.

NEED FORMULATION FOR FUNCTION PHI + DISCUSS (POSSIBLE IMBALANCE IN FLOW IF PSTAY IS GEOGRAPHICALLY HETEROGENEOUS)

## Statistical inference of model parameters

The parameters of the full model as presented in Section [sec:model] are:

* , the reproduction at time ,
* , the probability of staying in location , and
* , the exponent of the distance in the gravity model.

I WOULD REMOVE THIS BELOW? The reproduction number at time can be estimated from the incidence data for the previous 3 to 4 weeks using the package EpiEstim ((Cori et al. 2013)).

The other parameters can be estimated using maximum likelihood estimation or estimating the posterior distribution of the parameters using MCMC. Let the observed incidence time series at locations through and time be

$$I = \begin{bmatrix}
o\_{1,1} & o\_{1,2} & \dots & o\_{1,n} \\
o\_{2,1} & o\_{2,2} & \dots & o\_{2,n} \\
\hdotsfor{4} \\
o\_{t,1} & o\_{t,2} & \dots & o\_{t,n}
\end{bmatrix}$$

where is the observed incidence at time at location . Then the likelihood of the model parameters given the observations is proportional to the probability of the data given model parameters. The probability of given the model parameters is:

where is given by

Thus assuming that each observation is independent, the likelihood of the parameters is proportional to

In practice, we use past 2 to 3 weeks of incidence data to compute the likelihood.

NEEDS TO DESCRIBE BY GEOGRAPHICAL UNITS HERE... AND PROVIDE THE POSTERIOR DISTRIBUTION OF THE PARAMETERS

## Multi-model Comparison

We are exploring two possibilities:

* simple model selection procedure
* multi-model averaging.

While the first of these is easier, multi-model averaging better accounts for both parameter and model uncertainty.

Needs example (e.g. No gamma, i.e. Gamma=1) compare dic use best fit model

Or evaluate all models and use multi-model average (see (Burnham, Anderson, and Huyvaert 2011)): First evaluate dic for each model Compute relative importance of each model based on dic Predict for each model and weight predictions according to relative importance

## Predicting Future Incidence Pattern and Geographical Spread

# Implementation Details

## Software Package mRIIDS

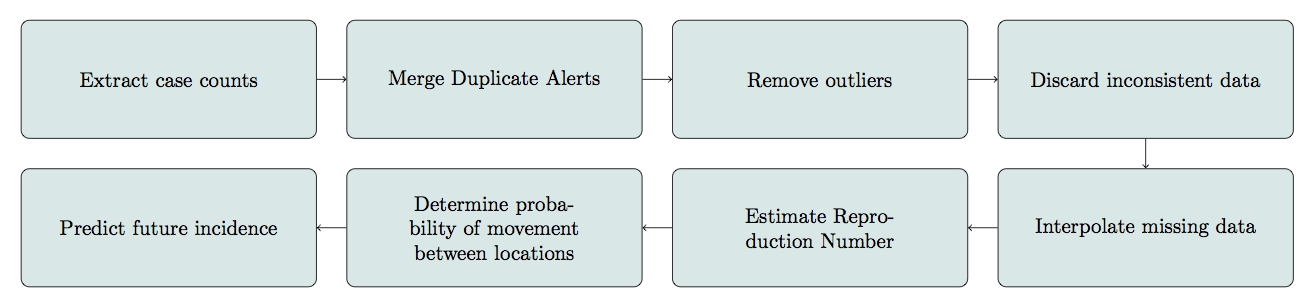
The general approach outlined above relies on several data streams, an inference framework, and a framework for projection. The code developed as part of the project is available as an open source R package that provide functions for pre-processing and collating the various data streams as well as plug the data into modules that will do the inference and the projection.

The software package will eventually be published on the R packages repository (CRAN). At the moment, it is available on GitHub (github.com/annecori/mRIIDSprocessData).

[fig:github]

The package will include extensive documentation in the form of user-friendly help files and vignettes.

## Collating data for each data stream



Workflow for the processing of the various data streams.

Figure [fig:workflow] summarizes the steps involved in collating the different data streams and in going from raw data to predictions. In Milestone 6, a step was added to the data pre-processing workflow to remove outliers from data. The removal of outliers was done using Chebyshev Inequality with sample mean (see (Saw, Yang, and Mo 1984)). Figure [fig:wf\_example] illustrates the results of each step in the pre-processing steps in the workflow.

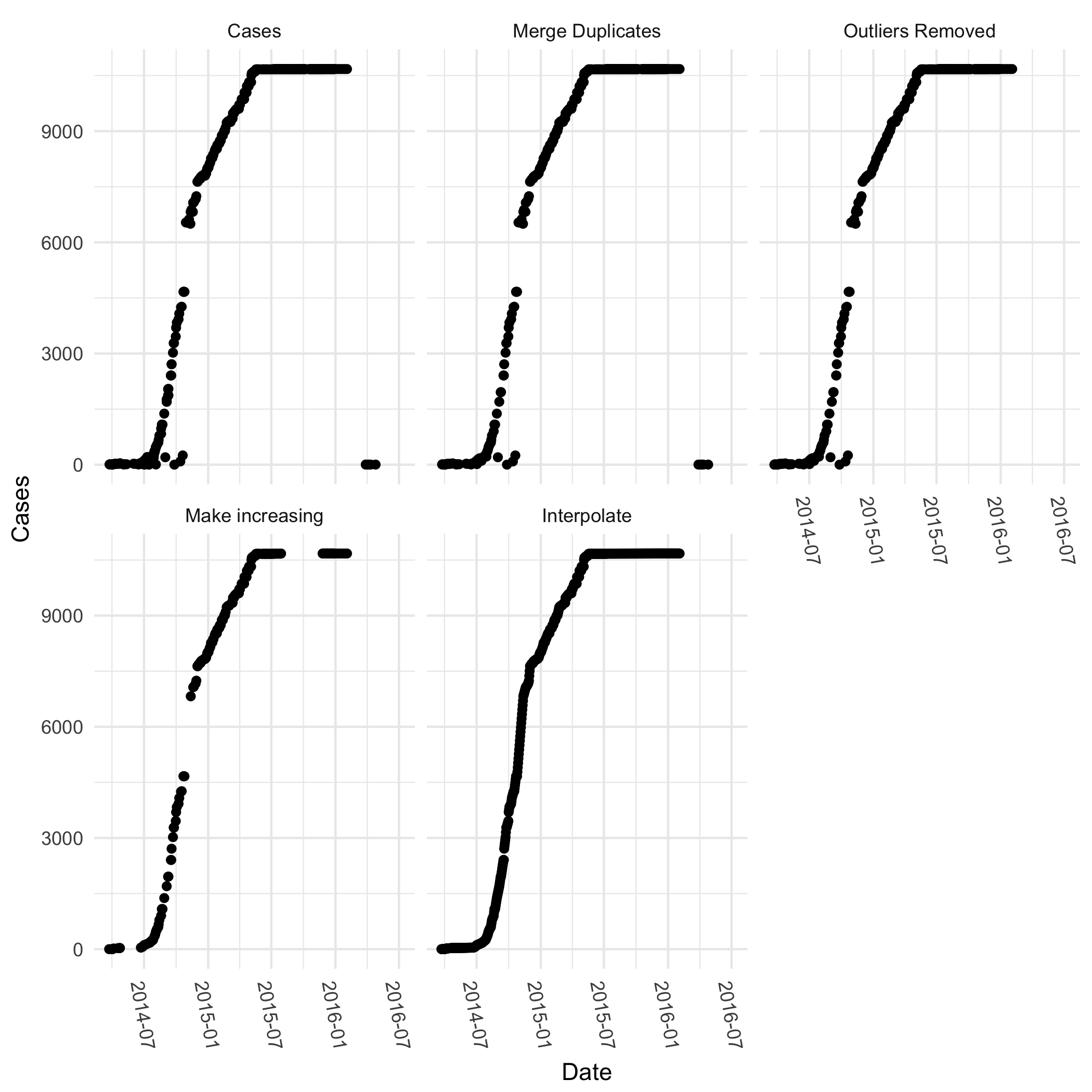


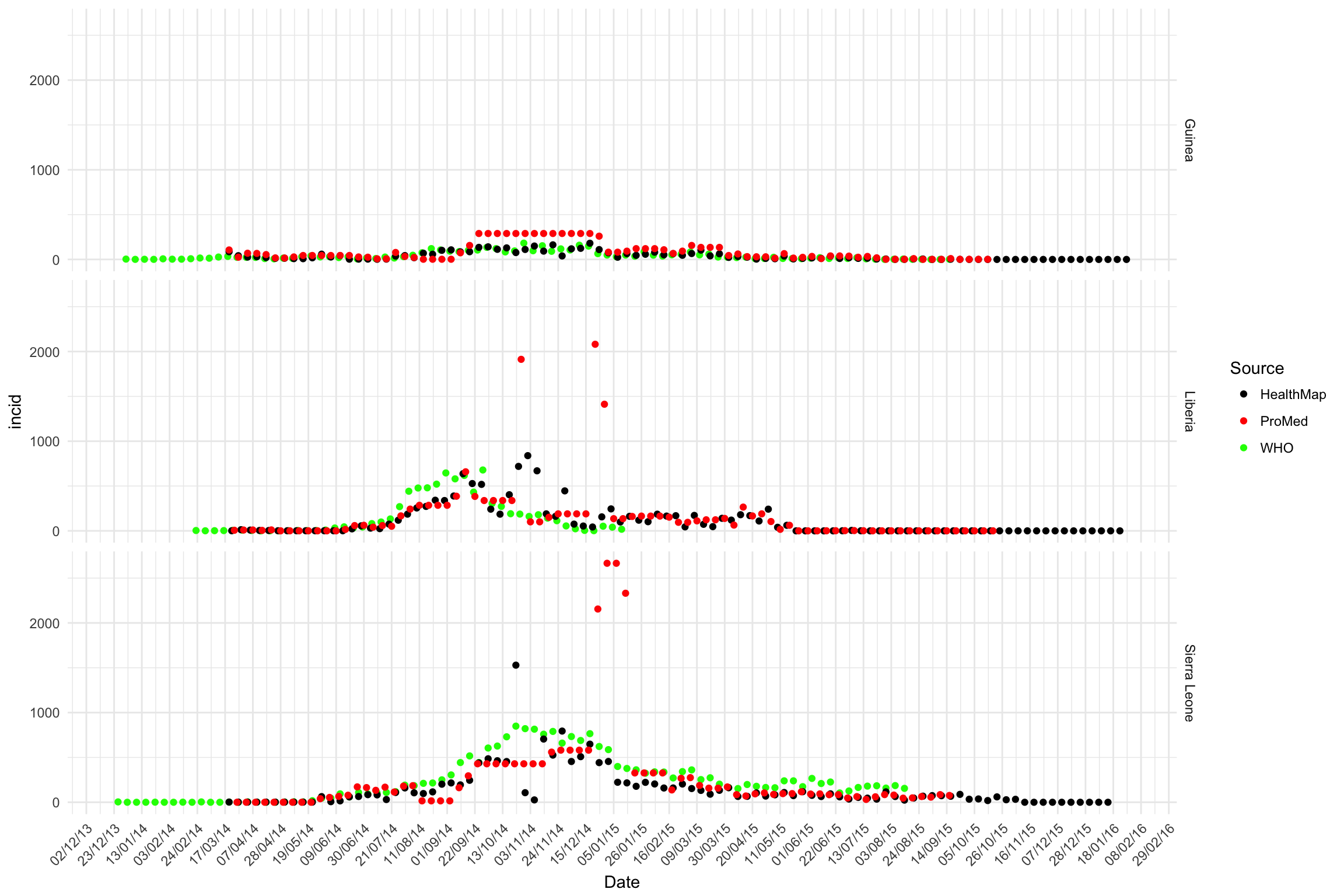
Illustration of the pre-processing steps on HealthMap incidence data for Liberia.

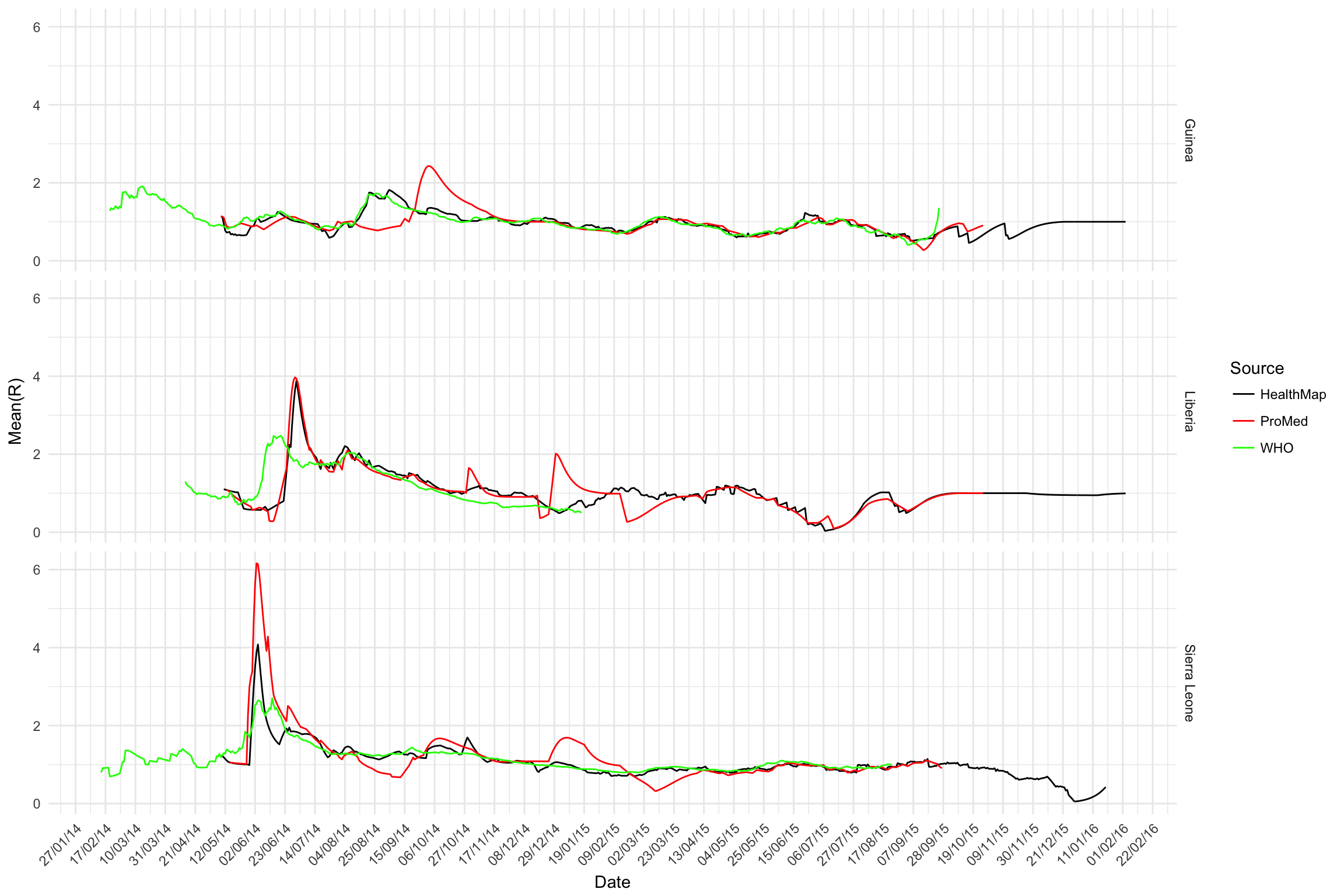
## Model training and validation using data from WHO

In the current iteration, the model was trained and validated the data on cases officially reported to the WHO during the 2013–2016 Ebola outbreak in Guinea, Liberia and Sierra Leone. This dataset was cleaned and published in (Garske et al. 2017) and it is this cleaned version of the data that were used in this work. This dataset consists of incidence reports at ADM2 level. Thus in using it, we were able to better validate the model at a finer spatial resolution than available with HealthMap/ProMed data. We refer to this dataset as WHO data throughout the rest of this document.

### Incidence trends from different data sources

We aggregated the WHO data to national level to compare the incidence trends derived from the three different data sources (WHO, HealthMap and ProMed). As can be seen in Figures [fig:incid\_comp] and [fig:r\_comp], the three data sources correlate well both in the incidence time series as well as the reproduction numbers estimated from the incidence data.

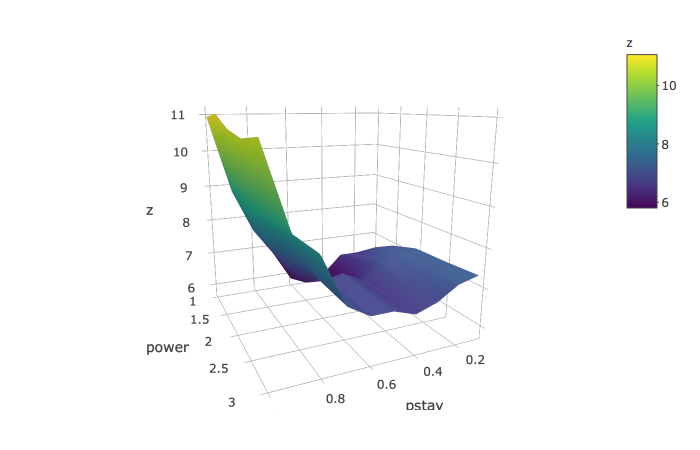
0.8 

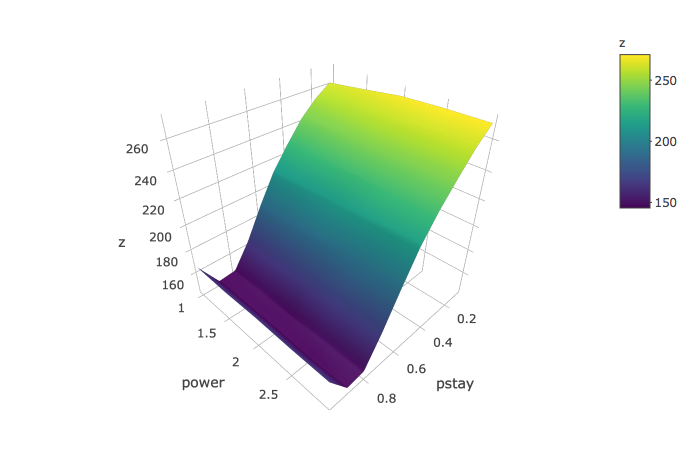
0.8 

## Inference of parameters

The parameters of the full model detailed in Section [sec:] are , , , and . The estimation os uses incidence data and is done using the EpiEstim package. In the interest of simplicity, we assume both and to be . Thus the set of parameters is now and . We explored the influence of these two parameters on the quality of fit of the model at various points in the epidemic. To assess the goodness-of-fit, we used the normalised root mean squared error (rms), which is the sum of squares of the differences between observed and predicted values. That is,

where is th observation, is the corresponding value predicted by the model and is the total number of observations.

0.4 

0.4 

# Output for Ebola in West Africa

## Estimating Transmissibility

## Predicting Future Cases

## Connectivity and Risk of International Spread

## Parameters Inference

[initial exploration of how and power, using rms]

[Formalization of the likelihood for joint estimation of R, , power]

This section should include figures

## Predicting future incidence and geographical spread

[present some early results – map of predictions/ incidence curve/ table at country level? ].

# Conclusions and next steps

# References

Burnham, Kenneth P, David R Anderson, and Kathryn P Huyvaert. 2011. “AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons.” *Behavioral Ecology and Sociobiology* 65 (1). Springer: 23–35.

Cori, Anne, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. 2013. “A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics.” *American Journal of Epidemiology* 178 (9). Oxford University Press: 1505–12.

Garske, Tini, Anne Cori, Archchun Ariyarajah, Isobel M. Blake, Ilaria Dorigatti, Tim Eckmanns, Christophe Fraser, et al. 2017. “Heterogeneities in the Case Fatality Ratio in the West African Ebola Outbreak 20132016.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 372 (1721). The Royal Society. doi:[10.1098/rstb.2016.0308](https://doi.org/10.1098/rstb.2016.0308).

Saw, John G, Mark CK Yang, and Tse Chin Mo. 1984. “Chebyshev Inequality with Estimated Mean and Variance.” *The American Statistician* 38 (2). Taylor & Francis Group: 130–32.