Mapping the Risk of International Infectious Disease Spread (MRIIDS)

27 November, 2017

Contents

1 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.

2 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.

For the model validation, we rely on both Promed/HealthMap incidence data, which are at the national scale, and WHO reported data which are at the district level. The advantage of using the more spatially refined WHO data is that it allows increased statistical power to infer the spatial parameters of the model (i.e. more movements occur at finer spatial scale, therefore the 'signature' of movement in inceidence data is more identifyable at finer scale). This exercise could be viewed as:

- demonstrating the flexible nature of our framework, i.e. the model is designed to be flexible in term of the choice of spatial scale, and
- a proof of concept to argue that reporting spatially refined incidence count can improve our ability to predict spatial spread.

3 Presentation of the model

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

$$I_{j,t} \sim Pois\left(\sum_{i=1}^{n}\left(p_{i\to j}R_{t,i}\sum_{s=1}^{t} I_{i,t-s}w_s\right)\right),$$

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

$$R_{t,i} := f(haq_i, R_0, t),$$

where haq_i is an index/score quantifying the health care capacity at location i, f denotes a function, R_0 is the basic reproduction number (data stream 2) and t 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 a gravity model. Under a gravity model, the flow of individuals from area i to area j, $\phi_{i\to j}$, is proportional to the product of the populations of the two areas, N_i and N_j and inversely proportional to the distance between them $d_{i,j}$, all quantities are raised to some power.

$$\phi_{i\to j} := \frac{N_i^{\alpha}N_j^{\beta}}{d_{i,j}^{\gamma}}.$$

In practice, α and β are assumed to be 1. The exponent γ modulates the effects distance on the flow of populations. A large value of γ indicates that the distances traveled by populations tend to be short.

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

$$r_{i\to j}^{spread} = \frac{\phi_{i\to j}}{\sum_x \phi_{i\to j}}.$$

The probability of movement from location i to location j is given by

$$p_{i\to j} = (1 - p_{stay}^{i})r_{i\to j}^{spread},$$

where p_{stay}^{i} is the probability of staying at location i. As the above equation indicates, by varying p_{stay}^{i}, we can capture the dynamics of population flow across spatial units. For instance, if p_{stay}^{i} is large, then the flow out of location i would be small. Thus, if this parameter is geographically heterogeneous, we obtain imbalanced flow of population (i.e. a source-sink dynamics).

3.1 Statistical inference of model parameters

The parameters of the full model as presented in Section 3 are:

- $R_{t,i}$, the reproduction at time t,
- p_{stay}, the probability of staying in location i, and
- γ, the exponent of the distance in the gravity model.

The 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 1 through n and time $1, 2 \dots t$ be

$$I = \begin{bmatrix} o_{1,1} & o_{1,2} & \cdots & o_{1,n} \\ o_{2,1} & o_{2,2} & \cdots & o_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ o_{t,1} & o_{t,2} & \cdots & o_{t,n} \end{bmatrix}$$

where $o_{i,j}$ is the observed incidence at time i at location j. Then the likelihood of the model parameters given the observations is proportional to the probability of the data given model parameters. The probability of $o_{j,t}$ given the model parameters is:

$$P(o_{j,t} \mid p^i_{stay}, \gamma, R_{i,t}) = e^{-\lambda_{j,t}} \frac{o_{j,t}^{\lambda_{j,t}}}{\lambda_{j,t}!},$$

where $\lambda_{j,t}$ is given by

$$\lambda_{j,t} = \sum_{i=1}^{n} \left(p_{i \to j} R_{i,t} \sum_{s=1}^{t} I_{i,s} w_{t-s} \right).$$

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

$$\mathcal{L} = P(\{o_{j,t}\} \mid \{p^i_{stay}\}, \gamma, \{R_{i,t}\}) = \prod_{t=1}^{t} e^{-\lambda_{i,t}} \frac{o_{i,t}^{\lambda_{i,t}}}{\lambda_{i,t}!}.$$

In practice, we estimate $R_{t,i}$ as an average over the past 2 or 3 week with sliding time windows.

Given this likelihood, we can write the joint posterior distribution of the parameter given the observaed data as:

$$P(\{p^i_{stay}\}, \gamma, \{R_{i,t}\} \mid \{o_{j,t}\}) \propto \mathcal{L} \times P(\{R_{i,0}\}) P(\{p^i_{stay}\}) P(\gamma).$$

Here, $P(\{R_{i,0}\})$ represents the prior distribution of the basic reproduction number. This prior distribution is influenced by data-stream 2 and the health capacity of the location i, as described above.

The other prior distributions, $P(\{p^i_{stay}\})$ and $P(\gamma)$, could in principle reflect the influence of additional data sources such as prior information derived from flight data.

3.2 Multi-model Comparison

Given the most general model formulation outlined above, multiple models could be formulated that can be viewed as simplification of the original model. For instance assuming $\{R_{i,t}\}$ to be constant across geographical units, or assuming unity for the parameter $zgamma$ of the gravity.

Variants of the model would have distinct number of parameters, for instance assuming $\gamma = 1$ would reduce the number of parameters to be estimated by 1.

Relying on data-driven and evidence based approaches, we seek to formulate the simplest model that account for patterns observed in the data. In such simplest model, all layers of complexity must be justified, i.e. it is justified to estimate a parameter γ (and not assume unity) if and only if the fit of our model to the observed data is significantly improved (in a statistical sense).

Relying on our MCMC estimation of the joint posterior, we can evaluate the goodness of fit of any models by calculating the Deviance Information Criterion (DIC). Such DIC is a well established measure of goodness of fit, and commonly used for model selection. it promotes models that can best reproduce observed patterns while penalizing for increased models complexity (i.e. increased number of parameters).

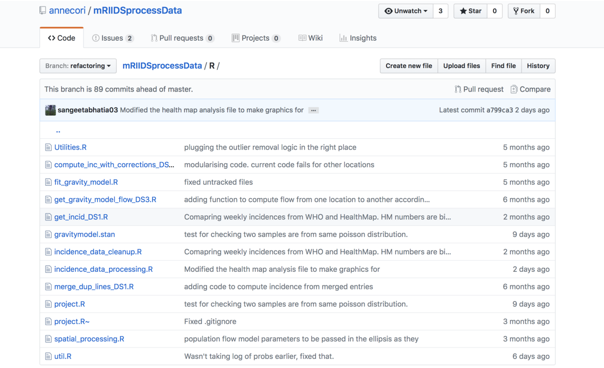
Using such 'model selection' approach, each model is evaluated in turn, DICs are obtained and compared, and we can rigorously selected the best model to produce the final predictions.

Alternatively, after evaluating each model, we can produce 'model averaged' predictions. Under this approach, the predictions of each model are weighted according to the statistical of the particular model. Such approach has the advantage of accounting for structural uncertainties, i.e. the very structure of the underlying model is treated as an unknown state.

The 'model averaged' predictions are finally obtained based on the 4 following steps:

- evaluate the goodness of fit (e.g. DIC) for each model,
- calculate the difference (Δ_k) between the DIC value of the best model and the DIC value for each of the other models,
- compute the relative weight for each model as $w_k = \frac{exp(\frac{1}{2}\Delta_k)}{\sum_{m=1}^{M} exp(\frac{1}{2}\Delta_m)}$, with M the total number of models evaluated,
- produce predictions by sequentially 1) randomly selecting a particular model according to its weight, 2) producing predictions based on the parameters' joint posterior distributions associated with this particular model.

We are currently in the process of implementing those approaches (simple model selection and multi-model averaging). Once implementation is finalized, we will validate and contrast both approach, recognizing that while selecting a model is easier to implement, multi-model averaging should better accounts for both parameters and model uncertainty, i.e. parametric and structural uncertainties.



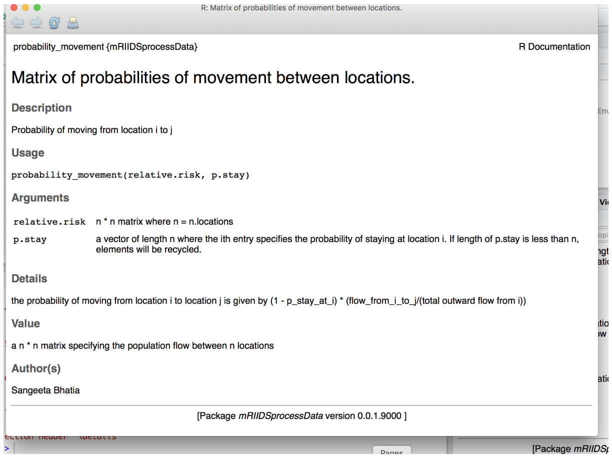
mRIIDSprocessData / R /

This branch is 89 commits ahead of master.

sangeetabhatia03 Modified the health map analysis file to make graphics for … Latest commit a799ca3 2 days ago

..		
Utilities.R	plugging the outlier removal logic in the right place	5 months ago
compute_inc_with_corrections_DS…	modularising code. current code fails for other locations	5 months ago
fit_gravity_model.R	fixed untracked files	5 months ago
get_gravity_model_flow_DS3.R	adding function to compute flow from one location to another accordin…	6 months ago
get_incid_DS1.R	Comapring weekly incidences from WHO and HealthMap. HM numbers are bi…	2 months ago
gravitymodel.stan	test for checking two samples are from same poisson distribution.	9 days ago
incidence_data_cleanup.R	Comapring weekly incidences from WHO and HealthMap. HM numbers are bi…	2 months ago
incidence_data_processing.R	Modified the health map analysis file to make graphics for	2 days ago
merge_dup_lines_DS1.R	adding code to compute incidence from merged entries	6 months ago
project.R	test for checking two samples are from same poisson distribution.	9 days ago
project.R~	Fixed .gitignore	3 months ago
spatial_processing.R	population flow model parameters to be passed in the ellipsis as they	3 months ago
util.R	Wasn't taking log of probs earlier, fixed that.	6 days ago

Figure 1: The software being developed for the project is available on GitHub.


R: Matrix of probabilities of movement between locations.

probability_movement {mRIIDSprocessData} R Documentation

Matrix of probabilities of movement between locations.

Description

Probability of moving from location i to j

Usage

```
probability_movement(relative.risk, p.stay)
```

Arguments

relative.risk	n * n matrix where n = n.locations
p.stay	a vector of length n where the ith entry specifies the probability of staying at location i. If length of p.stay is less than n, elements will be recycled.

Details

the probability of moving from location i to location j is given by (1 - p_stay_at_i) * (flow_from_i_to_j/(total outward flow from i))

Value

a n * n matrix specifying the population flow between n locations

Author(s)

Sangeeta Bhatia

[Package *mRIIDSprocessData* version 0.0.1.9000]

Figure 2: An example of the documentation for the R package mRIIDS

4 Implementation Details

4.1 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).

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

4.2 Collating data for each data stream

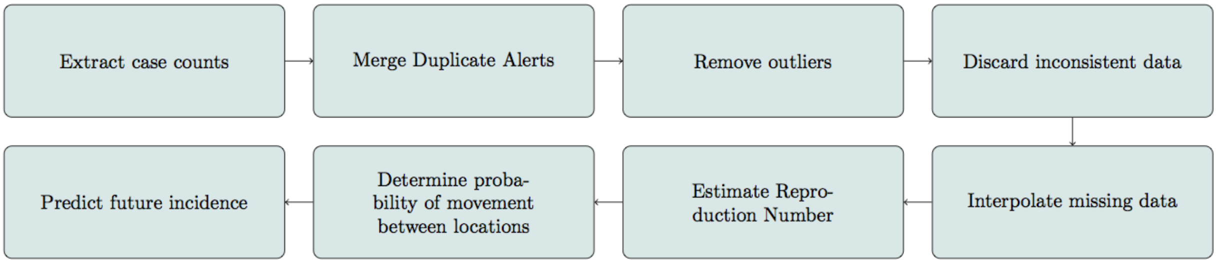

Extract case counts
Merge Duplicate Alerts
Remove outliers
Discard inconsistent data
Interpolate missing data
Estimate Reproduction Number
Determine probability of movement between locations
Predict future incidence

Figure 3: Workflow for the processing of the various data streams.

Figure 3 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 **?**). Figure 4 illustrates the results of each step in the pre-processing steps in the workflow.

4.3 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 [**?**] 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 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.

4.4 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 Figure 5, the three data sources correlate well in the incidence time series.

4.5 Inference of parameters

The parameters of the full model detailed in Section 3 are α, β, γ, p_{stay} and $R_{i,t}$. The estimation os $R_{i,t}$ uses incidence data and is done using the EpiEstim package. Figure 6 shows the reproduction number estimated using previous 28 days incidence data. The high degree of correlation in the estimates from the three different data sources shows that the estimation procedure is robust to slight variations in reported number.

In the interest of simplicity, we assume both α and β to be 1. The other two parameters are p_{stay} and γ. We explored the influence of these two parameters on the quality of fit of the predictions from the models 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

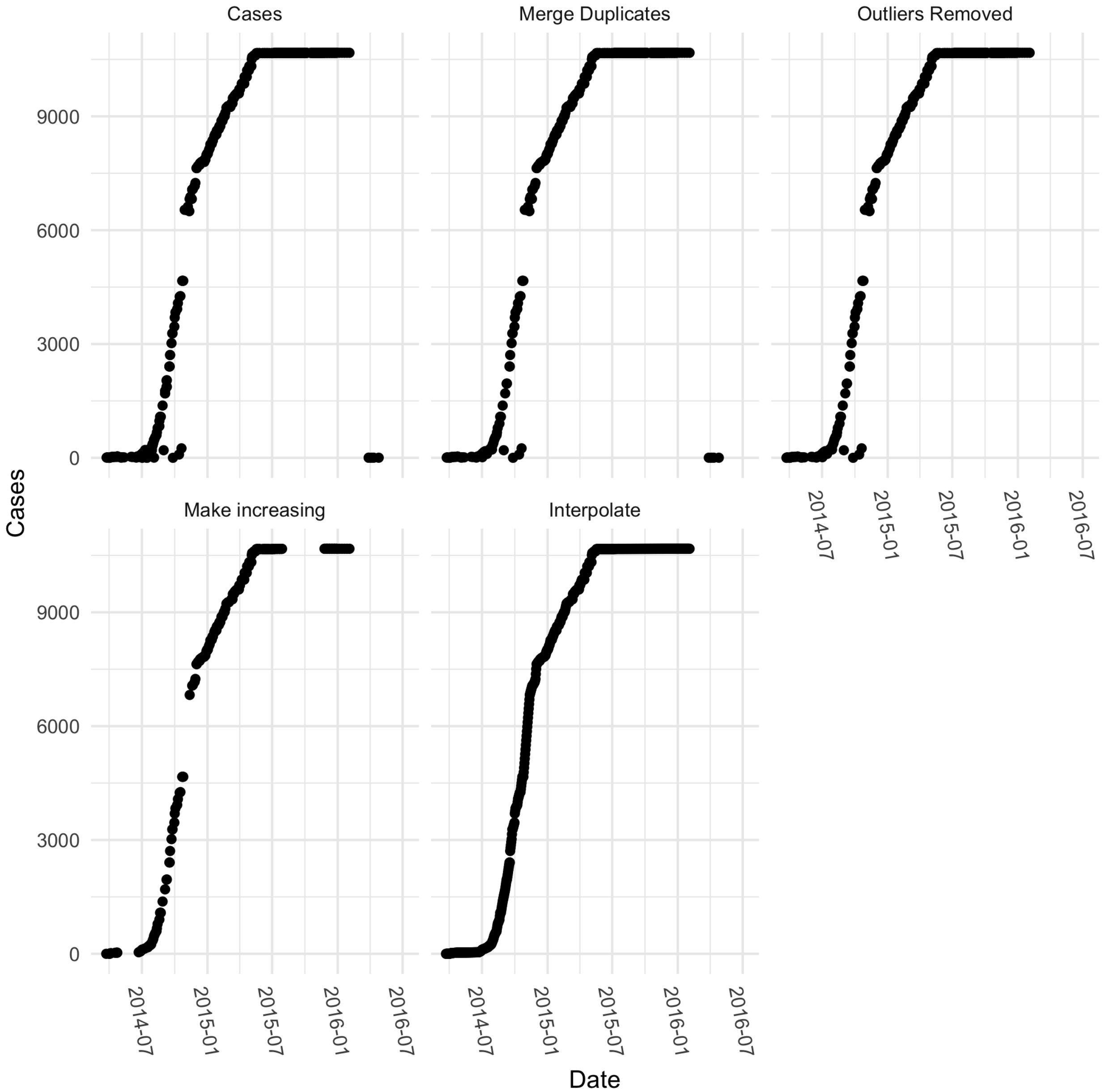


Figure 4: Illustration of the pre-processing steps on HealthMap incidence data for Liberia.

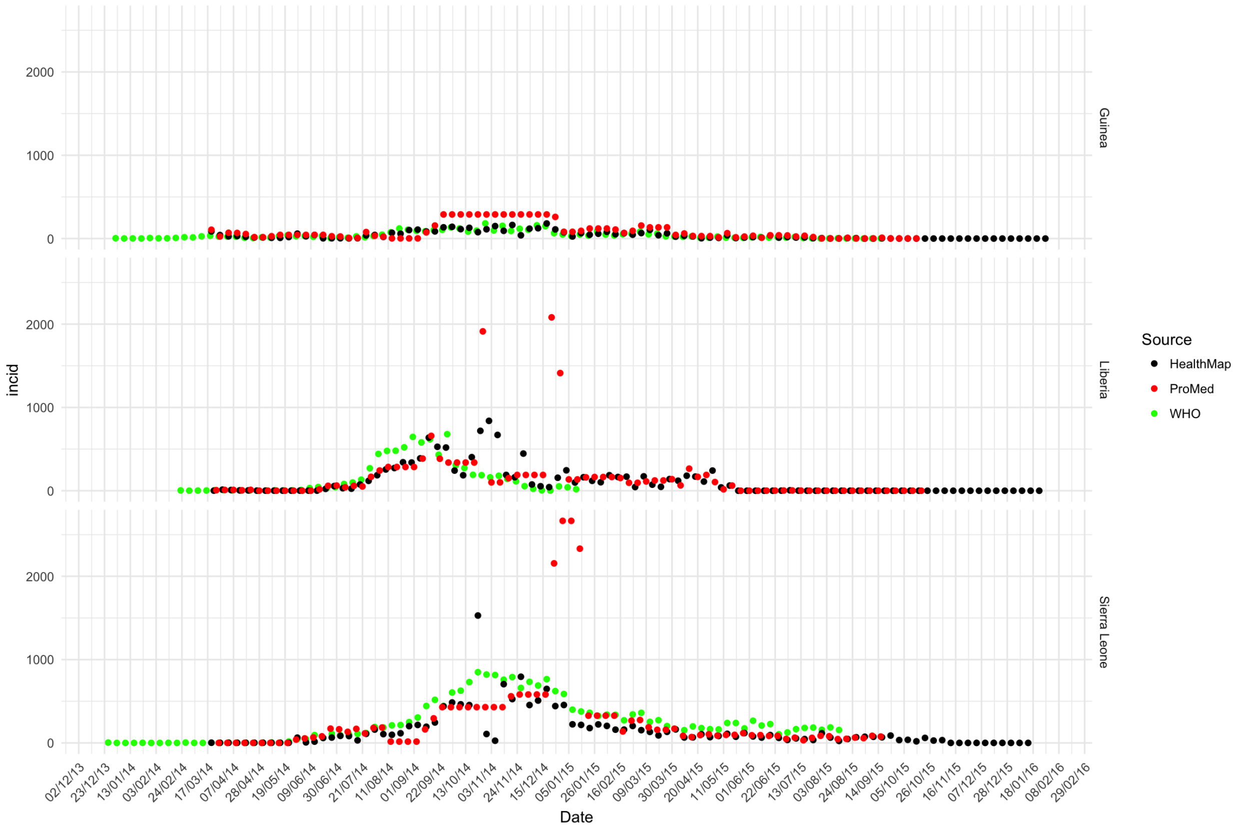


Figure 5: Comparison of incidence data from WHO, HealthMap and ProMed.

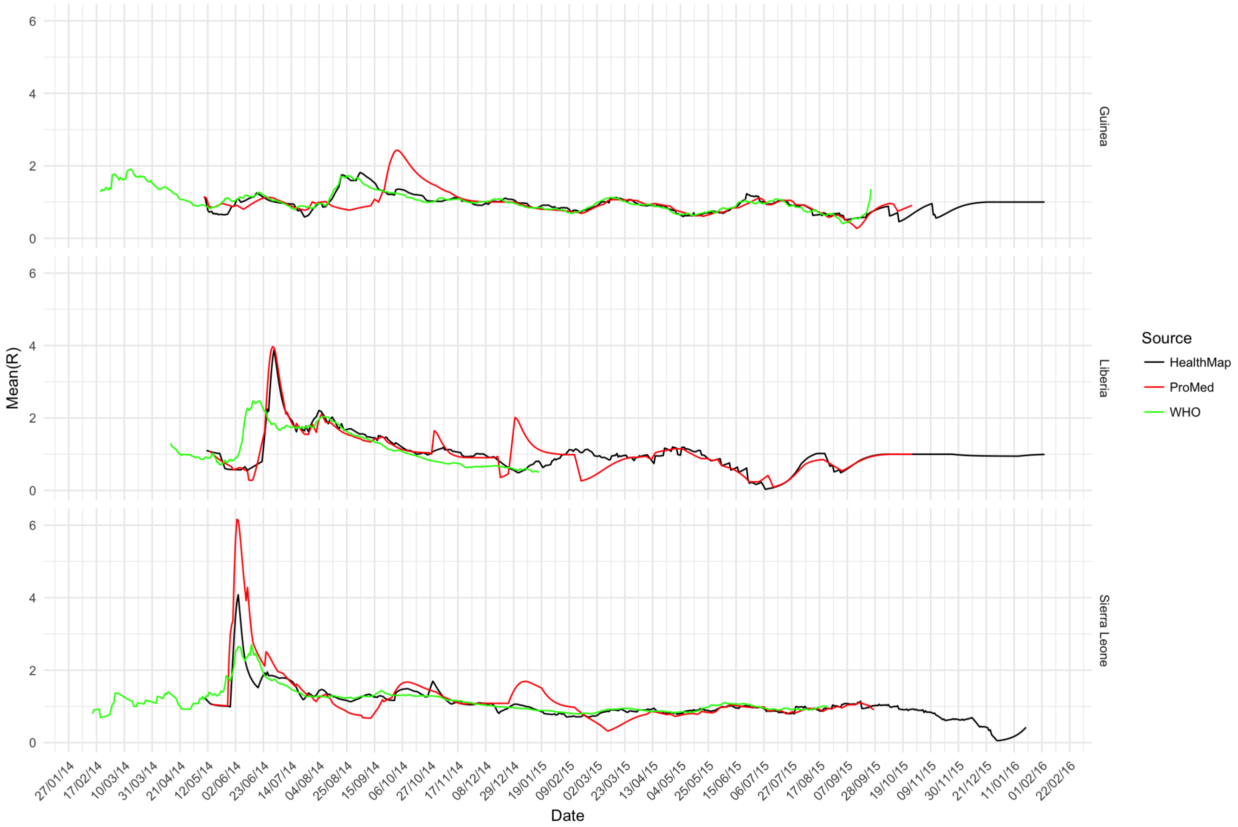
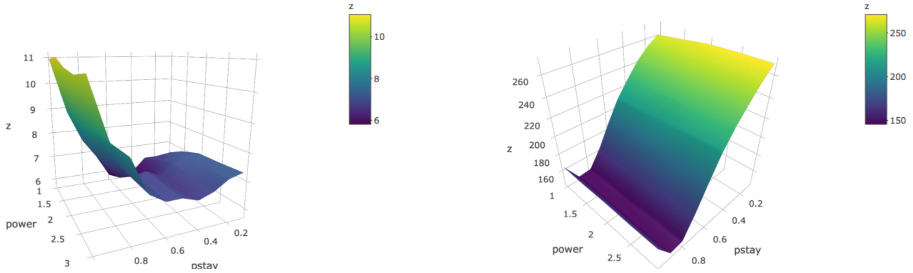


Figure 6: Comparison of the reproduction numbers estimated from the different sources of the incidence data.



(a) Root mean square errors for prediction for 5 weeks at 100 days from the start of the epidemic.

(b) Root mean square errors for prediction for 5 weeks at 300 days from the start of the epidemic.

Figure 7: Normalised root mean squared error as a function of the model parameters. The fit is assessed for prediction of 5 weeks at 100 and 300 days from the start of the epidemic. The fit is better for smaller values of the root mean square error. In the early phase of the epidemic, a better fit is obtained at a smaller value of p_{stay} while at the 300 days mark, a much higher value of p_{stay} is needed to obtain a good fit.

and predicted values. That is,

$$rms := \sum_{i=1}^{n} (o_i - p_i)^2,$$

where o_i is ith observation, p_i is the corresponding value predicted by the model and n is the total number of observations.

4.6 Predicting Future Cases

The WHO data consisted of incidence data at the district level. To validate the model, we carried out analysis at both the district level and country level.

4.6.1 Prediction at country level

4.6.2 Prediction at district level

5 Conclusions and next steps

References

References