Author's response to reviews

Title: Fuzzy Association Rule Mining and Classification for the Prediction of Malaria in South Korea

Authors:

Anna L Buczak (anna.buczak@jhuapl.edu)
Benjamin Baugher (benjamin.baugher@jhuapl.edu)
Erhan Guven (erhan.guven@jhuapl.edu)
Liane C Ramac-Thomas (liane.ramac-thomas@jhuapl.edu)
Yevgeniy Elbert (eugene.elbert@jhuapl.edu)
Steven M Babin (steven.babin@jhuapl.edu)
Sheri H Lewis (sheri.lewis@jhuapl.edu)

Version: 4 Date: 23 March 2015

Author's response to reviews: see over
We wish to thank the reviewers for their careful review of our manuscript, their comments and suggestions. We addressed the comments and suggestions one by one.

Reviewer #1: I enjoyed reading the manuscript and think the method may be of interest to people working in the field of malaria prediction. However, in order to be able to assess this, I think the manuscript needs to address the issues identified below.

1. Discretionary Revision for Title: "... prediction of malaria in Korea" As the prediction takes place on data from South Korea only (even though it makes use of some input from data from North Korea), it seems appropriate to state "South Korea" in the title.

As requested, we changed the title from Korea to South Korea.

2. Major Compulsory Revision: "Accurate prediction of malaria outbreaks" and also l 321 ".. when attempting to predict relatively rare disease outbreaks...": It is unclear how in the study how "outbreaks" are defined. With the classification of the number of weekly cases into low, medium and high, what constitutes an "outbreak"? Perhaps the word 'outbreak' is not appropriate here?

Outbreaks have been variously defined in the literature. Our definition is based on feedback from public health professionals in Korea and was the basis for the HIGH incidence classifier. We neglected to state explicitly that this was our definition of an outbreak so we have modified the text accordingly. By defining an outbreak, the PPV and NPV make sense. We thank the reviewer for pointing out this oversight.

3. In Background. Paragraph 1. 3 "Minor issue not for publication" "Discretionary Revision" l 24: replace "infected" by "infectious" l 26: replace "infected" by "infectious" l 28: replace "infected" by "infectious"

As suggested, we have made these changes.

4. "Major Compulsory Revision" l 32: The two main malaria vector species along the border of North Korea and South Korea are Anopheles lesteri and Anopheles sinensis. Neither of these species is known to favor breeding in uncovered containers.

We have now included a reference that states that these larvae are found in fresh-water ground pools, ditches, margins of streams and ponds, rice fields, marshes, swamps, lakes, and other impounded waters. While this may be interpreted to include uncovered containers, we have revised this sentence to instead mention the association of ditches, rice fields, and impounded waters due to nearby human populations.

5. "Discretionary Revision" l 33: For flight range, the following reference would be more appropriate for this manuscript: [Cho et al., A mark-release-recapture experiment with Anopheles sinensis in the northern part of Gyeonggi-do, Korea. Korean J Parasitol. Sep 2002; 40(3): 139–148.]

As suggested, we have added the reference.

6. Paragraph 2 6 "Minor Essential Revision" l 48: "P. vivax was isolated ... spread of the disease [8]". It is unclear why, in a paragraph describing the magnitude of the problem of malaria, rare local transmission in the United States is mentioned. This seems out of place.
The purpose was to emphasize how malaria may re-appear in areas in which it had long been eradicated. We have revised this sentence to make this point clearer.

7. Paragraph 7 "Major Compulsory Revision"

197: "There were studies that used all available data in model development, without using a separate set for assessing prediction accuracy". This statement is selective. Zinszer et al., 2012, write in their abstract "Model evaluation was typically based on a reserved portion". Therefore, MOST studies in their review DID measure prediction accuracy on reserved data. The text in line 97 and the last paragraph of the background section, the false impression is given that the present study is one of the first contributions to malaria prediction literature that measures prediction accuracy on reserved data.

The first statement quoted above simply says that there were studies that used all available data; that statement does not say anything about the relative number of studies that use reserved data. In the last paragraph of the background section, we mention the use of reserved data in the context of providing the user with a more conservative estimate of prediction accuracy than not doing so. We had no intention of implying that our study was the first to do this and we cannot understand how the reviewer inferred this. However, to avoid this misperception, we now cite the paper by Briet et al. as an example of a previous paper that used reserved data. Please note that, in the body of the paper, Zinszer et al. states “Some of the studies used all the available data to fit a model and did not reserve data for assessing forecast accuracy.” Also, please note that a more recent review by Corley et al. (2014) found only 6 out of 44 disease prediction model papers used separate datasets for testing. Both the Zinszer et al. and Corley et al. reviews emphasized the importance of reserving data for evaluation. Therefore, we feel that our statements are accurate and corroborated by our references. We have provided an additional citation of the Corley et al. reference along with the Zinszer et al. reference mentioned above.

8. "Major Compulsory Revision" 199: "They (Zinszer et al., 2012) concluded that prediction accuracy should always be measured on reserved data ... and common prediction measures should be used to allow model comparison." In fact, the conclusions of the abstract of Zinszer et al., 2012 were:

"Conclusions: Applying different forecasting methods to the same data, exploring the predictive ability of non-environmental variables, including transmission reducing interventions and using common forecast accuracy measures will allow malaria researchers to compare and improve models and methods, which should improve the quality of malaria forecasting." The crux is "Applying different methods to the same data .. will allow malaria researchers to compare and improve models and methods". Whereas this manuscript does compare the proposed FARM method to other machine learning methods 'Decision Tree', 'Random Forest' and 'Support Vector Machine', only 3 out of 29 malaria forecasting studies identified by Zinszer et al., 2012 were classified as using neural networks (also machine learning methods, but perhaps not necessarily comparable to Decision Tree, Random Forest and Support Vector Machine methods), the vast majority of publications applied GLM, ARIMA, Grey methods, and/or smoothing methods (including Holt-Winters, etc.). For malaria researchers to judge whether the FARM methods described in this manuscript contribute to the tool-kit in their field and are worth considering, the manuscript must compare the performance of the proposed FARM method to AT LEAST the performance of mono-variate predictions of Seasonal ARIMA and Holt-Winters and other easily and quickly implementable methods on the same data based on common prediction quality criteria.
The reviewer’s comment above mentions only one of the items listed by Zinszer et al. in their Abstract Conclusions. Here is the entire Zinszer et al. Abstract Conclusion: “Applying different forecasting methods to the same data, exploring the predictive capability of non-environmental variables, including transmission reducing interventions and using common forecast accuracy measures will allow malaria researchers to compare and improve models and methods, which should improve the quality of malaria forecasting.” This long sentence is the Conclusion of their abstract. While it is an adequate summary, it is important to examine the entire Zinszer et al. paper.

The reviewer then mentions that we should compare our method with SARIMA instead because this type of model is commonly used in published papers. In the review by Zinszer et al., it was noted that SARIMA models have widely different forecast accuracies and this is corroborated by the results for different regions in the Briet et al. paper. Therefore, the fact that many published papers have used SARIMA is more likely a result of it already being included in statistical software packages rather than a result of truly successful operational use.

The popularity of a technique in the literature does not mean that it is useful in practice. For example, the public health users in South Korea were not using SARIMA or Holt-Winters models. Nonetheless, we have responded to the reviewers comment by adding a comparison of the FARM results with the exponential smoothing by Holt-Winters and included the results in the paper. We developed separate exponential smoothing models with additive Holt-Winters for every region in the data set (64 regions). The values of α, β, and γ were optimized on the training data set, separately for each region (64 regions total) in such a way as to minimize the mean absolute relative error (MARE). The outputs of exponential smoothing models are continuous numbers and not categories. Therefore, in order to be able to compare these results with the three class results of FARM and other machine learning methods, the results were binned into the categories LOW, MEDIUM, and HIGH using the thresholds used for FARM, which were 3 and 17. Since the data were log transformed, we are actually using log(3+1) and log(17+1) as thresholds. In case of FARM, Random Forest, Decision Tree, and SVM, a single classifier is trained and used for all 64 regions. In case of Holt-Winters, a separate model was developed for every region (otherwise the results would have been much worse than presented here) making the comparison less than straightforward. Also, all the other models used multiple predictor variables, while only case counts were used in case of Holt-Winters. The metrics for MEDIUM and HIGH classes for Holt-Winters were the worst of all the methods, with the exception of PPV for Medium Class (worst result obtained by Random Forest).

We wish to emphasize that comparing our model with SARIMA and/or Holt-Winters is of limited benefit. We originally chose the machine learning models for comparison because they provided more of an “apples to apples” comparison. Although we have now added a Holt-Winters comparison as requested, we wish to emphasize the following points:

1) the metrics preferred by public health users (i.e., PPV and NPV) are never published for SARIMA or Holt-Winters models; 2) the Holt-Winters approach required creating a new model for each region, while our approach uses the same model for all the regions in South Korea; 3) SARIMA and Holt-Winters are temporal models only, while our model uses spatial information (e.g., distance from DMZ) as well as temporal; 4) the metrics used to evaluate Holt-Winters and SARIMA (e.g., MARE) are completely different than the PPV and NPV metrics used in our model; 5) our model is multivariate (we are using over 15 different variables along with 12 time lags for many of them; 6) Holt-Winters and SARIMA are usually used as univariate models; mutivariate
extensions exist, however we have not seen a Holt-Winters and SARIMA with more than 3 dimensions, and we have about 15 dimensions; the variables that we are using don’t have the same frequency: some are weekly, some change once a year. The variables with such a low frequency are not appropriate for inclusion in auto-aggressive models.

9. 2.1 ROK Malaria Case Data Paragraph 1. 9 "Major Compulsory Revision" | 136: "Malaria weekly data were obtained from the Korea Centers for Disease Control and Prevention." Were these cases confirmed by microscopy or RDT or were they suspected? How were these data 'obtained'? Were they downloaded from a website? Could the authors provide these data as a supplementary file? It is noticeable that none of the co-authors nor those acknowledged appear to be affiliated to a South or North Korean organization or institution. How is it ensured that this research has (any) local relevance? Are the categories for low, medium and high relevant for the Korea Centers for Disease Control and Prevention? Are the prediction horizons of 7-8 weeks (Figures 8 and 9) relevant for the Korea Centers for Disease Control and Prevention?

These data include confirmed malaria cases, but the data source does not specify how these were confirmed. These are the same malaria data used in their annual reports of confirmed cases. We have revised the paper to make this clearer. The data were downloaded from the KCDC (Korean CDC) website and we now provide this information in the paper. Thanks to the reviewer’s comment, we have discovered that Acknowledgements were missing from the paper. We consulted with numerous local public health professionals from the regions for which we made predictions and we now include this information in the Acknowledgements. Malaria is only a problem with those regions bordering the DMZ so those are the regions for which we made predictions. The reason we provided the results in classifications of LOW, MEDIUM, and HIGH is because the users requested these three classifications. The reason that our prediction horizons are 7-8 weeks is because the users requested this. They also requested that we implement these predictions on a website they can access, so we have done so. We have revised the text to make these points clearer. We regret that we did not include this information earlier and we thank the reviewer for prompting us to include it in the revised paper.

10. 3.4 Performance Metrics. Last paragraph. 10 "Major Compulsory Revision" | 410: Here, a multi-class problem is introduced. It is not until lines 450 - 458 that it is explained why. The reason appears to be 'ad hoc' because the two-class problem approach did not work well due to low numbers of cases in the data reserved for prediction (2013). I wonder if these value boundaries and classes in general have any relevance for the Korea Centers for Disease Control and Prevention, and if 2013 was simply a year with low case numbers, where any of the weekly numbers did not qualify as 'outbreaks'.

The point was that, due to low numbers of cases in the 2013 data, the two-class approach was predicting everything as a LOW. This prediction was 100% correct and the metrics were great. So the two-class approach worked well from that perspective, but obviously this had a limited usefulness as all predictions were always LOW. We added the MEDIUM category because of this and because the ROK users requested a third category that might include pre-outbreaks. We have revised the text to make this point clearer.

11. 4. Results 4.1 Incidence Prediction vs. Case Count Prediction Last paragraph 11 "Major Compulsory Revision"
Because of this problem using incidence rate, case counts were examined instead as a possible data source of determining the classes. The authors should describe the problems and biases associated with using case counts.

We have revised the text to explain better why we used case counts instead of incidence rates. We were not certain what the reviewer meant so we reviewed the Briet et al. paper because they used case counts. However, we could not find a discussion of any problems or biases with using them in their paper. Perhaps the reviewer thought we applied assumptions about the statistical distribution of case count data, but we had no reason to do so with our technique.

12. "Discretionary Revision" Considering the high cross-correlation between regions evident from Figures 6 and 7, would it not be preferable to demonstrate the technique on the spatially aggregated data?

The public health users wanted to see prediction results for their specific regions, so that is what we did. The data were spatially aggregated into the regions desired by the ROK users. The is easier to obtained good results for larger regions with more population and cases; so the fact that we used smaller regions, with smaller population and smaller number of cases, makes the problem that we solved more difficult.

13. Conclusions First paragraph 13 "Major Compulsory Revision" l 564-567: "The result is a model that ... predicts malaria cases ... using performance metrics that provide for more conservative and less biased estimates of model performance for the user". These performance metrics provide more conservative and less biased estimates than what? See also the comment under l 97.

Our performance metrics are more conservative and less biased than papers that do not use reserved data and consequently have performance numbers that are much higher than those that would be observed otherwise. We have revised this statement to make this point clearer.

14. Second paragraph. 14 "Major Compulsory Revision" l 568: "The data mining techniques in this method are general in the sense that they can use any data." It is unclear from the manuscript which covariate data contributed a lot, and which covariate data contributed very little or nothing to good prediction quality. If this is not quantified, users may be putting a lot of effort in collecting irrelevant data.

We have revised the text to remove this misunderstanding. The data mining techniques use all the data found relevant in the literature to create a model. The final model is chosen based on the criteria listed in the manuscript. These criteria select the predictor variables based on these criteria, so other predictor variables are dropped and not used in the process of creating the final model. All the variables used in the final model are useful and are therefore not irrelevant. However, if some of them are missing, the model still runs and produces a prediction, albeit with less accuracy. We have revised the text to make this clearer.

15. TABLES "Minor issue not for publication" 15 "Minor Essential Revision" Tables 1, 3 and 4: These results are for what forecasting horizon?
We have added the forecasting horizon to the Table captions.

16. FIGURE CAPTIONS "Minor issue not for publication" 16 "Minor Essential Revision" Figure captions numbering is wrong, it should number from 1 to 12.

We thank the reviewer for noting this oversight and have corrected the figure caption numbering in the manuscript.

17. "Minor issue not for publication" 17 "Minor Essential Revision" Figure 6: The vertical axis title should state the rate per population size.

We thank the reviewer for noting this oversight and have corrected the vertical axis title.

Reviewer #2: This manuscript is very informative and provides interesting model to malaria research. However, there are several points that the authors should elaborate more to make the concept clearer and easy for the reader to follow through the logic/algorithms used.

Major revision:
1. The authors listed and discussed about predictor variables used in this model (section 2, p6 to p11) quite extensively. However, the authors did not mention whether these set of variables are the same or different from those used in the previous model or not. The authors should provide some rationale for selection of such predictors to be used in this malaria model – whether/how they could be the same or different from the predictors used in previous dengue model. The malaria and dengue may or may not share the same predictor variables as the two diseases occur in different types of settings/environments.

We did not intend to imply that we used exactly the same variables for dengue as malaria. In fact, this comment has led us to identify a misunderstanding that we hope our revised text corrects. Our original text resulted in confusion between the model building part of the process and the model itself. In the revised manuscript, we mention that the first step in the model building process is to review the scientific literature to find the set of variables shown to be correlated with a specific disease. In the case of this paper, the disease was malaria and not dengue. The technique for building both models is very similar, with some enhancements noted in the text. However, the final model is not a dengue model applied to malaria. The final model is unique for malaria and unique for the selected regions of the Republic of Korea. We hope the revisions make this clearer.

2. One of the predictor variable mentioned in the abstract (p3 line 5) is socio-economic data. Which variable in the predictor variable set is the socio-economic - financial data for DRPK malaria control?
Yes, we used the term socio-economic data to include DPRK malaria control financial data and population density data. However, we have revised the manuscript to make this clearer.

3. In the model builder section (p 12), the initial and important part of modeling is about data pre-processing and fuzzification (line 277). But the authors simply referred to precious model for dengue according to reference #18. This should really be elaborated in this manuscript. Even though the malaria model may employ the same mechanism. The authors should review the method employed in this part clearly as it may or may not the same between malaria and dengue. Again, the authors should describe whether the same or different mechanism can be employed when applying from dengue model to malaria model.

We have revised the manuscript to make it clear that the dengue and malaria models differ. As noted above, it is the model building technique that is similar. This model building technique does not assume that the diseases are similar. The literature used to find predictor variables for malaria was not the same as that to find predictor variables for dengue.

4. In terms of data pre-processing in model builder using predictor variables in section 2, it is not quite clearly explained how the authors handle different types of predictor measurements. The predictors are based on different data units (daily, weekly, annually), different years (2004-2013, 2008-2013, 2013 only, 1981-2010, missing data since 2011, etc). Moreover in them methodology section, how/what the training dataset is used? And, what/how is, if any, the predicted/confirmed model dataset is developed?

The data pre-processing was described in considerable detail in the previous papers so we did not wish to duplicate the lengthy text. However, we revised this paper to mention in the introduction to the Predictor Variable section how these data were changed to weekly temporal and region spatial scales. We then specify these changes in each variable subsection.

Originally, we had begun talking about using training data to develop the model in Section 3.1 before we explained what the training dataset was in Section 4.2. Thanks to the reviewer, we realize this could easily cause confusion, so we now added text to the Methodology section to explain sooner in the paper how we divided the data into training and testing data sets.

5. The results section has clearly shown that the model used is quite a success and provides good statistics in terms of PPV and sensitivity. However, both statistics are good to predict LOW and HIGH, but not so good for MEDIUM. What could be the reason(s)? Since the authors did not have discussion section – this could, at least, be mentioned in the conclusion section.

We have revised the text to address this question. Because of the significantly higher threshold, there was much better class separation between the 2 classes used to train the HIGH classifier than between the 2 classes used to train the MEDIUM classifier. Thus, we would naturally expect any classification method to do better at distinguishing between HIGH and LOW than between MEDIUM and LOW, and this is borne out in the results seen in Table 1.
Although it is not as good as the HIGH classifier, the MEDIUM classifier actually performs quite well as a binary classifier. However, as a binary classifier, all points in its training set that are above the MEDIUM threshold are considered MEDIUMS even those that are actually also above the HIGH threshold. This is what causes the performance to drop when the 2 binary classifiers are fused into a ternary classifier. The MEDIUM classifier does really well on the points above the HIGH threshold as a binary classifier. However, when used as part of the fused classifier, it does not get credit for correctly classifying these points because these are classified by the HIGH classifier. Its performance on the remaining points is not as good (which is to be expected because there is less separation between these points and the LOW class), which causes the final performance to drop. It would be interesting in future work to investigate better methods of training the medium model so that the performance of the final ternary classifier on the MEDIUM class is closer to the performance of the binary MEDIUM classifier.