Disease SDMs with smaller extent

The folling demonstrates the potential pitfalls in modelling disease distributions using a virtual disease. The disease niche is set according to several environmental criteria and is modelled in a number of simulated scenarios. Some of the problems demonstrated below are well known in the ecological literature but the nature of these problems is often framed differently in epidemiological research and so should be highlighted in this particular context. The criteria for suitabile areas for this virtual disease will be as follows:

- Mean temperature between 18.0 and 22.5 degrees C
- Precipitation above 60 and below 170

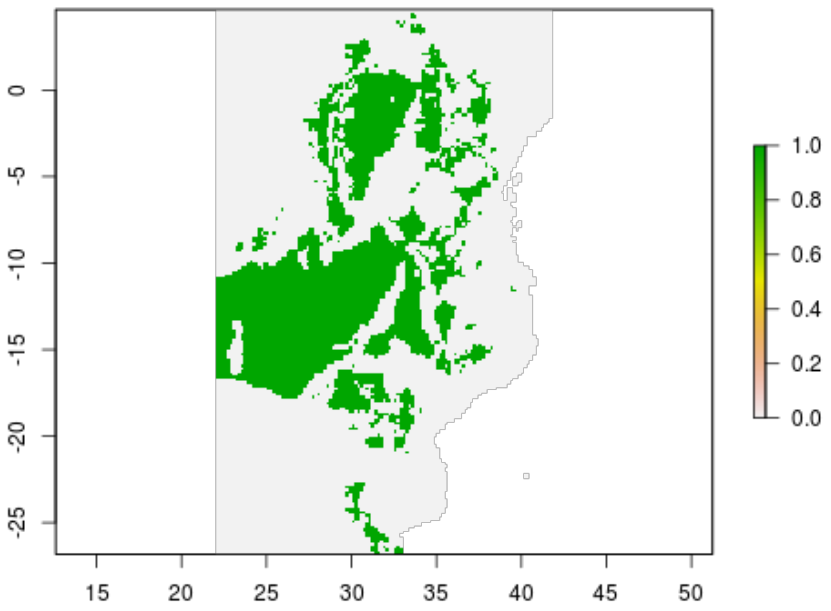

Figure 1: Suitable area for the virtual disease in africa

As the disease in this case is virtual, we are sure of the fundamental niche of this species and can assess model predictions for each of the scenarios against a known baseline (Figure 1). The environmental data used below are real data from Africa which is used as the geographic area for the examples. The environmental

predictors in this example come from WORLDCLIM and are cropped to the extent of the African continent (Figure 2). This is an appropriate choice as many of the diseases to which SDMs have been applied are present on the continent and it is likely that these mthods will continue to be applied to this region. In this demonstrated the following scenarios are simulated and assessed:

- *Full information* - The disease is in equilibrium with its environment and data is available for a spatially representative sample of its range
- *Heterogenous sampling effort* - The disease is in equilibrium with its environment but there is spatial bias in the detection of the disease (i.e. a heterogenous sampling effort)
- *Missing covariates* - The disease is in equilibrium with its environment and there is a spatially representative sample available but the covariates used for prediction do not fully reflect the species environmental constraints
- *Disease Control/Spreading* - The disease is not in equilibrium with its environment due to disease control or due to the current realised niche being smaller than the fundamental niche of the species as it spreads (i.e. there are FALSE negatives in the data)

Fieldwork for our modelling scenarios (Figure 3) constists of selecting random points as follows for the five scenarios:

- *Full information* - 300 random points are sampled from the true binary distribution of the disease at its full extent
- *Heterogenous sampling effort* - 200 random points are sampled from the true disease distribution within Kenya and a further 100 points from the rest of Africa
- *Missing covariates* - 300 random points are sampled from the true binary distribution of the disease at its full extent
- *Disease Control/Spreading* - 150 random points are selected from within Kenya and a further 150 random points from within Tanzania where the disease is not present (either due to eradication or the disease spreading). This model is then projected onto the whole of Africa

For each senario potential covariates were tested for collinearity. Generally the various measures of temperature were correlated strongly (Pearson's $c > 0.5$) with each other as well as precipitation. In these cases, precipitation was retained within the models as it was less strongly correlated with altitude (Pearson's $c < 0.5$ in all but one case). Generalised linear models are fitted for each of the scenarios using predictor variables as follows:

- *Full information* - Altitude and annual mean precipitation
- *Heterogenous sampling effort* - Altitude and annual mean precipitation
- *Missing covariates* - Altitude and mean temperature of the wettest quarter

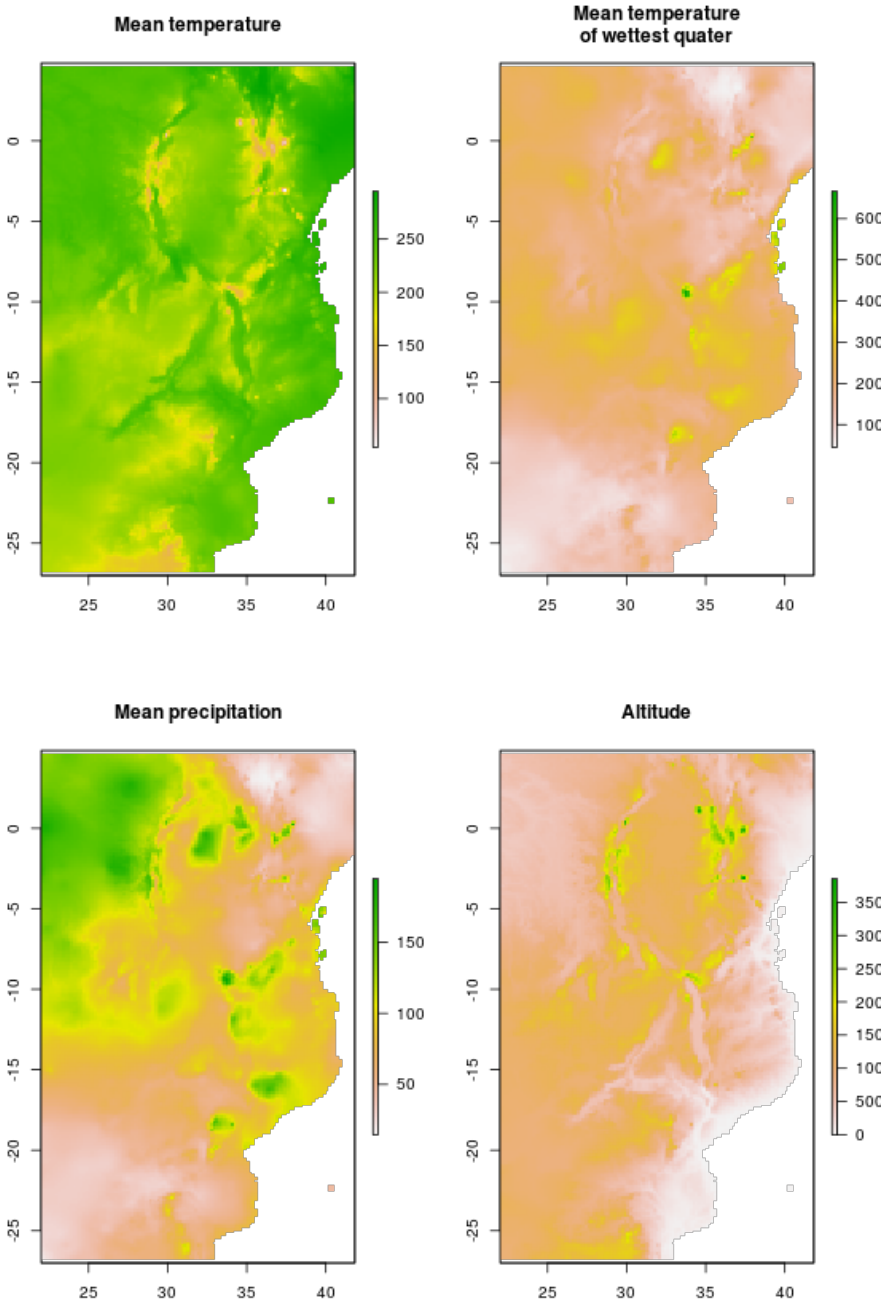

Figure 2: Temperature, rainfall and altitude surfaces for Kenya and Tanzania

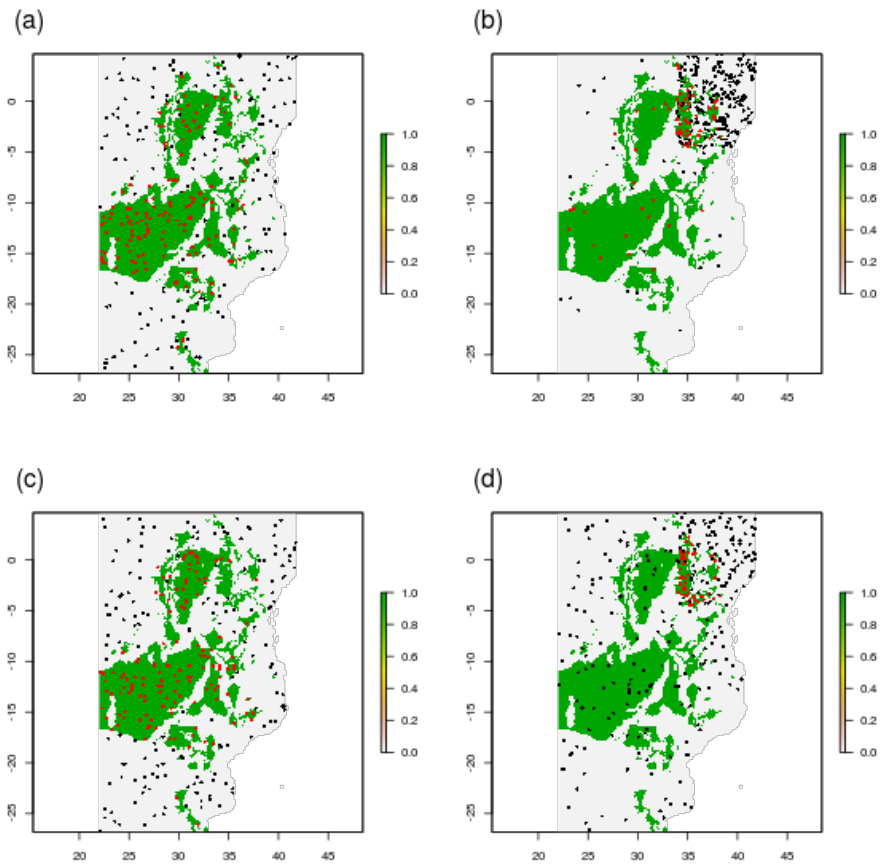

Figure 3: Fieldwork maps for the disease modelling scenarios. Positive records are shown in red and negative records in black

- *Disease Control/Spreading* - Altitude and annual mean precipitation

In all but one case the best model in terms of the lowest AIC score included an interaction between mean precipitation per year and altitude. In the *Disease Control/Spreading* scenario the model with the lowest AIC included only altitude as a predictor, however, the model inlcuding altitude and precipitation had delta AIC < 2 and was preferred for illustration purposes. Similarly, the best model for the *missing covariates* scenario included only the mean temperature opf the wettest quarter as a predictor however, the model including both altitude and mean precipitation in the wettest quarter had delta AIC < 2 so was chosen to maintain a level of consistency with the other scenarios.

Model evaluation is an important step in distribution modelling. In terms of prediction this should probably involve calculating more than one metric for evauluation but in this case, we will use one as we know the full distribution of the disease and have the luxury of being able to use random samples from the known distribution for testing rather than a subset of collected data. The receiver operating characteristic curve plots the rate of true positive results versus false positive. The area under this curve (AUC) provides a single value which represents the predictive performance of the model. The value is between 0 and 1 and bvalues above 0.5 represent better than random predictions with values over 0.7 conidered to indicate a "good" predictive model. The use of AUC values has been criticised in the literature recently but nevertheless is a commonly used method of model evaluation in the ecological literature (Fielding and Bell 1997). In each case the models are tested for predictive accuracy by converting their predictions into binary values (present or absent) and testing these against the true values for 1000 randomly selected points across Africa.

ROC curves and AUC scores for these modelling scenarios (Figure 4) suggest that the best performing model in this case is the *heterogenous sampling effort* scenarios, however, this is only slightly better than the *full information* scenario. The worst perfroming model by AUC score is the *control programme* scenario which performs worse than random. However, the *missing covariates* model only has an AUC of 0.5 becasue it produced no positive predictions (i.e. true positives = true negatives = 0).

By projecting the models onto data for the whole of Africa we can visually assess the performance of these models compared to the true distribution of the virtual disease (Figure 5). One feature of note is that the *missing covariates* model predicts a noticeably lower probability of presence and for a much wioder area than the other models. Also the model representing a *control programme* predicts a very low probability of presence in central Africa where we know the disease to be present. The model using *full information* predicts a wider area than the true distribution of the model to have a high proabability of presence. This could be due to the lqack of temeprature based predictors iin the model of becuse, as precipitation has a threshold over which it is suitable, the model would be imporved with the inclusion of nonlinear relationships. The *heterogenous*

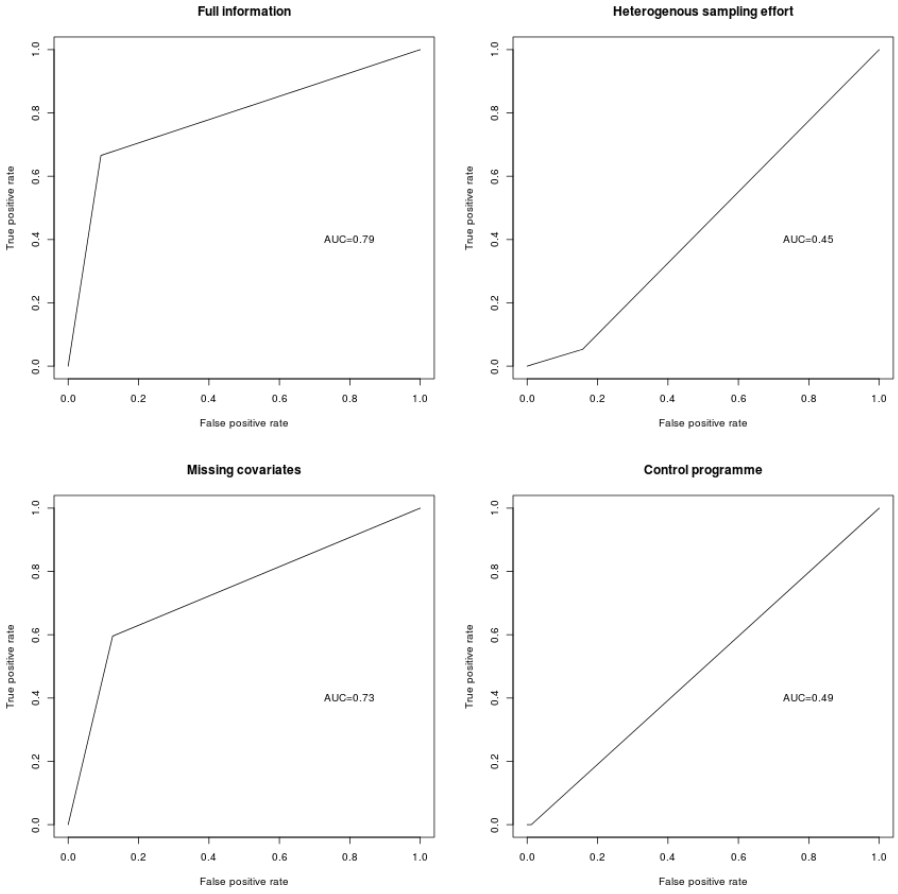

Figure 4: ROC curves for each modelling scenarios with AUC values given on each plot

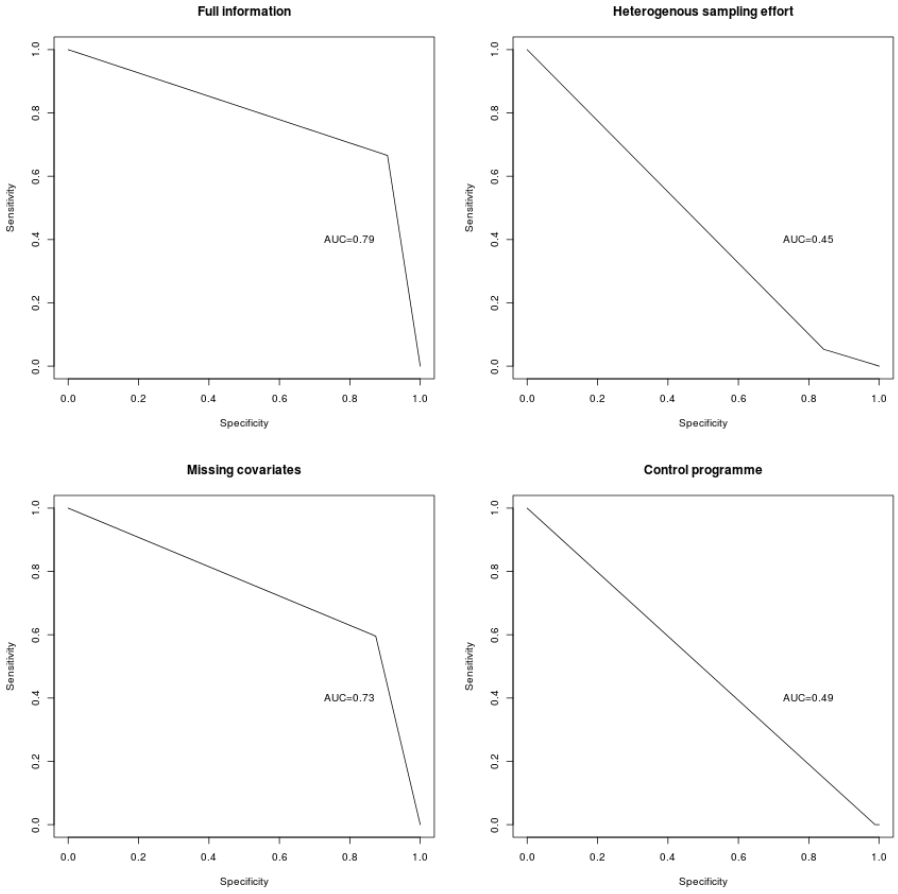

Figure 5: Sensitivity and specificity plot for each modelling scenarios with AUC values given on each plot

sampling effort model predicts a similar shaped area to the full information model but with a higher probability of presence outside of the true distribution. The *disease spreading* model predicts apatchy distribution covering some of the areas where the disease is known to be present but several further areas where the disease is absent including north Africa and the middle east.

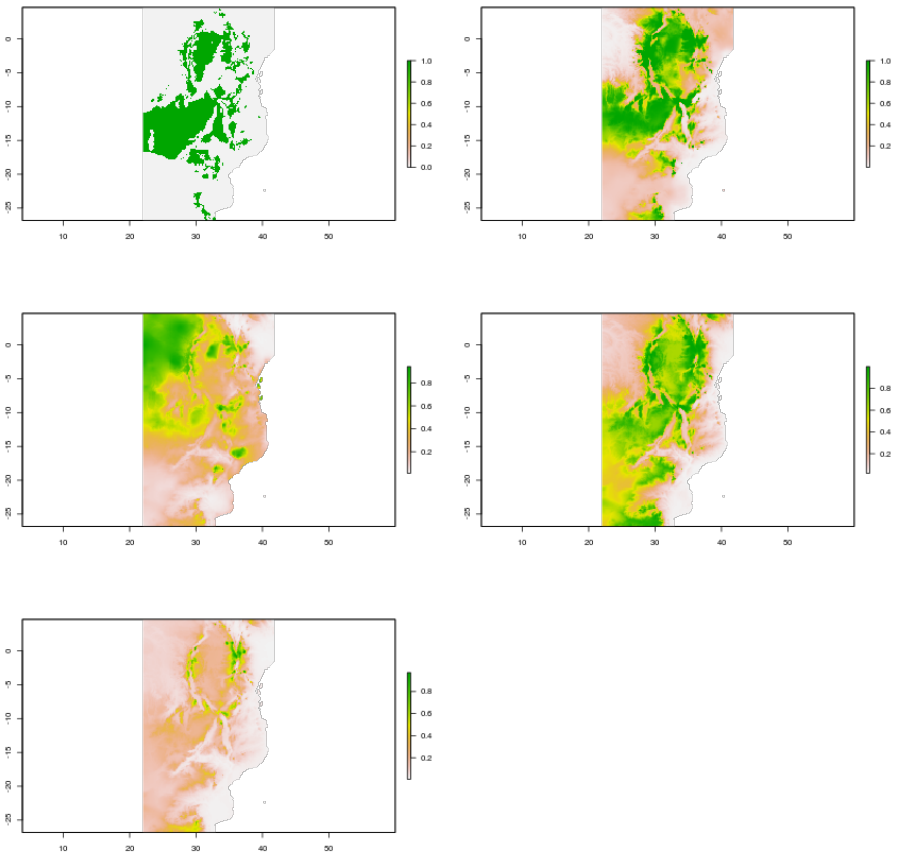

Figure 6: Predicted probability of disease presence across Africa for each of the modelling scenarios

The results of these projections are based on randomly placed points so there is the potential for variations in the modelling outcomes fo each scenario. To investigate the variability in these results and compare across scenarios the models are repeated 100 times each and assessed in terms of AUC azlong with the proportion of the study area correct, the proportion of false negative produced and the proportion of false positives.

Overall, the full model seems to perform the best in terms of the proportion of the study area predicted correct at the African extent (Figure 8a), for Kenya (Figure 9a) and for the combined Kenya and Tanzania extent (Figure 10a). Across Africa the missing covariates model performs better than might be expected in terms of proportion of the study area predicted correct (Figure 8a) but less well in terms of AUC score (Figure 8b). The full information model and the heterogenous sampling effort both have high AUC scores for the African extent (Figure 8b). The heterogenous sampling effort model also performs well at more limited extents with the full information model being noticeably less successful in Kenya (Figure 9b) and the combined Kenya extent where it is relatively low and Tanzania extent where it is highly variable (Figure 10b). At all extents the control measures model performs poorly in terms of AUC score and at the more limited extents the same is true for the missing covariate model. |The spreading disease model performs poorly across Africa (Figure 8) but considerably more successfully across both Kenya (Figure 9) and the combined Kenya and Tanzania extent (Figure 10).

Across Africa (Figure 8a & b) the full model and the heterogenous sampling effort model have low values for both false positive predictions (Figure 8a) and false negative predictions (Figure 8b). The missing covariate model has low fasle negatives but high false positives, the spreading diesease model has relativelt high false negatives and low false positives and the control measures model has high numbers of both. At the extent of Kenya both the full information model and the heterogenous sampling effort model have low numbers of false positives (Figure 9a) and moderate numbers of false negatives (Figure 9b) the proportion of which are highly variable for the heterogenous sampling effort model. The missing covariates model has a low number of false positives and a high number of false negatives. The spreading disease model and the control measures model both have high proportions of fasle positives but where the spreading disease model has low false negatives the control measures model has a larger proportion. In the combined Kenya and Tanzania region the full information model has highly variable proportions of false positives (Figure 10a) and a low proportion of false negatives (Figure 10b). The heterogenous sampling effort model and the spreading disease models have low proportions of both. The missing covariates and control measures model both have a low proportion of false positives but a high proportion of false negatives.

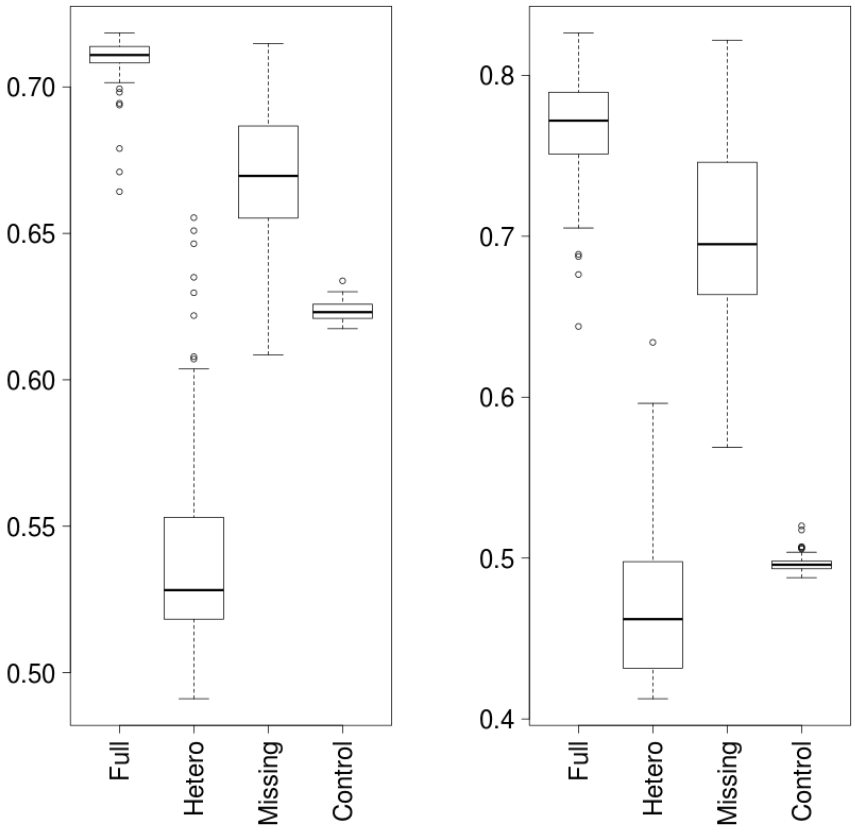

Figure 7: Results of 100 simulations for the five scenarios showing (a) Proportion of the study area correct, (b) AUC scores, (c) proportion of false positives and (d) proportion of false negatives across Africa

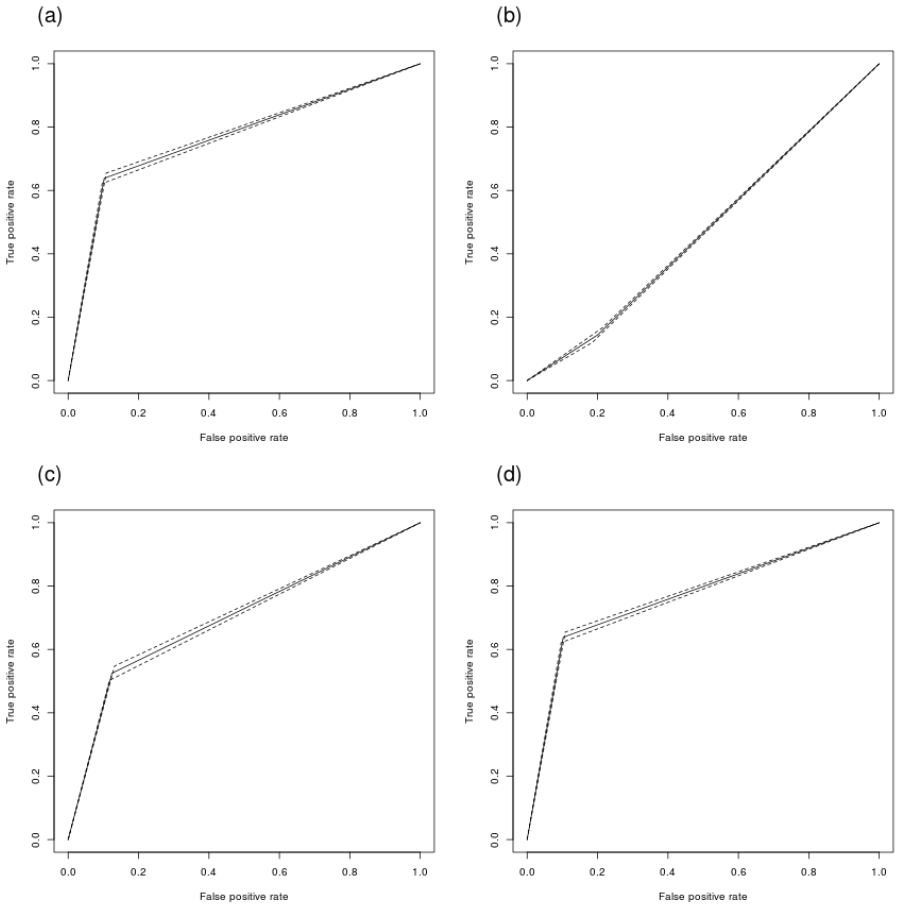

Figure 8: Mean ROC curves with 95% confidence intervals from 100 disease modelling simulations of the five scenarios showing (a) Full information, (b) Heterogenous sampling effort, (c) Missing covariates and (d) Disease spreading/Control programme

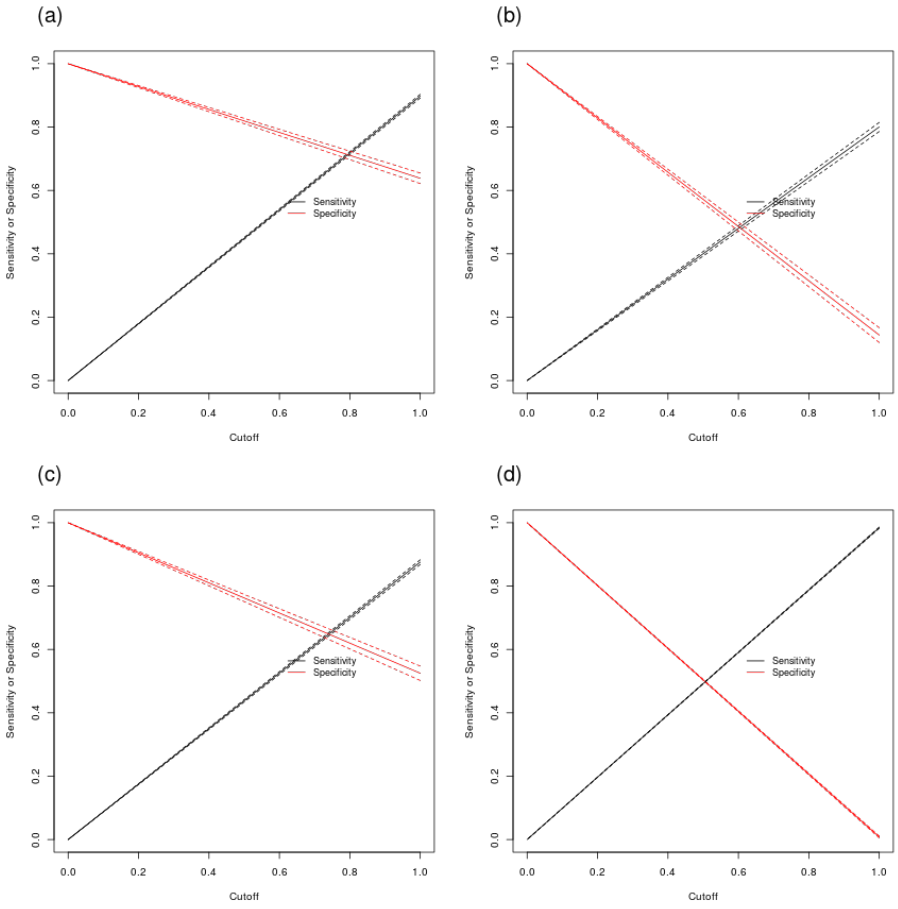

Figure 9: Mean sensitivity and specificity plots with 95% confidence intervals from 100 disease modelling simulations of the five scenarios showing (a) Full information, (b) Heterogenous sampling effort, (c) Missing covariates and (d) Disease spreading/Control programme