Proceedings of the Tenth International Conference on Urban Pests
Rubén Bueno,-Tomas Montalvo, and Wm. H Robinson (editors) 2022
CDM Creador de Motius S.L., Mare de Deu de Montserrat 53-59,
08930 Sant Adrià de Besòs, Barcelona, Spain

TOWARDS DEFINING THE IDEAL SAMPLE SIZE FOR SPATIAL DISTRIBUTION MODELING USING A VIRTUAL VECTOR

GUY HENDRICKX, ALIZÉE HENDRICKX, WESLEY TACK AND CEDRIC MARSBOOM
Avia-Gis, Risschotlei 33, 2980 Zoersel, Belgium

Abstract The objective of this work is to use a virtual disease vector at the scale of the European continent for the empirical testing of the number of spatial sampling points that are needed to develop stable spatial models to reliably map the distribution of disease pests: How many sampling points should one need to take to build stable spatial models? Despite the fact that this is crucial information to enable the cost-efficient planning of spatial sampling campaigns little has been published on this topic. To achieve this objective, we first designed a 'virtual vector' with a known distribution based on a set of climatic and environmental data at a 1X1 km grid. We then selected 10 areas of 400X400 km representative of different ecoclimatic settings of our virtual vector in Europe. In each of these areas we computed 500 models with sample sizes varying from 10 to 5000 randomly selected sampling points. In this paper we present preliminary results and discuss how these results may contribute to improve the quality and reduce the cost of spatial distribution models and guide our future research.

Key words Spatial model, spatial sampling

INTRODUCTION

Worldwide Emerging Infectious Diseases in general and mosquito transmitted vector-borne diseases in particular, are gaining importance (Jones et al., 2008). Obtaining high quality field information about the presence (and absence) of such disease vectors is notoriously costly and time-consuming. These costs can significantly be reduced by combining cost-efficient sampling strategies, remote sensing and spatial modelling techniques to compute spatial distribution maps of vector presence and/or abundance. Similar approaches also allow for the identification of high-risk zones for the introduction, establishment and spread of exotic species at a local or regional level.

Two important but often ignored factors are critical to develop reliable spatial distribution models: sample size and sample location. Most of the time spatial risk maps are based on an amalgamate of conveniently available historical data sets that were collected for other purposes. As a result, these data are rarely representative for the entire area covered resulting in (a) gaps in areas that are eco-climatically under-represented, and (b) clusters in some other areas which are over-represented. In this paper we focus on defining the number of randomly distributed samples that are needed within a given area to compute a stable spatial model.

MATERIALS AND METHODS

To answer our research question *'How many sampling points should one need to take to build stable spatial models?'*, we first generated a virtual vector within a clearly defined eco-climatic envelope in Europe. This was done based on a R package developed by Leroy et al. (2006). The ecoclimatic variables (1X1km grid) used to define the distribution limits included: altitude (DEM) and hill slope; de distribution of small ruminants (Gilbert et al., 2018); satellite data derived from the MODIS sensor (Salomonson, et al. 1989): snow cover, Fourier processed day-time surface temperature LST Day and Fourier processed vegetation index NDVI; climatic data derived from WorldClim (Fick and Hijmans, 2017): maximum temperature of warmest month – bio5, annual precipitation – bio12, precipitation seasonality – bio15, precipitation of wettest quarter – bio16. To avoid harsh (threshold) classifications a probabilistic response curve was computed for each predictor variable. An example of such a response is given in figure 1. By using the soft (probabilistic) classification we get a more realistic classification in edge zones between presences and absences. In Figure 2 the computed distribution model of the virtual vector is given, in this example a p value of 0.5 was used to discriminate between presence and absence.

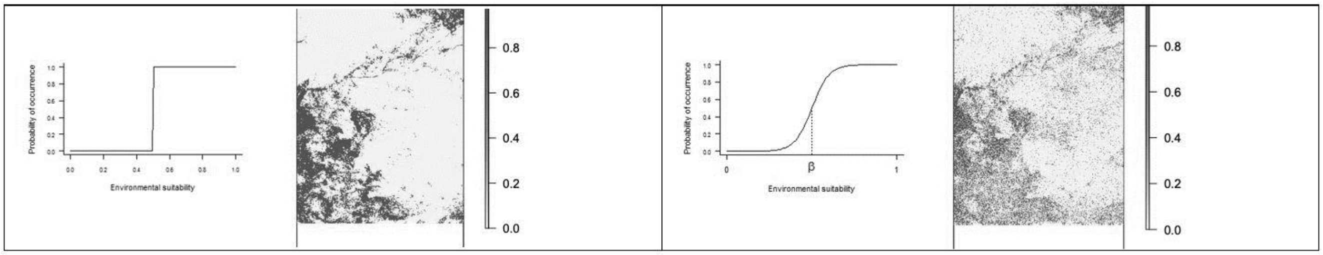

Figure 1 Effect of harsh (left) and soft (right) classification on distribution patterns.

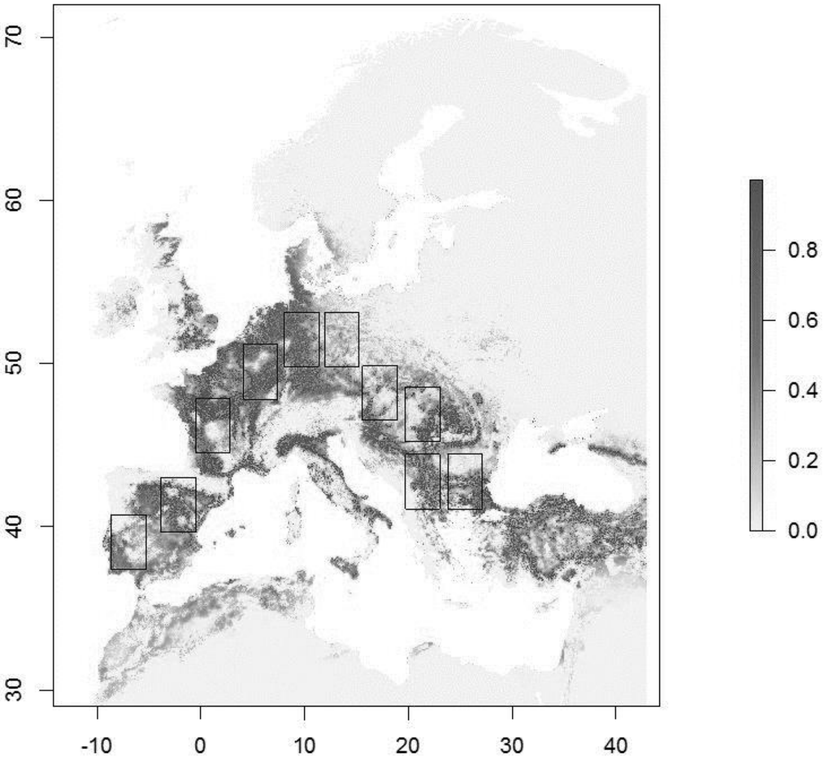

Figure 2. Computed distribution of virtual vector and selected test areas.
A p value of 0.5 was used to discriminate between presence and absence.

As a next step 10 areas of 400X400 km (i.e. 160 000km2) were selected that are representative of the various eco-climatic settings in which our virtual vector occurs (Fig.2). In each of these areas 500 models (regression forest) were run based on randomly selected sampling points ranging from 10 to 5000 samples, and a data set of eco-climatic predictor variables. Each time a comparison was made for results obtained with a balanced and un-balanced sample, respectively defined as equal number of presence and absence points and 10% of present points + 90% of absent points. Results were tested using a variety of standard statistic methods including: the area under the curve (AUC), the percentage correct predictions (PCC), the sensitivity (i.e. proportion of correctly predicted presence) and the specificity (i.e. proportion of correctly predicted absence).

RESULTS

The obtained results are summarised in Figure 3.

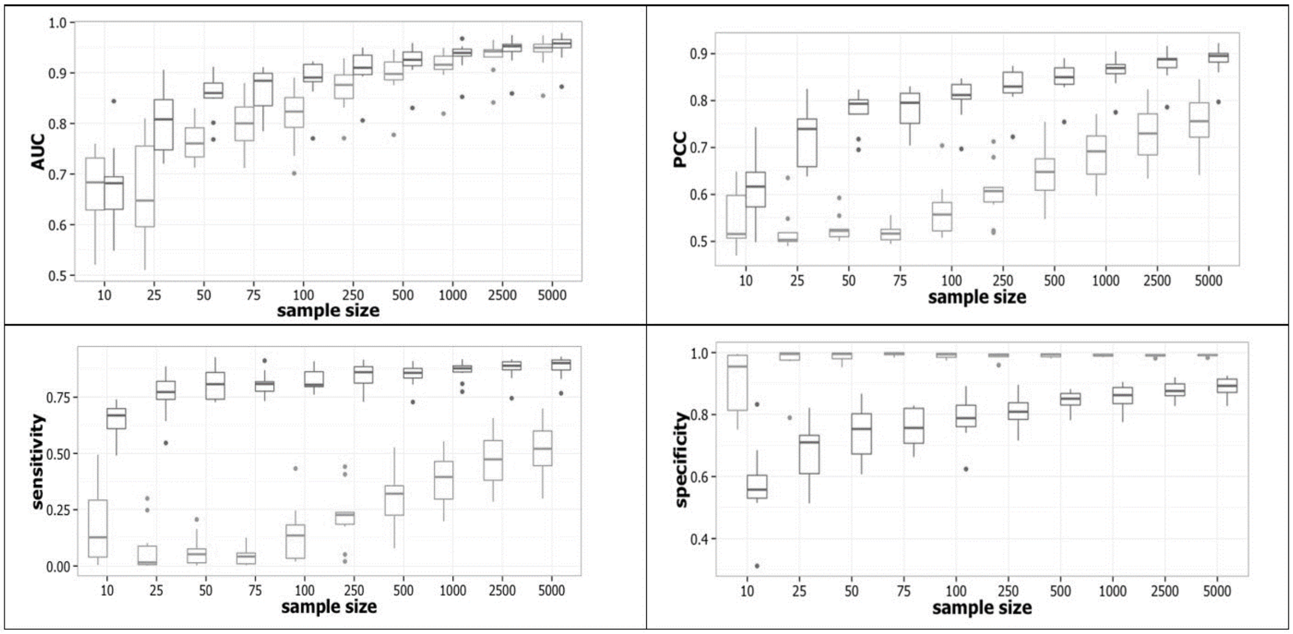

Figure 3. Box plots for respectively AUC, the percentage correctly classified grid cells, the sensitivity and the specificity. Each time respectively for balanced (blue) and unbalanced data sets.

In all, except one case, balanced data set perform better than unbalanced data sets. This is marginal for the AUC statistic, the difference is significant for the percentage of correct classified grids and for the sensitivity. The reverse is valid for the specificity, but this is due to the unbalanced data set is biased towards negative sample points.

In our results we use box plots to assess the stability of the computed model outputs: when the boxes are consistently narrow, the model outputs may be considered stable. Results obtained to date allow for the following preliminary qualitative assessment: For the AUC model stability is reached above a value of 0.90, this relates to a sample size between 100 and 250, for the PCC model stability is reached above a value of 0.80, this relates to a sample size between 100 and 250, for sensitivity model stability is reached above a value of 0.75, this relates to a sample size between 75 and 250, for specificity model stability is reached above a value of 0.80, this relates to a sample size between 100 and 500

DISCUSSION

In this paper we provide a preliminary qualitative analysis of obtained results. Work is ongoing to further quantify our analysis. The conclusions reached at this stage are summarized below. It is of critical importance to use balanced data sets with about the same number of presence and absence data points when computing spatial distribution models. This factor is often overlooked and raises the issue of filtering field data, and/or computing additional negative or positive data sets based on observed eco-climatic envelopes prior to modelling.

The sample size required to produce stable species distribution models for an area of 160K km2 is surprisingly low. A critical condition is to select all sampling points randomly within the given area. This is paramount to any spatial modeling exercise. Based on expert advice sampling strata may be identified, but within these strata sampling points should always be selected randomly. It is easier to correctly predict presence as opposed to absence. This may be due to the fact that we used a threshold of P0.5 to discriminate between presence and absence of our virtual vector and may be related to a classic issue in medical entomology: false negatives. More work is required to analyse this further. Ongoing work is aimed at (a) further refining and quantifying the obtained results, (b) evaluate the impact of spatial heterogeneity on spatial sample size and distribution, and (c) evaluate the impact of spatial data clusters in data bases used for spatial modeling. Our final objective is to develop an improved spatial sampling strategy module to integrate in the VECMAP® software package.

ACKNOWLEDGEMENTS

We are grateful to the Avia-GIS team and our VECMAP® champion users for constructive and challenge discussions on this (and other) topics.

REFERENCES CITED

Fick, S.E. and R.J. Hijmans 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 10.1002/joc.5086.

Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T.P., Vanwambeke, S., et. al. 2018. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. In: Scientific Data, Vol. 5, p. 180227.

Jones, K., Patel, N., Levy, M. et al. 2008. Global trends in emerging infectious diseases. Nature 451, 990–993 (2008). https://doi.org/10.1038/nature06536

Leroy, B. et al. 2016. Virtual species, an R package to generate virtual species distributions. Ecography. 39(6):599-607.

Salomonson, V. V., Barnes, W. L., Maymon, P. W., Montgomery, H. E., and Ostrow, H. 1989. MODIS: Advanced facility instrument for studies of the Earth as a system. IEEE Transactions on Geoscience and Remote Sensing, 27(2), 145-153.