ECDC TECHNICAL REPORT

Environmental risk mapping: *Aedes albopictus* in Europe

Proof-of-concept study for the
European Environment and Epidemiology Network

This report was commissioned by the European Centre for Disease Prevention and Control (ECDC), coordinated by Jonathan Suk and Jan Semenza, with input from Bertrand Sudre, Wim Van Bortel, Laurence Marrama and Hervé Zeller (all at ECDC), and produced by David Rogers and Willy Wint (University of Oxford, United Kingdom).

Suggested citation: European Centre for Disease Prevention and Control. Environmental risk mapping: *Aedes albopictus* in Europe. Stockholm: ECDC; 2013.

Stockholm, March 2013

ISBN 978-92-9193-449-2

doi 10.2900/78239

Catalogue number TQ-32-13-122-EN-C

Cover photo: John Beetham, Creative Commons-licensed content

Contents

Figures

Abbreviations

E3	European Environment and Epidemiology Network
EEA	European Environment Agency
MD	Mahalanobis distance
NLDA	Non-linear discriminant analysis
NUTS	Nomenclature of Territorial Units for Statistics

Introduction

Scope and purpose

This report has been designed to discuss the usefulness of environmental risk mapping for public health. The report demonstrates the application of non-linear discriminant analysis (NLDA) to examine how the following question could be answered by environmental risk mapping: 'How might we study and monitor *Aedes albopictus*, in order to be better prepared for outbreaks of diseases vectored by this species in the future?'

What is this document for?

This is a technical report published with the intent of informing readers about the ways in which the data contained in the ECDC European Environment and Epidemiology Network (E3) can be used for risk mapping of *Ae. albopictus* – as well as infectious disease vectors more generally – in continental Europe. The specific example of risk mapping given in this report is non-linear discriminant analysis.

Who is this document for?

This document has been prepared for public health institutions and scientists experts dedicated to the control of vector-borne diseases. This report should be of particular relevance to those working with vector-borne diseases in EU Member States and neighbouring countries, but it may also be relevant to a broader international audience. It may also be of relevance to those tasked with establishing surveillance protocols for invasive mosquito species.

Why was it commissioned?

This project was commissioned in 2010 in order to provide insight into two streams of ECDC work. One was to support the development of European Environment and Epidemiology (E3) Network, and the other is ECDC work focused on emerging and vector-borne diseases.

The project was financed solely by the European Centre for Disease Prevention and Control (ECDC), and the project was undertaken by the University of Oxford, United Kingdom, in collaboration with ECDC technical staff.

1 Background

Risk mapping with geospatial data is a rapidly growing area in public health. With the aim of facilitating information exchange on the risk mapping of infectious diseases in Europe, ECDC launched the European Environment and Epidemiology (E3) network in 2013. This network enables users to obtain the latest project results from ECDC and its partner organisations and provides access to hundreds of datasets relevant to the risk mapping of infectious diseases.

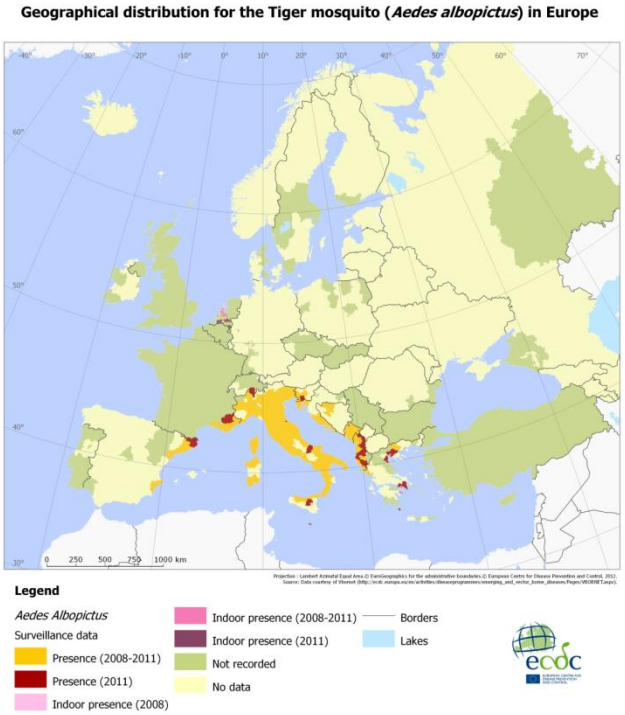
To evaluate and demonstrate the practice of risk mapping, this report describes the process involved in modelling the potential distribution of the Asian tiger mosquito (*Ae. albopictus*) in Europe. The main section describes the selected data sources, the formatting of environmental data, and the production of a series of risk maps. The discussion surveys the technical decisions made when producing risk maps, and the appendix provides a tutorial on risk mapping based upon non-linear discriminant analysis (NLDA).

Ae. albopictus, arrived in Europe in Albania in 1979, and then Italy in the early 1990s via trade in used tires[1]. *Ae. albopictus* is, today, principally present in the northwest Mediterranean basin (Figure 1)[2]. The expansion of the vector in Europe has been driven by global trade and travel between climatically similar regions[3], and it has been speculated that future European expansion of *Ae. albopictus* could be further facilitated by climate change, as altered warming and precipitation patterns might increase the number of suitable niches for the vector[4]. Irrespective of future climatic changes, a series of risk maps published by ECDC in 2012 suggest that *Ae. albopictus* has yet to fill its realisable niche in continental Europe[5].

Ae. albopictus is a less effective epidemic dengue vector than *Ae. aegypti* [6], but in recent years it has nonetheless been responsible for transmitting both dengue and chikingunya fever in continental Europe, including over 200 laboratory-confirmed cases of chikungunya in Italy in 2007[7] and local dengue transmission in Croatia and France.

Due to its potential public health significance, there are currently numerous activities in Europe to consider the impact that changing climatic and environmental variables could have on the distribution of *Ae. albopictus* in Europe. Similarly, organisations responsible for public health and vector control will need to be able to interpret the results from such initiatives so as to inform vector surveillance activities[8]. In this report, the processes for producing risk maps based upon the data stored in the ECDC European Environment and Epidemiology network (E3) and using NLDA methodologies will be discussed, as will the implications of the resultant risk maps for public health initiatives, for example in the areas of monitoring and early warning.

Figure 1. Recorded presence and absence of *Aedes albopictus* in continental Europe



Note: Indoor presence corresponds to presence recorded in greenhouses.
Data courtesy of VBORNET. Copyright © European Centre for Disease Prevention and Control, 2012

2 Selecting data for environmental risk mapping for non-linear discriminant analysis

2.1 Mosquito data

A number of recent initiatives have either gathered together existing information on European mosquitoes (e.g. the NBN gateway in the UK[i]; Tigermaps, funded by ECDC; or MosquitoMap in the USA[ii]) or have set out to gather information anew (e.g. MODIRISK in Belgium[iii] and a parallel initiative in the Netherlands).

ECDC has established the VBORNET initiative aimed at obtaining EU-wide vector distribution data, but in some instances the distribution data on disease vectors can be patchy or simply unavailable.

2.2 Environmental data

The studies described in this report received inputs from a wide range of environmental data from many different sources. These inputs needed to be standardised in terms of coverage and resolution, and thus required significant effort to acquire, process and archive. The availability of data archives such as the E3 Network therefore greatly facilitates the work flow and substantially enhances productivity. The following represents not only those datasets that have been used in these studies, but also identifies some of those that would be useful for work of this type.

Meteorological data

Interpolated meteorological data (temperature, rainfall, relative humidity/saturation deficit, wind, sunshine hours) at the highest available resolution (1/6th degree or better). These are available from a number of sources – notably Worldclim (www.worldclim.org) and the Climate Research Unit (http://www.cru.uea.ac.uk) – and processed versions suitable for modelling are available from the ECDC E3 network archive.

Low-resolution satellite data

Examples of low-resolution satellite data include MODIS, with a resolution down to 250m; SPOT multi-temporal, with a resolution down to 1km; AVHRR, with a resolution down to 1km; and MERIS data, with a resolution down to 300m. The MODIS and AVHRR data sets are available from the EDEN and E3 archives as Temporal Fourier Analysed outputs (see Annex 2 for processing details).

An additional dataset (CMORPH[iv]) has been used as an index of rainfall, using precipitation estimates that have been exclusively derived from low orbiter satellite microwave observations, and whose features are transported via spatial propagation information that is obtained entirely from geostationary satellite IR data. Note that this technique is not a precipitation estimation algorithm but a means by which estimates from existing microwave rainfall algorithms can be combined. Therefore, this method is extremely flexible such that any precipitation estimates from any microwave satellite source can be incorporated.

High-resolution satellite data

High-resolution satellite data (LANDSAT or SPOT, resolutions down to 10m). RGB versions of the LANDSAT imagery are held in the in the E3 archives.

Elevation data

Digital Elevation Models (DEM), notably the 90m-resolution datasets derived from the Shuttle Radar Topographic Mission are available within the E3 archives.

Land-cover data

Land-cover maps for Europe derived from satellites are available in several flavours, an example of which is the CORINE 100m-resolution land use map held at the European Environment Agency (EEA) for 1995, 2000, and 2005. These data, however, only cover the EU Member States and thus need to be supplemented with other datasets.

[i] http://data.nbn.org.uk

[ii] http://www.mosquitomap.org

[iii] http://www.modirisk.be/modirisk/GeneralSite/default.aspx?WPID=183&L=E&miid=174

[iv] Joyce RJ, Janowiak JE, Arkin PA, Xie P. CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydromet. 2004;5:487-503

Suitable candidates are GLOBCOVER from ESA, for which a new 2009 version is in preparation, or the Global Land Cover2000 from the UN system. These constituent elements are available in the E3 archives.

Soil data

Soil maps, especially those from which may be derived information about fertility, pH and particle size (affecting vegetation cover, pH and turbidity of water bodies). Some appropriate data are available from the European Soils Database, hosted at the JRC[v], and soil moisture information can be derived from METEOSAT satellite imagery and a number of climate change databases such as ENSEMBLE.

2.3 Environmental data processing

The environmental data were processed in several ways to produce data layers that may in some way be more suitable for modelling purposes. The overall strategy of such data processing is threefold:

- To produce orthogonal data sets from serially correlated multi-temporal data. PCA and temporal Fourier analysis (TFA) both achieve this objective, but the latter is preferable because it does so in more biologically interpretable ways.
- To determine environmental discontinuities that may act as a barrier to the spread of a vector and/or disease. Such discontinuities may be identified by wavelet analysis of single images, which may be regarded as the spatial equivalent of TFA. Wavelet analysis is preferable to Fourier analysis of single imagery (which is an option in many remote sensing packages) because it provides a local rather than global analysis. Since wavelet analysis is also a hierarchical technique, it is possible to produce multi-scale measures of environmental discontinuities, and at a scale appropriate for the species concerned.
- To determine the spatial structure of suitable habitat types. Packages such as FRAGSTATS[vi] are able to identify clumps or blocks of particular vegetation types, and to summarise this information in the form of mean area of block, mean perimeter, etc.

2.4 Ancillary data

Demographic data

Human population density and distribution at as fine a scale as possible, currently most widely available as 1km-resolution raster layers provided by the Global Rural Urban Monitoring Project (GRUMP) produced by the Gridded Population of the World project[vii], and held in the E3 archives.

Transport networks and travel

The current study provides a way of assessing the likelihood of spread that does not rely on knowledge of the transport infrastructure and the number of travellers using it. Such information may, however, prove to be useful additional indicators of spread risk for some diseases.

Layers for major road and rail networks, and ports of entry to the EU, can sometimes be acquired from public domain sources such as VMAP0, and in due course form the forthcoming higher resolution VMAP1, derived from the 1:250 000 Joint Operation Graphics (JOG) aerial navigation maps.

Host data

Data for the distribution and abundance of any hosts (birds; mammals both domestic and wild) likely to be fed upon by the mosquitoes, and information on the likelihood that such hosts might act as disease reservoir species. These are so far relatively difficult to find, especially at a continental level and at an acceptable resolution. They are likely, however, to be a focus of attention by the forthcoming FP7 EDENext project, which will collaborate closely with the ECDC E3 data archiving activities.

[v] http://eusoils.jrc.ec.europa.eu/ESDB_Archive/ESDB/index.htm

[vi] http://www.umass.edu/landeco/research/fragstats/fragstats.html

[vii] http://sedac.ciesin.columbia.edu/gpw/

3 Interpreting model results

3.1 Example model output: *Aedes albopictus* in Europe

As an example of the above exercise, this section highlights the conclusions of a map produced as part of the ECDC project modelling of *Aedes albopictus'* distribution[5], using firstly just one database of this species' invasion, so far, of Europe. The risk map is shown in Figure 2. Because the presence records are most numerous in, and more or less restricted to, northern Italy, the risk map highlights this area as being suitable for *Ae. albopictus*, with much of the rest of Europe apparently unsuitable.

The key variables across the 100 bootstrap models are shown in Table 1. Rainfall and temperature variables play a major role in this model. The WORLDCLIM rainfall phase 2 variable has a mean rank of 1.04 across all 100 bootstrap models, indicating that it was chosen as the first variable in virtually all 100 models. The second and subsequent variables have much lower mean ranks indicating that, although they were important, they were not consistently selected in the same rank order across the 100 models.

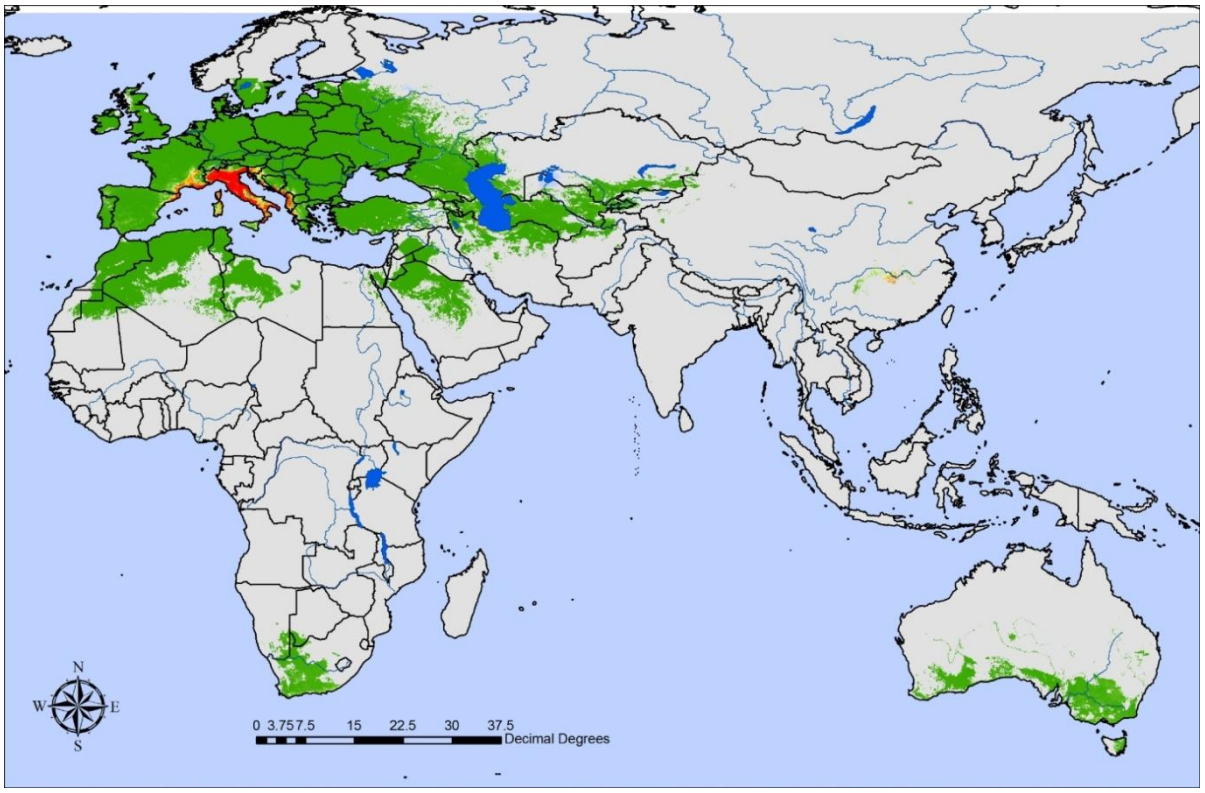
Table 1. Model 2, *Aedes albopictus*: overall ranking of variables

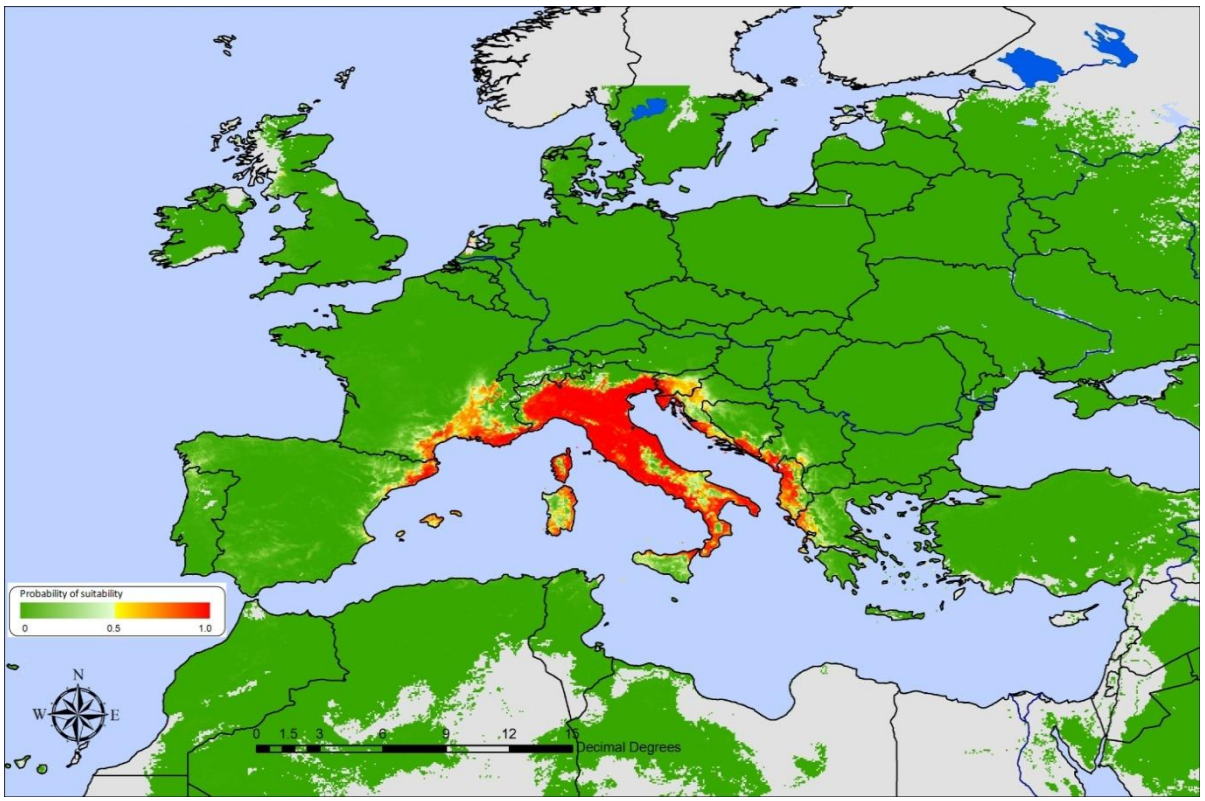
Variable	Rank	Description
wj0857p2	1	Rainfall from WORLDCLIM database (biannual phase)[9]
wj1557a0	2	Day time land surface temperature from MODIS (mean, 2001–2005)[10]
wj1508mx	3	Night time land surface temperature from MODIS (maximum)
DEM	4	Digital elevation model from SRTM
wj1507a0	5	Day time land surface temperature from MODIS (mean)
mmph50p2	6	Precipitation estimates processed acc. CMORPH (biannual phase)
wj0857a3	7	Rainfall from WORLDCLIM database (triannual amplitude)
wj1508p3	8	Night time land surface temperature (triannual phase)
wj0857p3	9	Rainfall from WORLDCLIM database (triannual phase)
wj1507mn	10	Day time land surface temperature from MODIS (mean)

Table shows the mean rank values of the top-ten ranked variables from all 100 bootstrap models.

Figure 2. Risk map for Model 1, *Aedes albopictus*

This risk map is the average of 100 bootstrap models each based on a sample of 200 presence and 200 absence pixels selected at random, with replacement from the training for this vector (see text for more details). Risk is on a probability scale from zero to 1.0. Probabilities from 0.0 to 0.49 are coloured green (darker to lighter) and indicate conditions not suitable for the vector (i.e. predicted absence of the vector). Probabilities from 0.50 to 1.0 are coloured yellow through to dark red, indicating conditions increasingly suitable for the vector. Grey indicates no prediction was undertaken.





3.2 *Aedes albopictus* in Europe: detailed analysis

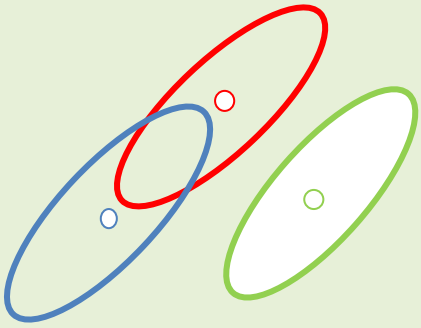
We can examine the performance of this model in more detail. Figure 3 shows a detail of the predictions centred on Italy. The upper figure reproduces part of Figure 2 but includes the presence points here identified by clusters. As mentioned above, clustering is a way of rendering non-multivariate data multivariate normal for discriminant analysis. The fact that here we see the three different clusters for presence arranged in an obvious geographical pattern illustrates the rather different habitats experienced in these different areas (for example, differences of more than 2.5 °C in mean daytime land surface temperature values across the three presence clusters).

The lower part of Figure 3 shows the same datapoints on top of the Mahalanobis distance (MD) (Box 1) image for this *Ae. albopictus* risk map. Once again, the MD image is the mean of 100 such images from the 100 bootstrap models. Inspection shows that none of the presence datapoints fall in places with MDs greater than about 40 (the MD colour scale is indicated in the legend on the lower part of Figure 3). In other words, an excellent fit to the European data for *Ae. albopictus* is obtained from a model that selected a suite of variables that could characterise the European presence data fairly precisely in multivariate space.

But is this what we want from a risk map of an invasive species? As the experience in North America shows, *Ae. albopictus* first established in limited and presumably very favourable conditions, and later spread into a much wider range of habitats. What will this 'wider range of habitats' be for *Ae. albopictus* in Europe?

What is a Mahalanobis distance and how can it be used in risk mapping?

The Mahalanobis distance between two sets of points (in the present case defining vector presence and absence) is the environmental distance between the two sets of points adjusted for the covariance of the variables. When Pythagoras defined the length of the hypotenuse of a right triangle, he was defining the linear or Euclidean distance between two points (two vertices of the triangle). Pythagoras' theorem can be used to determine the linear distance between any two points on a flat plane. However, when the two points define the mean or 'centroid' of two clusters of points, the Euclidean distance is not always the best measure of separation, as illustrated in the diagram below.

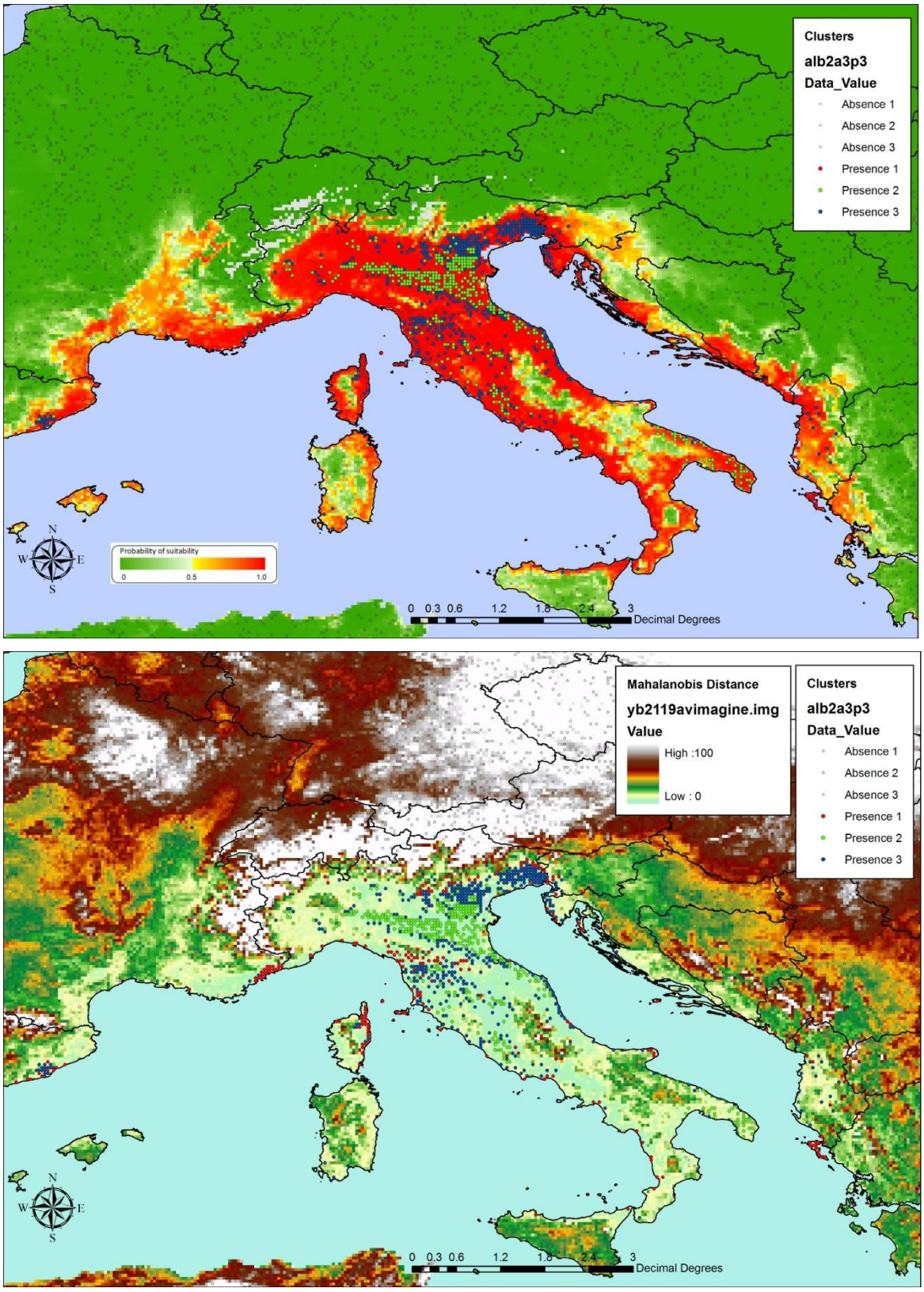


The three ellipses define the confidence intervals of three different clusters of points. The centroid of each ellipse is indicated by the small circle. The Euclidean distance between the red and green circle is the same as the Euclidean distance between the red and blue circles. But there is much more overlap between the red and blue ellipses than between the red and green ellipses. In some sense, the blue cluster is 'closer' to the red cluster than the green cluster. This greater closeness is captured by the alternative measure of separation in multivariate space – the Mahalanobis distance (MD) – which is defined as the distance between two clusters adjusted by the covariance of each cluster.

Prasanta Chandra Mahalanobis (1893–1972) was an Indian mathematician, founder of the Indian Statistical Institute. The Mahalanobis distance is arguably the best measure of separation in multivariate space of co-varying variables.

Figure 3. Model 2, *Aedes albopictus*

The upper figure shows a detail from the risk map in Figure 2. The individual clusters of presence are identified by the three colours shown in the legend (red, green, blue); the pseudo-absence points are indicated by dots. The lower figure shows the same points on the Mahalanobis distance image from the same model (the average of 100 such images from the 100 bootstrap models). The presence points fall within the lower range of MD values (approximately < 40)



3.3 Employing risk maps to predict future vector distributions

We can begin to answer the question posed at the end of the previous section by first looking at the global risk map for *Ae. albopictus* (Figure 4). This was produced by amalgamating five different regional and global risk maps for this species[5] and is shown in Figure 3, both with and without the global dataset for the presence of this species. Figure 5 shows the MD image for this combined model. This MD image was produced by combining the five averaged MD images for each of the five models. Combination involved selecting from among the five the lowest MD value. This is similar in concept to ensemble climate models which combine the outputs of several models into a single output – in this case indicating the minimum distance produced by any of the models where *Ae. albopictus* has been recorded. It can be seen in this global model that *Ae. albopictus* occurs up to higher values of Mahalanobis distances than was the case for Europe. This is hardly surprising given that the European risk map was regional whereas Figure 4 is global, but the significance is that if *Ae. albopictus* can inhabit a wider range of habitats globally, then it can also inhabit a wider range of habitats locally, within Europe, than it has done so already.

Details of the global map in Figures 4 and 5 are shown in Figure 6 for the region of northern Italy. The maps include the presence/absence dataset of the European model of Figure 3. The upper map in Figure 6 shows that far more of Europe appears suitable for this species, based on its global distribution data, than was apparent in Figure 4. This is supported by the MD image in the lower part of Figure 6. On the basis of *Ae. albopictus'* global distribution, the contrast between northern Italy and other parts of Europe is not so great environmentally as was suggested by Figure 3 (lower). This map suggests not only that *Ae. albopictus* is likely to spread beyond its present limits in southern Europe, but also the directions in which this spread is most likely. In other words we might be able to plan a monitoring and surveillance scheme for the future spread of this species based on the types of imagery and analysis shown in Figure 6.

Figure 7 (upper) shows the same MD image as in Figure 6 and (lower) which of the five contributory models these values came from (for clarity the data points are omitted). For large areas of Europe (red and green) the lowest MD values come from models based essentially on global datasets (Models 1 and 4, respectively). Relatively few areas are derived from the European-only model, Model 2 (pale yellow areas in Figure 7, lower).

Figure 4. Global model for *Aedes albopictus*

This is the final global map for this species produced by amalgamating the results of five different (global and regional) models. Upper with all the presence points used in all five models, lower without the presence points. Risk is on a probability scale from zero to 1.0. Probabilities from 0.0 to 0.49 are coloured green (darker to lighter) and indicate conditions not suitable for the vector (i.e. predicted absence of the vector). Probabilities from 0.50 to 1.0 are coloured yellow through to dark red, indicating conditions increasingly suitable for the vector. Grey indicates no prediction undertaken. (Reproduced from ECDC[5].)

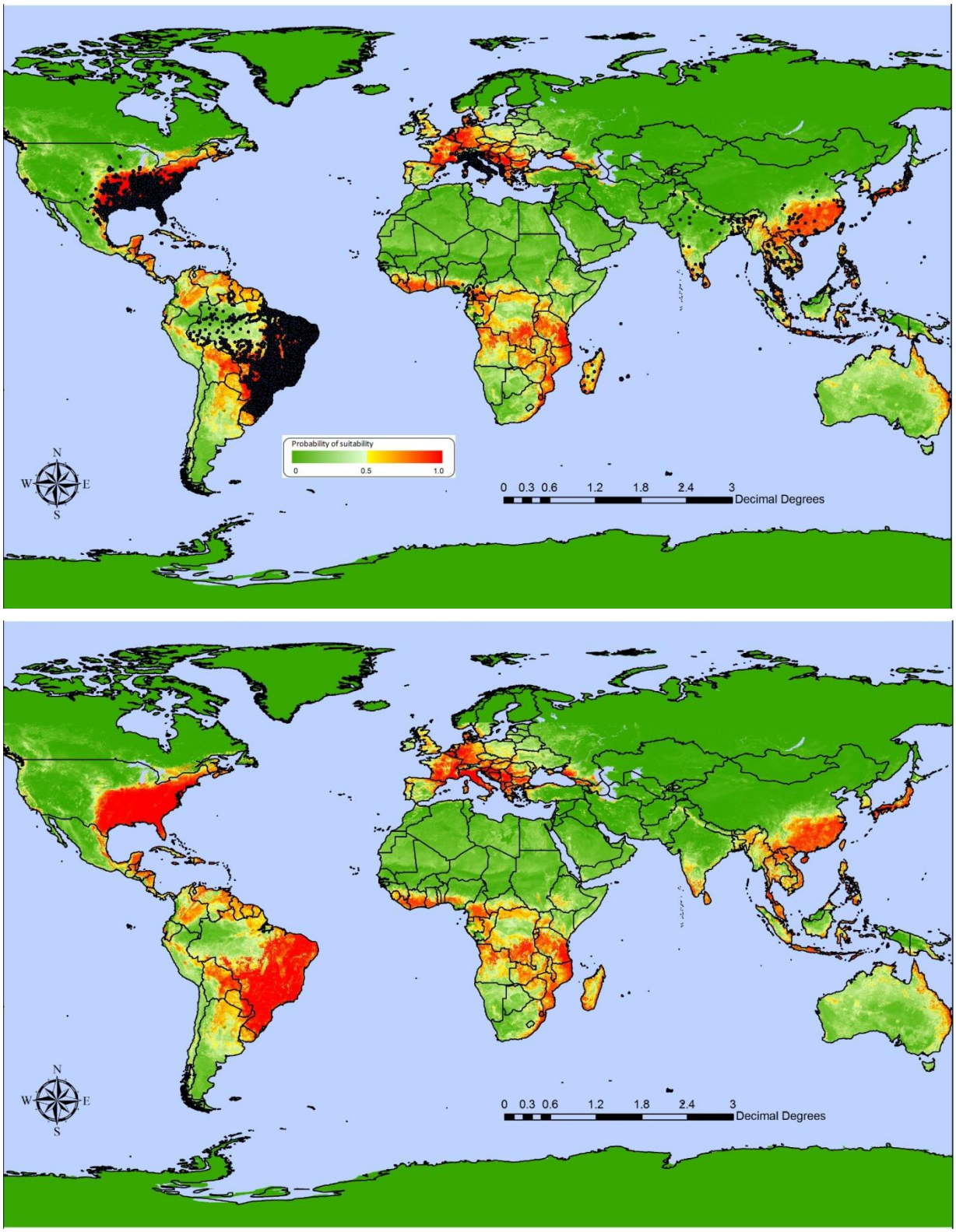
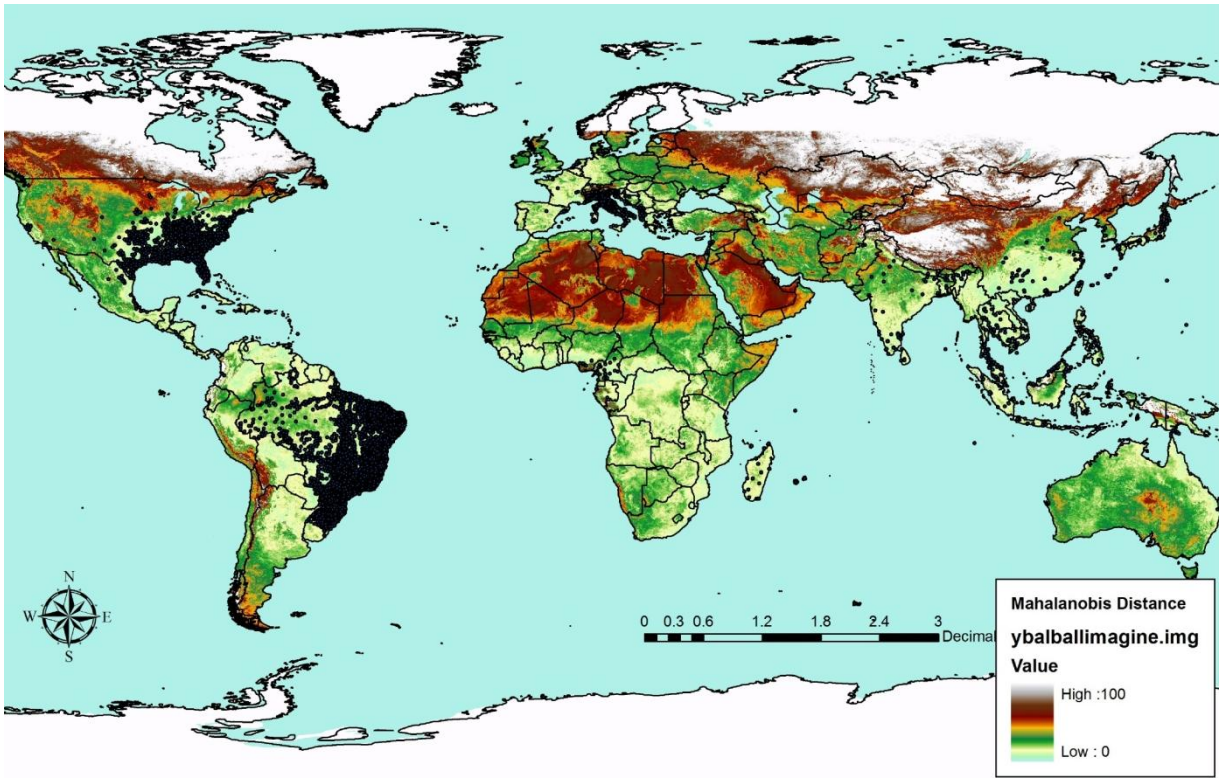


Figure 5. Mahalanobis distance of global model for *Aedes albopictus*

These maps show the Mahalanobis distance image for the combined global maps of Figure 4. Upper with all the presence points used in all five models, lower without the presence points.



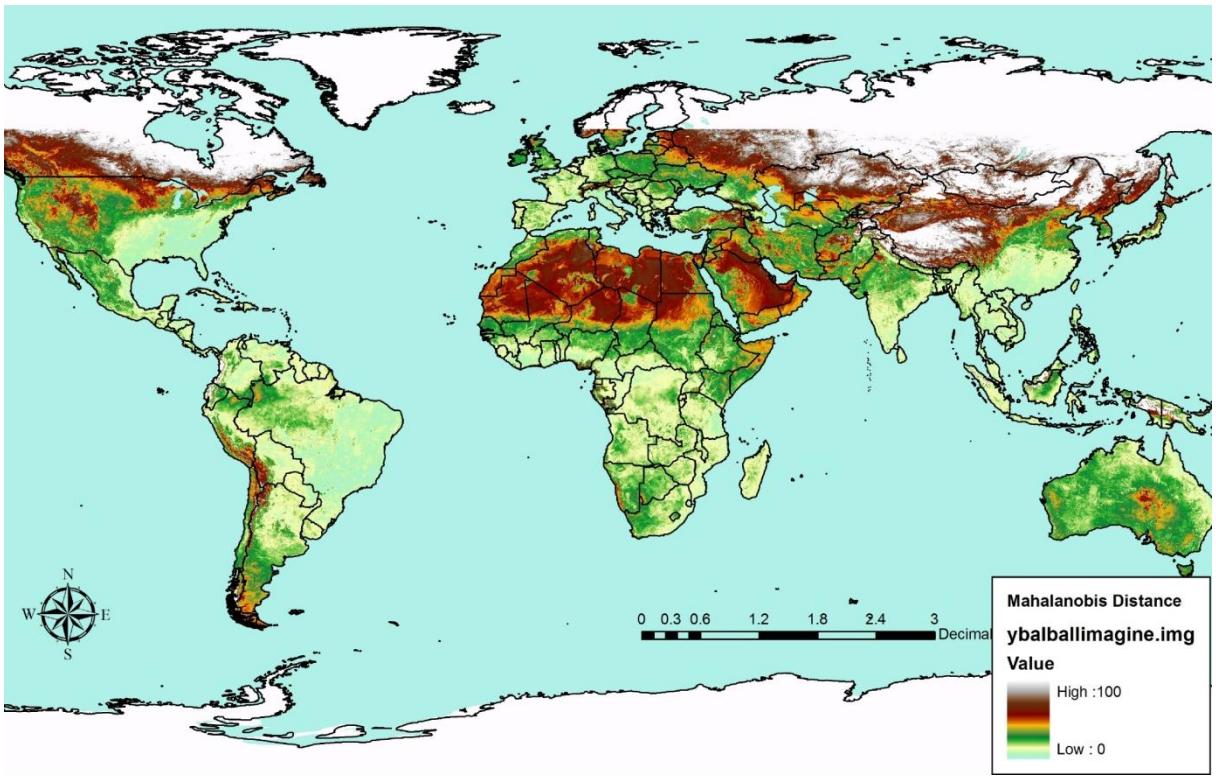


Figure 6. Details of global model for *Aedes albopictus*

The upper figure shows a detail from the risk map in Figure 4. The individual clusters of presence for the European (TigerMaps) dataset are identified by the three colours shown in the legend (red, green, blue); the pseudo-absence points are indicated by dots. The lower figure shows the same points on the Mahalanobis distance image from the global model, produced by selecting the lowest MD from the five MD images of the five different models for this species.

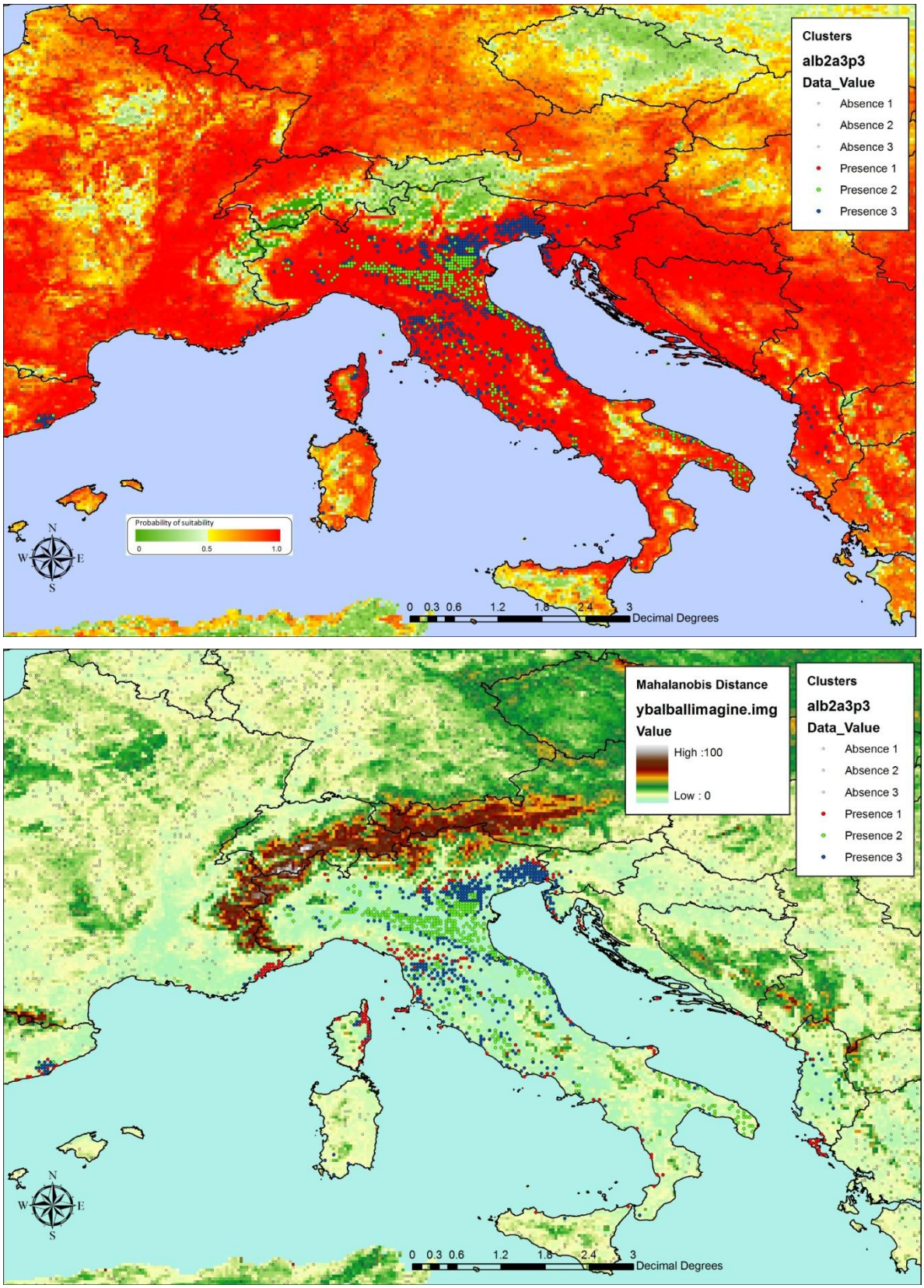
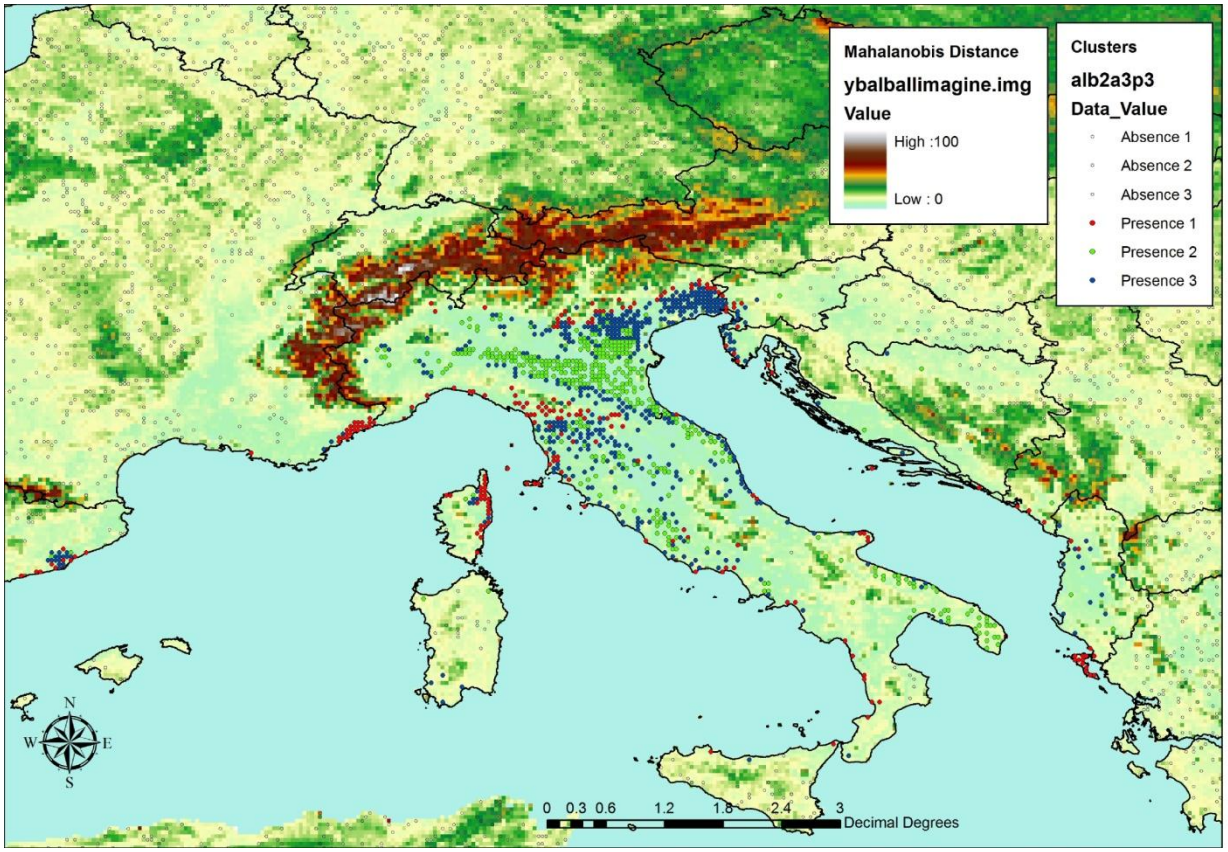
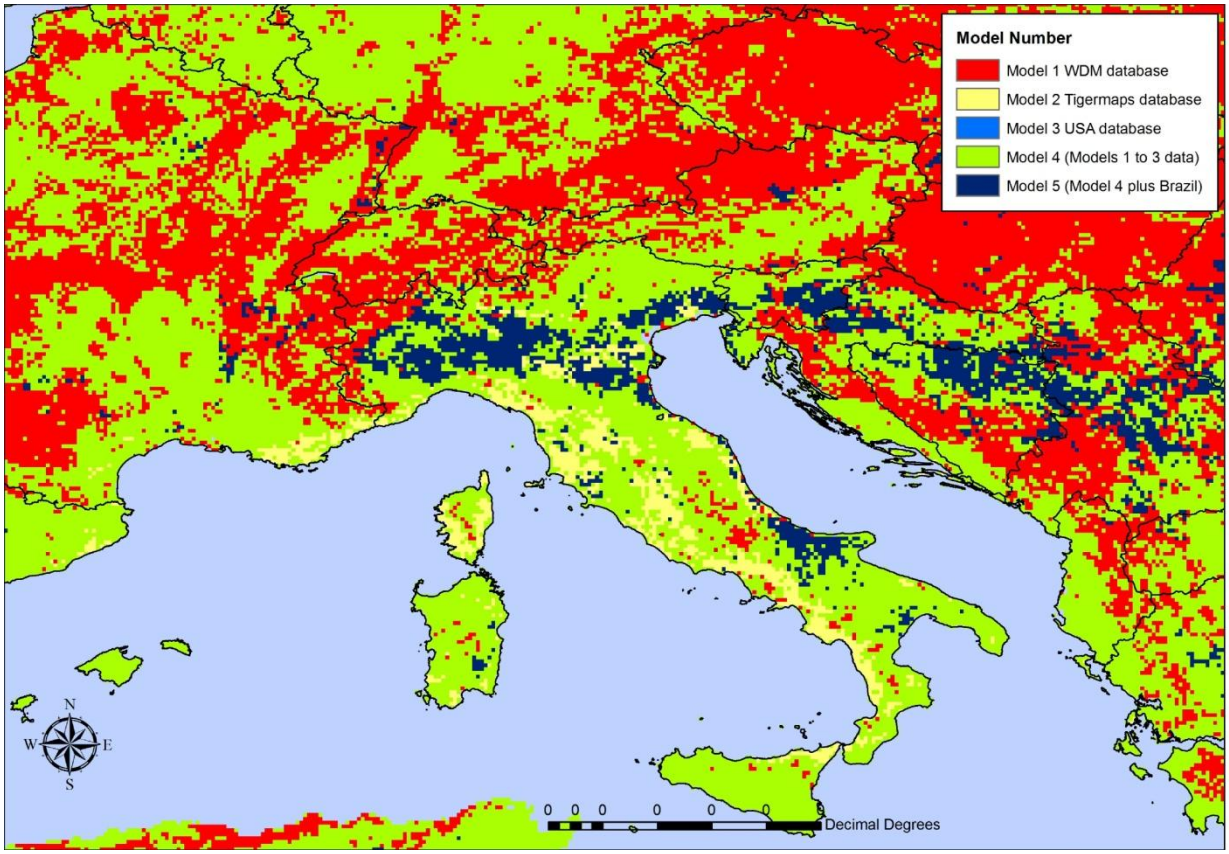


Figure 7. Details of Mahalanobis distance of global model for *Aedes albopictus*

The upper figure shows the Mahalanobis distance image from the global model, produced by selecting the lowest MD from the five MD images of the five different models for this species. The lower figure indicates which of the regional or global models (Models 1 to 5) contributed the MD values in the upper image (original).





Aedes albopictus in Europe: predicting future spread

We can predict the likely future direction of spread of *Ae. albopictus* in Europe by refining our view of the MD image shown in Figures 6 and 7. This is shown in Figures 8 and 9 for different regions of Europe. The same MD image has been used as before, but in this case the range of MD values from 0 to 75 is shown rather than the range 0 to 100. This has the effect of keeping in pale colours sites that are more similar to presently occupied sites. It is along these routes that we might expect *Ae. albopictus* to spread first. The very strong message from Figures 8 and 9 is that these spread routes for *Ae. albopictus* are along many parts of the northern Mediterranean coast and, inland, along extensive river systems or along road networks.

Aedes albopictus in Europe: a monitoring programme for future spread?

These maps suggest that a monitoring and surveillance system should concentrate on the areas within Europe highlighted as likely spread routes in Figures 8 and 9.

Many variables must be considered when deciding whether to implement a mosquito surveillance system and how to customise surveillance methods to the various local situations. ECDC has produced guidelines to deal with these questions[11].

Ovitraps are probably the simplest and most effective way of monitoring the presence of *Ae. albopictus* in any area, and the maps presented here could be used as a guide to where to place such traps. It is strongly recommended that such a sampling system not only monitors the spread of this mosquito, but also the accuracy of this sort of approach to vector and disease monitoring in general. This is best done by having a series of transects along the major predicted spread routes combined with transects at right angles to them that sample progressively less suitable areas. This would possibly be most easily done along the Rhone and Rhine rivers in France and Germany, with the cross transects heading up towards the Alps, where conditions appear to become rapidly less suitable for *Ae. albopictus.* However, large rivers and mountains can also be barriers for mosquito dispersion where they are less travelled by humans.

Examples of this suggested approach are shown in Figure 10, where parts of the imagery have been selected with rapid changes in MD with distance from water courses. What is particularly surprising about these images is that they were effectively derived from satellite data that does not 'see' water directly, but only in terms of the climatic and vegetation conditions associated with water. Even the rainfall layers (major contributors to the risk maps globally) would not on their own have produced these patterns, since rainfall does not fall only along water courses. Nevertheless the MD imagery shows low values along water courses. These areas are climatically more similar to sites inhabited by *Aedes albopictus* globally than any other areas in Europe.

In setting up the sampling programme suggested above it is recommended that the sort of imagery displayed here is combined with higher resolution Landsat, SPOT or similar imagery which shows more details of infrastructure (roads, fields, etc.) which will affect the placement of traps. For such activities, clearly, the higher resolution the better the results, as the placement of a few meters could result in big differences as concerns the collection of mosquitos.

While the MD images exploit information about habitat seasonality captured by temporal Fourier analysis (seasonality that is usually a vital determining factor for the presence or absence of species), higher resolution imagery contains more information on the fine-scale structure of habitats. It is a matter of debate whether such higher resolution imagery should be classified into land-cover types (fields, forests, farms, etc.) using supervised or unsupervised methods before they are used, or whether the raw data should be used directly. Geographers tend to favour classification; but any classification into land-cover types involves dividing continuous reflectance values into discontinuous categories. There are bound to be errors in this process and these are likely to be concentrated at ecotones – boundaries between land cover types – that are often important for invertebrates, both as areas in which to live and as corridors along which to move. Meanwhile, some advances have been made in classification techniques that take into account biological 'threshold' levels, which are important for models that are more biologically rather than statistically driven.

Unrealised habitat space for *Aedes albopictus* in Europe.

Confirmation of unutilised realised habitat space within Europe comes from Figure 11, which shows a series of frequency histograms of Mahalanobis distances (MDs) from the global *Ae. albopictus* map. Each frequency distribution is the distribution of MDs from the centroid defining *Aedes albopictus'* known presence somewhere on the globe. Figure 11 is best interpreted by remembering that low MDs indicate conditions similar to those in which *Aedes albopictus* has been reported, while high MDs indicate conditions quite different from known *Aedes albopictus* areas. The red curve shows the frequency distribution for the global dataset of presence points excluding the European database. In other words, it is the frequency distribution of MDs of individual points in the global dataset from their collective centroid. The blue curve shows the frequency histogram for the European

database. The mode of the blue curve is displaced to the left compared with the mode of the red curve. This is because MDs are on average less for the European dataset; if mosquitoes inhabit a smaller range of more similar habitats then their collective frequency distribution will contain more, lower values of MDs from their collective centroid. A comparison of the red and blue curves shows that globally *Ae. albopictus* inhabits a wider range of habitats (larger MDs) than it does just within Europe.

Therefore, if there are conditions in Europe corresponding to the MD values associated with the bulk of the red curve, they represent areas within which *Ae. albopictus* can survive in Europe even though the species has not yet reached them. That this is the case is shown by the green curve which is the frequency distribution of MD values taken at regular intervals (Figure 11). This is the frequency distribution of the selected European points from the centroid of *Aedes albopictus'* global distribution. While much of northern Europe is clearly unsuitable for the species (i.e. many parts of Europe have high MD values) there are sufficient areas elsewhere that appear to be suitable. In fact, there are proportionately fewer areas on the green curve corresponding to the mode of the blue curve (current European *Aedes albopictus'* distribution) than there are areas with slightly higher MD values coinciding with the mode of the red curve (*Aedes albopictus'* global distribution). The conclusion from these three histograms therefore is that there are proportionately more areas in Europe that appear to be suitable for *Ae. albopictus* than have currently been invaded by it. We should expect this species to spread very much farther in Europe than it has done so already.

Figure 8. Future spread of *Aedes albopictus* in south-western Europe

Predicted future routes of spread of Ae. albopictus *in France (upper) and Spain (lower). The MD images are stretched over a range of values to reveal areas of greatest similarity to presently occupied areas. Clearly, river systems are the likely major spread routes in Europe (original).*

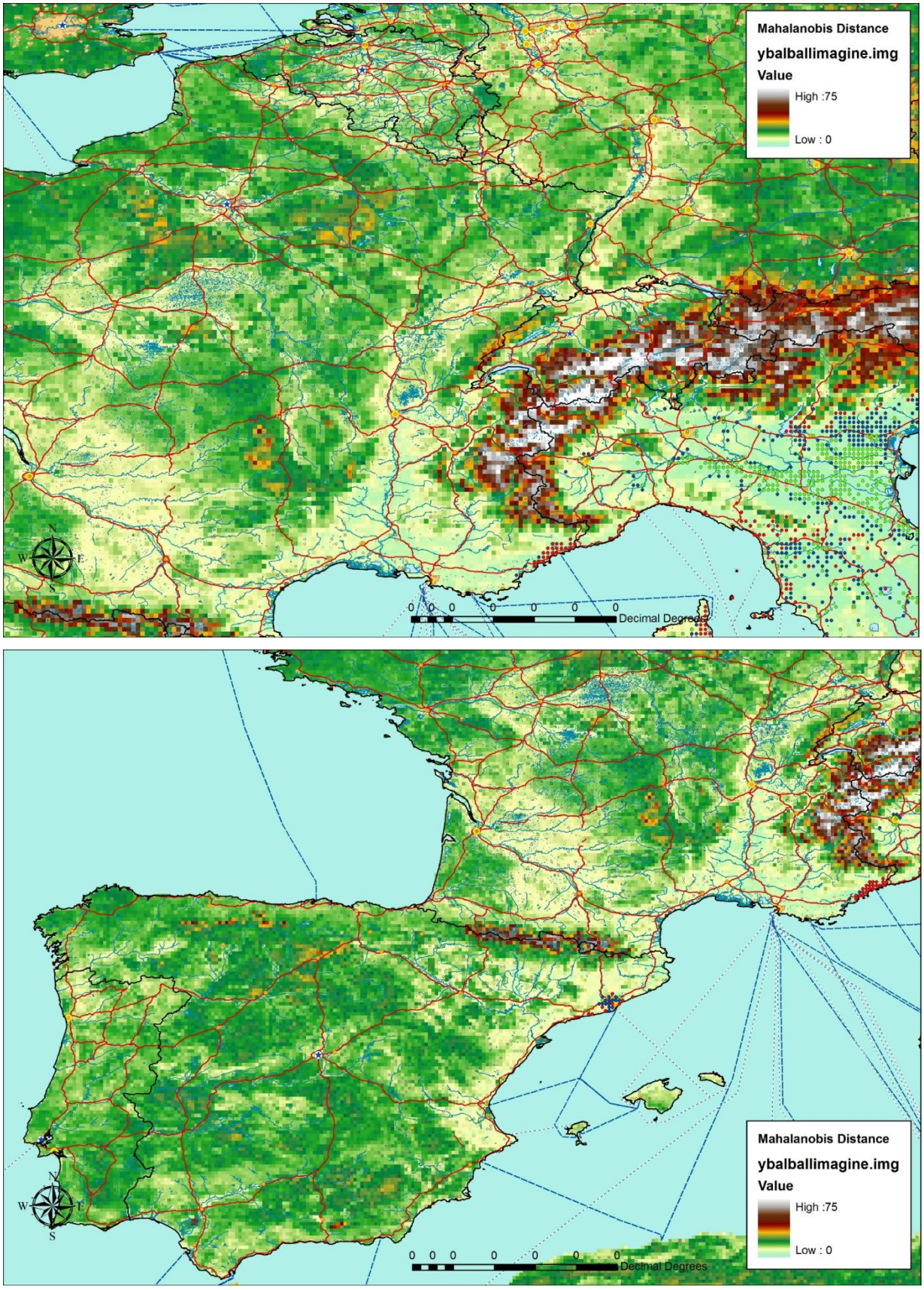


Figure 9. Future spread of *Aedes albopictus* in south-eastern Europe

Predicted future routes of spread of Ae. albopictus *in the northern Adriatic region (upper) and in Greece (lower). The MD images are stretched over a range of values to reveal areas of greatest similarity to presently occupied areas. Once again, river systems are the likely major spread routes (original).*

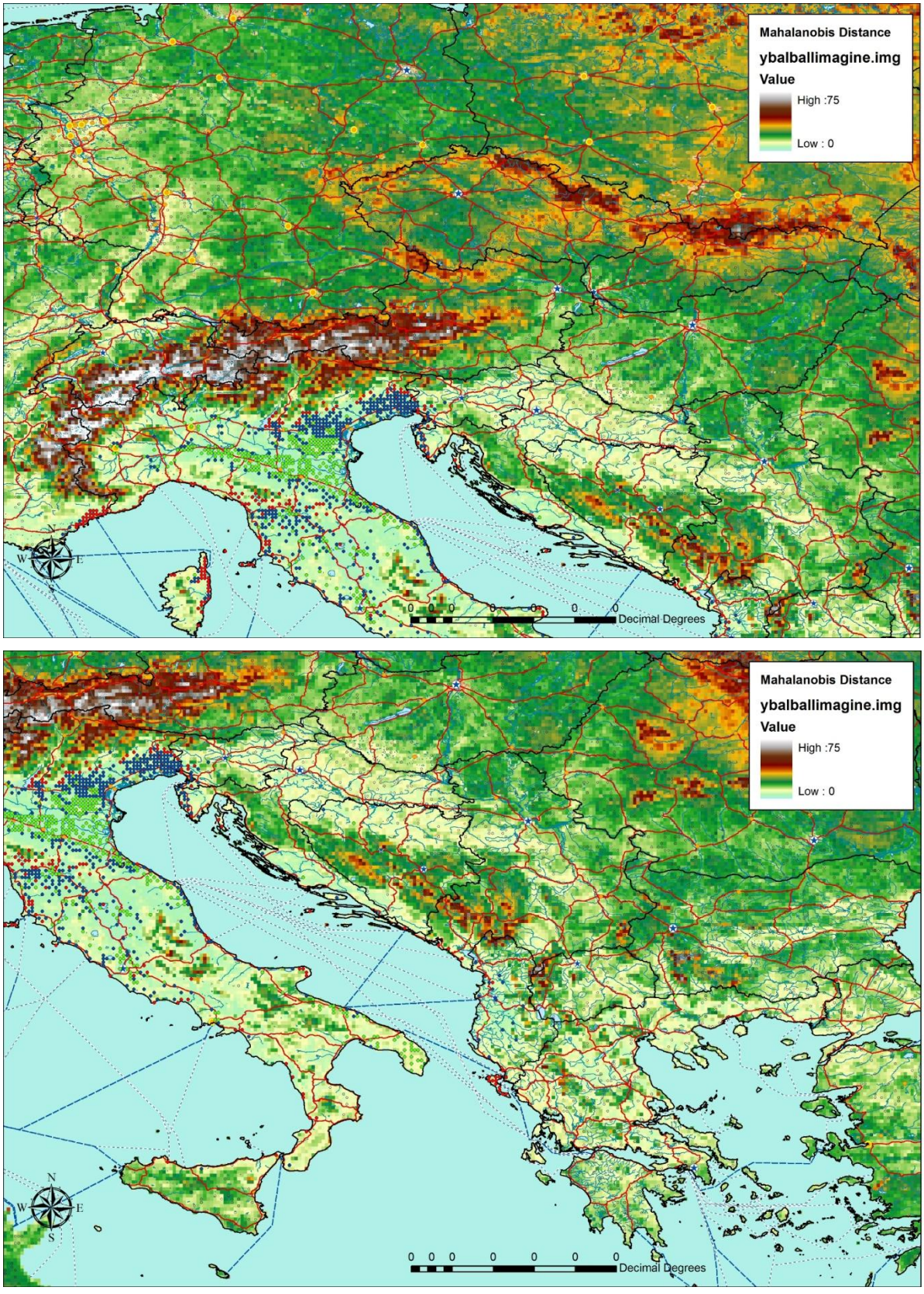


Figure 10. Future spread of *Aedes albopictus* in central Europe

Details from Figure 7 for two regions showing the rapid increase of MDs with distance away from water courses. Upper figure, region south-west of Paris; lower figure, border area of France/Switzerland/Germany (original).

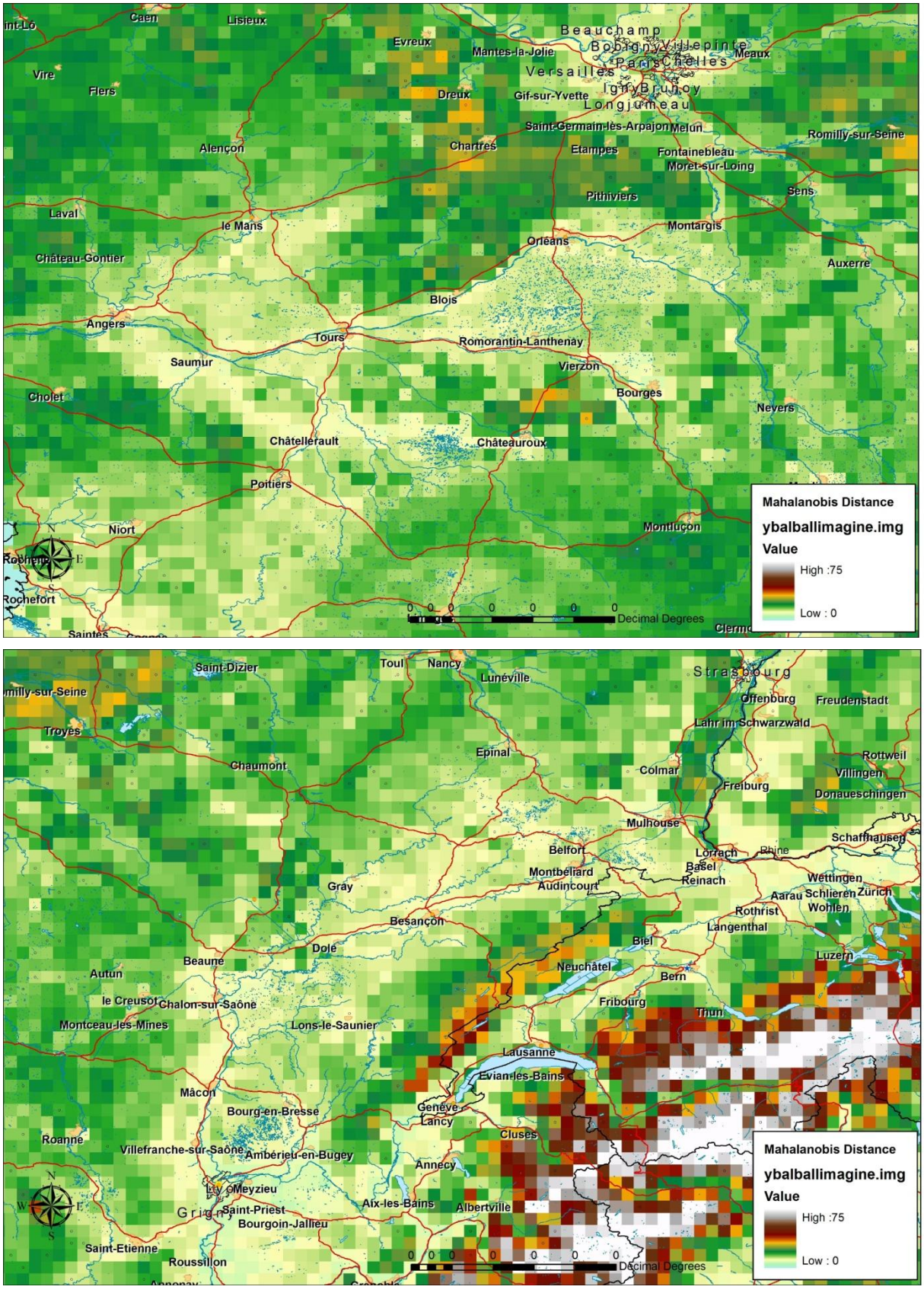
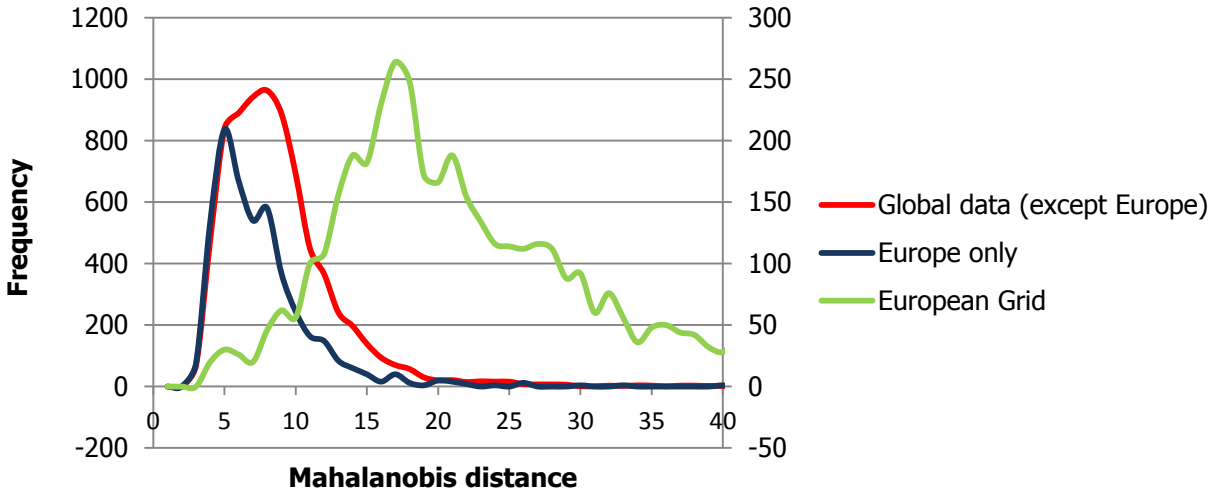


Figure 11. Frequency distributions of Mahalanobis distances

The red curve is the frequency distribution of MDs from the global model but excluding the data for Europe. The blue curve is the frequency distribution for just the European presence data (from Tigermaps[4]). Notice that the mode of the global curve is shifted to the right compared with the blue curve, indicating that Ae. albopictus *inhabits globally a wider range of conditions than it does in Europe. The green curve is the frequency distribution of values extracted from a regular grid placed over Europe and sampling at regular intervals within a window running from 35N to 60N and from 11W to 26E. The mode of this curve is shifted to the right compared with the other two curves because many areas of northern Europe, sampled by the grid, are unsuitable for this species. However, there are, proportionately speaking, more areas of Europe not yet occupied by* Ae. albopictus *that show conditions similar to those inhabited by it globally than there are areas already occupied by this species in Europe (because this part of the green curve is gradually increasing towards the modal value). The relative heights of the different histograms are functions of sampling intensity and are hence not a good guide to the total area that is suitable for this species in Europe. Imagine the blue curve reduced to fit inside the green curve, since it represents a subset of European environmental space, sampled more fully by the green curve (original).*



3.4 Understanding model limitations

The outcome of any risk map is dependent on the type of modelling approach used, the input variables, the time-frames of the variables used, and many other factors. It is essential that public health practitioners understand that models are nearer to representing 'best estimates' rather than the 'truth'. One limitation to the NLDA approach described here is that 'external' validation of the models is very complicated. Methodologically, the bootstrapping approach helps to ensure the 'internal' validity of the model, but it is not possible to assess whether the variables modelled are the best ones – this often comes down to the expert judgement of the modeller and their collaborators. Similarly, there are no independent datasets that could prove or disprove the results of this model, given that we are modelling places likely to be suitable habitats for *Aedes albopictus* – it is only if or when *Aedes albopictus* comes into these habitats that the question of the suitability will be answered.

Conclusion

In this handbook we have taken a risk map produced via the non-linear discriminate analysis methodology and interpreted it in several ways to show the potential for *Ae. albopictus* to spread further in Europe. These interpretations show the places that are likely to be suitable and the potential spread routes that the vector might use to reach such places. In turn these provide the potential for targeting monitoring resources as efficiently as possible, to check on the spread of this species in the near future.

It should be emphasised that all of these conclusions are based on models which may be wrong, and that there are multiple different possible approaches to risk mapping that have not been discussed here. Therefore a decent monitoring scheme should not only be positioned according to current predictions and be able to capture spread if the model is correct, but also be set up to test whether the model is correct in the first place. It can do so by sampling not only areas of predicted greatest suitability, but also areas of predicted greatest change in suitability. Thus model validation and application occur simultaneously. New records feed back into the modelling process progressively to improve the models locally.

Risk maps can indicate which sites are more likely for local transmission to occur. They are an increasingly important tool in public health, helping to address questions about the possible impact of climatic and environmental change on the potential distribution of vector-borne diseases, and on where vector- or disease-surveillance initiatives might be most wisely implemented. ECDC intends to facilitate such activities through initiatives such as VBORNET and the E3 network.

Annex. Tutorial exercise for NLDA modelling

This tutorial serves as an introduction to the eRiskMapper software for creating disease risk maps. Operation of the software is described using a set of remotely sensed (MODIS 1km) climatic variables as environmental data and some vector presence and absence datasets (mosquito distributions in south-eastern Europe). The first exercise runs through the creation of a non-linear discriminant analysis model to produce a risk map from some known disease presence points. Following this, the processing of presence data known only to administration district level is dealt with. The tutorial concludes with a series of exercises to attempt once you have mastered the basic functions of the software.

A 1.1 Overview

eRiskMapper is a user-friendly application for creating disease risk maps. A risk map shows the modelled distribution of a particular disease or vector over a region and is created by applying an algorithm to a set of training data (of known disease presence or abundance and absence) associated with a set of predictor variables often derived from climatological databases or earth-orbiting satellites. Risk maps identify areas of possible disease risk across the entire area covered by the training set and often regions well beyond the training set's geographic limits. eRiskMapper is introduced here by means of a series of worked examples. This software is currently still a beta release; please send any bug reports, requests or comments to: david.morley@zoo.ox.ac.uk.

A 1.2 Getting started

Ensure that you have downloaded and installed the latest version of eRiskMapper from the FTP site: ftp://tala-ftp.zoo.ox.ac.uk/EDEN/EDENLDA/ (username: EDENLDA, password: Ox4ordUn1), or you have copied the data across from an Oxford University hard drive.

For this exercise we will be using MODIS images covering Europe. These data are temporal Fourier transformed (TFA) and are at a spatial resolution of 1km. Ensure that you have downloaded/copied the folders 'MODIS Imagery' and 'ESA Test data' to a local drive on your computer.

The disease data we are using is this demonstration are the distribution of a species of mosquito (*Anopheles labranchiae*), as recorded by Kuhn et al[12]. These data have been extracted from literature published between 1927 and 2002. The dataset covers the whole of Europe but to make this tutorial run faster, we are going to focus on a subset of these data and investigate a region covering south-eastern Europe.

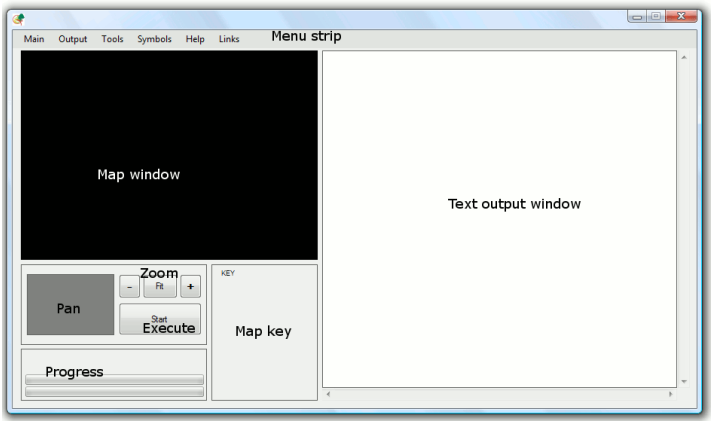
A list of the files needed to complete this tutorial is given below:

Files needed	Description
EG******.rst (and .rdc)	MODIS imagery: ~90 European images, IDRISI raster format (~180 files)
A_Lab_PA.csv	Text file of *A. labranchiae* point presence data
A_Lab_PREV.csv	Text file of *A. labranchiae* point prevalence and absence data
ItalyLVL2.rst (and .rdc)	IDRISI raster image of Italian level 2 administrative district boundaries (two files)
A_Lab_ADMIN.csv	Text file of *A. labranchiae* administrative level presence and absence data

A 1.3 The basics

eRiskMapper is an entirely GUI-based application. Upon loading, the main display window is presented (Figure A1). Model results will be plotted in the upper-left panel, while a log of the model making process will be kept in the right-hand panel. The menu strip contains various tools to convert between raster data types, to display data and view model results.

Figure A1. The main eRiskMapper interface



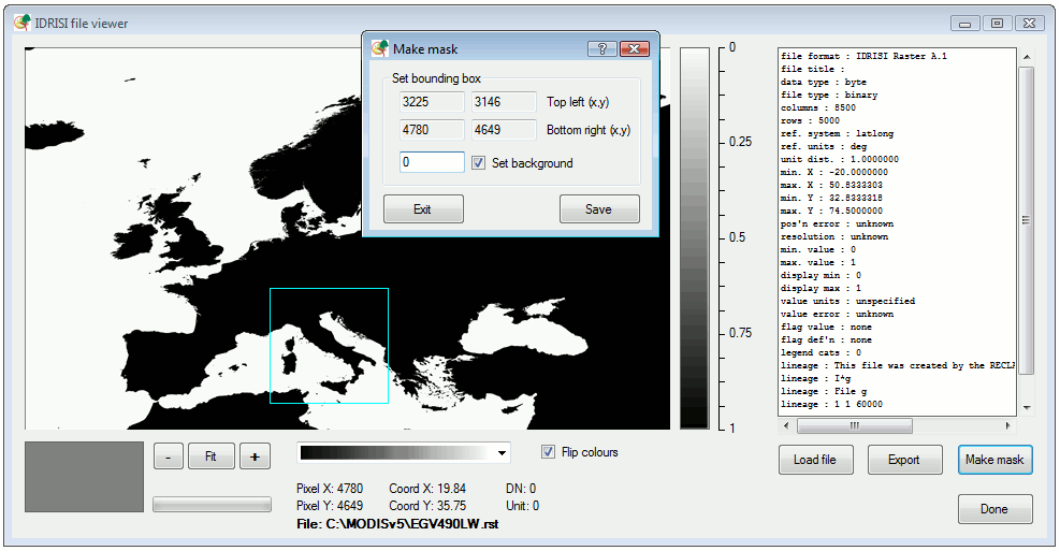
Making a mask file

As our environmental data cover the whole of Europe, we must create a mask file to restrict the analysis to our area of interest over south-eastern Europe.

To create a mask, use the IDRISI file viewer:

- Start eRiskMapper.
- Go to, Tools → IDRISI File Viewer, in the main menu strip
- The file viewer window opens. Click, 'Load file'.
- Select, 'EGV490LW.rst' from your MODIS imagery folder. This will display the European-wide land/water mask. Our new mask will be a subset of this image.
- Click on 'Make mask'; a small window containing bounding box co-ordinates will appear.
- Ensure that 'Set background' is checked with value 0.
- On the map, click and hold the mouse to draw a box around a region containing (approximately) only Italy and the Balkan countries (see Figure A2 as an example).
- Ensure that the top-left and bottom-right x,y co-ordinates are similar to those shown in the 'Make mask' box in Figure A2.
- If you are not happy with your box, click on the image and start again.
- Once you have a suitable box, click 'Save' and call the file 'ItalyMSK.rst' (filename must include 'MSK' or 'MASK' to be recognised as such) and save it in the same folder as the MODIS imagery (important!).
- The mask will now save; do not do anything until you see a message box telling you that the mask file has been saved.
- View the new mask by clicking 'Load file' and selecting 'ItalyMSK.rst'

Figure A2. Creating a mask covering Italy and surrounding areas



A 1.4 Setting up an NLDA model

For the first exercise we will create a model using non-linear discriminant analysis (NLDA) with bootstrapping using the *A. labranchiae* data, and focused on south-eastern Europe.

We are now ready to start running the model:

- Click, 'Start' on the main window. This will launch the 'Model details' dialogue.
- For this exercise, we would like a non-linear discriminant analysis model. Ensure that this option is selected.
- The data type is point/pixel level, as we have precise x,y coordinates of our disease presence locations. The second section of this tutorial will explain how to deal with data recorded only to administrative district or polygon level.
- The data scaling should be set as 'Probability scale'. Here, we only know that the disease is either present or absent; there is no information about prevalence or abundance.
- In directories, click 'Add' and navigate to the directory containing the MODIS imagery. Also set a directory in which to store all the model output to be created. It is a good idea to create a new directory for each new model run. Click 'Done' when you are finished.

Next, we need to specify the imagery:

- All available imagery sets in the folder(s) selected above will be displayed here. For this exercise, there should be just one set – the MODIS imagery with dimensions 8500 x 5000 pixels and ~115 images in total. Ensure that this set is selected using the checkbox.
- Ensure that the box 'Replace drop-out values' remains unchecked.
- Use the drop-down list to select the mask we have just created, 'ItalyMSK.rst', as the primary mask. Leave the secondary mask box as 'None'.
- Click on the 'Model options' tab at the top of the dialogue box. Accept all the default values, but give the model a new name if you like. This name will prefix all output files created.
- Click, 'Done' to continue.

The 'Point presence/absence box' will pop-up. Here, we must specify our response data.

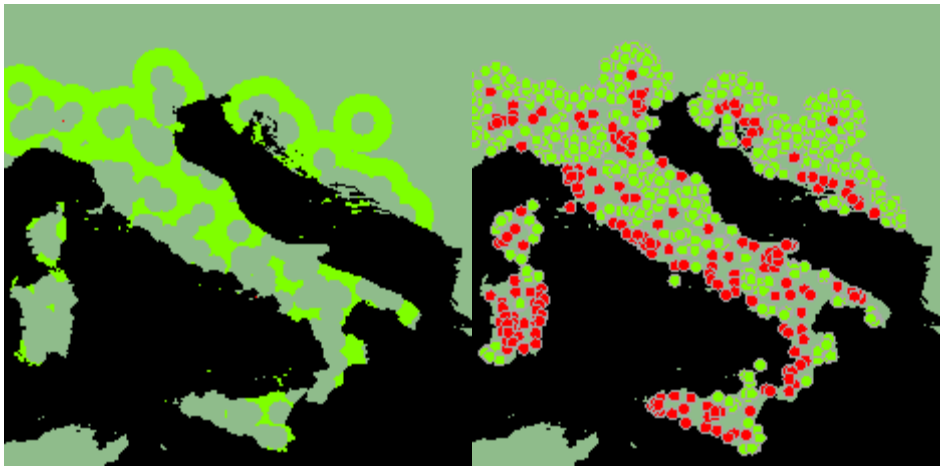
- We are creating a new data table, rather than loading an existing one. Ensure that, 'Create new datatable' is checked.
- The presence data for this example are in the form of a text file giving a list of the precise latitude/longitude locations of known presence points. Select 'Presence data from coordinate list' and navigate to the file 'A_Lab_PA.csv'.
- Although recorded, this dataset contains too few absence points over our masked region; we must instead generate pseudo-absence locations. Select 'Generate pseudo-absences' and click on the 'Options' tab at the top of the dialogue.
- Specify the minimum and maximum distance that an absence point must be from any recorded presence location as 0.3 and 0.75 respectively (these are distances in decimal degrees). There are also options to confine absences further, using an environmental distance, and to screen for input errors. We will not be using these functions in this example.
- If you have time later, experiment changing the minimum and maximum buffer distances for pseudo-absence locations. Notice the effect that this has on the final model. As a result, it is important to consider the choice of these distances carefully when creating your own models.
- Click, 'Done' to proceed.
- Note that the text window on the main form is keeping track of all options selected during the modelling process. On completion, this log is saved as a record of the details of the model created.

Select number of points

eRiskMapper has now plotted the presence locations and all possible pseudo-absence locations on screen (Figure A3). The presence/absence counts box shows that we have 206 presence and 257 409 possible absence locations.

- This is obviously far too many absence locations, so we are going to pick a random sub-sample from these. In the sample size box, enter 206 for presence (that is, we keep all presence locations), but enter 600 for absence (to take a random 600 from the total of 257409).
- Click, 'Done' to continue. The sampled points will now be displayed on screen.
- Use the + and – buttons, the scroll bars and scroll box in the main window to zoom and pan the map.
- If you want to, go to: symbols → change symbols, to change the style and size of the map symbols.

Figure A3. All possible presence (red) and pseudo-absence (light green) locations, before (left) and after (right) sub-sampling



Extract environmental data

- Click, 'Import RS data' on the main window to continue. This displays a list of all the environmental variables available. Each one corresponds to an individual raster layer.
- Variables in the left panel will be considered as candidate explanatory variables during modelling. Variables in the right panel will be removed from all further analysis. To move variables, highlight them with the mouse then use the '>>>' and '<<<' buttons to move them between panels.
- For this example, remove all variables ascertaining to the percentage variability in the annual, bi-annual, tri-annual and annual to tri-annual cycles. You should have removed a total of 24 variables.
- Click, 'Done' to continue.
- The next window will show DN conversion factors and drop-out value flags. This is used to convert values stored as integers in raster layers back into their real values and also to detect any missing data. Conversion factors are automatically read from a raster layer's documentation file, you do not need to alter these values, so just click 'Done' to proceed.
- Environmental values for each presence and absence location will now be extracted from each raster layer. This may take a couple of minutes.

Data clustering

The data must now be split into a number of multivariate normal clusters to ensure greater accuracy of the discriminant analysis routine. Discriminant analysis requires input distributions to be multivariate normal. As a disease may be found under a variety of environmental conditions, the single distribution summarising these conditions is unlikely to be multivariate normal. Splitting the complete presence (or absence) distribution into a number of smaller clusters has the effect of creating a new set of distributions each covering a shorter environmental gradient. As a result they are more likely to be multivariate normal.

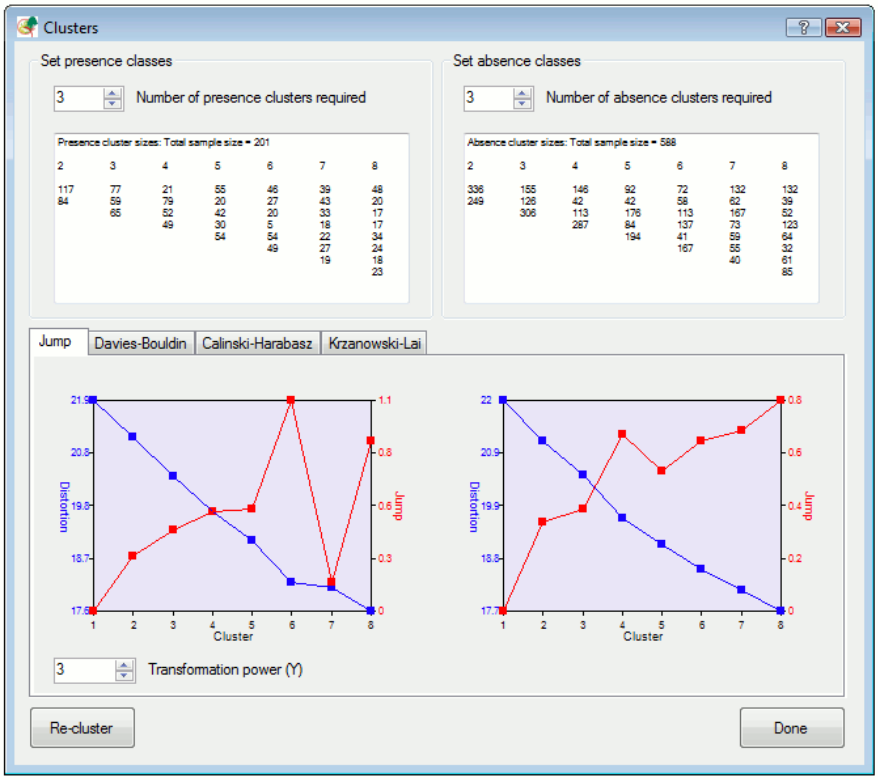
Clustering is achieved using the k-means algorithm:

- Click, 'k-means cluster' in the main window.
- Set the parameters for the k-means algorithm. Ensure that you use 'Smart Centroids (k-means++)', this will make an optimal clustering more likely. Accept the default values for all other options.
- On clicking 'Done', you will be asked to select the variables used to cluster the data (this will be a subset of all the variables available). This dialogue box functions in the same way as the one used in Section A 1.4 (Extract environmental data). To cluster the data, we will use the DEM and all components (except the three phases) of daytime LST, NDVI and middle infrared. You should now have 22 variables selected.

The k-means algorithm will now run and cluster both the presence and absence data into two to eight clusters inclusively. You must now decide how many clusters are in the dataset. This is a difficult choice, but several indices are calculated to aid this decision. See the help files for more details.

- Results for the k-means clustering are now shown (Figure A4: Results of k-means clustering.). The upper panels show the size of potential clusters if the data are partitioned into two to eight clusters. The lower panel shows the results of the optimal clustering indices.
- Look for the greatest CHANGE in the jump method and put the MAXIMUM in all other indices to indicate the most likely number of clusters. Results will vary as we are using a random subset of datapoints. You will rarely need more than four or five clusters for either absence or presence. For this example, use one presence and two absence clusters.

Figure A4. Results of k-means clustering



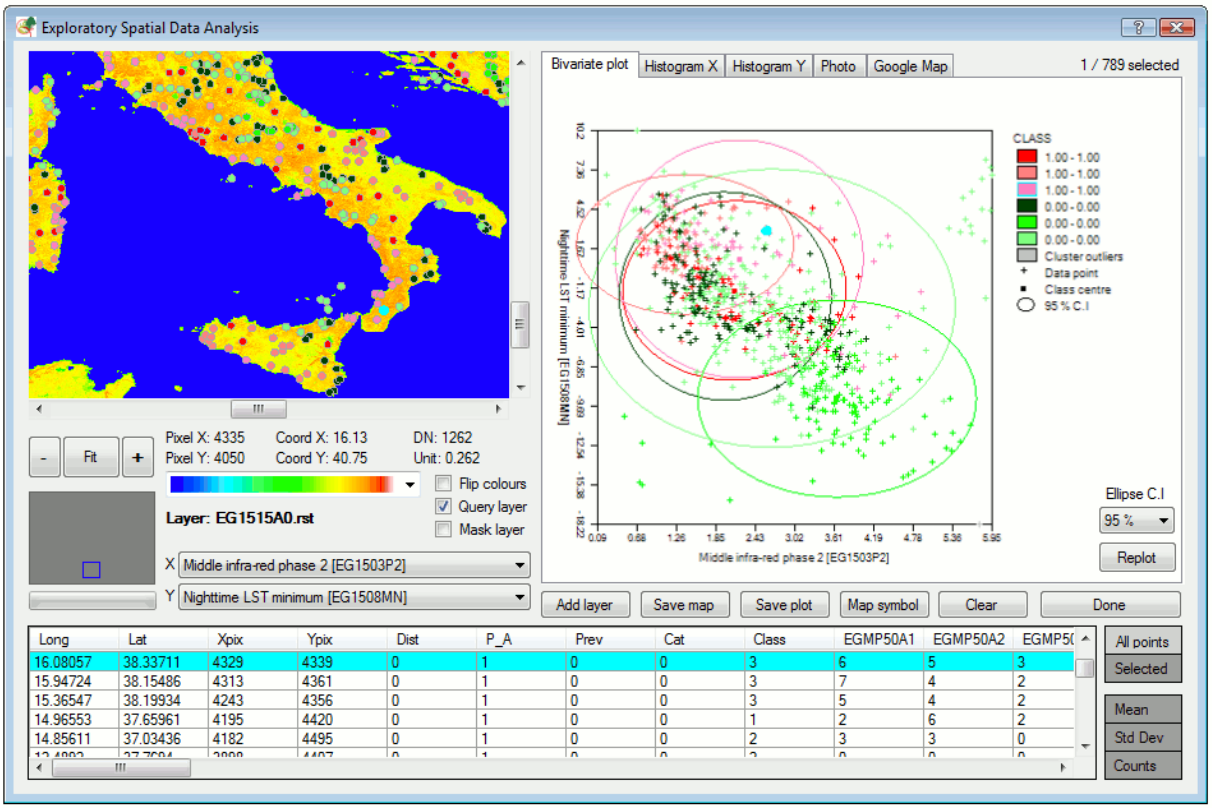
Presence is on the left, absence on the right. The upper panels show the resultant class sizes for possible clustering of two to eight classes. The lower panel shows graphs of indices attempting to define the optimal clustering. Use the spinners at the top of the form to select the number of presence and absence clusters needed.

Exploratory spatial data analysis (ESDA)

eRiskMapper provides a tool to carry out ESDA. This enables users to fully understand their dataset and view it in the context of environmental data, identify outliers or trends, and highlight areas for future data collection.

- In the main menu, select Tools → ESDA. Click, 'Yes' when asked if you want to calculate error ellipses. The ESDA window will launch (Figure A5).
- Use the X and Y drop-down boxes to select two variables to plot against each other. Note that each cluster defined in Section A 1.4 is given a different colour. Click on the legend in the bivariate plot to change these colours. Summary statistics for each cluster are also given in the datatable at the bottom of the window.
- Click on a point in the graph. You will notice that it is now highlighted. The location of this point is now shown on the map, as well as its details in the datatable. Clicking on a point in the map or table will also highlight the same point in the graph or table too.
- To select multiple points, click and hold the mouse button to drag a box around a cluster of points. Alternatively, hold the 'Ctrl' key and click on multiple points.
- The tabs above the bivariate plot enable the graph to be viewed as histograms.
- Select a single point on the map or the bivariate plot, then click on the 'Google Map' tab. The location will automatically be displayed using Google Maps.
- Click on 'Add layer' and select an image. This will place your selected environmental raster layer as the base map for the training set locations. Checking 'Query layer' will allow you to click on the map to obtain the co-ordinates and variable value for the click location.
- When you are finished, click 'Done' to close this window.

Figure A5. The ESDA viewer



Non-linear discriminant analysis

We are now ready to run the model and produce the risk map. First, the options for variable selection need to be selected.

- Click, 'NLDA' on the main window to continue.
- In the variable selection tab, keep all the default settings. We are creating a model via NLDA, using standardised data, separate covariance matrix, equal prior probabilities, and with variables chosen by stepwise selection based on minimising the AICc (threshold value 5).
- Also check the box to create Mahalanobis distance images and select 'to nearest presence class'. This will create images showing the similarity of each pixel to the conditions described by presence locations in the training set.
- In the bootstrapping tab, ensure that the box to carry out bootstrapping is checked and that we want to bootstrap both presence and absence data.
- Usually we use 100 bootstraps, but this can take many hours to process. For the sake of this tutorial reduce this to two bootstraps.
- Uncheck, 'Save each bootstrap raster'. Only use this option if you want to keep a raster image of each intermediate bootstrap model. However, doing this can quickly use a lot of hard disk space if running many bootstraps. The memory usage panel will warn you if this may be a problem.
- Select, 'Bootstrap both presence and absence' and choose to bootstrap with 75 presence and 75 absence points and ensure that the checkbox 'Carry out validation' is checked with a percentage of 50%. This will use half the points not selected for a particular bootstrap as a validation dataset.
- We do not want to force any variables into the model, so there is no need to go to the 'Force variables' tab.
- Click 'Done' to start running the model. Note that this may take several minutes.

Model results

During model running

- Results for each bootstrap will be displayed as they are created. This consists of a map and a table of accuracy statistics and details of the variables chosen by the model (Figure A6).
- Look at the model details displayed on screen: You will see which 10 variables were included in the model and their rank of importance. The correlation matrix shows the relationship between these 10 variables. The classification matrix shows how well the model classified the observed data. For a perfect model, all values

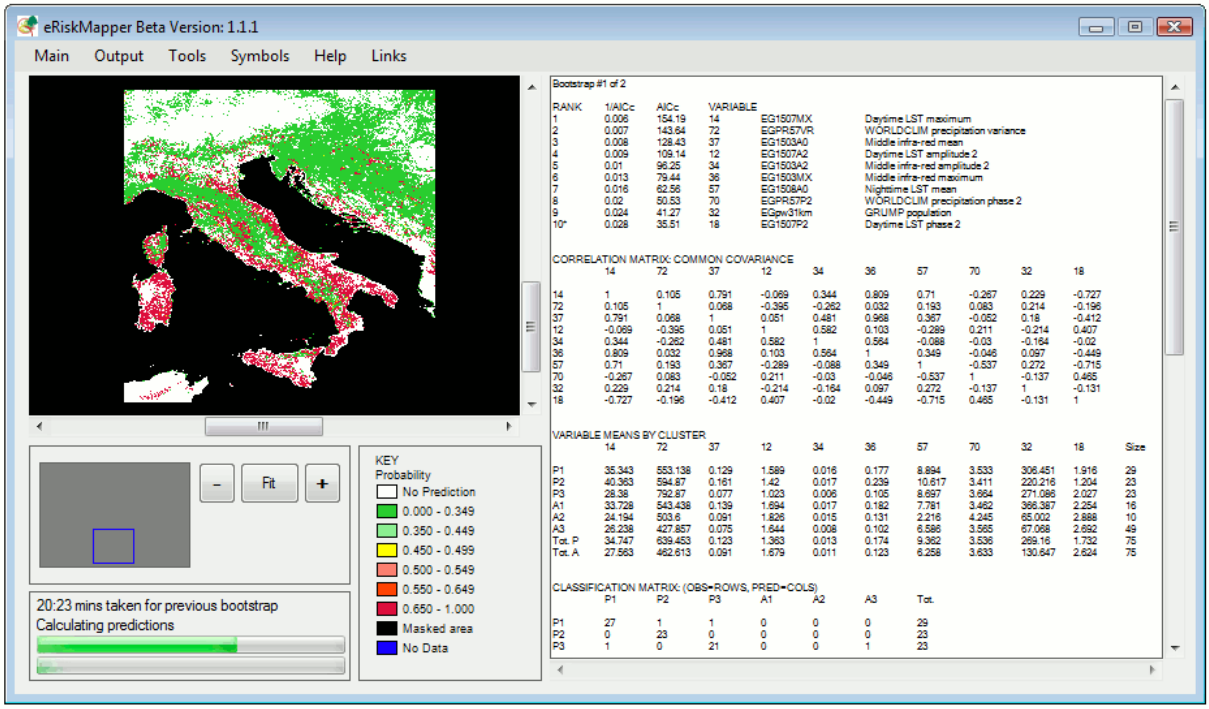
fall on the main diagonal of the matrix. Also shown are the probability that a point from a certain class will be assigned to each of the others and a table of producer and consumer errors.

- Note also the sensitivity and specificity. A value of 1 for sensitivity means all actual presences have been recognised (0 would indicate none have been recognised). A value of 1 for specificity means all actual absences have been recognised.
- Values of Cohen's kappa and Klecka's tau also range from 0 to 1. Models with a perfect fit will have a value of 1 for each of these.

At the end of the model run, after all bootstraps have been carried out:

- A rainbow plot is created showing the variables selected over all bootstraps. This is useful to identify the most common variables. (It is not so useful here as we only have two or three models to compare). Click on the help button for this window to see a full rainbow plot.
- The final risk map (average of all the bootstraps) is shown along with the model log.
- To view individual model results, go to: Output → Model output in the main menu and enter the number of the model you wish to see results for.
- The model log contains details of the average rank of each variable over all models, the number of times each variable was included and an overall assessment of the model performance.

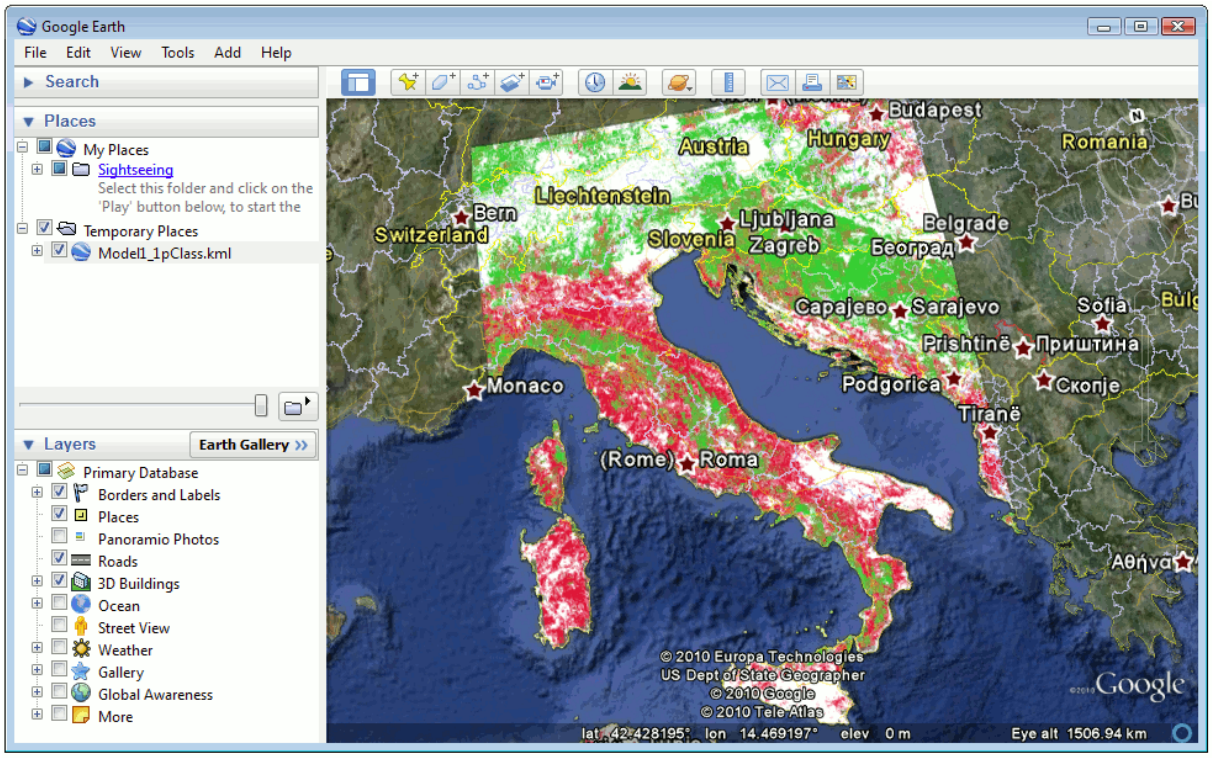
Figure A6. The results of a single-bootstrap model displayed on screen



Saved results:

- All results are automatically saved to the directory specified in step 0.
- All raster files are compatible with IDRISI, ArcGIS and other GIS software. The final risk map is called 'Model1_AVPROBCLASS.rst' (classified into six probability bins) and 'Model1_AVPROB.rst' (the raw probabilities ranging from 0 to 1).
- Various other files are saved: the model log, accuracy statistics for each bootstrap, jpegs for each bootstrap and the final map, the training set as a datatable, a table of the model classification errors, raster layers of Mahalanobis distances, cluster memberships and KML files readable by Google Earth.
- If you have Google Earth installed, double-click on a KML file to open it. The risk map will automatically be draped over the Google Earth globe (Figure A7).

Figure A7. Risk map overlay in Google Earth using KML



The first part of this tutorial has shown you how to create a risk map using NLDA and presence data reported to specific latitude/longitude point locations and view the results. The next part demonstrates how to deal with data recorded only at administrative district level. This is problematic because, unlike point reported data, a presence location cannot be related to a single pixel in our environmental data. To demonstrate this, we will start by creating a new model.

A 1.5 Processing presence data at administrative level

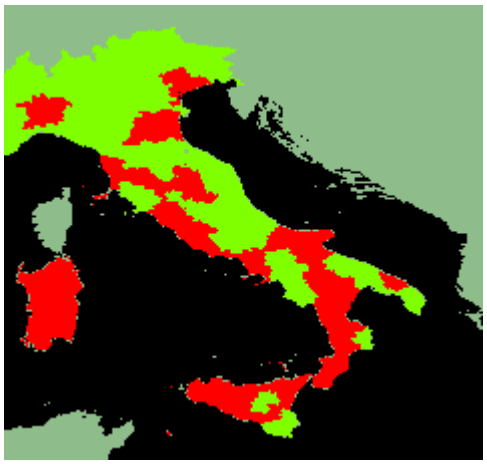
Restart eRiskMapper.

- Click on the 'Start' button on the main interface.
- Enter the model details as before but be sure to pick 'District (polygon)' as the data type and create a new folder for the model outputs. Click, 'Done' to continue.
- Ensure that the 8500 x 5000 MODIS imagery is selected. Enter the mask you created in step 0 as the primary mask file. Leave the secondary mask option as 'None' as before. Click 'Done'.

Set the data sources

- Enter the raster layer defining administrative districts as 'ItalyLVL2.rst'. This is an image of the level 2 administration boundaries in Italy.
- Enter the district code and data list as 'A_Lab_ADMIN.csv'. This specifies the number of presence cases in each district. For the purposes of this exercise, these data have been derived by degrading the real data recorded at point level over Italy by Kuhn et al.[12] to polygon level.
- Enter the presence threshold as 1. This means that we will treat districts with low presence (one case or less) as absence districts.
- Ensure the method selected is 'Clustering'. Click 'Done' to continue.
- You will be presented with a list of all available districts listed by their codes (this list should start at 915). Click 'Done' as we do not want to exclude any districts. This function would be useful if we wanted to remove any districts that correspond to urban areas, for example.
- The location of all presence and absence districts will be plotted on screen (Figure A8).

Figure A8. *A. labranchiae* presence and absence



Presence (red) and absence (light green), level 2 administration districts in Italy

Set the clustering options:

- In 'Clustering strategy', check, 'Cluster a sample' and enter the sample value as 10%. As clustering is a calculation-intensive process, this option will take a random sample of just 10% of the entire dataset and cluster it. The remaining 90% of the dataset will then be classed by assignment to their nearest clusters as defined by the initial 20% sample.
- Accept all other defaults. Click 'Done' to proceed.
- Click 'Select k-means variables' on the main window. You will now be asked to choose with which variables you want to cluster the data.
- Select the variables to cluster with as the DEM and the means and variances of daytime LST, NDVI and middle infrared.
- Click 'Done' to accept the default conversion factors. Then click, 'k-means cluster' on the main window to begin the clustering. When prompted, select the number of clusters in which to cluster the data to be 100.
- The clustering will now run. This may take several minutes to complete.

Identifying absence pixels

All presence pixels have now been clustered into one of 100 possible clusters. The rationale here is that clusters that are found in a large proportion of the presence districts are indicating common environmental conditions and therefore probable favourable conditions for disease presence. Clusters indicating environmental conditions found in only a few of the presence districts indicate less common habitats and thereby cannot be associated with disease presence in the study region. As a result, pixels in these less widespread clusters can be used to define disease absence.

- On completion, a summary of the 100 clusters is presented. They are ranked according to the percentage of the original presence districts in which pixels belonging to this cluster are found. A cluster with a value of 100% would be represented in all presence districts. The number of pixels in each cluster is also given.

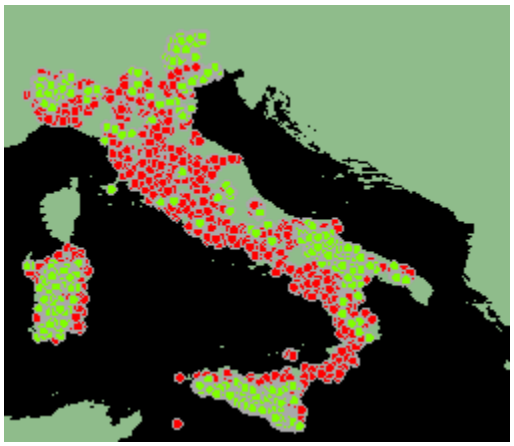
Cluster	%	Pixels
1	13.115	395
2	14.754	16
3	26.23	2072
4	26.23	1383
5	27.869	2681
6	29.508	987
7	31.148	721
8	31.148	2234

  For example, using the above table, a threshold of 15% would remove clusters 1 and 2 thereby reclassing 411 (395 + 16) pixels to absence.

- A dialogue box will ask for an elimination threshold. All pixels belonging to clusters with a percentage below this threshold will be reclassified as absence points. For this exercise, enter a threshold of 35% and click 'Done' to continue. Usually, you would use a lower threshold (5 to 15%) and repeat the clustering process to gradually remove pixels. As this takes time, for this exercise it is more convenient (but less accurate) to remove a large number of pixels at once.

- You will be asked if you want to repeat this process to remove more pixels. In this case we do not, so click 'No' to continue.
- You will now be shown the 'Presence and absence counts' dialogue box. Here the total number of presence pixels and absences are shown (note that the two values sum to the initial total of presence pixels, 11163). Select 1000 presence and 2000 absence. eRiskMapper will select these at random from the total available and plot them on screen (Figure A9).
- The remainder of the modelling process proceeds as from section A 1.4 above ('Extract environmental data'). You can either practise and carry on to obtain a risk map for these data, or restart eRiskMapper and move to the next exercise. If you choose to complete this model, try experimenting and change the number of clusters, variable selection method and number of samples taken at each bootstrap.

Figure A9. Distribution of sub-sampled presence and absence points after filtering by clustering



Note that due to the stochastic nature of this method, your results may be slightly different.

Additional exercises

If you have time, attempt one or all of the exercises below.

Exercise 1: Point sampling of administrative level data

Using the MODIS imagery and the administrative districts defined in the 'ItalyLVL2.rst' layer and the text file 'A_Lab_ADMIN.csv', create an NLDA (or random forest) using the point sampling approach rather than the clustering method described in section A 1.4.

Hints

- Enter all data and select the options as you did for the first part of section A 1.6. However, be sure to select 'Point sampling' as the method in the 'District sampling options' dialogue.
- This time try removing a possible outlier polygon. District number 1604 is small and corresponds to the urban area of Napoli. If you want to remove this district, ensure you move polygon 1604 to the 'Districts excluded' box when prompted.
- Select to sample the same number of points from within each district. Set this figure to be 15. The sampled points are then plotted on screen.
- After random points have been selected, the modelling approach will follow that of the standard NLDA approach with point presence data as described above.

Exercise 2: Using abundance data

Using the MODIS imagery and the file 'A_Lab_PREV.csv', create a risk map based on disease abundance/prevalence. Using abundance scaling, the final risk map is not an image of the probability of disease presence, but rather an image showing the most likely of a set of classified zones. These zones correspond to a discrete class of abundance (e.g. high abundance, 10 to 20 cases, or low abundance 1 to 10 cases).

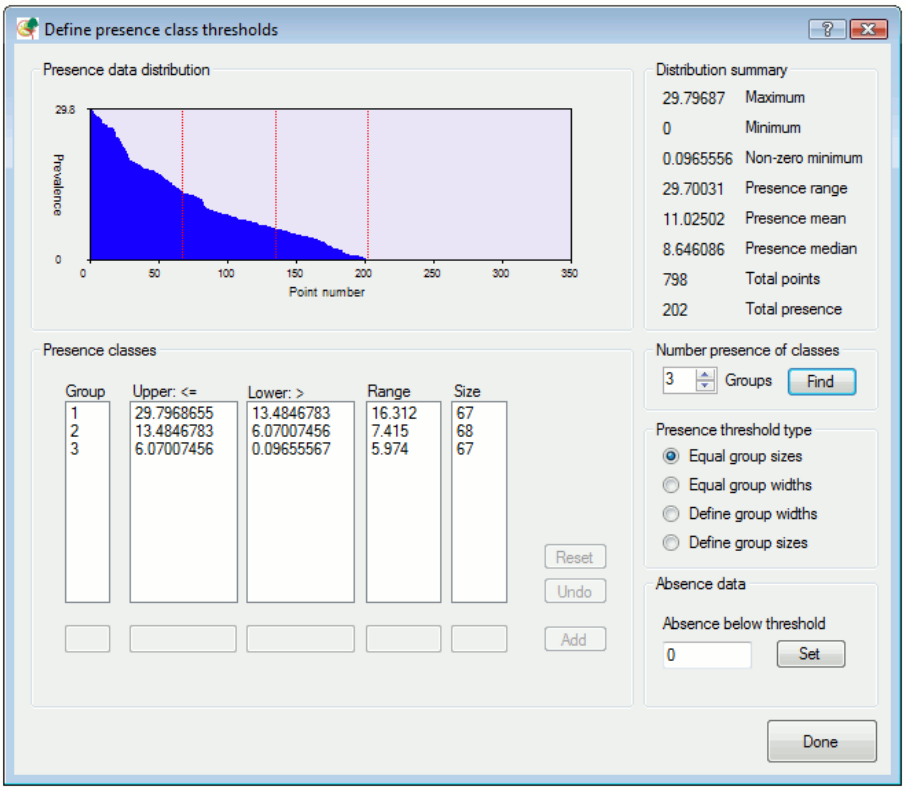
Hints

- Be sure to select, 'Abundance scale' in the model details dialogue, otherwise the data will only be treated as presence/absence (probability scale) as before.
- This data set contains known absences (n = 596), so there is no need to generate pseudo-absence points. (Select, 'Absences specified with presence data list' when entering your data files.). The file 'A_Lab_PREV.csv' also contains 202 presence points along with a value for their disease prevalence.

- You will be asked to define the prevalence thresholds (Figure A10). This basically splits your presence data into categories of abundance (e.g. high, medium and low = three classes); the final risk map is essentially a map of these categories. Use the spinner (on the centre-right of the dialogue box) to select the number of classes you want and click 'Find'. The boundaries will be displayed on the distribution histogram and as a list.

Figure A10. Defining the class thresholds for prevalence/abundance data

The distribution of the data is plotted as a histogram (ordered from high to low abundance). The selected number of presence classes is represented as a series of red lines. The corresponding boundaries and resultant class sizes are shown in the table below the graph.



Hints

- For this exercise, use 'Equal group sizes' and pick three classes. Make sure you click the 'Find' button to accept this change.
- Absence data cannot be thresholded in this way (since all values are 0); clustering here is done by k-means as in part A 1.4. Pick two absence classes.
- The final risk map is plotted according to the greatest likelihood of a pixel belonging to one these presence or absence categories. See the model log for the key to these categories.

Exercise 3: Combining the administrative level data methods

Use results from parts A 1.5 and A 1.7.2 to combine presence points from presence districts, but randomly sampled locations from known absence administrative districts. To do this you will need to edit the saved datatables (in the output directory you specified at the start of each model).

Hints

- Using Excel or similar, open up the datatable (Model1_DATATABLE.csv) for the model created in part A 1.5. Cut the columns for Longitude, Latitude and P_A and paste into a new document. Make sure you only copy across the presence points (coded as 1 in the P_A column)
- Open the datatable from the model created in part A 1.7. This time, cut the columns for absence (coded as 0) and paste into the new document directly underneath the presence data. Your new sheet should have three columns and contain both presence and absence data like this:

Long	Lat	P_A
32.5	55.4	1
34.6	55.7	1
36.3	54.6	1

38.6	53.1	0
32.4	54.1	0
34.4	54.8	0

- Save this new sheet as a CSV file.
- Run eRiskMapper and create a model using this new CSV table, with the data type as 'point (pixel)' and specify that absences are specified with presences (do not generate pseudo-absences).

---

[1] Dalla Pozza GL, Romi R, Severini C. Source and spread of *Aedes albopictus* in the Veneto region of Italy. J Am Mosq Control Assoc. 1994 Dec;10(4):589-92.

[2] Medlock JM, Hansford KM, Schaffner F, Versteirt V, Hendrickx G, Zeller H, Van Bortel W. A review of the invasive mosquitoes in Europe: ecology, public health risks and control options. Vector Borne Zoonotic Dis. 2012 Jun;12(6):435-47. doi: 10.1089/vbz.2011.0814. Epub 2012 Apr 20.

[3] Tatem AJ, Hay SI, Rogers DJ. Global traffic and disease vector dispersal. Proc Natl Acad Sci U S A. 2006 Apr 18;103(16):6242-7. Epub 2006 Apr 10.

[4] European Centre for Disease Prevention and Control. Development of *Aedes albopictus* risk maps. Stockholm: ECDC; 2009. Available from: http://www.ecdc.europa.eu/en/publications/Publications/0905_TER_Development_of_Aedes_Albopictus_Risk_Maps.pdf

[5] European Centre for Disease Prevention and Control. The climatic suitability for dengue transmission in continental Europe. Stockholm: ECDC; 2012.

[6] Lambrechts L, Scott TW, Gubler DJ. Consequences of the expanding global distribution of Aedes albopictus for dengue virus transmission. PLoS Negl Trop Dis. 2010 May 25;4(5):e646. doi: 10.1371/journal.pntd.0000646.

[7] Rezza G, Nicoletti L, Angelini R, Romi R, Finarelli AC, Panning M, et al. Infection with chikungunya virus in Italy: an outbreak in a temperate region. Lancet. 2007 Dec 1;370(9602):1840-6.

[8] European Centre for Disease Prevention and Control. Meeting Report. ECDC expert consultation on mosquito surveillance guidelines. Stockholm: ECDC; 2011.

[9] According to the TF harmonics decomposition, the third and the forth digits correspond to the temporal coverage of the time series where harmonics where extracted from (in this case is 2000–2008).

[10] Time series 2001–2005

[11] European Centre for Disease Prevention and Control. Guidelines for the surveillance of invasive mosquitoes in Europe. Stockholm: ECDC; 2012.

[12] Kuhn KG, Campbell-Lendrum DH, Davies CR. A continental risk map for malaria mosquito (Diptera: Culicidae) vectors in Europe. J Med Entomol. 2002 Jul;39(4):621-30.