ecdc
EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

TECHNICAL REPORT

A spatial modelling method for vector surveillance

www.ecdc.europa.eu

ECDC TECHNICAL REPORT

A spatial modelling method for vector surveillance

This report was commissioned by the European Centre for Disease Prevention and Control (ECDC) under Specific Contract No. 5 – ECD.3962 implementing Framework Service Contract ECDC/2009/018, coordinated by Wim Van Bortel and Olivier Briët, and produced by Neil Alexander and William Wint, Environmental Research Group Oxford; Guy Hendrickx, Veerle Versteirt and Els Ducheyne, Avia GIS. Hector Honrubia built the web-map application.

Authors

Neil Alexander, Wim van Bortel, Guy Hendrickx, Veerle Versteirt, Els Ducheyne and William Wint

Contributors

Albieri, Alessandro	Centro Agricoltura Ambiente 'Giorgio Nicoli', Bologna, Italy
Alten, Bulent	Hacettepe University, Ankara, Turkey
Alves, Maria Joao	Minesterio da Saude, Lisbon, Portugal
Antunes, Ana	Faculdade de Medicina Veterinária - Universidade de Lisboa , Lisbon, Portugal
Aranda, Carles	Consell Comarcal del Baix Llobregat, Servei de Control de Mosquits, Barcelona, Spain
Beeuwkes, Jacob	Laboratory of Entomology, Wageningen, The Netherlands
Bødker, Rene	National Veterinary Institute (DTU), Frediksberg, Denmark
Bucher, Edith	Biological Laboratory, Laives, Italy
Bueno Mari, Ruben	Laboratorios Lokímica, Valencia, Spain
Collantes, Francisco	Universidad de Murcia, Murcia, Spain
Dikolli, Enkelejda	Institute of Public Health, Tirana, Albania
Eritja, Roger	Consell Comarcal del Baix Llobregat - Servei de Control de Mosquits, Barcelona, Spain
Estrada-Pena, Augustin	University of Zaragoza, Zaragoza, Spain
Falcuta, Elena	Cantacuzino Institute, Bucharest, Romania
Fontenille, Didier	IRD / Directeur de l'Institut Pasteur du Cambodge, Cambodia
Gewehr, Sandra	Ecodevelopment, Thessaloniki, Greece
Gunay, Filiz	Hacettepe University, Ankara, Turkey
Hansford, Kayleigh	Public Health England, Porton Down, UK
Hristovski, Slavco	Faculty of Natural Sciences and Mathematics, Skopje, North Macedonia
Hubalek, Zdenek	Institute of Vertebrate Biology, Academy of Sciences, Brno, Czech Republic
Hufnagl, Peter	Austrian Agency for Health and Food Safety (AGES), Vienna, Austria
Ibañez-Justicia, Adolfo	Centre for Monitoring of Vectors, Wageningen, the Netherlands
Ivovic, Vladimir	University of Primorska, Koper, Slovenia
Jaenson, Thomas	Uppsala University, Uppsala, Sweden
Kalan, Katja	University of Primorska, Koper, Slovenia
Kampen, Helge	Friedrich-Loeffler-Institut, Greifswald - Insel Riems, Germany
Karakus, Mehmet	Hacettepe University, Ankara, Turkey
Kasap, Ozge Erisoz	Hacettepe University, Ankara, Turkey
Kavur, Hakan	Cukurova University, Dept of Medical Parasitology, Adana, Turkey
Klobucar, Ana	Institute of public health 'Dr. Andrija Stampar', Zagreb, Croatia
Krüger, Andreas	Berhard Nocht Institut für Tropenmedizin, Hamburg, Germany
Medlock, Jolyon	Public Health England, Porton Down, UK
Miranda Chueca, Miguel Angel	University of the Balearic Islands, Department of Biology, Palma de Mallorca, Spain
Montalvo, Tomas	Agència de Salut Pública de Barcelona, Barcelona, Spain
Mosca, Andrea	IPLA, Turin Area, Italy
Ognyan, Mikov	National Centre of Infectious and Parasitic Diseases, Parasitology and Tropical Medicine, Sofia, Bulgaria
Oguz, Gizem	Hacettepe University, Ankara, Turkey
Ozbel, Yusuf	Ege University Faculty of Medicine Department of Parasitology, Izmir, Turkey
Pajovic, Igor	University of Montenegro, Biotechnical Faculty, Montenegro
Perrin, Yvon	Centre National d'Expertise sur les Vecteurs, Montpellier, France
Petric, Dusan	Faculty of Agriculture, University of Novi Sad, Serbia
Piazzi, Mauro	IPLA, Turin Area, Italy
Plenge-Bönig, Anita	Div. Hygiene and Infectious Diseases, Institute for Hygiene and Environment of the City of Hamburg, Hamburg, Germany
Prioteasa, Liviu	Cantacuzino Institute, Bucharest, Romania
Regan, Eugenie	National Biodiversity Data Centre, Ireland
Saska, Aleksandra	Science and Research Center, Koper, Slovenia,
Schaffner, Francis	Francis Schaffner Consultancy, Riehen, Switzerland; & University of Zurich, Zurich, Switzerland
Sousa, Carla A.	Instituto de Higiene e Medicina Tropical, Lisbon, Portugal
Sulesco, Tatiana	Academy of Sciences of Moldova, Chisinau, Moldova
Vatansever, Zati	Kafkas University, Kars, Turkey
Vaux, Alexander	Public Health England, Porton Down, UK
Walder, Gernot	Medizinische Universität Innsbruck, Division of Hygiene and Medical Microbiology, Innsbruck, Austria
Zamburlini, Renato	University of Udine, Dept. of Agricultural and Environmental Science, Udine, Italy
Zygutiene, Milda	Centre for Communicable diseases and AIDS, Vilnius, Lithuania

Suggested citation: European Centre for Disease Prevention and Control. A spatial modelling method for vector surveillance. Stockholm: ECDC; 2019.

Stockholm, November 2019

ISBN 978-92-9498-389-3
doi: 10.2900/633757
Catalogue number TQ-02-19-879-EN-N

Cover photo: Philippe Garcelon via flickr; attribution 2.0 Generic (CC BY 2.0)

Contents

Figures

Tables

Abbreviations

AIC	Akaike information criterion
BRT	Boosted regression trees
DEM	Digital elevation model
EVI	Enhanced vegetation index
GAUL	Global administrative unit layer
GIS	Geographic information system
GLM	General linear model
LST	Land surface temperature
MODIS	Moderate resolution imaging spectrometer
NDVI	Normalised difference vegetation index
NLDA	Non-linear discriminant analysis
NUTS3	Nomenclature of units for territorial statistics, level 3
RF	Random forest

Executive summary

Vector-borne diseases are a specific group of infectious diseases that are a (re-)emerging threat to Europe.

One important aspect of preparedness for vector-borne diseases is the surveillance of the introduction, establishment and spread of the main disease vectors. ECDC regularly publishes updated vector distribution maps at the NUTS3 level.

This document describes a methodology to estimate the vector distribution status for those NUTS3 units for which observations are not yet available. These estimates are produced with spatial modelling techniques, using the currently available distributions to calibrate the modelling process. This document provides an overview of gap analysis procedures, sets out the full methodology, and also provides details of which methodological components were used with each output provided.

The model outputs presented in this document are available on the ECDC website at: https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/prevention-and-control/vector-distribution-modelling.

1. Background

ECDC has a mandate to strengthen the European Union's capacity to prevent and control infectious diseases. Vector-borne diseases are a specific group of infectious diseases that are a (re-)emerging threat to Europe, requiring particular attention. The continuous increase of international travel and trade is one important risk factor for the introduction of new pathogens and vectors onto the continent, as is the extensive travel between mainland Europe and European overseas territories. Furthermore, changes in global climate may enhance the probability of previously absent vectors appearing in Europe and increase the further spread of vectors previously only present in limited numbers. All these factors could contribute to an increased risk for vector-borne disease transmission, representing a threat for outbreaks and the health of European citizens.

One important aspect of preparedness for vector-borne diseases is the monitoring or surveillance of the introduction, establishment and spread of the main disease vectors. The level of organisation and responsibility of vector surveillance activities differs between the EU Member States, and there are several stakeholders at European and international levels. In order to ensure a coordinated approach and strengthen preparedness for vector-borne diseases, ECDC launched a call for the establishment of a collaborative network of entomologists and other public health professionals. VBORNET, the ensuing network, started its activities in September 2009. The activities of the VBORNET network continue under the extended network VectorNet. More information on VectorNet is available from: http://ecdc.europa.eu/en/healthtopics/vectors/VectorNet/Pages/VectorNet.aspx.

Through VectorNet, ECDC maintains a database on the presence and distribution of vectors in Europe. The Centre regularly publishes updated distribution maps at the NUTS3 unit level. Substantial progress has been made in acquiring and validating available vector distributions, yet vector species maps at the NUTS3 level remain incomplete. Therefore, a methodology has been developed – subsequently referred to as 'gap analysis' – to provide estimates of the distribution status for those administrative units for which no information is currently available. These estimates are produced with spatial modelling techniques, using the currently available distributions to calibrate (or train) the modelling process. The 1-km-resolution outputs are then translated to values for each NUTS3 administrative unit.

This document provides an overview of the gap analysis procedures. The development of the methods used (and the input data) has been, and continues to be, an iterative process. This document sets out the full methodology, as developed to date, but also provides details of which methodological components were used with each output provided.

2. Methods

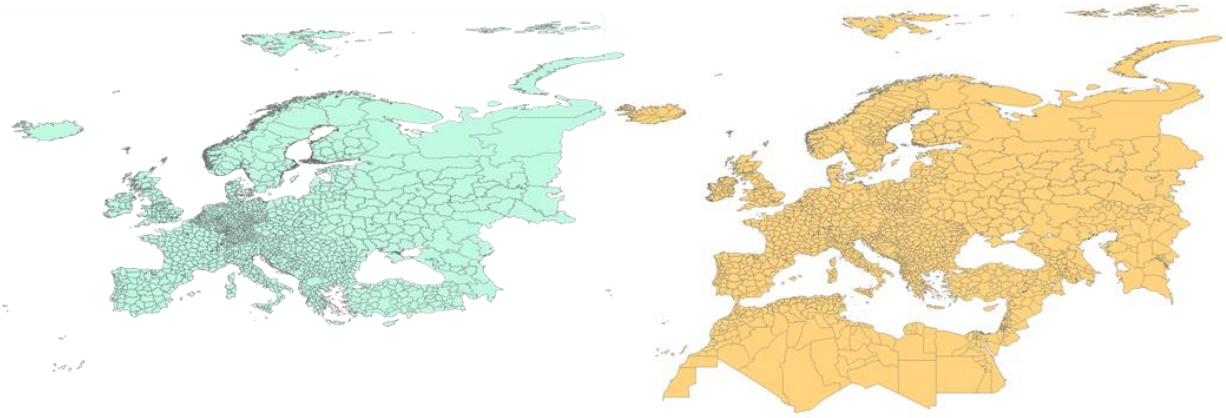
Overview

The basic modelling process involves establishing a statistical relationship between vector distribution data (the known presence, or absence status of a geographic area based on observations or expert opinion) and the values of a series of selected predictor covariates. These relationships are calculated for a set of sample locations, and the estimated equations are then applied to maps of the covariates which provide values at a pixel resolution. This results in a modelled spatial distribution with the probability of presence at the resolution of the covariate maps (standardised at 1 kilometre).

A number of challenges related to the available data in the VBORNET/VectorNet database needs to be addressed:

- The vector distribution status data were provided as presence, absence, or no data records for each NUTS3 unit within the project areas shown in Figure 1. Though these exclusively polygon-level records have more recently been complemented by the addition of point location records, the techniques have been developed to accommodate both polygon and point data.
- The modelling methods require both presence and absence points to calibrate the models. When there are only few or no absence points, they have to be inferred from other data, either by using known distribution limits from other sources or from a map of known habitat suitability from which unsuitable areas can be extracted as known absences.
- These raw 'medium-resolution' probability maps do not, however, provide a clear picture of vector presence or absence for each administrative unit and are thus difficult to compare with the original input data. Procedures were, therefore, developed to summarise the modelled outputs for each administrative polygon in order to provide an output format in which input and modelled data can be readily compared.

Figure 1. VBORNET (left) and VectorNet (right) project areas



An important aspect of these procedures has been the close involvement of international experts who provided the vector distribution data during all stages of the modelling project and who were critical in defining habitat suitability and validating output models.

The analyses cover a series of discrete stages that are expanded upon below:

- Selection and prioritisation of species to be modelled
- Extraction and evaluation of the known species distributions
- Provision of habitat suitability maps to provide absence data, if required
- Selection of covariates
- Extracting covariates data from sample points
- Spatial modelling
- Evaluation and selection of model outputs
- Conversion of selected modelled probability of presence to polygon level maps

Species prioritisation

While the primary prioritisation may be dependent on epidemiological relevance or strategic importance, it is only possible to model a species if sufficient information about its distribution is available to calibrate or train the modelling process. Ideally, these data should be evenly distributed over the entire area of interest, so that the known status locations are representative of the whole region to be modelled. Where this is not the case, it should be noted that the further away a location is from the nearest confirmed status location, the less likely it is to be reliably modelled. An initial selection process would therefore discard species with insufficient known data or with known data that are too clustered in parts of the project area.

There are a number of additional technical factors that need to be considered. One such factor is the feasibility of defining habitat suitability or range limits if no known absence data are available. The definition of habitat suitability for the model relies on either information from other sources or quantitative knowledge of environmental limiting factors or suitable habitat types. Another factor is whether new data are likely to be provided in the short term, which would suggest the modelling be delayed until these data are available.

The current report covers gap analysis modelling over a five-year period for the following species:

- Five species were selected for a first tranche (Phase 1): *Ixodes ricinus, Aedes vexans, Culex modestus, Phlebotomus perniciosus* and *Phlebotomus tobbi.* These were used to develop and test various possible methodologies.
- Once this process was complete, a further five species were selected for the second phase of analysis: *Dermacentor reticulatus, Hyalomma marginatum, Anopheles plumbeus, Phlebotomus papatasi* and *Phlebotomus ariasi*. The analyses were initially performed in late 2012 and early 2013. Extensive data collections allowed revisions of the tick models in early 2016.

Known distributions

The known European distribution data of the vector species used as input for the models are shown in Figures 2–4. The dates are shown in the figure titles. It is evident from these maps that the completeness of the known distribution data of each species varies widely, and in a number of different ways, each presenting different challenges to the modelling process:

- The phlebotomine data are relatively extensive and have many absence records, but are very regional, therefore the gaps are largely at the edges of the known distributions. If these edges were large relative to the known areas, the degree to which known distributions could be extrapolated by the modelling procedures could have been compromised.
- The mosquito records are the least complete, particularly in the central and eastern areas, though they do have both presence and absence records. Unlike the phlebotomine data sets, the areas with unknown distribution status often extend very far from areas with known presence or absence, and model predictions are likely to come with large uncertainties in these large areas.
- The tick distributions generally have quite wide-ranging presence data, but rather little absence data, and in the case of *Ixodes,* none at all. This represents a potential problem for spatial modelling techniques of presence and absence as they generally rely on having both presence and absence records in approximately similar proportions within the training data sets. Efforts (as described later) were made to add absence points to ensure that the total numbers of each category are balanced to be approximately equal for a number of reasons, firstly not to limit the spatial modelling techniques available for use, secondly to prevent the over-prediction of the species distribution, and thirdly to include higher resolution data to identify areas of absence within coarse regions recorded as positive. It is notable also that the more recent records of some of the tick species include both point and polygon data. This will become standard in future.

Figure 2. Known sandfly distribution data in 2013

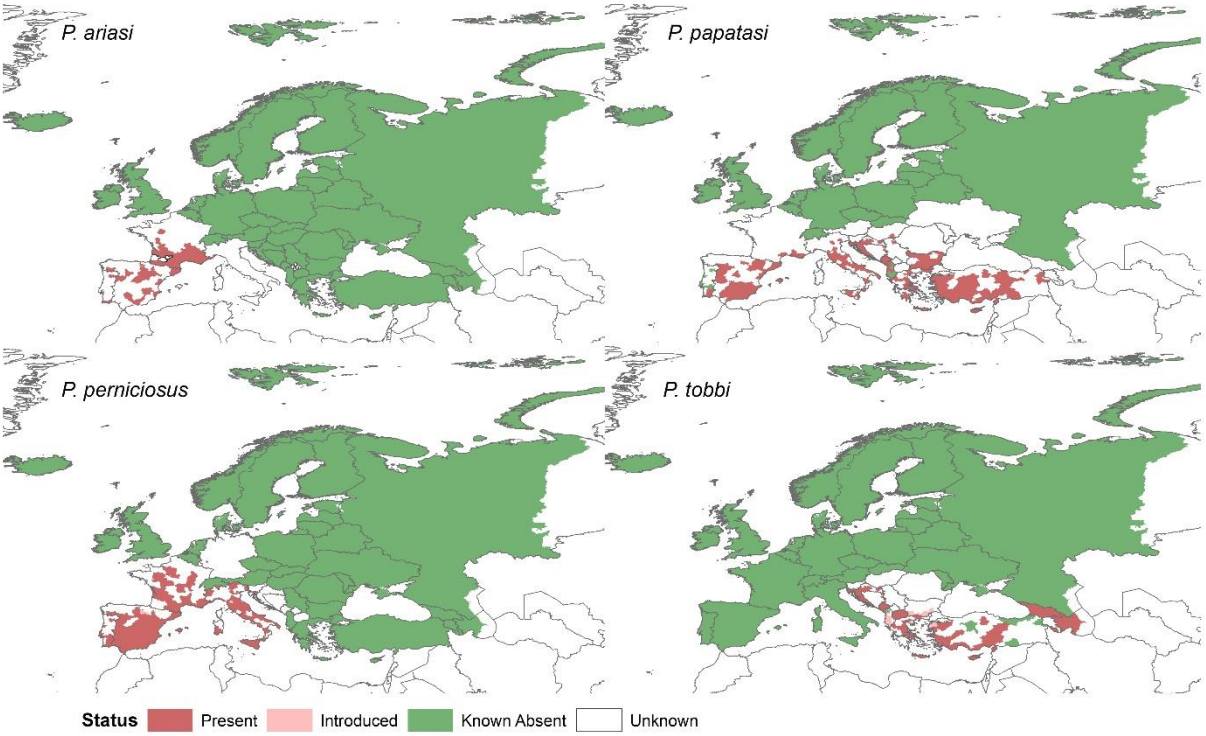


Figure 3. Known mosquito distribution in 2013

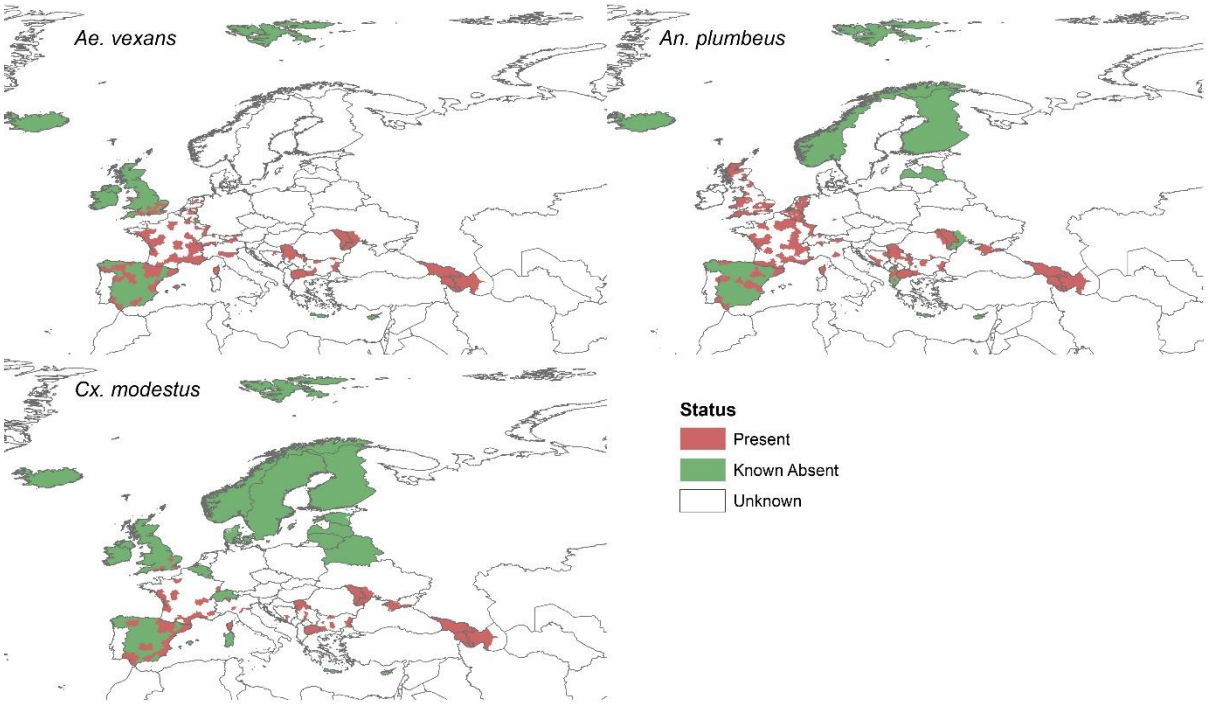
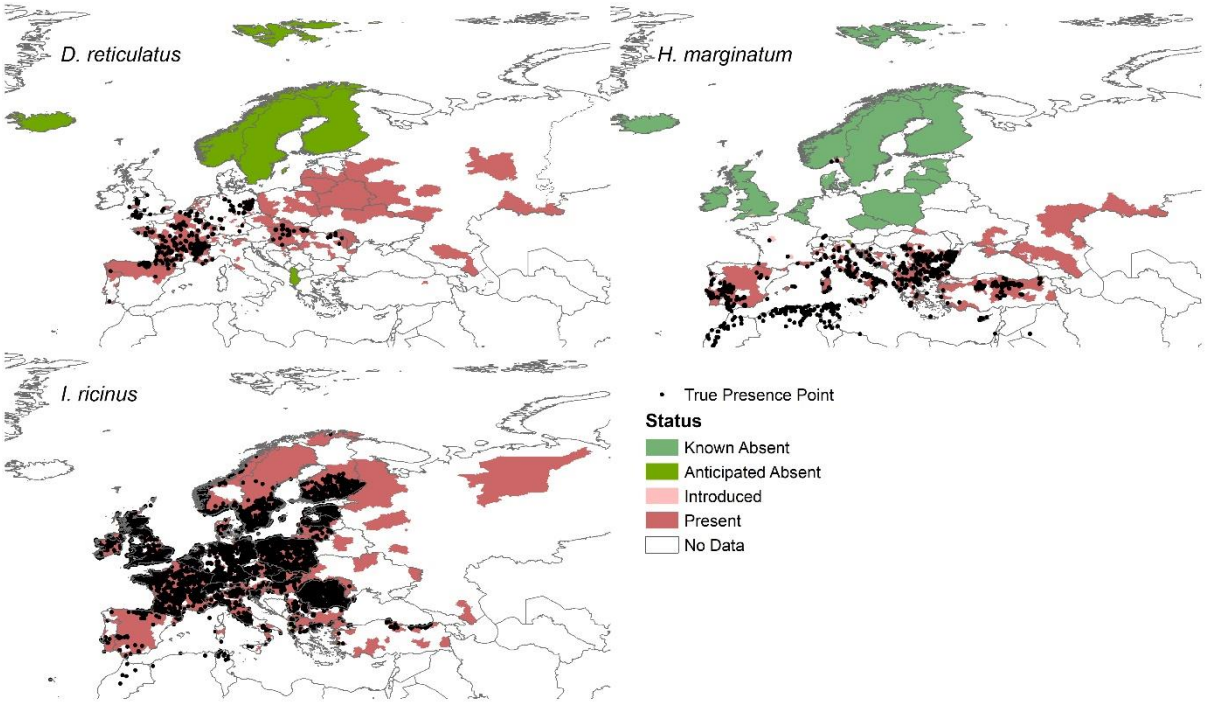


Figure 4. Known tick distribution data in 2015



One approach to the problem of incompleteness in these data sets is to wait until there are sufficient additional data gathered in order to model the data using standardised spatial distribution modelling techniques. This condition is, however, unlikely to be met in the near future, and there will always be patterns of missing data that require special treatment. Therefore, a number of strategies are required to deal with all three categories of known distribution data (presence point locations, presence area locations, and absence locations). In short, presence points are added to the model training data, they are also compared with the presence area (polygon) data, which are updated if required before being used to generate additional presence training data. Absence data are compared with habitat data, where available and as described in the next section.

Defining absences

The most persistent challenge has been to define absence data where these records are sparse or missing from the project archives, as exemplified by the data for the tick *Ixodes ricinus* (Figure 4). One widely used way to do this is to define 'pseudo absences' or 'background points' [1,2] so that there are both presence and absence data to train the modelling. These pseudo absences are generally defined by the distance from known presence points, anything further away from a known presence than a set distance is set to absent, while the background points are defined by the selection of points within a species range determined purely by their position in relation to known presences. Such approaches may be appropriate when the available data are reasonably complete within the known range of a target species, and it is reasonable to assume that the absences will be around the edge of a distribution. It is less certain that such a method is appropriate when the data are known to be rather incomplete throughout the potential range, and so where the known presences are rather sparsely located within what is likely to be a continuous distribution: the sparser the known presence data, the more likely it is that pseudo absences will be defined in areas where the target species is, in fact, present.

One way around this is to use information about the factors which limit the distribution of the target species rather than purely geostatistical approaches; examples are habitat preferences and range limits imposed by climatic factors [3]. Areas defined as unsuitable can then be set as 'inferred absent', and standard presence/absence modelling techniques can then be used. An advantage over the standard pseudo-absence methods is that such inferred absences are not distance dependent and, as they do not assume reasonably complete sample coverage, can be located within the general range of recorded presence. The approach is especially applicable to polygon-level presence records, as limiting conditions within these polygons can be used to introduce absence points within areas that would otherwise be assigned a status of (entirely) present. Thus, for example, a polygon may be assigned a status of present, while the unsuitable habitats within it can be set to absent.

The habitat types each vector species prefers were defined accordingly. Experts were asked to define primary, secondary and unsuitable habitat types from two land cover maps: CORINE 2006, which covers the EU and Turkey, but not the eastern- or southernmost parts of the project coverage [4], and GLOBCOVER 2009, which is a global product [5]. From these, a combined suitability map for each vector species was produced, using both land use

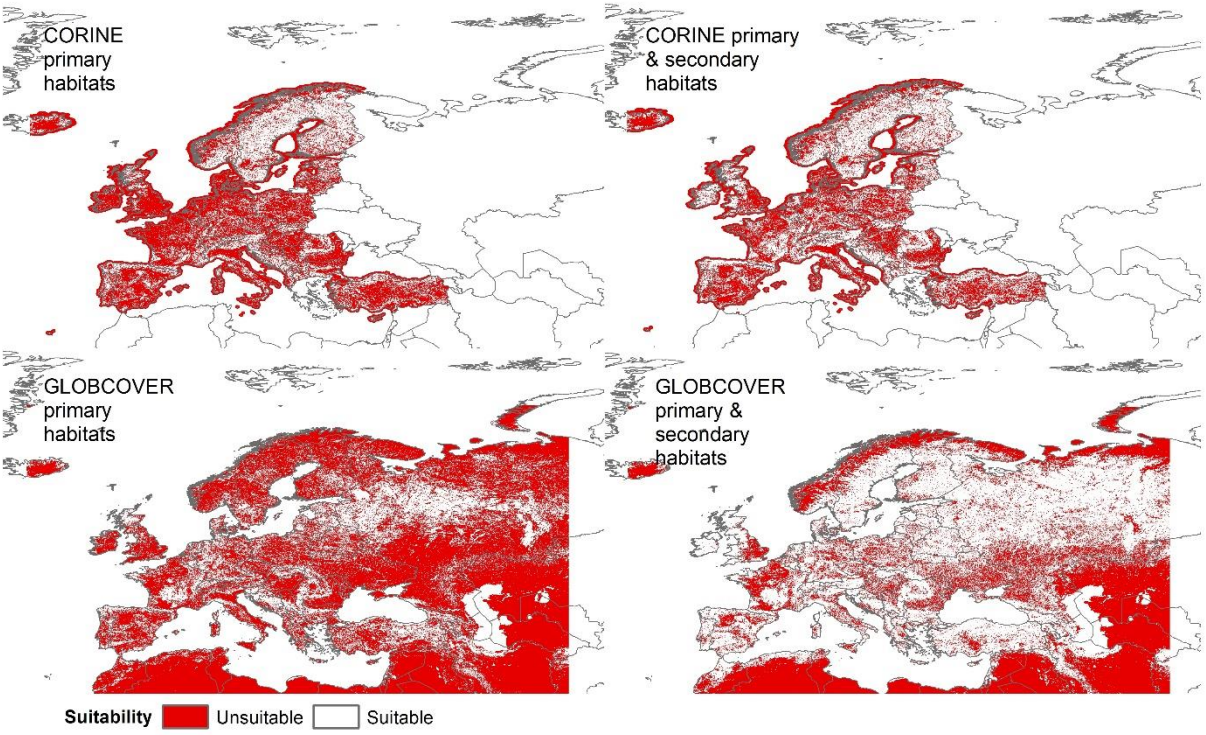
definitions. This process was iterative: all versions of the habitat maps were evaluated at local and regional level by sending the complete range to the relevant experts and asking them to select the habitat map which they felt best represented an accurate map of the suitable habitats and potential distribution of the vector(s) in which the experts specialised.

A number of issues became apparent during this process:

- Experts found the CORINE land classes easier to assign to the suitability categories, and so this system was given default priority where its coverage was available, and the GLOBCOVER was used only where CORINE data were not available.
- Though different experts did, by and large, agree (occasionally after some negotiation) on the definitions, it became clear that the same habitats may assume different priorities in different areas. This was most evident for *I. ricinus*, for which natural grassland and moorland was defined as primary habitat in the UK, but secondary in northern continental Europe. Fortunately, this did not affect the outcome of these analyses, as the relevant habitats were comparatively rare in continental Europe, though subsequent revisions of these procedures and analyses of different species may need to take such regional variations into account.

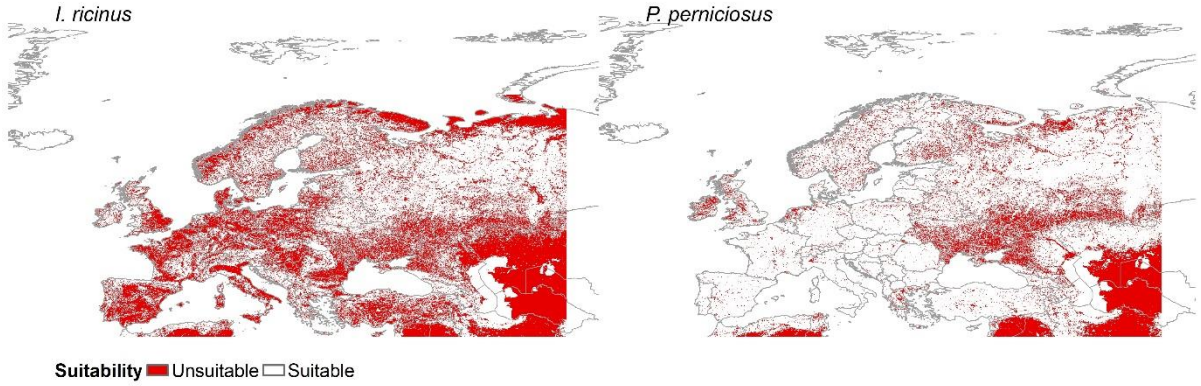
These various habitat maps illustrated for *I. ricinus* in Figure 5 were used in a number of ways, most importantly to define unsuitable habitats to identify inferred absences – as discussed above – and also to help interpret the model outputs, using the assumption that primary habitats are more likely to be associated with a high predicted probability of presence. Thus, a model with extensive mismatches between predicted probability values and the expert-defined category could be assumed to be less reliable than one with no such disparities.

Figure 5. Range of habitat-masked distributions defined for *Ixodes ricinus*



In a number of cases, it was possible to also use environmental limiting factors to define further absence regions. In the case of *I. ricinus,* for example, additional information about environmental limits was available and could be readily mapped, and so the habitat threshold was combined with ecological thresholds reflecting the fact that the tick is only present in areas with fewer than 175 days of snow cover per year and where the vegetation period was greater than 180 days. Altitude and temperature thresholds were also used for some of the other vector species. The details are given in Annex 1 and the masks combining habitat and limiting factors are shown in Annex 2. The habitat suitability maps themselves are provided online and available from: https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/prevention-and-control/vector-distribution-modelling. Examples generated for *I. ricinus* and *P. perniciosus* are provided in Figure 6.

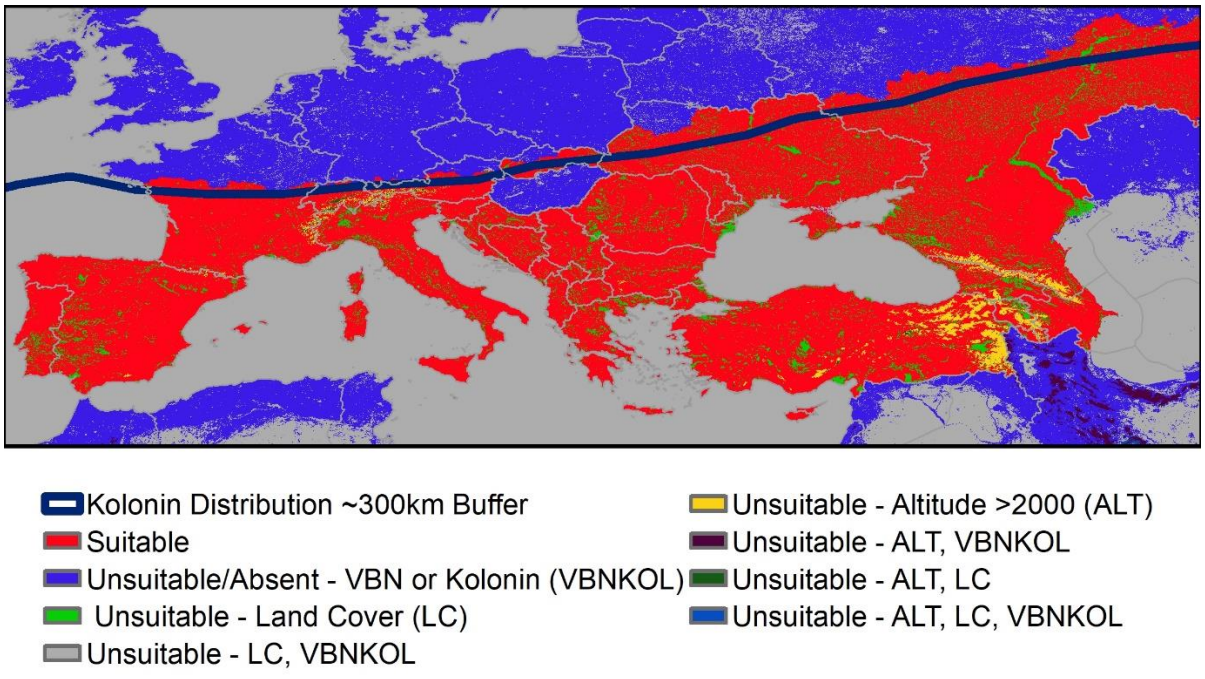
Figure 6. Habitat suitability based on habitat types and environmental limiting factors for *Ixodes ricinus* and *Phlebotomus perniciosus*



Another source of absence information included other mapped data sets – which in the case of ticks, is exemplified by the global distributions provided by the data held in the webpages on fauna of ixodid ticks of the world [6] (www.kolonin.org which was available until 2014, but has since been closed). Data such as these can be used to define areas in which presences are to be expected. It can also be assumed that the species are unlikely to be found far from the edges of a defined presence polygon: in these analyses, appropriate data were available for *D. reticulatus* and *H. marginatum* (but not *I. ricinus* as its presence extended beyond the project areas) and absence was defined at 300 km or more from the presence polygon boundary. The resulting absence data were, therefore, taken from locations deemed to be unsuitable for the species within the project 'present', 'unknown' or 'no data' areas, as well as 300 km beyond the presence polygons.

The presence data sample locations were taken only from the suitable habitat within the project administrative units defined as 'present' (see Figure 7), and, where available, from point location data. This process ensured that presence points were taken only from ecologically or climatically suitable locations within a polygon designated as 'present' and not from unsuitable areas.

Figure 7. Presence and absence data of *Hyalomma marginatum* based on VBORNET status 2013



Covariates

The covariates tested in the modelling procedures were drawn from a standardised set of ecological parameters, in particular a suite of Fourier-processed MODIS satellite images which provides a range of biologically interpretable variables related to levels and seasonality of temperature and vegetation related factors. These covariates have been widely used in species distribution modelling [7, 8] since their initial production in 2005. These analyses used

covariate time series between 2001 and 2012. The covariates are summarised in Table 1, and all are available to registered members of the Spatial Data Website (www.spatialdatasite.com).

Table 1. Covariates offered to modelling procedures.

Covariate	Covariate
1. EDYY03A0: Middle infra-red mean	39. EDYY14P3: NDVI phase 3
2. EDYY03A1: Middle infra-red amplitude 1	40. EDYY14VR: NDVI variance
3. EDYY03A2: Middle infra-red amplitude 2	41. EDYY15A0: EVI mean
4. EDYY03A3: Middle infra-red amplitude 3	42. EDYY15A1: EVI amplitude 1
5. EDYY03MN: Middle infra-red minimum	43. EDYY15A2: EVI amplitude 2
6. EDYY03MX: Middle infra-red maximum	44. EDYY15A3: EVI amplitude 3
7. EDYY03P1: Middle infra-red phase 1	45. EDYY15MN: EVI minimum
8. EDYY03P2: Middle infra-red phase 2	46. EDYY15MX: EVI maximum
9. EDYY03P3: Middle infra-red phase 3	47. EDYY15P1: EVI phase 1
10. EDYY03VR: Middle infra-red variance	48. EDYY15P2: EVI phase 2
11. EDYY07A0: Daytime LST mean	49. EDYY15P3: EVI phase 3
12. EDYY07A1: Daytime LST amplitude 1	50. EDYY15VR: EVI variance
13. EDYY07A2: Daytime LST amplitude 2	51. EDBC2K12: BIOCLIM Annual Precipitation
14. EDYY07A3: Daytime LST amplitude 3	52. EDBC2K13: BIOCLIM Precipitation of Wettest Month
15. EDYY07MN: Daytime LST minimum	53. EDBC2K14: BIOCLIM Precipitation of Driest Month
16. EDYY07MX: Daytime LST maximum	54. EDBC2K15: BIOCLIM Precipitation Seasonality (Coefficient of Variation)
17. EDYY07P1: Daytime LST phase 1	
18. EDYY07P2: Daytime LST phase 2	55. EDBC2K16: BIOCLIM Precipitation of Wettest Quarter
19. EDYY07P3: Daytime LST phase 3	56. EDBC2K17: BIOCLIM Precipitation of Driest Quarter
20. EDYY07VR: Daytime LST variance	57. EDBC2K18: BIOCLIM Precipitation of Warmest Quarter
21. EDYY08A0: Nighttime LST mean	58. EDBC2K19: BIOCLIM Precipitation of Coldest Quarter
22. EDYY08A1: Nighttime LST amplitude 1	59. EDV590AS: DEM (Aspect)
23. EDYY08A2: Nighttime LST amplitude 2	60. EDV590EL: DEM (Elevation)
24. EDYY08A3: Nighttime LST amplitude 3	61. EDV590RG: DEM (Ruggedness)
25. EDYY08MN: Nighttime LST minimum	62. EDWC57A0: WORLDCLIM precipitation mean
26. EDYY08MX: Nighttime LST maximum	63. EDWC57A1: WORLDCLIM precipitation amplitude 1
27. EDYY08P1: Nighttime LST phase 1	64. EDWC57A2: WORLDCLIM precipitation amplitude 2
28. EDYY08P2: Nighttime LST phase 2	65. EDWC57A3: WORLDCLIM precipitation amplitude 3
29. EDYY08P3: Nighttime LST phase 3	66. EDWC57MN: WORLDCLIM precipitation minimum
30. EDYY08VR: Nighttime LST variance	67. EDWC57MX: WORLDCLIM precipitation maximum
31. EDYY14A0: NDVI mean	68. EDWC57P1: WORLDCLIM precipitation phase 1
32. EDYY14A1: NDVI amplitude 1	69. EDWC57P2: WORLDCLIM precipitation phase 2
33. EDYY14A2: NDVI amplitude 2	70. EDWC57P3: WORLDCLIM precipitation phase 3
34. EDYY14A3: NDVI amplitude 3	71. EDWC57VR: WORLDCLIM precipitation variance
35. EDYY14MN: NDVI minimum	72. EDXXGRPD: GRUMP Population density
36. EDYY14MX: NDVI maximum	73. EDXXGRPW: GRUMP Population weighted
37. EDYY14P1: NDVI phase 1	74. EDXXJRCA: Travel time (Joint Research Centre)
38. EDYY14P2: NDVI phase 2	75. EDXXLPG1: Length of Growing Period LGP

LST = land surface temperature. NDVI = normalised difference vegetation index; EVI = enhanced vegetation index. DEM = digital elevation model. All files starting with EDYY are Fourier-processed MODIS satellite images produced by the TALA Research Group Department of Zoology, University of Oxford. All analyses prior to 2013 used a time series running from 2001–2008. Thereafter, this suite was replaced with outputs derived from a 2001–2012 time series. Files with Bioclim and Worlclim in the filename are derived from WORLCLIM data sets (www.wordclim.org). GRUMP is derived from population layers produced by http://sedac.ciesin.columbia.edu/. JRC Accessibility downloaded from http://bioval.jrc.ec.europa.eu/products/gam/index.htm; Length of growing period is derived from data provided by FAO, Rome. All layers extracted and standardised by ERGO for EDENext and VMERGE (www.edenextdata.com).

Modelling

Spatial models quantify the statistical association between response variables (e.g. vector presence or absence) and predictor covariates for a number of sample locations [9]. Once calibrated, these models are used to make predictions for the response variable in locations where only data on one or more predictor variables are available. Different spatial modelling methods are optimised to address specific limitations and bias of input data, for example dispersion, patchiness, relative numbers of vector presence and absence. In this analysis, a comparison of established methods was made by both statistical methods and expert opinion.

Extraction of sample points

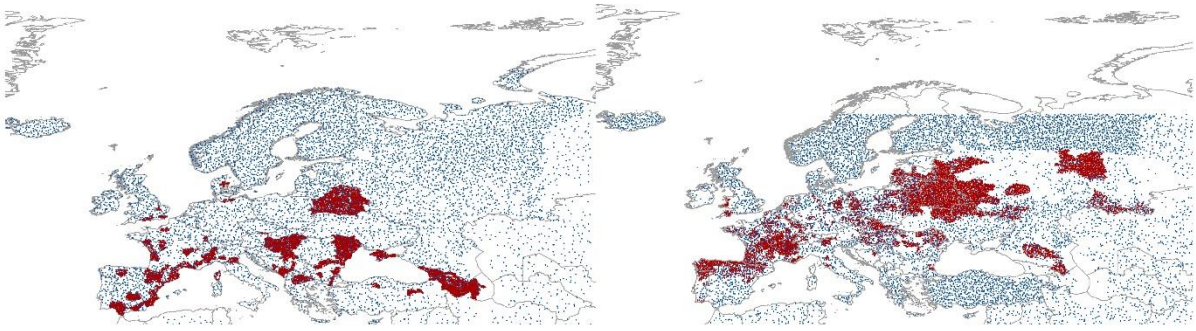
Once the maps of known presence and absence status have been prepared, it is necessary to define a series of sample points for which to extract covariate values for each category. As there are typically response values for both polygons and point locations, two separate protocols are required:

- For polygon locations, the number of points is set according to the areas of the training data coverage to provide a sample point every 50–100 km.
- By default, all locations with point data are used.

These combined samples are then screened to establish the ratio between the number of recorded presence and absence points. Additional absence points are subsequently defined within the entire study area to ensure that the total numbers of each category are balanced to be approximately equal. During consultations with vector experts it was confirmed that balancing the presence/absence ratio in this manner results in improved output maps.

Examples are provided in Figure 8 for the *Cx. modestus* and *D. reticulatus* 2015 iterations, which illustrate one of the hurdles to be overcome in generating the sample location data set. The presence data are restricted to a relatively small area in central and southern Europe, but outposts in northern Europe indicate that it is not advisable to produce a regional model. Also, the extent of the absence area is rather larger than that of the known presence. Further, the geography of habitat suitability is such that absence regions within the presence polygons are quite small. One way to deal with this is to define absence and presence point locations separately using the 'create spatially balanced point' geoprocessing tool in Esri's ArcMap. This extracts a user-defined number of sample points from raster probability images, with the number of locations defined by the probability value. The most straightforward implementation is to create separate binary images for presence and absence status and set an equal number of points to be sampled from each image, which can then be concatenated into a single sample location data set.

Figure 8. Presence (red) and absence (blue) sample points, *Culex modestus* (left), *Dermacentor reticulatus* (right), 2015 iterations



Modelling extent

For several species – notably *P. tobbi*, *P. papatasi, D. reticulatus,* and *H. marginatum* – the known distribution status (particularly presences) was restricted to fairly contiguous regions of Europe, and the known absence status was also rather spatially coherent. For these species, it was not necessary to build continental models, and the model extents were confined to the area of known presence, plus all administrative units with unknown distribution status and an extensive buffer area (> 200 km) into the known absence areas.

Modelling methods

The range of modelling techniques used to provide the candidate outputs for expert assessment included non-linear discriminant analysis (NLDA), random forest (RF), general linear model (GLM) regression-based methods (including logistic regression) and boosted regression trees (BRT). All these were implemented using the VECMAP [13] modelling system.

Earlier analyses focussed on NLDA and RF [10, 11]. More recently, BRT has replaced NLDA because models based on BRT provide a stronger contrast than those produced with the RF technique. BRT also accommodates non-linear relationships more effectively than NLDA, so it is less limiting in the potential range of extrapolation. Models produced with NLDA may also be more affected by zonation than BRT, which, by its nature, accounts for clustering in geographic and covariate space within its operation. Recent RF models also include a zoned element in the analysis, whereby separate models were calculated for distinct spatial stratification zones related to ecosystem type. This is intended to let RF produce models that are more closely tailored to local conditions.

In order to reduce the impact of certain methodological idiosyncrasies, it was decided to produce a range of models by using at least two distinct methods (Table 2). These could then be assessed and compared to identify and select the best one. Model outputs were evaluated using standard, and extensive, accuracy metrics (e.g. ROC, AUC, Kappa, Confusion matrices) as provided by the VECMAP software [13]. All models were run on a random 75% of the sample data set to ensure variability of replicates, and automatic covariate variable reduction was applied. Model validation was implemented for the training data locations. All models presented to the experts for validation or incorporated into the ensembles had AUC values of 0.85 or better, which indicates a highly significant model fit. Ground truthing via field sampling has yet to be formally completed on any of the models beyond the expert validation based on local knowledge, as mentioned elsewhere in this document.

In addition to the statistical analysis of predictive accuracy, experts and vector data providers were asked to help select or reject candidate models from the range provided, based on comparisons with known detailed local and regional distributions. In this process, assessments of unexpected results, such as false negatives and false positives, were especially useful. At times this led to changes in data used for identifying environmental limits, for example to improve predictions.

An alternative could have been to select the model outputs solely based on accuracy levels, but, as the metrics of all selected models indicated high reliability, these were not seen to provide sufficient discriminatory power, and expert evaluation was considered to be needed to select the definitive outputs.

As a results of this approach, usually more than one model output was found to be statistically reliable and judged to be accurate by the experts. In this event, there was no obvious way to select one model over another and so, following recent trends, the 'validated' models were averaged (ensembled) to provide a 'consensus' product, which would be less prone to anomalies in predicted distributions, biases in the input data, or vagaries in the statistical methodologies. Relevant experts also evaluated these ensembles.

Masking

All model outputs are masked by the suitability masks (see Annex 1) used to define 'inferred absences'. These masks do not, however, incorporate the bounding distributions used to help define absences beyond the extents of known presence status, as this allows the modelling process to identify potential areas with presence in suitable habitats beyond currently defined ranges. These locations may well be the most likely areas for the vectors to spread to in the future.

Conversion of pixel-level probabilities to NUTS3 outputs

After producing and selecting a preferred candidate from the range of model outputs assessed by the experts, one final step remained: converting the 1-kilometre-resolution model of probability of vector presence to the standard project format of NUTS3 areas.

Spatial models rarely produce zero-probability predictions and thus will generally not predict absolute 'absences' even at the pixel level, and certainly not when summarised probabilities for entire NUTS3 units are calculated. Further, a mean NUTS3 probability of, for example, 0.3 may be derived from either a narrow range of values around 0.3 or a combination of high and low values. With a probability of 0.5 as the threshold value for defining modelled presence, the former would be interpreted as entirely absent, but the latter as a combination of present and absent.

Therefore, a more discerning method was developed to describe the level of vector presence within each unit: All model pixel outputs, after unsuitability masking, were converted to binary presence or absence, in accordance with the modelled probability, using a 50% threshold value to indicate presence. The percentage of each NUTS3 polygon area – with the vector predicted to be present – was calculated and mapped. The resulting values were then categorised as 'negligible predicted risk' (<1%), 'low predicted risk' (1—25%), 'medium predicted risk' (25—75%), and 'high predicted risk' (>75%). These estimates were combined with the original project input distributions to run the final gap analysis and generate NUTS3-level outputs.

3. Results

The full set of outputs is detailed in Table 2, which provides the model technique, modelled extent, analysis date, and revision date, if applicable. The mapped outputs are available online from: https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/prevention-and-control/vector-distribution-modelling.

Table 2. Model details

Species	Model type	Latest date	Extent	AUC
Ticks				
Ixodes ricinus	*Zoned RF*	Spring 2016	Europe	0.9396
	BRT			0.9871
	Ensemble			0.9918
Dermacentor reticulatus	*Zoned RF*	Autumn 2013	Europe	0.923
	NLDA			0.958
	Ensemble			0.964
Hyalomma marginatum	*Zoned RF*	Spring 2016	Europe	0.91
	BRT			0.904
	Ensemble			0.9143
Mosquitoes				
Anopheles plumbeus	*Zoned RF*	Autumn 2013	Europe	0.981
	NLDA			0.926
	Ensemble			0.994
Culex modestus	RF	Spring 2016	Europe	0.955
	BRT			0.943
Aedes vexans	RF	Autumn 2012	Europe	>0.85
	NLDA			>0.85
Sandflies				
Phlebotomus ariasi	*Zoned RF*	Autumn 2013	South-west EU	0.92
	NLDA			0.99
	Ensemble			0.989
Phlebotomus papatasi	*Zoned RF*	Autumn 2013	Southern EU	0.94
	NLDA			0.975
	Ensemble			0.99
Phelbotomus tobbi	**NLDA**	Autumn 2012	South-west EU	>0.85
Phlebotomus perniciosus	**RF**	Autumn 2012	Southern EU	>0.85
	NLDA			>0.85

RF = random forest, BRT = boosted regression trees, NLDA = non-linear discriminant analysis, bold = selected model, italics = ensembled components. Sample point number min. 200, max. 20 000

This section uses three case studies to illustrate the results obtained from different types of input distribution. These are:

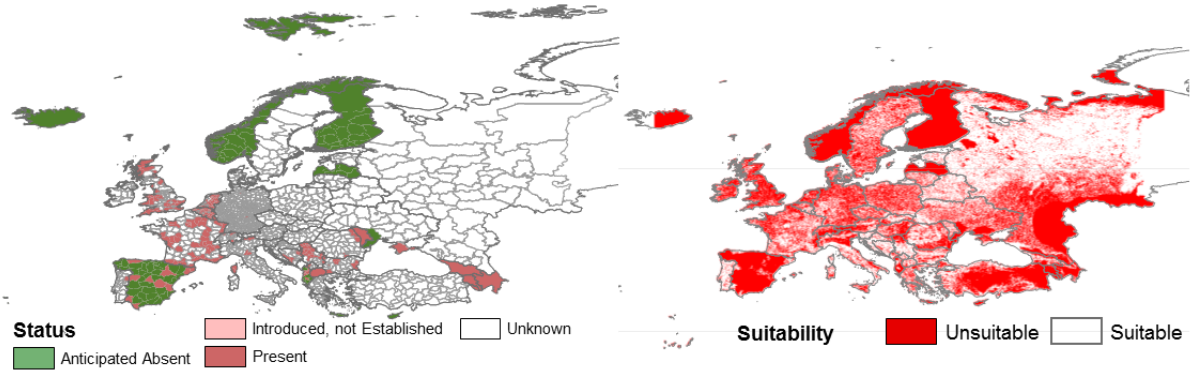
- *A standard model typified by* Anopheles plumbeus*;*
- *A model with no absent data initially provided illustrated by the case of* Ixodes ricinus*; and*
- *A regional model as illustrated by* Phlebotomus tobbi

A standard model – *Anopheles plumbeus*

The input data presented in Figure 9 show the 2012 version of recorded presence/absence at the NUTS3 unit level (left) with the unsuitability mask (right). The recorded distributions are quite sparse – the majority of the polygons have no data associated with them – but what data there are (both presence and absence) are quite well distributed to cover the whole of continental Europe. There is also an approximate balance between presence and absence records.

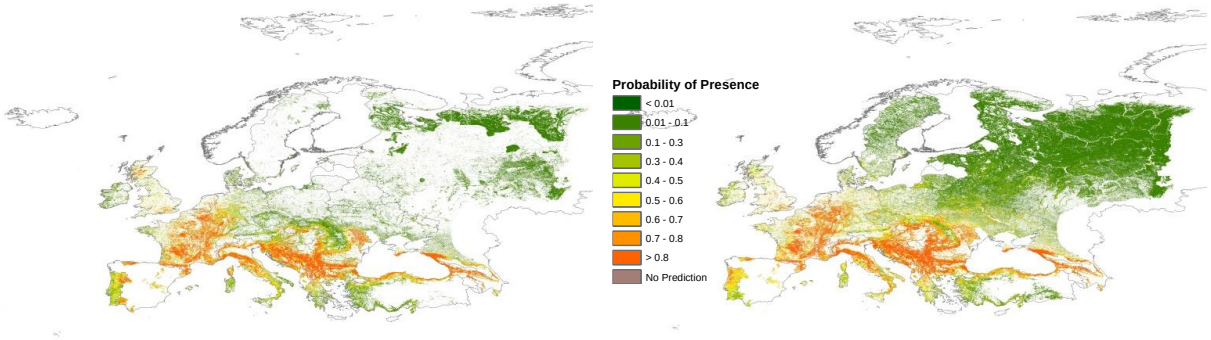
The suitability map shows a large area of unsuitable habitat. The suitable areas that were defined consist largely of the habitat categories containing at least some deciduous forest, while the unsuitable habitats without deciduous forest are quite extensive. When combined with the polygons with known absence, this provides a good coverage for known or inferred absence across the whole project area.

Figure 9. Distribution status as of 2012 (left) and habitat suitability (right), *Anopheles plumbeus*



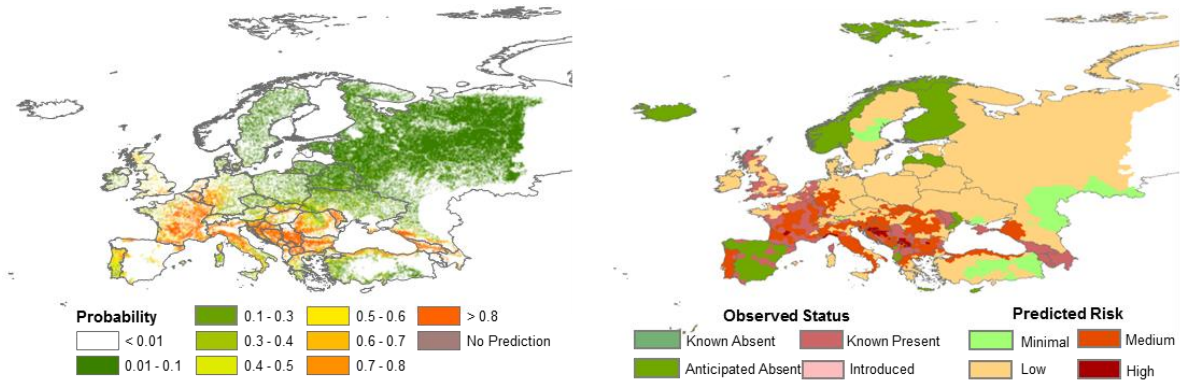
Both NLDA and RF models were implemented for this species (Figure 10). At first sight, these appear quite different as there is much more green in the RF model. This green class, however, represents the lowest predicted probability class (lower than 1% probability) and thus is more or less certain to represent absence like the white class in the NLDA map. More importantly, the yellow and orange classes inferring presence are very similar in both models.

Figure 10. NLDA (left) and RF (right) model outputs for *Anopheles plumbeus* (as of 2012)



The experts were unable to select the best output between the two, and the final model generated was an ensemble of the two (Figure 11). The NUTS3 output derived from this model is shown on the right of the same figure, overlain with the original input polygon records to produce a combined NUTS3-level map of known and predicted risk. The match between the two measures is fairly straightforward: the predicted risk – as indicated by the proportion of each NUTS3 unit that is predicted to support the vector – is highest in central European areas near the recorded presences. It is, by contrast, lower in the UK even though there are recorded presences there, which most probably reflects the much higher proportion of unsuitable land. The risk in Portugal is predicted to be medium, despite the extent of the neighbouring recorded absences, which is likely to reflect the difference in environmental conditions between Portugal and Spain.

Figure 11. Ensembled model (left) and derived NUTS3 unit risk overlaid with known distribution status (right) for *Anopheles plumbeus* (2012)

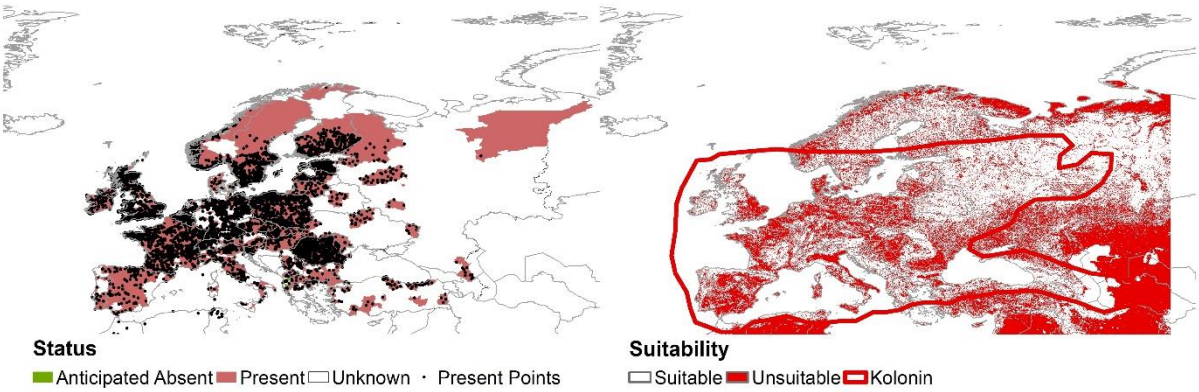


A model with no recorded absences – *Ixodes ricinus*

The data for *I. ricinus* represent an example for which there are plenty of occurrence records signifying presence, but there are few, if any, records of absence. *I. ricinus* is a very widely recorded tick vector of several significant diseases such as tick-borne encephalitis and Lyme borreliosis, and so is an important species from a public health perspective. As a result, it was high on the priority list for modelling and regular updating, despite the unbalanced recorded data sets.

The data for this species included presence records as polygons and as recorded geographic locations. Absence data were very limited in number and only defined by project fieldwork designed to establish the northern edge of the species range (where the vector has been recorded as introduced but not established (Figure 12). This is not sufficient to allow any of the modelling to produce a reliable prediction, and the majority of absence records offered to the model were therefore taken from the unsuitable areas defined by the habitat mask (derived from land cover data sets and snow cover data) (Figure 12).

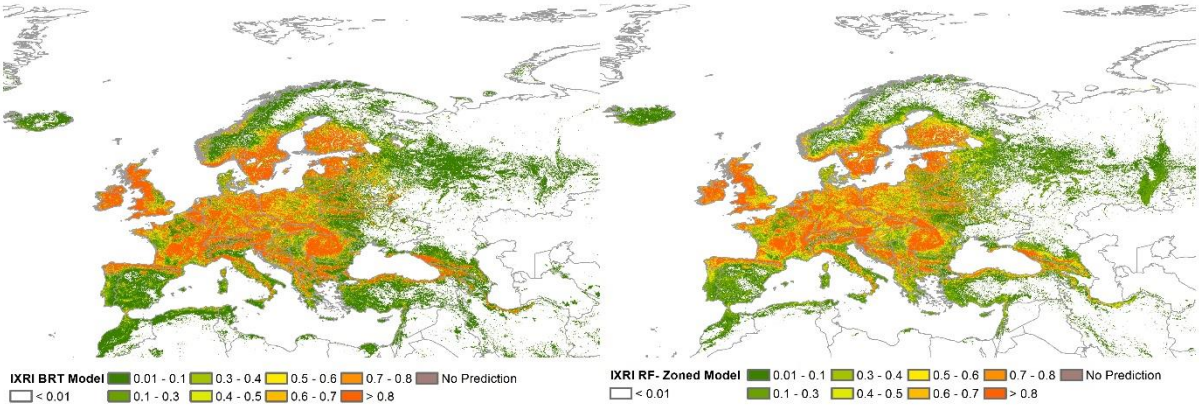
Figure 12. Status (left) and habitat (right) suitability, *Ixodes ricinus* (spring 2016)



The habitat mask graphic also shows the boundary of the vector's distribution according to Kolonin (buffered with 300 km), which, in this case, was not used to add additional absence points to the model as experts indicated the mask was representative of unsuitability on its own. The coincidence between the Kolonin boundaries and the model outputs in Figure 13 does, however, suggest that they are both likely to be fairly accurate.

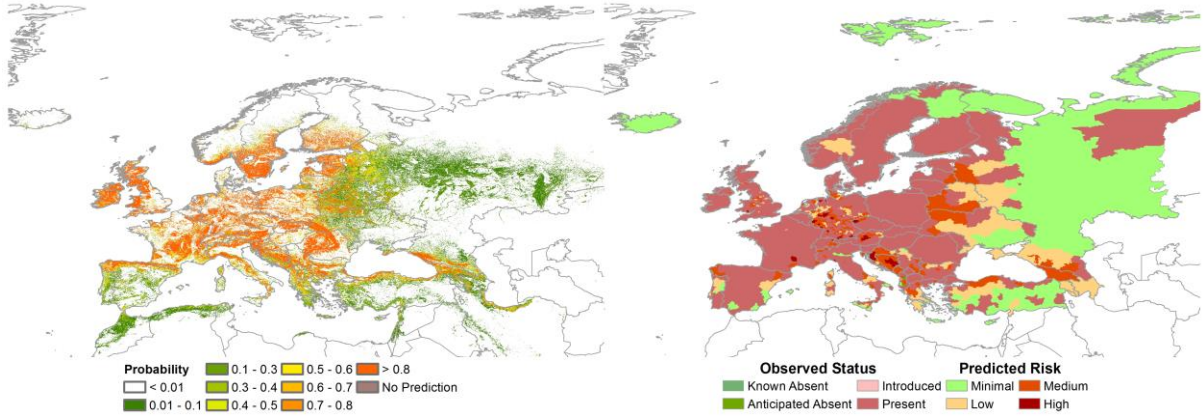
Two model methods were used for this species, namely zoned RF and BRT, with 100 and 1000 prediction trees, respectively (Figure 13). As with the analyses of *An. plumbeus* presented above, the outputs are very similar for the high predicted probabilities, but differed substantially for the low probability classes. Experts who were asked to assess and validate the models expressed no clear preference for one over the other, and the two candidates were therefore ensembled and then masked to provide the final output model.

Figure 13. BRT (left) and RF (right) unmasked models, *Ixodes ricinus* (spring 2016)



The final output masked ensemble model and its derived NUTS3 unit risk maps are shown in Figure 14. A key feature of the model is that it shows extensive areas of low probability within the NUTS3 units that are recorded as positive. The two maps, therefore, provide a somewhat different impression of the vector spatial distributions, especially in areas such as Spain or northern Scandinavia, where the predicted areas of presence within the present polygons are very restricted. This shows that the admin unit level maps do indeed disguise considerable detail, especially at the edges of the vector's distribution.

Figure 14. Ensembled and masked model (left) and derived NUTS3 unit risk overlaid with known distribution status (right), *Ixodes ricinus* (spring 2016)

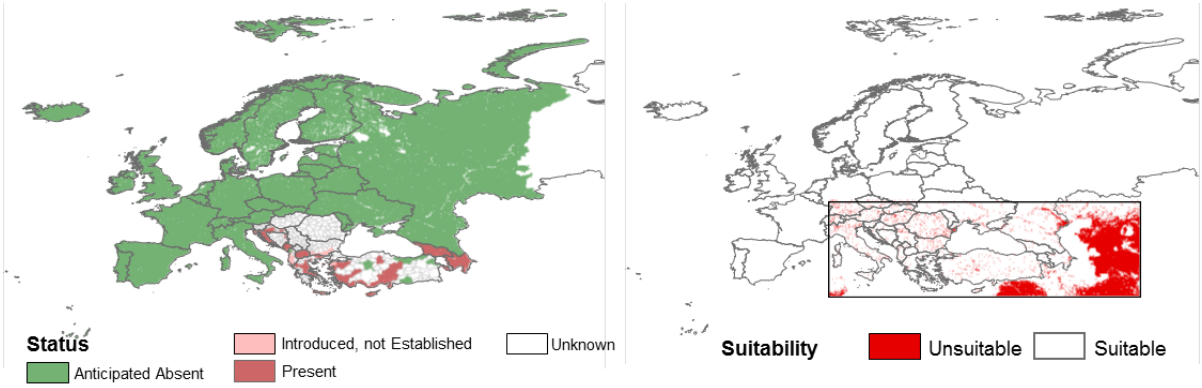


The predicted probability of presence values (Figure 14, right) compares well with the known distribution status data: the predicted values are 'medium risk', which means that between a quarter and three-quarters of the NUTS3 unit area are predicted to support the vector.

A regional model – *Phlebotomus tobbi*

Some of the known distribution status data cover only parts of the project area with regard to recorded presences and known or anticipated absences confined to geographically coherent regions. This means that the model may be restricted to a limited area: it includes a suitable number of known absences and presences, but does not need to be extended to the whole project area.

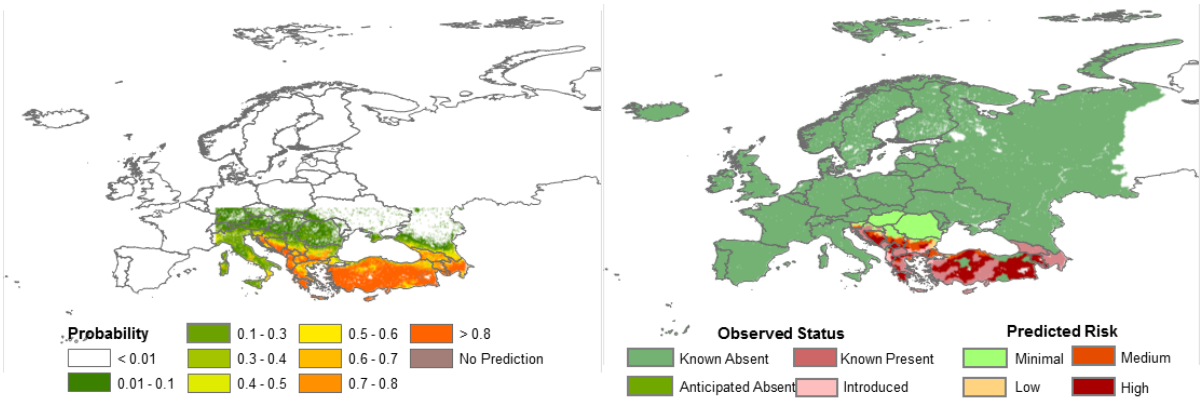
Figure 15. Polygon and habitat suitability of *Phlebotomus tobbi* (as of 2012)



This is the case for the sandfly *P. tobbi,* which is only recorded as present in the south-east quadrant of the project area, and known or anticipated to be absent elsewhere. The models were restricted accordingly to cover a rectangle which contains all known presences, a similar surface area of absences, and a 'no data' area in-between (Figure 15). There are several important benefits to this regional approach: it is quicker to calculate because the area is smaller; it avoids the risk of extrapolating too far from the known presence data, which could lead to anomalous results such as the prediction of spurious presences in the areas most remote from the recorded occurrences where the covariate values may be very different to those in the core range.

For *P. tobbi,* emphasis was placed on detailed examination of a single model rather than a number of models. In particular, the apparent contradictions between recorded and predicted distributions were investigated. Examples of these contradictions are found in central Turkey, northern Turkey, and southern Italy, for which the model suggested that the vector is present – though only in a few regions and at comparatively low probabilities – while the recorded data showed absence. In a number of cases (mostly in Turkey), the status assigned to these NUTS3 units was within the VectorNet database, which was re-evaluated by experts and amended if justified.

Figure 16. Selected model and NUTS3 unit level risk overlaid with known distribution status of *Phlebotomus tobbi* (autumn 2012)



Data availability

Figures of all selected models and the contributory habitat masks are provided online. https://www.ecdc.europa.eu/en/all-topics-z/disease-vectors/prevention-and-control/vector-distribution-modelling.

The models for sandflies and mosquitos are available as data papers [10, 11] and provide access to all digital data and summary graphics.

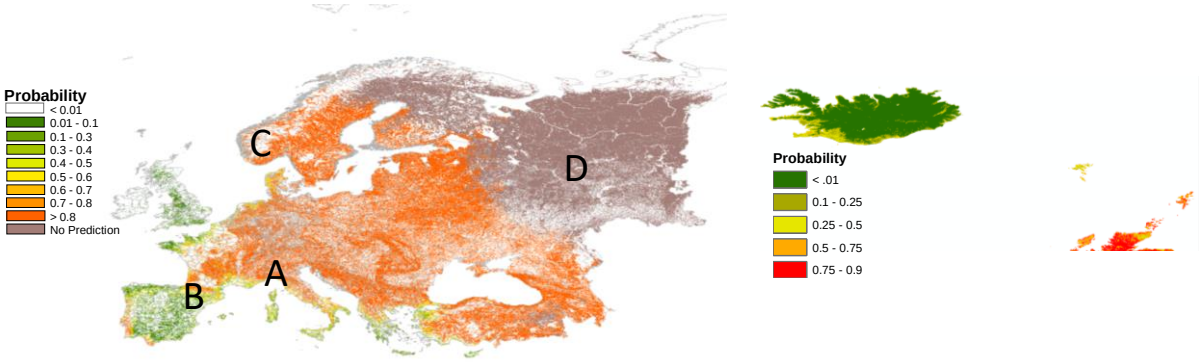
Using the modelled outputs

While the primary objective of producing gap-free and continental-scale vector maps was to 'fill the gaps' and show an index of disease risk for the whole project area, there are also a number of additional applications. Perhaps the most obvious is to target additional field work – both to validate the models and to gather more specific information about a vector. Examples, as illustrated in Figure 17 (left panel) for *Ae. vexans*, might include the following:

- Defining locations for epidemiological surveys at sites which are predicted to have a high probability of presence but for which no field data are yet available (A in Figure 17);
- Defining transects along predicted edges of distributions to confirm range limits, or identifying locations for abundance surveys, or monitoring spread or contraction (B in Figure 17);
- Defining locations for seasonal additional surveys, and for confirming absence (C in Figure 17, left panel);
- Identifying where additional field data are needed to produce a prediction. This is particularly relevant to species for which known records are restricted to a small part of the likely distribution so that the model algorithms cannot provide precise predictions for the whole project area because the covariate values in some parts are too different from those in the known areas to allow for reliable predictive relationships (D in Figure 17).

There are also likely to be occasions when a model prediction contradicts the observed data, most often when models predict presence in areas that the known distribution status data or the masks used suggest the vector to be absent (as discussed for *P. tobbi* above) or along the edges of spreading distributions that have not been surveyed for some time.

Figure 17. Using the models for *Aedes vexans* (left) and for *Ixodes ricinus* (right)



Another example: *I. ricinus* in Iceland illustrates detailed sample targeting. Field surveys in 2015 were required to establish the northern limits of the species distribution. A number of broad locations were identified, namely the Shetland, Orkney and Faroe Islands, northern Scandinavia and Iceland.

The gap analysis shown for Iceland, the Faroes and the Orkney Islands in Figure 17 (extracted from Figure 13; right panel) predicted relatively few areas of presence in Iceland, and, by contrast, relatively few areas of absence in the Islands. Sampling was guided by the relative probabilities, rather than absolute, with sampling focused on the highest of probability areas in each of the three locations. This approach proved successful; in Iceland, for example, the results of field sampling guided by these models proved accurate in ~75% of the locations [14].

4. Conclusions and potential implications

The work described here was designed to develop a strategy for enhancing the NUTS3 polygon level distribution maps of selected species taken from the three main vector groups that VBORNET efforts address: sandflies, ticks and mosquitoes. There are a number of data sets for other species, which have not yet been subjected to these gap filling procedures. In addition to applying these methods to the remaining species, there are a number of methodological steps that could be further developed, most notably:

- Additional refinements are needed to define the suitable habitats by investigating regional variations, incorporating a wider range of limiting threshold environmental or climatic values, and possibly assessing the merits of using one or more of the ecological classification data sets (e.g. Olson [12]). This process could be formalised within project recording systems, with a similar process of expert contribution and specialist validation. It could also take place as a regular contributory session at project meetings with network members.
- The only validation – beyond the purely statistical screening through accuracy metrics – is external evaluation by vector experts. This could be enhanced further by considering the model outputs in relation to the defined habitats and the reported distributions to find out if there are any systematic errors that could be rectified.
- In some cases, the modelling process was unable to produce statistically reliable predictions because the known presence (or absence) was too far removed, either in environmental or geographic terms. This phenomenon could be used to prioritise data collection efforts which would then provide more complete prediction surfaces for the model.
- The final outputs are provided only as NUTS-related values. Depending on the uses to which these maps are put, it may be desirable to provide additional types of maps, for instance maps combining NUTS3-level project data with pixel resolution modelled outputs. This may facilitate the use of the maps for higher resolution targeting of field surveillance to validate the maps or use them in response to some epidemiological event.
- Reliance on polygon-level data has, until recently, been a matter of pragmatic necessity: for the earlier models there were few, if any, point data available. This also implied that the sample values used to calibrate the models was based on chance. This is acceptable for absence records where the species is known to be absent from every point in a polygon. It is, however, less reliable for presence records as there will be unsuitable places (as illustrated by the habitat maps) where the species is absent. While this can be mitigated through the use of suitability masks, as applied here, it is clearly preferable to use geo-referenced data and, where possible, known locations.
- The number of covariates tested in the modelling process is rather high. While this may improve the likelihood of building good models, there is an argument to be made for using a reduced number of predictors to mitigate overfitting, which is always a danger for multivariate modelling. An alternative could be to introduce another modelling step once the full models have been selected, and produce a final product of a model where the covariates are limited to, for example, the top ten predictors. This may also mean that using the results to identify which covariates are driving the distributions is more difficult, as collinearity between predictors may affect the priority of covariates in the model equations. Therefore covariates listed in the top ten may be more a matter of chance than biology.

The earlier models were restricted to EU countries and relied on NUTS3 boundaries. Later analyses covered an extended area, for which NUTS3 units were not available. It was therefore necessary to use the UN Global Administrative Unit Layer (GAUL) boundaries. One of the drawbacks with the NUTS3-only data set was the fact that the unit size varied dramatically from country to country: Germany for example, has very small NUTS3 regions, while Scandinavia has very large ones. It is preferable to have equal-sized map units for the whole study area so that all the sampling, summarising and displaying procedures work with similar constraints. As a result, the current administrative unit data set is a combination of both levels, each chosen in an effort to standardise the NUTS3 unit area. The more standardised size of polygons used in the VectorNet project area displayed in Figure 1 is quite obvious if compared to the exclusive use of NUTS3 units in VBORNET.

References

1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, et al. The global distribution and burden of dengue. Nature. 2013 Apr 25;496(7446):504-7.

2. Kraemer MU, Sinka ME, Duda KA, Mylne AQ, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. Elife. 2015 Jun 30;4:e08347.

3. Estrada-Peña A, Alexander N, Wint GR. Perspectives on modelling the distribution of ticks for large areas: so far so good? Parasit Vectors. 2016 Mar 31;9:179.

4. European Environment Agency. *CORINE* land cover 2006 raster data [Internet]. Copenhagen: EEA; 2014 [accessed 9 Nov 2019]. Available from: http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3

5 Arino O, Ramos Perez JJ, Kalogirou V, Bontemps S, Defourny P, Van Bogaert E. ESA GlobCover Version 2.3 2009 300m resolution land cover map [Internet]. ESA and Université catholique de Louvain: Paris and Louvain; 2011 [accessed 11 Oct 2019]. Available from: http://www.edenextdata.com/?q=content/esa-globcover-version-23-2009-300m-resolution-land-cover-map-0

6. Kolonin GV. Fauna of *Ixodid* ticks of the world (*Acari, Ixodidae*). [No year, no place] [accessed May 2014; site was discontinued in 2014]. Formerly available from: http://www.kolonin.org.

7. Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW, et al. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. PLoS One. 2008 Jan 9;3(1):e1408.

8. Robinson TP, Wint GR, Conchedda G, Van Boeckel TP, Ercoli V, Elisa Palamara, et al. Mapping the global distribution of livestock. PLoS One. 2014 May 29;9(5):e96084.

9. Elith J, Graham, C, Anderson R, Dudík M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006 29(2):129-151

10. Schaffner F, Verstiert V, Van Bortel W, Zeller H, Wint W, Alexander NS. VBORNET gap analysis: Mosquito vector distribution models utilised to identify areas of potential species distribution in areas lacking records. Journal of Open Health Data. 2016. 4(1):e6. Available from: https://openhealthdata.metajnl.com/article/10.5334/ohd.27/.

11 Alten B, Verstiert V, Van Bortel W, Zeller H, Wint W, Alexander NS. VBORNET gap analysis: Sand fly vector distribution models utilised to identify areas of potential species distribution in areas lacking records. Journal of Open Health Data. 2016. 4(1):e5. Available from: https://openhealthdata.metajnl.com/article/10.5334/ohd.26/

12 Olson JS. Olson's world ecosystem complexes map (GeoLayer), 4 June 2015 [Internet]. Food and Agriculture Organization of the United Nations. Rome; 2015 [accessed 4 Nov 2019]. Available from: http://193.43.36.146/map?entryId=be34dd90-88fd-11da-a88f-000d939bc5d8&tab=about.

13 European Space Agency. Modelling module for the VECMAP system. Paris: ESA [no year]. Available from: https://www.avia-gis.com/vecmap and https://business.esa.int/sites/default/files/KruijffIAC-11-B5%201%2010%20-%20VECMAP%20-%20v2.2.pdf.

14 Alfredsson M, Olafsson E, Eydal M, Unnsteinsdottir ER, Hansford K, Wint W, et al. Surveillance of *Ixodes ricinus* ticks (Acari: Ixodidae) in Iceland. Parasit Vectors. 2017 Oct 10;10(1):466.

Annex 1. Habitat suitability data

Table 3. *CORINE* habitat preferences defined by experts for all species

CORINE label	*Aedes vexans*	*Culex modestus*	*Ixodes ricinus*	*Phlebotomus tobbi*	*Phlebotomus perniciosus*	*Phlebotomus ariasi*	*Phlebotomus papatasi*	*Anopheles plumbeus*	*Hyalomma marginatum*	*Dermacentor reticulatus*
Continuous urban fabric	3	3	3	3	3	3	1	3	3	3
Discontinuous urban fabric	3	3	3	3	3	3	1	3	3	3
Industrial or commercial units	3	3	3	3	3	3	1	3	3	3
Road and rail networks and associated land	3	3	3	3	3	3	2	3	3	3
Port areas	3	3	3	3	3	3	2	3	3	3
Airports	3	3	3	3	3	3	2	3	3	3
Mineral extraction sites	3	3	3	3	3	3	2	3	3	3
Dump sites	3	3	3	3	3	3	2	3	3	3
Construction sites	3	3	3	2	2	2	1	3	3	3
Green urban areas	2	3	3	2	2	2	1	1	3	2
Sport and leisure facilities	3	3	3	3	3	3	2	1	3	3
Non-irrigated arable land	3	3	3	2	2	2	1	3	2	3
Permanently irrigated land	3	2	3	2	2	2	1	3	3	3
Rice fields	2	1	3	2	3	3	2	3	3	3
Vineyards	3	3	3	2	2	2	2	3	1	3
Fruit trees and berry plantations	3	3	3	2	1	1	2	3	2	2
Olive groves	3	3	3	2	2	2	2	1	2	3
Pastures	2	3	2	2	3	3	1	3	2	2
Annual crops associated with permanent crops	3	3	3	2	2	2	1	3	2	3
Complex cultivation patterns	3	3	3	2	2	2	1	3	2	3
Agriculture with significant natural vegetation	2	2	2	2	2	2	1	1	2	2
Agro-forestry areas	2	2	3	2	2	2	2	1	3	2
Broad-leaved forest	2	3	1	2	1	1	3	1	3	1
Coniferous forest	3	3	1	2	2	2	2	3	2	1
Mixed forest	2	3	1	2	2	2	2	1	2	1
Natural grasslands	2	3	1	2	3	3	2	3	1	2
Moors and heathland	2	3	1	2	1	1	1	3	1	1
Sclerophyllous vegetation	3	3	3	1	1	1	2	3	1	3
Transitional woodland-shrub	3	3	1	2	2	2	2	1	1	1
Beaches, dunes, sands	3	3	3	3	3	3	3	3	3	3
Bare rocks	3	3	3	2	2	2	3	3	3	3
Sparsely vegetated areas	3	3	3	3	2	2	2	1	2	3
Burnt areas	3	3	3	3	3	3	3	3	3	3
Glaciers and perpetual snow	3	3	3	3	3	3	3	3	3	3
Inland marshes	1	1	3	3	3	3	3	3	3	3

CORINE label	*Aedes vexans*	*Culex modestus*	*Ixodes ricinus*	*Phlebotomus tobbi*	*Phlebotomus perniciosus*	*Phlebotomus ariasi*	*Phlebotomus papatasi*	*Anopheles plumbeus*	*Hyalomma marginatum*	*Dermacentor reticulatus*
Peat bogs	2	3	3	3	3	3	3	3	3	3
Salt marshes	3	3	3	3	3	3	3	3	3	3
Salines	3	3	3	3	3	3	3	3	3	3
Intertidal flats	3	3	3	3	3	3	3	3	3	3
Water courses	3	3	3	3	3	3	3	3	3	3
Water bodies	3	3	3	3	3	3	3	3	3	3
Coastal lagoons	3	3	3	3	3	3	3	3	3	3
Estuaries	1	2	3	3	3	3	3	3	3	3
Sea and ocean	3	3	3	3	3	3	3	3	3	3

Note: 1. Primary habitat = land classes providing most suitable habitat for a species and providing the likelihood of greatest vector numbers; 2. Secondary habitat = land classes where a species may still be found but less likely and in much lower numbers than above; 3. Unsuitable land = land classes where a species in unlikely to be found except in exceptional circumstances.

Table 4. GLOBCOVER habitat preferences defined by experts for all species

GLOBCOVER label	*Aedes vexans*	*Culex modestus*	*Ixodes ricinus*	*Phlebotomus tobbi*	*Phlebotomus perniciosus*	*Phlebotomus ariasi*	*Phlebotomus papatasi*	*Anopheles plumbeus*	*Hyalomma marginatum*	*Dermacentor reticulatus*
Post-flooding or irrigated croplands (or aquatic)	1	1	3	2	3	3	2	3	3	3
Rainfed croplands	3	3	3	2	3	3	2	3	3	3
Mosaic cropland (50–70%) / vegetation (grassland/shrubland/forest) (20–50%)	2	2	2	1	1	1	1	3	2	2
Mosaic vegetation (grassland/shrubland/forest) (50–70%) / cropland (20–50%)	2	2	1	2	1	1	1	1	1	1
Closed to open (>15%) broad-leaved evergreen or semi-deciduous forest (>5m)	2	3	1	2	2	2	2	1	3	1
Closed (>40%) broad-leaved deciduous forest (>5m)	2	3	1	2	1	1	2	1	3	1
Open (15–40%) broad-leaved deciduous forest/woodland (>5m)	3	3	1	2	2	2	2	1	2	1
Closed (>40%) needle-leaved evergreen forest (>5m)	3	3	2	2	2	2	2	3	2	2
Open (15–40%) needle-leaved deciduous or evergreen forest (>5m)	2	3	1	2	2	2	2	1	2	2
Closed to open (>15%) mixed broad-leaved and needle-leaved forest (>5m)	2	3	1	2	2	2	2	1	1	1

GLOBCOVER label	*Aedes vexans*	*Culex modestus*	*Ixodes ricinus*	*Phlebotomus tobbi*	*Phlebotomus perniciosus*	*Phlebotomus ariasi*	*Phlebotomus papatasi*	*Anopheles plumbeus*	*Hyalomma marginatum*	*Dermacentor reticulatus*
Mosaic forest or shrubland (50–70%)/grassland (20–50%)	2	2	1	2	2	2	2	3	1	1
Mosaic grassland (50–70%)/forest or shrubland (20–50%)	2	2	1	2	2	2	2	3	1	1
Closed to open (>15%) (broad-leaved or needle-leaved, evergreen or deciduous) shrubland (<5m)	3	3	1	3	2	2	2	3	1	1
Closed to open (>15%) herbaceous vegetation (grassland, savannahs or lichens/mosses)	2	3	2	3	1	1	2	3	3	3
Sparse (<15%) vegetation	3	3	3	3	2	2	2	3	2	3
Closed to open (>15%) broad-leaved forest regularly flooded (semi-permanently or temporarily) – Fresh or brackish water	1	2	3	3	3	3	3	3	3	2
Closed (>40%) broad-leaved forest or shrubland permanently flooded – Saline or brackish water	3	3	3	3	3	3	3	1	3	3
Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil – Fresh, brackish or saline water	1	1	3	3	3	3	3	3	3	3
Artificial surfaces and associated areas (Urban areas >50%)	3	3	3	2	2	2	1	1	3	3
Bare areas	3	3	3	3	3	3	3	3	3	3
Water bodies	3	2	3	3	3	3	3	3	3	3
Permanent snow and ice	3	3	3	3	3	3	3	3	3	3
No data (burnt areas, clouds,…)	3	3	3	3	3	3	3	3	3	3

Note: 1. Primary habitat = land classes providing most suitable habitat for a species and providing the likelihood of greatest vector numbers; 2. Secondary habitat = land classes where a species may still be found but less likely and in much lower numbers than above; 3. Unsuitable land = land classes where a species in unlikely to be found except in exceptional circumstances.

Annex 2. Environmental limiting factors

Table 5. Limiting factors applied to habitat suitability masks

Species	Environmental limiting factor
Phlebotomus ariasi	Altitude min <1700 m; BIOCLIM Tmax >15 °C <32 °C
Phlebotomus papatasi	Altitude min <2000 m; BIOCLIM Tmean >20 °C <30 °C
Anopheles plumbeus	Altitude mean <1200 m; precipitation >450 mm annual
Hyalomma marginatum	Altitude min <2000 m
Ixodes ricinus	Snow days <150; vegetation period >145 days

European Centre for Disease Prevention and Control (ECDC)

Gustav III:s Boulevard 40, 16973 Solna, Sweden

Tel. +46 858601000
Fax +46 858601001
www.ecdc.europa.eu

An agency of the European Union
www.europa.eu

Publications Office
of the European Union