VectorNet European Network for Medical and Veterinary Entomology

EXTERNAL SCIENTIFIC REPORT

ERGO Environmental Research Group Oxford | Francis Schaffner Consultancy Surveillance and management of biting insects | Institute of Tropical Medicine Antwerp

APPROVED: 15 April 2020

doi:10.2903/sp.efsa.2020.EN-1847

RVF vector spatial distribution models: vector abundance

William Wint[1], Dušan Petrić[2], Wim Van Bortel[3], Neil Alexander[1], Francis Schaffner[4]

1. Ergo – Environmental Research Group Oxford, Oxford, United Kingdom
2. University of Novi Sad, Faculty of Agriculture, Novi Sad, Republic of Serbia
3. Institute of Tropical Medicine, Antwerp, Belgium
4. Francis Schaffner Consultancy, Riehen, Switzerland

Abstract

EFSA has commissioned the VectorNet consortium to undertake a series of spatial distribution models for seven potential mosquito vectors of Rift Valley fever virus, namely *Aedes albopictus*, *Aedes caspius*, *Aedes detritus*, *Aedes japonicus*, *Aedes vexans*, *Culex pipiens* and *Culex theileri*. The modelling used the distribution data held within the VectorNet archive (as of September 2019), updated by literature searches to acquire new records available since 2016. The modelling has been implemented in three phases: (i) data collection, collation and standardisation; (ii) spatial modelling for presence and absence, and the calculation of presence metrics at the country level to be compatible with the MintRisk utilities; and (iii) the spatial modelling of vector abundance, dependent on the data available. This document presents the results of the abundance modelling due for delivery in March 2020. Sufficient data were amassed to produce statistically reliable spatial models for all species except *Ae. detritus*. Data for abundance models were extracted and the abundance values standardised for sampling effort where possible. Additional corrections were implemented that attempted to standardise for trapping methods and life stages, using expert opinion to define the conversion factors. The models were implemented at 1 km resolution covering the whole of continental Europe, using standard modelling techniques (Boosted Regression Trees and Random Forest) implemented through the VECMAP software suite. The models were evaluated according to expert opinion and the degree to which they matched the presence/absence models. Whilst all models produced were statistically reliable – so represented the input data effectively – not all were judged to reflect the field situation, implying that the input data were not sufficiently complete or extensive to feed continental scale distribution models. Training data and outputs for selected models are supplied in ESRI compatible format.

Key words: Rift Valley fever virus, probability of presence, spatial distribution model, mosquito vectors, Europe, abundance

Question number: EFSA-Q-2020-00285

Correspondence: ALPHA@efsa.europa.eu

Contributors: This work would not have been possible without the data provided to the VectorNet archives by a large number of contributors.

Acknowledgement and contributors: Firstly we thank all VectorNet (and formerly VBORNET) contributors, and in particular for the present work: Carles Aranda (Spain), Dominique Bicout, Gilles Besnard, Rémi Foussadier, Benoit Frances & Grégory L'Ambert (France), Jolyon Medlock & Alex Vaux (UK), Spiros Mourelatos (Greece), Lusine Paronyan (Armenia), Frantisek Rettich (Czech Republic), and Bouchra Trari (Morocco).

We also thank the VectorNet consortium and our counterparts with the VectorNet funding agencies for their support and input. The authors are also grateful to colleagues who provide invaluable inputs to the assessment of the results, to the report editing and to the strategic analysis decisions made during the course of the work. These include Miguel Miranda, Marieta Braks and Hein Sprong. We thank Renate Smallegange from Wageningen Academic Publishers for editing this report. We acknowledge Adwine Vanslembrouck and Jonas Torfs for the support in data extraction from the literature.

Suggested citation: Wint W., Petric D., Van Bortel W., Schaffner F. 2020 RVF vector spatial distribution models: Vector Abundance. EFSA supporting publication 2020:EN-1847. 37 pp. doi:10.2903/sp.efsa.2020.EN-1847

ISSN: 2397-8325

Summary

The VectorNet consortium received an *ad hoc* request for technical support from the European Food Safety Authority (EFSA) in September 2019. One of the activities required to contribute to the risk assessment is an estimation of the distribution of potential RVF arthropod vectors within each member state. The vectors specified are *Aedes albopictus*, *Aedes caspius*, *Aedes detritus*, *Aedes japonicus*, *Aedes vexans*, *Culex pipiens* and *Culex theileri*. This activity included spatial distribution modelling of each of the vectors.

This is the second phase of spatial distribution modelling which focuses only on vector abundance assessment, and follows an initial phase that focussed on presence/absence modelling. Sufficient data were available for all species except *Ae. detritus.*

Data for abundance models were extracted and the abundance values standardised for sampling effort where possible. Additional corrections were implemented that attempted to standardise for trapping methods and life stages, using expert opinion to define the conversion factors. The models were implemented at 1 km resolution covering the whole of continental Europe, using standard modelling techniques (Boosted Regression Trees and Random Forest) implemented through the VECMAP software suite.

The models were evaluated according to expert opinion by submitting all model outputs maps to at least four VectorNet Core Group experts familiar with the mosquito distributions, and also the degree to which they matched the presence/absence models. Whilst all models produced were statistically reliable – representing the input data effectively, not all were judged to reflect the field situation, implying that the input data were not sufficiently complete or extensive to feed continental scale distribution models. All training data and model outputs are supplied in ESRI compatible format.

The clear limitation is the heterogeneity of the training data – it is extremely heterogeneous both in terms of methods and seasonality of the sampling, and, largely speaking, does not cover the likely extent of the vector species' ranges.

This is not surprising, perhaps, given the fact that most of the data are derived from literature rather than targeted field sampling, and that acquiring data suited to quantitative abundance monitoring is a relatively recent focus of the VectorNet project. Whilst substantial field data has been collected over the past few years, they are mostly not yet sufficient in extent to provide a reliable continental perspective.

This does not yet mean the cause is lost. Data recording procedures for both field survey and literature extraction have been thoroughly revised, with much tighter and more comprehensive definitions of sample effort metrics that need to be recorded to feed abundance models.

The preparatory work done here may also be used to identify regions within Europe where additional data may most effectively contribute to the training dataset; to further elaborate on the most suitable data sampling methods needed to acquire useful pan-European abundance data sets; and to improve the extraction procedures from literature – as now encapsulated in the new VectorNet Search protocols.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

This contract/grant was awarded by EFSA to: VectorNet

Contractor/Beneficiary: VectorNet

Contract/Grant title: RVF vector spatial distribution models: Abundance

Contract/Grant number: SPECIFIC CONTRACT No 01/EFSA implementing framework contract NO ECDC/2019/020

The overall aims of this scientific report are to produce spatial models of the probability of presence and the abundance of a number of Rift Valley fever (RVF) arthropod vectors in Europe using spatial distribution modelling techniques.

RVF is a mosquito-borne viral disease affecting mainly ruminants. It causes abortion in pregnant susceptible ruminants and high mortality in new-borne animals. During epidemics, it can have a high impact on public health and the economy in the affected regions. The disease is caused by the RVF virus, a virus of the family Bunyaviridae and genus *Phlebovirus*. The virus has been isolated from more than 30 mosquito species. Mosquitoes belonging to the *Aedes* and *Culex* genera are considered to be the main vectors. The disease is widespread in Africa and it spread to the Arabian Peninsula in 2000-2001.

The VectorNet consortium received an *ad hoc* request for technical support from the European Food Safety Authority (EFSA) to quantify the risk of RVF virus to the EU in September 2019. One of the activities required to contribute to the risk assessment is an estimation of the distribution of the main potential RVF mosquito vectors within each member state. The vectors specified are *Aedes albopictus*, *Aedes caspius*, *Aedes detritus*, *Aedes japonicus*, *Aedes vexans*, *Culex pipiens* and *Culex theileri*. The first phase of distribution modelling focused on presence and absence assessment and was finalised in January 2020 (Wint et al., 2020). This second phase focuses on vector abundance.

2. Methodologies and Data

2.1. Methodologies

The techniques used to provide spatial distribution of vector abundance for main RVF arthropod vectors within each member state were based on the methodologies outlined in the recent European Centre for Disease Prevention and Control Technical Report entitled "A spatial modelling method for vector surveillance" (Alexander et al., 2019), associated data papers (Schaffner et al., 2016), and Boosted Regression Trees and/or Random Forest modelling (Breiman, 2001). These methods estimate statistical relationships between known georeferenced observations of presence, absence or abundance and a series of predictor covariates at a series of sample locations. These relationships are then used to calculate the predicted abundance (or probability of presence) for the whole area of interest at a fixed resolution, in this case 1 kilometre.

Spatial modelling requires the training data to be fairly evenly spread throughout the potential range of each target species. If the records are too clustered or too sparse then modelling (for presence and absence or as well as abundance) is not possible. Sufficient data were available for all species except *Ae. detritus*.

As indicated above, spatial modelling was implemented with Boosted Regression Trees and Random Forest Techniques, executed through the VECMAP software suite (AVIA-GIS, 2020). Models were only accepted with correlation coefficients between observed and predicted values at the training data locations equivalent to $P<0.01$. While both techniques are well established and widely used (Kraemer

et al., 2015; Nicolas et al., 2016), each produces slightly different outputs. As there is no basis for choosing one over the other, outputs for each method were ensembled to provide a mean consensus prediction. These consensus predictions were then masked using the probability of presence surfaces produced in the first phase of this modelling effort (Wint et al., 2020). The threshold of predicted probability used to define presence was set at 0.5 for all species, below which abundance values were masked out.

2.2. Covariate data

The covariates offered to the modelling procedures were drawn from a standardised set of agricultural, climatic and environmental parameters, and a suite of Fourier processed MODIS satellite imagery (Scharlemann et al., 2008) which provides a range of biologically interpretable variables related to levels and seasonality of temperature and vegetation during the period 2001-2015. These are available to registered members of the PALE-Blu Data Website (www.palebludata.com) and are summarised in Table 1. The predicted probability of presence for each species produced in phase one was also offered to each modelling attempt.

Table 1: Covariates offered to the modelling procedures.

Source	Description	Processing/Reference
MODIS time series 2001-2016	Land Surface Temperature (LST)	Fourier processed indicators produced (Hijmans et al., 2005; Scharlemann et al., 2008)
	Normalised Difference Vegetation Index (NDVI)	
	Middle Infra Red Channel 3	
	Enhanced Vegetation Index (EVI) series	
Worldclim (http://www.worldclim.com)	Annual Precipitation, Time series 2005-2007	
STRM (https://www2.jpl.nasa.gov/srtm/)	DEM Digital Elevation	
GRUMP (http://www.ciesin.org/)	Human population density derived from population layers produced	(Center for International Earth Science Information Network (CIESIN) et al., 2011)
Earthenv consensus land cover product (http://www.earthenv.org)	Landcover % for snow, water, urban extent, broadleaved and needle leaved evergreen forest, deciduous broadleaved forest, managed land, shrubland, herbaceous cover, other land cover, bare ground, flooded land, water	

2.3. Available vector data

The methods depend heavily on the availability of known observations (training data) which are used to calibrate the spatial models. For these RVF vectors, training data were available from a number of sources, namely the VectorNet archive, as of September 2019; a series of previous studies by Avia-GIS on RVF vector distribution produced for EFSA in 2013 (Ducheyne et al., 2013; Versteirt et al., 2013), and data collated from an extensive literature review for all specified species from 2016 to late 2019 The VectorNet archive contains data at both polygon and georeferenced point levels, and contains records of simple presence and absence, as well as records of abundance. Of these, only records that are georeferenced at the point level, and contain abundance records (including confirmed zero values) can, in principle, be used to train abundance models.

Simple abundance numbers are however not sufficiently precise for such modelling as they may represent cumulative numbers for extended sampling efforts using one or more traps, deployed for one or more days, over several sampling sessions. Unless, therefore, ancillary data are available that can be used to standardise for sampling effort – initially to provide numbers per trap per unit time (day) – either data must be discarded, or if they are to be retained, then it must be assumed that the values represent a single sampling event. A further complication arises from repeated sampling at a single location: most vector populations vary during the year, often rising to one or more seasonal peaks. Including abundance values from different times of the year is almost certain to mean that values are taken from different phases of the seasonal population cycle, which in turn means any difference in abundance between locations are as likely to reflect geographical as seasonal variation.

Abundance modelling requires therefore that the input data are standardised both in terms of sample effort and also of seasonal variation. This could be done by for example taking a mean value per location – but this would rely on the availability of samples for the whole season. An alternative, used successfully in abundance modelling of *Culicoides* vectors (Balenghien et al., in preparation), is to use the maximum recorded number per location. This measure is less dependent on regular longitudinal sample regimes, though may still be unreliable if the peak abundance season is not sampled. This approach can be further extended if samples are combined to produce spatially aggregated summaries at for example 1 or 10 km resolution.

A further level of complication is that even the abundance records with standard trap times and known trap deployment regimes were obtained by a wide range of active and passive trapping methods: ovitraps, larval dipping, baited and unbaited adult traps, human landing catches and so on. The variety of trap/catch types is shown in Table 2:.

Table 2: Number abundance records for each species with specified collection methods. BG = Biogents (DE) Sentinel trap, CAA = Centro Agricoltura Ambiente (IT) trap, CDC = Centers for Disease Control and Prevention (US) trap, EVS = Encephalitis Vector Survey trap, BioQuip (US), IMT = Insect Mosquito Trap, Gennico (IT), Mag = Mosquito Magnet trap® (US). Note: CDC and EVS generally baited with CO_2; BG can be baited with CO_2 or BG Lure only or a combination of both attractants.

Collection method	*Ae. albopictus*	*Ae. caspius*	*Ae. detritus*	*Ae. japonicus*	*Ae. vexans*	*Cx. pipiens*	*Cx. theileri*
BG	153			10			
BG baited	12	36			7		49
CAA	12						
CDC	5	3			6		
CDC baited	49		1		49	78	
EVS	49	23	1	2	24		
EVS baited	62	275	310	56	123	20	24
Magnet baited	10						
Unspecified baited	57	63	47	56	65	79	
Unspecified unbaited					29	20	
Citizen	1						
Aspirator	8		4	27			
Gravid	2						
Human landing	21	40	28	24	3	16	
IMT	6						
Larval sampling	18	82	99	913	9	797	156
Ovitrap	1,691			240		2	
Other / Unknown	62	627	657	13	6		9

All of these will produce different sample numbers from a given population of vectors, and are often heavily dependent on the weather conditions, trap siting, and other variables. The standardised data needed for modelling should therefore either use values from all trap types converted in some way to a single index measure, or use records from the different trap types in separate models. Note also that whilst there are a significant number of records for *Aedes detritus*, these were all focused in far too few locations to model.

A final constraint centres on the spatial distribution of the available data. The current objectives require the production of EU-wide models. Spatial modelling techniques are primarily designed to fill the gaps (interpolate) between known training data values, and to identify areas beyond the extent of the known data (extrapolate) that are similar to the areas for which training data are available.

Extrapolating distributions from known areas in France to totally different areas in e.g. Scandinavia is at best dangerous, and at worst completely inaccurate. Training data for continental models therefore needs to be continental in extent, or at least cover a large proportion of the likely range of vectors with regional rather than continental distributions. This condition is not fulfilled for the majority of the species modelled here.

The simple fact remains that only a small proportion of the data currently in the VectorNet archive was collected with the specific objective of providing the training data needed for reliable abundance modelling. Most of the data came from the published literature which often do not provide complete details of sampling regimes and tend to report abundance numbers in bulk (i.e. for the whole study).

This has left a stark set of alternatives: either waiting until sufficient standardised data are available at a continental level, or trying to standardise what data are available and see what models these data generate. This is the approach that has been taken here.

2.4. Vector data processing

Considerable time was spent in trial and error modelling of the raw recorded values, and of categorical classes (e.g. high, medium, low) derived from them. These both completely failed to produce statistically reliable model outputs. As set out above, the available data need to be standardised in order to provide usable inputs for modelling.

Table 3: Number equivalent and conversion factors for different collection methods and species. BG=Biogents Sentinel, CDC=Centers for Disease Control, EVS= Encephalitis Vector Survey, red indicates standardisation base.

	Ovitraps	Larval dips	BG traps		CDC traps		EVS traps		Human landing
			unbaited	baited	Unbaited	baited	unbaited	baited	
	Number Equivalents								
Ae. albopictus	100		3	8	1	3	1	3	10
Ae. caspius				10	5	100	1	40	50
Ae. japonicus	15		1	2			1	2	3
Ae. vexans			5	10	5	100	1	40	50
Cx. pipiens		5	5	10		50		20	6
Cx theileri		5		10				20	
	Conversion Factors								
Ae. albopictus	1		33.3	12.75	100	33.3	100	33.3	10
Ae. caspius				4	8	0.4	40	1	0.8
Ae. japonicus			2	1			2	1	0.67
Ae. vexans			8	4	8	0.4	40	1	0.8
Cx. pipiens		4	4	2		0.4		1	3.3
Cx theileri		4		2				1	

Three levels of standardisation were attempted as follows and set out in Table 3 above:

1) Correcting the raw abundance numbers to numbers per sample (per trap per day). If no sample effort data were provided, the sample number was set to 1.

Standardising the recorded values for trap type. This process was based on conversion factors between one trap type and another. The conversion factors thus rely on the relative efficiency of each trap method as well as natural attrition between life stages. These factors were defined by expert opinion in consultation with at least four mosquito biologists from the VectorNet Core Group and from AIMCOST project members. They were based on instar mortality AND differences in trapping efficiency between life stages. No attempt was made to adjust these to regional or

habitat differences. This attempted to account for the fact that sample efforts of some life stages – particularly larvae – tend to be much less well quantified than others. For all species except *Ae. albopictus*, the raw numbers were standardised to Encephalitis Vector Survey (EVS) baited catch numbers as these were overall the most recorded collection methods (

2) Table 2:). By the same token *Ae. albopictus* numbers were standardised to ovitrap catch numbers as that was the predominant trap type for this species. The values used are shown in Table 3.

2.5. Models run

Models were run for three sets of corrected numbers for each species as follows:

1) Simple numbers per sample (logged values without applying the conversion factors).
2) Log numbers per sample corrected for trap type and life stage using the conversion factors in Table 3. This interim set models were initially run with all records with attributable trap types, corrected to ovitrap numbers. Using the same conversion logic as set out in Table 3 for baited EVS catch numbers.
3) After some deliberation it was decided that sampling effort recorded for larval dipping was too unreliable to merit inclusion, and despite the fact that larval sampling made up a significant proportion of records for several of the species, a third set of models excluding larval samples was implemented (for all species except *Ae. albopictus* which had few larval samples), standardised to ovitrap catch numbers (*Ae. albopictus*) or baited EVS or catch numbers (all other species).

In an attempt to further standardise for repeated seasonal samples, and to base the modelling on a standard annual metric, the three metrics listed above for each point location were aggregated to the maximum value recorded for each 1 and/or 10 kilometre grid. This has the effect of combining all records not only for a single location, but for all location within the aggregation grids. As repeated samples from different years are therefore combined and the maximum value calculated, this removes zeros from leach locations where there are positive values, and also makes some limited attempts to identify the seasonal peak values. Aggregation increases the chances that a particular maximum value actually represents a seasonal peak.

Finally it should be noted that *Ae. albopictus* and *Ae. japonicus* are invasive species, and so the models produced will reflect the current state of the invasions, as discussed further in the section of selected models below

3. Results

3.1. Models produced

The accuracy metrics of the 5 sets of models runs for each species are set out in **Error! Not a valid bookmark self-reference.**, below. VECMAP provides various such metrics by default: for BRT the correlation coefficient of the training data with the predicted values ("BRT R"), and the correlation coefficient of a 10 times cross validation using the 25% holdback data against model predictions ("BRT XVALR"). For RF models it provides the Pseudo R correlation coefficient derived for each tree used in the model process (RF PseudoR). Given the number of sample points in each model, a guideline value for these metrics to be significant at the 1% level is around 0.15. A value of 0.5 infers significance of at least 0.001%, i.e. a very reliable model, in which the training data are reflected very well.

Error! Not a valid bookmark self-reference. therefore suggests that all bar two models (in italics, on the bottom row) are highly significant in statistical terms. This implies that the training data are well modelled, at least in the geographical areas and wherever the numerical ranges of the predictor variables are similar to those where there are training data. This further suggests that the interpolated values are likely to be statistically reliable. It does not, however necessarily mean that the extrapolated predictions are accurate as this relies on the fact that not only are the relationships of the abundance values with predictor covariates for the training data locations also apply in the extrapolated areas, but also that the interactions with multiple predictors remain relatively similar.

Table 4: Number equivalent and conversion factors for different trap types and species.

Model run	Accuracy metric	*Ae. albopictus*	*Ae. caspius*	*Ae. japonicus*	*Ae. vexans*	*Cx. pipiens*	*Cx. theileri*
Trap Corrected no/sample, no larvae, 10 km	BRT R	na	0.92	0.95	0.99	0.99	0.99
	BRT XVALR	na	0.66	0.79	0.53	0.77	0.85
	RF PseudoR	na	0.35	0.59	0.19	0.52	0.64
Trap corrected no/sample, no larvae, 1 km	BRT R	na	0.93	0.97	0.96	0.97	0.99
	BRT XVALR	na	0.73	0.82	0.6	0.76	0.84
	RF PseudoR	na	0.51	0.6	0.28	0.57	0.63
Trap corrected no/sample, with larvae, 10 km	BRT R	0.99	0.99	0.95	0.99	0.98	0.99
	BRT XVALR	0.76	0.72	0.7	0.58	0.78	0.8
	RF PseudoR	0.48	0.44	0.35	0.29	0.42	0.65
Trap corrected no/sample, with larvae, 1 km	BRT R	0.97	0.96	0.96	0.99	0.97	0.97
	BRT XVALR	0.79	0.78	0.71	0.62	0.79	0.87
	RF PseudoR	0.4	0.54	0.3	0.4	0.51	0.66
Number per sample, 10 km	BRT R	0.98	0.83	0.91	0.95	0.78	0.88
	BRT XVALR	0.73	0.41	0.66	0.32	0.54	0.52
	RF PseudoR	0.42	*0.11*	0.45	*0.081*	0.21	0.17

3.2. Models selected

Whatever the statistical arguments, it is clear that the majority of the models for all metrics accurately reflect the input data. Using accuracy measures to choose between them is therefore unlikely to be effective, and some form of expert assessment was needed. All models were therefore presented for assessment to at least four VectorNet Core group experts and several AIMCOST project members with specific knowledge of mosquito distributions. Experts were asked separately to judge models (e.g. maps 1 km and/or 10 km aggregation, numbers per sample or standardised) by requesting their opinion about: over or under-estimation of abundance in general, over-representation or under-representation in some key regions in Europe (i.e. wetlands), and usefulness for risk assessment based on the representation of abundance among Member States, considering both native and invasive species.

Maps of the abundance records, habitat suitability, and surfaces for the predicted probability of presence were also provided, on the assumption that abundance models that produced similar distribution patterns to the presence/absence models were more likely to be realistic. Model selection was based on which the majority of experts judged to be the best, which were consistently the models that excluded the records from larval sampling and were corrected for collection method. Whilst this consistency reduces the uncertainty as to which models are considered best, it should be understood that these outputs represent an abundance index rather than a true measure of abundance.

For *Ae. albopictus* and *Cx. pipiens* the selected models were those aggregated at 1 km, corrected to egg numbers and Baited EVS catch numbers per sample, respectively. For all the other species, the

models aggregated at 10 km corrected to baited EVS catch number per sample were selected. All model values are log10 transformations of the raw values. The list of the top ten predictor covariates for each selected model are given in Appendix B. Figures of all training data as well as unmasked and masked distribution models are provided in Appendix C.

As discussed above all selected models are statistically highly significant (Table 4) and are good representation of the training data. This does not mean, however, that they are entirely accurate representations of the actual situation on the ground, and the consulted experts had a number of comments for each of the models. For details of the presence/absence masks used, reference should be made to the companion report on probability of presence modelling (Wint et al., 2020).

Aedes albopictus. Figure 1 to .

Figure 3. Whilst the predicted probability of presence and abundance match reasonably well in Mediterranean and western Europe, the abundance model clearly underestimates some areas such as the Po valley in Italy. Generally speaking predicted abundances are quite low. Predicted abundance is also significant in northern Europe (Scandinavia, UK and even Iceland), which is clearly erroneous, but is removed by the probability of presence masking.

Aedes caspius. Figure 6 to Figure 6. The high recorded densities in the Carmargue and southern Spain are captured effectively. In most areas the masked outputs are considered reasonably feasible, though with some notable exceptions such as the Danube Delta and the north eastern margins of the Caspian Sea. The unmasked predicted abundances are however significant throughout Russia, Sweden and north Africa, in areas that are certainly unsuitable for this vector (not shown on the output map).

Aedes japonicus. Figure 7 to Figure 9 The predicted probability of presence and predicted abundances match quite well. The predicted abundance values are relatively low (0-2 as opposed to 0-3 for other species). This is one of the species with a lot of larval samples that were discarded, but these discarded records spatially overlapped the others. This perhaps suggests that the training data for this species is relatively complete – i.e. it covers a good proportion of the areas where the vector is currently present in Central Europe, France and Northern Spain.

Bearing in mind that both *Ae. albopictus* and *Ae. japonicus* are invasive species it should be emphasised that these predicted distributions reflect the current situation, and so identifies areas that are environmentally similar to the current extent of the invasion. The models will need to be re-run if the species spreads into environmentally dissimilar regions. However, the fact that there are relatively few areas predicted outside the current distributions may suggest that these vectors will need to be able to adapt to different environmental conditions if it is to move beyond its current range.

Aedes vexans. Figure 10 to Figure 12. The training data were within a fairly wide band running from Belgium and South east France to Armenia. No training data were available for most of France or Spain, despite the fact that the presence/absence data show this vector to be widespread in France and northern Spain, if perhaps limited by habitat suitability (see companion report on presence and absence models). The unmasked abundance models do identify regions with high predicted numbers in south central France, and eastern Spain, but these are masked out by the presence mask. The mask also removes predicted abundance in North Africa and Scandinavia. Expert opinion was divided as to whether the mask used was too strict as numbers in Scandinavia may, in reality, be high.

Culex pipiens. Figure 13 to Figure 15. For this species high abundance levels are predicted throughout Russia and its neighbours, the output presented here is thus heavily constrained by the presence mask. This is one of the species for which much of the available abundance data were obtained from larval sampling and so discarded. Unlike *Ae. japonicus*, for which a lot of larval samples that were also discarded, though mainly from locations that overlapped other samples, many of the rejected records for *Cx. pipiens* were from areas like the Belgium, UK and Morocco for which data from other sampling techniques were not available.

Culex theileri. Figure 16 to Figure 18. This is another species for which the predicted abundance is modified significantly by the probability of presence mask – which 'removes' widespread low

abundance predictions in northern Europe, and some seemingly anomalous high densities in central Spain and North Africa. As the training data are very patchy and dispersed, and many were lost to the standardisation process this is no great surprise. The false positives in Scandinavia are so far beyond what could be reasonably expected to be extrapolated from Mediterranean data points that they can comfortably be discounted. The recorded 'hotspots' in southern Spain and Armenia are, however fairly well captured.

4. Data provided

All geographic data – predicted surfaces and updated point and polygon data surfaces are provided in an ArcMap 10.4 compatible format 'package' *vnrvfabundancemodelsmarch20.mpk* which is available for download at the following link: https://drive.google.com/drive/folders/13mk5-SrZS6LAUOKAMxFrORSs5Khe4R-F?usp=sharing

This dataset also includes a surface for the distribution of vector hosts, provided for a previous EFSA study, to allow for the calculation of the overlap between hosts and predicted vector presence. The package includes an ArcMap document (*vnrvfabundancemodelsmarch20*.mxd) file which displays all layers with an explanatory legend and simplified file label. Actual filenames to which these file labels refer are given in the file list in Appendix A, and can be accessed through the source tab in the layer properties dialogue box.

5. Conclusions

Considerable and somewhat labyrinthine efforts have been made to generate standardised abundance metrics that can drive reliable spatial distribution models. Whilst statistically largely effective, the models produced do not meet with universal approval for every species from the VectorNet Core Group experts who were tasked with assessing more than 100 model outputs produced during the iterative development process.

Nevertheless, there is a consensus that some progress has been made in producing abundance models for some of the species, that apply for most or all of the likely vector ranges within Europe and its neighbours: *Ae. albopictus*, *Ae. japonicus, Ae. caspius* and *Ae. vexans.* Most of these (*Ae. albopictus*, *Ae. caspius* and *Ae. vexans)* have some obvious faults, however though they may be more feasible for some regions. The models for the *Culex* species are less convincing - though again sub regional portions of the models may be more reliable.

The clear limitation is the training data – it is extremely heterogeneous both in terms of methods and seasonality of the sampling, and, largely speaking, does not cover the likely extent of the vector species' ranges.

This is not surprising, perhaps, given the fact that most of the data are derived from literature rather than targeted field sampling, and that acquiring data suited to quantitative abundance monitoring has been a focus of the VectorNet project for only the last 4 years. Whilst substantial field data has been collected over this period, they are not yet sufficient to provide a reliable continental perspective.

This does not mean the cause is lost. Data recording procedures for both field survey and literature extraction within the VectorNet project have been thoroughly revised, with much tighter and more comprehensive definitions of sample effort metrics that need to be recorded to seed abundance models. It would be very helpful if authors of peer reviewed papers could be sensitised to the advantages of contributing to continental scale dataset, and encourage to present more complete details of their data. There are a number of EU wide initiatives which are currently advocating a minimum standard for field sampling and data reporting, which, if adopted, will help considerably.

The preparatory work done here may also be used to identify regions within Europe where additional data may most effectively improve the training data, and to further elaborate on the most suitable data sampling methods needed to acquire useful pan-European abundance data sets. In addition, the

concept of conversions to standardise for trap types and life stages needs to be much better substantiated.

6. References

Alexander N, Van Bortel W, Hendrickx G, Versteirt V, Ducheyne E and Wint W, 2019. A spatial modelling method for vector surveillance. European Centre for Disease Prevention and Control, Stockholm, Sweden. 23 pp.,doi: 10.2900/633757 Available online: https://www.ecdc.europa.eu/sites/default/files/documents/spatial-modelling-method-vector-surveillance.pdf

AVIA-GIS, 2020. VECMAP: the one-stop-shop for data collection and risk mapping. AVIA-GIS, Zoersel, Belgium.

Breiman L, 2001. Random Forests. Machine Learning, 45, 5-32. doi:10.1023/A:1010933404324

Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI), The World Bank and Centro Internacional de Agricultura Tropical (CIAT), 2011. Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Count Grid. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY, USA.

Ducheyne E, Versteirt V and Hendrickx G, 2013. Abundance of Rift Valley fever vectors in Europe and the Mediterranean Basin. EFSA Supporting Publications, 10, EN-420. doi:10.2903/sp.efsa.2013.EN-420

Hijmans RJ, Cameron SE, Parra JL, Jones PG and Jarvis A, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25, 1965-1978. doi:10.1002/joc.1276

Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, Moore CG, Carvalho RG, Coelho GE, Van Bortel W, Hendrickx G, Schaffner F, Elyazar IRF, Teng H-J, Brady OJ, Messina JP, Pigott DM, Scott TW, Smith DL, Wint GRW, Golding N and Hay SI, 2015. The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. elife, 4, e08347. doi:10.7554/eLife.08347

Nicolas G, Robinson TP, Wint GRW, Conchedda G, Cinardi G and Gilbert M, 2016. Using Random Forest to improve the downscaling of global livestock census data. PLoS One, 11, e0150424. doi: https://doi.org/10.1371/journal.pone.0150424

Schaffner F, Versteirt V, Van Bortel W, Zeller H, Wint W and Alexander N, 2016. VBORNET gap analysis: mosquito vector distribution models utilised to identify areas of potential species distribution in areas lacking records. Open Health Data, 4, e6. doi:10.5334/ohd.27

Scharlemann JPW, Benz D, Hay SI, Purse BV, Tatem AJ, Wint GRW and Rogers DJ, 2008. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. PLoS One, 3, e1408. doi:10.1371/journal.pone.0001408

Versteirt V, Ducheyne E, Schaffner F and Hendrickx G, 2013. Systematic literature review on the geographic distribution of Rift Valley fever vectors in Europe and the neighbouring countries of the Mediterranean Basin. EFSA Supporting Publications, 10, 412E. doi:10.2903/sp.efsa.2013.EN-412

Wint W, Van Bortel W and Schaffner F, 2020. RVF vector spatial distribution models: probability of presence. EFSA Supporting Publications, 17, 1800E. doi:10.2903/sp.efsa.2020.EN-1800

Appendix A - Spatial data file list

All GIS data are provided as an ARCMAP 10.4.1 'package' *vnrvfabundancemodelsmarch20. mpk*, visualised using the ARC Document file supplied (*vnrvfabundancemodelsmarch20.mpk*). The package can be downloaded from the following link https://drive.google.com/drive/folders/13mk5-SrZS6LAUOKAMxFrORSs5Khe4R-F?usp=sharing

Species	Value	File
Ae. albopictus	Training data, corrected to log10(egg numbers per sample)	albomxper1xyoveqjan20v2data.shp
	Predicted abundance masked with predicted presence. 1 km aggregation	alb1xylnoepsmeanmaskedPA.tif
	Predicted abundance, unmasked 1 km aggregation	albo1xyymxlnoepsPSRFBRTMEANJan20ok.tif
Ae. caspius	Training data, corrected to log10(baited EVS catch per sample)	casmxper10xyevbjan20data.shp
	Predicted abundance masked with predicted presence. 10 km aggregation	cas10xylnevsbpsmeanmaskedPA.tif
	Predicted abundance, unmasked 10 km aggregation	cas10xylnevbpsBRTRFMEANjan20ok.tif
Ae. japonicus	Training data, corrected to log10(baited EVS catch per sample)	japmxper10xyevbjan20data.shp
	Predicted abundance masked with predicted presence. 1 km aggregation	jap10xylnevsbpsmeanmaskedPA.tif
	Predicted abundance, unmasked 1 km aggregation	jap10xylnevbpsBRTRFMEANjan20ok.tif
Cx. pipiens	Training data, corrected to log10(baited EVS catch per sample)	pipmxper1xyevbjan20data.shp
	Predicted abundance masked with predicted presence. 10 km aggregation	pip1xylnevsbpsmeanmaskedPA.tif
	Predicted abundance, unmasked 10 km aggregation	pip1xylnevbpsbrtRFMEANjan20ok.tif
Cx. theileri	Training data, corrected to log10(baited EVS catch per sample)	themxper10xyevbjan20data.shp
	Predicted abundance masked with predicted presence. 10 km aggregation	the10xylnevsbpsmeanmaskedPA.tif
	Predicted abundance, unmasked 10 km aggregation	the10xylnevbpsBRTRFMEANjan20ok.tif
Cx. theileri	Training data, corrected to log10(baited EVS catch per sample)	vexmxper10xyevbjan20data
	Predicted abundance masked with predicted presence. 10 km aggregation	vex10xylnevsbpsmeanmaskedPA.tif
	Predicted abundance, unmasked 10 km aggregation	vex10xylnevbpsBRTRFMEANjan20ok.tif
Shape file columns	lat10k or lat1k	1 km or 10 km grid centre Latitude
	lon10k or lat1k	1 km or 10 km grid centre Longitude
	Numpersamp	uncorrected number per sample
	mxlnoeps (albopictus)	maximum value per grid corrected to log10 (egg numbers per sample)
	lnmxevsb (other species)	maximum value per grid corrected to log10 (baited EVS catch numbers)

Appendix B - Top ten predictor covariates, selected models

Aedes albopictus (ovitrap catch number)				*Aedes caspius (baited EVS catch number)*			
BRT variable and metric		**RF variable and metric**		**BRT variable and metric**		**RF variable and metric**	
Mean Rainfall, 2005/2007	13.96	Prob Presence	8.08	Elevation	29.58	Elevation	27.43
Daytime LST amplitude 2	10.47	Rainfall 05/07 Minimum	6.73	Rainfall 05/07 Minimum	23.20	Rainfall 05/07 Minimum	21.45
Rainfall 05/07 Amplitude 1	5.30	Mean Rainfall, 2005/2007	5.85	Nighttime LST phase 2	3.03	Prob Presence	7.59
Prob Presence	4.93	Nighttime LST phase 2	3.47	NDVI mean	3.02	Rainfall 05/07 Phase 1	6.25
Daytime LST phase 2	4.47	GRUMP Population density	3.19	Nighttime LST phase 1	2.61	Nighttime LST phase 1	5.26
Rainfall 05/07 Minimum	4.24	Daytime LST phase 2	3.03	Daytime LST amplitude 2	2.53	NDVI mean	4.83
Nighttime LST minimum	3.95	Nighttime LST phase 3	2.23	Rainfall 05/07 Phase 1	2.36	Daytime LST maximum	4.38
Rainfall 05/07 Phase 1	3.76	Rainfall 05/07 Phase 1	2.08	Middle infra-red phase 2	1.98	Mean Rainfall, 2005/2007	3.23
Nighttime LST phase 2	3.54	NDVI phase 1	2.02	Rainfall 05/07 % Var 1	1.66	Nighttime LST minimum	3.22
Nighttime LST maximum	3.54	Rainfall05/07 Maximum	2.01	NDVI minimum	1.64	NDVI maximum	3.19
Aedes japonicus (baited EVS catch number)				*Aedes vexans (baited EVS catch number)*			
BRT variable and metric		**RF variable and metric**		**BRT variable and metric**		**RF variable and metric**	
Prob Presence	46.76	Prob Presence	8.25	Nighttime LST phase 1	5.79	Middle infra-red amplitude	7.63
Rainfall 05/07 Minimum	8.09	Rainfall 05/07 Minimum	2.31	Middle infra-red phase 3	4.69	Nighttime LST phase 1	7.42
Rainfall 05/07 Amplitude 2	6.11	Nighttime LST amplitude 1	1.50	Nighttime LST maximum	4.17	EVI % var. annual cycle	5.83
Nighttime LST amplitude 1	5.96	NDVI phase 1	1.27	EVI % var. annual cycle	3.72	Rainfall 05/07 Maximum	5.29
Rainfal 05/07 Amplitude 1	2.55	Rainfall 05/07 Amplitude 2	0.92	Daytime LST phase 1	3.45	Middle infra-red phase 1	5.10
GRUMP Population density	2.53	GRUMP Population density	0.90	Rainfall 05/07 Amplitude	2.99	Middle infra-red phase 3	4.88
Nighttime LST phase 3	1.67	Nighttime LST phase 3	0.85	Daytime LST amplitude 1	2.93	Middle infra-red phase 2	4.85
Mean Rainfall, 2005/2007	1.56	Daytime LST maximum	0.76	Daytime LST mean	2.90	Elevation	4.28
Rainfall 05/07 Maximum	1.12	Rainfall 05/07 Amplitude 1	0.70	Middle infra-red amplitu	2.74	Nighttime LST maximum	4.14
Rainfall 05/07 % Var 1	0.99	Mean Rainfall, 2005/2007	0.55	Nighttime LST amplitude	2.74	Daytime LST mean	4.08
Culex pipiens (baited EVS catch number)				*Culex theileri (baited EVS catch number)*			
BRT variable and metric		**RF variable and metric**		**BRT variable and metric**		**RF variable and metric**	
Rainfall 05/07 Maximum	16.91	Rainfall 05/07 % Var 1	0.86	Rainfall 05/07 Phase 1	17.05	Elevation	18.22
Prob Presence	13.20	Rainfall 05/07 Amplitude 2	0.91	Elevation	14.82	Daytime LST % var. annual	12.32
Rainfall 05/07 Phase 1	9.62	Rainfall 05/07 Amplitude 1	0.98	Daytime LST % var. annua	11.43	Rainfall 05/07 Phase 1	9.97
Mean Rainfall, 2005/2007	5.28	Nighttime LST amplitude 3	1.05	Rainfall 05/07 % Var 2	9.71	Rainfall 05/07 Amplitude 2	4.91
Rainfall 05/07 Amplitude 3	5.24	Mean Rainfall, 2005/2007	1.19	Daytime LST minimum	6.78	Daytime LST phase 2	4.84
GRUMP Population density	5.10	Daytime LST % var. annual c	1.19	Rainfall 05/07 Amplitude	6.64	Rainfall 05/07 % Var 2	3.65
NDVI phase 1	4.93	GRUMP Population density	1.32	Rainfall 05/07 % Var 1	5.26	Daytime LST minimum	3.03
Rainfall 05/07 Minimum	3.24	Nighttime LST minimum	1.34	Daytime LST maximum	2.92	Nighttime LST phase 1	2.52
Rainfall 05/07 % Var 1	3.05	Nighttime LST phase 1	1.40	GRUMP Population densi	1.81	Daytime LST amplitude 1	2.30
Rainfal 05/07 Amplitude 1	2.82	Nighttime LST phase 3	1.59	Rainfall 05/07 Minimum	1.76	Rainfall 05/07 Minimum	1.77

See Table 1 for more information about the covariates.

Appendix C - Maps of point data and predicted values for selected models

This appendix contains figures of the recorded training data and the predicted abundance values selected for each species by expert opinion. The models are presented with and without masks of the probability of presence produced during the first phase of this work. The masked maps display the masks in a light green to identify them. Clicking on the title will move the page view to the relevant figure.

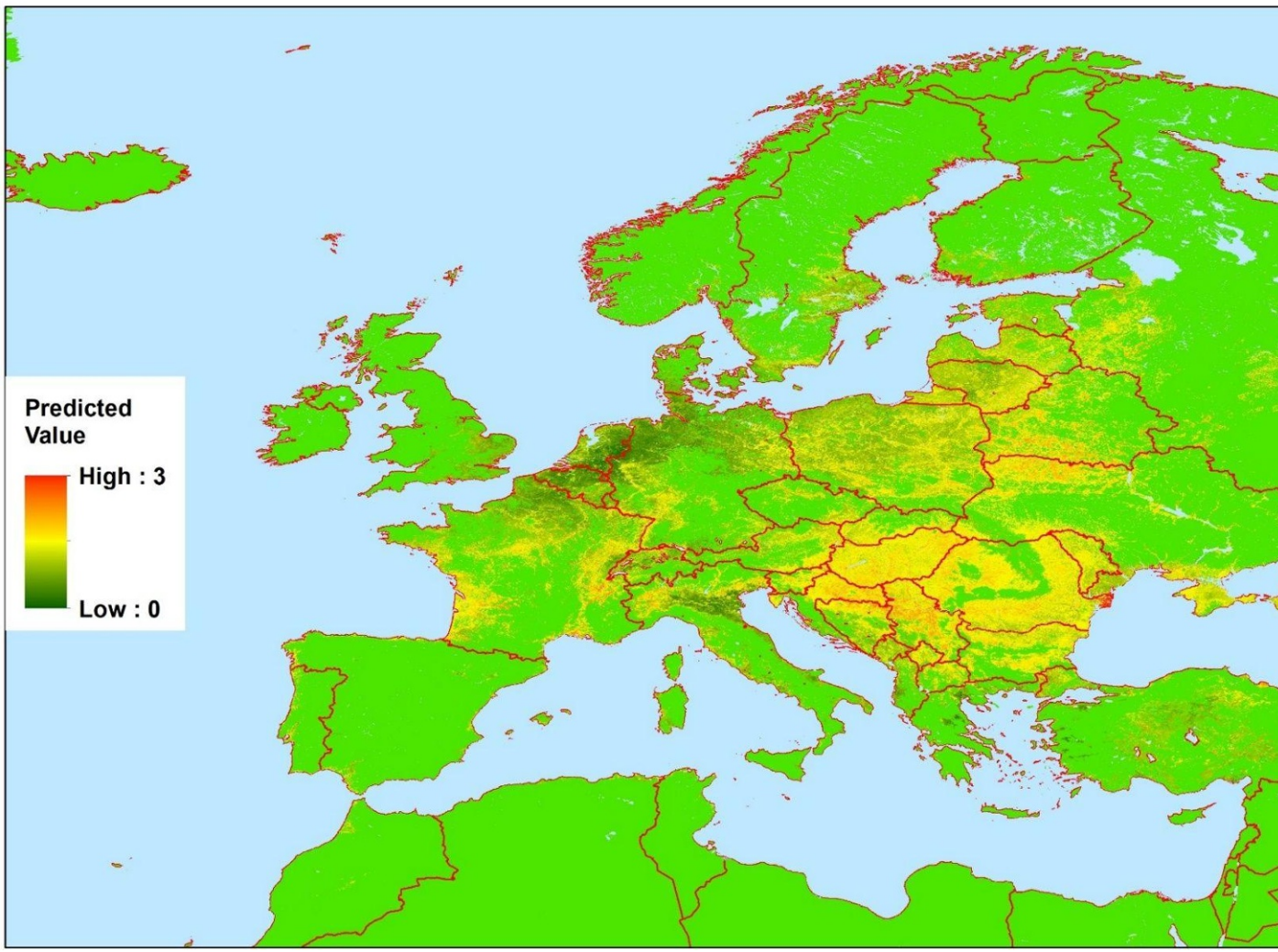

Figure 12: Predicted abundance (log EVS baited equivalent), masked, 10 km *aggregation for Aedes* vexans.

Figure 13: Available point abundance data (log EVS baited equivalent), 1 km aggregation for *Culex pipiens.*

Figure 14: Predicted abundance (log EVS baited equivalent), unmasked, 1 km aggregation, *for Culex pipiens.*

Figure 15: Predicted abundance (log EVS baited equivalent), masked, 1 km aggregation, *for Culex pipiens.*

Figure 16: Available point abundance data (log EVS baited equivalent), 10 km aggregation for *Culex theileri.*

Figure 17: Predicted abundance (log EVS baited equivalent), unmasked, 10 km aggregation, *for Culex theileri.*

Figure 18: Predicted abundance (log EVS baited equivalent), masked, 10 km aggregation, *for Culex theileri.*

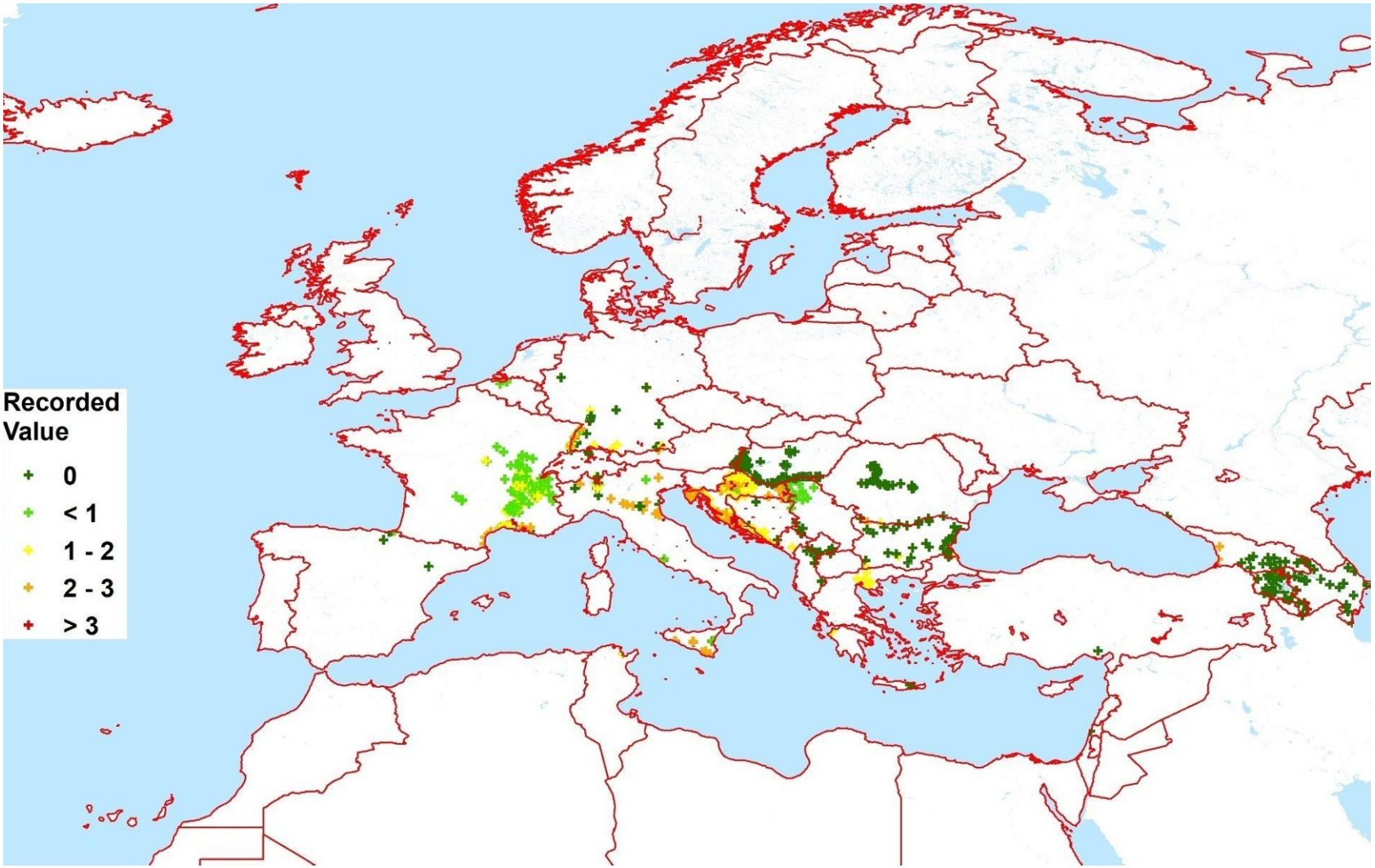

Figure 1: Available point abundance data (log egg equivalent), 1 km aggregation for *Aedes albopictus*.

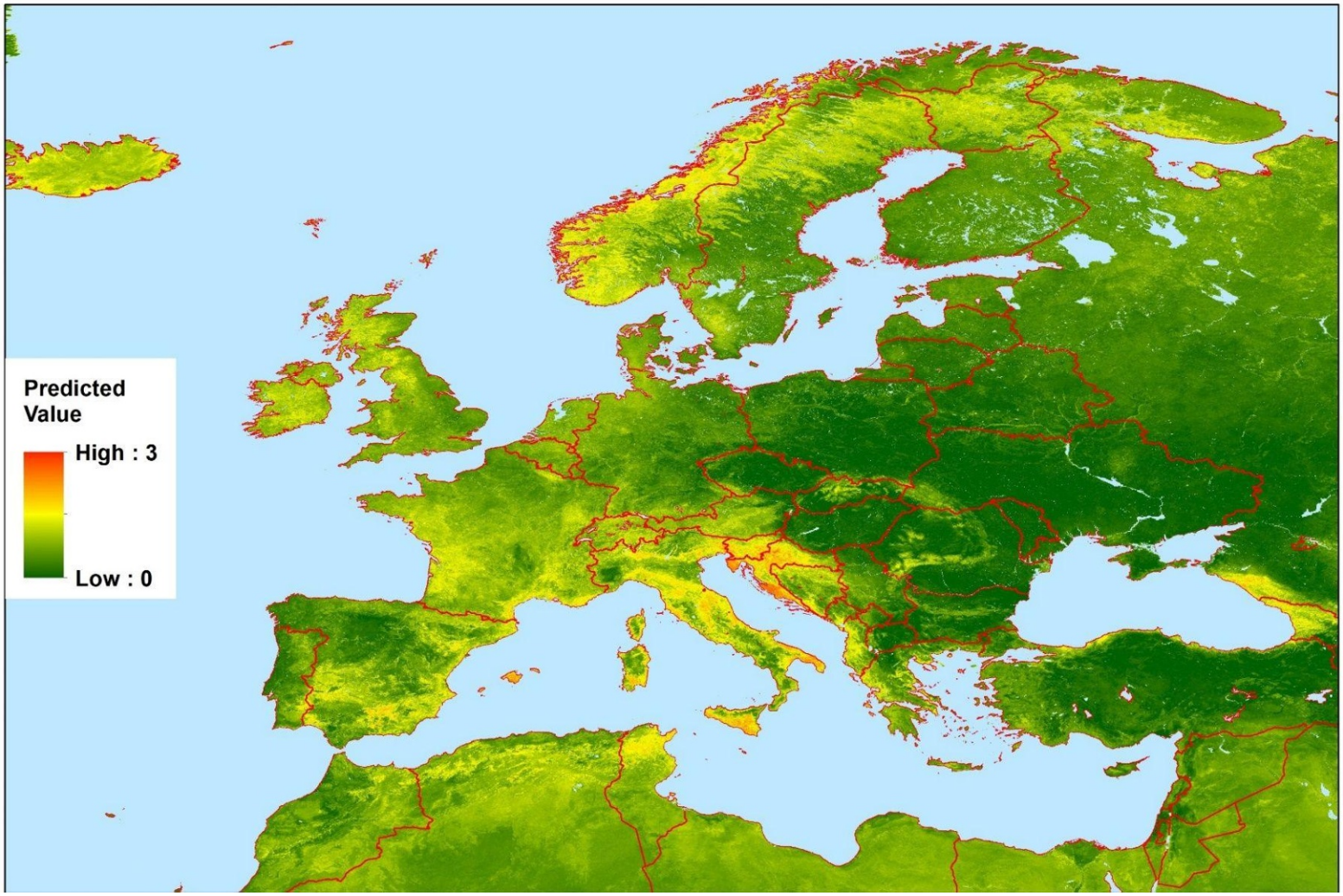

Figure 2: Predicted abundance (log egg equivalent), unmasked, 1 km aggregation for *Aedes albopictus*.

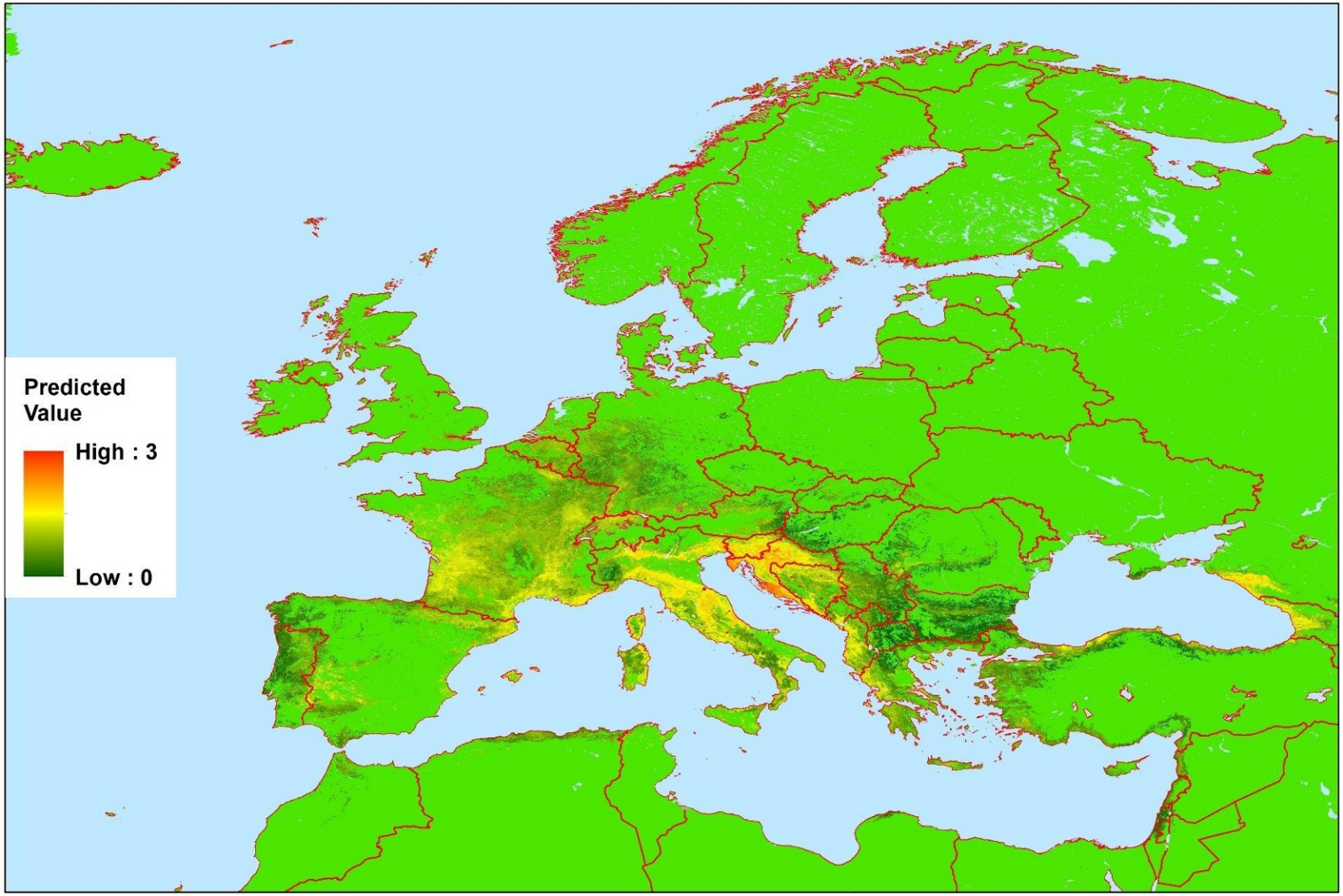

Figure 3: Predicted abundance (log egg equivalent), masked, 1 km aggregation for *Aedes albopictus*.

Figure 4: Available point abundance data (log EVS baited equivalent), 10 km aggregation for *Aedes caspius*.

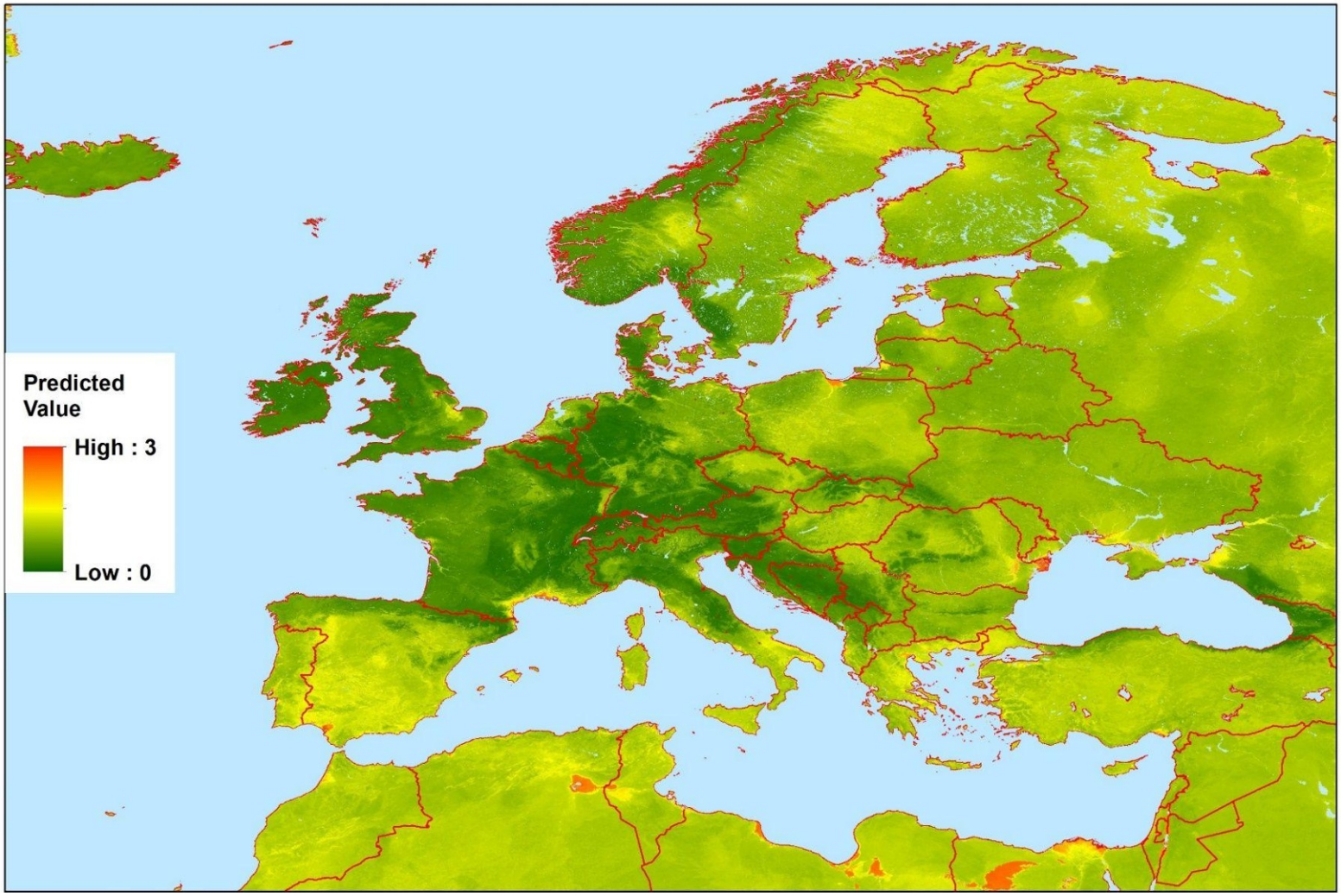

Figure 5: Predicted abundance (log EVS baited equivalent), unmasked, 10 km aggregation for *Aedes caspius*.

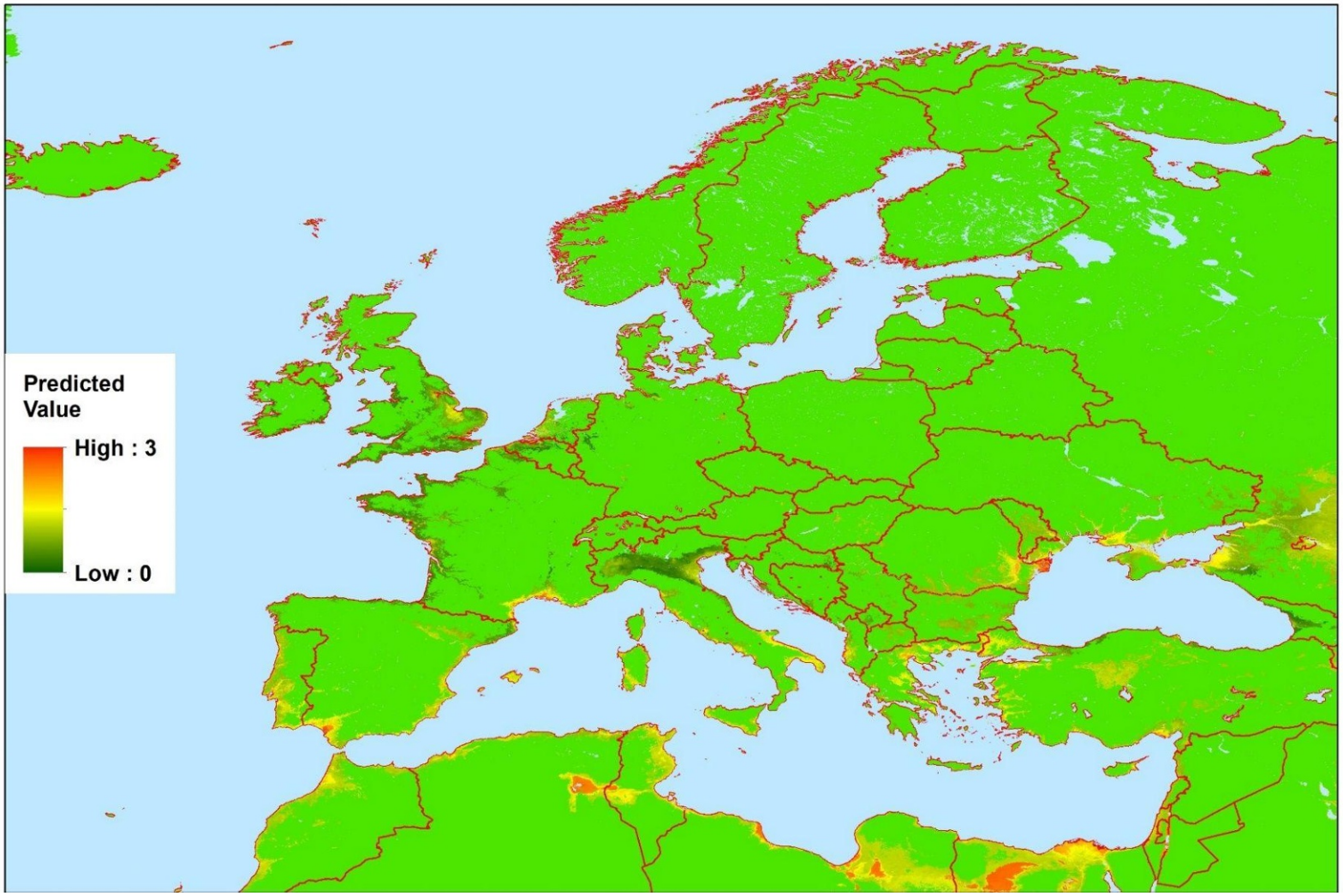

Figure 6: Predicted abundance (log EVS baited equivalent), masked, 10 km aggregation for *Aedes caspius*.

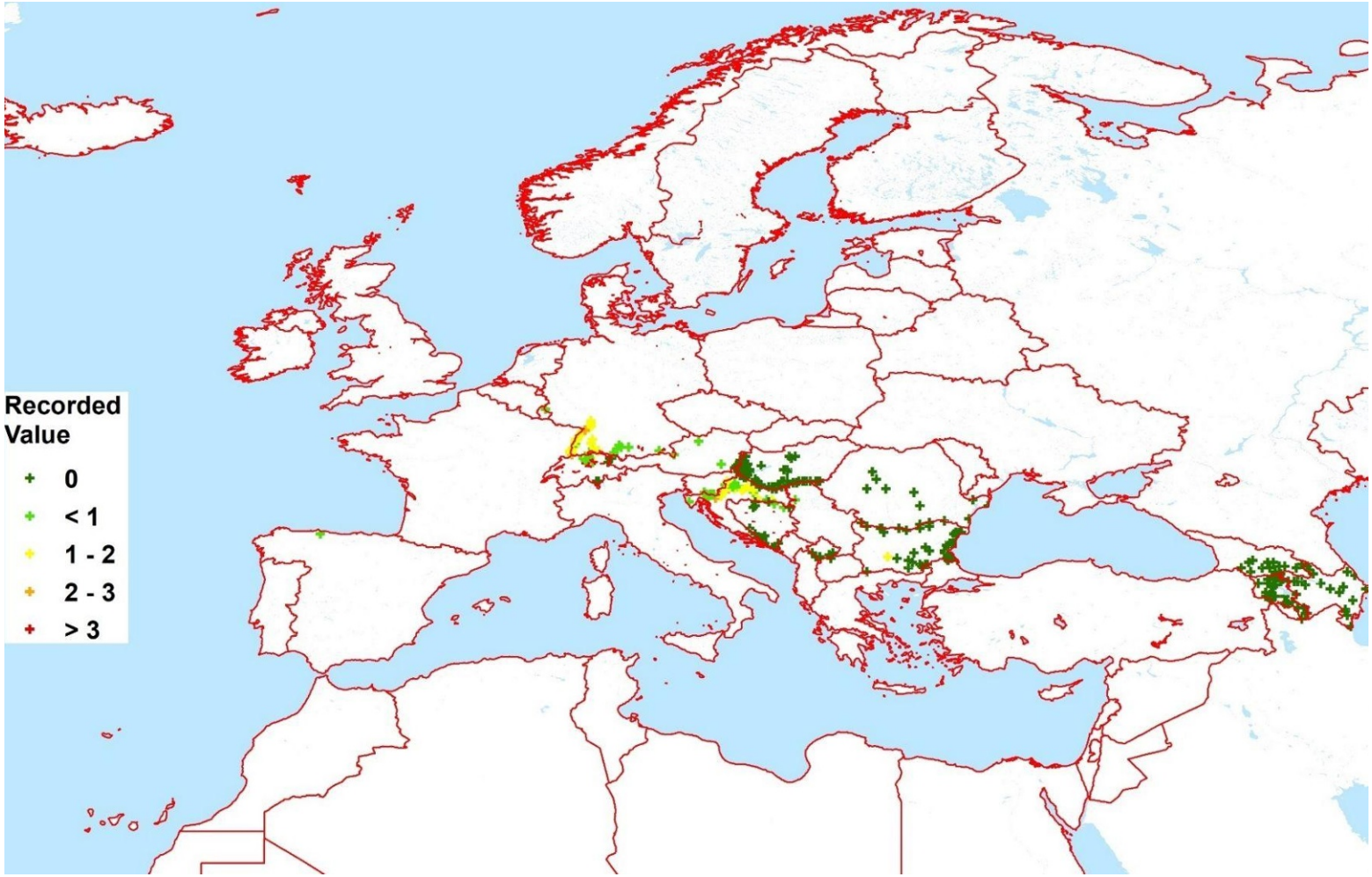

Figure 7: Available point abundance data (log EVS baited equivalent), 10 km aggregation for *Aedes japonicus*.

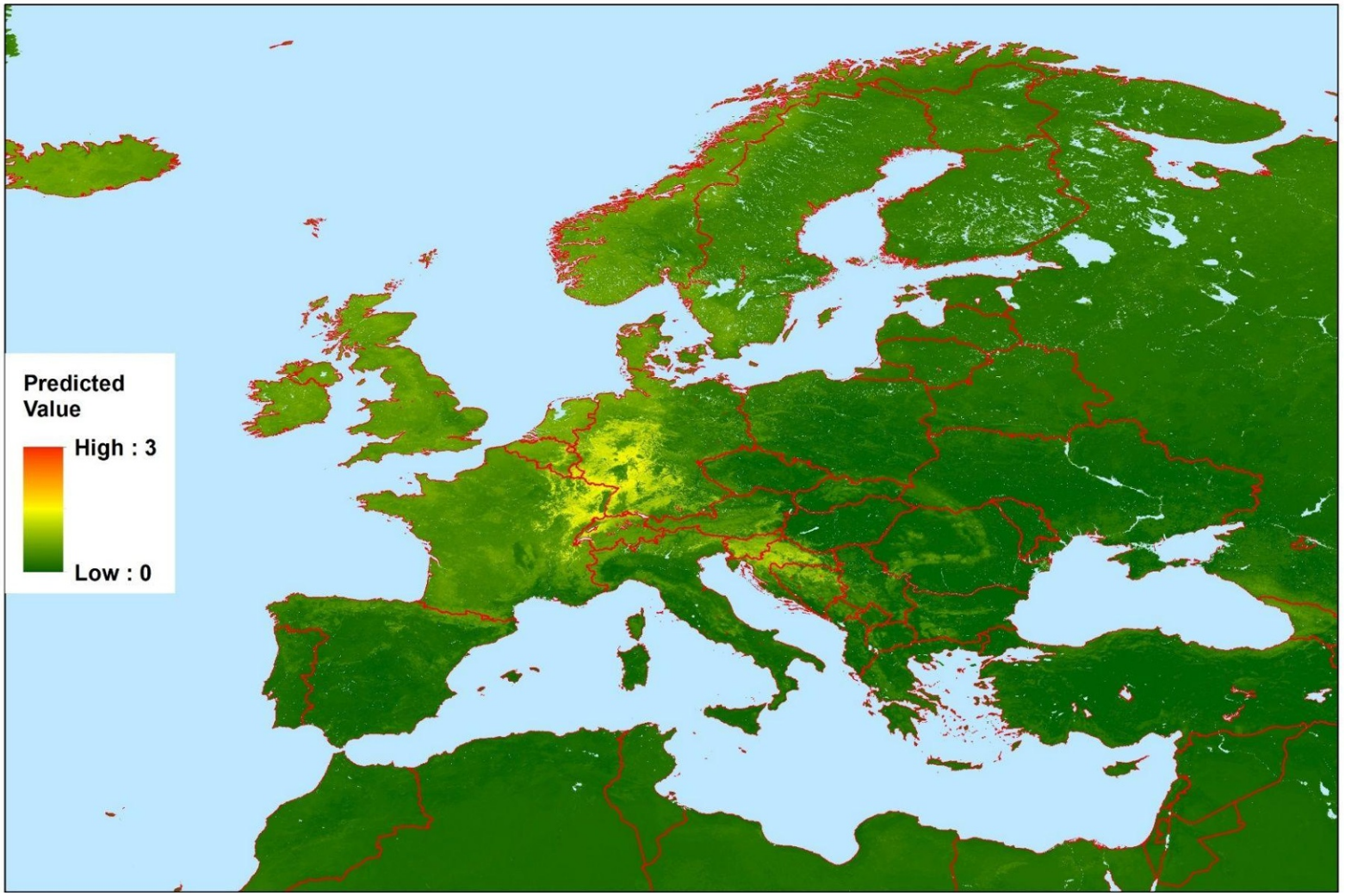

Figure 8: Predicted abundance (log EVS baited equivalent), unmasked, 10 km aggregation for *Aedes japonicus*.

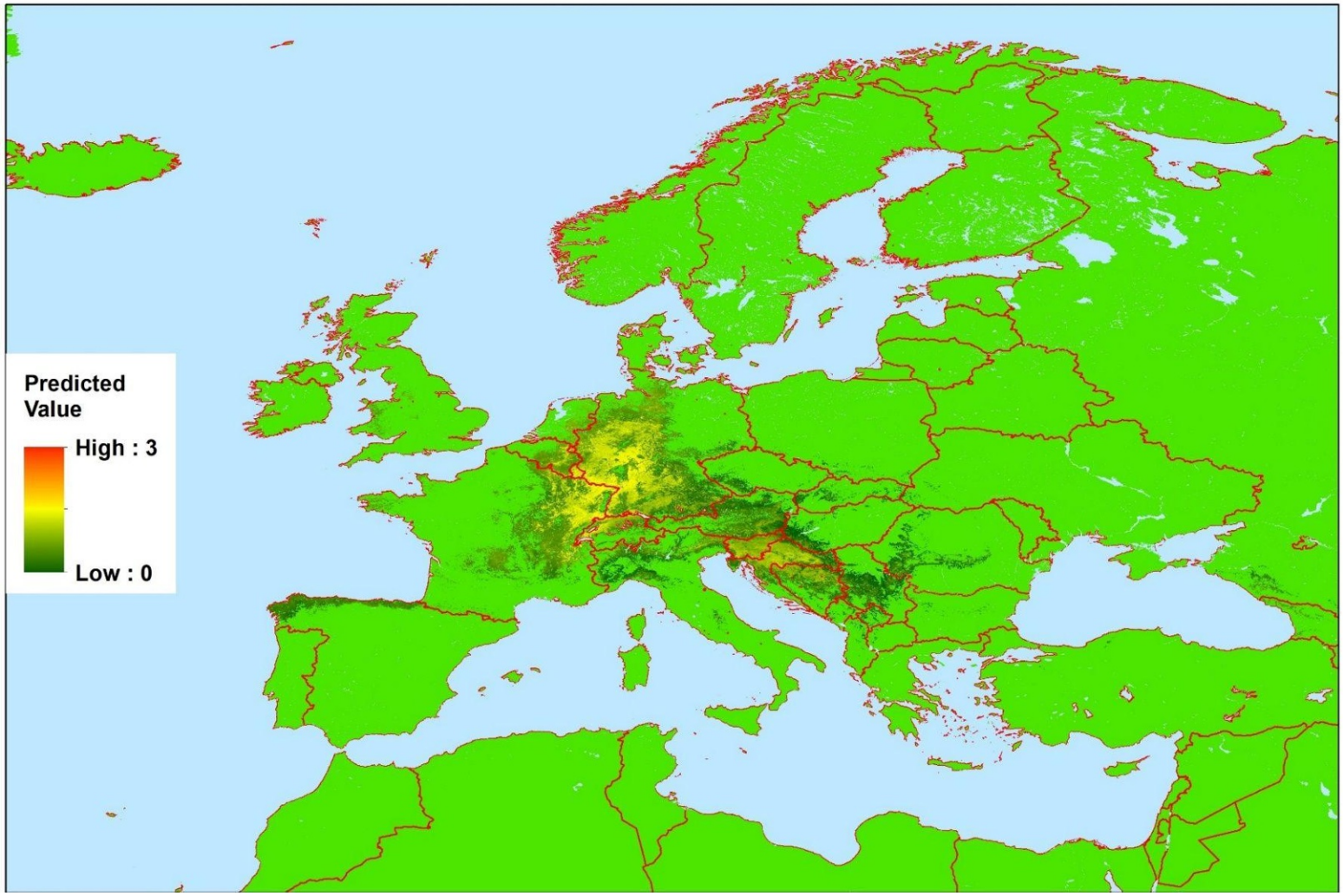

Figure 9: Predicted abundance (log EVS baited equivalent), masked, 10 km aggregation for *Aedes japonicus*.

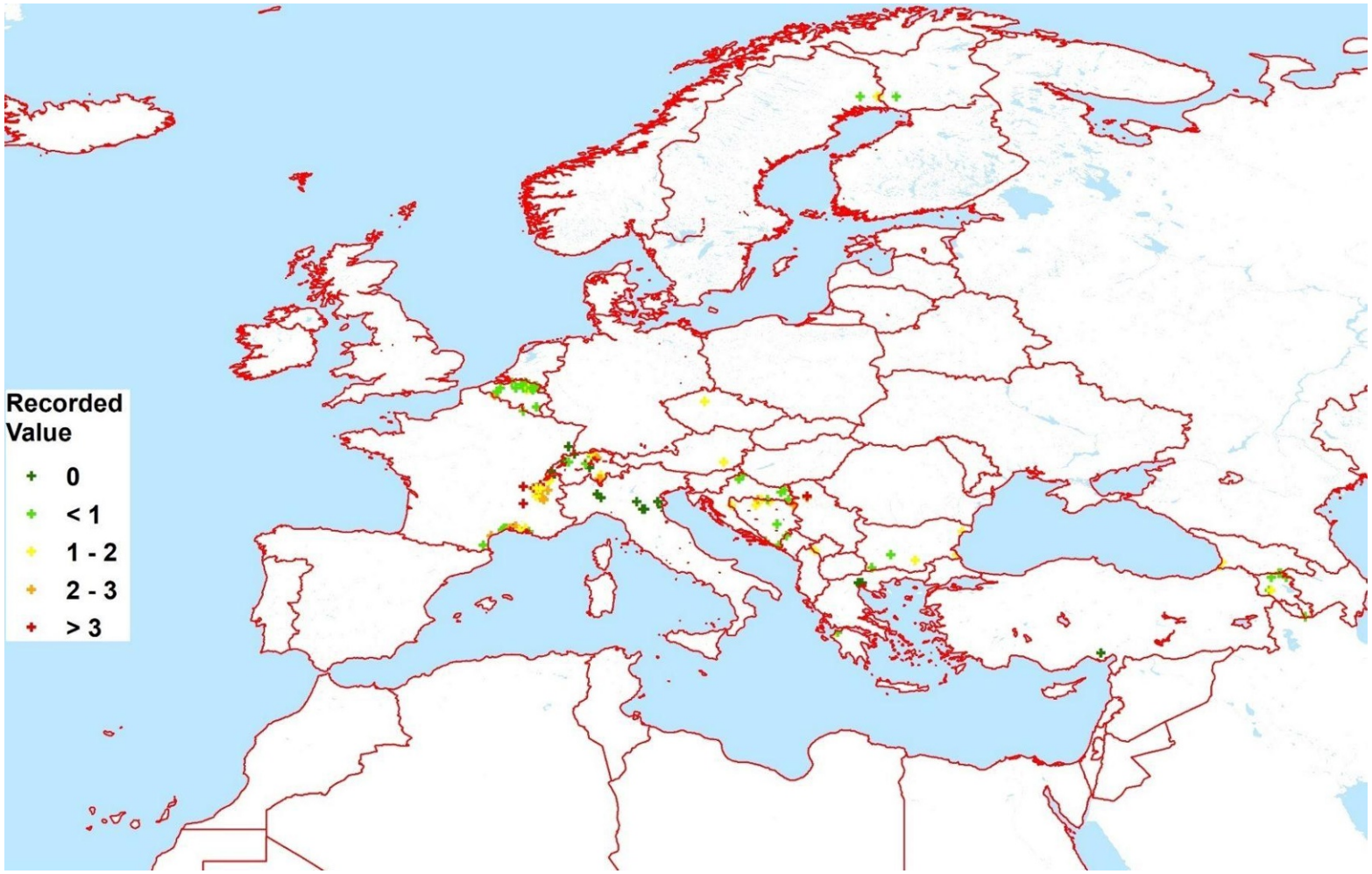

Figure 10: Available point abundance data (log EVS baited equivalent), 10 km aggregation for *Aedes vexans*.

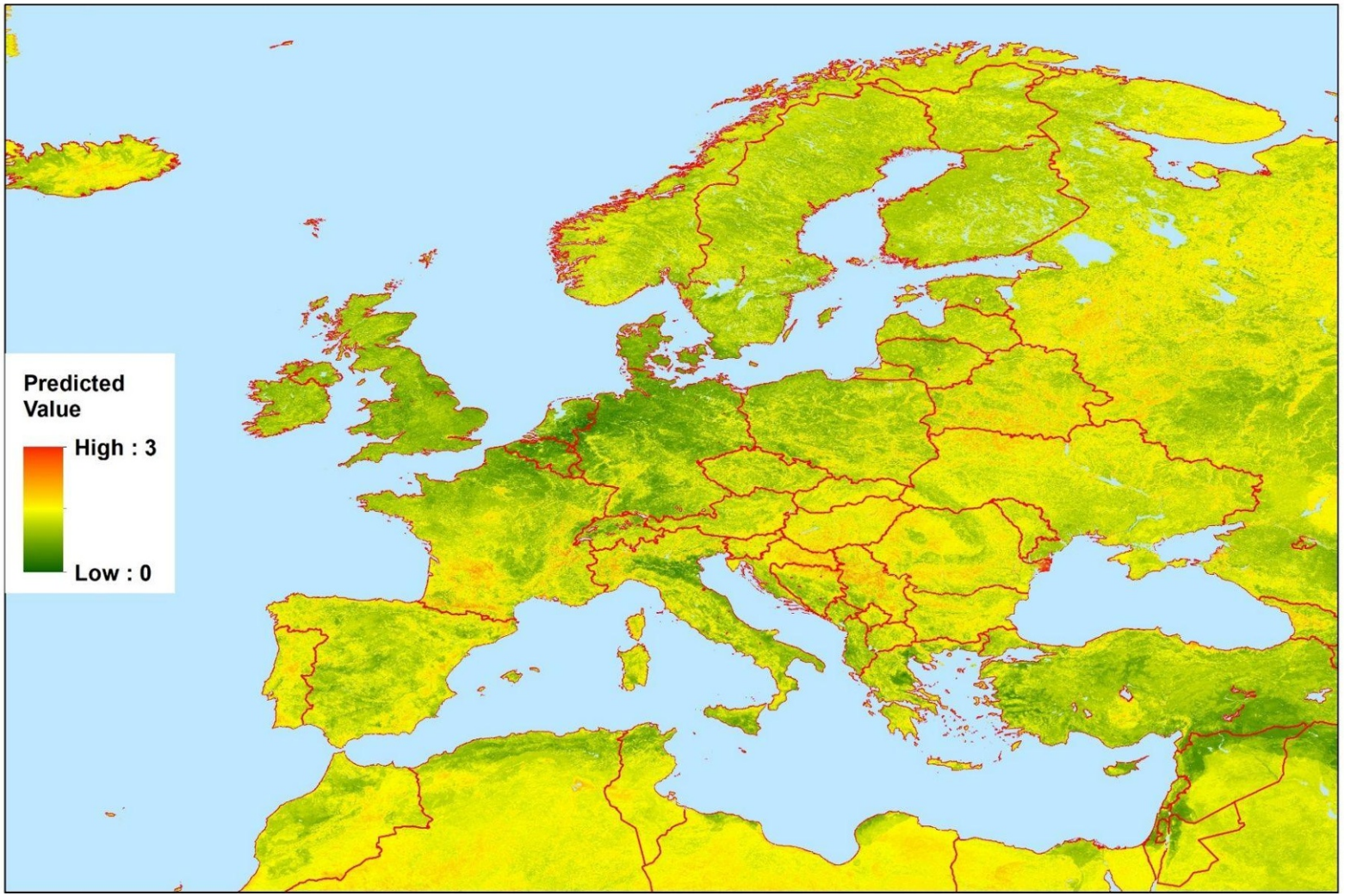

Figure 11: Predicted abundance (log EVS baited equivalent), unmasked, 10 km aggregation for *Aedes vexans*.

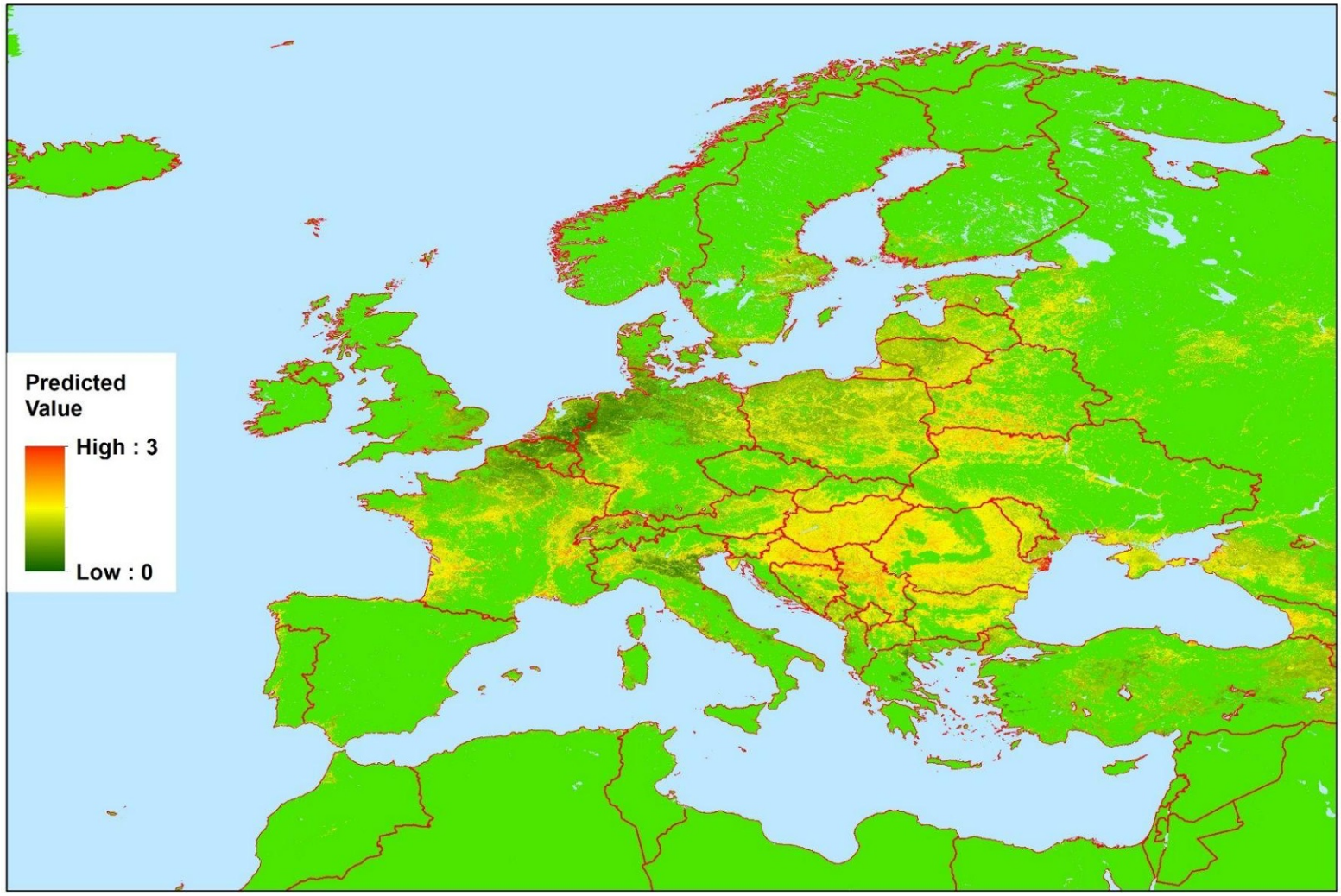

Figure 12: Predicted abundance (log EVS baited equivalent), masked, 10 km aggregation for *Aedes vexans*.

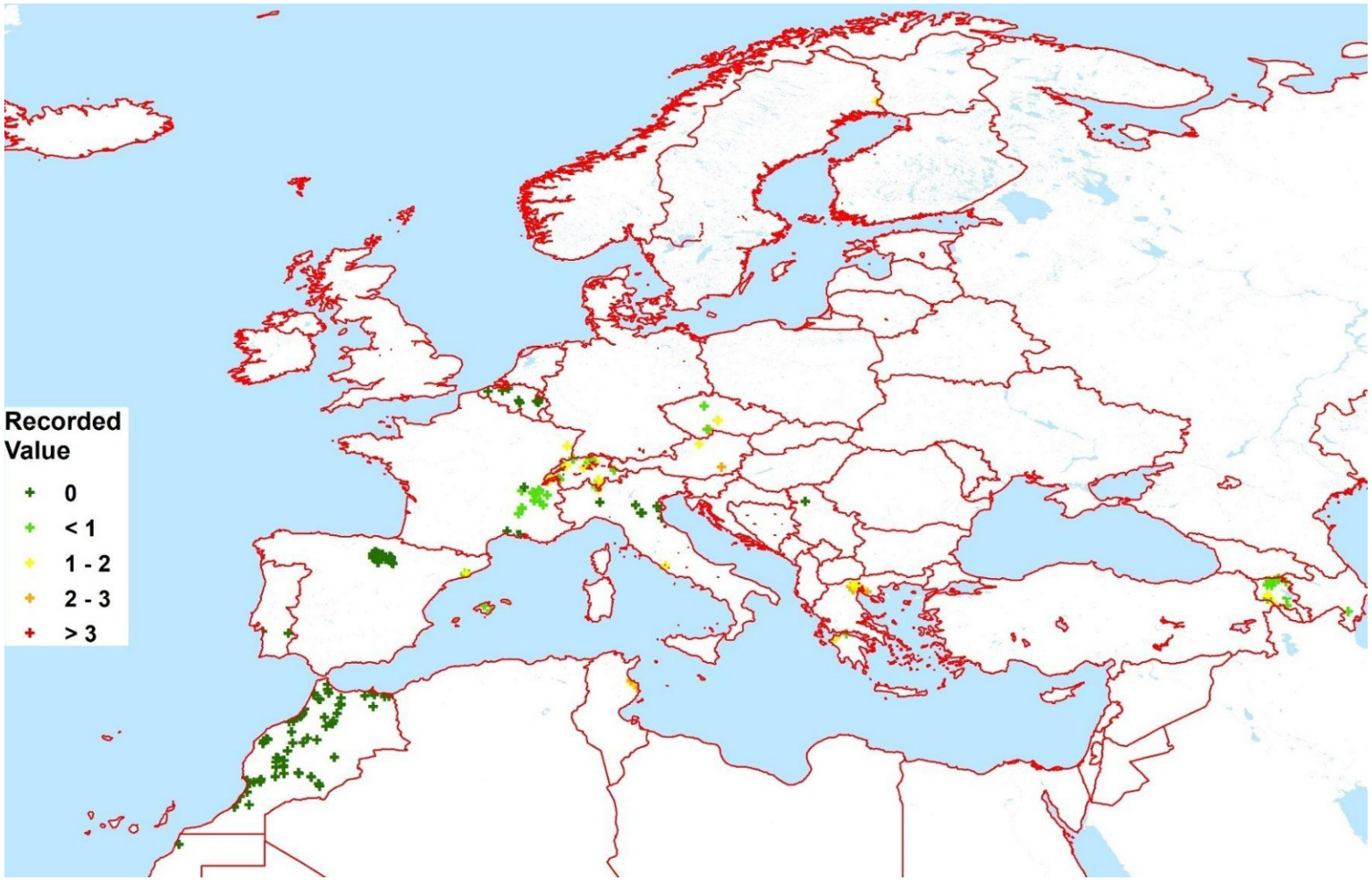

Figure 13: Available point abundance data (log EVS baited equivalent), 1 km aggregation for *Culex pipiens*.

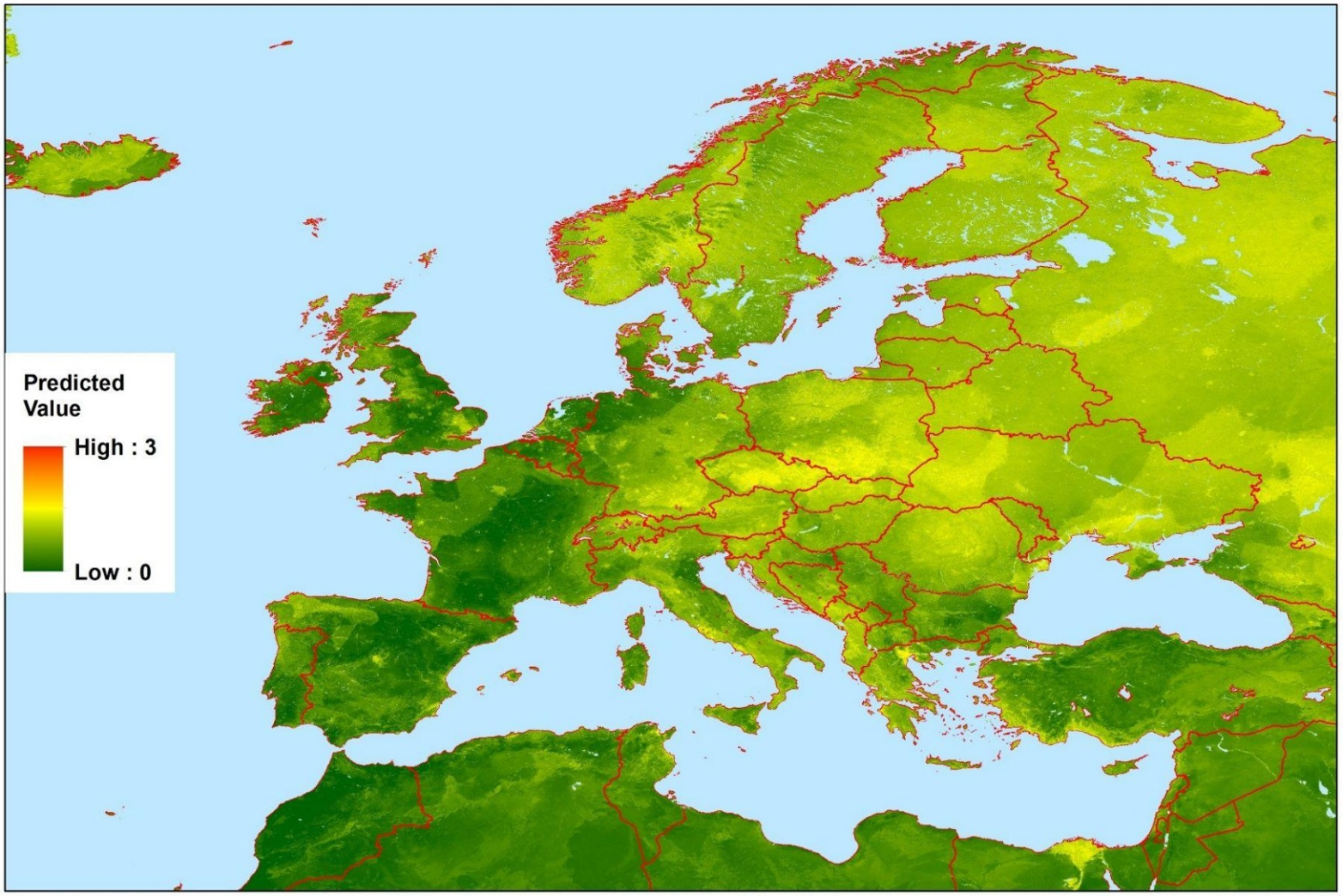

Figure 14: Predicted abundance (log EVS baited equivalent), unmasked, 1 km aggregation, for *Culex pipiens*.

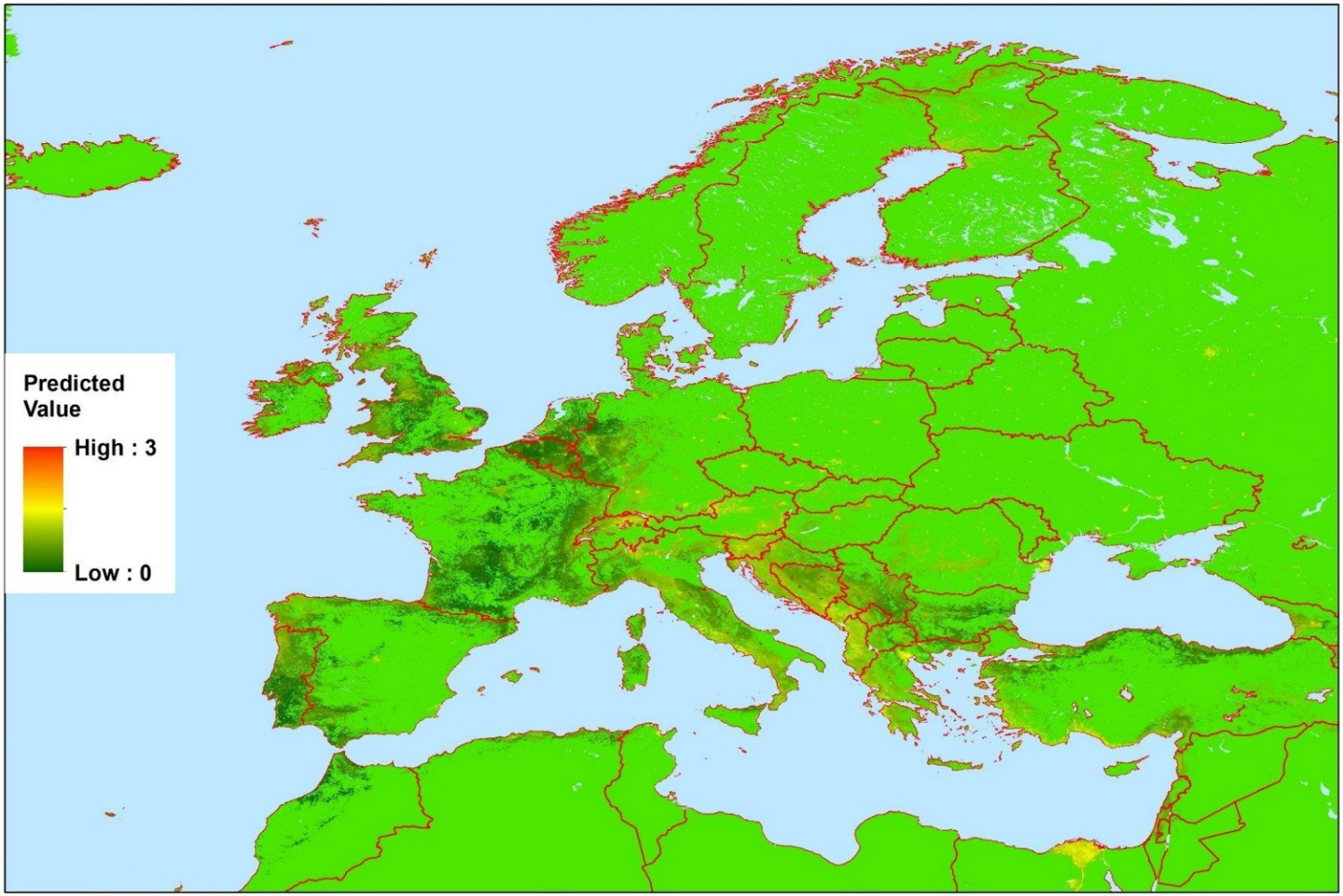

Figure 15: Predicted abundance (log EVS baited equivalent), masked, 1 km aggregation, for *Culex pipiens*.

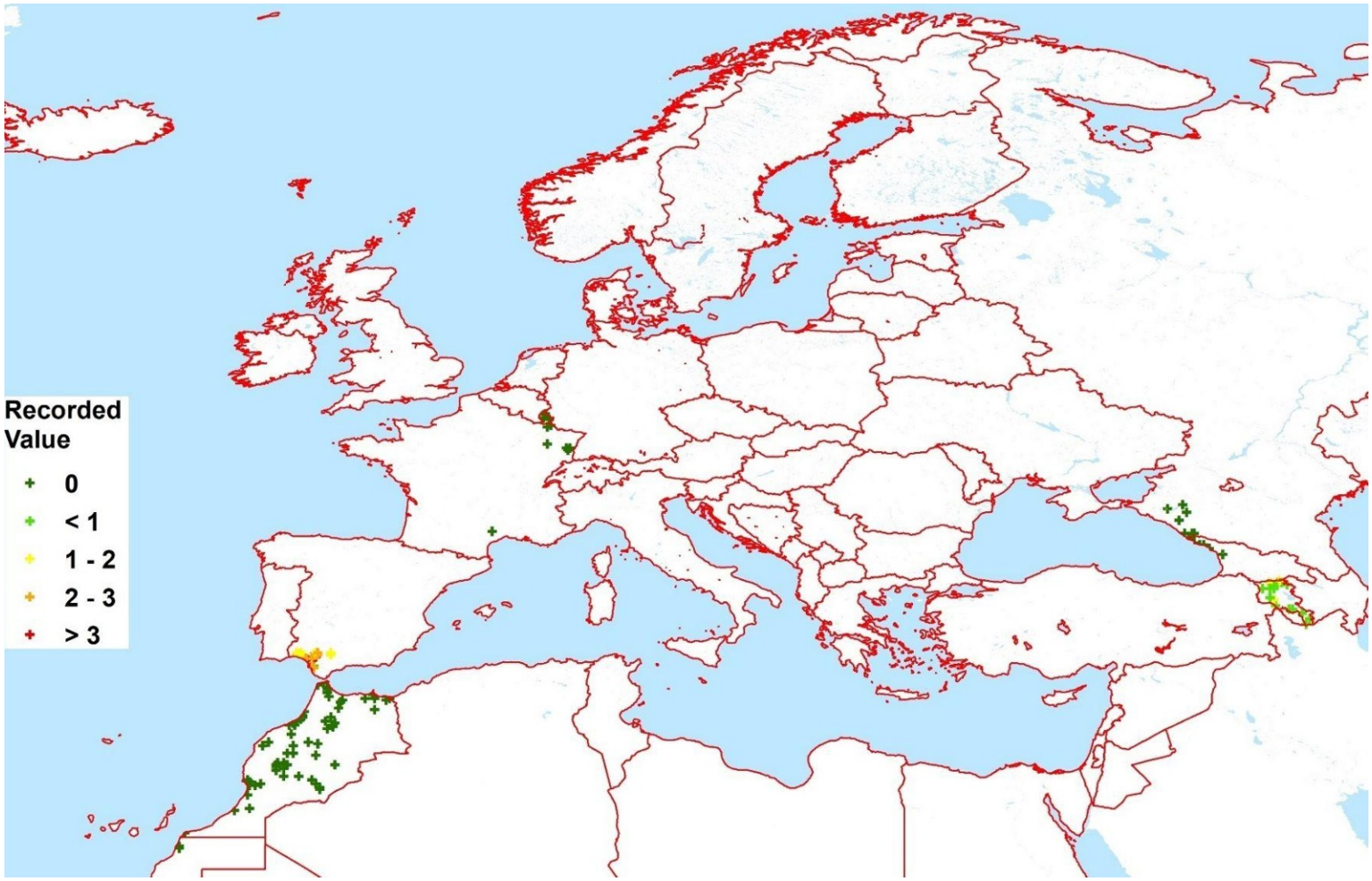

Figure 16: Available point abundance data (log EVS baited equivalent), 10 km aggregation for *Culex theileri*.

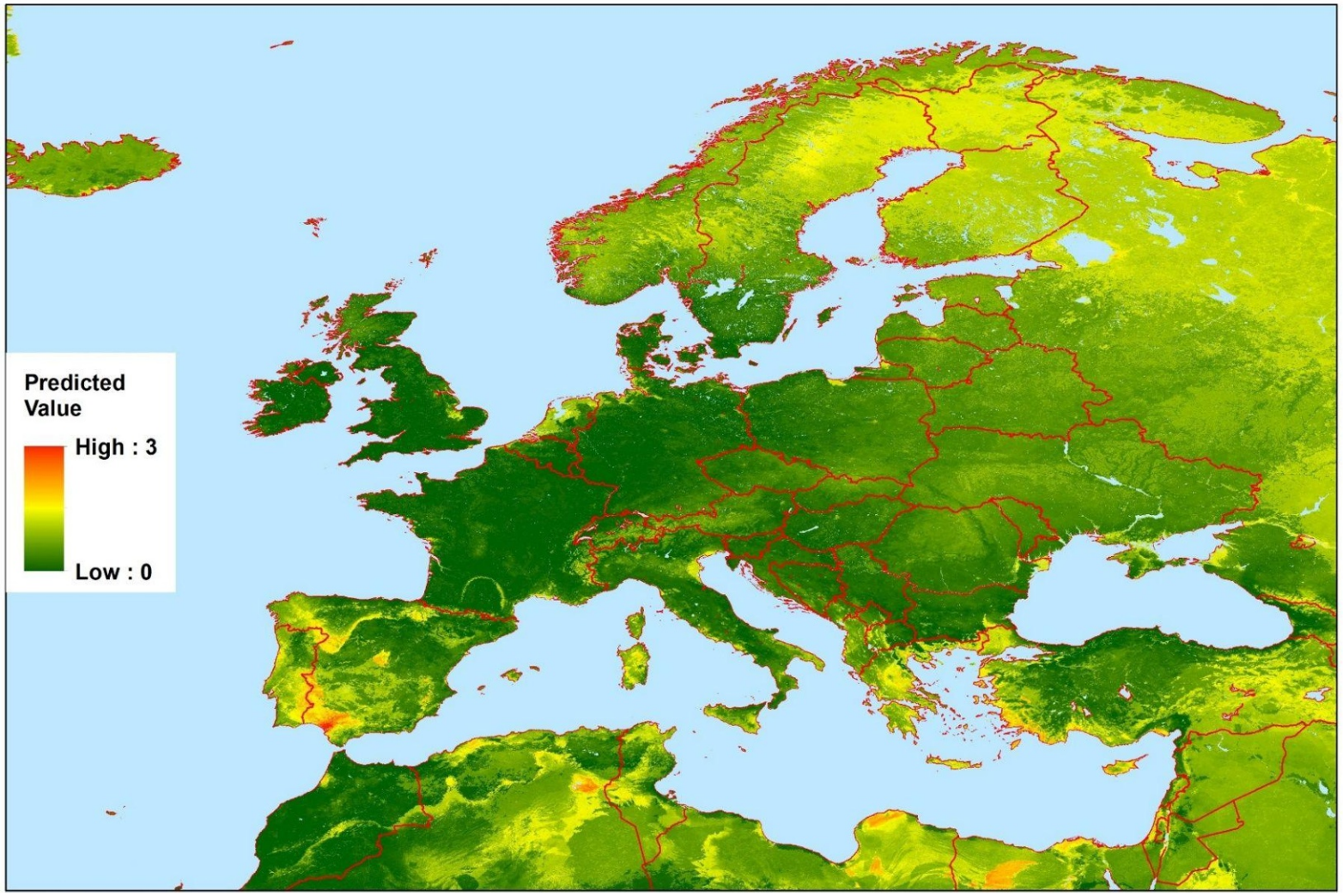

Figure 17: Predicted abundance (log EVS baited equivalent), unmasked, 10 km aggregation, for *Culex theileri*.

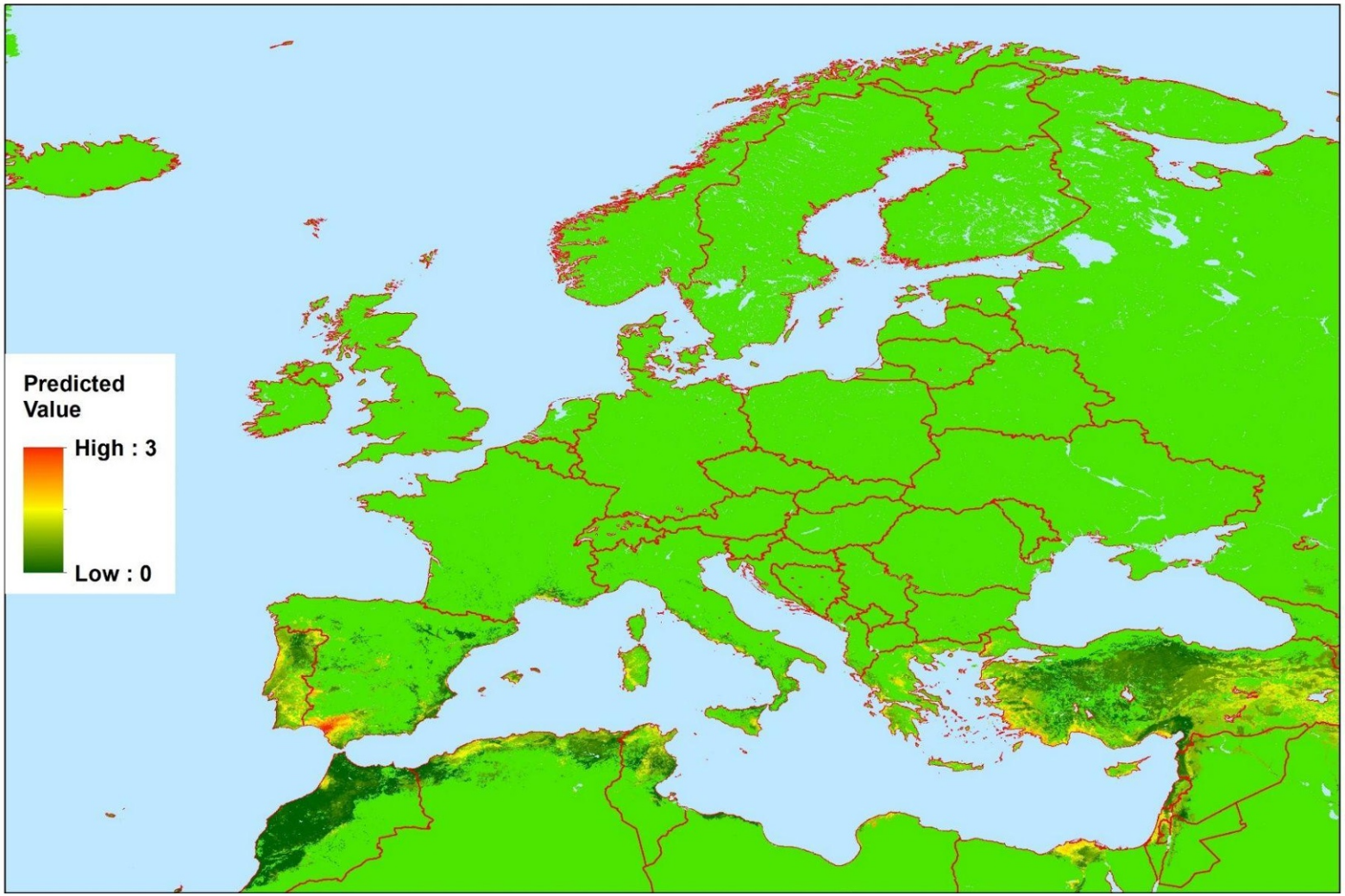

Figure 18: Predicted abundance (log EVS baited equivalent), masked, 10 km aggregation, for *Culex theileri*.