Geospatial Health 8(1), 2013, pp. 241-254

Abundance modelling of invasive and indigenous *Culicoides* species in Spain

Els Ducheyne[1], Miguel A. Miranda Chueca[2], Javier Lucientes[3], Carlos Calvete[4], Rosa Estrada[3], Gert-Jan Boender[5], Els Goossens[1], Eva M. De Clercq[1], Guy Hendrickx[1]

[1]Avia-GIS, Zoersel, Belgium; [2]Laboratory of Zoology and Emerging Diseases, University of the Balearic Islands, Mallorca, Spain; [3]Unidad de Sanidad y Produccion Animal, Centro de Investigacion y Tecnologia, Agroalimentaria, Zaragoza, Spain; [4]Departamento de Patalogia Animal, Universidad de Zaragoza, Zaragoza, Spain; [5]Central Veterinary Institute, Wageningen University, Lelystad, The Netherlands

Abstract. In this paper we present a novel methodology applied in Spain to model spatial abundance patterns of potential vectors of disease at a medium spatial resolution of 5 x 5 km using a countrywide database with abundance data for five *Culicoides* species, random regression Forest modelling and a spatial dataset of ground measured and remotely sensed eco-climatic and environmental predictor variables. First the probability of occurrence was computed. In a second step a direct regression between the probability of occurrence and trap abundance was established to verify the linearity of the relationship. Finally the probability of occurrence was used in combination with the set of predictor variables to model abundance. In each case the variable importance of the predictors was used to biologically interpret results and to compare both model outputs, and model performance was assessed using four different accuracy measures. Results are shown for *C. imicola*, *C. newsteadii*, *C. pulicaris* group, *C. punctatus* and *C. obsoletus* group. In each case the probability of occurrence is a good predictor of abundance at the used spatial resolution of 5 x 5 km. In addition, the *C. imicola* and *C. obsoletus* group are highly driven by summer rainfall. The spatial pattern is inverse between the two species, indicating that the lower and upper thresholds are different. *C. pulicaris* group is mainly driven by temperature. The patterns for *C. newsteadii* and *C. punctatus* are less clear. It is concluded that the proposed methodology can be used as an input to transmission-infection-recovery (TIR) models and R_0 models. The methodology will become available to the general public as part of the VECMAP^{TM} software.

Keywords: spatial abundance modelling, medium resolution, *Culicoides*, random Forests, Spain.

Introduction

Bluetongue, a vector-borne arboviral (Orbivirus; Reoviridae) infectious disease listed by the World Organization for Animal Health (OIE), is transmitted by *Culicoides* biting midges spp. (Diptera: Ceratopogonidae). Twenty-four different bluetongue virus (BTV) serotypes are currently known worldwide. In European outbreaks, five serotypes have been identified in the Mediterranean biome (BTV1, 2, 4, 9, 16) and four in the temperate biome (BTV1 and 8 plus two alleged vaccine strains BTV6 and BTV11). Depending on the serotype, BTV causes high morbidity and mortality in certain breeds of sheep and other domestic and wild ruminants (Elbers et al., 2008; Le Gal et al., 2008; Allepuz et al., 2010). The disease has a high economic impact on livestock, e.g. the total economic losses of the recent BTV8 epidemic in the Netherlands amounted to € 32 million in 2006 and € 164-175 million in 2007 (Velthuis et al., 2010).

In the European Mediterranean biome the BTV spread that started in the late 1990s was largely, but not solely, related to the originally tropical vector *C. imicola*. At its distribution margins BTV was also related to indigenous European *Culicoides* species, most notably species of the *C. obsoletus* complex (Purse et al., 2008). Whether the presence of *C. imicola* in the Mediterranean biome is related to a recent invasion followed by the incursion of BTV serotypes or it has already been present for a longer period of time is still an open question (Conte et al., 2009), but recent work has shown that the theory of a recent introduction is highly unlikely (Mardulin et al., 2013).

Until 2006, bluetongue remained limited to the areas around the Mediterranean, i.e. the Mediterranean biome. However, from August 2006, in the absence of *C. imicola*, an unprecedented introduction, establishment and spread in the temperate biome of the non-Mediterranean BTV8 serotype, solely based

Corresponding author:
Els Ducheyne
Avia-GIS
Risschotlei 33, 2980 Zoersel, Belgium
Tel. +32 3 458 2979
E-mail: educheyne@avia-gis.be

on indigenous *Culicoides* species, have been observed north of 50° N in Benelux, Germany and France (Saegerman et al., 2008). In the following 2 years, this serotype spread very rapidly in the temperate part of Europe, including the temperate tip of Sweden and Norway. The "restriction zone" imposed by the European Union (EU) now covered an area up to approximately 2.3 million km^2, i.e. 43% of the total European territory.

Whilst BTV1 had been restricted to the European eastern Mediterranean biome until 2006, a newly introduced BTV1 strain originating from Morocco invaded the *C. imicola* range in 2007 in Spain (OIE, 2007a) and in Portugal (OIE, 2007b). From the latter country it spread to the northern part of Spain and also southern France (OIE, 2007c) into areas where *C. imicola* is absent and where indigenous European midge species of the *C. obsoletus* complex and *C. newsteadii* are responsible for transmission of the BTV. The potential further expansion northwards was stopped in 2009 by a massive, countrywide vaccination campaign in France. In addition, BTV8 also spread southwards into the European Mediterranean biome, invaded Spain and also appeared in Italy, creating a large geographical overlap between both serotypes in Europe.

Several authors have modelled the probability of the occurrence of *Culicoides* species in the Mediterranean basin, relating the observed presence/absence or abundance of the midges to meteorological and environmental variables mainly derived from satellite imagery. These relationships are established using either statistical techniques such as non-linear discriminant analysis as used by Tatem et al. (2003) in Portugal, logistic regression (Calvete et al., 2008) and more recently, data-mining techniques such as random Forests (Peters et al., 2011).

In order to develop transmission-infection-recovery (TIR) models (Szmaragd et al., 2009, 2010) or basic reproduction number (R_0) models (Hartemink et al., 2009), abundance data of *Culicoides* spp. are an essential parameter. Some abundance models yield abundance classes as output (Tatem et al., 2003), while other studies established a direct linear relationship between probability of occurrence and abundance measured in the traps (Calvete et al., 2008; Guis et al., 2011). The first group of methods could be used as input for TIR models or for R_0 models (e.g. Hartemink et al., 2011) but because the output is categorical map outputs that indicate risk in indirect manner; the second category assumes that there is a linear relationship between probability of occurrence and abundance, a relationship that remains to be validated for *Culicoides*. In this paper, we go one step further and present a new methodology to estimate abundance data using random Forests and a wide set of meteorological and environmental data sets. Outputs are generated as a continuous raster at a spatial resolution of 5 x 5 km. First, we estimate the probability of occurrence, in a second step a direct regression between the probability of occurrence and trap abundance is established to verify the linearity of the relationship; finally, probability of occurrence is used in combination with an additional set of predictor variables to estimate the abundance.

Material and methods

Entomological data

Data collected on mainland Spain and the Balearic Islands under the Spanish Bluetongue National Surveillance Programme in 2007 (Calvete et al., 2008) were used in this study. *Culicoides* spp. specimens were caught using ultraviolet light traps, fitted with a suction fan and a collection vessel containing ethanol and ethylene glycol in water to preserve the samples. The traps were positioned outside selected farms with a minimum of 10 large ruminants and not further than 30 m away from livestock. The traps were operational for one night per week in each farm. Monthly aggregated data showing the maximum catch per farm were available for this study. A hand-held global positioning system receiver recorded the coordinates of the sample locations (Fig. 1). The red dots indicate traps where specimens were found, while the green dots represent traps where no specimens were caught. The abundance of a species was calculated and reported as $log_{10}(n+1)$, where n is equal to the number of individuals caught in a trap. The radius of the red circle in the figure is a measure of the abundance.

Trapped *Culicoides* spp. were identified as described by Calvete et al. (2008) resulting in abundance data for the following species: *C. imicola, C. pulicaris, C. punctatus, C. newsteadii* and the *Obsoletus* group containing *C. obsoletus, C. scoticus, C. montanus, C. dewulfi* and *C. chiopterus*.

Meteorological and environmental data

Variables describing environmental conditions for *Culicoides* were selected based on expert knowledge and a literature review (Conte et al., 2007; Calvete et al., 2008). This study included a total of 74 variables

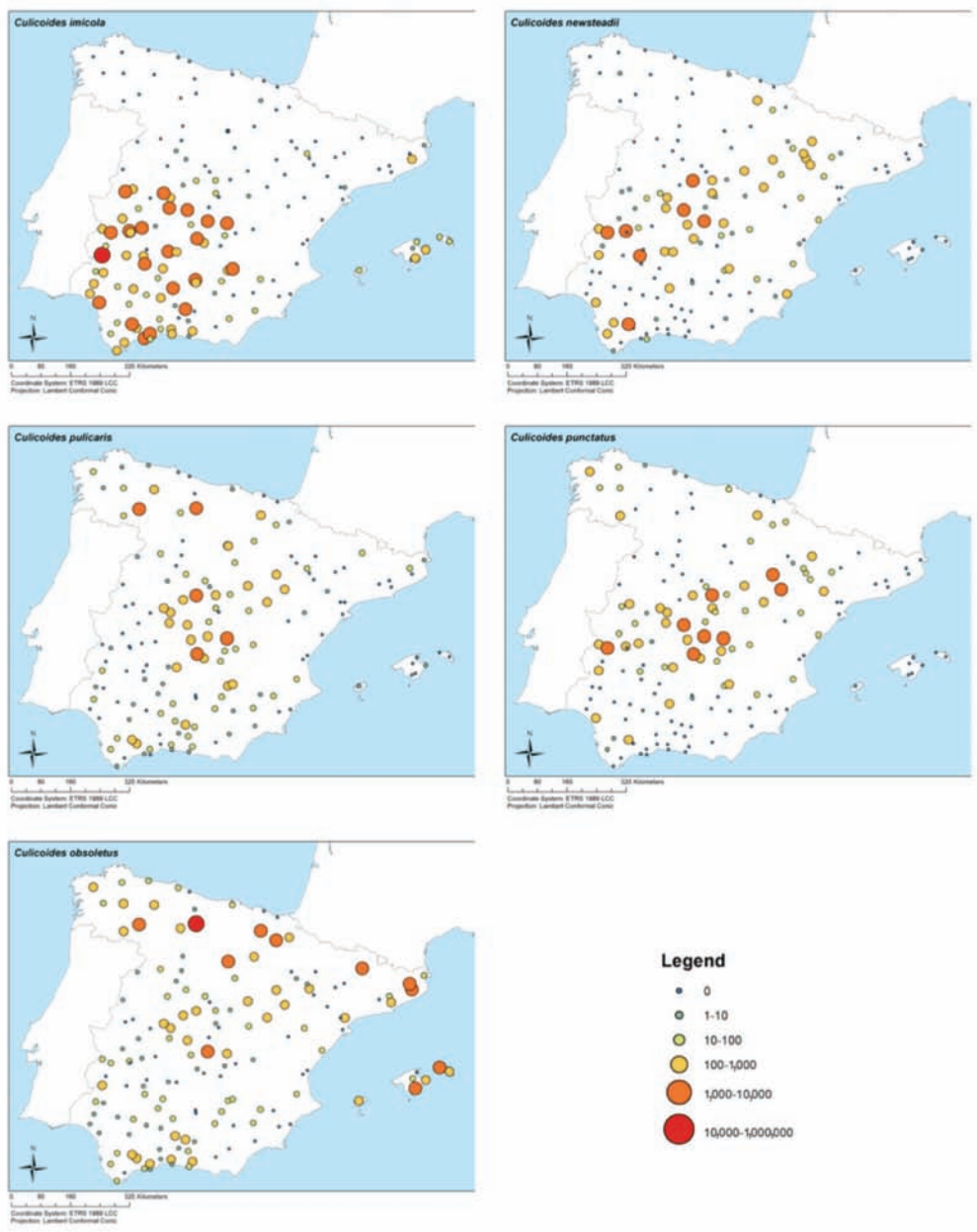

Fig. 1. Location and abundance of five different *Culicoides* species in Spain (2007).

belonging to the following categories: land cover, ground water-related variables, precipitation and land surface temperature (LST).

The land cover information was derived from the CORINE dataset (JRC, 2005). The percentages of the three main land cover classes relevant for *Culicoides*, i.e. urban, agriculture and natural, within a 5 x 5 km pixel were computed. The percentage cover was determined for land cover classes particularly favourable for *Culicoides,* i.e. pastures and forest. Within the forest class, cover percentage was also computed for three forest types: broadleaved, coniferous and mixed forest. Human population pressure on the landscape was assessed by the number of inhabitants/km^2 (GWPv3.0 compiled as described by Wint (2005)). Other environmental data layers included the water capacity of the topsoil (JRC, 2009), distance to waterways (GfK geoMarketing, 2009) and the GTOPO30 elevation (USGS, 1996). The mean total yearly precipitation and the mean monthly mean precipitation were obtained from the WORLDCLIM dataset (Hijmans et al., 2005). Data from the MODIS sensor (http://modis.gsfc.nasa.gov) were used to derive additional variables such as the LST during daytime and at night-time as well as two vegetation indices, i.e. the enhanced vegetation index (EVI) and

Table 1. Number of traps and catches of *Culicoides* spp. in Spain.

Species	Positive traps	Negative traps	Total
C. imicola	96	62	158
C. newsteadii	78	80	158
C. pulicaris	112	48	158
C. punctatus	74	84	158
C. obsoletus	122	36	158

Table 2. Number of traps and catches of used to model the occurrence probability of the various *Culicoides* spp. in Spain.

Species	Positive traps	Negative traps	Total
C. imicola	62	62	124
C. newsteadii	78	78	154
C. pulicaris	48	48	96
C. punctatus	74	74	148
C. obsoletus	36	36	72

the normalized difference vegetation index (NDVI) (Gao et al., 2000). Time series covering the years 2004, 2005, 2006, 2007 and 2008 were used to detect yearly trends. After Fourier transformation, the first three harmonics were included as predictor variables (Scharlemann et al., 2008). These harmonics describe seasonal cycles, e.g. in temperature and vegetation variables, by approximating the temporal signal with cosinusiodal waves. The amplitude of the cosinusoidal wave represents the magnitude and the phase the timing of when the maximum amplitude is reached. For real-life temperatures, for example, the amplitude is the maximum difference between the lowest and the highest recorded temperature, whilst the phase indicates when the peak (maximum of the highest temperature) is achieved. The first harmonics thus represent the major seasonal differences in temperature; subsequent harmonics describe secondary and tertiary seasonal phenomena. Finally, the number of days with a mean temperature above (i) 0°C; (ii) 5 °C; and (iii) 12.5°C were derived from the MODIS LST time series for the year 2007 for which data on midges were available.

All data layers were clipped to the extent of the study area i.e. mainland Spain and Portugal. Geographical information systems (GIS) manipulations were performed using ArcGIS, 9.3 (ESRI, 2009).

Modelling framework

Random Forest modelling

Models were generated using the random Forest (RF) approach (Breiman, 2001). This is a robust ensemble learning technique, which can be applied either to model probability maps using a random classification forest or abundance maps through a random regression forest. The technique consistently outperforms traditional modelling techniques such as logistic regression (Cutler et al., 2007; Peters et al., 2007). Random classification forests have been used to assess if temperature and precipitation affect the minimum infection rate of *Culex* species for the West Nile virus in Illinois (Ruiz et al., 2010) and to model the current spatial distribution of *Aedes albopictus* in Europe using a wide set of predictor variables (ECDC, 2009).

RF allows both internal and external validation through a bootstrapping procedure. For each classification or regression tree, the full data set is bootstrapped, i.e. a number of data points are sampled from the complete data set with replacement. From the bootstrapped sample approximately one third of the data are excluded. This set of the excluded data is referred to as the "out-of-bag" (OOB) dataset for the tree; each tree will have a different OOB dataset. Since these datasets are not used to build the tree, they constitute an independent validation dataset for the tree in absence of autocorrelation.

To measure the classification error of the random classification forest, the OOB data for each tree are classified and the classification error is computed. The error values for all trees in the forest are averaged to give the overall classification error. In case of random regression forests, the error is expressed as the mean squared error between the predicted values for the OOB data and the observed data.

Probability of occurrence maps

For each species, the abundance data were classified into presence and absence classes. If a site were negative over the entire year, it was classified into the absent class; all other sites were classified into the present class. In order to maximize model accuracy (McPherson et al., 2004), equal numbers of presence and absence sites were randomly selected.

The probability modelling was based on a balanced set of presence and absence observations at the sample sites. The presence/absence traps were selected at random from the entire dataset of observations points. The number of traps used for each species can be found in Table 2. The performance of the probability model was assessed using four accuracy measures: per-

centage of correctly classified instances (PCC), sensitivity, specificity and "area under the receiver operating curve" (AUCOC). The AUCOC can be roughly interpreted as the probability that a model will correctly distinguish a true presence and a true absence at random. For example, a value of 0.8 for the AUCOC means that for 80% of the time a random selection from the positive group (presence) will have a score greater than a random selection from the negative class (absence) (Fielding and Bell, 1997). Predictor variable importance is assessed through the measurement of the decline in performance if the model is run without the variable. This performance decline is expressed as the mean decrease in GINI index (Breiman, 2001). The GINI index is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous) *versus* the contribution of each variable. The changes in GINI are summed for each variable and normalised at the end of the calculation.

Abundance maps

Abundance data were $\log_{10}$ transformed according to the following formula $log_{10}(n+1)$, where n is the maximum number of individuals caught for the site in question. The species data for all sites were used in the abundance modelling.

In many cases, the predicted probability of occurrence is a surrogate of habitat suitability, and is therefore used as an indirect predictor of species abundance (Osborne et al., 2001; Boyce et al., 2002; Gibson et al., 2004; Chefaoui et al., 2005; Calvete et al., 2008). To test this hypothesis, in a first step, the Pearson correlation coefficient was established between the predicted probability of occurrence and the observed abundance data. In a second step, the probability of occurrence was added as a predictor variable for modelling the abundance of a species using RF. Accuracy was assessed quantitatively through the calculation of the mean squared error (MSE) and the coefficient of determination (cor) between the $\log_{10}$-transformed observed and predicted abundances. The importance of the predictors for the abundance modelling was assessed using the "increase in node purity" (INP). This measure shows how much the impurity, i.e. a measure for inaccuracy, increases when that variable is omitted from the model. Important variables have a high value. Statistical analysis and modelling were performed in the R2.10.1 statistical language environment (R Development Core Team, 2006) using of the R-package "rgdal", version 0.6-25, and "randomForest", version 4.5-34.

Results

Observed presence and abundance data

For this Spanish study area, data of 158 traps were used. For each species, the number of positive traps, negative traps as well as the total number of traps can be found in Table 1. The observed data show three distinct patterns for *Culicoides* species in Spain (Fig. 1).

C. imicola was mainly present in the drier central and south-western part of continental Spain and mostly absent from the northern more humid part. In addition some specimens were also caught along the Ebro Valley and along the north-eastern Mediterranean coast. The species was also present in the Balearic Islands.

C. newsteadii and *C. punctatus* both have a comparable distribution pattern along a north-east/south-west axis with distinct areas of absence north and south of this axis. Some more positive sites were found for *C. newsteadii* in the southern part of the country and for *C. punctatus* in the North.

Finally, *C. pulicaris* and the *C. obsoletus* group are the most widespread *Culicoides* groups in Spain suggesting they may adapt to a wider range of eco-climatic circumstances than the other species. The distribution pattern of the *C. obsoletus* group was the inverse of the pattern observed with *C. imicola*. Though *C. obsoletus* was present in most of the area covered by *C. imicola*, low densities were recorded, whilst no *C. imicola* were found in northern Spain, where the highest densities of *C. obsoletus* have been recorded.

Probability of occurrence maps

Fig. 2 depicts the probability of occurrence for the *Culicoides* species in Spain. *C. imicola* is predicted to occur in southwest Spain. It has a crisp distinction between the high and low probability zones, which closely resembles the observed distribution pattern. This is reflected in high accuracy indices: PCC = 0.81 and AUCOC = 0.88 (Table 3). *C. newsteadii* and *C. punctatus* exhibit a similar predicted probability of occurrence. The gradient between the higher and lower probability areas is smooth. A large zone shows a medium probability. The accuracy measures are similar for both species: the specificity is fair, 0.78 and 0.72 for *C. newsteadii* and *C. punctatus,* respectively) but the sensitivity is lower, indicating that more sites are falsely classified as being present. This is even more pronounced for the distribution of *C. pulicaris*

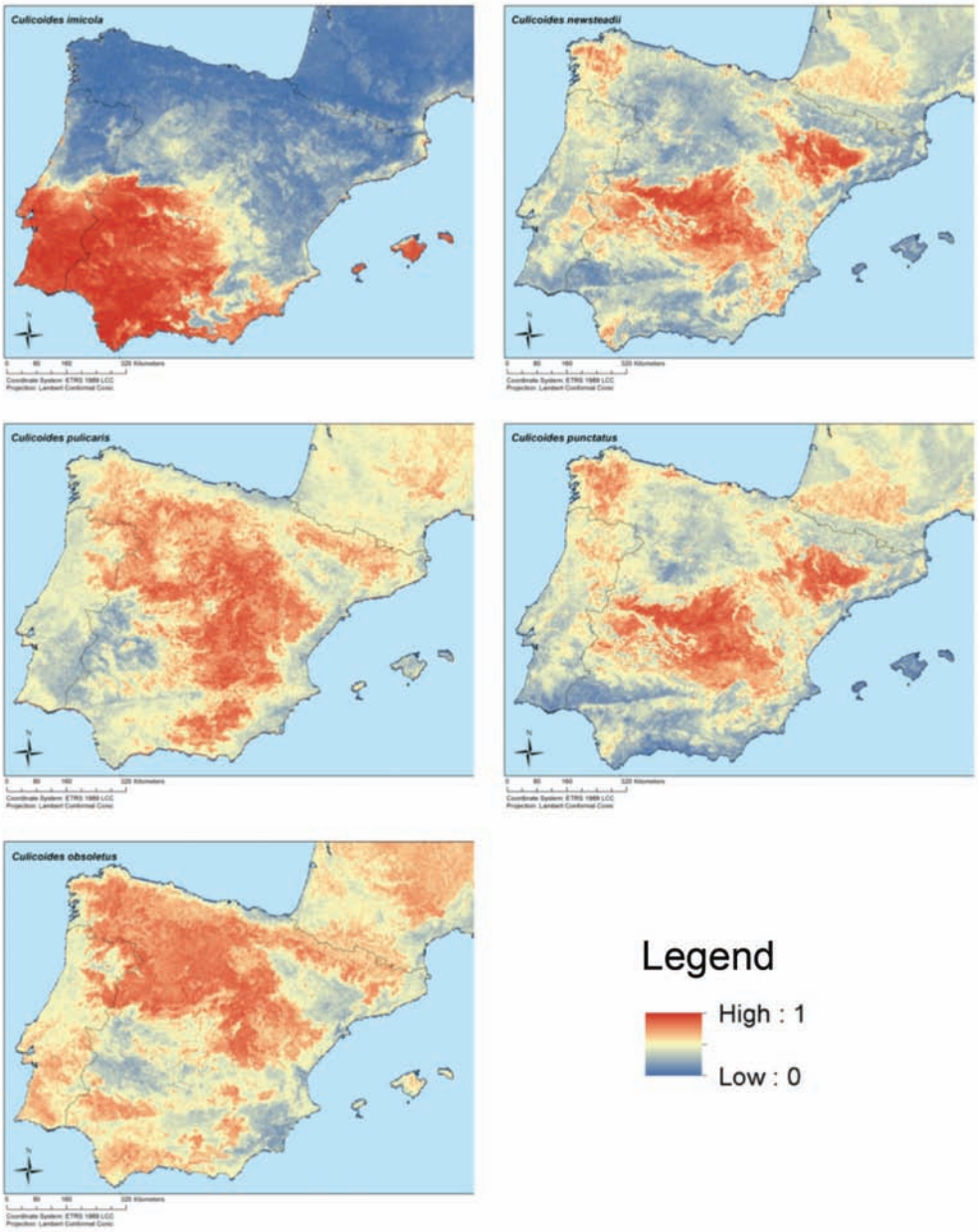

Fig. 2. The predicted probability of occurrence of *C. imicola*, *C. newsteadii*, *C. pulicaris* group, *C. punctatus* and *C. obsoletus* group in Spain.

and *C. obsoletus*, which are generally more ubiquitous resulting in a lower PCC = 0.51 and AUCOC = 0.59 (Table 3).

Table 4 summarises the 10 most important variables driving the model. The occurrence of *C. imicola* is most strongly driven by precipitation. The five most important factors were found to be precipitation variables such as the precipitation level of the driest month and the monthly mean precipitation of the summer months from June until September. The timing of the greening of the vegetation (phase 1 of the NDVI and phase 1 of the EVI) and the number of days with daytime LTS greater than 12 °C were also important. Indeed, in 2008, the temperature variables were found to be one of the 10 most important variables. Spring precipitation was the 10th most important factor.

The occurrence of *C. newsteadii* is determined by the night-time LTS through the first amplitude (peak of the annual cycle) and the third phase (the timing of the peak of the tri-annual cycle), forest cover and the winter precipitation variables (maximum precipitation, plus the precipitation in December and November).

The occurrence of the *C. pulicaris* group turned out to be mainly determined by temperature: seven out of

Table 3. Accuracy measures of the probability of occurrence modelling for *Culicoides* spp. in Spain.

Species	PCC	Sensitivity	Specificity	Kappa	AUCOC
C. imicola	0.81	0.76	0.85	0.61	0.88
C. newsteadii	0.75	0.72	0.78	0.50	0.81
C. pulicaris	0.62	0.59	0.65	0.24	0.58
C. punctatus	0.70	0.69	0.72	0.41	0.80
C. obsoletus	0.51	0.47	0.56	0.03	0.59

10 variables are temperature-related, both day- and nigh temperatures. These variables were: the peak of the bi-annual and tri-annual daytime temperature cycle, the number of days with night temperatures in each year greater than 5 °C, the number of days above the freezing point in 2008 and the peak of the tri-annual night temperature cycle. Elevation, a proxy for temperature, was also retained. Additionally, the timing of the bi-annual (phase 2) and tri-annual vegetation peak (phase 3) of the EVI was found to drive the probability of occurrence. Thus, the crucial factors for the occurrence of *C. punctatus* are a combination of different types of variables with the main factors being temperature (first and third amplitude of night-time LST, the number of days above 0 °C in 2008, the number of days above 12 °C in 2007 and the mean yearly night-time land surface temperature), distance to water and precipitation. However, elevation and forest cover both play a role.

Finally, the occurrence of the *C. obsoletus* group is clearly determined by temperature, both in the day and during night: seven out of the 10 variables were

Table 4. The 10 most important predictors indicated by the variable importance (VARIMP) for the probability of occurrence of the *C. imicola*, *C. newsteadii*, *C. pulicaris* group, *C. punctatus* and *C. obsoletus* groups.

C. imicola		*C. newsteadii*		*C. pulicaris* group		*C. punctatus*		*C. obsoletus* group	
Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP
Minimum precipitation	6.19	Nighttime LST, amplitude 1	2.51	Day LST, amplitude 2	1.52	Nighttime LST, amplitude 1	2.82	EVI, phase 3	1.22
July precipitation	5.63	Forest cover (% per pixel)	2.3	Nights in 2007 with LST >5 °C	1.35	Distance to water	2.57	Nights in 2008 with LST >5 °C	0.98
September precipitation	3.46	Nighttime LST, phase 3	2.13	EVI, phase 3	1.33	June precipitation	2.32	EVI, amplitude 3	0.96
August precipitation	3.22	December precipitation	2.13	EVI, phase 2	1.25	December precipitation	2.1	Nighttime LST, phase 3	0.92
June precipitation	3.08	Population density	2.05	Elevation (from the DTM model)	1.08	Nighttime LST, amplitude-3	1.84	NDVI, phase 1	0.88
NDVI Phase 1	2.63	Distance to water	1.86	No. of days in 2008 with LST >0 °C	1.08	No. of days in 2008 with temp. >0 °C	1.8	Nights in 2007 with LST >5 °C	0.86
Daytime LST, phase 2	2.30	Maximum precipitation	1.83	No. of nights in 2006 with LST >5 °C	1.04	Days in 2007 with LST >12 °C	1.73	Nighttime LST, mean	0.84
EVI, phase 1	1.80	Total precipitation	1.83	Daytime LST, amplitude 3	1.02	Forest cover (% per pixel)	1.57	Days in 2008 with LST >12 °C	0.80
Days in 2008 with daytime LST >12 °C	1.66	November precipitation	1.70	No. of nights in 2008 with LST >5 °C	1.01	Elevation (from the DTM model)	1.52	Days in 2006 with LST >12 °C	0.79
May precipitation	1.56	June precipitation	1.64	Nighttime LST, amplitude 3	0.99	Nighttime LST, mean	1.44	Day LST, phase 1	0.73

DTM: digital terrain model; EVI: enhanced vegetation index; LST: land surface temperature; NDVI: normalized difference vegetation index.

Table 5. Pearson correlation between the predicted probability of occurrence and $log_{10}(n+1)$-observed abundance of *Culicoides* spp. in Spain.

Species	Correlation[a]
C. imicola	0.06
C. newsteadii	0.28
C. pulicaris	0.97
C. punctatus	0.39
C. obsoletus	0.53

[a]Coefficient of determination.

found to be temperature-related. The timing of the annual and tri-annual peak of vegetation was also important but precipitation does not have an influence on the distribution modelling for this species group.

Development of abundance models

In Table 5 the correlation is given between the predicted probability of occurrence and the $log_{10}(n+1)$-observed abundance. This relationship is shown in Fig. 3. Whilst all reported correlation coefficients were highly significant, it is clear that correlation is mainly achieved because of the good match between negative traps and zero probabilities of occurrence. When removing the zero observations, no correlation was observed (not shown here). A simple linear relationship will thus not allow modelling the abundance of the different *Culicoides* species.

The output of abundance RF models is shown in Fig. 4. *C. imicola* is most abundant (from >200 individuals up to 2,200 individuals) in Extremadura (central Spain). Andalucía in south-western Spain featured a medium abundance, while other regions had abundance lower than 100 specimens. Very low densities are predicted in northern Spain, concurrent with the observed data. *C. newsteadii* and *C. punctatus* were most abundant on the Mediterranean coast and in central to northern Spain. The Pyrénées are predicted to have a very low abundance of this species. The *C. pulicaris* group, and also the *C. obsoletus* group, is predicted to occur very abundantly in the entire country with the exception of the driest areas, i.e. Extremadura and Andalucía. A visual comparison of the predicted abundance map of *C. obsoletus* (Fig. 4) and the observed data (Fig. 1) suggests a considerable underestimation of abundance in the overlapping area with *C. imicola*, whilst higher abundances were correctly predicted in the northern part of Spain. This is also the case for *C. pulicaris*, though to a lesser extent. Results for species with distinct presence/absence

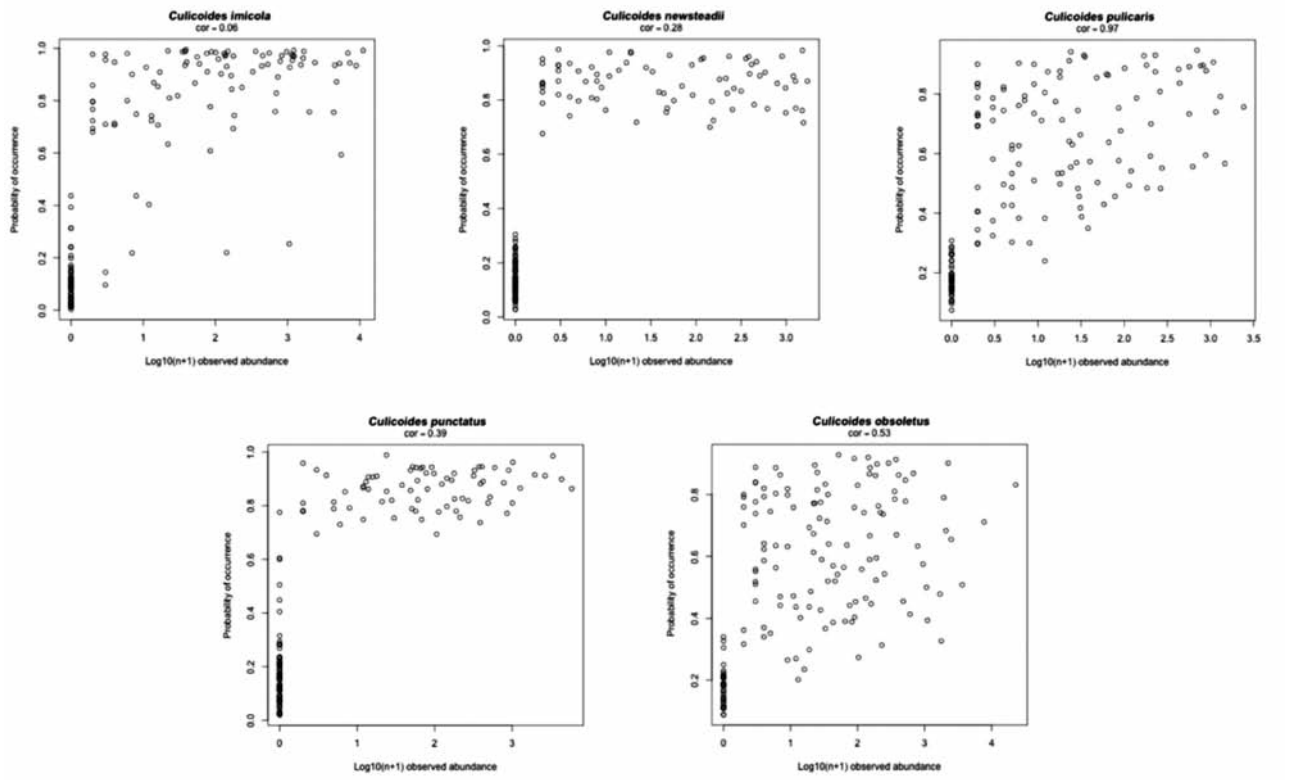

Fig. 3. The relation between the $log_{10}(n+1)$ observed abundance and the predicted probability of occurrence of *Culicoides* spp. in Spain.

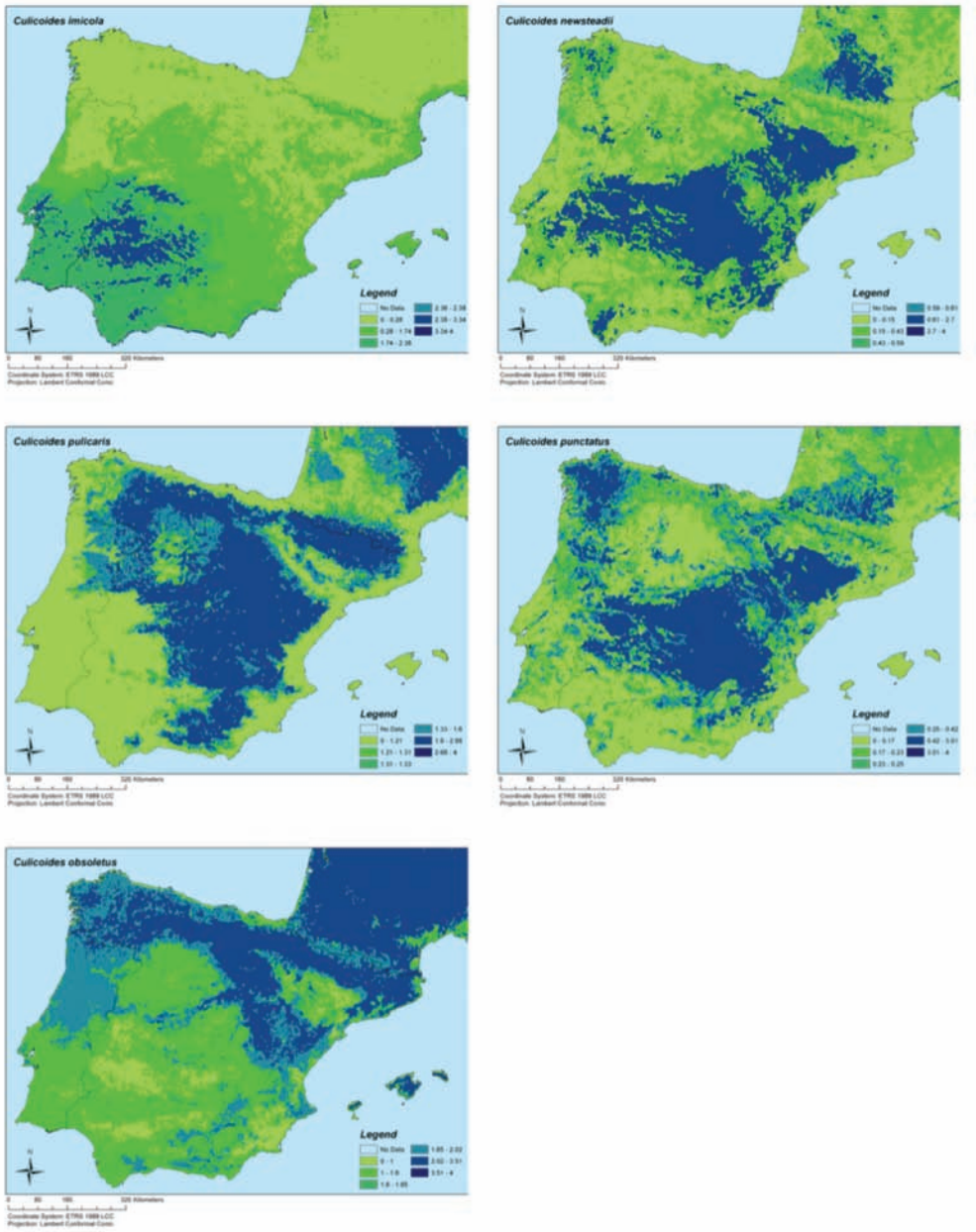

Fig. 4. Predicted abundance of *Culicoides* spp. in Spain (abundance is represented as $log_{10}(n+1)$ where n = number of individuals).

areas, i.e. *C. imicola*, *C. newsteadii* and *C. punctatus*, showed a more crisp distinction between high and low abundance areas.

The interpretation of predictions obtained outside the study area (i.e. Spain) reflects a similar pattern. In southern France, the absence of *C. imicola* and the presence of *C. obsoletus* group and *C. pulicaris* were correctly predicted as compared to the results from the Epizone-Dynvect project (Balenghien et al., 2010). Results obtained in Portugal, however, were less clear: *C. imicola* was correctly predicted, *C. obsoletus* to a lesser extent so, while the discriminating power between absence and presence zones was less for *C. pulicaris*. Results obtained with *C. newsteadii* and *C. punctatus* in these countries could not be compared since no data were available to us.

In Table 6 the MSE of the abundance model for each species is given. In case of *C. imicola*, the MSE on the $log_{10}(n+1)$ data amounts to 0.73. This is even lower for the other species, which is also reflected in the determination coefficient. Fig. 5 shows the relationship between the observed and predicted abundance data.

Table 7 lists the 10 most important predictor variables together with their variable importance for the abundance modelling of the different *Culicoides*

Table 6. The mean of squared error (MSE) and Pearson correlation coefficient with intercept and regression coefficient between the observed and predicted data for *Culicoides* spp. in Spain.

Species	MSE	Cor[a]	Intercept	Coefficient
C. imicola	0.81	0.76	0.85	0.88
C. newsteadii	0.75	0.72	0.78	0.81
C. pulicaris	0.62	0.59	0.65	0.58
C. punctatus	0.70	0.69	0.72	0.80
C. obsoletus	0.51	0.47	0.56	0.59

[a]Coefficient of determination.

species. For all species the most important predictor variables determining the abundance is the probability of occurrence. This is concurrent with the previously identified correlation between the predicted probability and the observed $log_{10}(n+1)$-transformed abundance data. The probability can therefore be considered a necessary condition for the abundance, and the additional predictor variables will define the level of abundance within the presence zone. For *C. imicola*, *C. newsteadii* and *C. pulicaris* these predictor variables are similar to those found in the probability mapping. Abundance of *C. imicola* is mostly driven by summer precipitation, abundance of *C. newsteadii* by temperature and precipitation and *C. pulicaris* by number of days below the 5 °C threshold and number of days with temperatures above freezing. *C. punctatus* abundance seems to be determined more by temperature variables (five variables), albeit different variables than be found in the probability of occurrence model. Additionally, precipitation in June, population density and the peak of the annual vegetation cycle influence the abundance. *C. obsoletus* abundance is relative to summer precipitation as well as the mean yearly daytime temperature and the peak of annual cycle of the temperature at night. Summer precipitation did not influence the probability of occurrence.

Discussion

Aggregation of time series

The Spanish dataset originates from a network of traps, which have been sampled on a regular basis during one year. The available aggregated data (i.e. maximum catch per month) allow reducing the risk of false negative trap sites and enable the use of annual maximum catch figures per trap site as a measure of abundance leading to improved apparent density estimates. It is important to note that this apparent density is not the same as the true population density, because the

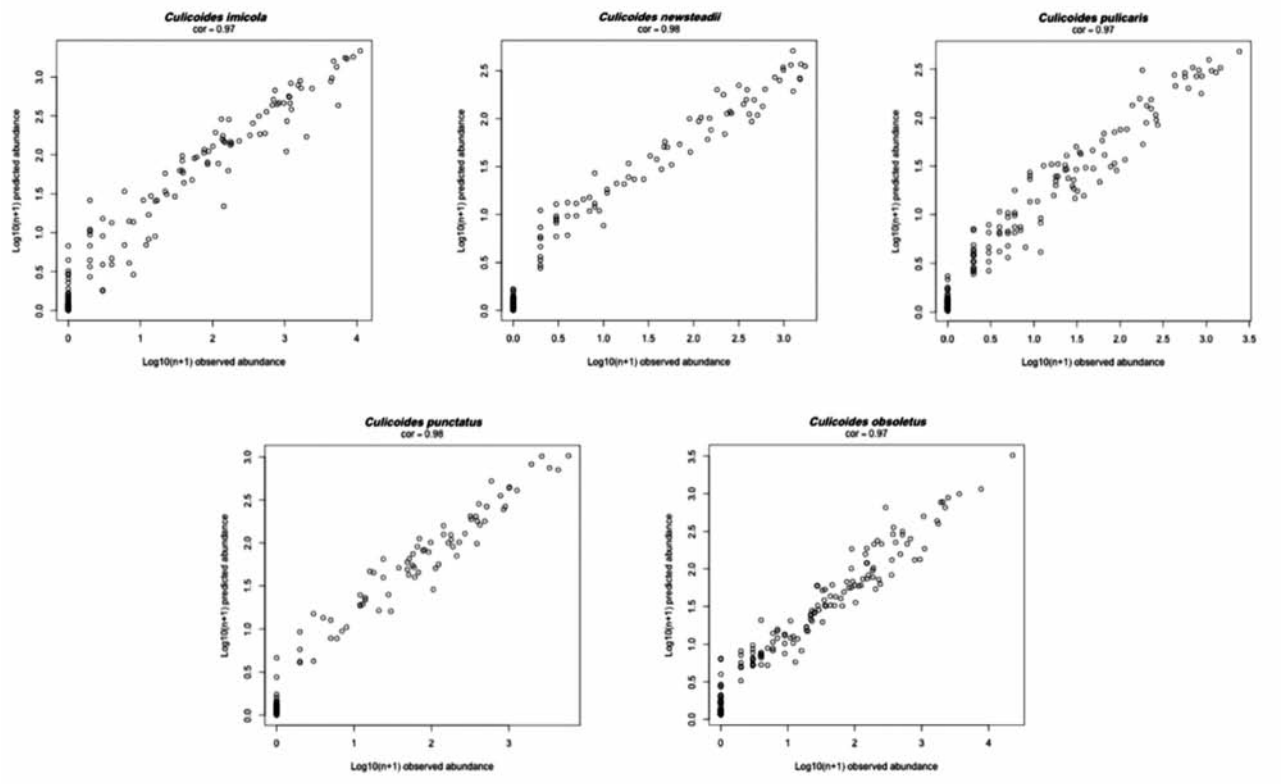

Fig. 5. The relation between the observed abundance of *Culicoides* spp. in Spain and the predicted abundance (represented as $log_{10}(n+1)$ where n = number of individuals).

Table 7. The 10 most important predictors indicated by the variable importance (VARIMP) for the probability of occurrence of the *C. imicola*, *C. newsteadii*, *C. pulicaris* group, *C. punctatus* and *C. obsoletus* groups.

C. imicola		*C. newsteadii*		*C. pulicaris* group		*C. punctatus*		*C. obsoletus* group	
Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP	Predictor	VARIMP
Probability of occurrence	50.88	Probability of occurrence	67.80	Probability of occurrence	39.15	Probability of occurrence	86.72	Probability of occurrence	44.62
Minimum precipitation	29.18	Nighttime LST, amplitude 1	10.32	Elevation (from the DTM model)	10.58	Nighttime LST, amplitude 1	9.39	August precipitation	7.55
July precipitation	27.66	Population density	6.94	Nights in 2007 with LST >5 °C	8.67	Probability of occurrence	4.78	September precipitation	4.71
September precipitation	18.61	Distance to water	4.14	No. of days in 2008 with temp. >0 °C	6.29	Nighttime LST, phase 1	4.71	Minimum precipitation	3.90
August precipitation	15.17	Total precipitation	4.04	No. of nights in 2006 with LST >5 °C	4.23	No. of days in 2007 with temp. > 0°C	3.36	Daytime LST, amplitude 0	3.86
June precipitation	9.88	December precipitation	3.19	No. of nights in 2008 with LST >5 °C	3.90	Nighttime LST, phase 3	3.26	July precipitation	3.51
Daytime LST, phase 2	5.85	NDVI phase 3	2.78	Nighttime LST, amplitude 2	3.07	June precipitation	3.20	NDVI, phase 2	3.47
Daytime LST, phase 1	3.63	Nighttime LST, phase 1	2.78	Distance to water	2.20	Total precipitation	2.98	Nighttime LST, amplitude 1	3.42
May precipitation	2.87	Nighttime LST, amplitude 3	2.45	No. of days in 2006 with temp. >0 °C	2.04	EVI, amplitude 0	2.75	Total precipitation	3.32
Nighttime LST, phase 3	2.82	Nighttime LST, amplitude 3	2.26	EVI, amplitude 2	2.00	No. of days in 2007 with LST >12 °C	2.29	EVI, phase 3	3.29

DTM: digital terrain model; EVI: enhanced vegetation index; LST: land surface temperature; NDVI: normalized difference vegetation index.

apparent density is dependent on a wide range of factors such as trap type, location type, presence of hosts nearby and the weather. The data did not allow correcting for these potential biases.

Comparing abundance data

It is possible to directly use abundance data as input to the model because the abundance was measured using a standardised protocol (trap type, trap frequency, morphological identification). A wide range of trapping techniques is in use for the collection of adult mosquitoes (Kline, 2006). They differ in design and vary greatly in effectiveness and usefulness (Campbell, 2003). Comparing the performance of three trap mechanisms (using the "CO_2-baited mosquito magnet liberty plus trap", the "BG sentinel trap" and the "gravid trap"), Versteirt (2012) noted that the species caught and the abundances per species differed greatly between the different traps and thus that abundance values are not comparable. This leaves obviously the issue that the proposed methodology cannot be directly applied to continental-wide mapping if the observed data are not uniformly collected and illustrates that it is essential to design a sampling strategy tailored to the study objectives prior to the field work. When working with historical datasets from different sources and collected in different ways, methods will have to be designed to compensate for discrepancies.

Predictors for Culicoides *probability of occurrence and abundance*

In this study, *Culicoides* species with a clear distinction between areas of presence and absence scored considerably better in models predicting the probability of occurrence than species with a wider distribution range. This suggests that at a pixel resolution of 5 x 5 km for the latter (i.e. *C. pulicaris* and the *C. obsoletus* group) may not contain enough information to discriminate between unsuitable regions in Spain. Additional predictor data will therefore be needed to improve the discriminating capacity of the models. This is also reflected when analysing the correlation between the predicted probability of occurrence and the observed abundance. While the RF approach does not allow investigating the direction of influence, it does permit assessing the importance of each variable. Future research with techniques such as boosted regression trees (Elith et al., 2009) may overcome this disadvantage.

Precipitation, especially summer rainfall (June-September), was the most influential factor determining the distribution of *C. imicola*. Field observations indicate that the population of *C. imicola* peaks in the September-October period and it seems that the summer rainfall has a direct impact on the population. This is concurrent with the observations from Calvete et al. (2008, 2009), who noted that the coefficient of variation and the total amount of precipitation was retained in all logistic regression models for *C. imicola*. In contrast, models produced by Wittmann et al. (2001) did not include any variable related to rainfall.

Our research shows that temperature and temperature variation, expressed either through Fourier-transformed variables or through a coefficient of variation, are also important. Work by Wittmann et al. (2001), Purse et al. (2004) and Calvete et al. (2008, 2009) confirms this relationship. In the case of the *C. obsoletus* group, the main factors were related to temperature (seven out of the 10 most important factors) and vegetation. The population of this species group peaks in summer time. Most of the temperature variables are related to days colder than a 0 °C (which has to do with overwintering), 5 °C or 12 °C (temperature when insect activity starts) respectively, but the first phase of land surface daytime temperature is also included in the model. This is in agreement with the findings of Calvete et al. (2008), which indicate that the mean temperature and the coefficient of variation were the significant factors (in addition to the mean NDVI).

Models for *C. newsteadii*, *C. pulicaris* and *C. punctatus* are sparse and little information on the driving factors for these species is known. From these results it is concluded that the developed methodology enables to produce sufficiently accurate abundance models for *Culicoides* species in Spain. Whilst it is clear that the approach may still be improved, it still provides a good basis for further work.

Using probability of occurrence as a predictor for abundance

In all the abundance models, the probability of occurrence is the most important factor. This factor was added as a predictor given that many authors indicated that there is a (non)-linear relationship between probability of occurrence and abundance. In our results, the linear relationship proved to be too weak to create abundance maps directly from the probability of occurrence maps and therefore the probability of presence was added into the set of predictors for the abundance model using random regression forest. While this would introduce correlation in traditional statistical modelling techniques, and thus may cause considerable variable inflation, this is allowed when using the RF technique, which is not affected by the predictor variables correlation. For all species, the RF model delineates the zones of probability of presence (the necessary condition), and within that zone the level of abundance is determined by the remaining predictor variables.

Conclusion

The methodology for the creation of abundance models, and its application for multiple *Culicoides* species in Spain and Portugal using RF, is shown. The results indicate that this modelling approach is robust and that the predictor variables that are retained by the model are concurrent with existing studies for the mapping of probability of occurrence. The development of abundance models using a continuous output has not been attempted before and it is shown here that using a combination of probability of occurrence maps and a set of dedicated predictor variables, an accurate output can be obtained and used as input to TIR or R_0 models.

Acknowledgements

This work was sponsored by the EU network of Excellence, EPIZONE (contract nr FOOD-CT-2006 016236) and the out-

come of two internal call projects (IC 6.6 BT EPIDEMIOLGY and IC 6.7 BT-DYNVECT). Additional funding was provided by Central Veterinary Institute (The Netherlands) under contract BO-08-010-021.

References

Allepuz A, Garcia-Bocanegra I, Napp S, Casal J, Arenas A, Saez M, Gonzalez MA, 2010. Monitoring bluetongue disease (BTV-1) epidemic in southern Spain during 2007. Prev Vet Med 96, 263-271.

Balenghien T, Bodker R, Kiel E, De Deken R, Chirico J, Lucientes J, Carpenter S, Elbers ARW, Calistri P, Miranda M, Staubach C, Van der Stede Y, Guis H, 2010. Dynvect's overview of the *Culicoides* surveillance systems in the EU and distribution maps of key species. In: 4th Annual Meeting EPIZONE, Saint-Malo, France, 7-10 June 2010, Saint-Malo, France.

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA, 2002. Evaluating resource selection functions. Ecol Model 157, 281-300.

Breiman L, 2001. Random Forests. Mach Learn 45, 5-32.

Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J, 2008. Modelling the distributions and spatial coincidence of bluetongue vectors *Culicoides imicola* and the *Culicoides obsoletus* group throughout the Iberian Peninsula. Med Vet Entomol 22, 124-134.

Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J, 2009. Ecological correlates of bluetongue virus in Spain: predicted spatial occurrence and its relationship with the observed abundance of potential *Culicoides* spp. vector. Vet J 182, 235-243.

Campbell CB, 2003. Evaluation of five mosquito traps and a horse for West Nile vectors on a north Florida equine facility (MS Thesis). University of Florida, Gainesville.

Chefaoui RM, Horal J, Lobo JM, 2005. Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. Biol Conserv 122, 327-338.

Conte A, Gilbert M, Goffredo M, 2009. Eight years of *Culicoides imicola* survey in Italy: can we detect range expansion? J Appl Ecol 46, 1332-1339.

Conte A, Goffredo M, Ippoliti C, Meiswinkel R, 2007. Influence of biotic and abiotic factors on the distribution and abundance of *Culicoides imicola* and the *Obsoletus* complex in Italy. Vet Parasitol 150, 333-344.

Cutler R, Edwards TC, Bear KH, Cutler A, Hess KT, Gibson J, Lawler JJ, 2007. Random Forests for classification in ecology. Ecology 88, 2783-2792.

ECDC, 2009. Development of *Aedes albopictus* risk maps. ECDC, Stockholm, 52 p.

Elbers ARW, Backx A, Mintiens K, Gerbier G, Staubach C, Hendrickx G, van der Spek A, 2008. Field observations during the bluetongue serotype 8 epidemic in 2006 II. Morbidity and mortality rate, case fatality and clinical recovery in sheep and cattle in the Netherlands. Prev Vet Med 87, 31-40.

Elith J, Leathwick JR, Hastie T, 2009. A working guide to boosted regression trees. J Anim Ecol 77, 802-813.

ESRI, 2009. ArcGIS 9.3 software.

Fielding AH, Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24, 38-49.

Gao X, Huete AR, Ni W, Miura T, 2000. Optical-biophysical relatinships of vegetation spectra without background contamination. Remote Sens Environ 74, 609-620.

GfK, 2009. Available at: http://www.gfk-geomarketing.de/ (accessed on February 2011).

Gibson LA, Wilson BA, Cahill DM, Hill J, 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. J Appl Ecol 41, 213-223.

Guis H, Caminade C, Calvete C, Morse AP, Tran A, Baylis M, 2011. Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. J R Soc Interface 9, 339-350.

Hartemink NA, Purse BV, Meiswinkel R, Brown HE, de Koeijer A, Elbers ARW, Boender GJ, Rogers DJ, Heesterbeek JAP, 2009. Mapping the basic reproduction number (R_0) for vector-borne diseases: a case study on bluetongue virus. Epidemics 1, 153-161.

Hartemink NA, Vanwambeke SO, Heesterbeek H, Rogers D, Morley D, Pesson B, Davies C, Mahamdallie S, Ready P, 2011. Integrated mapping of establishment risk for emerging vector-borne infections: a case study of canine leishmaniasis in south-west France. PLoS One 6, e20817.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A, 2005. WORLDCLIM: Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25, 1965-1978.

JRC, 2005. IMAGE2000 and CLC2000, products and Methods. Available at: http://ies.jrc.cec.eu.int/ (accessed on February 2011).

JRC, 2009. European soil database. Available at: http://eusoils.jrc.ec.europa.eu/ESDB_Archive/ESDB/index.htm (accessed on February 2011).

Kline DI, Patnaude M, Barnard DR, 2006. Efficacy of four trap types for detecting and monitoring *Culex* spp. in North Central Florida. J Med Entomol 43, 1121-1128.

Le Gal MC, Dufour B, Geoffroy E, Zanella G, Rieffel JN, Pouilly F, Moutou F, 2008. La fièvre catarrhale ovine due au sérotype 8 dans les Ardennes françaises en 2007: taux de morbidité, mortalité, létalité et signes cliniques observeés chez les bovines et les ovins. Ann Med Vet 152, 240-245.

Mardulyn P, Goffredo M, Conte A, Hendrickx G, Meiswinkel R, Balenghien T, Sghaier S, Lohr Y, Gilbert M, 2013. Climate change and the spread of vector-borne diseases: using approximate Bayesian computation to compare invasion scenarios for

the bluetongue virus vector *Culicoides imicola* in Italy. Mol Ecol in press.

McPherson JM, Jetz W, Rogers DJ, 2004. The effect of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artifact? J Appl Ecol 41, 811-823.

OIE, 2007a. Bluetongue Spain (immediate notification) 26/07/2007. Paris: World Organisation for Animal Health.

OIE, 2007b. Bluetongue Portugal (immediate notification) 21/09/2007. Paris: World Organisation for Animal Health.

OIE, 2007C. Bluetongue France (immediate notification) 23/11/2007. Paris: World Organisation for Animal Health.

Osborne PE, Alonso JC, Bryant RG, 2001. Modelling landscape-scale habitat using GIS and remote sensing: a case study with great bustards. J Appl Ecol 38, 458-471.

Peters J, De Baets B, Van doninck J, Calvete C, Lucientes J, De Clercq E, Ducheyne E, Verhoest NEC, 2011. Absence reduction in entolomological studies surveillance data to improve niche-based distribution models for *Culicoides imicola*. Prev Vet Med 100, 15-28.

Peters J, De Baets B, Verhoest N, Samson R, Degroeve S, De Becker P, Huybrechts W, 2007. Random Forests as a tool for ecohydrological distribution modelling. Ecol Model 207, 304-318.

Purse BV, Brown HE, Harrup L, Mertens PPC, Rogers DJ, 2008. Invasion of bluetongue and other orbivirus infections into Europe: the role of biological and climatic processes. Rev Sci Tech Oie 27, 427-442.

Purse BV, Tatem AJ, Caracappa S, Rogers DJ, Mellor PS, Baylis M, Torina A, 2004. Modelling the distributions of *Culicoides* bluetongue virus vectors in Sicily in relation to satellite-derived variables. Med Vet Entomol 18, 90-101.

R Development Core Team, 2006. R stats v2.10.1.

Ruiz MO, Chaves LF, Hamer GL, Sun T, Brown WM, Walker ED, Haramis L, Goldberg TL, Kitron U, 2010. Local impact of temperature and precipitation on West Nile virus infection in *Culex* species mosquitoes in northeast Illinois, USA. Parasit Vectors 3.

Saegerman C, Berkvens D, Mellor PS, 2008. Bluetongue epidemiology in the European Union. Emerg Infect Dis 14, 539-544.

Scharlemann JPW, Benz D, Hay S, Purse BV, Tatem A, Wint W, Rogers DJ, 2008. A novel algorithm for temporal Fourier processing MODIS data for ecological and epidemiological applications. PLoS One 3, e1408.

Szmaragd C, Wilson AJ, Carpenter S, Wood JLN, Mellor PS, 2009. A modeling framework to describe the transmission of bluetongue virus within and between farms in Great Britain. PLoS One 4, e7741.

Szmaragd C, Wilson AJ, Carpenter S, Wood JLN, Mellor PS, 2010. The spread of bluetongue virus serotype 8 in Great Britain and Its Control by Vaccination. PLoS One 5, e9353.

Tatem AJ, Baylis M, Mellor PS, Purse BV, Capela R, Pena I, Rogers DJ, 2003. Prediction of bluetongue vector distribution in Europe and North Africa using satellite imagery. Vet Microbiol 97, 13-29.

USGS, 1996. Available at: GTOP030,http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30_info (accessed on February 2011).

Velthuis AGJ, Saatkamp HW, Mourits MCM, de Koeijer AA, Elbers ARW, 2010. Financial consequences of the Dutch bluetongue serotype 8 epidemics of 2006 and 2007. Prev Vet Med 93, 294-304.

Versteirt V, 2012. Taxonomic and functional biodiversity of indigenous and exotic mosquito species (Culicidae) in Belgium (PhD Thesis). University of Antwerp, Belgium.

Wint W, 2005. Global World Population v3.0 compiled for EDEN DMT. Available at: http://ergodd.zoo.ox.ac.uk/eden/ (accessed on February 2011).

Wittmann EJ, Mellor PS, Baylis M, 2001. Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe. Rev Sci Tech Oie 20, 731-740.