Ecological Informatics 22 (2014) 69–80

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Predicting spatio-temporal *Culicoides imicola* distributions in Spain based on environmental habitat characteristics and species dispersal

J. Peters [a], W. Waegeman [a], J. Van doninck [b], E. Ducheyne [c], C. Calvete [d], J. Lucientes [e], N.E.C. Verhoest [b], B. De Baets [a,*]

[a] *KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Gent, Belgium*
[b] *Laboratory of Hydrology and Water Management, Department of Forest and Water Management, Ghent University, Coupure links 653, 9000 Gent, Belgium*
[c] *Avia-GIS, Risschotlei 33, 2980 Zoersel, Belgium*
[d] *Unidad de Sanidad y Producción Animal, Centro de Investigación y Tecnología, Agroalimentaria (CITA), Gobierno de Aragón, 50059 Zaragoza, Spain*
[e] *Departamento de Patología Animal, Universidad de Zaragoza, 50071 Zaragoza, Spain*

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 14 May 2014
Accepted 19 May 2014
Available online 27 May 2014

Keywords:
Species dispersal
Species distribution model (SDM)
Random Forests
Model testing
Variable importance
MODIS
Bluetongue
Veterinary epidemiology

ABSTRACT

The use of distance variables expressing the likelihood of species occurrence at a given site in relation to the distance to observed species presence is demonstrated to improve species distribution models, especially when combined with environmental variables which relate species occurrence to the environmental habitat characteristics. In this study we developed models to predict the spatio-temporal distribution of *Culicoides imicola*, which is the main transmission vector for the bluetongue virus in the Mediterranean region. We investigated (i) the importance of the environmental habitat characterization by means of bioclimatic variables, (ii) the effect of different distance variables to model the dispersal process, and (iii) the suitability of two different parameter identification procedures to determine the distance variables for species distribution modeling. Results showed that niche-based species distribution models, which only use environmental data, could estimate the occurrence of *Culicoides imicola* accurately, given that environmental data of the period of high species abundance (April until October) was included. The use of these models may therefore be hampered for predictive risk assessment aiming to estimate the probabilities and magnitude of undesired effects caused by the occurrence of *C. imicola*. Species distribution models accounting for species dispersal in addition to the environmental habitat characteristics, i.e. hybrid models, did provide accurate predictions of *C. imicola* distributions well before the onset of the season of high species abundance. A Gaussian or negative exponential function of the distance to presence locations was most suitable to model insect dispersal. The enhanced predictive capacity of these models potentially leads towards an increased model applicability in risk assessment and disease control.

1. Introduction

Niche-based species distribution models (SDMs) determine the occurrence of a species based on its habitat requirements. The formulation of a SDM is generally empirical, and based on geo-referenced species occurrence data, presence only or absence–presence, that are projected on environmental data layers. The backbone of an SDM is a formal description of the relationship between the environmental site characteristics (explanatory variables) and the species occurrence (dependent variable). Several techniques have been applied to formalize this relationship, among which generalized linear models (e.g. logistic regression, Hosmer and Lemeshow (2000)), linear and non-linear discriminant analysis (Rogers et al., 1996), tree-like classification algorithms (e.g. CART, Breiman et al. (1984) and Random Forests, Breiman (2001)), and maximum entropy models (e.g. MAXENT, Phillips et al. (2006)) are the most popular (see Elith et al. (2006) for a comparison of techniques). SDM calibration is performed on species distribution records which reflect the environmental habitat suitability and the species dispersal, colonization and survival processes that are involved in the establishment of viable populations in these habitats. The range of environments and biotic processes captured by the calibration data depends on the species survey extent (Anderson and Martínez-Meyer, 2004), where more extensive surveys lead to a tighter representation of the realized niche (sensu Hutchinson (1957)), and more accurate model predictions of the actual species distribution. However, since the biological processes involved in population establishment are generally not included explicitly as explanatory variables in SDMs, model predictions are not restricted to the actual species distribution. Those parts of the potential distribution that intersect the environmental range of the calibration data will be qualified as suitable habitat as well. Therefore, SDMs may be used to model the potential geographical extent of invasive species in novel landscapes (Peterson, 2003; Peterson et al., 2003; Sutherst and Bourne, 2009) and to model changes in

---

* Corresponding author. Tel.: +32 9 264 59 41; fax: +32 9 264 62 20.
E-mail address: bernard.debaets@ugent.be (B. De Baets).

http://dx.doi.org/10.1016/j.ecoinf.2014.05.006
1574-9541/© 2014 Elsevier B.V. All rights reserved.

geographical species ranges under changing environmental conditions (Berry et al., 2002; Engler et al., 2009; Pearson, 2006; Thomas et al., 2004).

The importance of spatial autocorrelation in the geographical distribution of species has been demonstrated from ecological data (Legendre and Fortin, 1989). Typically, species occurrence is positively autocorrelated, such that the occurrence at nearby locations tends to be more similar than would be expected by chance (Lichstein et al., 2002). This pattern is often driven by multiple causes that may be exogenous (e.g. autocorrelated environment or disturbance) and/or endogenous (e.g. conspecific attraction, dispersal limitation, demography) (Sokal and Oden (1978); Legendre (1993), cited in Lichstein et al. (2002)). Methodologies have been proposed to incorporate spatial autocorrelation in SDMs. Augustin et al. (1996) developed an autologistic regression model as an extension of a logistic regression model that includes a distance-weighted function derived from neighborhood responses as an extra covariate. Other methods, including spatial models based on eigenvector mapping (Dray et al., 2006), generalized least square and autoregressive models, and generalized estimating equations (Liang and Zeger, 1986), are reviewed and compared by Dormann et al. (2007) for species distribution mapping. Hengl et al. (2009) combined spatial techniques with environmental factor analysis (ENFA) to obtain spatial prediction maps of species occurrence.

Recent studies added distance variables, sometimes referred to as dispersal or distance constraints (Václavík and Meentemeyer, 2009), expressing the likelihood of a species' occurrence at a site in relation to the distance of observed species presence (Allouche et al., 2008), as explanatory variables, and showed an improved prediction accuracy and ecological understanding of distribution models (Allouche et al., 2008; Meentemeyer et al., 2008; Václavík and Meentemeyer, 2009). These models are called hybrid models because they combine the niche-based approach with additional distance variables. The use of distance variables has the benefit of being suitable for presence-only models. These variables are consequently often applied for distribution modeling of invasive species (iSDM) in novel landscapes, where its absence does not necessarily indicate unsuitable habitat conditions but could also result from dispersal limitations (Václavík and Meentemeyer, 2009). A similar approach is adopted to model the spread of pathogens in epidemiological literature. Boender et al. (2007) and Szmaragd et al. (2009) developed models for the transmission of avian influenza and bluetongue, respectively, including transmission kernels to describe the spread of the virus. Mintiens et al. (2003) used kernel estimation of the intensity of pig herds in a neighborhood as a risk factor for the transmission of the CSF virus.

Niche-based and hybrid models are increasingly applied in epidemiological research, where they are frequently used to assess the spatial spread of pathogen vectors. Distribution models based on Random Forests have been shown to be highly suitable for these applications, as they are less affected by false absences (i.e. catch failures) in the training data (Peters et al., 2011). In veterinary epidemiology, the transmission vector for the bluetongue virus (BTV), *Culicoides imicola* Kieffer (Diptera: Ceratopogonidae), has thereby received substantial research attention. In the Mediterranean basin, it is considered that the geographical distribution of *C. imicola* is constrained by bioclimate. Several studies demonstrated the ability to predict its presence and/or abundance based on bioclimatic information acquired from both meteorological stations and satellite imagery (e.g. Baylis et al., 2001; Calistri et al., 2003; Calvete et al., 2008; Conte et al., 2007; Purse et al., 2004; Tatem et al., 2003; Wittmann et al., 2001). The strong relationship between bio-climate and the geographical distribution of *C. imicola* resulted in a northward range expansion from source points in Northern Africa during the last two decades linked to climate change (Mardulyn et al., 2013). The *Culicoides* dispersal models developed by Ducheyne et al. (2007) and Hendrickx et al. (2008) further analyzed the role of wind in the dispersion pattern of different *Culicoides* species, including *C. imicola* and indigenous European species (primarily the *Culicoides obsoletus* complex and the *Culicoides pulicaris* complex). Prevailing winds were linked to the medium/long-distance spread while short range spread was mainly driven by active *Culicoides* flight. The dispersal pattern is thus driven by different factors: longer distance (up to several hundreds of kilometers) migratory movements are mediated by winds, in contrast to the active short distance movements towards resources, both up- and downwind. These dispersal mechanisms are explained in Sellers (1992) and Reynolds et al. (2006).

The four research questions addressed in this study are:

1. Does the model performance change when the environmental habitat characteristics are determined by environmental data covering longer periods?;
2. Is the model performance improved when accounting for species dispersal by empirical distance variables?;
3. What is the effect of the identification procedure to parameterize these distance variables?; and
4. Which distance variable(s) is/are most appropriate to describe the dispersal process of *C. imicola*?

2. Material and methods

2.1. Study site and species data collection

A year-round collection of *Culicoides* spp. in the period 2004–2008 was carried out on mainland Spain and the Balearic Islands under the Spanish Bluetongue National Surveillance Programme (for details, see Calvete et al. (2006)). Throughout this region, *Culicoides* spp. were caught using 4 W ultraviolet light traps, fitted with a suction fan and a collection vessel containing ethanol and ethylene glycol in water to preserve the samples (miniature blacklight model 1212, J.W. Hock Co., Gainesville, FL, U.S.A.). The traps were installed outside selected farms with a minimum of 10 large livestock animals, not further than 30 m away from livestock. The coordinates of the sample locations were recorded by a hand-held GPS receiver, and samples were taken during one or two consecutive nights. Catches were analyzed on *Culicoides* spp. abundance. For this study, however, we focused on *C. imicola* occurrence exclusively, and the abundance data were transformed into absence–presence data (zero and non-zero catches). Given the high seasonality in *C. imicola* activity and abundance (Miranda et al., 2004; Ortega et al., 1998), the data set was further reduced by only selecting data obtained between April and October, which is the period of maximal *C. imicola* activity and abundance. As such, a data set containing *C. imicola* absence–presence records at 284 (year 2004), 382 (year 2005), 885 (year 2006), 151 (year 2007) and 71 (year 2008) farms was compiled.

2.2. Environmental variables

The environmental characterization of the *Culicoides* sampling locations was derived from multispectral optical and thermal satellite imagery from MODIS (http://modis.gsfc.nasa.gov) by means of the daytime and nighttime land surface temperature (LST_{day} and LST_{night}), and the normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI). The selection of these environmental variables is based on their demonstrated influence on the geographical *C. imicola* distribution (e.g. Baylis et al., 2001; Peters et al., 2011; Purse et al., 2004, 2007; Tatem et al., 2003) and their availability throughout the study period. The use of the highly correlated NDVI and EVI together as independent variables is not problematic for the modeling technique used in this study which is capable of dealing with highly correlated features by its variable selection mechanism. We decided to use 8 day and 16 day composite images, for LST and the vegetation indices respectively, to reduce the impact of cloud cover. A temporal spline-interpolation to a five-day temporal resolution resulted in 73 images for each year of the five year long study period, that were subsequently averaged to obtain monthly averages throughout the 5 year long study