Big data - modelling of midges in Europa using machine learning techniques and satellite imagery


Published in: NKVet Symposium 2017 - abstract book

Publication date: 2017

Document Version
Peer reviewed version


DTU Library
Technical Information Center of Denmark

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Biting midges (Diptera, Ceratopogonidae) of the genus *Culicoides* are important vectors of pathogens causing diseases in free living and production animals and can lead to large economic losses in many European countries. In Europe, *Culicoides imicola* and the *Obsoletus* group are considered to be the main vectors of bluetongue virus that mostly affects ruminants such as cattle and sheep. Spatio-temporal modelling of vector distribution and abundance allows us to identify high risk areas for virus transmission and can aid in applying effective surveillance and control measures.
We used presence-absence and monthly abundance data of *Culicoides* from 1005 sites across 9 countries (Spain, France, Denmark, Poland, Switzerland, Austria, Poland, Sweden, Norway) collected between the years 2007 and 2013. The dataset included information on the vector species abundance (number of specimens caught per night), GPS coordinates of each trap, start and end dates of trapping. We used 120 environmental predictor variables together with Random Forest machine learning algorithms to predict the overall species distribution (probability of occurrence) and monthly abundance in Europe. We generated maps for every month of the year, to visualize the abundance of *C. imicola* and Obsoletus group in Europe as well as distribution maps showing the probability of occurrence.

We were able to create predictive maps of both *Culicoides* sp. occurrence and abundance using Random Forest models, and although the variance was large, the predicted abundance values for each site had a positive correlation with the observed abundance. We found relatively large spatial variations in probability of occurrence and abundance for both *C. imicola* and the Obsoletus group. For *C. imicola* probability of occurrence and abundance was higher in southern Spain, where as the Obsoletus group had higher probability of occurrence and abundance in central and northern Europe such as France and Germany. Temporal variation was also observed with higher abundance occurring during summer months and low or no abundance during winter months for both *C. imicula* and the Obsoletus group, although abundance was generally higher for a longer period of time for *C. imicula* than for the Obsoletus group.

Using machine learning techniques, we were able to model the spatial distribution in Europe for *C. imicola* and the Obsoletus group in terms of abundance and suitability (probability of occurrence). Our maps corresponded well with the previously reported distribution for *C. imicola* and the Obsoletus group. The observed seasonal variation was also consistent with reported population dynamics for *Culicoides*, as it depends on environmental factors such as temperature and rainfall. Longer seasonal abundance for *C. imicula* compared to the Obsoletus group can be explained by the species distribution, as *C. imicula* is limited to the southern parts of Europe where the warm season lasts longer, whereas the Obsoletus group is found further north. The outputs obtained here will be used as input for epidemiological models and can be helpful for determining high risk areas for disease transmission.