Geographic potential of the world’s largest hornet, *Vespa mandarinia* Smith (Hymenoptera: Vespidae), worldwide and particularly in North America

– ODMAP Protocol –

Claudia Nuñez-Penichet, Luis Osorio-Olvera, Victor Gonzalez, Marlon E. Cobos, Laura Jiménez, Devon A. DeRaad, Abdelghafar Alkishe, Rusby G. Contreras-Díaz, Angela Nava-Bolaños, Kaera Utsumi, Uzma Ashraf, Adeola Adeboje, A. Townsend Peterson, Jorge Soberón

2020-08-10

## Overview

#### Authorship

Contact : jsoberon@ku.edu

<Study link>

#### Model objective

Model objective: Inference and explanation

#### Focal Taxon

Focal Taxon: *Vespa mandarinia*

#### Location

Location: North America, Southeast Asia, world

#### Scale of Analysis

Spatial extent: -180, 180, -90, 83.623596 (xmin, xmax, ymin, ymax)

Spatial resolution: 10’ resolution (~18 km at the Equator)

Temporal extent: 1990-2020

Temporal resolution: N/A

Boundary: natural

#### Biodiversity data

Observation type: citizen science, field survey, GPS tracking, standardised monitoring data

Response data type: presence-only

#### Predictors

Predictor types: climatic

#### Hypotheses

Hypotheses: Dispersal simulations based on invasive records and areas detected to be suitable using ecological niche models allow detection of patterns of potential invasion of *V. mandarinia* in North America.

#### Assumptions

Model assumptions: After data cleaning and thinning, species data are free of bias. Predictors are free of bias after initial steps of selection.

#### Algorithms

Modelling techniques: Maxent

Model complexity: A process of variable selection (based on correlation and biological relevance) and principal component analysis was used to control the complexity of the environmental space, limit the number of dimensions involved, and reduce collinearity.

Model averaging: The medians of distinct replicates of Maxent models created with different parameter settings and sets of variables were used to summarize all results.

#### Workflow

Model workflow: Model calibration was done within a 500 km buffer around species’ occurrences. To consider uncertainty deriving from specific treatments of occurrence records and environmental predictors, we calibrated models via four distinct schemes: (1) raw variables and distance-based thinned occurrences, (2) PCs and distance-based thinned occurrences, (3) raw environmental variables and country-density thinned occurrences, and (4) PCs and country-density thinned occurrences. For each scheme, we calibrated models five times, each time randomly selecting 50% of occurrences for training models, and using the remaining records for testing. We assessed model performance using partial ROC (for statistical significance), omission rates (E = 5%, for predictive ability), and Akaike Information Criterion corrected for small sample sizes (AICc). We selected models with delta AICc ≤2 from among those that were statistically significant and had omission rates below 5%. We created models with the selected parameter values, using all occurrences after the corresponding thinning process, with 10 bootstrap replicates, cloglog output, and model transfers to the world using three types of extrapolation (free extrapolation, extrapolation and clamping, and no extrapolation). As a final evaluation step, we tested whether each replicate of the selected models was able to anticipate the known invasive records of the species in the Americas (British Columbia, Canada; Washington, USA). We created two types of consensus: (1) a median of the medians obtained for each parameterization, and, (2) the sum of all suitable areas derived from binarizing each replicate using a modified least presence (5% omission) threshold. We used the mobility-oriented parity metric (MOP) to detect areas where strict or combinational extrapolation risks could be expected, and we used those areas to trim our binary results. Areas detected as suitable (excluding areas of strict extrapolation) were used as the base for simulation of potential invasion patterns of this hornet in North America using a cellular automaton dynamic model. Simulations were done for results of all four schemes of data processing.

#### Software

Software: Maxent 3.4.1; R 3.6.2 (packages: bam, biosurvey, ellipsenm, kuenm, ntox, raster, rgdal, and rgeos)

Code availability: <https://github.com/townpeterson/vespa>

Data availability: <http://hdl.handle.net/1808/30602>

## Data

#### Biodiversity data

Taxon names: *Vespa mandarinia*

Taxonomic reference system: N/A

Ecological level: species

Data sources: Occurrence data for *V. mandarinia* were downloaded from the Global Biodiversity Information Facility database (GBIF; <https://www.gbif.org/>).

Sampling design: Spatial thinning and maximum density per country thinning

Sample size: 18 (country-density thinned occurrences); 49 (distance-based thinned occurrences)

Clipping: east Asia (native range of distribution of this species)

Scaling: N/A

Cleaning: We kept records from the species’ native range separate from non-native occurrences facilitated by human introduction. We cleaned occurrences from the native distribution following Cobos et al. (2018) by removing duplicates and records with doubtful or missing coordinates. To avoid model overfitting derived from spatial autocorrelation and overdominance of specific regions due to sampling bias, we thinned these records spatially in two ways: by geographic distance and by density of records per country. In the first case (distance-based thinning), we excluded occurrences that were <50 km away from another locality. In the second thinning approach (country-density thinning), we randomly reduced numbers of occurrences in countries with the densest sampling, namely Japan, Taiwan, and South Korea (from 30, 6, and 5, to 6, 2, and 2 occurrences, respectively), to match an approximate reference density of India, Nepal, and China.

Absence data: N/A

Background data: A buffer of 500 km around species records for spatial delimitation of calibration areas. Records from 1990 onwards.

Errors and biases: N/A

#### Predictor variables

Predictor variables: We used two types of predictors: 1) Six raw bioclimatic variables selected based on correlation levels and species natural history criteria: isothermality (BIO3), maximum temperature of warmest month (BIO5), minimum temperature of coldest month (BIO6), temperature annual range (BIO7), specific humidity of most humid month (BIO13), and specific humidity of least humid month (BIO14). 2) The first four principal components (PC) axes of a PCA analysis done with 15 variables (all variables of MERRAclim but Bio8, Bio9, Bio18, and Bio19), as they explained 97.9% of the cumulative variance.

Data sources: MERRAclim database (Vega, Pertierra & Olalla-Tárraga, 2018)

Spatial extent: 72.856955, 148.618463, 18.190443, 48.987058 (xmin, xmax, ymin, ymax)

Spatial resolution: 10’ resolution (~18 km at the Equator)

Coordinate reference system: +proj=longlat +datum=WGS84 +no\_defs

Temporal extent: 2000-2010

Temporal resolution: N/A

Data processing: N/A

Errors and biases: N/A

Dimension reduction: N/A

#### Transfer data

Spatial extent: -180, 180, -90, 83.623596 (xmin, xmax, ymin, ymax)

Spatial resolution: 10’ resolution (~18 km at the Equator)

Temporal extent: 2000-2010

Temporal resolution: N/A

## Model

#### Variable pre-selection

Variable pre-selection: We excluded four of the “bioclimatic” variables because they are known to contain spatial artifacts as a result of combining temperature and humidity information (Escobar et al., 2014): mean temperature of most humid quarter, mean temperature of least humid quarter, specific humidity mean of warmest quarter, and specific humidity mean of coldest quarter. The 15 variables remaining were masked to an area for model calibration. These 15 variables were submitted to a principal component analysis (PCA) to reduce dimensionality and multicollinearity. To select a set of raw variables, we reduced them to a subset with Pearson’s correlation coefficients (*r*) ≤ 0.85, choosing the most biologically relevant or interpretable variables based on our knowledge of AGH natural history. The PCA was calibrated using environmental variation across the M area, and transferred to the whole world.

#### Multicollinearity

Multicollinearity: We did a principal components analysis to reduce multicollinearity. We also used Pearson’s correlation coefficient values to select a subset of non-correlated variables.

#### Model settings

Maxent: Feature Set (eight feature classes (lq, lp, lqp, qp, q, lqpt, lqpth, lqph, where l is linear, q is quadratic, p is product, t is threshold, and h is hinge)), Regularization Multiplier Set (10 regularization multiplier values (0.10, 0.25, 0.50, 0.75, 1, 2, 3, 4, 5, 6)), Preictor Set (All combinations of more than 2 predictors of the first 4 PCs. All combinations of more than 2 variables of the six selected biolimatic predictors.)

#### Model estimates

Coefficients: Median

Parameter uncertainty: N/A

Variable importance: N/A

#### Model selection - model averaging - ensembles

Model selection: N/A

Model averaging: N/A

Model ensembles: N/A

#### Analysis and Correction of non-independence

Spatial autocorrelation: N/A

Temporal autocorrelation: N/A

Nested data: N/A

## Assessment

#### Performance statistics

Performance on training data: True positive rate, AIC, AUC

#### Plausibility check

Response shapes: No plausibility checks conducted