Examining and Addressing the Wallacean shortfall: Species distribution modelling and biodiversity patterns of Hawkmoths in the Old World

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Liliana Ballesteros Mejia

Aus Kolumbien

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von
PD. Dr. Jan Beck , Prof. Dr. Peter Nagel und Dr. Carsten Bruehl.

Basel, den 13 November 2012

Prof. Dr. J. Schibler
Dekan

TABLE OF CONTENTS

CHAPTER 1

1.1. General Introduction

Explaining how biodiversity is spatially and temporally distributed across our planet has been a central topic in biology since the time of Alexander von Humboldt (Hawkins, 2001). Over 200 years later, understanding biodiversity patterns remains a major topic of investigation in biogeography and macroecology (Guisan and Rahbek 2011).

Unfortunately, our current knowledge of biodiversity is very incomplete, we are still uncertain about how many species are there on our planet, and for those described, knowledge about their ecology and distribution is very scarce. These two phenomena are the major drawbacks in current study of biodiversity: known as the Linnean and the Wallacean shortfalls (Brown and Lomolino 1998). The first refers to the fact that a vast majority of species diversity remains undescribed (e.g. from tropical arthropods for example only 30% are described Hamilton *et al.* 2010), taking into account that recent estimates predict about 8.7 millions of species in the world (Mora *et al.* 2011). The second refers to the fact that the geographical distribution of most species is only incompletely, if at all, known (Lomolino 2004, Bini *et al.* 2006). We are in need for these data to be able to appreciate and understand the full taxonomical and functional diversity range that currently exist, but even more because biodiversity is threatened at the very core; global warming, land use changes, among others factors are driving species to extinction at a very alarming speed.

Large-scale analyses both temporal and spatial, trying to capture emergent patterns are part of the research agenda of macroecology as a response to the realization that focusing on local scales and or single or few species did not fully explain neither abundance nor distribution of the species (Gaston and Blackburn, 2000). These analyses have benefited from the current development of sophisticated statistical techniques that have open their way into ecological applications (Heisey *et al.* 2010). Analyses of species distribution and richness patterns have become technically feasible with the availability of remotely sensed environmental data, and the development of Geographic Information System (GIS) (Brown *et al.* 1996) and seems the way to disentangle causal and collinear driver of the observed biodiversity patterns (i.e. global and/or regional) (Beck *et al* 2012).

These technical developments allowed the possibility to delimit species potential distributions based on correlations with environmental parameters at sampled locations across space, to produce species range maps. Grid-based analysis overlays range species maps and allow addressing questions such as, which environmental factors (e.g. temperature, primary productivity, water and energy availability) provide better explanations to the observed patterns of biodiversity at different scales and extents (i.e. global, continental, regional) as well as among different taxa (Hawkins *et al.* 2003)

Also spatial models of distributions can be used to analyze similarities and differences between species niche, and even design networks of protected areas and forecast what species will be found at a given site (Kremen *et al.* 2008)

However, despite the advances in techniques and methodologies for analyses, there are substantial data deficiencies in this field of research. Species distribution, species traits and phylogenetic data would be needed to allow more comprehensive analyses (Beck *et al.* 2012). The majority of the large-scale analyses have been biased towards a limited set of relatively well-known taxa (i.e. birds, mammals and plants; Rahbek and Graves 2001, Kreft and Jetz 2007, 2010, Tittensor *et al.* 2010) whereas studies on groups like invertebrates, particularly herbivore insects are scarce, despite being the most species-rich groups (Beck *et al*. 2012).

The gap in knowledge on species distributions has prompted an awareness of the potential importance of Natural History collections – data that are generally available yet practically not accessible without substantial effort. Accordingly, these institutions and interested users have promoted endeavours for compiling such data in electronic databases to make them more widely available (Graham *et al.* 2004a). Projects include the Global Biodiversity Information Facility (GBIF; http://www.gbif.org/) that was established in 2001 and facilitates the access to biodiversity data comprising so far more than 338 million records (accessed October 21, 2012). However a large proportion of records are still not electronically available (O'Connell *et al.* 2004, Newbold 2010; *Chapter 5* of this thesis), leaving an enormous task to digitize such information.

An innovative computer-based tool that have seen a rapid development in the recent years aiming to generate range information of species based on distributional records (Guisan and Thuiller 2005, Phillips *et al.* 2006, Elith *et al.* 2010), called Species Distribution Modelling (SDMs). It has been used for different purposes and to address interesting ecological questions for example in conservation management (Thorn *et al.* 2009), for predicting past distributions of species (Peterson *et al.* 2004), distributions under future climate or land use scenarios (Araujo et al 2004), or the

ecological and geographic differentiation of closely related species (Graham *et al.* 2004b). However, its application to large scales, many species, and to relatively poor and biased distributional records is still sparse. Additionally, some methodological issues need to be sorted out before reliable range estimates can be retrieved, which is one of the aims of this dissertation and will be discussed below.

1.2. Study region

The study has almost a global spatial extent, excluding only the Americas. Reasons for excluding these were the need to reduce species richness due to time constraints, and the very low species overlap between the Americas and the rest of the world which made this split feasible (see also Kawahara et al. 2009). In addition to the Old World (i.e., Europe, Africa and Asia) this study also includes Melanesia, Australia and the Pacific. (From 25W° to 180E° and from 89N° to 49S°).

This extensive geographical area includes two complete latitudinal ranges (South Africa to Scandinavia; New Zealand to Siberia). Furthermore, the region includes altitudinal gradients ranging from coastal lowlands to heterogeneous mountain landscapes including alpine and nival landscapes. Very distinct ecosystem types, from deserts to rainforest, occur in spatial replications on different continents. It also includes various geographical structures (i.e. isolated islands to continents) that vary in size, isolation, geology and geographical history. Such variation facilitates correlation analysis that aims to uncover general global patterns and thereby gives hints towards the mechanisms causing them.

1.3. Sphingid moths as model taxa

Commonly referred to as hawkmoths, Sphingidae is a family of the Lepidoptera, placed phylogenetically within the Bombicoidea superfamily. It is a taxon of moderate species richness with <1500 species known globally, of which 982 are recognized within the study area (see *Chapter 5* for detail). Their large body size and great beauty have made them very appealing to both amateur and professional collectors for over two centuries. In consequence, this group of moths has been sampled relatively well and is well represented in collections worldwide. The abundance of specimens and relatively low diversity (for an insect family) has probably contributed to the fact that its taxonomy and phylogeny is among the best-known invertebrate taxa (Kitching and Cadiou 2000), although there are still many details to be resolved.

Hawkmoth larvae, referred to as "hornworms", are folivorous with a low degree of hostplant specialization (i.e. specialization below family is rare; for example, the Oleander hawkmoth *Daphnia nerii* feeds mainly on the toxic oleander (*Nerium oleander*) but also on other plants of the family Apocynaceae) (Pittaway 1997-2012 , Mazzei *et al* 1999-2012).

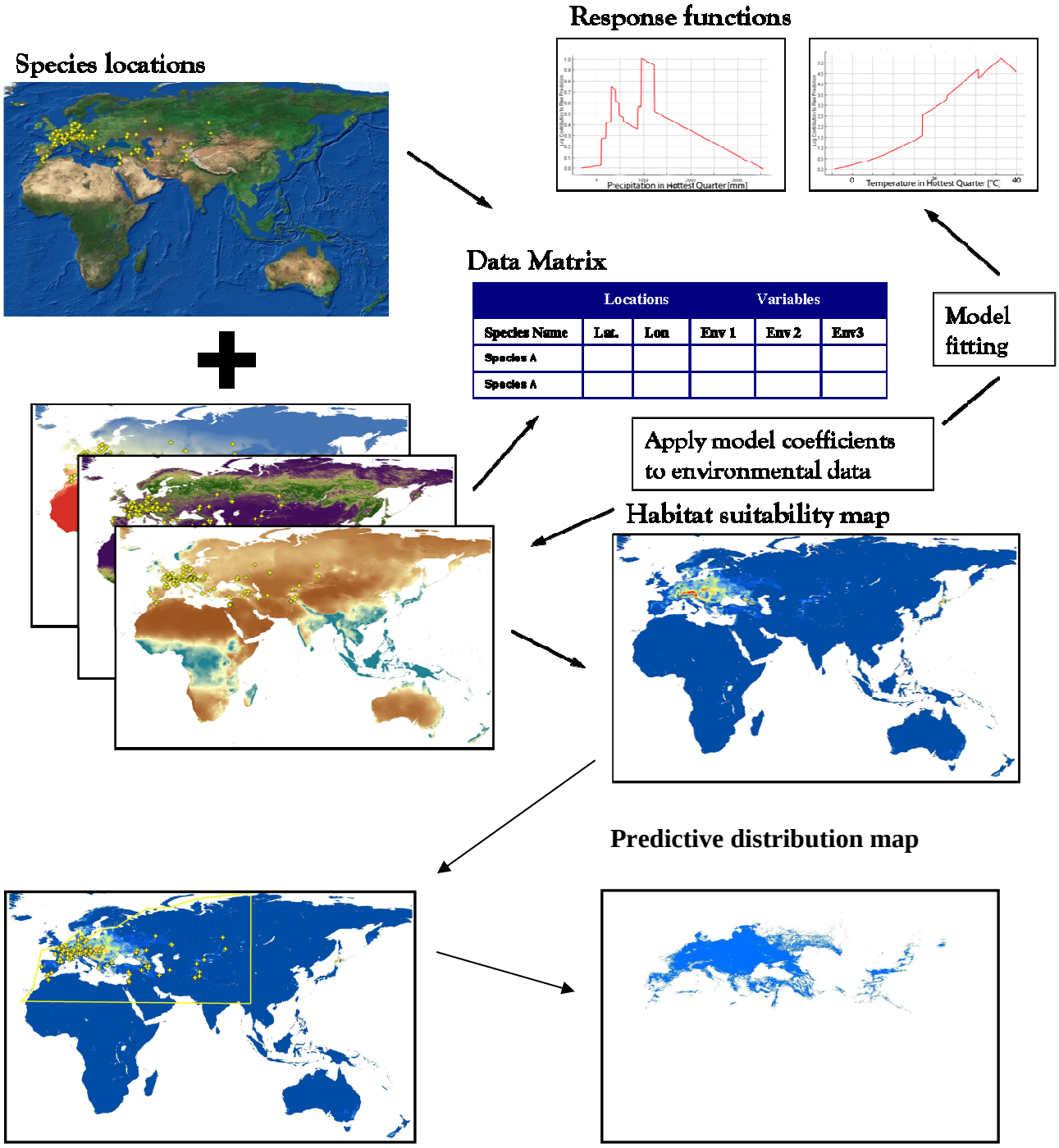
Most of the adults are nocturnal although there are some diurnal genera (*Hemaris, Sataspus, Macroglosum, and Hayesiana*). Hawkmoths show a great variability in traits, such as life history strategies, adult resource use (from non-feeding, flower nectar feeding, bee-nest parasites), egg maturation and mobility (Beck *et al.* 2006). Overall this interesting diversity of species traits makes them very suitable for evolutionary ecological studies (Janzen, 1984).

1.4. Species distributions: statistical modelling and ecological theory

In recent decades interest in knowing the geographical distribution of biodiversity on Earth increased, sparked by the alarming speed of losing biodiversity due to global warming, land use changes, and other anthropogenic effects.

Scientists often use locally collected data to then assess change at different spatial extents: (i.e. landscape, regional or global) and often use statistical or simulation models to extrapolate those data in space (Peters *et al.* 2004). A technique that has became popular nowadays using statistics models to extrapolate collected data is species distribution modelling (SDM). This technique allows characterizing the environmental conditions that are suitable for the species to live and then identify where such environmental conditions are distributed in space. To fit models, it links observations of the occurrence of the species with environmental conditions at these sites, focusing on variables that are thought to influence habitat suitability and therefore the distribution of the species (Pearson 2007). These correlative models provide insights on the species' environmental tolerance and preferences, with the potential of being extrapolated in time and space. Figure 1.1. illustrates the steps towards an SDM-based distribution map.

Figure. 1.1. Diagram depicting the steps towards producing a distributional map. Species locations are linked to the values of environmental predictors at those locations coordinates. A modelling algorithm is applied then to describe the relationship between the species' locations and the predictors. Parameters derived from such models are extrapolated to environmental data available as grid-based maps to produce a geographical prediction for habitat suitability. After accounting for a biogeographical reasonable expectation, and setting a threshold for transforming those continuous predictions into a binary presence-absence, we get a predictive distribution map. Adapted from (Franklin 2009)



1.4.1. Niche concepts and other theories in environmental and geographical space

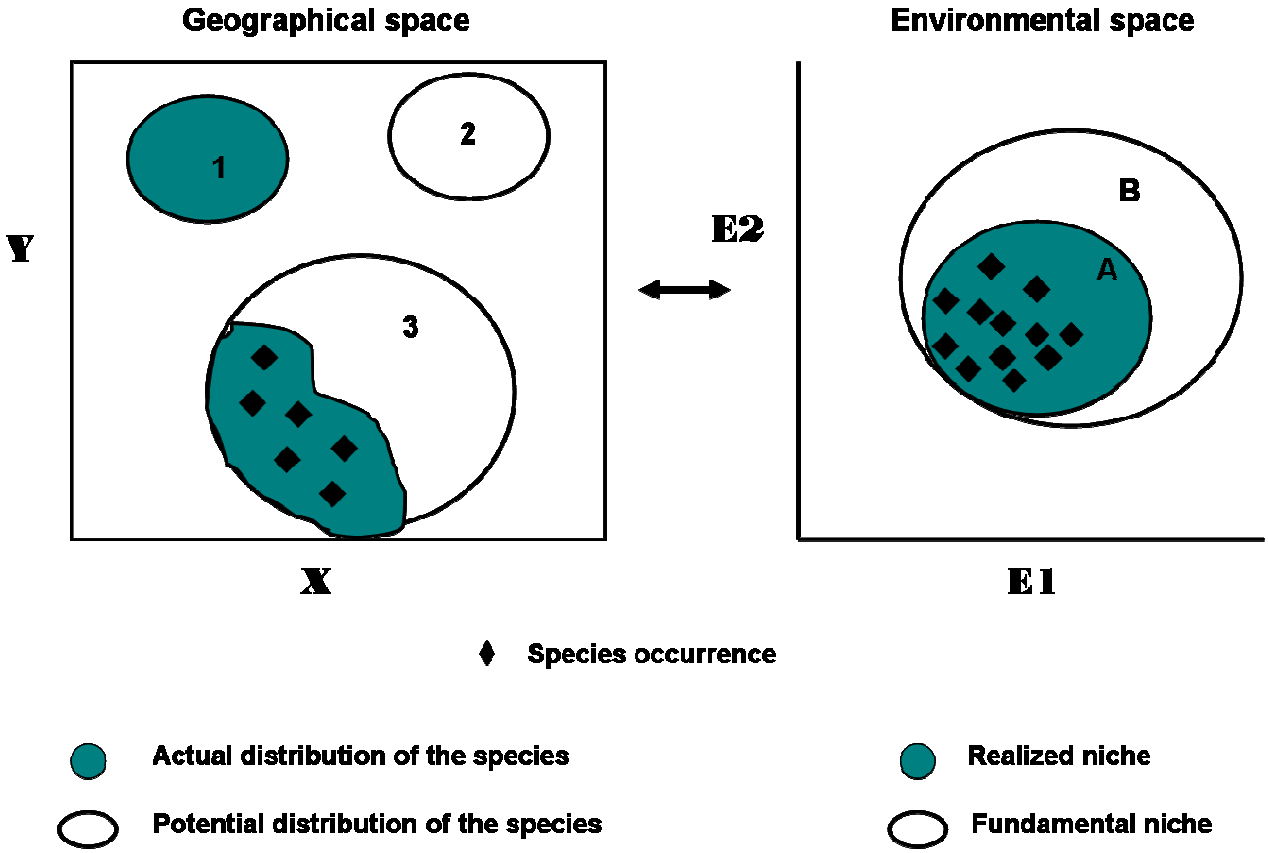
SDMs have their grounds in niche theory (Soberon and Peterson 2005). In recent years several authors have discussed the relationship between ecological niche concepts and SDMs (Austin 2002, Guisan and Thuiller 2005, Kearney 2006).

The species niche concept has changed over time and has several interpretations (Chase and Leibold 2003). A major distinction is between (a) the functional concept (Elton 1927), i.e. the position or functional role of the species in the community; and (b) the ecological concept (Grinnell 1917), i.e. the set of environmental factors within which the species can survive and reproduce. Hutchinson (1957) made a distinction between *fundamental* and *realized* niche. The *fundamental* (potential) niche is the space in a n-dimensional hypervolume formed by a set of environmental variables where the species can survive and reproduce. However, because of biotic interactions (e.g. competition, predation, facilitation etc.) a species can be excluded from some parts of that fundamental niche. This reduced hypervolume is called *realized* (actual) niche (Whittaker *et al.* 1972). Species Distribution Models deal with environmental niches and therefore with Grinellian rather than Eltonian niches (Peterson *et al.* 2006). There is discussion within the SDM community whether SDMs model fundamental or realized niches (see below). In any case, the variables usually available for SDM represent only a subset of all the possible environmental factors that might influence the distribution of the species. They mostly represent abiotic factors, often constrained by availability at the desired extent and resolution. A majority of SDM approaches is heavily or entirely focused on climatic environmental variables (Carpenter et al 1993, Pearson et al 2002). Figure 1.2 shows a species' distribution in environmental as well as in geographical space to visualize the concepts defined above.

Apart from abiotic and biotic factors, Pulliam (2000) stressed the importance of including measures of fitness when identifying species; niches, and pointed out that source sink dynamics and metapopulation dynamics (Hanski 1999) might help to explain the relationship between distribution of species and suitable habitat. A species might not occupy a suitable habitat due to local extinction resulting from population dynamics or dispersal limitation (i.e. metapopulation theory). Source-sink dynamics refers to the situation where an area ("sink") does not provide suitable conditions to support a viable population but is frequently colonized by individuals coming from an area which does it ("source"), so that a species can be recorded in an unsuitable place. This particular consideration should be taken into account when applying SDMs since they rely on occurrence records of species. Ideally, only occurrences known from breeding populations should be used, but practically such information is often unavailable (Beale & Lennon 2012). Furthermore,

observations are probably more frequent from "source" population rather than "sink", which may justify to a certain degree that this problem is usually overlooked (Pearson 2007).

Figure 1.2. Diagram representing the relationship between geographic distribution and environmental niche. Geographical space refers to the spatial location that the species occupy whereas environmental space refers to what can be considered Hutchinson niche (i.e. hypervolume, represented here only within two dimensions). Black diamonds represent the occurrence of the species. In geographical space green areas represent the actual distribution of the species, which in environmental space is the realized niche. Region 1 in geographical space and region A in environmental space both represent areas where the species has not been detected yet. Region A in geographical space and Region B in environmental space are both within the fundamental niche of the species but is not occupied because of some biotic factors such competition or dispersal limitation. Region 2 in geographical space depicts that part of the niche that the species could live in (i.e. appropriate environmental conditions) but it has not been able to disperse to. (Diagram extracted from Pearson, 2007)



1.4.2. Other considerations when model species distributions

As outlined above, SDMs combines occurrence of species with environmental factors in the area of interest, and has undergone a rapid development in modelling techniques in recent years (Stockwell and Noble 1992, Breiman 2001, Phillips *et al.* 2006) as well as increased popularity. It has been a lot of recent discussion about exactly what component of the niche is used for SDMs. On the one hand, some authors argued that due to the absence of variables that involve biotic interactions or dispersal limitation the fundamental part of the niche is modeled (Soberon and Peterson 2005,

Soberón 2007, 2010), although some progress has been made in trying to include such variables (Warren *et al.* 2010, Wilson *et al.* 2010). On the other hand, it can be argues that SDMs identify the realized niche of the species even without including biotic interaction variables, as they use actual (i.e. realized) distributional data to build the model (Guisan and Zimmermann 2000, Austin 2002, 2007, Pearson and Dawson 2003). Personally, I consider that SDM is modelling realized niches by producing a model that closely resemble realized distributions of species based on observations where the species were actually found. Despite of this debate, SDM seems to be able to capture a significant amount of the ecological signature even when biotic data is often lacking in the models

1.4.3. SDMs and spatial dependency

There are multiple factors operating in a hierarchical way at different scales both spatial and temporal to shape the distribution of a species and patterns of species richness (Levin 1992). However, the extent to which those factors affect the observed pattern of distribution depends on the resolution (grain size) and the extent (area) of the study (Elith and Leathwick, 2009).

Soberon & Peterson (2005) present an interesting framework to analyze this issue. In there, they recognized three important factors: (1) *Abiotic factors* usually determine the size and shape of distributions at continental and even regional scales but become less important the smaller the scale gets (Hortal *et al.* 2010). These factors are often responsible for physiological constraints and climatic responses. (2) *Biotic factors* affect fitness in a regulatory way (predation, competition, facilitation). These factors show the opposite pattern to the abiotic factors, i.e. they are often less important at continental scales but become increasingly important the smaller the scale gets. At large scales, biotic factors only will have a determining role for extreme specialists species (e.g. butterflies which distribution is strongly linked to that of their host-plant) (Araújo and Luoto 2007). (3) *Movement related factors* are determining the spatial movement of individuals or populations. They can be divided into two categories, biogeographic and occupancy factors, and they also have a mixed strength of their influence on distributions at different scales. Theoretically, biogeographic factors have major effects on distribution patterns at large scales, though they are not easy to account for and their effects could be idiosyncratic (i.e. vary from species to species). Progress has been made but still there is a lot to do in that field (Wallace, 1869; 1876; Matthew, 1915; MacArthur and Wilson, 1967; Kreft and Jetz, 2010; Soberon 2010). Occupancy factors are the result of multiple demographic factors coming together. Metapopulation dynamics, short-distance dispersal and localized disturbances (Hanski 1999, Pulliam 2000) can have effects on a small scale, affecting how individuals or populations aggregate.

There is compelling evidence for these three factors acting together to shape species distributions (Leathwick and Austin 2001, Mackey and Lindenmayer 2001, Heikkinen *et al.* 2007). However,

there is still a lot of work to do to try to incorporate all three factors in SDM despite of the progress made.

1.5. Available species distribution data

Numerous endeavours have been reported in recent time towards mapping the distribution of species for different taxa from occurrence records around the world, i.e. trying to collect, compile and make available such data for various purposes (Graham *et al.*, 2004a; Soberon and Peterson, 2004). Available data typically stemmed from highly non-random observations and surveys both in space and time, and a common output of their use is a set of distribution maps. These maps vary enormously in three aspects: (1) Data type (i.e., presence-absence data per grid cells, based on surveys; model predictions of occurrence; expert-opinion range maps or focal species point occurrences), (2) resolution (grain size) and (3) extent. Table 1.1 illustrates examples of large-extent species distribution data available for biodiversity analysis. It is evident that the amount of data is not impressive when compared to global species richness. Furthermore, it is evident the biased towards vertebrates (14 out 24 databases are exclusively dedicated to them) and temperate regions. Table 1.1 highlights the necessity for providing more large-extent, high-resolution distributional data for taxa in the tropics, particularly insects. Similar biases in published macroecological studies (Beck et al. 2012) are almost certainly due to this lack of data.

Despite the increased availability of data in recent years, the greatest demand for data is for conservation planning and the global change (climate or land use) analyses that cannot wait until all sites have been surveyed and detailed presence-absence data are ready to use. It is here where SDMs are valued the most, providing an alternative approach to expand the use of direct observation data and helping us to understand patterns in species distributions.

1.6. On this thesis

This work is the result of a collaborative project that aims at retrieving distributional information for a complete family of insects, and some of the first analysis that such a dataset allows. In particular, the compilation of raw distributional records, and their processing until they could be utilized for SDM, were mostly carried out by I.J. Kitching (Natural History Museum London) and J. Beck (Univ. Basel) over a time period of >10 years. My part in this project (i.e., this thesis) was restricted to utilize these data for SDM and other analyses, with the particular aim of providing high resolution GIS-based range data for all species. Due to the aim of providing chapters for stand-alone publication in scientific journals, it is necessary to describe and discuss all aspects of this work. Therefore I will refer to “we” throughout much of the text.

In the following parts of the thesis, **Chapter 2** introduces SDMs as a tool for providing species distributions and evaluates which algorithm (from the most commonly used), was the most suitable for modelling while considering some intrinsic properties of the species and data. In **Chapter 3**, we assess the value of different raw data sources, i.e. by comparing an independent compilation of occurrence data and the GBIF database, with special attention to the information on geographical distribution and climatic niche that they provide. **Chapter 4** reports species diversity patterns based on numerical estimators methods in a fraction of our study region (sub-Saharan Africa), in relation to their main environmental correlates. We also provide an assessment of inventory completeness for that particular region. **Chapter 5** contains a detailed documentation of data acquisition, processing and modelling procedures. Furthermore, here we present patterns of species diversity for the whole family in the complete region and report achievements, challenges and limitations of the project. **Chapter 6** focuses on one specific application of SDMs, the prediction of the range of an invasive species (*Agrius cingulata*). **Chapter 7** provides a general discussion and conclusions, including some preview on further studies on this dataset that are likely to be done in the future.

At the time of official submission of this thesis:

Chapter 2 is a manuscript re-submitted after "Major revision" to Ecological Modelling

Chapter 3 was published in 2013 Diversity and Distribution

- Beck, J., **Ballesteros-Mejia, L.**, Nagel, P., Kitching, I.J. (2013). Online solutions and the "Wallacean shortfall": What does GBIF contribute to our knowledge or species' ranges?. Diversity and Distributions, early view (doi:10.1111/ddi.12083).

Chapter 4 was published in 2013 in Global Ecology and Biogeography

- Ballesteros-Mejia, L., I. J. Kitching, W. Jetz, P. Nagel, and J. Beck. (2013). Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography 22: 586-595.

Chapter 6 was published in 2011 in the International Journal of Pest Management.

- Ballesteros-Mejia, L., I. J. Kitching, and J. Beck. 2011. Projecting the potential invasion of the Pink Spotted Hawkmoth (Agrius cingulata) across Africa. International Journal of Pest Management 57:153 – 159.

In addition, associated with this thesis, the following electronic data are submitted at the electronic network drive of the university computing centre (\\nlu-jumbo.nlu.p.unibas.ch\nlu-gis$\GIS).

- Raw model outputs from the random forest (RF) models; RandomForest_Models
- Raw model outputs from the Maxent models: Maxent Models_raw data
- Raw model outputs expert-edited for dispersal limitation plus the polygons used for editing.
- Thresholded maps (in WGS1984 geographical coordinates)
- Threshold output models (Projected into equal area grid: Mollwide at 5 x 5 km resolution)
- Threshold output models (Projected into equal area grid: Mollwide at 200 x 200 km resolution)
- Biodiversity maps:
    - Maps at 5 x 5 km resolution of the total species richness as well as for each one of the 7 tribes of the family.
    - Maps at 200 x 200 km resolution the total species richness as well as for each one of the 7 tribes of the family.
    - Map of beta diversity at 200 x 200 km resolution for the total species.

1.7. References

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16:743–753.

Austin, M. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157:101–118.

Austin, M. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200:1–19.

Beck, J., L. Ballesteros-Mejia, C. M. Buchmann, J. Dengler, S. a. Fritz, B. Gruber, C. Hof, F. Jansen, S. Knapp, H. Kreft, A.-K. Schneider, M. Winter, and C. F. Dormann. 2012. What's on the horizon for macroecology? Ecography 35:1–11.

Beck, J., I.J. Kitching, and K.E. Linsenmair. 2006. Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. Ecotropica 12:1–13.

Bini, L. M., J. A. F. Diniz-Filho, T. F. L. V. B. Rangel, R. P. Bastos, and M. P. Pinto. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. Diversity and Distributions 12:475–482.

Breiman, L. 2001. Random forests. Machine learning 45:5–32.

Brown, J., and M. Lomolino. 1998. Biogeography. Sinauer Press, Massachusetts.

Brown, J. H., G. C. Stevens, and D. M. Kaufman. 1996. The geographic range: Size, Shape, Boundaries, and Internal Structure. Annual Review of Ecology and Systematics 27:597–623.

Carpenter, G. A.N. Gillison, J.Winter. 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2: 667-680

Chase, J. M., and M. A. Leibold. 2003. Ecological Niches: Linking Classical and Contemporary Approaches. Universtity of Chicago Press, Chicago.

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1:330–342.

Elith, J., J.L. Leathwick. 2009. Species Distribution Models: Ecological explanation and prediction across space and time. Annual review of ecology, evolution and systematics 40: 677-697

Elton, C. 1927. Animal Ecology. Sidgwick & Jackson, London.

Franklin, J. 2009. Mapping species distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge.

Galster, S., N. D. Burgess, J. Fjeldsa°, L. A. Hansen, and C. Rahbek. 2007. One degree resolution databases of the distribution of 1085 mammals in Sub-Saharan Africa.

Gasc, J. P., A. Cabela, J. Crnobrnja-Isailovic, D. Dolmen, K. Grossenbacher, P. Haffner, J. Lescure, H., T.S. Martens, M. Veith, and A. Zuiderwijk. 1997. Atlas of amphibians and reptiles in Europe. Societas Europaea Herpetologica & Museum National d'Histoire Naturelle, Paris.

Gaston, K. J. & Blackburn, T. M. 2000. Pattern and process in macroecology. Oxford, UK: Blackwell Science.

Godinho, R., J. Teixeira, R. Rebelo, P. Segurado, and A. Loureiro. 1999. Atlas of the continental Portuguese herpetofauna : an assemblage of published and new data. Revista Espanola de Herpetologia 13:61–82.

Graham, C., S. Ferrier, F. Huettman, and C. Moritz. 2004a. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19:497–503.

Graham, C. H., S. R. Ron, J. C. Santos, C. J. Schneider, and C. Moritz. 2004b. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58:1781–93.

Grinnell, J. 1917. The Niche-Relationships of the California Thrasher. The Auk 34:427–433.

Guisan, A., and C. Rahbek. 2011. SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography 38:1433–1444.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., and N. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Hamilton, A. J., Y. Basset, K. K. Benke, P. S. Grimbacher, S. E. Miller, V. Novotný, G. A. Samuelson, N. E. Stork, G. D. Weiblen, and J. D. L. Yen. 2010. Quantifying uncertainty in estimation of tropical arthropod species richness. The American naturalist 176:90–5.

Hansen, L. A., N. D. Burgess, J. Fjeldsa°, and C. Rahbek. 2007a. One degree resolution databases of the distribution of 739 amphibians in Sub-Saharan Africa.

Hansen, L. A., J. Fjeldsa°, N. D. Burgess, and C. Rahbek. 2007b. One degree resolution databases of the distribution of 1789 birds in Sub-Saharan Africa.

Hanski, I. 1999. Metapopulation Ecology. Oxford Universtiy Press, Oxford UK.

Hawkins BA (2001) Ecology's oldest pattern? Trends in Ecology and Evolution 16, 470.

Hawkins, B. A., R. Field, H. V. Cornell, D. J. Currie, J.-F. Guegan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. O'Brien, E. E. Porter, and J. R. G. Turner. 2003. Energy,water, and broad-scale geographic patterns of species richness. Ecology 84:3105–3117.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography 16:754–763.

Heisey, D. M., E. E. Osnas, P. C. Ross, D. O. Oly, J. A. Langenberg, and M. W. Miller. 2010. Rejoinder : sifting through model space. Ecology 91:3503–3514.

Hortal, J., N. Roura-Pascual, N. Sanders, and C. Rahbek. 2010. Understanding (insect) species distributions across spatial scales. Ecography 33:51–53.

Huntley, B., R. E. Green, Y.C. Collingham and S.G. Willis. (2007) A climatic atlas of European breeding birds.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposium On Quantitative Biology 22:415–427.

IUCN 2012. The IUCN Red List of Threatened Species. Version 2012.2. <http://www.iucnredlist.org>.

Janzen, D. H. 1984. Two ways to be a tropical big moth: Santa Rosa saturniids and sphingids. Pages 85–140 *in* R. Dawkins and M. Ridley, editors. Oxford surveys in Evolutionary Biology, 1st edition. Oxford University Press, Oxford.

Jetz, W., and C. Rahbek. 2002. Geographic range size and determinants of avian species richness. Science 297:1548–51.

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? Oikos 115:186–191.

Kitching, I., J., and J. M. Cadiou. 2000. Hawkmoths of the world. The Natural History Museum & Cornell University Press, London.

Kreft, H., and W. Jetz. 2007. Global patterns and determinants of vascular plant diversity. Proceedings of the National Academy of Sciences (B) 104:5925–30.

Kreft, H., and W. Jetz. 2010. A framework for delineating biogeographical regions based on species distributions. Journal of Biogeography 37:2029–2053.

Kremen, C., A. Cameron, A. Moilanen, S. J. Phillips, C. D. Thomas, H. Beentje, J. Dransfield, B. L. Fisher, F. Glaw, T. C. Good, G. J. Harper, R. J. Hijmans, D. C. Lees, E. Louis, R. a Nussbaum, C. J. Raxworthy, A. Razafimpahanana, G. E. Schatz, M. Vences, D. R. Vieites, P. C. Wright, and M. L. Zjhra. 2008. Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. Science 320:222–226.

Leathwick, J.R., and M. Austin. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forest. Ecology 82:2560–2573.

Levin, S. 1992. The Problem of Pattern and Scale in Ecology. Ecology 73:1943–1967.

Lomolino, M. V. 2004. Conservation Biogeography. Pages 293–296 *in* M. V. Lomolino and L. R. Heaney, Eds. Frontiers of Biogeography: New directions in the geography of Nature. . Sinauer Associates, Inc. Publishers, Sunderland, MA.

MacArthur, R.H. and E. Wilson. 1967 The Theory of Island Biogeography. Princeton University Press, New Jersey.

Mackey, B. G., and D. B. Lindenmayer. 2001. Towards a hierarchical framework for modelling the spatial distribution of animals. Journal of Biogeography 28:1147–1166.

Matthew, W.D. 1915. Climate and evolution. Annals of the New York Academy of science 24: 171:318.

Mazzei, P., D. Morel,R. Panfili., I. Pimpinelli, D.Reggianti. 1999-2012. Moths and butterflies of Europe and North Africa. http://www.leps.eu.

Mora,C., D.P.Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How many species are there on Earth and in the ocean? PLoS biology 9:e1001127.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography 34:3–22.

Opler, P. A., K. Lotts, T. Naberhaus. Coordinators. 2012. Butterflies and Moths of North America.

O'Connell, A. F. J., A. T. Gilbert, and J. S. Hatfield. 2004. Contribution of Natural History Collection Data to Biodiversity Assessment in National Parks. Conservation biology 18:1254–1261.

Pearson, R. G. 2007. Species' Distribution Modeling for conservation Educators and Practitioners. Synthesis. American Museum of Natural History: 50.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12:361–371.

Pearson, R. G., T. P. Dawson, and P.M. Berry, P.A. Harrison. 2002. SPECIES: A spatial evaluation of climate impact on the envelope of species. Ecological Modelling 154:289-300

PESI (2012). Pan-European Species directories Infrastructure. Accessed through www.eu-nomen.eu/portal.

Peters, D. P. C., J. E. Herrick, D. L. Urban, R. H. Gardner, and D. D. Breshears. 2004. Strategies for ecological extrapolation. Oikos 106:627–636.

Peterson, A. T., E. Martínez-meyer, and C. González-salazar. 2004. Reconstructing the Pleistocene geography of the Aphelocoma jays (Corvidae). Diversity and Distributions 10:237–246.

Peterson, A. T., V. Sánchez-Cordero, E. Martínez-Meyer, and A. G. Navarro-Sigüenza. 2006. Tracking population extirpations via melding ecological niche modeling with land-cover information. Ecological Modelling 195:229–236.

Phillips, S., R. Anderson, and R. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Pittaway, A. R. 1997-2012. Sphingidae of the Western Palaearctic.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters 3:349–361.

Rahbek, C., and G. R. Graves. 2001. Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences (B) 98:4534–9.

Rasmussen, J. B., L. A. Hansen, N. D. Burgess, J. Fjeldsa°, and C. Rahbek. 2007. One degree resolution databases of the distribution of 467 snakes in Sub-Saharan Africa.

Settele, J., O. Kudrna, A. Harpke, I. Kuehn, C. van Swaay, R. Verovnik, M. Warren, M. Wiemers, J. Hanspach, T. Hickler, E. Kühn, I. van Halder, K. Veling, A. Vliegenthart, I. Wynhoff, and O. Schweiger. 2008. Climatic Risk Atlas of European Butterflies. BIORISK – Biodiversity and Ecosystem Risk Assessment. . Pensoft, Sofia.

Soberón, J. and A.T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society, London (B) 359, 689–698.

Soberón, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species distributional areas. Biodiversity Informatics 2:1–10.

Soberón, J. 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecology letters 10:1115–1123.

Soberón, J. M. 2010. Niche and area of distribution modeling: a population ecology perspective. Ecography 33:159–167.

Stockwell, D. R. B., and I. R. Noble. 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Mathematics and Computers in Simulations 33:385–390.

Thorn, J. S., V. Nijman, D. Smith, and K. a. I. Nekaris. 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). Diversity and Distributions 15:289–298.

Tittensor, D. P., C. Mora, W. Jetz, H. K. Lotze, D. Ricard, E. V. Berghe, and B. Worm. 2010. Global patterns and predictors of marine biodiversity across taxa. Nature 466:1098–101.

Wallace, A. R. 1869. The Malay Archipelago. Oxford in Asia Hardback Reprint (1986). Oxford University Press, Oxford.

Wallace, A.R. 1876. The Geographical Distribution of Animals,Macmillan

Warren, M., M.P. Robertson, and J.M.Greeff. 2010. A comparative approach to understanding factors limiting abundance patterns and distributions in a fig tree-fig wasp mutualism. Ecography 33:148–158.

Whittaker, R.H., S.A. Levin, and R.B. Root. 1972. Niche, Habitat and Ecotope. The American naturalist 107:321–338.

Wilson, R.J., Z.G. Davies, and C.D. Thomas. 2010. Linking habitat use to range expansion rates in fragmented landscapes: a metapopulation approach. Ecography 33:73–82.

Table 1.1. List of available databases (online or in atlases) compiling information about distribution of organisms.

Data of publication (Year)	Description	URL	Reference
The Reptile Database. 1995	Data type: Resolution: Maps based on TDWG standart (but not a precise distribution map)	http://www.reptile-database.org	Uetz, P. & Etzold, T. 1996
EBCC Atlas of European Breeding birds. 1997	Data type: Survey Maps for 495 bird species Resolution: 50x50 km	http://s1.sovon.nl/ebcc/eoa/	Huntley *et al.* 2007
BirdLife International	Data type: Expert drawn maps Resolution:100 – 200 km	http://www.birdlife.org/datazone/info/spcdownload	
Atlas of amphibians and reptiles in Europe. 1983	Data type: Presence only Resolution:50 x 50 km UTM		Gasc *et al.* 1997
Copenhagen database for African Mammals. 2007	Data type: Presence only and expert opinion data Resolution: 1 degree	http://130.225.211.158/subsaharanafrica/subsaharan.htm	Galster *et al.* 2007
Copenhagen database for African Birds. 2007	Data type: Presence only and expert opinion data Resolution: 1 degree	Data type: Presence only and expert opinion data Resolution: 1 degree	Hansen *et al.* 2007a
Copenhagen database for African Amphibians. 2007	Data type: Presence only and expert opinion data Resolution: 1 degree	http://130.225.211.158/subsaharanafrica/subsaharan.htm	Hansen *et al.* 2007b
Avian distribution database	Data type: Survey data Resolution: 1 degree	www.sciencemag.org/cgi/content/full/297/5586/1548/DC1	Online supplementary material in Jetz & Rahbek 2002.
Plant database 2007	Data type: Inventory data Resolution: 1 degree		Kreft & Jetz 2007
Climatic risk Atlas of European Butterflies	Data type: Presence only data Resolution: 1 degree	Collect the data from the project "Mapping European Butterflies Project" (MEB: www.european-butterflies.eu)	Settele *et al.* 2008
The Sphingidae of Southeast-Asia. 2004-2008	Data type: Range maps	http://www.sphin-sea.unibas.ch	
Sphingidae of the Western Palaearctic 1997-2012	Data type: Presence only	http://tpittaway.tripod.com/sphinx/list.htm	Pittaway, A. R. (1997-2012)
Sphingidae of the Eastern Palaearctic (including Siberia, the Russian Far East, Mongolia, China, Taiwan, the Korean Peninsula and Japan). 2000-2012	Data type: Presence only	http://tpittaway.tripod.com/china/china.htm	Pittaway, A. R., and I. J. Kitching. (2000-2012).
IUCN database for Birds	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.

IUCN database for Reptiles	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Amphibians	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Mangroves	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Corals	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for See grasses	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Parrotfish	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Angelfish	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
IUCN database for Wrasses	Data type: Expert-drawn maps Resolution: 100-200 km	http://www.iucnredlist.org/technical-documents/spatial-data	IUCN, 2012.
A Pan-European Species-directories infrastructure (PESI), European taxa	Data type: Occurrence data and expert opinion. Resolution: 100 – 200 km	http://www.eu-nomen.eu/portal/	PESI, 2012.
The Butterflies of North America: A Natural History and Field Guide. 1986	Data type: Expert-drawn maps Resolution: 100-200 km		Scott, 1986

CHAPTER 2

What factors influence the accuracy of distribution models?

Liliana Ballesteros-Mejia[1*], Ian J. Kitching[2], Peter Nagel[1], Jan Beck[1]

[1] University of Basel, Department of Environmental Science (Biogeography), St. Johanns-Vorstadt 10, 4056 Basel, Switzerland

[2] The Natural History Museum, Department of Entomology, Cromwell Road, London SW7 5BD, UK.

*Author for correspondence: Tel.: +41-2670803, E-mail: liliana.ballesteros@unibas.ch

Manuscript submitted to: *Ecological Modelling* (*re-submitted*)

Abstract

Accurately predicting species' distribution has become a key factor for many aspects in ecology, evolution and conservation. Species distribution modelling (SDM), a widely used technique, aims to explain observed patterns of occurrence and predict geographic and ecological distributions. However, there is still disagreement on what method(s) to use. In particular, it is unclear whether different methods simply differ in quality (in which case one should use the best method a priori), or whether they perform differently depending on input data and the ecology of the species involved (in which case quality-weighted model averaging, for example, may be advisable). We investigated the performance of eight commonly applied SDM methods while also considering intrinsic characteristics of the species and their distributions (i.e., sample size, range size, climatic zone of occurrence and phylogenetic association), using a representative sample of species from the lepidopteran family Sphingidae (hawkmoths) and presence-only data. We used three criteria to evaluate the accuracy of models: Area under the receiver-operating characteristic (AUC), minimal predicted area (MPA), and expert opinion. Our results showed that *maximum entropy* modelling followed by *random forest* were the best methods. We did not find consistent effects of taxonomic association or range properties (climatic zone, range size) on model quality, nor did sample size (ranging from 3 to 889) allow good prediction of model performance. Our study is a relevant extension to previous modelling techniques comparisons as our test species are representative of a higher taxonomic group (i.e., family) regarding major distribution types, phylogeny and range of sample sizes, rather than being chosen for data availability. We show that the choice of modelling method is highly relevant whereas claims for effects of species or data properties could not be confirmed.

Keywords: AUC, BIOMOD, Expert opinion, Lepidoptera, Maxent, Niche modelling

2.1.Introduction

Species distribution models (SDMs) are correlative models that use environmental information to explain the observed patterns of species occurrence and predict their geographical and ecological distributions (Elith and Leathwick 2009). Accuracy in knowing species distributions is essential to understand emerging patterns of biodiversity and the processes that shape them (Ferrier et al. 2002).

SDMs are widely used for purposes such as conservation planning (Ferrier 2002), invasive species predictions (Peterson & Vieglais 2001; Thuiller et al. 2005; Ballesteros-Mejia et al. 2011), or predicting responses to climate change (Yates et al. 2010). They rely on the availability of point distributional records, but for a vast majority of species such data are sparse and biased taxonomically, ecologically and geographically (Boakes et al. 2010; Jetz et al. 2012; Beale & Lennon 2012; Ballesteros-Mejia et al. 2013).

While presence-absence data as a result of systematic surveys are ideal to use with SDMs, for the majority of species only presence records, if any, are available (i.e., true absence and sampling deficit cannot be distinguished; Elith and Leathwick, 2009). Natural history collections and faunistic publications are the primary sources of distributional information (Elith and Leathwick 2007; Newbold 2010) although for many species great advances have been made recently to compile and make such data available online (e.g., Global Biodiversity Information Facility, GBIF). However, all these data usually stem from opportunistic sampling, which can affect the quality of SDMs (Phillips et al. 2009).

In a landmark study, Elith et al. (2006) compared the performance of different SDM algorithms across a large number of taxa, guiding users on which methods were likely to perform better than others, based on their performance using the same databases. One important conclusion was the finding that maximum entropy models (MAXENT; Phillips et al. 2006; Phillips and Dudík 2008) outperformed other modelling methodologies. In combination with its easy-to-use software, MAXENT has since become a very popular method of SDM despite a widespread feeling of a lack of transparency of the method and software (Joppa et al. 2013).

However, some design details of the study by Elith et al. (2006) imply the need for further study. Models were fine-tuned to each species with great knowledge and attention to detail regarding the properties of each method as well as the ecology of modelled species (e.g., choice of relevant environmental variables). While this is obviously the ideal approach to SDM it may not reflect the majority of applications. In particular, SDMs are often applied to taxa about which very little is known (hence the need to estimate distributions via SDM), and relatively unspecific application to a broad range of taxa is required when advocating SDMs to address the ‘Wallacean shortfall’ (i.e.,

our poor knowledge of geographical distributions of most species; Lomolino 2004) on a broad scale (Jetz et al. 2012). Related to that, Elith et al. (2006) used taxa with independently available presence-absence data. This represents the "gold standard" in empirical model testing, but it also enforces a non-random selection of species that are relatively well-studied (implying a non-random selection of ecological traits, among them abundance).

It is not only important to understand the properties of these methods under ideal conditions (i.e., abundant data, good understanding of species' ecology, perfectly adjusted implementation of methodology, etc.), but we also need to know how robust they are under conditions of non-ideal implementation, which may represent the majority of cases. In analogy, risk assessments of new products (e.g., cosmetics: Larner 2008) also need to consider the chance and magnitude of damage due to likely occasional misuse. We do not want to advocate incorrect use of distribution models, but we need to acknowledge that in many cases there simply is not enough data and background knowledge available to guarantee perfect application.

SDM quality and accuracy can also vary between species (Newbold et al. 2009b). How different characteristics of studied taxa affect the performance of SDM techniques is a critical topic. Species with narrow niches (i.e., better defined climatic and/or habitat requirements) were found to be easier to model than those with a wider niche (Pearce et al. 2001; Newbold et al. 2009b), range-restricted species better than widespread species (Segurado and Araújo 2004). Model accuracy was also found to be influenced by the number of presence records used for model building (Pearce and Ferrier 2000), and predictions based on few records are often seen as weaker than those based on a larger number of samples (Hernandez et al. 2006; Wisz et al. 2008; Mateo et al. 2010b). Only few studies have investigated how phylogenetic relationships are linked to the quality of SDMs (Pöyry et al. 2008) although there is evidence for phylogenetic conservatism among niche parameters (Hof et al. 2010) and range characteristics (Beck et al. 2006a; Jablonski 2008) of species. Different model performance ranking under different conditions of input data would lend support to techniques of quality-weighted model averaging (Araujo and New 2007), whereas one should use the single best modelling technique if there was little data-driven variation in the performance rank of different methods.

In the present study we investigated the performance of eight commonly applied SDM methods in their standard software implementations (see Joppa et al. 2013 for associated problems) while considering some intrinsic characteristics of species and their distributions as covariates. Crucially, we used a sample of species selected to be representative for the Lepidoptera family Sphingidae across the Old World, based on a combination of three different criteria (see Methods), rather than

hand-picking taxa with good and abundant data. We hypothesized that species from climatic zones with supposedly limiting climatic factors, such as cold temperatures in temperate regions or drought in arid regions, will be better modelled in climate-based SDMs than those from humid tropical regions (or with mixed distribution). We also hypothesized that larger sample size is beneficial to model quality, whereas range size should reduce model quality (after controlling for sample size; Segurado and Araújo 2004; Newbold et al. 2009a). We expected differences between phylogenetic lineages, which in sphingids imply considerable life history variation with regard to mobility and dispersal, resource use, reproductive biology, habitat preference and other ecological traits (Beck et al. 2006a, b, c; Beck and Kitching 2007). We evaluated modelling accuracy using three independent methods: Area under the receiver-operating characteristic (AUC), minimal predicted area (MPA), and expert opinion.

2.2. Methods

2.2.1. Species data

We based our study on occurrence records for the Sphingidae from the Old World + Australia/Pacific region. Out of 982 taxa known from the region, we selected taxa (Appendix A) in a stratified design according to three criteria: (a) rarity, quantified in three classes of record numbers (5-10, 11-50, >50; record numbers are lognormal-distributed), (b) climatic zone of occurrence (four classes: humid-tropical, arid, temperate, mixed) and (c) membership in one of seven systematic tribes. Tribal placements were based on a recent molecular phylogeny (Kawahara et al. 2009) and had been shown to impact distribution in earlier analyses of the family (Beck et al. 2006a). For all possible combinations of these criteria (i.e., classes) we randomly selected one species if available. This process led to the choice of 64 species that represent family-wide data variability in these characteristics. Computational limitation prevented us from including more species into the study.

Distributional data were compiled from museums and private collections, correspondence with collectors, publications (including online databases such as GBIF, www.gbif.org) and own fieldwork. All data were carefully checked for reliability of taxonomy (i.e., synonyms, misidentification, etc.) and locality information (i.e., coordinates associated with the locality of each record). Potentially erroneous records (e.g., highly unlikely localities, likely misidentifications) were excluded, and for the purposes of this analysis we also excluded all records that could not be reliably georeferenced to at least 1° latitude/longitude (ca. 110 km; most records were georeferenced with an estimated error <<0.1° latitude/longitude). A “record” is here defined as a unique combination of species, locality, year and collector (or source). Record numbers may hence contain replicates regarding distribution modelling (in time or space, depending on the modelling resolution), and a few occurred at sites outside the environmental data grids used for fitting and prediction. We use the term “sample size” for the number of distributional data that actually entered SDMs.

2.2.2. Environmental data for distribution modelling

We compiled sixteen variables for use as predictors in SDMs (Appendix B). Twelve climatic variables and altitude were extracted from the WorldClim database (v. 1.4; www.worldclim.org; accessed Feb. 2009). This compilation based on interpolations of monthly climate averages from 1950-2000 is commonly used in SDM. In addition, we used vegetation cover data from MODIS continuous fields indicating percent tree, herb and bare ground cover (http://glcf.umiacs.umd.edu/data/vcf; accessed Feb. 2009). All layers were used in a spatial resolution of 2.5 arc-minutes (ca. 5 x 5 km).

2.2.3. Species distribution modelling

From the broad variety of currently available modelling techniques we selected eight SDM algorithms for our comparison: Generalized Linear Models (GLM), Generalized Additive Models (GAM), Generalized Boosting Models (GBM), Classification Tree Analysis (CTA), Artificial Neural Network (ANN), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF) and Maximum Entropy (MAXENT). Some of these (i.e., MAXENT, GBM) were among the top-scorers in the comparison of Elith et al. (2006).

These methods fall into two distinct categories: 1) Regression-type methods (GLM, GAM and MARS) and 2) machine-learning methods (ANN, RF, CTA and MAXENT) (Thuiller 2003; Phillips et al. 2006; Hastie et al. 2008; Marmion et al. 2009). Note that MARS can also be viewed as a simple machine learning method. All methods except MAXENT were calculated within the BIOMOD platform, implemented in R (http://www.r-project.org; Thuiller et al. 2009). For MAXENT we used software provided by Phillips et al. (2006; version 3.3.3e).

SDMs usually require presence-absence data for model fitting and testing, but reliable absences were not available (as for most SDM applications). A commonly used solution is the generation of pseudo-absences (Ferrier et al. 2002; i.e., selected locations are used as absences based on the

assumption that the species really does not occur there)t is important to keep in mind that the measures to evaluate performance will not represent a distinction between presence and absence but rather between presence and random. From the different strategies for generating pseudo-absences incorporated in BIOMOD, we chose to generate a random sample of 10000 points across the research region (Elith et al. 2006) constrained to not fall within in a radius of 40 km around recorded occurrence points (following advice in Mateo et al. 2010a).

MAXENT fits models by using background points instead of pseudo-absences (background points are a random sample across the landscape and may include presence sites). The choice of background samples can be refined by using an externally supplied bias distribution. We used a bias file based on the "target-group absences" approach (Mateo et al. 2010a), i.e. a kernel density distribution grid of our sphingid moth database for all Old World species. This accounts for the fact that some sites are much better sampled than others, and hence a lack of presence for a given species at such sites is much more meaningful than at rarely sampled or entirely unvisited sites. We fitted MAXENT models with both methods of background sampling (random and target-group) and compared results. Only MAXENT models with target-group sampling were used for across algorithm-comparisons.

Various studies have pointed out that SDMs can be sensitive to the choice of research region because it affects background or pseudo-absence selection as well as the predicted area, and an informed *a priori* choice of where the species occurs was advised (VanDerWal et al. 2009; Barve et al. 2011). However, if the aim of SDMs is to provide estimates of geographic distributions for species with very little ecological information available, this demand becomes circular (i.e., we need a SDM to make a good choice), and subsequently decisions are, to a certain degree, arbitrary. Furthermore, because we know more about some species than about others, the uncertainty and error in the *a priori* choice of research regions will differ between species, which further complicates comparison. Using a rather broad calibration area might be more useful when exploring species' unknown distribution (Giovanelli et al. 2010). Moreover, one of our evaluation criteria

(AUC, see below) is sensitive to change in modelling extent (Barve et al. 2011), which would introduce additional variability to model comparison. For these reasons and because uncertainty regarding such decisions probably reflects the typical state for the majority of organisms (i.e., tropical invertebrates) and, by design, all automatized applications of SDM (Guralnick and Hill 2009), we based our main comparisons on models fitted and evaluated across the entire research region (i.e., Old World + Australia). Hence, we compare models under the somewhat naïve approach that we know nothing about the biogeography of species. All final models are averages of 5 replicate models, using a random selection of 75% of occurrence records for model fitting ("training"), and 25% for testing (see below).

2.2.4. Assessment of model quality

In absence of independent presence-absence data for modelled species we used three different approaches to compare predictions generated by the different models.

(1) The area under the receiver-operating curve (AUC) is a widely used but also widely criticized method (Lobo et al. 2008; Jiménez-Valverde 2011) to estimate predictive accuracy independently of threshold (Pearce and Ferrier 2000; Elith et al. 2006). Following standard procedures, we used a cross-validation to retrieve "independent" AUC (25% test data). The averages of AUC from five replicates were used for analysis. AUC ranges between 0 and 1 with 0.5 indicating a random prediction (high AUC indicates good models).

(2) Minimal predicted area size and cut-off threshold (MPA; Engler et al. 2004) is based on the idea that a good map should predict a species' range as small as possible while including most recorded occurrences (i.e., 90% of records). MPA does not require absence data and therefore is independent of pseudo-absence generation (Engler et al. 2004; Rupprecht et al. 2011). MPA was calculated in ArcGIS 10 after transforming data to an equal-area projection by using a specific cut-off threshold for every map. Low MPA indicates good models (note that no direct comparison can be made

across species with different range extent). MPA was calculated for final, averaged models (see above).

(3) Expert opinion is the basis of many commonly used global-scale distributional data sets (BirdLife International 2000; Baillie et al. 2004).The knowledge of experts has also been incorporated at different stages of SDM approaches (Seoane et al. 2005; Murray and Goldizen 2009). One of us (IJK) is a leading expert on the taxonomy, systematics and distribution of the Sphingidae, and we used his opinion as evaluation criterion of the plausibility of SDM range predictions. Although we compiled and processed raw data together, he was not involved in SDM generation. To further increase the independence of assessments, IJK was usually presented range maps anonymously with regard to the SDM algorithm employed. Unlike AUC and MPA, expert opinion excluded model predictions that were clearly far outside a biogeographically reasonably expectation. Additionally to the presence records used for modelling the expert also considered records unknown to the model (e.g. because they were excluded due to low precision of georeferencing), ecological traits of the species (e.g., host plant associations), and reliable absences from very well sampled localities. Models were graded numerically (1-6, with 6 representing best models). Grades were given to final, averaged models (see above).

2.2.5. Evaluating differences in model performance

We analysed the variation of model quality with generalized linear mixed models (GLMMs), using quality metrics (i.e. AUC_{test} (arcsine-transformed), MPA or grades) as response variables (separate GLMM for each quality metric). We fitted as fixed effects model algorithm (8 types), climatic zone (4 types), systematic association (7 tribes), sample size and range size (both continuous). We used a model-independent estimate of range size, the product of latitudinal and longitudinal extent of records (Beck et al. 2006a). Species identity was fitted as random effect (random intercept). In the analyses presented below, no interactions were modelled. However, in preliminary analyses we included, in various combinations, the most interesting potential interactions (e.g., habitat

type*modelling algorithm, sample size*modelling algorithm). These GLMMs did not gain in explanatory power if weighted against increased model complexity (higher deviance information criterion (DIC), all ΔDIC >>50), hence we did not further consider them.

Analyses were performed using *MCMCglmm* package in R (Hadfield 2010), which uses a Bayesian framework with Markov Chain Monte Carlo algorithms. This approach has the advantage of being highly flexible and accurate. We specified a non-informative uniform prior for all the parameters, which is equivalent to a GLMM fitted with a maximum-likelihood approach (Bolker et al. 2009). Comparison between the modelling methods for every model quality metric was summarized after 160000 iterations (burn-in period of 40000). Results are summarized by the mean of the posterior distribution and their 95% confidence intervals, indicating direction, strength and significance of effects.

2.3. Results

Of 512 SDMs, 6 did not converge (i.e., no prediction could be derived; 3 for MARS, 2 for CTA and 1 for GAM; these were not further considered in comparisons). The three metrics of model quality did not lead to consistent assessments when we correlated raw data (pairwise linear regressions: n = 64, all r^2 <0.01); however, in a pairwise GLMM analysis with species as a random effect significant yet weak correlations between the three were recovered in the expected direction (Table 2.1). Weak correlations between these metrics are not unexpected, but they undermine the idea that the "quality" of model can reliably assessed from these metrics.

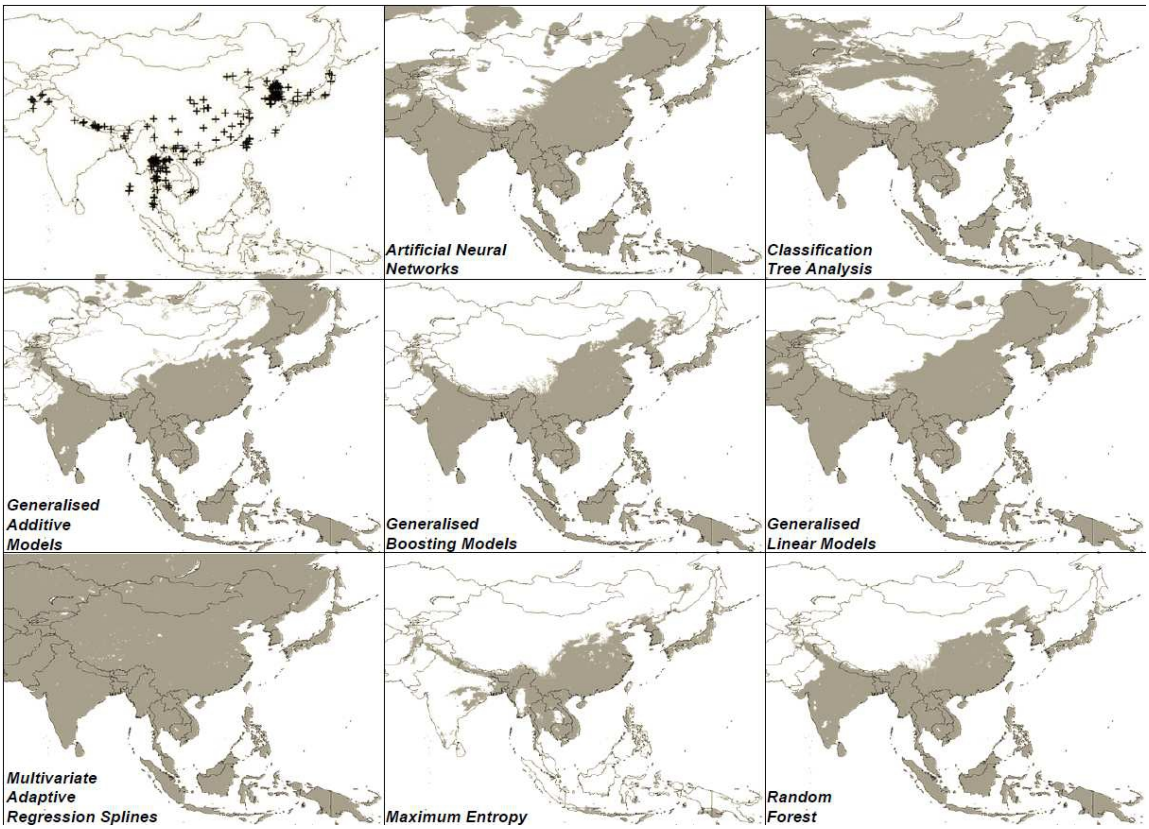
Table 2.1 Correlation between the different quality metrics (normalized and standardized) in pairwise GLMMs with species as random effect: Mean in the posterior distribution (slope), pMCMC (*<0.05, ***<0.01, ***<0.001). Note that good models are indicated by high AUC, high grades, and low MPA.

	AUC	MPA	Grades
AUC		-0.164*	0.192**
MPA	-0.087**		-0.240**
Grades	0.219**	-0.623**	

2.3.1. Differences between model algorithms

Results of GLMMs showed that there were significant differences between the algorithms utilized (see Figure 2.1 for an example). MAXENT was the best method according to MPA and expert-assessed grades for plausibility (see Methods; 'grades' from here on) followed by RF, while ANN was approximately equal with RF and MAXENT according to AUC (Figure 2.2). MARS performed worst according to all criteria. Variation across the data set was large (see confidence intervals in Figure 2.2), which restricts our major results to MAXENT being judged better than most other methods except RF.

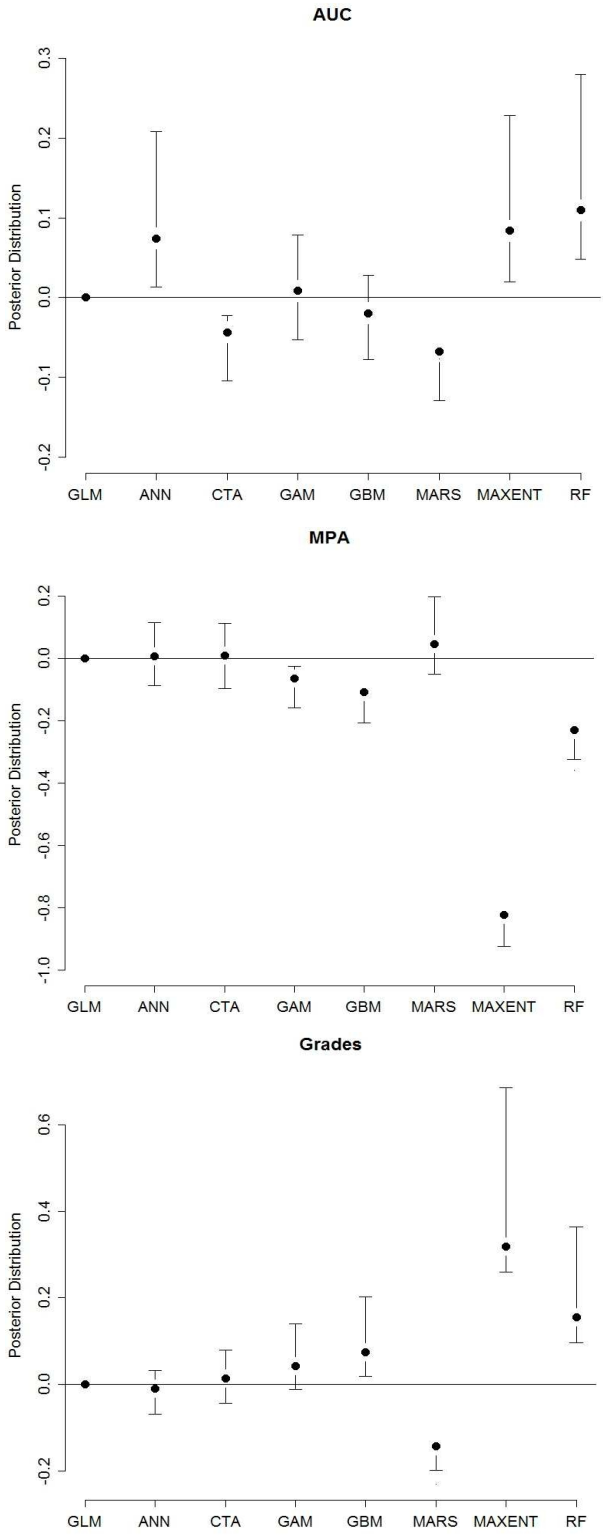
Figure 2.1 Distribution records (black crosses; upper left map) and range estimates (grey) derived from eight SDM techniques, shown exemplarily for the Southeast-Asian species *Psilogramma increta*. Predicted probabilities of occurrence were transformed into a binary presence/absence maps for display, using the MPA cut-off value as threshold. (i.e. predicting at least 90% of records correctly; see Methods).



2.3.2. Effects of distribution region

Model performance did not vary consistently and significantly with the climatic zones of distribution of species (Figure 2.3). Furthermore, the variation contrasted our initial hypotheses of better performance in regions with assumed climatic constraints (i.e., arid and temperate). However, some interesting patterns emerged when analyzing modelling performance across regions from raw data. Regression methods (GLM, GAM and MARS) tended to performed poorer within mixed habitats, whereas machine learning methods seem to perform much better there. When modelling species from arid zones, however, regression methods, especially MARS, performed consistently better than some machine learning methods (CTA, GBM). Contrary, general performance of these methods was poor when modelling species with temperate distribution with the exception of MAXENT that consistently outperformed the others (Appendix 2.3).

Figure 2.2 Posterior distributions (±95% confidence intervals) from GLMMs across different modeling algorithms for (a) AUC (high value = good model), (b) MPA (low value = good model), (c) expert grades (high value = good model). GLM was arbitrarily taken as zero.



2.3.3. Differences between taxonomic groups

Effects of tribal association of species were inconsistent across the three quality assessments and largely non-significant. Species belonging to the tribe Acherontiini (many of which are generalists in larval feeding and have large ranges), in particular, had worse models according to AUC and MPA but not to expert grades (Figure 2.4).

2.3.4. Effects of range and sample size

We did not observe a significant decline in model quality with increasing range size, only a non-significant trend for MPA (mean in the posterior distributions: AUC = 0.033, P = 0.405; MPA = 0.334, P = 0.100; grades = -0.073, P = 0.742). Surprisingly, we did also not find significant effects of (log)sample size on any of the metrics of model quality despite a range from 3 to 889 (in total, four species had a sample size <5; means of the posterior distributions: AUC = 0.029, P = 0.640; MPA = 0.314, P = 0.305; grades = 0.112, P = 0.716). However, when removing one of these collinear variables (Appendix D), simplified GLMMs recovered effects for AUC and MPA (models did not change assessment for grades, nor did any of the other effects change in any model). When removing range size, we found positive effects of sample size on model quality for AUC (posterior distribution = 0.069, P = 0.045) but negative ones for MPA (posterior distribution = 0.743, P <0.001); when removing sample size, however, we found positive effects of range size on model quality for AUC (posterior distribution = 0.049, P = 0.032) but negative ones for MPA (posterior distribution = 0.532, P <0.001). Thus, recovered effects were functionally inconsistent and are therefore most likely artefacts of collinearity. Plots of raw data indicated wedge-shaped relationships of sample size with AUC and MPA, respectively (i.e., large samples are associated with good models, while for small samples sizes there were both bad and good models; Appendix 2.5). However, no such pattern was evident for grades.

We observed large variability in model quality across the 64 species studied; Appendix 2.1). Although some methods occasionally failed badly (i.e., making antithetical predictions) or did not converge, MAXENT and RF were quite consistently well-performing.

.

2.3.5. Effects of biased absences within MAXENT

Regarding the issue of whether models should be fitted with random background points or alternatively by selecting the background points from an external bias file (see Methods), we observed significantly higher AUC with bias file only for training data (t = 3.642, df = 63, P <0.001) but not for test data in cross validation (t = 0.746, df = 63, P >0.45).

Figure 2.3 Posterior distributions (±95% confidence intervals) from GLMMs for distribution types. “Mixed” habitat was arbitrarily taken as zero.

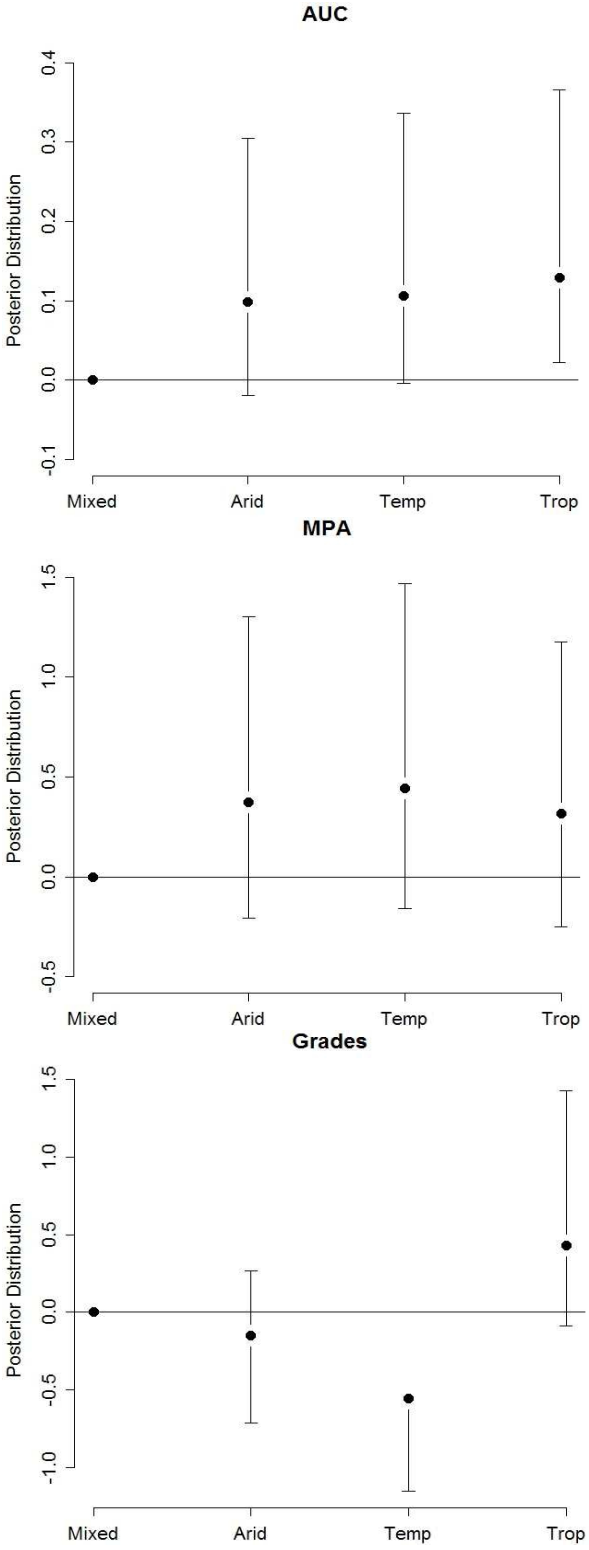
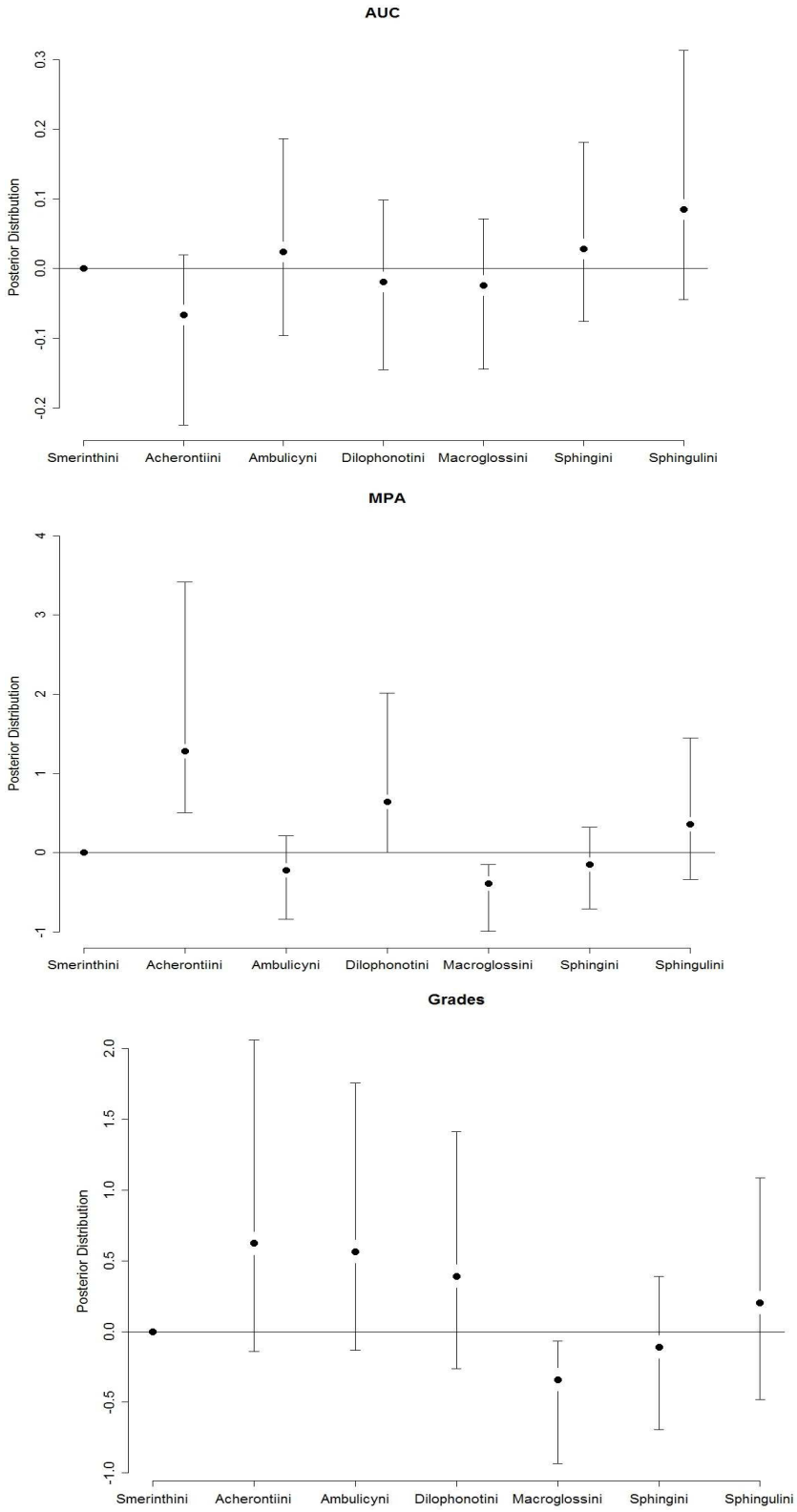


Figure 2.4 Posterior distributions (±95% confidence intervals) from GLMMs for phylogenetic associations (7 tribes). Smerinthini (the tribe with phylogenetically most basal characteristics) was taken as zero.



2.4. Discussion

Our analyses demonstrated major differences in the predictive performance of a range of widely used modelling techniques. In particular, MAXENT, and to a lesser degree *random forest* (Figure 2.2), outperformed other methods, which is consistent with earlier studies (e.g., Elith et al. 2006; Pearson et al. 2007; Cutler et al. 2007; Graham et al. 2008; Philips and Dudik 2008; Giovanelli et al. 2010). However, as our comparison does not strictly compare model algorithms but rather their implementation in available software, we cannot dissect which aspect of MAXENT causes the good performance (as software may also differ in other aspects). Another well-performing method in Elith et al. (2006) was GBM (termed BRT there), which in our analysis had an intermediate performance. Regression-type methods were intermediate and quite similar to each other in terms of performance. Earlier studies also suggested that methods with non-linear fits (GAM, ANN, GBM) are comparable in terms of performance and are usually superior to simple classification trees (like CTA; Muñoz and Felicísimo 2004; Segurado and Araújo 2004). Contrary to Elith et al. (2006) we only used single-species SDM methods in our comparison.

However, despite high between-species variation (controlled as a random effect in analyses; Appendix 2.6) we did not find strong and consistent evidence for the hypothesized effects of taxonomic groupings (i.e., tribal association) and climatic zone of occurrence in multivariate analysis. Several studies have reported negative correlations of modelling performance and range size, proportion of habitat used by the species, area of occupancy or high ecological tolerance (Hepinstall et al. 2002; Segurado and Araújo 2004; Newbold et al. 2009b). The species modelled here spanned a wide range of distributions from almost cosmopolitan to highly localized, and range size was found to correlate with the width of other niche dimensions such as diet breadth in sphingids (Beck and Kitching 2007). Nevertheless, we could not observe consistent links between range size and model quality. Hence, the variation between species (or data sets) that were "good to model" and those that were not seems quite idiosyncratic.

Even more surprising was that we did not find clear and strong effects of sample size. More data may not necessarily lead to better models, although raw data plots indicated that bad models tend to be among those with small sample sizes (Appendix 2.5). Conflicting results (depending on inclusion or exclusion of range or sample size from GLMMs) may be due to the collinearity of these two variables. However, positive range-abundance relationships (Beck et al. 2006d for sphingids) make this collinearity in interspecific comparisons inevitable. While some studies suggested that low sample size negatively influence model performance (Hernandez et al. 2006; Mateo et al. 2010b), Elith et al.(2006) could not corroborate this either. Possibly geographical

sampling bias is more relevant with large sample size, so that the advantages of more information are counterbalanced by the disadvantage of containing less representative information (Loiselle et al. 2008).

An important additional finding was that the three methods of evaluation only correlated with each other after accounting for species-specific differences, and even then agreement was not particularly strong (Table 2.1). Without presence-absence data we have no objective means to know which SDM is closest to the true geographic distribution of a species (however, we tend to put most trust in expert grades). AUC has been criticized for several undesired properties such as dependence on modelling extent and sample size (Barve et al. 2011). AUC calculated on the basis of pseudo-absences might be additionally problematic (Lobo et al. 2008). Despite this, the main results of our study were consistent for all three metrics. The most relevant difference between evaluation criteria was that RF and ANN appeared similar in quality with MAXENT when evaluated with AUC, but less so with MPA and grades.

The SDM community is divided over the question of whether it is necessary to choose a "best" modelling algorithm or if model averaging (Araujo & New 2007, Thuiller et al. 2009) offers an easy way around this decision. With the repeated finding (this study and others) that some methods are consistently better than others, model averaging can only be useful if weighted by model quality (BIOMOD, e.g., offers averaging weighted by AUC). However, if model quality assessment itself is not reliable (i.e., not consistent across criteria) and AUC in particular must be viewed as problematic in this respect, we see much potential for getting worse instead of better predictions from model averaging, without even realizing it. Given the weak, if any, additional predictability of model quality (i.e., no species characteristics effects), our results lead us, at least, to recommend restricting model averaging to generally well-performing methods (e.g., RF and MAXENT). However, we note that this assessment is based on our results regarding best range prediction. As outlined above, regression-type approaches such as GLM may have advantages if the aim is defining or testing niche dimensions of species (Austin 2002). Also, we have not considered, in the absence of good biological knowledge of the test species, how correct ecologically the modelled responses are. Modelling realistic responses are of utmost importance for prediction into new regions or climates (Elith et al. 2010; Svenning et al. 2011).

Choice of spatial scale in SDMs may be confounded by a trade-off of more precise environmental data for fine-grained models on the one hand, and less error in georeferencing (particularly for older museum data) and higher computing speed in coarser grain on the other hand (for general discussion input data-driven uncertainty see Beale and Lennon 2012). A preliminary comparison of SDMs based on different grain size (ca. 1 x 1 km vs. 5 x 5 km, using only MAXENT; data not shown) indicated no relevant differences (AUC and visual comparison), which tentatively supported

coarser-grain modelling to facilitate shorter computing times. We did find, however, support for better models with background sample choice constrained by a bias file to account for unequal sampling across the region. Hence, we tentatively recommend against random background point selection if relevant data are available (Philips et al. 2009). Collector bias in distribution records compilations can be high and relevant for observed patterns (Boakes et al. 2010, Ballesteros et al. 2013 [Chapter 4 in this thesis]), which could considerably weaken SDM quality.

Our comparison of SDM methods has some unique properties, such as considering a broad geographic scale, major distribution types, phylogenetic variation and a wide range of sample sizes with a data set that can be considered representative for an entire systematic group. However, it also has some drawbacks such as lack of true presence-absence data for model evaluation or the need, for the sake of comparability, of modelling across larger regions than would be ideal for some taxa. Given these limitations, however, our data set nonetheless represents the majority of species distribution data that is becoming available with increasing digitization of records from natural history collections. For invertebrates, i.e. the bulk of biodiversity, we often do not have more ecological knowledge than a name and some sites of occurrence, yet these taxa are most in need of SDMs to get justified and detailed estimates of their geographic distributions.

2.5. Conclusions

With a set of species chosen to be representative for an insect family, we can confirm the superior position of the MAXENT method for SDMs (Elith et al. 2006), while we note that the *random forest* method also performed quite well. In light of this and of the finding that model evaluation criteria used for weighting seem not very reliable (i.e., lack of congruence among each other) we suggest to restrict model averaging, if employed, to algorithms that can *a priori* be expected to provide good SDMs. We have not explicitly tested averaged models against single models, but averages of good and bad models must necessarily lead to a weaker performance than that of good models alone. We did not find consistent differences between taxonomic groups (as a proxy of life history variation) or range properties (climatic zone, range size) on model quality, nor did sample size seem to strongly affect model performance. Rather, idiosyncratic differences between taxa or data sets seem to make some species' ranges easier to approximate by SDMs than others.

2.6. Acknowledgements

We are grateful to all the collectors that made their data available for our project (too many to mention here). M. Curran, M. Kopp, R. Hagmann, S. Widler and S. Lang helped with databasing and georeferencing. We received financial support from the Swiss National Science Foundation (SNF, grant no. 31003A_119879).

2.7. References

Araujo MB & New M (2007) Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47

Austin M (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157: 101-118

Baillie J, Hilton-Taylor C, Stuart SN (2004) *IUCN* Red List of Threatened Species: a Global Species Assessment. World Conservation Union, IUCN Glad, Switzerland and Cambridge

Ballesteros-Mejia L, Kitching IJ, Beck J (2011) Projecting the potential invasion of the Pink Spotted Hawkmoth (*Agrius cingulata*) across Africa. International Journal of Pest Management 57:153-159

Ballesteros-Mejia L., Kitching I.J., Jetz W., Nagel P., Beck J. (2013) Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography 22, 586-595

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, Soberón J, Villalobos F (2011) The crucial role of the accessible area in ecological niche modelling and species distribution modelling. Ecological Modelling 222: 1810-1819

Beale CM, Lennon JJ (2012) Incorporating uncertainty in predictive species distribution modeling. Philosophical Transactions of the Royal Society B 367:247-258

Beck J, Kitching IJ (2007) Correlates of range size and dispersal ability: a comparative analysis of sphingid moths from the Indo-Australian tropics. Global Ecology and Biogeography 16: 341-349

Beck J, Kitching IJ, Linsenmair KE (2006a) Measuring range sizes of South-East Asian hawkmoths (Lepidoptera: Sphingidae): effects of scale, resolution and phylogeny. Global Ecology and Biogeography 15: 339–348

Beck J, Kitching IJ, Linsenmair KE (2006b) Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. Ecotropica 12: 1–13

Beck J, Kitching IJ, Linsenmair KE (2006c) Effects of habitat disturbance can be subtle yet significant: biodiversity of hawkmoth-assemblages (Lepidoptera: Sphingidae) in Southeast-Asia. Biodiversity and Conservation 15: 465–486

Beck J, Kitching IJ, Linsenmair KE (2006d) Extending the study of range – abundance relations to tropical insects: sphingid moths in Southeast Asia. Evolutionary Ecology Research 8: 677-690

BirdLife International (2000) Threatened Birds of the World. BirdLife International and Lynx Editions, Cambridge and Barcelona

Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8: e1000385

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24: 127–135

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88: 2783-92

Elith J, Leathwick J (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 265-275

Elith J, Leathwick J (2009) Species Distribution Models: Ecological Explanation and Prediction across Space and Time. Annual Reviews of Ecology, Evolution and Systematics 40: 677-697

Elith J, Graham C, Anderson R, Dudık M, Ferrier S, Guisan A, Hijmans R, Huettmann F, Leathwick J, Lehmann A, Li J, Lohmann, L, Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton Jm, Peterson A, Phillips S, Richardson K, Scachetti-Pereira R, Shapire R, Soberon J, Williams S, Wisz M, Zimmermann N (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods in Ecology and Evolution 1: 330-342

Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263-274

Ferrier S (2002) Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? Systematic Biology 51: 331-363

Ferrier S, Watson G, Pearce J, Drielsma M (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling Biodiversity and Conservation 11: 2275-2307

Giovanelli JGR, de Siqueira MF, Haddad CFB, Alexandrino J (2010) Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. Ecological Modelling 221: 215-224

Graham CH, Elith J, Hijmans RJ, Guisan A, Townsend PA, Loiselle BA (2008) The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45: 239-247

Guralnick R, Hill A (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25: 421-8.

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22

Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York.

Hepinstall JA, Krohn WB, Sader SA (2002) Effects of niche width on the performance and agreement of avian habitat models. In: Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, Samson FB (eds), Predicting species occurrences,. Island Press, Covelo (CA), pp. 593-606

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modelling methods. Ecography 29: 773-785.

Hof C, Rahbek C, Araújo MB (2010) Phylogenetic signals in the climatic niches of the world's amphibians. Ecography 33: 242-250.

Jablonski D (2008) Species selection: theory and data. Annual Reviews of Ecology, Evolution and Systematics 39: 501-524.

Jetz W, McPherson JM, Guralnick RP (2012) Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology and Evolution 23:151-159

Jiménez-Valverde A (2011) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography 21:498–507.

Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan D., Emmott, S., 2013. Troubling trends in scientific software use. Science 340, 814-815.

Kawahara AY, Mignault AA, Regier JC, Kitching IJ, Mitter C (2009) Phylogeny and biogeography of hawkmoths (Lepidoptera: Sphingidae): evidence from five nuclear genes. PloS One 4: e5719

Larner, J., 2008. Safety assessment of cosmetics: an EU perspective. In: Chilcott R and Price S (Eds), Principles and Practice of Skin Toxicology. John Wiley and Sons, England, pp 311-330

Lobo J, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17:145-151

Loiselle BA, Jørgensen PM, Consiglio T, Jiménez I, Blake JG, Lohmann LG, Montiel OM (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35: 105-116

Lomolino MV (2004) Conservation Biogeography. In: Lomolino MV, Heaney LR (eds) Frontiers of Biogeography: New directions in the geography of Nature. Sinauer Associates, Inc. Publishers, Sunderland, MA, pp 293-296

Marmion M, Hjort J, Thuiller W, Luoto M (2009) Statistical consensus methods for improving predictive geomorphology maps. Computers & Geosciences 35: 615–625

Mateo RG, Croat TB, Felicísimo ÁM, Muñoz J (2010a) Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Diversity and Distributions 16:84-94

Mateo RG, Felicísimo AM, Muñoz J (2010b) Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. Journal of Vegetation Science 21:908-922

Murray J, Goldizen A (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies *Petrogale penicillata*. Journal of Applied Ecology 46:842-851

Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15:285–292

Newbold T (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography 34:3-22

Newbold T, Gilbert F, Zalat S, El-Gabbas A, Reader T (2009a) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. Journal of Biogeography 36:2085-2095

Newbold T, Reader T, Zalat S, El-Gabbas A, Gilbert F (2009b) Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. Biodiversity and Conservation 18:3629-3641

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225-245

Pearce J, Ferrier S, Scotts D (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. Journal of Environmental Management 62:171–184

Pearson R, Raxworthy C, Nakamura M, Peterson A (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34:102-117

Phillips SJ, Dudík M (2008) Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161-175

Phillips SJ, Anderson R, Schapire R (2006) Maximum entropy modelling of species geographic distributions. Ecological Modelling 190: 231-259

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19: 181-197

Pöyry J, Luoto M, Heikkinen RK, Saarinen K (2008) Species traits are associated with the quality of bioclimatic models. Global Ecology and Biogeography 17:403-414

Rupprecht F, Oldeland J, Finckh M (2011) Modelling potential distribution of the threatened tree species Juniperus oxycedrus: how to evaluate the predictions of different modelling approaches? Journal of Vegetation Science 22:647-659

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. Journal of Biogeography 31:1555-1568

Seoane J, Bustamante J, Diaz-Delgado R (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology 19:512–522

Svenning JC, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand,S (2011) Applications of species distribution modelling to paleobiology. Quaternary Science Reviews 30:2930-2947

Thuiller W (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9: 1353-1362

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD–A platform for ensemble forecasting of species distributions. Ecography 32 369-373

VanDerWal J, Shoo LP, Graham C, Williams SE (2009) Selecting pseudo-absence data for presence-only distribution modelling: How far should you stray from what you know? Ecological Modelling 220 589-594

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008) Effects of sample size on the performance of species distribution models. Diversity and Distributions 14:763-773

Yates CJ, Elith J, Latimer AM, Le Maitre D, Midgley GF, Schurr FM, West AG (2010) Projecting climate change impacts on species distributions in megadiverse South African Cape and Southwest Australian Floristic Regions: opportunities and challenges. Austral Ecology 35: 374–391

2.8. Appendix

Appendix 2.1 Names, systematic classification (tribe), climatic zone of occurrence, number of records in ca. 5 x 5 km grid cell resolution (i.e. sample size for SDMs) of the 64 sphingid moth species used in analyses. We also report modelling performance for each species, summarized across methods and for three quality metrics (area under the receiver-operator characteristic, AUC; minimal predicted area, MPA; expert grades. Note that our nomenclature results from some as yet unpublished taxonomic revisions.

Species name	Tribe	Climatic zone	Sample size	AUC			MPA (km^2)			GRADES		
				Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Agnosia orneus	Smerinthini	Mixed	3	0.5	0.6925	0.8801	1.36713E+12	4.51424E+12	7.05E+12	1	3.125	5
Agrius convolvuli	Acherontiini	Tropical	889	0.78	0.885	0.93	4.87658E+13	1.35167E+14	1.68E+14	1	3.875	6
Agrius godarti	Acherontiini	Mixed	9	0.5	0.593763	0.833	1.4302E+12	3.87033E+12	5.52E+12	1	2.75	5
Akbesia davidi	Ambulycini	Arid	33	0.98	0.985719	0.9957	6.82169E+11	1.97351E+12	2.96E+12	2	3.125	6
Ambulyx kuangtungensis	Ambulycini	Mixed	48	0.61	0.92765	0.997	3.86917E+12	7.33246E+12	1.22E+13	2	3.375	5
Ambulyx lahora	Ambulycini	Temperate	5	0.5	0.8321	0.9988	62397277909	6.74112E+12	1.57E+13	2	4.125	6
Ambulyx maculifera	Ambulycini	Temperate	8	0.8	0.952313	0.998	1.97175E+11	3.58348E+12	1.74E+13	1	4	6
Ambulyx rudloffi	Ambulycini	Tropical	6	0.86	0.945675	0.9964	12950000000	22069214060	2.56E+10	4	5	6
Ambulyx wildei	Ambulycini	Tropical	30	0.94	0.967825	0.9896	9.9743E+11	1.67742E+12	4.66E+12	3	4.5	6
Amplypterus mansoni	Ambulycini	Tropical	46	0.92	0.970313	0.997	9.21325E+11	8.24986E+12	1.77E+13	2	2.625	6
Apocalipsis velox	Sphingini	Temperate	19	0.9	0.97975	0.995	8.52394E+11	4.34269E+12	1.26E+13	2	3.375	6
Barbourion lemaii	Ambulycini	Temperate	17	0.87	0.9634	0.992	1.99385E+12	6.65659E+12	1.26E+13	1	2.5	6
Callosphingia circe	Acherontiini	Arid	15	0.91	0.964725	0.995	3.80018E+12	4.93458E+12	5.8E+12	1	3	5
Cephanodes banksi	Dilophonotini	Tropical	18	0.86	0.975988	0.998	1.80199E+11	2.4552E+11	3.36E+11	1	5.25	6
Cephanodes hylas	Dilophonotini	Mixed	443	0.82	0.918925	0.998	3.53416E+13	6.76742E+13	1.15E+14	1	3.25	6
Cephanodes janus	Dilophonotini	Tropical	9	0.55	0.98225	0.999	2.3649E+12	6.35521E+13	1.15E+14	1	3	6
Cephanodes kingii	Dilophonotini	Mixed	7	0.88	0.931988	0.988	6.82816E+11	2.18154E+12	7.72E+12	2	5.25	6
Cephanodes rufescens	Dilophonotini	Tropical	10	0.92	0.965663	0.99	2.80975E+11	4.72569E+11	5.59E+11	1	4.125	6
Coelonia brevis	Acherontiini	Tropical	18	0.93	0.979463	0.998	3.5415E+11	6.33008E+13	1.15E+14	5	5.125	6
Coelonia solani	Acherontiini	Tropical	22	0.98	0.986213	0.9927	2.1365E+11	6.32832E+13	1.15E+14	2	5.125	6
Cypa kitchingi	Smerinthini	Tropical	3	0.5	0.797	0.995	31025000000	89787500000	1.33E+11	1	3.625	6
Cypa latericia	Smerinthini	Arid	15	0.7	0.919225	0.996	4.9465E+11	2.67479E+12	3.83E+12	2	2.625	6
Dolbina grisea	Sphingulini	Temperate	29	0.5	0.988329	0.999	1.83578E+12	8.66015E+12	1.18E+13	1	2.5	6
Dolbina inexacta	Sphingulini	Mixed	119	0.96	0.971688	0.995	3.28228E+12	1.1415E+13	1.86E+13	1	3.125	5
Dolbina schnitzleri	Sphingulini	Tropical	4	0.995	0.99754	0.999	24225000000	99449284030	1.47E+11	3	4.25	6

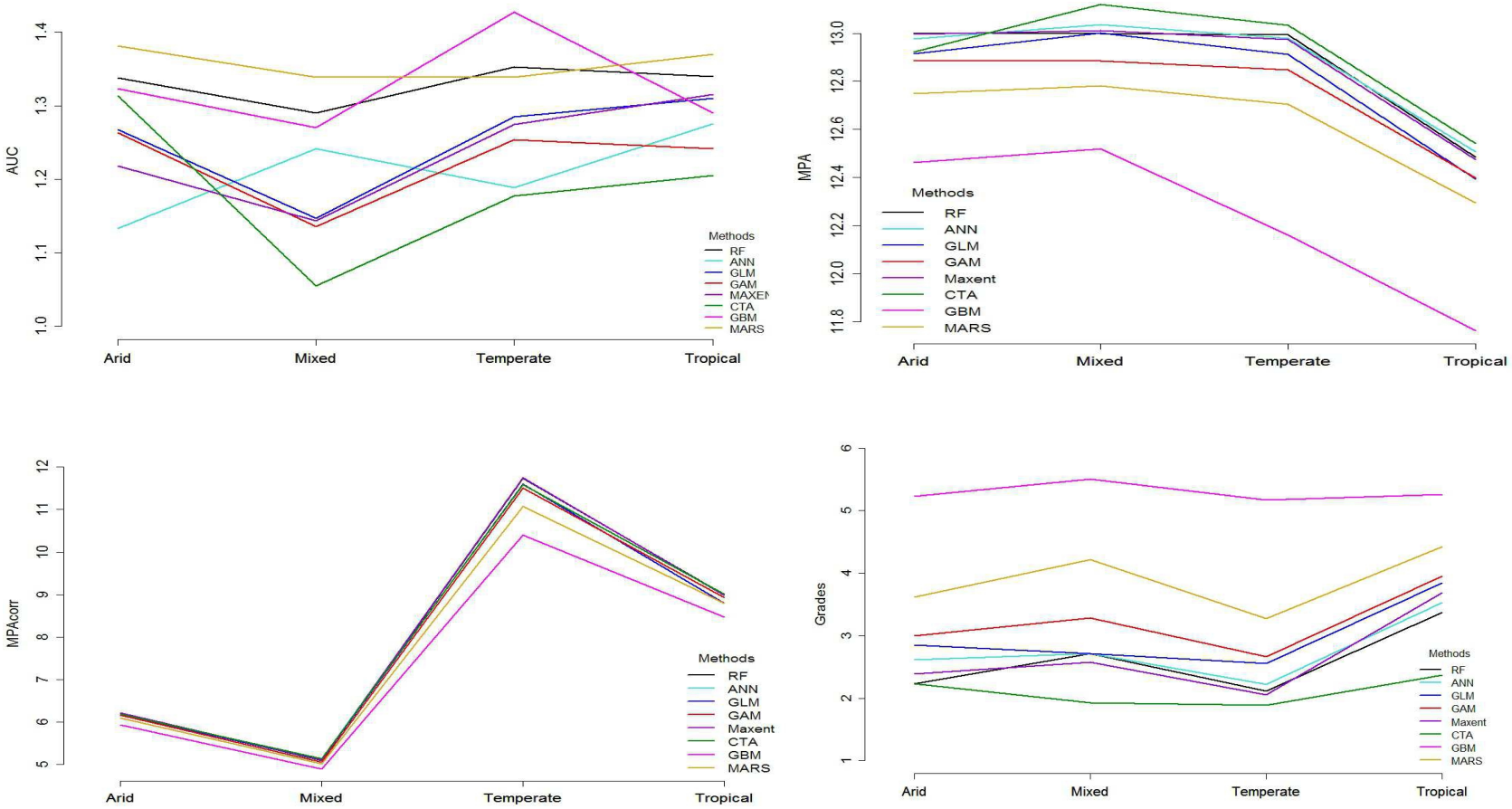
Falcatula svaricki	Smerinthini	Arid	8	0.97	0.988538	0.998	1.5706E+12	3.75661E+12	5.07E+12	2	3.5	5
Hemaris ottonis	Dilophonotini	Temperate	10	0.89	0.953325	0.995	3.75155E+13	7.36959E+13	1.15E+14	1	2.875	6
Hemaris rubra	Dilophonotini	Temperate	7	0.5	0.764686	0.9791	88300000000	1.76642E+13	2.13E+13	1	1.75	6
Hemaris tityus	Dilophonotini	Temperate	889	0.94	0.975538	0.999	3.75155E+13	7.03075E+13	1.15E+14	2	2.75	4
Hippotion scrofa	Macroglossini	Mixed	18	0.91	0.961125	0.997	3.02673E+12	6.30987E+12	8.25E+12	1	3.25	6
Hopliocnema brachycera	Sphingulini	Arid	6	0.75	0.924638	0.999	3.75155E+13	7.21837E+13	1.15E+14	1	3.25	6
Hyles centralasiae	Macroglossini	Temperate	28	0.75	0.83823	0.9786	2.47588E+12	8.17187E+12	1.51E+13	1	2.5	6
Hyles siehei	Macroglossini	Arid	8	0.9	0.981838	0.99	1.50835E+12	8.75443E+12	1.2E+13	1	2	3
Hyles tithymali	Macroglossini	Arid	59	0.82	0.895763	0.978	5.45728E+12	1.98347E+13	2.65E+13	2	2.625	5
Kentrochrysalis streckeri	Sphingulini	Temperate	52	0.89	0.972588	0.997	3.75155E+13	7.76032E+13	1.15E+14	1	2.875	6
Leuucophlebia lineata	Smerinthini	Mixed	156	0.9	0.966088	0.998	3.75155E+13	7.03075E+13	1.15E+14	1	3.875	6
Mapcropoliana gessi	Sphingini	Mixed	5	0.5	0.84125	0.998	1.258E+11	3.99941E+12	5.44E+12	1	2.375	6
Mimas christophi	Smerinthini	Temperate	45	0.93	0.978413	0.998	1.63215E+12	4.73953E+12	6.33E+12	2	3.25	6
Neoogurelca hyas	Macroglossini	Mixed	108	0.9	0.954325	0.998	1.84875E+12	1.48392E+13	4.29E+13	2	3	5
Nephele joiceyi	Macroglossini	Tropical	8	0.92	0.96315	0.9982	2.63325E+11	1.07621E+12	1.23E+12	2	2.625	5
Nephele lannini	Macroglossini	Arid	9	0.74	0.824163	0.9494	1.87613E+12	8.88115E+12	1.05E+13	1	2.5	4
Oligographa juniperi	Sphingini	Arid	8	0.74	0.9525	0.998	1.7065E+11	2.73498E+12	4.41E+12	2	2.5	6
Pantophaea favillacea	Sphingini	Arid	49	0.89	0.973475	0.998	3.75155E+13	7.03075E+13	1.15E+14	3	4.5	6
Pantophaea jordani	Sphingini	Mixed	18	0.95	0.974225	0.991	9.9088E+12	1.94715E+13	2.51E+13	2	3.375	6
Panogena lingens	Sphingini	Tropical	26	0.79	0.943475	0.9958	1.44875E+11	4.35472E+11	5.62E+11	2	3.75	6
Platyshinx bouyeri	Smerinthini	Arid	7	0.75	0.8555	0.985	9.03425E+11	6.5283E+12	1.37E+13	2	3.375	6
Platysphinx_phyllis	Smerinthini	Mixed	18	0.75	0.9126	0.996	8.89128E+12	1.29794E+13	1.59E+13	2	2.875	6
Polyptychus carteri	Smerinthini	Tropical	109	0.75	0.923263	0.971	5.217E+12	1.45242E+13	2.37E+13	2	3.625	6
Polyptychus girardi	Smerinthini	Tropical	26	0.9	0.957863	0.98	5.673E+11	1.25733E+13	2.02E+13	2	3.375	6
Praedora leucophaea	Sphingini	Arid	9	0.79	0.892763	0.988	4.67278E+12	7.31052E+12	8.57E+12	1	2.25	5
Proserpinus proserpina	Macroglossini	Temperate	229	0.96	0.981325	0.996	5.73161E+12	7.76954E+12	8.57E+12	2	3.375	5
Psilograma argos	Sphingini	Tropical	11	0.88	0.9335	0.976	9.96204E+11	3.07099E+12	7.51E+12	1	3.625	6
Psilograma increta	Sphingini	Mixed	225	0.83	0.9463	0.981	8.0873E+12	1.38109E+13	2.49E+13	1	2.75	6
Psilograma salomonis	Sphingini	Tropical	16	0.84	0.926875	0.999	4.16128E+12	7.12136E+12	8.45E+12	2	2.875	6
Rhodambulyx schnitzleri	Smerinthini	Temperate	5	0.93	0.965725	0.998	7.62975E+11	7.49705E+12	1.22E+13	2	2.25	4
Rhodoprasina winbrechlini	Smerinthini	Temperate	8	0.92	0.9634	0.9882	6.54575E+11	7.12278E+12	1.21E+13	1	2.125	4
Sphingulus centrosinaria	Sphingini	Temperate	7	0.5	0.947714	0.995	5.6445E+11	6.93437E+12	9.06E+12	1	2.125	4
Sphingulus maurorum	Sphingini	Temperate	44	0.93	0.964575	0.998	4.80725E+11	1.96661E+12	2.94E+12	1	2.25	5

Sphingulus mus	Sphingulini	Temperate	26	0.94	0.970725	0.988	8.29925E+11	4.96971E+12	7.45E+12	1	2.125	4
Temnora_nitida	Macroglossini	Tropical	8	0.73	0.84175	0.975	4.075E+11	5.12288E+11	5.66E+11	1	2.75	5
Tetrachroa edwardsi	Sphingulini	Arid	8	0.81	0.96475	0.998	5.73161E+12	7.4432E+12	8.57E+12	1	4	6
Theretra cajus	Macroglossini	Mixed	10	0.74	0.8955	0.988	4.403E+11	2.14033E+12	3.01E+12	2	2.5	6
Theretra griseomarginata	Macroglossini	Temperate	6	0.94	0.984438	0.998	68100000000	1.32132E+12	2.17E+12	2	2.625	5
Theretra jugurtha	Macroglossini	Tropical	78	0.83	0.950975	0.995	8.9025E+12	1.71266E+13	2.32E+13	1	3.125	5

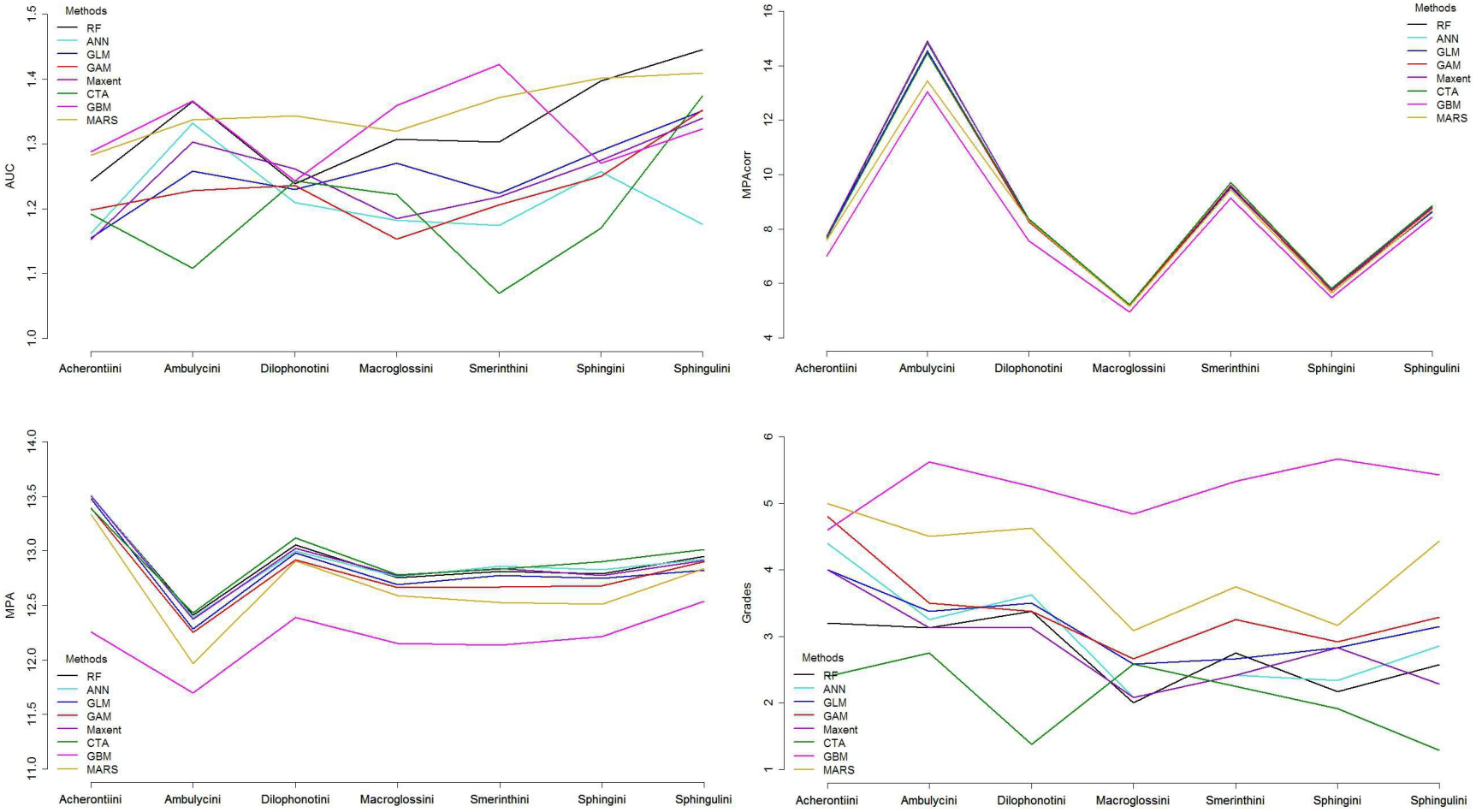
Appendix 2.2 List of predictor variables used in the models.

Variables	**Abbreviation**	**Units**
Altitude	alti	m
Annual Temperature range	antemr	0.1 °C
Annual precipitation	yprecip	mm
Annual Temperature	ytem	0.1 °C
Bare ground cover	bare	%
Herb cover	herb	%
Mean Temperature of the coldest quarter	mtcq	0.1 °C
Mean Temperature of the warmest quarter	mthq	0.1 °C
Mean Temperature of the wettest quarter	mtwq	0.1 °C
Mean Temperature of the driest quarter	mtdq	0.1 °C
Precipitation of the coldest quarter	pcq	mm
Precipitation of the warmest quarter	phq	mm
Precipitation of the wettest quarter	pwq	mm
Precipitation of the driest quarter	pdq	mm
Precipitation seasonality	pseas	mm
Tree cover	tree	%

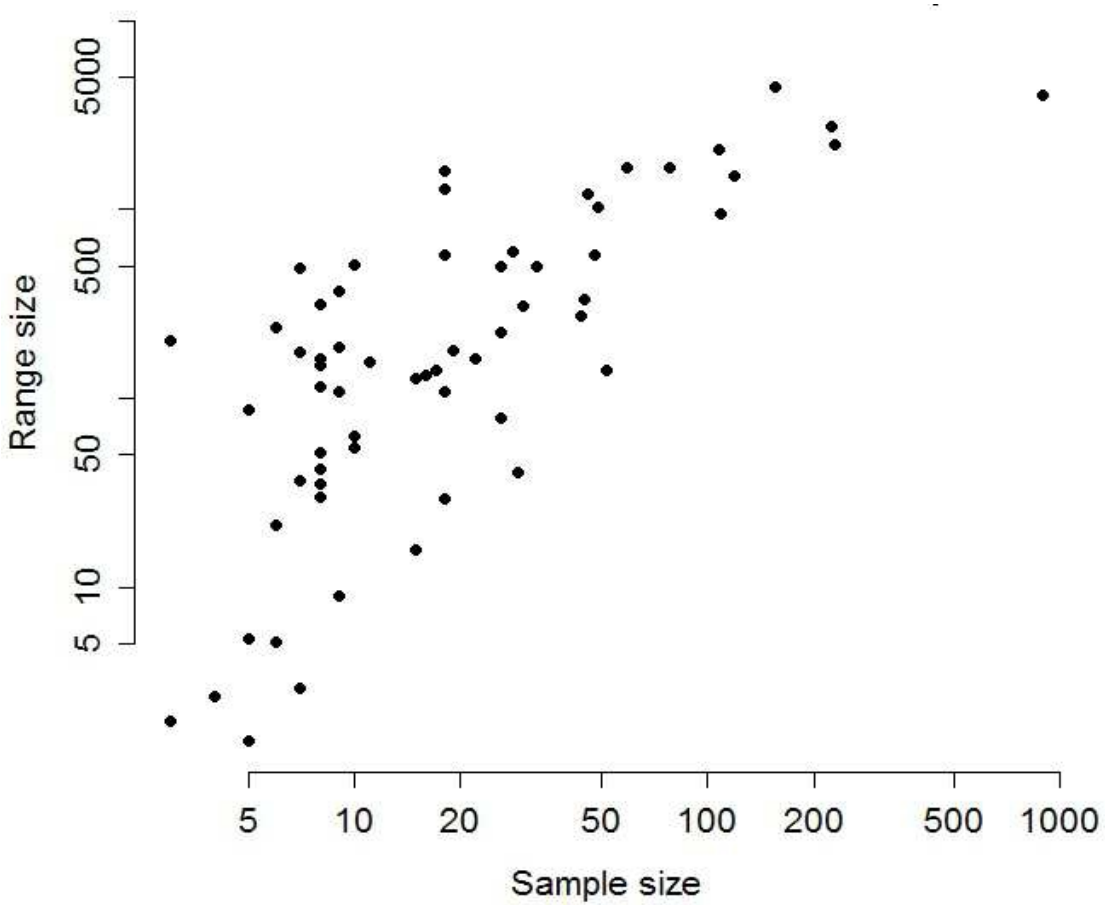
Appendix 2.3 Univariate comparison of modelling performance across the different habitat distribution, for the eight methods used. Habitat distributions are sorted by the mean of each model quality measure i.e. AUC, Grades and MPA across all the species. As MPA varies with the (true) range of species and therefore also with sample size, we used an approximate correction (MPA_{corr}) by dividing MPA through our range size estimate (latitude x longitude extent).



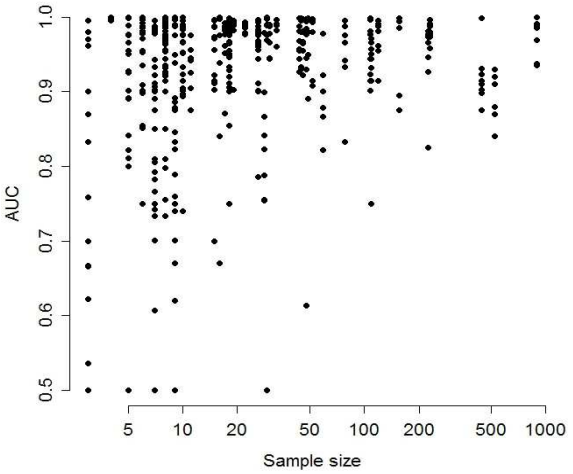
Univariate comparison of modelling performance across the eight different tribe associations, for the eight methods used. Tribes are sorted by the mean of the model quality measure across all the species.

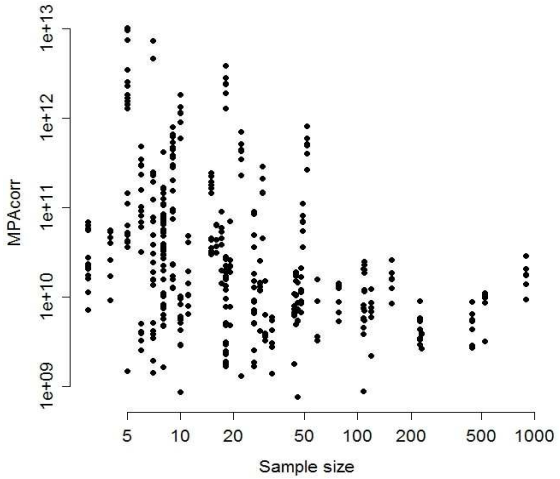


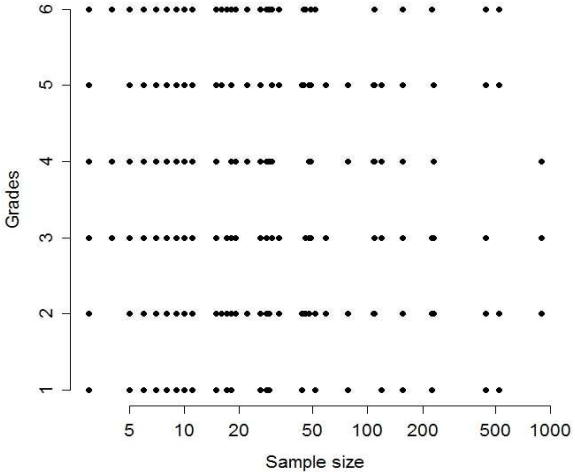
Appendix 2.4 Relationship between range size (latitudinal x longitudinal extent) and sample size (number of 5 x 5km cells with records) of species used in the model. The axes are log-transformed.



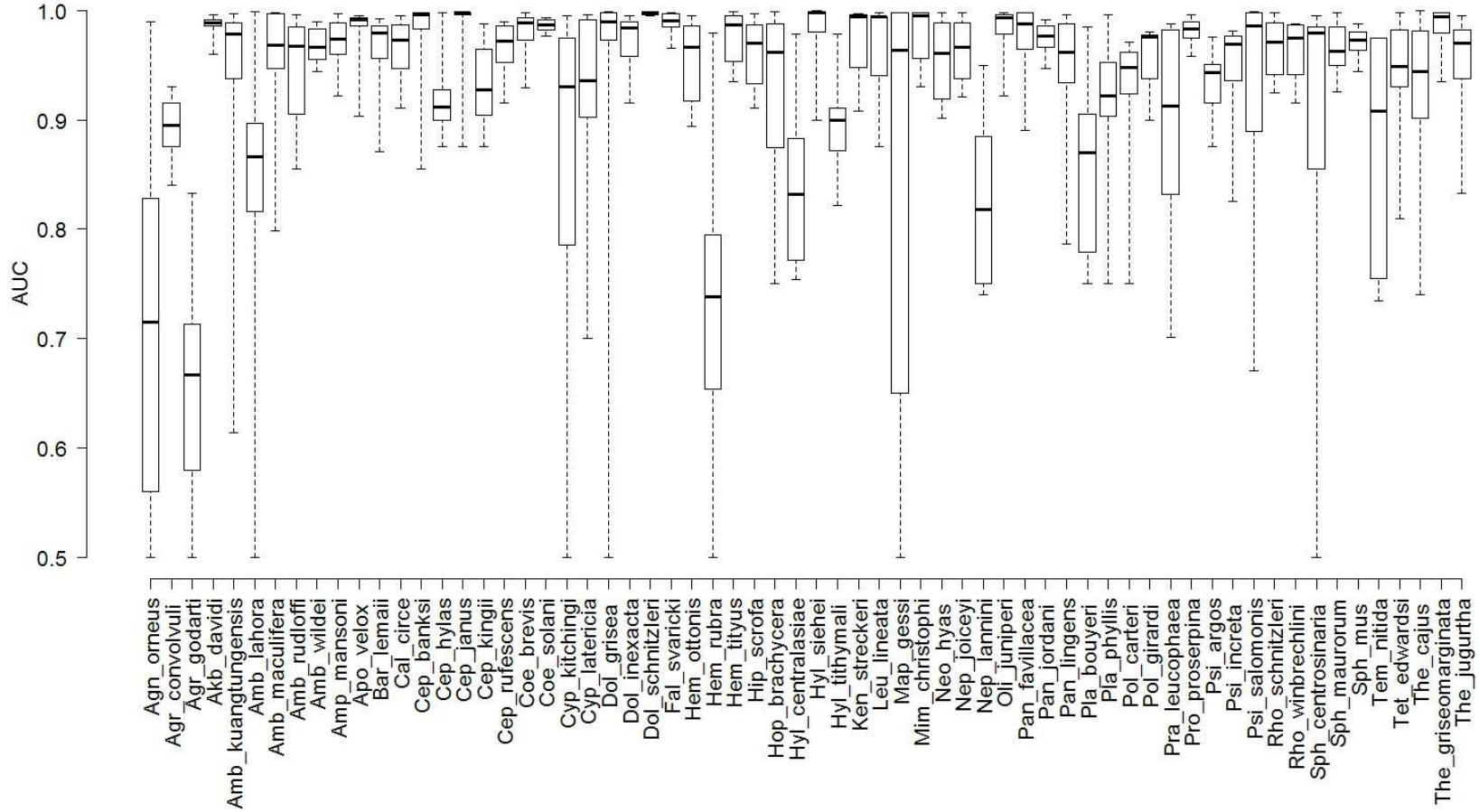
Appendix 2.5 Variation of AUC, MPA and grades with sample size. As MPA varies with the (true) range of species and therefore also with sample size, we used an approximate correction (MPA_{corr}) by dividing MPA through our range size estimate (latitude x longitude extent).

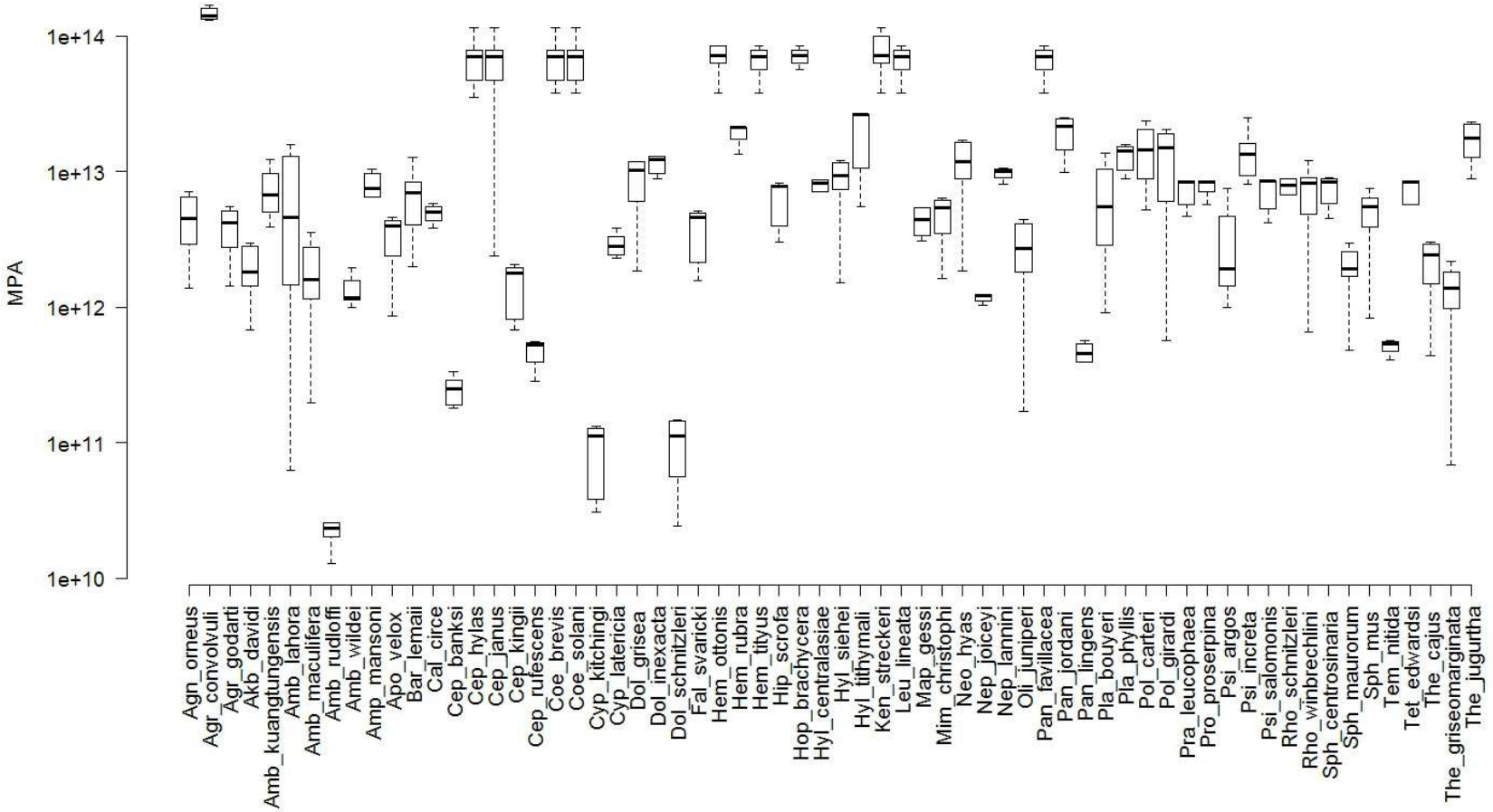


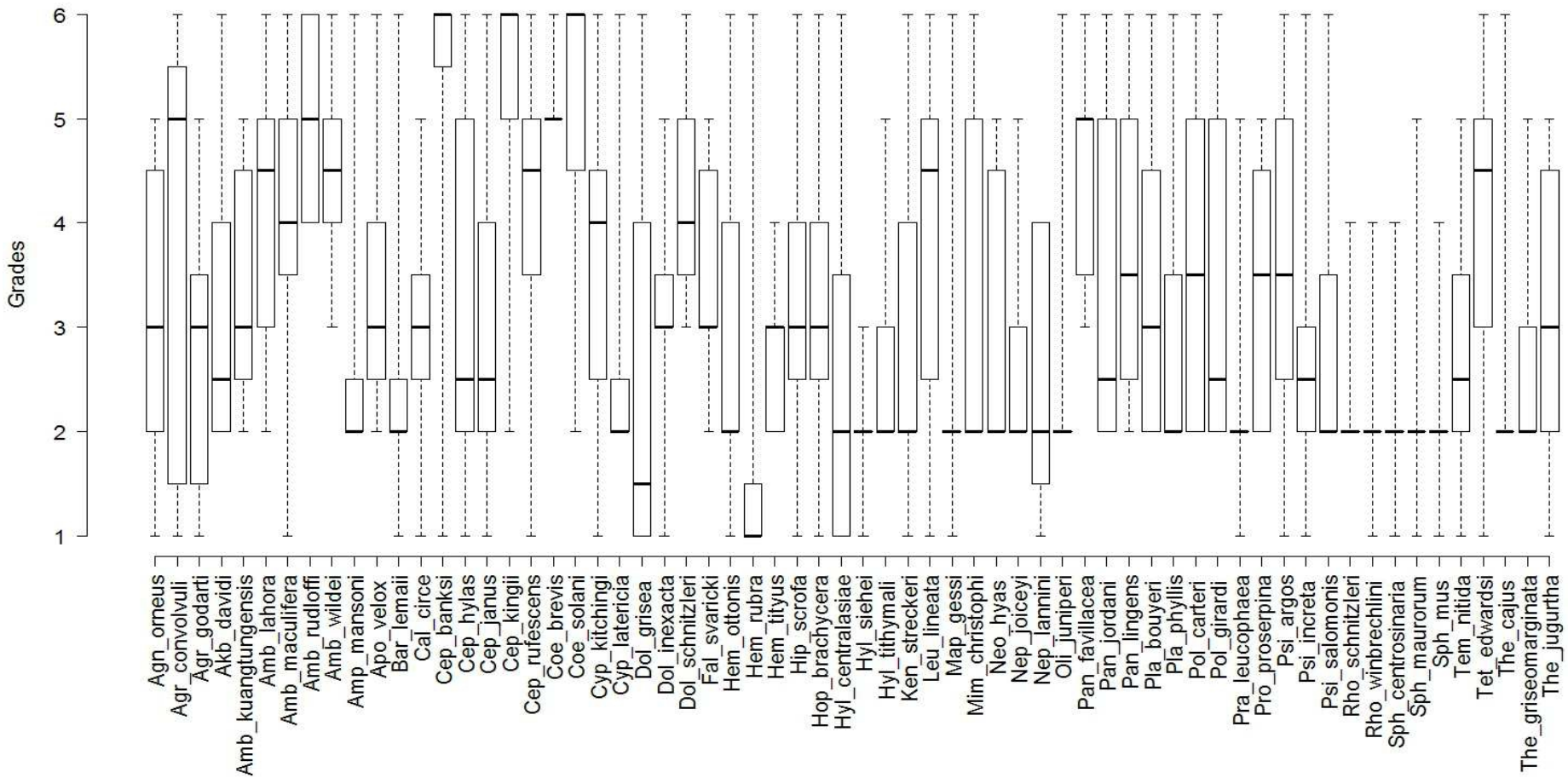




Appendix 2.6 Random effects in GLMMs: Boxplots (median, quartiles, range) of the variation of the three model quality metrics (AUC, MPA, grades) across all species.




MPA
1e+14
1e+13
1e+12
1e+11
1e+10
Agn_orneus
Agr_convolvuli
Agr_godarti
Akb_davidi
Amb_kuangtungensis
Amb_lahora
Amb_maculifera
Amb_rudloffi
Amb_wildei
Amp_mansoni
Apo_velox
Bar_lemaii
Cal_circe
Cep_banksi
Cep_hylas
Cep_janus
Cep_kingii
Cep_rufescens
Coe_brevis
Coe_solani
Cyp_kitchingi
Cyp_latericia
Dol_grisea
Dol_inexacta
Dol_schnitzleri
Fal_svaricki
Hem_ottonis
Hem_rubra
Hem_tityus
Hip_scrofa
Hop_brachycera
Hyl_centralasiae
Hyl_siehei
Hyl_tithymali
Ken_streckeri
Leu_lineata
Map_gessi
Mim_christophi
Neo_hyas
Nep_joiceyi
Nep_lannini
Oli_juniperi
Pan_favillacea
Pan_jordani
Pan_lingens
Pla_bouyeri
Pla_phyllis
Pol_carteri
Pol_girardi
Pra_leucophaea
Pro_proserpina
Psi_argos
Psi_increta
Psi_salomonis
Rho_schnitzleri
Rho_winbrechlini
Sph_centrosinaria
Sph_maurorum
Sph_mus
Tem_nitida
Tet_edwardsi
The_cajus
The_griseomarginata
The_jugurtha


Grades
1
2
3
4
5
6
Agn_omeus
Agr_convolvuli
Agr_godarti
Akb_davidi
Amb_kuangtungensis
Amb_lahora
Amb_maculifera
Amb_rudloffi
Amb_wildei
Amp_mansoni
Apo_velox
Bar_lemaii
Cal_circe
Cep_banksi
Cep_hylas
Cep_janus
Cep_kingii
Cep_rufescens
Coe_brevis
Coe_solani
Cyp_kitchingi
Cyp_latericia
Dol_grisea
Dol_inexacta
Dol_schnitzleri
Fal_svaricki
Hem_ottonis
Hem_rubra
Hem_tityus
Hip_scrofa
Hop_brachycera
Hyl_centralasiae
Hyl_siehei
Hyl_tithymali
Ken_streckeri
Leu_lineata
Map_gessi
Mim_christophi
Neo_hyas
Nep_joiceyi
Nep_lannini
Oli_juniperi
Pan_favillacea
Pan_jordani
Pan_lingens
Pla_bouyeri
Pla_phyllis
Pol_carteri
Pol_girardi
Pra_leucophaea
Pro_proserpina
Psi_argos
Psi_increta
Psi_salomonis
Rho_schnitzleri
Rho_winbrechlini
Sph_centrosinaria
Sph_maurorum
Sph_mus
Tem_nitida
Tet_edwardsi
The_cajus
The_griseomarginata
The_jugurtha

CHAPTER 3

Online solutions and the 'Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges?

Jan Beck[1*], Liliana Ballesteros-Mejia[1], Peter Nagel[1], Ian J Kitching[2]

1) University of Basel, Department of Environmental Science (Biogeography Section), Basel, Switzerland

2) The Natural History Museum, Department of Entomology, Cromwell Road, London SW7 5BD, UK

*) corresponding author: jan.beck@unibas.ch

Published in: Diversity and Distributions (2013). 1:8. DOI: 10.1111/ddi.12083.

Abstract

Aim: To investigate the contribution to range filling, range extent and climatic niche space of species of information contained in the largest databank of digitized biodiversity data: the Global Biodiversity Information Facility (GBIF). We compared such information with a compilation of independent distributional data from natural history collections and other sources.

Location: Europe.

Methods: We used data for the hawkmoths (Lepidoptera, family Sphingidae) to assess three aspects of range information: 1) Observed range filling in 100 x 100 km grid cell squares, 2) observed European extent, and 3) observed climatic niche. Range extents were calculated as products of latitudinal and longitudinal extents. Areas derived from minimum convex polygons drawn onto a 2-dimensional niche space representing the two main axis of a principal component analysis (PCA) were used to calculate climatic niche space. Additionally, record-based permutation tests for niche differences were carried out.

Results: We found that GBIF provided many more distribution records than independent compilation efforts, but contributed less information on range filling, range extent, and climatic niches of species.

Main conclusions: Although GBIF contributed relevant additional information, it is not yet an alternative to manual compilation and databasing of distributional records from collections and literature sources, at least in lesser-known taxa such as invertebrates. We discuss possible reasons for our findings, which may help shape GBIF strategies for providing more informative data.

Keywords: Climatic niche space, Global Biodiversity Information Facility (GBIF), Lepidoptera, Natural history collections, Species' range extent, Sphingidae.

3.1. Introduction

Knowledge on species' distributions is, for most of the organisms on Earth, very scarce - a situation that has been dubbed the 'Wallacean shortfall' (Lomolino 2004). Furthermore, much of the existing distributional data are scattered throughout a multitude of sources, such as taxonomic publications, checklists, and natural history collections. As such the problem is part of the wider fragmentation of the taxonomic and systematic information knowledge base (Godfray *et al.* 2007; Scoble *et al.* 2007; Clark *et al.* 2009). This leads to considerable input in time and effort being necessary to compile data comprehensively. With increasing technological development of computing and analytical tools to make use of such "presence-only" distributional information, such as species distribution modelling (SDM; Elith & Leathwick 2009), there is a high demand to make such data more easily and quickly available (Jetz *et al.* 2012). Invertebrates, particularly insects, are heavily underrepresented in macroecological studies despite their major contribution to global biodiversity (Beck *et al.* 2012), which is almost certainly due to data shortage. Successfully addressing the 'Wallacean shortfall', therefore, will be, to a large degree, about providing data on insect distributions.

The Global Biodiversity Information Facility (GBIF) provides free access to digitized ecological data from different sources (e.g. museum collections, survey programs, etc.) as a result of collaborative endeavours between data providers and taxonomists across many institutions. This information is collated online into a searchable database. Being the largest initiative of its kind, GBIF will certainly play an important role in scientists' attempts to close the gap in species distributional knowledge. On the one hand, accessing GBIF data is comparably fast in comparison to compiling data from original sources, making large-scale multi-taxon analyses feasible in relatively short timeframes. On the other hand, GBIF content has also been strongly criticized due to, for example, data quality issues (Soberon *et al.* 2002; Graham *et al.* 2004; Yesson *et al.* 2007). A particular strength of GBIF is its easy combination with SDM, where potential ranges of species are calculated from climate-based correlations. An important aspect of GBIF data is therefore how well it represents the occurrence of species in climatic niche space.

In this study, we aim to contribute to the evaluation and improvement of GBIF from the perspective of an exemplar insect taxon. We investigate how much distributional data the Initiative provides in a "quick and easy" manner compared to the much more laborious compilation of data from original sources. In particular, we investigated contributions to knowledge of range filling, range extent and climatic niche space of species.

We used sphingid moths (Lepidoptera: family Sphingidae) as model taxa for these analyses, as we have already compiled distributional data independently from GBIF that allow us to compare these

two approaches. Sphingids are among the best-known insect groups, but their distributional data are nevertheless much more incomplete than, for example, European vascular plants or birds. Currently, data on invertebrates are very scarce in GBIF for non-industrialized countries, so we restricted our comparison to Europe.

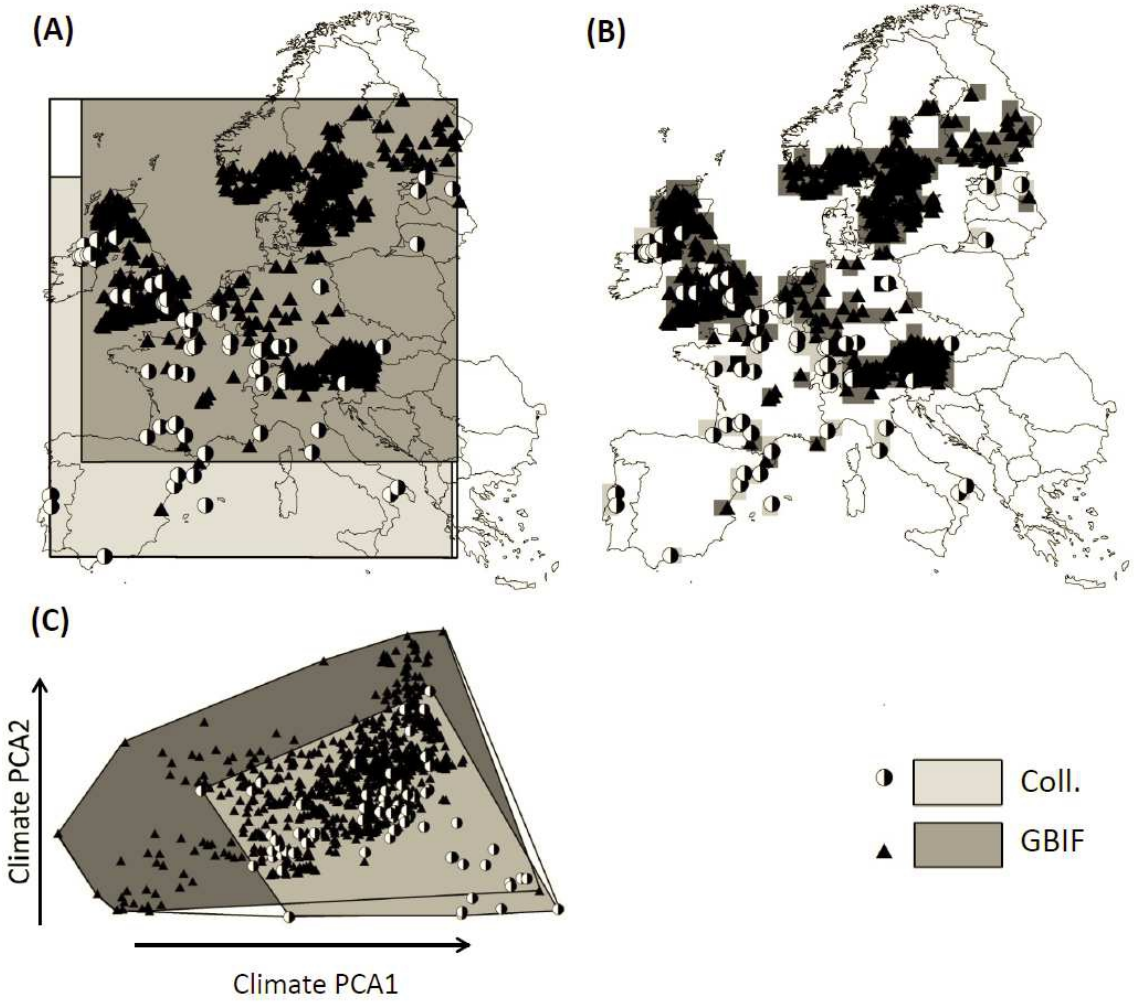
3.2. Methods

We compiled distributional records from a large number of sources, such as private and public natural history collections (see Appendix 3.1. for detail) and faunistic publications. We carefully checked data for credibility, taxonomy and nomenclature, and georeferenced locality data using atlases, gazetteers and websites such as Google Earth. For brevity, we call this compilation *independent compilation* data because the large majority comprises previously unpublished data from collections. In November 2009, we downloaded all available GBIF records for the family Sphingidae and processed these data in the same way, i.e. checking nomenclature and georeferencing, and excluding records deemed to be erroneous or where missing locality information could not be supplemented with reasonable effort. We acknowledge that more data may subsequently have become available from GBIF, but quality control is of utmost importance and requires considerable time to undertake. Data from both sources covered a time-frame of >170 years, with GBIF data being on average a bit older (median [lower, upper quartile] = 1984 [1958, 1995] than collection data (median = 2001 [1965, 2007]). Although this difference is statistically significant (no details shown), it is probably not relevant for the topics studied here. For the purposes of our analyses, we excluded all data that could not be georeferenced with a precision <0.1° latitude/longitude (i.e., ca. 11 km at most), as a coarser resolution may distort SDM attempts on such data. This excluded, in particular, coarse-scale data (>1°) provided in monographs such as Danner *et al.* (1998).

GBIF data for sphingids (as for many other invertebrate groups) make no substantial contribution outside the industrialized countries (Newbold 2010), which are currently GBIF's main data providers. To make the comparison between the two data sources as fair as possible, we therefore restricted our analyses to Western and Central Europe (excluding Iceland, Cyprus, the Canary Islands and the Azores; Fig. 3.1). We considered 32 species of sphingids found in the region (Table 1) while discarding records of Afrotropical *Polyptychus trisecta* (Gibraltar: probably transported by ship) and *Leucophlebia edentata* (Merjenje, Slovenia: probably location error). However, some species are based on records from the extreme edge of otherwise non-European distributions (e.g., African *Theretra osiris* in Gibraltar), or they are only summer visitors in most of their European range (e.g., *Agrius convolvuli*). We therefore repeated analyses using only the 25 autochthonous

European species. The taxonomy of some species is still tentative pending further molecular and morphological study (i.e., the *Hyles euphorbiae*-complex; Hundsdörfer *et al.* 2009, 2011).

Figure 3.1. Measurement of range extent (A), range filling (B) and climatic niche (C), exemplified by data for *Deilephila elpenor*. (A) Records from GBIF and *independent compilation* (Coll.) in Europe, and latitudinal x longitudinal extents. Combined extent is indicated by the large white rectangle in the background. (B) Filled 100 x 100 km cells for both data types. Cells filled by both types of records are drawn in black. (C) Minimum convex polygons in 2-dimensional climatic niche space based on record types (based on PCA, see Methods). Combined extent is indicated by the large white rectangle in the background.



We assessed three aspects of range information (Fig. 3.1): observed range filling, observed European extent, and observed climatic niche space. We measured observed range filling as the number of 100 x 100 km squares from which records are known. This resolution of analysis was chosen for three reasons. (1) Range information based on "expert opinion" is typically at (implicit) resolutions between 100-200 km (Jetz *et al.* 2012), and (2) many studies publish or analyse such data at scales between 50 x 50 and 200 x 200 km (e.g., Danner *et al*. 1998; Settele *et al*. 2008; Jetz & Fine 2012; Ballesteros *et al*. 2013). Furthermore (3), our analyses of range modelling based on GBIF data (M. Böller, W. Schwanghart & J. Beck, unpublished data) indicated that a higher density of records can produce models of lower rather than higher quality due to spatial bias in sampling.

Comparisons of datasets with different degrees of spatial auto-correlation are generally scale-dependent (e.g., Wiens, 1989; Legendre *et al.* 2002; Schwanghart *et al.*, 2008), and we can expect that below a certain grain size comparisons between GBIF and *independent compilation* data would be mainly driven by record numbers, not by the spatial coverage of ranges. We counted the number of cells known exclusively from *independent compilation* or from GBIF data respectively. We expressed these figures as a percentage of the total number of European cells known for each species (i.e., *independent compilation* plus GBIF).

We measured observed extent as the product of longitudinal (X) and latitudinal (Y) range (measured in km) within Europe (cf. Beck *et al.* 2006, for correlation with more detailed range metrics). Again, we expressed the observed extent according to *independent compilation* and GBIF as a percentage of the extent observed from both sources. To check how comprehensively these point-locality data covered "true" range extents, we visually compared the combined range extent (GBIF and *independent compilation*), which we used as reference for true range extent here, with available coarser-scale distributional data in Danner et al. (1998) for five widespread European species (*A. convolvuli, D. nerii, D. elpenor, H. celerio, H. livornica*). Our combined GBIF and *independent compilation* data had quite similar range extents as published range data (indicating range edges further in the north, whereas those shown in Danner et al. (1998) were further southeast for some species).

To measure observed climatic niche space, we first extracted data for eight climatic variables (annual mean temperature; temperature in the hottest month; temperature in the coldest month; annual temperature range; annual precipitation; precipitation in the wettest month; precipitation in the driest month; precipitation seasonality) from WorldClim (www.worldclim.org) at a 5 x 5 km resolution for all of Europe. Climatic means, extremes and variability can be assumed to affect species' distributions. We performed a principle components analysis (PCA; conducted in R package *ade4*, Chessel *et al.* 2004), the first two axes of which explained two-thirds (39.0% and 27.6%, respectively) of original data variability. After mapping these two axes across Europe, we extracted values for species records and plotted these in two-dimensional niche space for each species. We drew minimum convex polygons around these records based on all data, *independent compilation* data only and GBIF data only. We then measured the proportions of polygon areas from *independent compilation* and GBIF data respectively, in comparison to the 'all data'-polygons. However, while this analysis of climatic niches is intuitive, it is possibly over-simplistic. It ignores the density of records in niche space and is driven by most extreme records, which may be affected by sample size. We reanalysed niche comparison with a more sophisticated method that utilized information on the density of records, expressed as density kernel in PCA-space of climatic variables (Broennimann *et al.* 2011). Randomization tests allowed comparing the equivalence of

niches (i.e., *independent compilation*, respectively GBIF, vs. combined data). If GBIF and *independent compilation* were random samples from the same data pool, we expect no significant rejection of niche equivalence.

We compared differences between *independent compilation* and GBIF data for these three aspects of distribution. Because some data were not normal distributed, we based most analyses on rank data. For statistical testing of median differences between *independent compilation* and GBIF contributions across species, we applied Wilcoxon matched pair tests.

3.3. Results

An overview of the raw data is given in Table 1. For 32 species, we had 3,537 records from *independent compilation* and 23,986 from GBIF. Per species, many more records were available from GBIF compared to *independent compilation* (Fig. 3.2; Wilcoxon test: N = 32, Z = 2.73, p = 0.006; restricted to European species *sensu stricto* (see Methods): N = 25, Z = 2.46, p = 0.014). However, *independent compilation* contributed substantially more to the observed range filling (higher percentage of 100 x 100 km cells exclusively known from *independent compilation*; N = 32, Z = 2.00, p = 0.046). This result was not significant if data were restricted to European species *s.s.* (N = 25, Z = 1.74, p = 0.083). *Independent compilation* also led to larger observed European range extents (as a fraction of total known European extent for each species; N = 32, Z = 3.49, p <0.001; restricted to European species *s.s.*: N = 25, Z = 2.97, p = 0.003). Lastly, *independent compilation* data make a larger contribution to observed European niche space than GBIF data (N = 31, Z = 2.63, p = 0.009; restricted to European species: N = 25, Z = 2.33, p = 0.020).

Niche comparisons based on kernel densities of records are shown in Table 2. As density kernels of combined data are strongly affected by the much more numerous GBIF data, it is unsurprising that we found, in apparent contrast to niche polygon analyses (above), a higher niche overlap of GBIF with combined data, rather than *independent compilation* with combined data (Wilcoxon test: N = 22, Z = 2.42, p = 0.016; European species: N = 17, Z = 2.49, p = 0.013). Nevertheless, our previous conclusions are supported insofar as niche equivalence of both *independent compilation* and GBIF with combined data was rejected for most species. That is., neither data type resembles combined data; for some species they are not more similar than a random draw. Plots of data in niche space

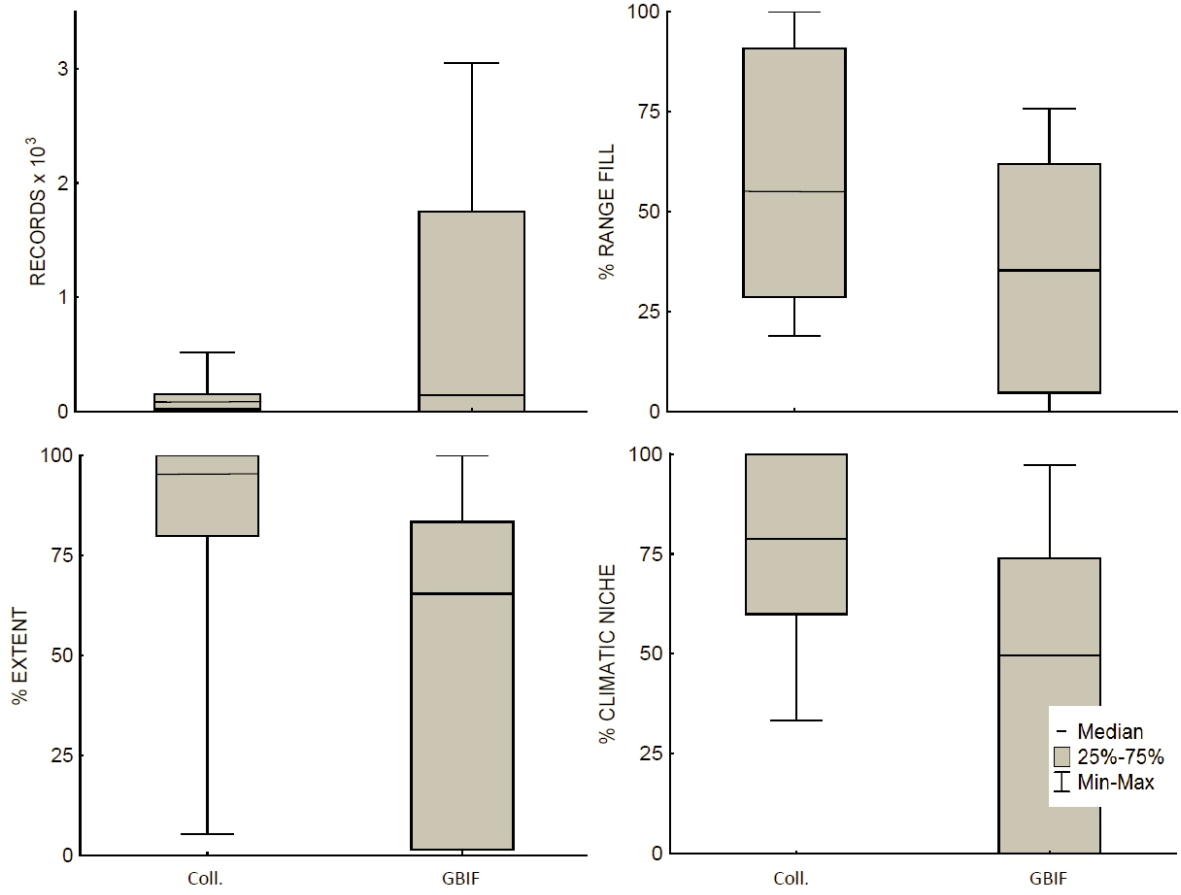
(Electronic Supplements) indicated for many species broader observed niches for *independent compilation* despite their lower sample sizes.

Despite the significantly larger contribution of *independent compilation* data to most of these aspects of known distributions, Fig. 3.2 indicates that GBIF does make considerable contributions to range filling and niche space. Only for range extent was *independent compilation* data close to the total known (European) extent of the species (with the exception of a few outliers; see e.g. median and quartiles).

Table 3.2. Results from density kernel-based analyses of climatic niches. For species with sufficient records, *independent compilation* and GBIF data, respectively, are compared to combined data. The *D* metric measures niche overlap (0 = no overlap, 1 = complete overlap). Two types of null model simulations were carried out and tested for significance (* indicates $p < 0.05$) in 100 replicate runs. $D_{sim}(eq)$ is the expected niche overlap for equivalent niches (significant rejection means data are not equivalent to combined data). $D_{sim}(sim)$ simulates a random draw from niche space, test results indicate whether observed data are significantly more similar to combined data than expected from chance. Collection and GBIF data were significantly non-equivalent to each other for all tested species (not shown).

	Coll.			**GBIF**		
Species	**D_{obs}**	**$D_{sim}(eq)$**	**$D_{sim}(sim)$**	**D_{obs}**	**$D_{sim}(eq)$**	**$D_{sim}(sim)$**
A. atropos	0.495	0.931*	0.103*	0.704	0.929*	0.124*
A. convolvuli	0.441	0.921*	0.318	0.743	0.926*	0.411*
D. nerii	0.505	0.841*	0.091*	0.557	0.858*	0.068*
D. elpenor	0.321	0.928*	0.040*	0.900	0.951*	0.116*
D. porcellus	0.505	0.921*	0.245	0.935	0.940	0.248*
H. fuciformis	0.235	0.919*	0.073*	0.885	0.943*	0.147*
H. tityus	0.475	0.921*	0.128*	0.980	0.949*	0.125*
H. celerio	0.871	0.857	0.061*	0.335	0.844*	0.221
H. euphorbiae	0.312	0.919*	0.232	0.770	0.933*	0.061*
H. gallii	0.270	0.924*	0.269	0.947	0.946	0.254*
H. hippophaes	0.959	0.837*	0.202*	0.179	0.781*	0.085
H. livornica	0.752	0.910*	0.088*	0.428	0.907*	0.091*
H. vespertilio	0.587	0.831*	0.189*	0.618	0.845*	0.194*
L. amurensis	0.410	0.817*	0.047	0.914	0.859*	0.022*
L. populi	0.349	0.938*	0.150*	0.947	0.961*	0.193*
M. stellatarum	0.442	0.929*	0.298*	0.812	0.945*	0.266*
M. quercus	0.838	0.869	0.064*	0.459	0.863*	0.155*
M. tiliae	0.366	0.922*	0.034*	0.812	0.941*	0.104*
P. proserpina	0.566	0.892*	0.231	0.756	0.887*	0.186*
S. ocellata	0.315	0.922*	0.239	0.838	0.944*	0.245*
S. ligustri	0.365	0.919*	0.213	0.949	0.943	0.325*
S. pinastri	0.431	0.935*	0.216*	0.947	0.959	0.249*

Figure 3.2. Number of records per species (*upper left*); proportion of 100 x 100 km cells containing exclusively records from *independent compilation* (Coll.) or GBIF, respectively (*upper right*); proportion of European range extent (longitudinal x latitudinal extent; *lower left*); proportion of climatic niche representation (based on polygon areas in 2-dimenasional niche space; *lower right*). All differences are statistically significant.



3.4. Discussion

The geographic range of a species is a basic and important unit of information in biogeography (Brown *et al.* 1996). GBIF can doubtless play a major role in making huge amounts of distributional records easily accessible, and it is rightfully seen as an important stepping stone to addressing the 'Wallacean shortfall' (Jetz *et al.* 2012). Through our analysis we hope to provide some constructive answers as to how GBIF can be made more useful to potential end-users of these data, i.e. macroecologists and biogeographers.

Our study, based on a taxonomic family for which we had relevant data available, was biased towards GBIF in two aspects. First, we restricted our analyses to Europe, for which GBIF has a much better coverage than for many other regions (e.g., for Africa, GBIF provided 42 sphingid records, for Southeast-Asia only 28; see also Yesson *et al.* 2007). Second, our independently assembled database of European records from natural history collections was a rather cursory by-product of studies focused primarily on collating distribution records of the tropical representatives of the family. Many more data could have been included for Europe if we had made it a priority in these studies.

Despite this, and despite the fact that many more records were available from GBIF than from *independent compilation*, we found that *independent compilation* data contributed more to our total knowledge of ranges (filling and extent) as well as covering the climatic niches of species more comprehensively. Thus, with the caveat that the situation may be different in other taxonomic groups with a much better coverage in GBIF, we conclude that for any detailed biogeographical study GBIF data cannot yet be viewed as an alternative to laborious data compilation from primary sources, at least for lesser-known taxa such as invertebrates. However, our analysis also showed that GBIF contributed significant amounts of information (on range filling and niche space in our definitions) - *independent compilation* data alone did not provide the whole picture. Combined with the ease of accessing GBIF data, this highlights its usefulness as a supplement to independent data compilations.

We can only speculate on the reasons for the rather surprising result that GBIF provides more, but less informative, data compared to *independent compilation*. An even spatial representation is probably an important feature for the biogeographic relevance of distribution data. GBIF, due to its country-based policies of funding and dataset contribution, shows large inequalities in regional data availability (Yesson *et al.* 2007) – for example, species-poor southern Scandinavia is very densely sampled, whereas the Balkans (which are rich in endemics for many taxa) are not (see Electronic Supplements). As a consequence, in our study taxon, GBIF data has gaps particularly in the rare, locally more restricted taxa found in south-eastern Europe (see Electronic Supplements).

Furthermore, species observations in local surveys contribute considerable to GBIF (Electronic Supplements). Guralnick & Van Cleeve (2005) pointed out that museum collections may over-represent rare taxa, hence giving more complete information on species richness at lower specimen numbers, whereas survey tend to contain a lot of data on common species.

When compiling our independent distribution data from original sources, we often found it useful to shift attention among species and regions depending on data availability, considering that costs (and work-time) per specimen are the same but databasing rare species from an undersampled region contains more novel information than duplicating records for well-known species and regions. Implementing similar strategies for allocating funds is admittedly much more challenging in a huge collaborative project like GBIF, but it may be one path towards attaining more comprehensive geographical coverage. We also found private collections to be highly valuable sources of distribution records for poorly sampled taxa and regions (even more so in the tropics), and we recommend incentives be developed for private collectors to publish their data. Finally, encouragements to publish faunistic data in databases (instead of on paper only), and web-crawling applications that can search for 'informal' data on the internet (e.g., community-run picture-sharing sites, specimen sales sites) may be interesting additions to GBIF and/or related data providers.

Natural history collections and faunistic literature are an enormous storehouse of raw biodiversity and distributional data. The internet provides the technological opportunity to make these data available for broad-scale biodiversity research (Soberon & Peterson 2004). GBIF is currently the leading platform for publishing such information (see Jetz *et al.* 2012, Beck *et al.* 2012 for other initiatives) but there is still room for improvement. Apart from data quality and its documentation (which was not the topic of this study), geographic representation rather than sheer quantity of data should be a focus for further data input.

3.5. Acknowledgments

Discussions with R. Ernst, U. Fritz and D. Nogues-Bravo encouraged us towards this study. Data are part of a larger compilation, and we thank countless people who provided data and/or made accessible collections under their custody (see Electronic Supplements). Our work was financially supported by the EU Synthesys program and by the Swiss National Science Foundation (SNF, grant 31003A_119879).

3.6. References

Ballesteros, L., Kitching, I. J., Jetz, W., Nagel, P. & Beck, J. (2013) Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. *Glob. Ecol. & Biogeography* 22: 586-595.

Beck, J., Ballesteros-Mejia, L., Buchmann ,C.M., Dengler, J., Fritz, S., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft H., Schneider, A-K., Winter, M. & Dormann, C.F. (2012) What's on the horizon of macroecology? *Ecography* 35, 1-11.

Beck ,J., Kitching, I.J. & Linsenmair, K.E. (2006) Measuring range sizes of South-East Asian hawkmoths (Lepidoptera: Sphingidae): effects of scale, resolution and phylogeny. *Global Ecology and Biogeography* 15, 339–348.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W, Fortin ,M.-J., Randin, C., Zimmermann, N.E., Graham, C.H. & Guisan, A. (2011) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography* 21, 481-497.

Brown, J.H., Stevens, G.C. & Kaufman, D.M. (1996) The geographic range: Size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics* 27, 597–623.

Chessel, D., Dufour, A.B. & Thioulouse, J. (2004) The ade4 package –I: One-table methods. *R News*. 4, 5-10.

Clark, B.R., Godfray, H.C.J., Kitching, I.J., Mayo, S.J. & Scoble, M.J. (2009) Taxonomy as an e-Science. *Philosophical Transactions of the Royal Society A* 367, 953-966.

Danner, F., Eitschberger, U., & Surholt, B. (1998) Die Schwärmer der westlichen Palaearktis. Bausteine zu einer Revision (Lepidoptera: Sphingidae). *Herbipoliana* 4, 1-368, 1-720.

Elith, J. & Leathwick, J. (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics* 40, 677–697.

Godfray, H.J.C., Clark, B.R., Kitching, I.J., Mayo, S.J. & Scoble, M.J. (2007) The web and the structure of taxonomy. *Systematic Biology* 56, 943-955.

Guralnick, R. & Van Cleeve, J. (2005) Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. *Diversity and Distributions.* 11, 349–359.

Graham, C., Ferrier, S., Huettman, F. & Moritz, C. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* 19, 497–503.

Hundsdörfer, A.K., Mende, M.B., Kitching, I.J., & Cordellier, M. (2011) Taxonomy, phylogeography and climate relations of the Western Palaearctic spurge hawkmoth (Lepidoptera, Sphingidae, Macroglossinae). *Zoologica Scripta* 40, 403–417.

Hundsdörfer, A.K., Rubinoff, D., Attié, M., Wink, M., & Kitching, I.J. (2009) A revised molecular phylogeny of the globally distributed hawkmoth genus *Hyles* (Lepidoptera: Sphingidae), based on mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution* 52, 852–865.

Jetz, W. & Fine, P. V. A. (2012) Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. *Plos biology* 10(3), e1001292.

Jetz, W., McPherson, J.M., & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution* 27, 1–9.

Legendre, P., Dale, M. R. T., Fortin, M.-J., Gurevitch, J., Hohn, M. & Myers, D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* 25, 601–615.

Lomolino, M.V. (2004) Conservation biogeography. *Frontiers of Biogeography: new directions in the geography of nature* (eds. Lomolino MV & Heaney LR), pp 293–296. Sinauer Associates, Sunderland, Massachusetts.

Newbold, T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* 34, 3–22.

Settele, J., Kudrna, O., Harpke, A., Kuehn, I., van Swaay, C., Verovnik, R., Warren, M., Wiemers, M., Hanspach, J., Hickler, T., Kühn, E., van Halder, I., Veling, K., Vliegenthart, A.,

Wynhoff, I. & Schweiger, O. (2008) *Climatic Risk Atlas of European Butterflies. BIORISK – Biodiversity and Ecosystem Risk Assessment*. Pensoft, Sofia, 710 pp.

Schwanghart, W., Beck, J. & Kuhn, K. (2008) Measuring population densities in a heterogeneous world. *Glob. Ecol. Biogeogr.* 17, 566-568.

Scoble, M.J., Clark, B., Godfray, H.C.J., Kitching, I.J. & Mayo, S. (2007) Revisionary taxonomy in a changing e-landscape. *Tijdschrift voor Entomologie* 150, 305-317.

Soberon, J., Arriaga, L. &Lara, L. (2002) Issues of quality control in large, mixed-origin entomological databases. *Towards a global biological information infrastructure* (eds. Saarenmaa H, Nielsen ES), pp. 1–72. European Environment Agency, Copenhagen.

Soberón, J. & Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society, London (B)* 359, 689–698.

Wiens, J.A. (1989). Spatial scaling in ecology. Functional Ecology. 3, 385-397.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. & Culham, A. (2007): How global is the Global Biodiversity Information Facility? *PloS One* 2, e1124.

Table 3.1. European sphingid species and properties of their known geographic ranges. Only records with an estimated precision <1° latitude/longitude were considered for both Global Biodiversity Information Facility data (GBIF) and those compiled from multiple sources (Coll.). European records for some species are at the extreme edge of their distribution (E), and some are mainly found in Europe in non-permanent summer populations (S). X x Y is the product of longitudinal and latitudinal European range extent [in 10^6 km^2]. "Cells" are 100 x 100 km cells, percentages (in italics) refer to the total known from *independent compilation* and GBIF. Note that areas of minimum convex polygons in climatic niche space are dimensionless data based on PCA axes.

SPECIES	Records Coll.	Records GBIF	% Cells only known from Coll.	% Cells only known from GBIF	Cells Total	Cells only known from Coll.	Cells only known from GBIF	%X*Y Coll.	%X*Y GBIF	X*Y Coll.	X*Y GBIF	Total niche polygon area	Exclusive niche polygon area Coll.	Exclusive niche polygon area GBIF	% niche polygon Coll.	% niche polygon GBIF
Acherontia atropos S	516	570	*51.2*	*35.9*	170	87	61	*80.0*	*75.0*	78.4	73.5	58.2	30.5	48.2	*52.4*	*82.9*
Agrius convolvuli S	391	468	*48.5*	*37.1*	194	94	72	*97.1*	*85.2*	91.8	80.5	77.8	61.3	50.2	*78.8*	*64.5*
Daphnis nerii S	31	66	*51.2*	*43.9*	41	21	18	*80.0*	*35.4*	50.0	22.1	28.2	17.8	8.2	*63.6*	*29.4*
Deilephila elpenor	128	2240	*19.0*	*71.7*	184	35	132	*82.8*	*82.8*	62.4	62.4	44.8	22.5	41.5	*50.3*	*92.6*
Deilephila porcellus	168	2054	*24.4*	*62.8*	180	44	113	*84.7*	*84.0*	57.2	56.7	36.5	24.1	27.0	*66.0*	*74.0*
Dolbina elegans E	3	0	*100.0*	*0.0*	1	1	0	*100.0*	*0.0*	0.1	0.0	0.00	0.00	0	*100*	*0*
Hemaris croatica	20	0	*100.0*	*0.0*	7	7	0	*100.0*	*0.0*	4.8	0.0	5.2	5.2	0	*100*	*0*
Hemaris fuciformis	62	1763	*19.7*	*75.8*	132	26	100	*77.8*	*66.3*	56.7	48.3	35.6	20.0	24.5	*56.2*	*68.7*
Hemaris tityus	81	1752	*20.8*	*67.4*	144	30	97	*78.6*	*64.3*	57.2	46.8	47.0	28.2	35.1	*59.9*	*74.7*
Hippotion celerio S	82	35	*72.6*	*19.4*	62	45	12	*80.8*	*81.0*	44.1	44.2	23.3	18.7	6.4	*78.0*	*27.5*
Hippotion osiris E	1	2	*50.0*	*50.0*	2	1	1	*100.0*	*100.0*	0.1	0.1	NA	NA	NA	*NA*	*NA*
Hyles dahlii	21	2	*85.7*	*14.3*	7	6	1	*100.0*	*2.9*	3.5	0.1	2.4	2.0	0	*83.5*	*0*
Hyles euphorbiae	163	687	*63.4*	*24.1*	112	71	27	*73.9*	*90.3*	50.4	61.6	31.5	28.1	15.7	*89.3*	*49.7*
Hyles gallii	96	739	*24.7*	*67.9*	162	40	110	*73.7*	*60.4*	70.2	57.5	46.3	17.5	39.3	*37.9*	*85.0*
Hyles hippophaes	44	5	*85.7*	*9.5*	21	18	2	*100.0*	*4.7*	44.8	2.1	23.1	23.1	0.8	*100*	*3.3*
Hyles livornica	222	165	*71.8*	*12.9*	124	89	16	*100.0*	*59.5*	75.6	45.0	37.9	36.0	17.0	*95.0*	*44.9*
Hyles nicaea	21	0	*100.0*	*0.0*	11	11	0	*100.0*	*0.0*	3.6	0.0	5.1	5.1	0	*100*	*0*
Hyles tithymali	8	0	*100.0*	*0.0*	4	4	0	*100.0*	*0.0*	3.2	0.0	0.3	0.3	0	*100*	*0*
Hyles vespertilio	42	85	*58.8*	*29.4*	34	20	10	*100.0*	*6.0*	23.4	1.4	15.1	13.2	6.7	*87.2*	*44.3*
Laothoe amurensis	15	132	*31.8*	*63.6*	22	7	14	*5.4*	*100.0*	3.6	66.7	0.9	0.3	0.6	*33.2*	*65.7*
Laothoe populi	200	3053	*26.0*	*60.3*	219	57	132	*93.3*	*80.9*	84.0	72.8	61.5	41.9	41.4	*68.0*	*67.3*
Macroglossum stellatarum S	354	1900	*33.6*	*44.4*	232	78	103	*100.0*	*82.6*	108.5	89.6	61.5	41.8	49.0	*67.9*	*79.7*

Marumba quercus	85	61	*58.7*	*34.8*	46	27	16	*60.0*	*51.6*	27.9	24.0	18.2	11.2	10.4	*61.7*	*56.9*
Mimas tiliae	146	1631	*31.2*	*61.0*	141	44	86	*79.9*	*95.7*	44.1	52.8	37.3	25.2	31.4	*67.5*	*84.1*
Proserpinus proserpina	146	182	*61.3*	*26.3*	80	49	21	*100.0*	*74.0*	62.0	45.9	20.5	13.6	12.4	*66.6*	*60.6*
Rethera komarovi	11	0	*100.0*	*0.0*	3	3	0	*100.0*	*0.0*	0.2	0.0	0.3	0.3	0	*100*	*0*
Smerinthus ocellata	115	1747	*28.6*	*66.9*	154	44	103	*88.9*	*76.1*	67.2	57.5	38.3	34.6	17.9	*90.3*	*46.6*
Sphingonaepiopsis gorgoniades	16	0	*100.0*	*0.0*	8	8	0	*100.0*	*0.0*	9.9	0.0	6.1	6.1	0	*100*	*0*
Sphinx ligustri	79	1828	*19.1*	*73.0*	141	27	103	*69.5*	*85.7*	50.6	62.4	41.1	22.5	29.1	*54.8*	*70.8*
Sphinx maurorum	81	2	*96.0*	*0.0*	25	24	0	*100.0*	*0.0*	9.0	0.0	13.2	13.2	0	*100*	*0*
Sphinx pinastri	181	2817	*28.8*	*50.3*	177	51	89	*88.8*	*96.4*	57.2	62.1	51.6	30.6	50.2	*59.2*	*97.2*
Theretra alecto E	8	0	*100.0*	*0.0*	7	7	0	*100.0*	*0.0*	2.0	0.0	2.7	2.7	0	*100*	*0*

Appendix 3.1 Sources of GIBF data (search on Sept 16th, 2009) for European sphingid moths. Spellings were edited an interpreted as original download from GBIF contained many font errors (due to use of characters not contained in the English alphabet). Ca. 52% of records are based on observations, 37% on specimen records, and 11 % from unknown sources.

Data provider (% of records)	**Datasets** (sorted by contribution, largest to smallest)
UK National Biodiversity Network (31.2)	Dorset Environmental Records Centre - Dorset Hawkmoths - NBN South West Pilot Project Case Studies Joint Nature Conservation Committee - Scarce Macro Moth Review Data (historical) Highland Biological Recording Group - HBRG Lepidoptera dataset Natural England - Invertebrate Site Register - England. Take a Pride in Fife Environmental Information Centre - Records for Fife from TAPIF EIC Scottish Natural Heritage - Invertebrate Site Register, Scotland East Ayrshire Countryside Ranger Service - East Ayrshire Species Database Environment and Heritage Service - EHS Species Datasets Lothian Wildlife Information Centre - Lothian Wildlife Information Centre Secret Garden Survey
Biologiezentrum der Oberoesterreichischen Landesmuseen (27.0)	Biologiezentrum Linz
GBIF-Sweden (17.2)	Bugs (GBIF-SE:Artdatabanken) Lepidoptera (Observations) Lepidoptera (Specimens NRM) Lund Museum of Zoology - Insect collections (MZLU)
University of Helsinki, Department of Applied Biology (5.2)	European Moth Nights Lepidoptera collection of Hannu Saarenmaa Lepidopterological Society of Finland European Lepidoptera Observations by Donald Hobern
Jyvaskyla University Museum - The Section of Natural Sciences (4.9)	Invertebrate collection of Jyvaskyla University Museum
Natural History Museum, University of Oslo (4.1)	Norwegian Lepidoptera working group Norwegian Lepidoptera collection, Oslo Arthropod collection, Tromsø Museum
inatura - Erlebnis Naturschau Dornbirn (3.7)	inatura - Erlebnis Naturschau Dornbirn
European Environment Agency (3.3)	EUNIS
Banc de dades de biodiversitat de Catalunya (0.9)	Banc de dades de biodiversitat de Catalunya-ArtroCat

NLBIF (0.8)	Natural History Museum Rotterdam (NMR)
GEO-Tag der Artenvielfalt (0.5)	Artenvielfalt auf der Weide - GEO-Hauptveranstaltung in Crawinkel Danielsberg (Mölltal, Kärnten) GEO Hauptveranstaltung Tirol (Innsbruck) GEO-Hauptveranstaltung (Insel Vilm) Pilstingermoos Artenfülle um das Schalkenmehrener Maar 4. Tag der Artenvielfalt, Naturschutzgebiet Hockenheimer Rheinbogen Fels- und Weinbergsflächen in Hatzenport/Terrassenmosel Gelände des IVL (Zeckern) GEO-Hauptveranstaltung (NLP Harz / Hochharz) Gurgltal (Tarrenz) Neckartalsüdhang (Horb) Schulhof Goethe-Gymnasium (Emmendingen) BUND - Dassower See (Lübeck/Dassow) Erlengraben/Lipp-Tal (├ûstringen) Schlern - (Bozen) B?G Bannwald Burghauser Forst Faberpark (Nürnberg/Stein) GNOR-Projekt "Halbwilde Weidehaltung zwischen Kamp-Bornhofen und Kestert" und Umland Halbwilde Weidehaltung zwischen Kamp-Bornhofen und Kestert sowie Umland Laubenheimer Bodenheimer Ried - von Stromtalwiesen und Flutrasen Perchtoldsdorfer Heide Streuobstwiese RSG (Cham) Sudeniederung (Amt Neuhaus) Weinberge und angrenzende Felsflächen (Drieschen) in Hatzenport/Terrassenmosel 3. Tag der Artenvielfalt Hockenheim 5.Tag der Artenvielfalt: Thema Stadtbiotope Aussenfeuerstelle Königsbol (Hartheim/Messstetten) Biologische Station im Kreis Wesel Biosphärenpark Wienerwald - Wiener Steinhofgründe Biosphärenreservat Münsinger Alb Borstgrasrasen um die Burg Baldenau im Oberen Dhrontal Brander Wald (Stolberg) Eppingen und Umgebung

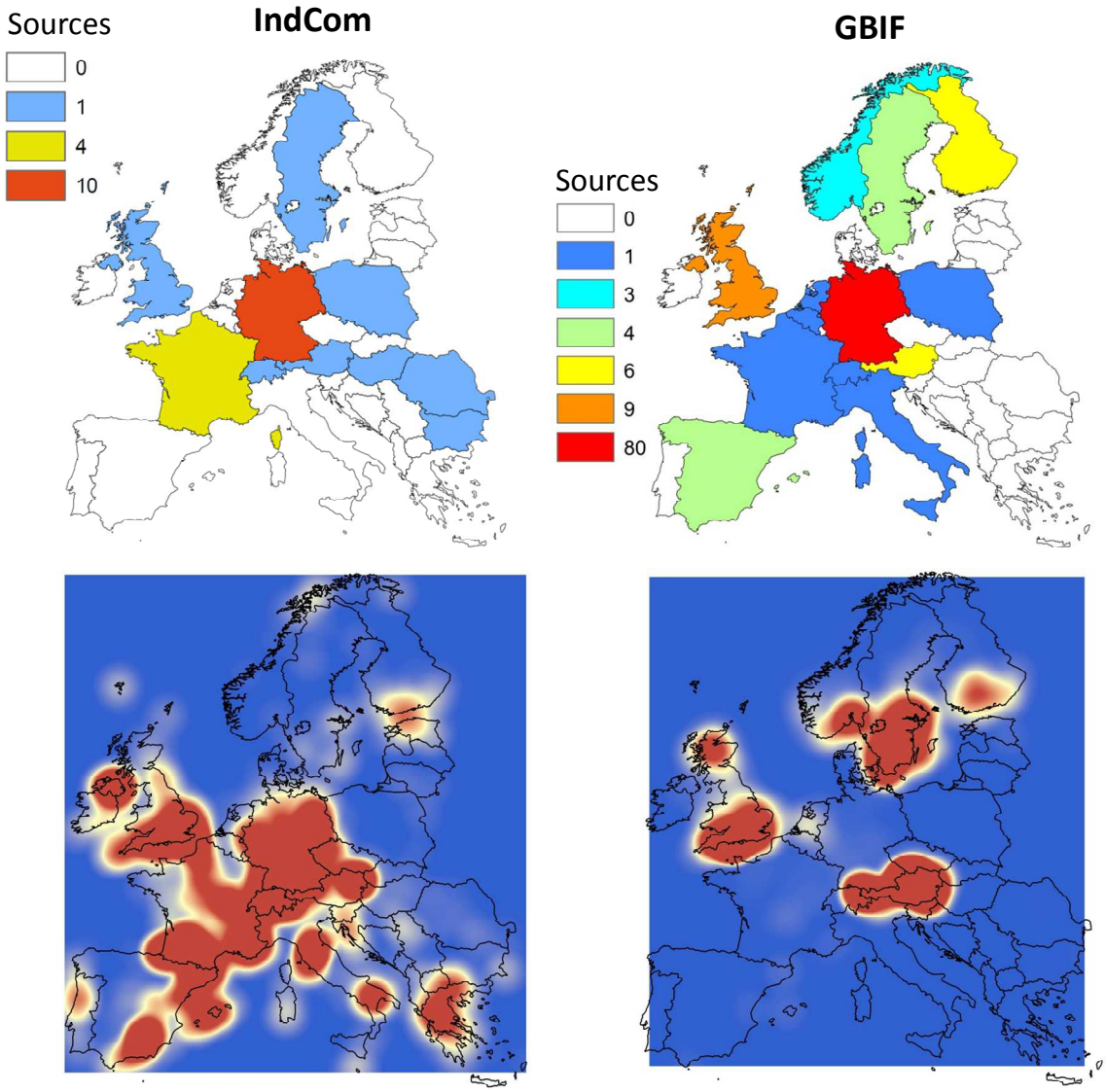
GEO-Tag der Artenvielfalt (0.5) *(continued)*	Feuchtbiotop, Wildtier- und Artenschutzstation Sachsenhagen, Sielmanns Natur-Ranger
	FFH-Gebiet "Calwer Heckengäu"
	Flora und Fauna am Mittelrheintal
	Freiburger Tag der Artenvielfalt
	Freigelände Naturschutzscheune Reinheimer Teich (Kreis Darmstadt-Dieburg)
	Gemeinde Sursee
	Gemeindegebiet Weikendorf (Marchfeld)
	GEO-Hauptveranstaltung (Duisburg)
	GEO-Hauptveranstaltung im Nationalpark Bayerischer Wald
	Geo-Tag der Artenvielfalt Süssen
	Hornwiesen-Grundschule
	Geschützter Landschaftsbestandteil - GLB "Troppach"
	Heinersdorfer Sumpfwiese
	Hintere Halde
	Innenstadt Göttingen - Natur Zuhause
	Kiesbagger (Mittelhausen)
	Knechtweide (Kohlfurth)
	LaBoOb02
	Langes Tannen
	Lillachtal mit Kalktuffquelle bei Weissenohe
	Lustadter Wald .
	NABU Naturschutzhof Netttetal (Sassenfeld) e.V.
	Natur aus zweiter Hand am Muldestausee
	Naturschutzgebiet Bausenberg
	Naturschutzgebiet Sistig-Krekeler-Heide
	Naturschutzstation Schmidsfelden
	NSG Hülenbuch Hörnle (Tieringen/Messstetten)
	NSG Leist bei Ziegenhain
	Pöhlberg bei Annaberg
	renaturierter Main (Kemmern bei Bamberg)
	Riedensee
	Rohrmeistereiplateau und angrenzendes Gebiet
	Rund um das LUGY
	Rund um den Eichwald, Schulhof Friedrich Fröbel Gymnasium- Bad Blankenburg
	Schule Sulzbach (Oberegg)
	Schulhof (Bad Waldsee)
	Schulhof der Astrid-Lindgren-Schule und Umgebung (Elmshorn)
	Spandau HBO
	Streuobstwiese Kugelberg (Ulm)
	Tage der Artenvielfalt rund um die

	Naturschutzstation Molsberg
	Teich Berlin Wuhlheide
	Trockenhang Greinhartsberg Edelfingen
	Umgebung der Gesamtschule Hamburg-Winterhude
	Umgebung des Spalatin Gymnasium Altenburg
	Verwilderter Hausgarten mit angrenzendem Gelände (Laufenburg-Hochsal)
GEO-Tag der Artenvielfalt (0.5) *(continued)*	Von A(horn) bis Z(ecke) des WWP Chemnitz
	Waldränder der Frankenhöhe (Rothenburg ob der Tauber)
	Weinberg Reichersdorf
	Wiese am Waldrand (Gurtweil)
	Zwei Flüsse - eine Stadt (Villingen-Schwenningen)
Finnish Museum of Natural History (0.4)	Hatikka Observation Data Gateway
Service du Patrimoine naturel, Museum national d'Histoire naturelle, Paris (0.4)	Inventaire national du Patrimoine naturel (INPN)
	Laboratorio de Entomologia y Control de Plagas del Instituto Cavanilles de Biodiversidad y Biologia Evolutiva de la
GBIF-Spain (0.1)	Universidad de Valencia: ENV
	BDBCV BioBlitz in Penyagolosa (Castellon, Spain)
	University of Ghent - Zoology Museum -
BeBIF Provider (0.1)	Invertebratacollectie
SysTax (0.1)	Lobbecke Museum Dusseldorf
	SysTax
Staatliches Museum für Naturkunde Stuttgart (0.0)	EDIT - ATBI in Mercantour/Alpi Marittime (France/Italy)
University of Navarra, Museum of Zoology (0.0)	Museum of Zoology, University of Navarra
Institute of Nature Conservation PAS (0.0)	National System of Protected Areas
NatureServe (0.0)	NatureServe Network Species Occurrence Data

Appendix 3.2 Primary sources of *independent compilation* data for European Sphingidae.
These public or private collections were either databased by one of us or data were communicated to us. Collections are listed in declining order of records contribution. The first three collections together made up ca. 70% of records. Published literature (not listed in detail) added another ca. 9 % of total records.

Natural History Museum, London, UK
J. Haxaire collection, Laplume, France
Muséum national d'Histoire naturelle, Paris, France
R. Brechlin collection, Pasewalk, Germany
Hungarian Natural History Museum, Budapest, Hungary
Carnegie Museum of Natural History, Pittsburgh, Pennsylvania, USA
Museum für Naturkunde, Leibnitz-Institut für MNHU Evolutions- und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Germany
J. Beck collection, University of Basel, Switzerland (incl. observations)
U. Eitschberger collection, Marktleuthen, Germany
Institut für Pharmazie und Molekulare Biotechnologie, Heidelberg, Germany
Muséum d'Histoire naturelle de Dijon, Dijon, France
S.V. Beschkow collection, Sofia, Bulgaria
A.K. Hundsdörfer research collection, Dresden, Germany
Museum Thomas Witt, Munich, Germany
Zoölogisch Museum Amsterdam, Amsterdam, The Netherlands
Zoologische Staatssammlung des Bayerischen Staates, München, Germany
J. Bury collection, Poland
Landessammlungen für Naturkunde, Karlsruhe, Germany
Staatliches Museum für Tierkunde in Dresden, Dresden, Germany
Y. Estradel collection
MSc thesis by Hauke Koch (unpublished)
H. Falkner collection, Karlsruhe, Germany
Naturhistorisches Museum, Vienna, Austria
Naturhistoriska Riksmuseet, Stockholm, Sweden
R. Paul collection, Romania

Appendix 3.3 Maps of source and record distributions. Distribution of data sources across European nations (*upper maps*) and relative distributions of record densities (density kernel, 200 km search radius, mapped in 100 x 100 km cells; *lower maps*). Overall densities for GBIF are much higher (see main text), colour stretch follows equal rules per data set (1 SD). Note that *independent compilation* data (IndCom) covers species-rich south-eastern Europe better than GBIF.

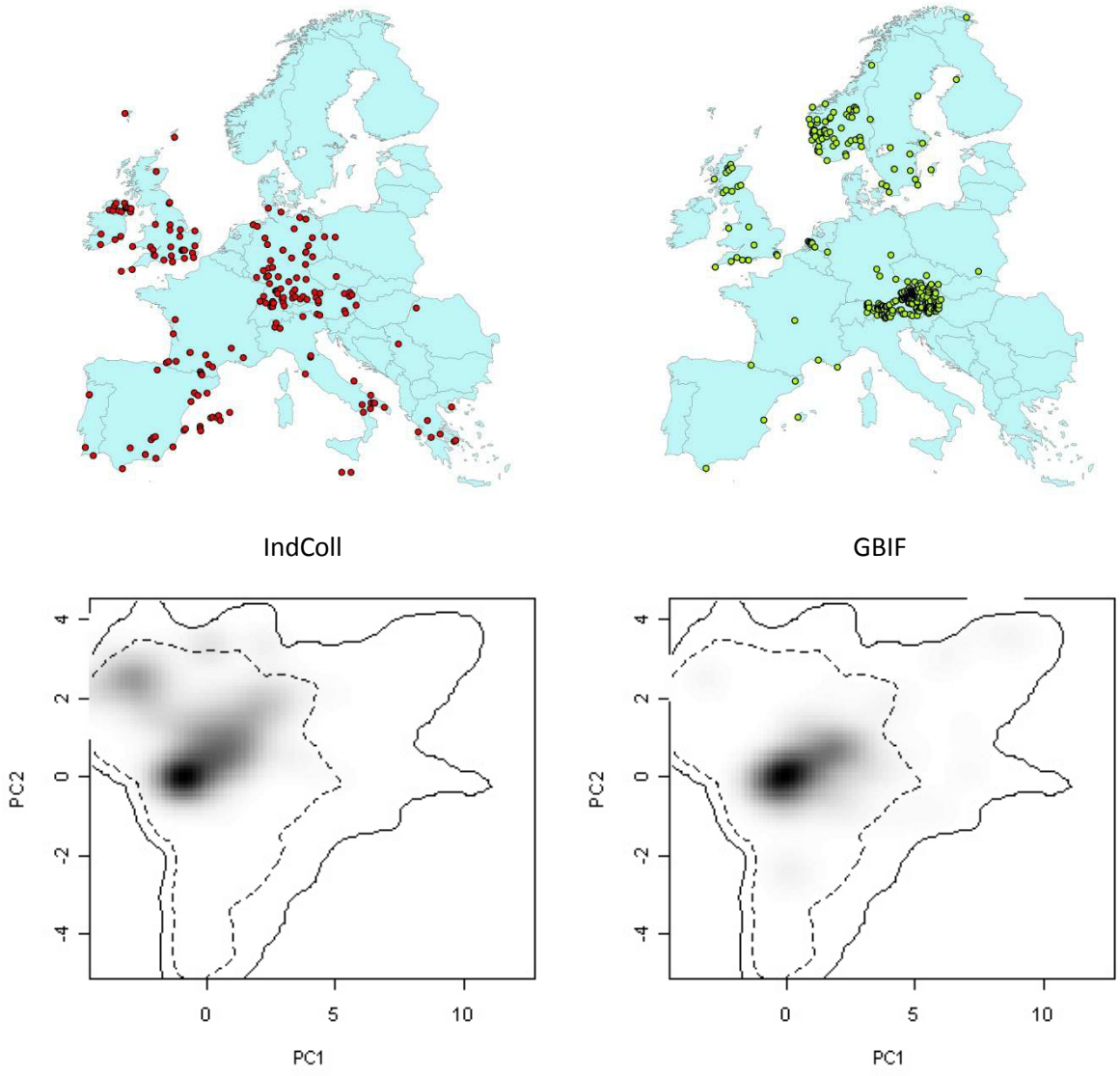


Appendix 3.4 Records in geographic and climatic niche space.

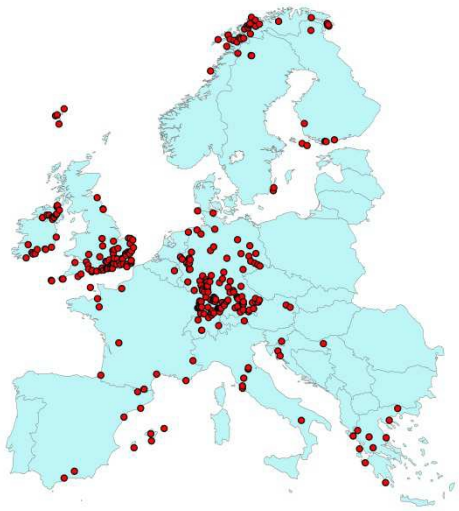
Maps of records for all species in analysis (green = GBIF, *right side*; red = *independent compilation* (IndCom), *left side*). See main text, Table 1, for some details on species and data. Geographical maps are in Mollweide equal area projections.

For species with sufficient numbers of records we also show record density kernels in 2-dimensional climatic niche space underneath each corresponding geographical map. The first two axes from a principle component analysis (PCA) of climate are shown, dark colour indicate high densities of records.

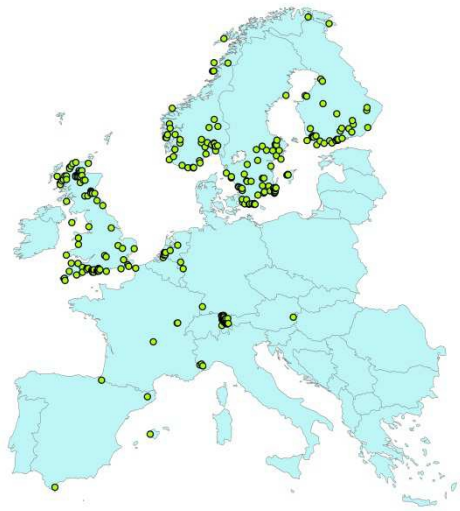
Acherontia atropos



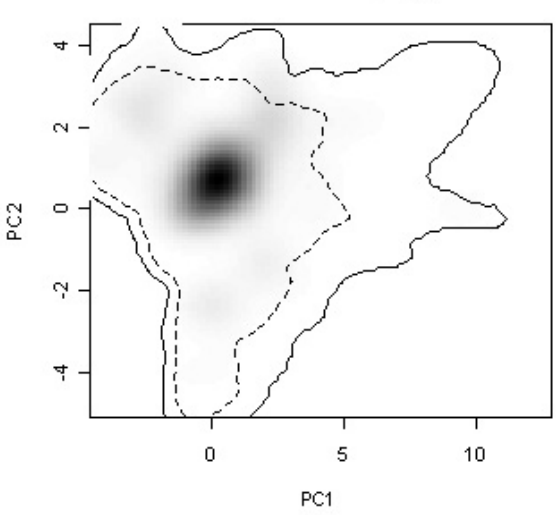
Agrius convolvuli

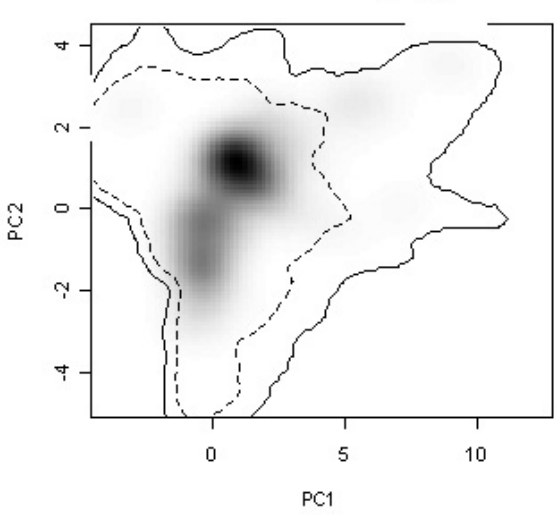


IndColl



GBIF

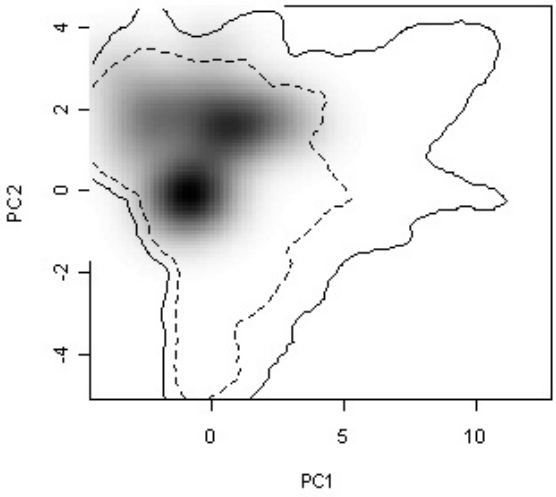


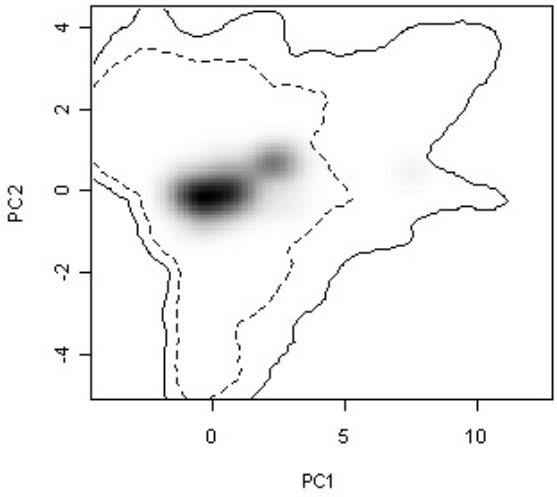


Daphnis nerii

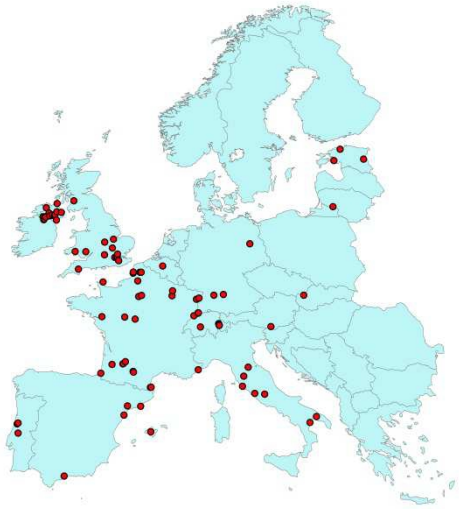
IndColl

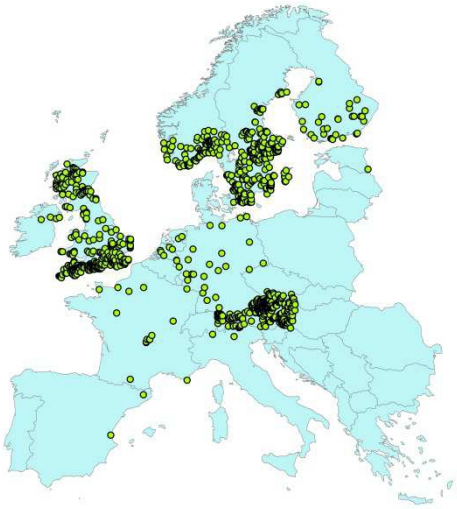
GBIF





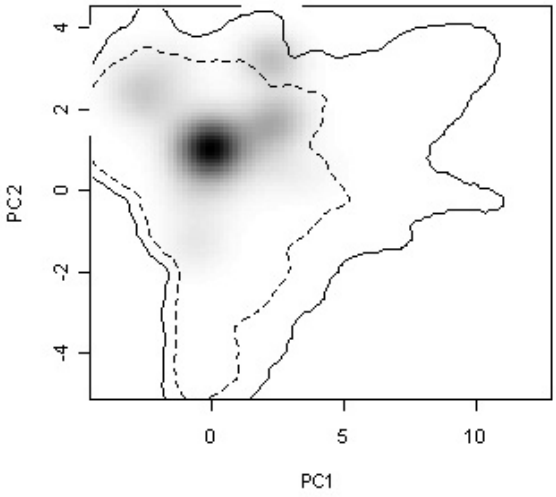
Deilephila elpenor

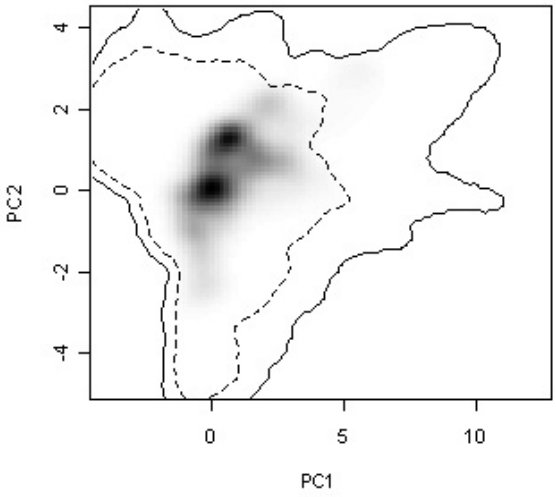




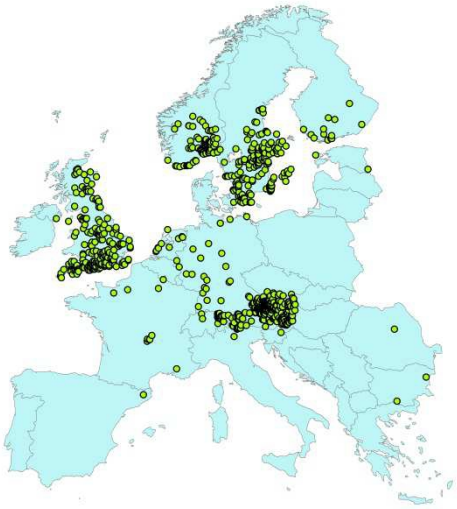
IndColl

GBIF



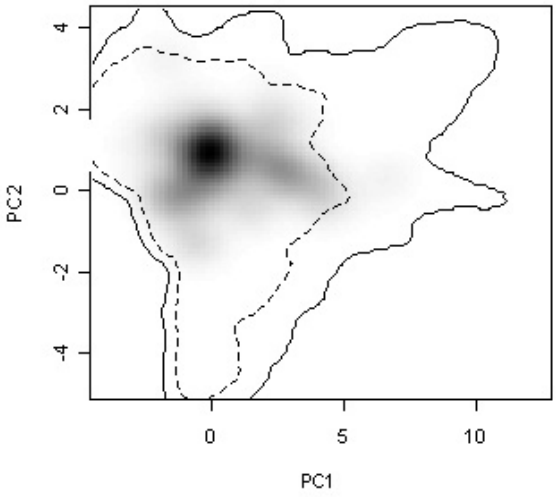


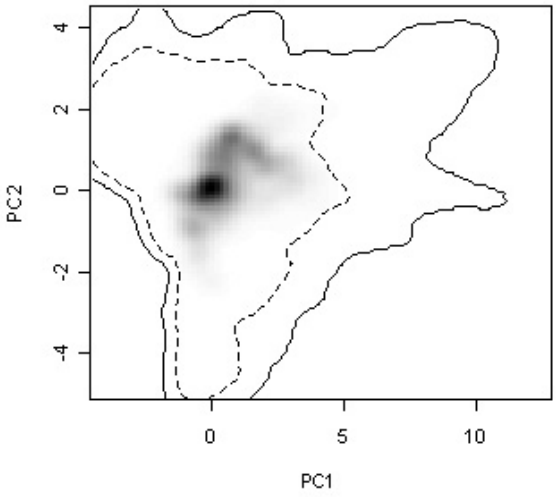
Deilephila porcellus



IndColl

GBIF

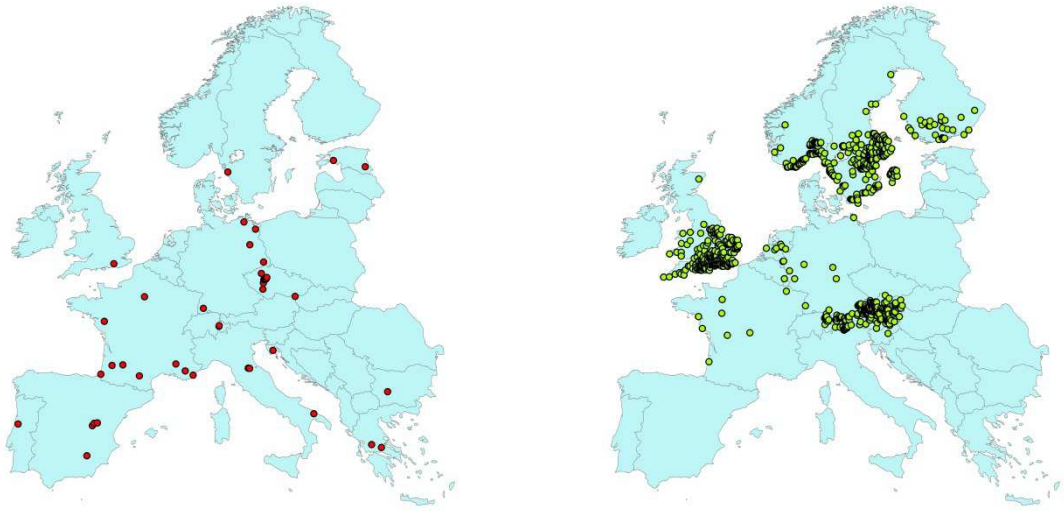




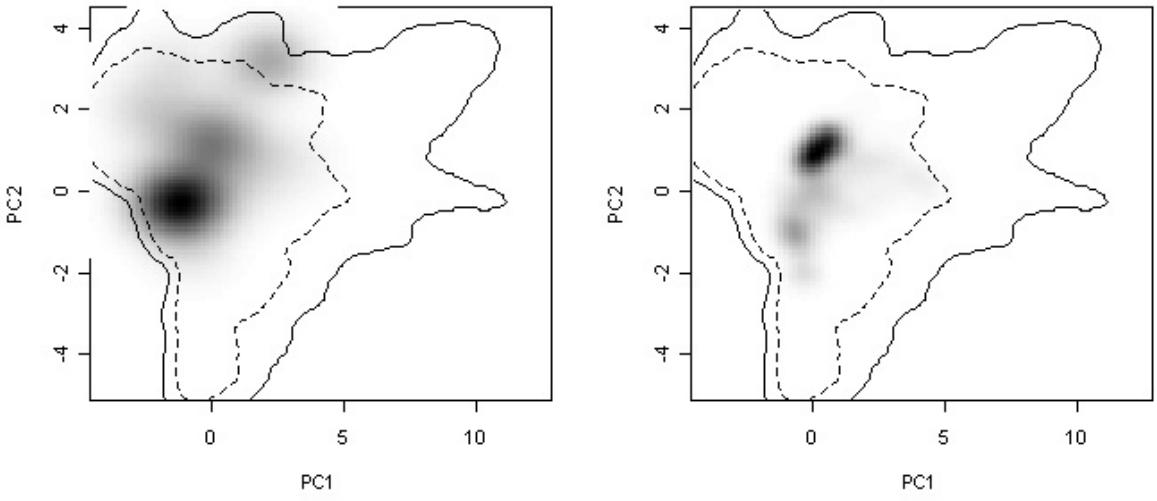
Dolbina elegans

Hemaris croatica

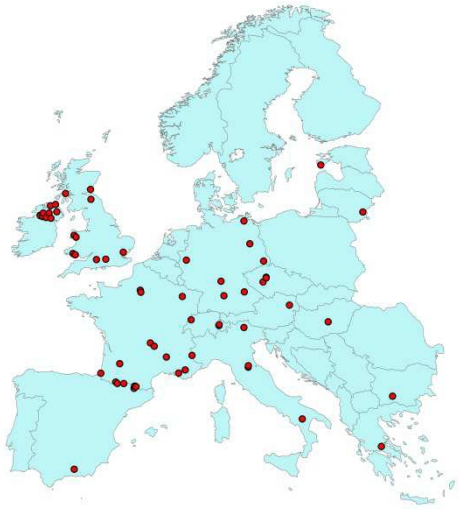
Hemaris fuciformis



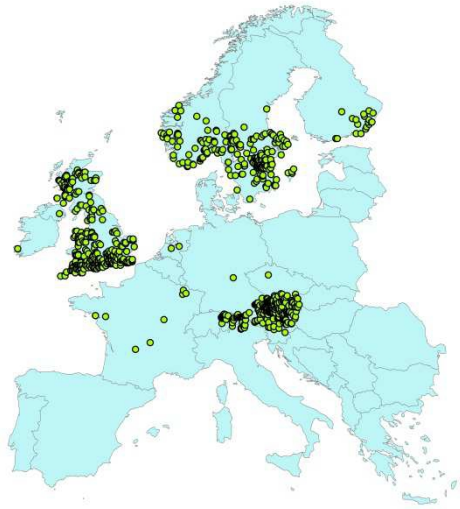
IndColl GBIF



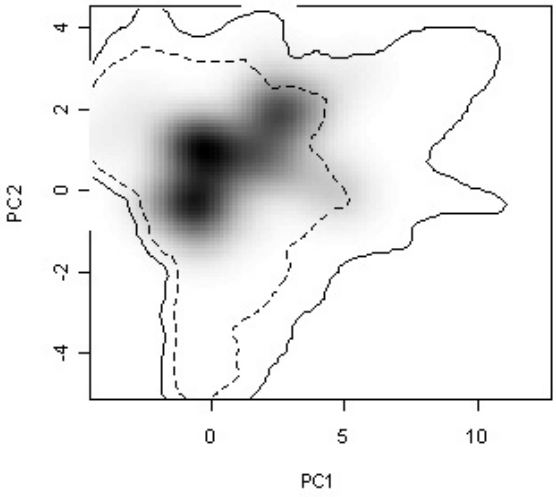
Hemaris tityus

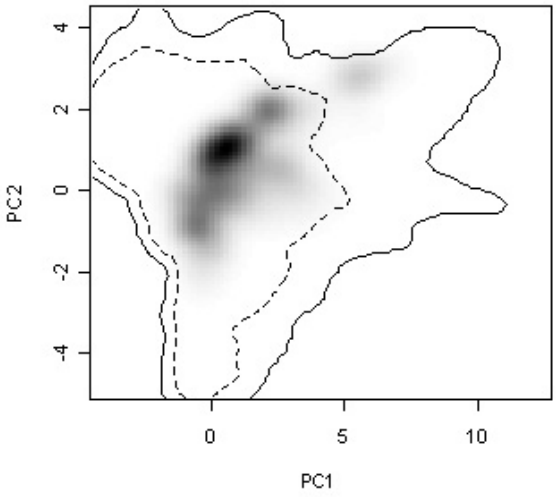


IndColl

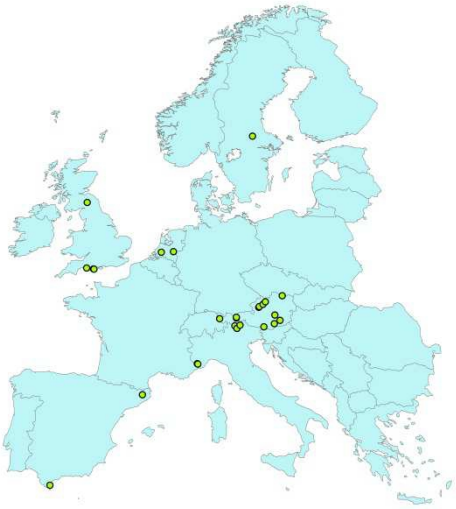


GBIF



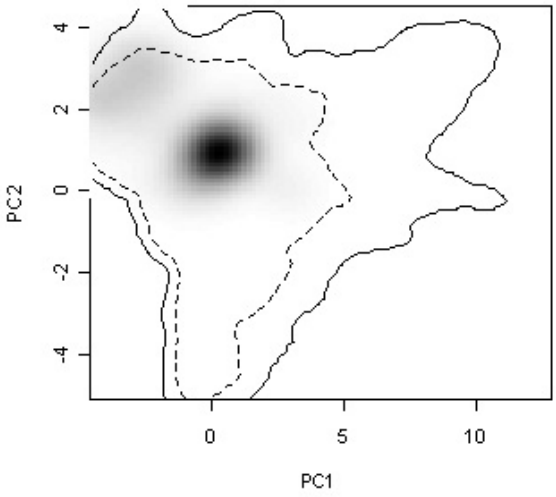


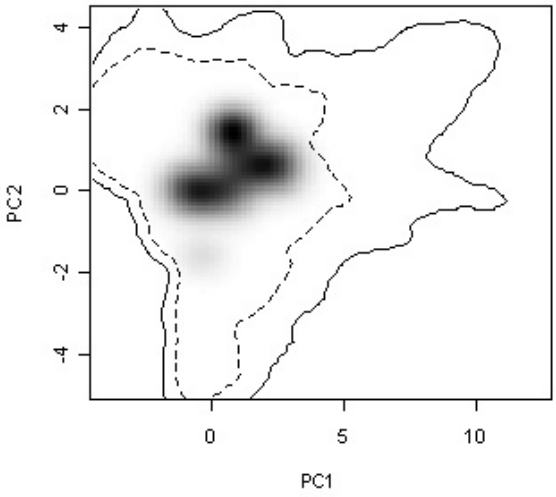
Hippotion celerio



IndColl

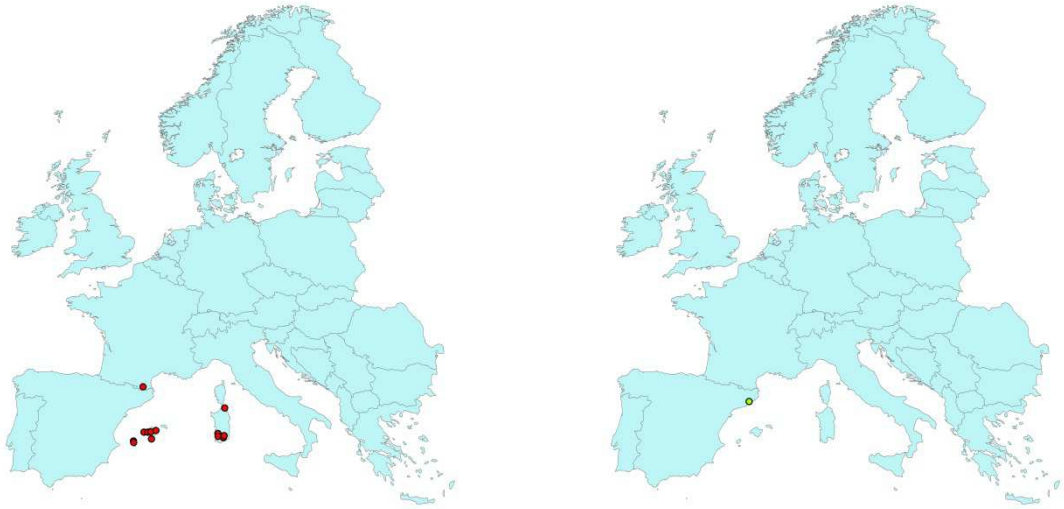
GBIF



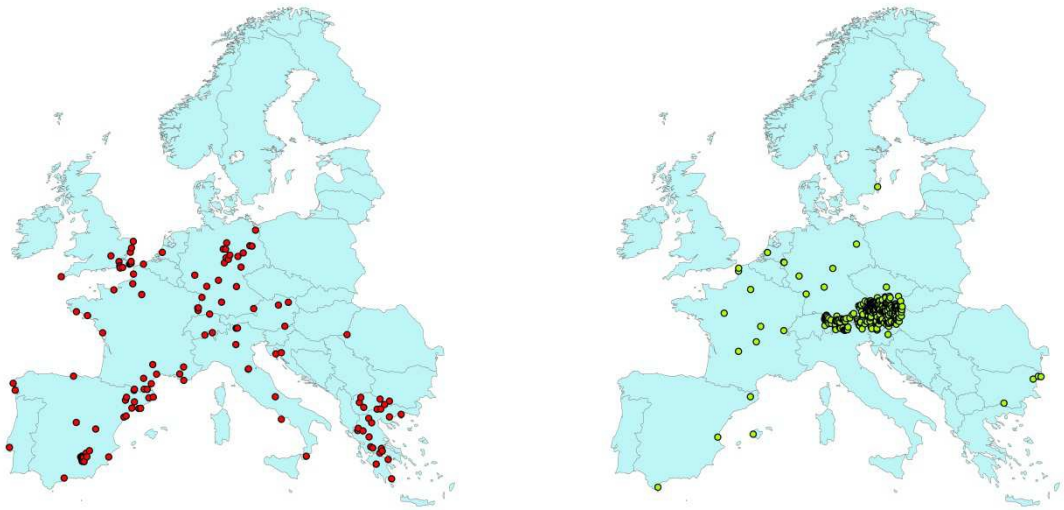


Hippotion osiris

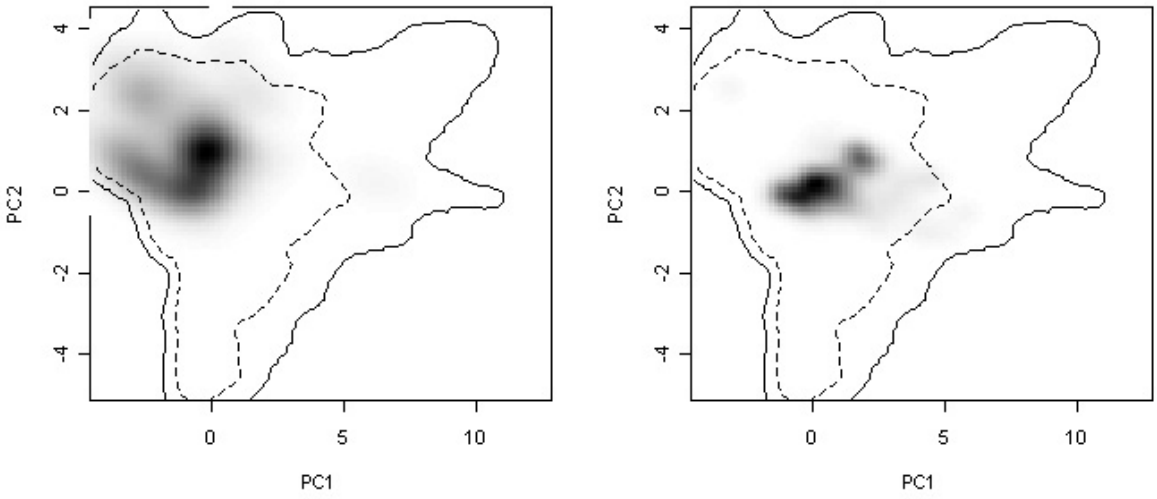
Hyles dahlia



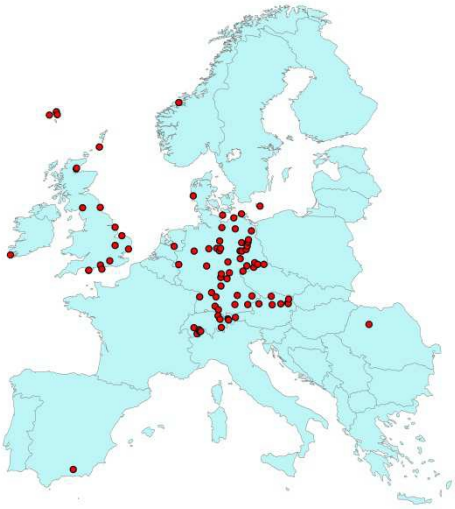
Hyles euphorbiae

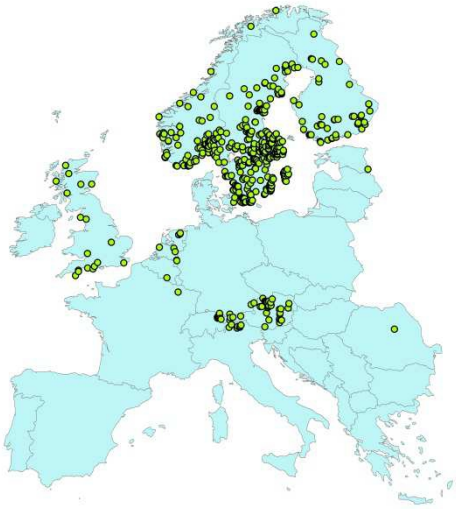


IndColl GBIF



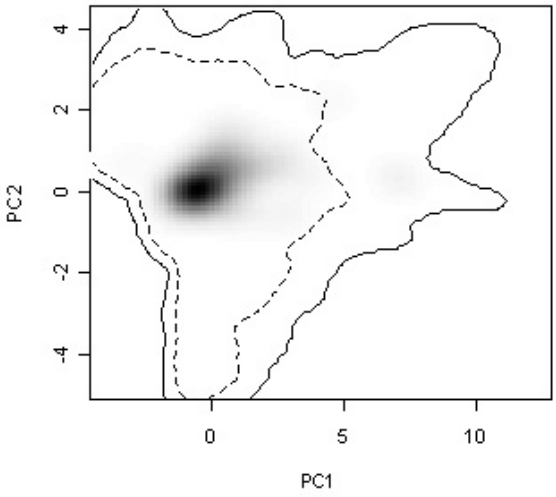
Hyles gallii

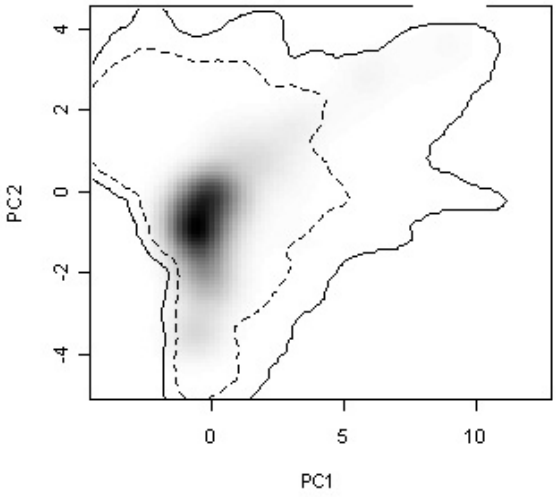




IndColl

GBIF

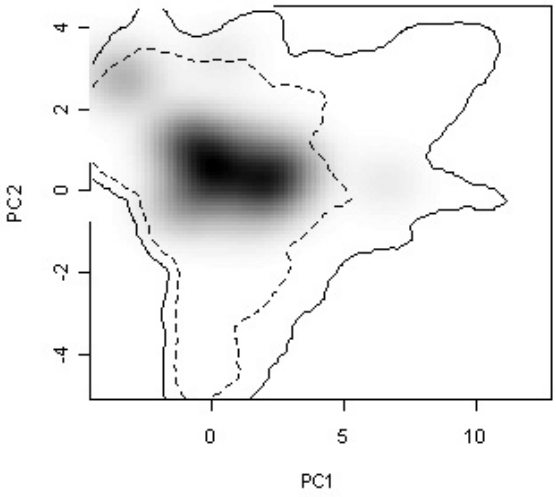


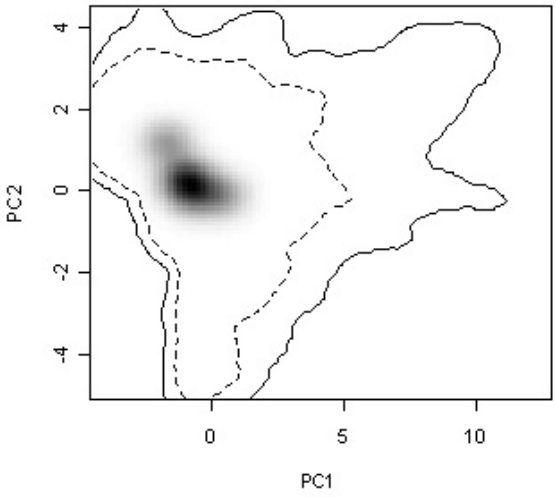


Hyles hippophaes

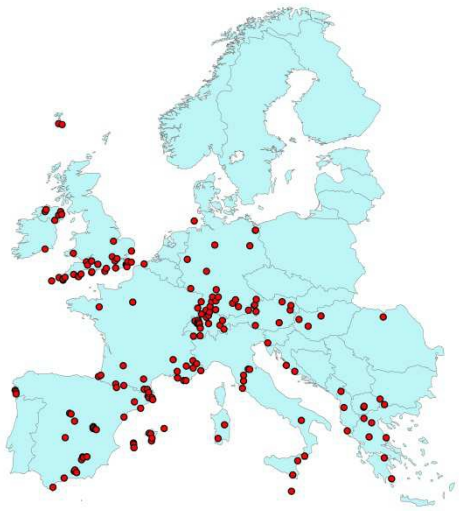
IndColl

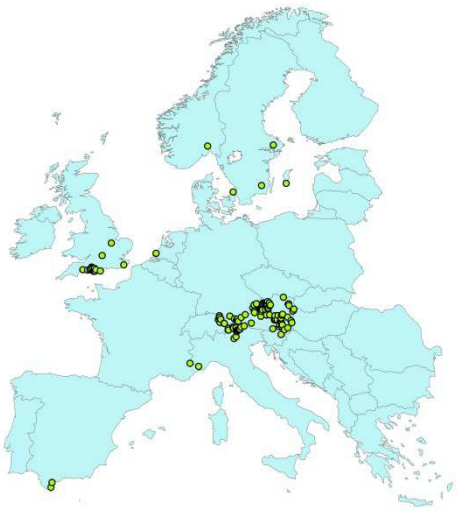
GBIF





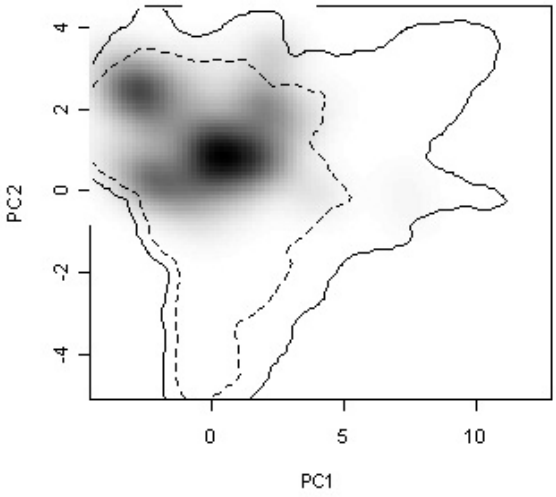
Hyles livornica

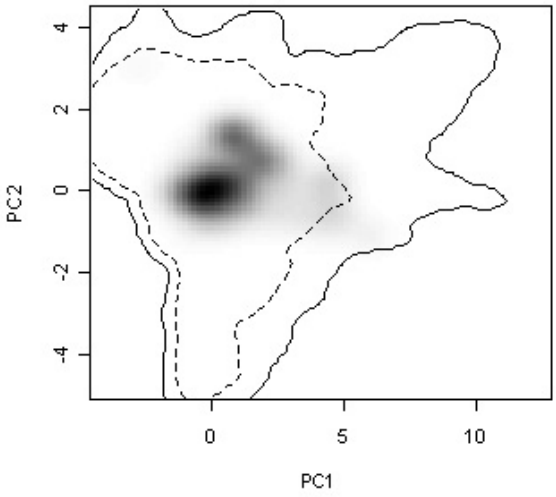




IndColl

GBIF

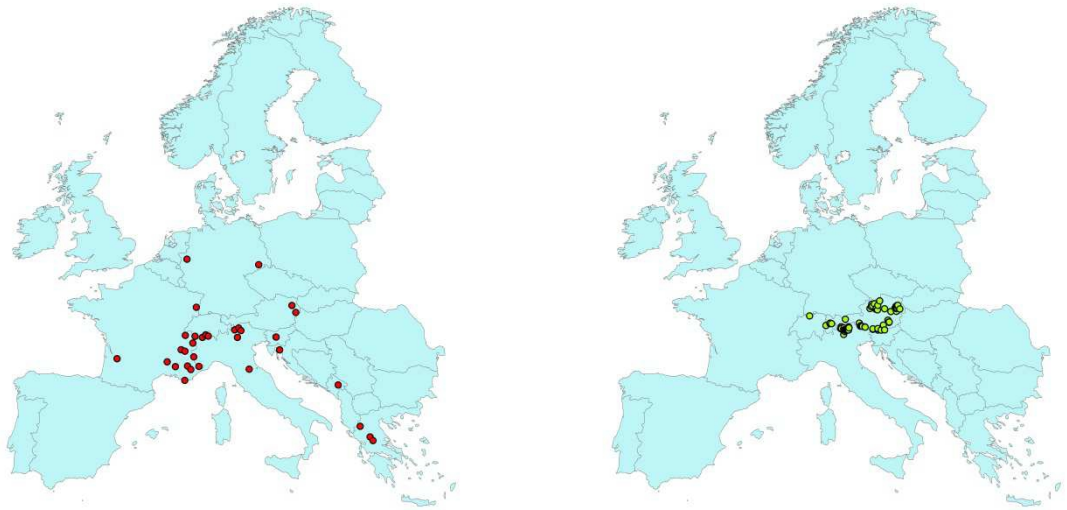




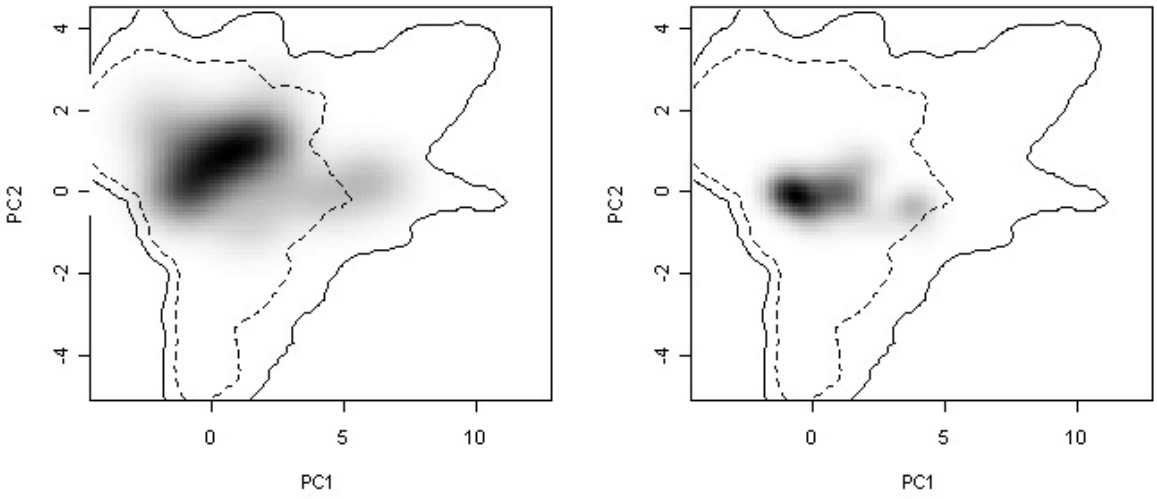
Hyles nicaea

Hyles tithymali

Hyles vespertilio



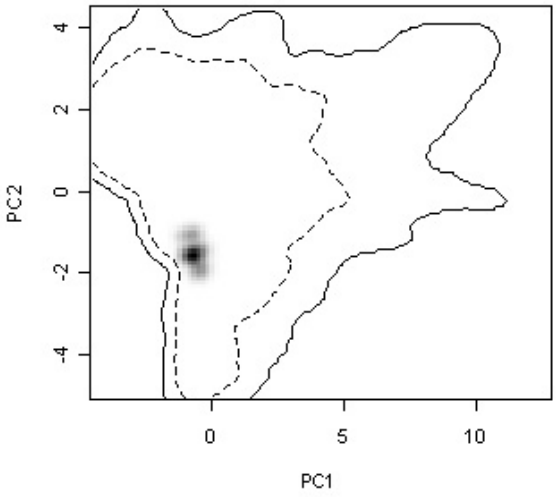
IndColl GBIF

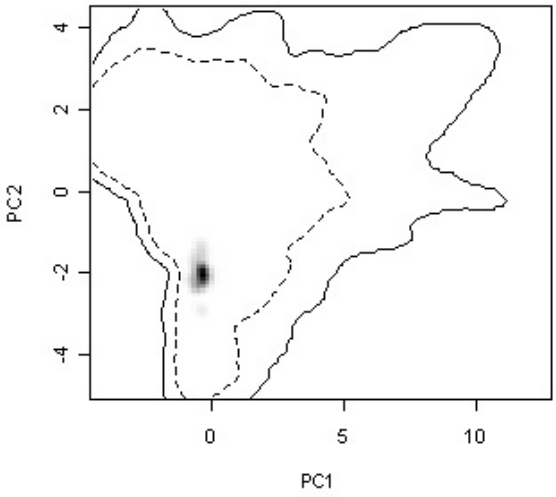


Laothoe amurensis

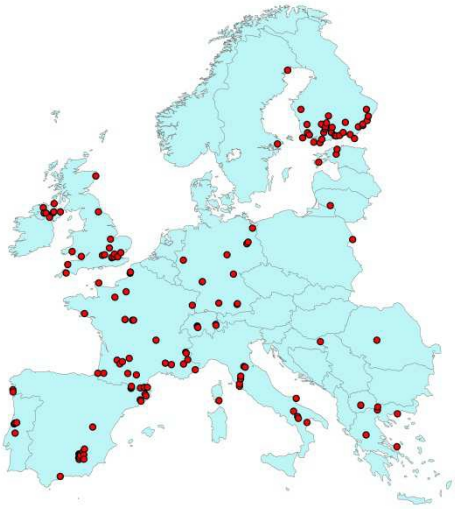
IndColl

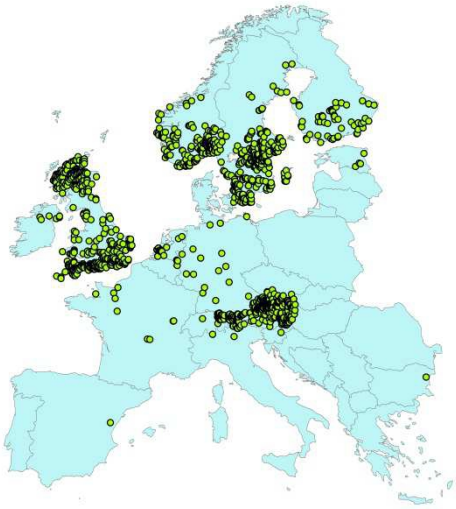
GBIF





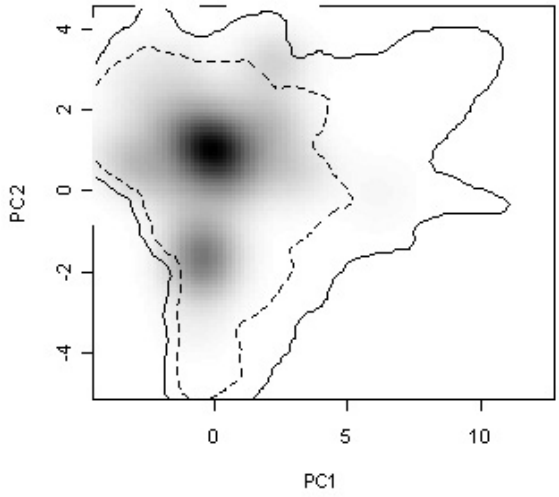
Laothoe populi

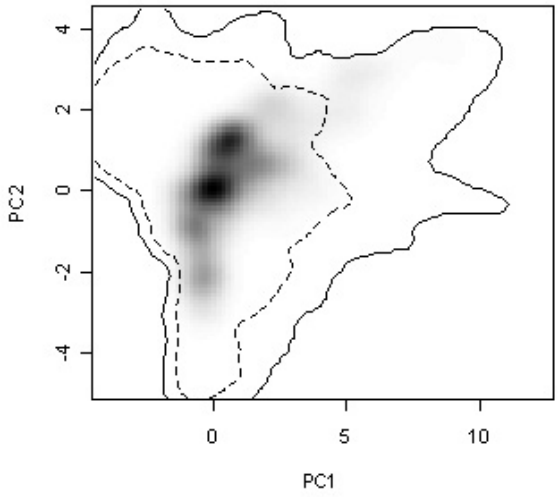




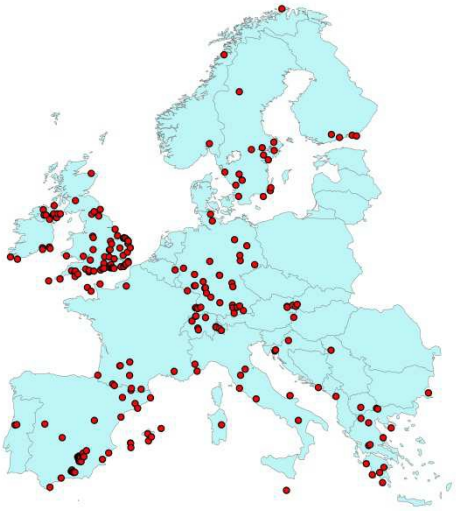
IndColl

GBIF

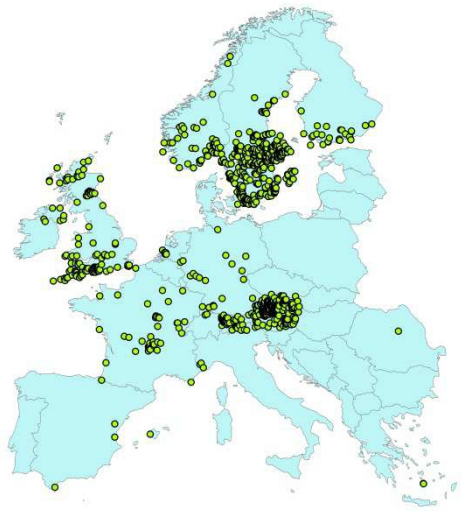




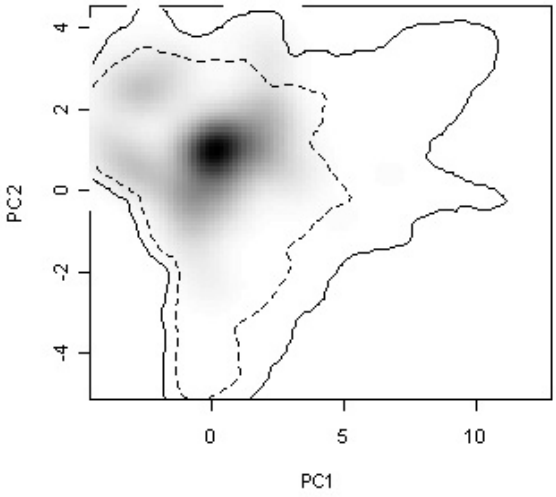
Macroglossum stellatarum

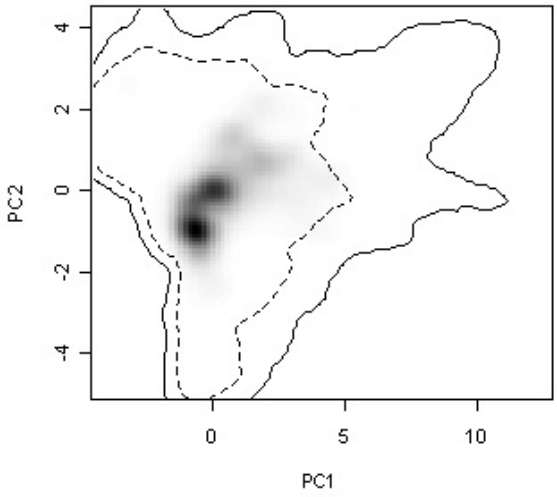


IndColl



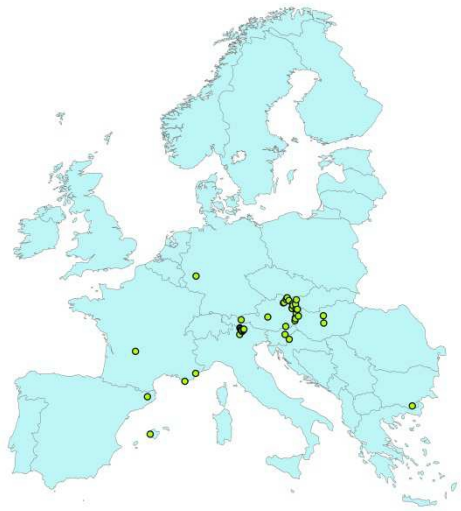
GBIF



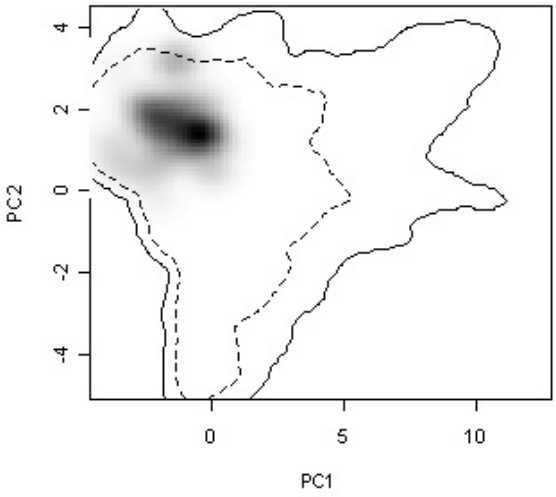


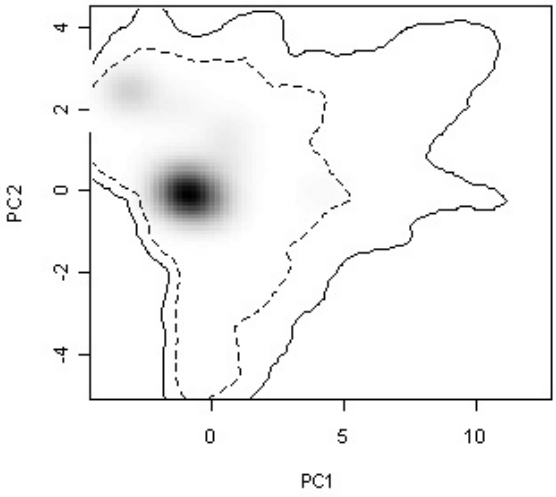
Marumba quercus

IndColl



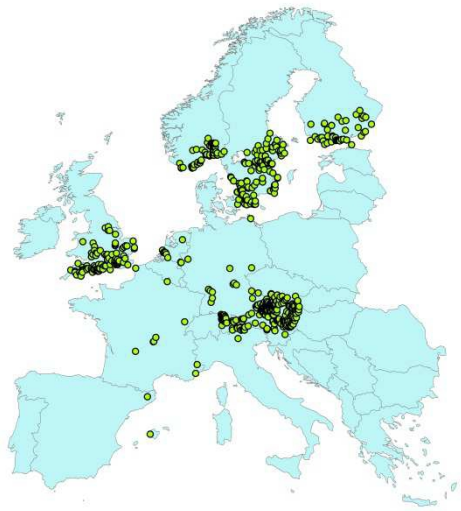
GBIF



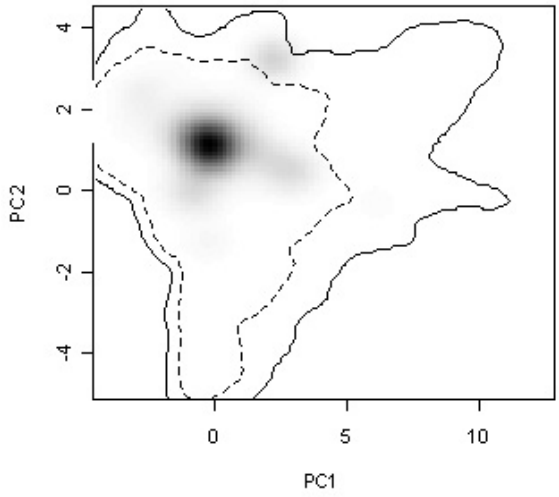


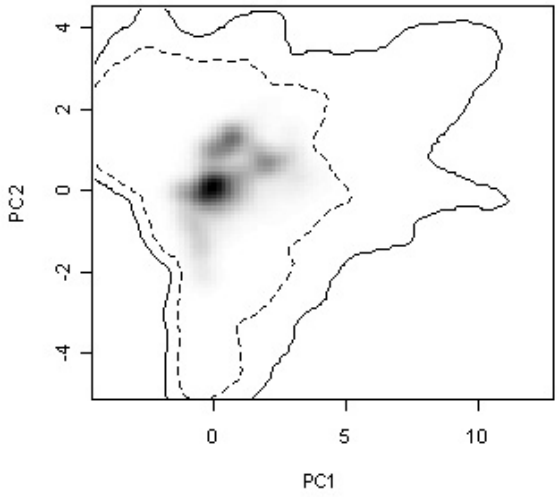
Mimas tiliae

IndColl

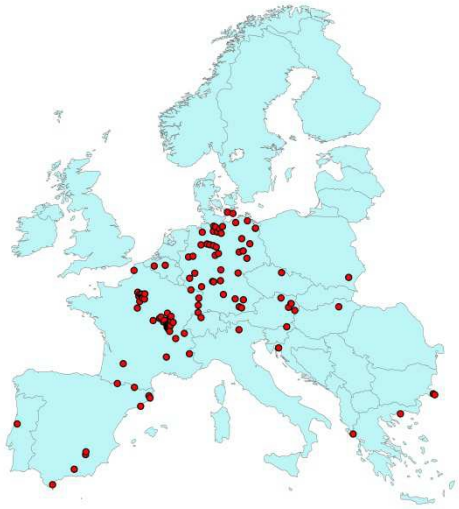


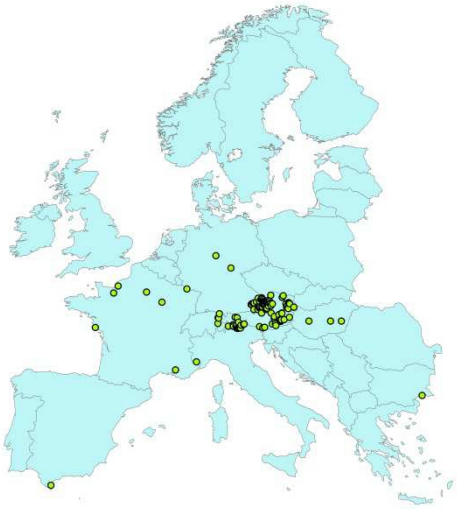
GBIF





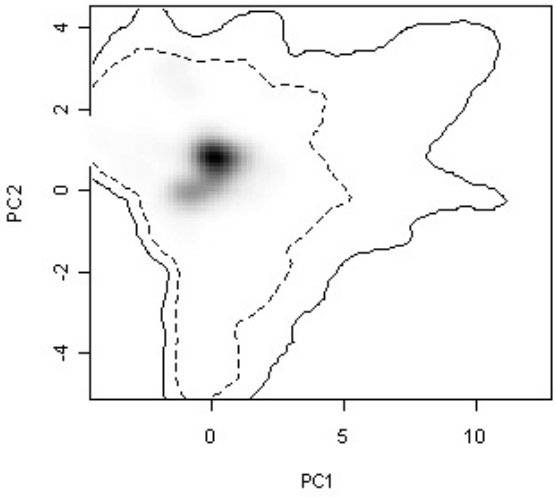
Proserpinus proserpina

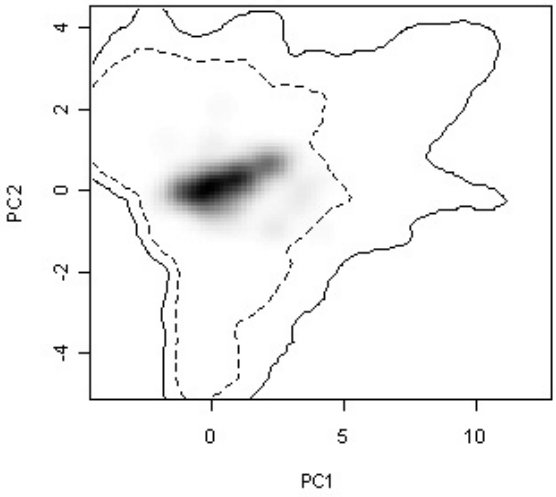




IndColl

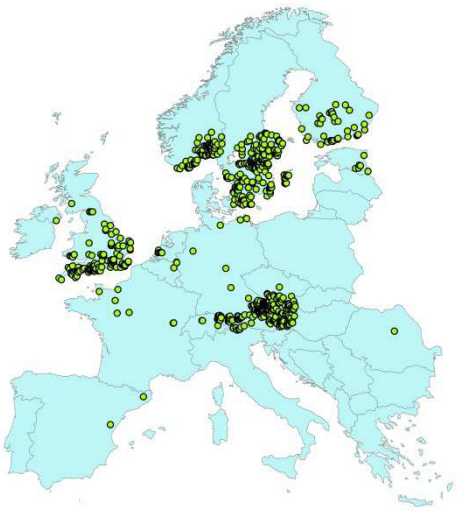
GBIF





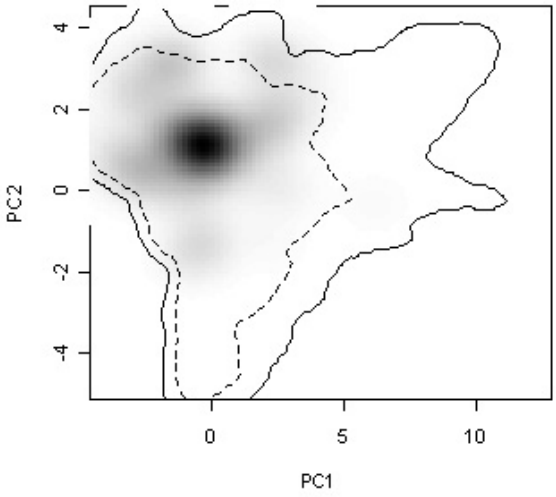
Rethera komarovi

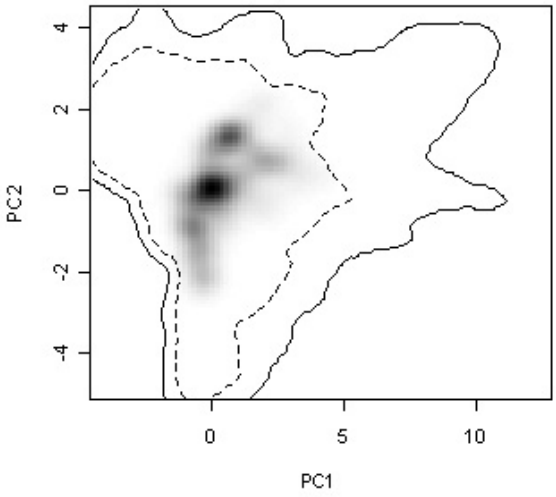
Smerinthus ocellata



IndColl

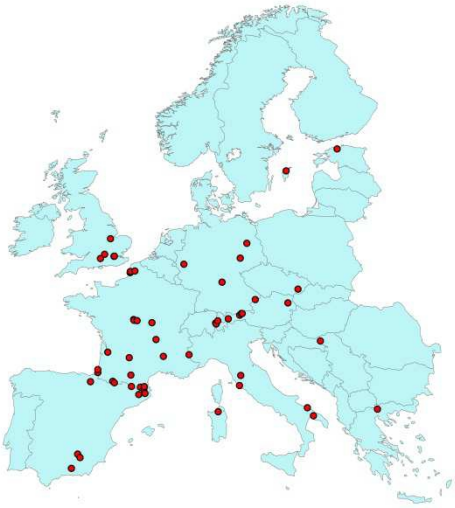
GBIF

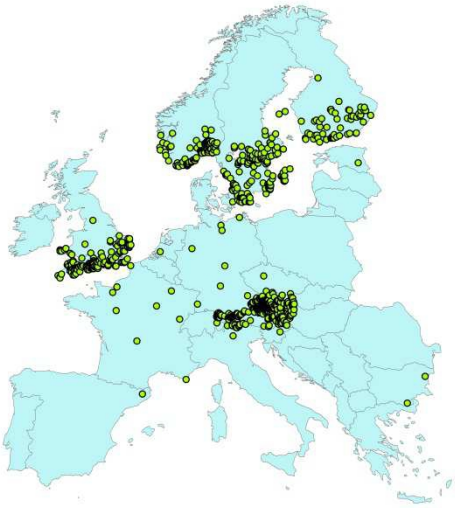




Sphingonaepiopsis gorgoniades

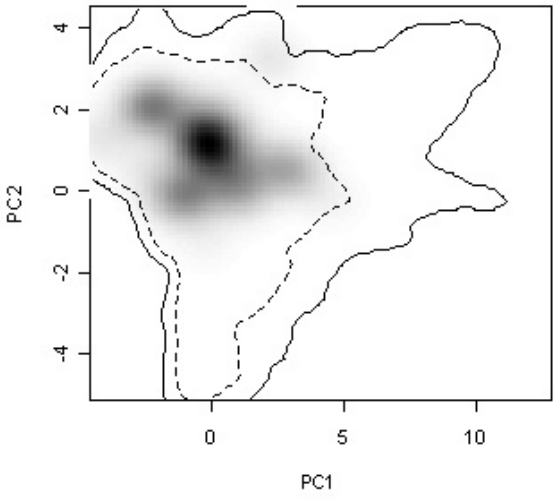
Sphinx ligustri

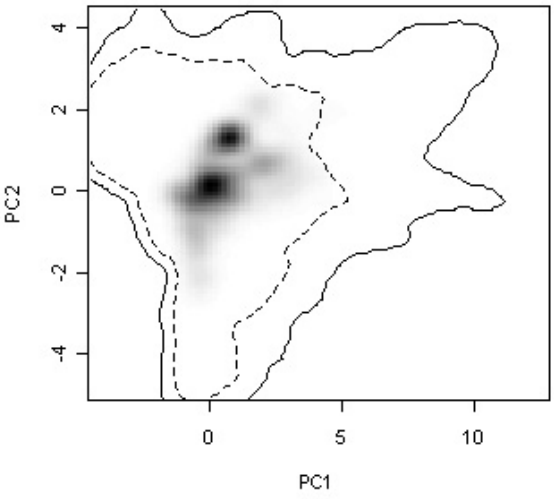




IndColl

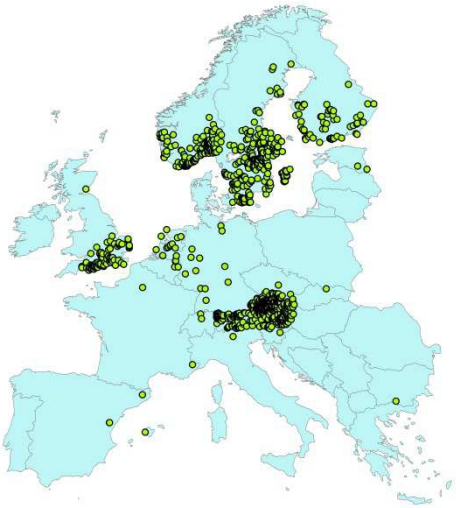
GBIF





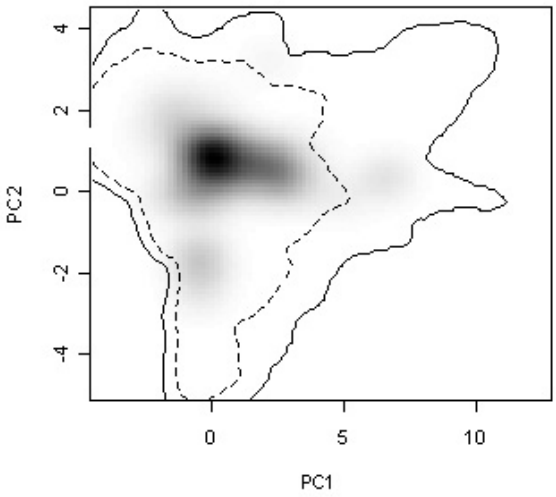
Sphinx maurorum

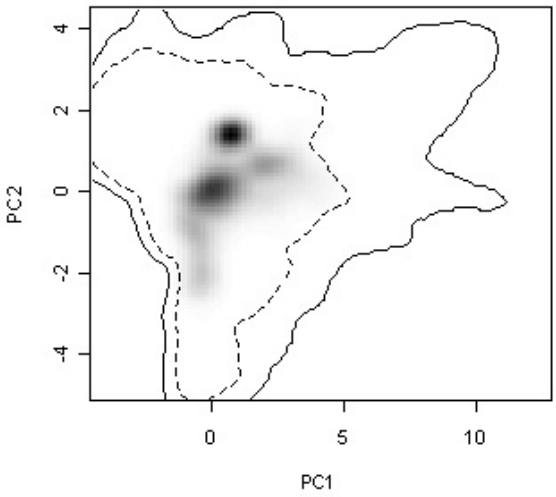
Sphinx pinastri



IndColl

GBIF





Theretra alecto

CHAPTER 4

Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths.

Liliana Ballesteros-Mejia[1*], Ian J Kitching[2], Walter Jetz[3], Peter Nagel[1] & Jan Beck[1]

1) University of Basel, Department of Environmental Science (Biogeography), St. Johanns-Vorstadt 10, 4056 Basel, Switzerland

2) The Natural History Museum, Department of Life Sciences, Cromwell Road, London SW7 5BD, UK

3) Yale University, Department of Ecology and Evolutionary Biology, 165 Prospect Street, New Haven, CT 06520, USA

*) Author for correspondence: Tel.: +41-2670803, E-mail: liliana.ballesteros@unibas.ch

Manuscript for: *Global Ecology and Biogeography* (2013). 22, 586-595

Abstract

Aim: Many taxa, especially invertebrates, remain biogeographically highly understudied and even baseline assessments are missing, with too limited and heterogeneous sampling as key reasons. Here we set out to assess the human geographical and associated environmental factors behind inventory completeness for all hawkmoths of Sub-Saharan Africa. In particular we aim to separate the potential causes of differential sampling from those affecting gradients of species richness to illustrate a potential general avenue for advancing spatial diversity knowledge in understudied groups.

Location: Sub-Saharan Africa

Methods: Using a database of distributional records of hawkmoths, we computed rarefaction curves and estimated total species given sufficient sampling across 200 x 200 km grid cells. We fitted multivariate models to identify environmental predictors of species richness and used environmental co-kriging to map region-wide diversity patterns. We estimated cell-wide inventory completeness from observed and estimated data, and related these to human geographic factors.

Results: Observed geographic patterns of hawkmoth species richness are strongly determined by the number of available records in grid cells. Both show spatially structured distributions. Variables describing vegetation type emerge as important predictors of estimated total richness, and variables capturing heat, energy availability and topographic heterogeneity all show a strong positive relationship. Patterns of interpolated richness identify three centers of highest diversity: Cameroon coastal mountains, and the northern and southern East African montane areas. Inventory completeness is positively influenced by population density, accessibility, protected areas, and colonial history. Species richness is still under-recorded in the western Congo basin and southern Tanzania/Mozambique.

Main conclusions: Sampling effort is highly biased and controlling for it in large-scale compilations of presence-only data is critical for drawing inferences from our still limited knowledge of invertebrate distributions. Our study shows that a baseline estimate of broad-scale diversity patterns in understudied taxa can be derived from combining numerical estimators of species richness, models of main environmental effects, and spatial interpolation. Inventory completeness can be partly predicted from human geographic features and such models may offer fruitful guidance for prioritization of future sampling to further refine and validate estimated patterns of species richness.

Keywords: Co-kriging interpolation, Hawkmoths, Lepidoptera, sampling effort, spatial pattern, Sphingidae.

4.1. Introduction

The compilation and mapping of species richness over large spatial extents have, over the past decade, considerably advanced our understanding of global gradients of diversity and underlying processes (e.g., Jetz & Rahbek, 2002; Currie *et al.*, 2004; Kreft & Jetz, 2007; Field *et al*. 2008). Maps of species richness also offer an important first, if limited (Jetz & Rahbek 2002), guide to identifying regions of potential conservation value (Beck *et al.*, 2011 for a tropical insect example). However, broad-scale studies of diversity gradients are spatially biased (toward well-studied continents such as North America, Europe) and even more so taxonomically, with tropical invertebrates in particular receiving much less attention than their contribution to global biodiversity would dictate (Godfray *et al.*, 1999; Boakes *et al.*, 2010; Beck et al. 2012). Among recent studies on insects on continental to global extents, Jenkins *et al.* (2011) and Guénard *et al.* (2012) have investigated global ant diversity patterns, Beck *et al.* (2006a) have investigated Southeast-Asian sphingid moths, and several additional taxa have been studied at regional scale in the temperate zone (e.g., Hawkins & DeVries, 2009; Kumschick *et al*., 2009; Kundra *et al*., 2011; Hortal *et al*., 2011).

Both geographic and taxonomic biases appear to be a direct function of sampling activity and data availability (Boakes *et al.*, 2010; Beck *et al.*, 2012; Jetz *et al*., 2012), which will depend to some degree on (and correlate with) factors of human geography. Incomplete knowledge of the spatial occurrence of taxa has thus usually prevented the reliable documentation of species richness patterns. Several techniques have been developed to make use of incomplete local inventory data (Colwell & Coddington, 1994) and successfully applied to provide estimates of species richness at larger extents (Beck & Kitching, 2007; Mora *et al.*, 2008; Tittensor *et al.*, 2010). These approaches (and further refinements) combined with increasingly mobilized and integrated distribution information (Jetz *et al.*, 2012; Beck *et al.*, 2012) open up new and exciting prospects for the use of natural history collections data.

Hawkmoths (Lepidoptera, family Sphingidae) are among the most well-known insects with regard to their taxonomy and distribution (Kitching & Cadiou, 2000), and therefore represent an ideal model taxon to study insect macroecology at a global scale. Nevertheless, shortage of distributional data for tropical species has so far prevented detailed, grid-based analyses of their broad-scale species richness patterns in relation to environmental factors (but see Beck *et al*., 2006a). Based on findings for other taxa (Field *et al*., 2008), we expect climatic variables and resulting patterns of

habitat productivity to explain some variation in species richness. Given the herbivorous lifestyle of sphingid caterpillars, we also hypothesize that vegetation type affects their diversity.

However, inventory completeness (which may also affect observed species richness) is ultimately determined by collectors' decisions on where to engage in field sampling. While geographic patterns of sampling intensity will necessarily be partly idiosyncratic (e.g., high record density near places of residence of particular collectors), we also expect some generalities to emerge (Reddy & Dávalos, 2003; Martin *et a*l., 2012). For example, high human population density and dense infrastructure (i.e., traffic accessibility, tourism) should have a positive effect on sampling effort, whereas regions of armed conflict have probably been avoided by collectors (Balmford *et al.*, 2001). Given the impact of European colonialism on Sub-Saharan Africa even after formal political independence of countries, we also expect effects of colonial history. This sort of political history may explain past sampling activity as well as mobilization and data access to date.

Here, using an extensive, expert-validated data compilation, we provide a first quantitative assessment of sphingid moth species richness patterns across Sub-Saharan Africa, using a variety of estimators and specifically addressing sampling effort. We use environmental predictors to identify and model the main correlates of spatial variation in sphingid richness and combine them with spatial interpolation techniques to provide a full sub-continental map of species richness. To assess the robustness of these findings, we specifically quantify patterns of survey completeness (see Moerman & Estabrook, 2006; Guénard *et al.*, 2012; Zagmajster *et al.*, 2010; for relevance to biodiversity research and conservation) and model their potential human geographical determinants. Using African hawkmoths as continental study system, we illustrate how separating the causes of species richness and its sampling facilitates a more rigorous documentation and understanding of the geographic diversity patterns of the many remaining understudied groups.

4.2. Methods

4.2.1. Distribution data

We compiled distribution records for all Sphingidae of Sub-Saharan Africa (south of ca. N17° latitude, including Madagascar), based on an extensive search of published literature and the internet (e.g., Lepidoptera blogs, specimen trading sites, the Barcode of Life Database (http://www.barcodinglife.org), the Global Biodiversity Information Facility (http://www.gbif.org)), as well as correspondence with a large number of professional and amateur collectors, our own field sampling, and through databasing several major natural history collections (e.g., museums in London, Berlin, Paris, Munich, Tervuren and Pittsburgh). We took the utmost care to exclude or correct confirmed or likely errors in locality and species identity. We georeferenced localities as precisely as feasible and applied a unified nomenclature of taxa (following Kitching & Cadiou 2000 and recent, in parts yet unpublished updates; see also Boakes *et al*., 2010). For the purposes of this study, we ignored all locality records that could not be allocated with a precision of at least 1° latitude/longitude (~ 110 km). We defined a record as a unique combination of species, locality, year and collector. Hence, a record may contain between one and many specimens caught at the same time, whereas temporal replicates (e.g., a species being caught repeatedly in different years at the same site) would be considered as separate records. While the oldest data originated from the late 19th century, the vast majority of data were collected later than 1950 (and most from 1980 onward).

We mapped numbers of records (N) and observed species richness (S_{obs}) in an equal area Mollweide projection, aggregated in raster grids with a cell size of 200 x 200 km. Preliminary analyses identified this cell size as the best compromise between resolution and number of cells and data quality within cells.

4.2.2. Correcting for incomplete species inventories

We applied three approaches, all based on the distribution of records and species per grid cell, to attempt to control for variable sampling effort and ultimately provide an estimate of actual grid cell richness values: (1) We calculated rarefaction curves (i.e., randomized accumulation of species with records; S_{rar}) for each grid cell, which allows estimation of how many species would have been observed in a cell if only a given number of specimens had been sampled (Gotelli & Colwell, 2001). Thus, rarefaction allows standardization of sampling effort across cells but outputs are not estimates of the complete species richness of cells (unlike the following methods). We used 25 records as a standard to compare estimated S_{rar}. (2) We fitted several asymptotic functions (i.e.,

Michaelis-Menten, negative exponential, asymptotic, Chapman-Richards, Rational, Weibull; see Mora et al. 2008 for details) to the rarefaction curves to derive estimates of the total species richness expected with infinitely large sampling effort. Each of these functions was evaluated separately for each grid cell using Akaike's information criterion (AIC) for its fit with the rarefaction curves, and AIC-weighted average estimates of species richness (S_{asym}) were calculated. Asymptotic estimators have recently been used by Mora et al. (2008) in a similar context. (3) As an alternative estimator of 'true' species richness we calculated a non-parametric metric, *Chao1* (Chao 1984), that makes use of the ratio of species recorded only once, or exactly twice, per cell (S_{Chao}). Because this method yields results similar to S_{asym} we only mention important data for S_{Chao} in the main text but present details in the Appendix section.

Output from these three approaches varied in quality and reliability between cells, and we applied some 'pruning' rules to remove highly unreliable cell estimates, at the cost of reducing number of cells available for analysis. We present here data for cells where at least 25 records were available, and where coefficients of variation (i.e., standard error of estimate / estimate) for species richness estimators were <0.2. We also repeated our analyses using more (>50 records per cell) and less rigorous (>10, >15 records) pruning rules (i.e., affecting numbers of cells available vs. reliability of estimates), but this did not affect the main conclusions.

4.2.3. Environmental effects on species richness patterns

We investigated the effect of some environmental variables that have often been found or assumed to affect species richness patterns at large extents and grain sizes on $\log_{10}$-transformed estimates of species richness. In particular, we investigated effects of potential evapotranspiration (PET; from http://edit.csic.es/Climate.html) as a measure of energy input into the ecosystem (Hawkins *et al.*, 2003), actual evapotranspiration (AET; from http://edit.csic.es/Climate.html) as a measure of primary productivity (Currie *et al.* 2004), topographic heterogeneity (altitudinal range within cells) as proxy of habitat variability and consequent beta diversity (Ruggiero & Hawkins, 2008), and vegetation structure (herb and tree cover from MODIS Vegetation Continuous Fields, http://glcf.umiacs.umd.edu/data/vcf/, means for 200 km cells). For sphingids, as herbivorous insects, we expected functional links with vegetation type although most species are not particularly host-specific (i.e., specialization below plant family level is rare; Beck *et al.*, 2006b). MODIS data are based on satellite imagery taken in 2000-2001 and hence include aspects of human-induced changes to the landscape. Vegetation data are correlated with AET estimates (tree cover: r^2= 0.62; herb cover: r^2= 0.30), which may affect interpretation of results (see below). Coastal raster cells may appear to harbour reduced species richness due to smaller area alone. However, in our data set

coastal regions often contained well-sampled and species-rich cells, and given this sampling pattern the effect of land area on observed richness was weak (Spearman rank correlation, 405 cells: r_s = -0.105). We thus included coastal cells down to 5.0% land area in the analysis to avoid loss of critical information. We tested model residuals for spatial autocorrelation (software SAM, v.4; 999 permutations), finding significant Moran's I >0.1 for lag-distances up to ca. 500 km. We used spatially explicit multivariate generalised least square (GLS) models to account for spatial autocorrelation in the data (Beale *et al.*, 2010; spherical variogram structure; software R.2.13.1, *nlme* package).

For all three response variables (i.e., S_{rar}, S_{asym}, S_{Chao}), we evaluated full models (all listed variables, no interactions) and various simplified models using Akaike's information criterion (AIC); we used only the best (lowest AIC) for further analyses. We calculated the pseudo-R^2 of models as a correlation of predicted vs. observed values. GLS model coefficients were used to extrapolate species richness estimates across Sub-Saharan Africa, allowing intuitive evaluation of the consistency of patterns derived from the three estimation methods. We also applied co-kriging (i.e., spatial interpolation of raw estimates based on their autocorrelation, the autocorrelation of environmental model predictions, and the cross-correlation between them) for mapping (Kreft & Jetz, 2007). Co-kriging was carried out in ArcGIS 10 software, assuming anisotropic variogram structures. Estimates were optimized by cross-validation, and we report final root mean square errors (RMSE) of interpolation predictions.

4.2.4. Quantifying and analysing inventory completeness

Using co-kriging estimates of 'true' species richness (S_{asym}), we determined cell-wide species inventory completeness as S_{obs}/S_{asym} and yet unrecorded species richness as S_{asym}-S_{obs}. Cells without any data consequently had an inventory completeness of zero. In some cells estimates of S_{asym} were lower than S_{obs} due to imperfect function fitting; for these we defined inventories as complete (i.e., S_{obs}/S_{asym} = 1) and set the number of unrecorded species to zero (S_{asym}-S_{obs}= 0).

We related the geographic patterns of inventory completeness to human factors such as road and tourism infrastructure, habitat encroachment, population density, armed conflict, and colonial history (see Appendix 4.1 for details and sources). We hypothesized that each of these factors may play a role in affecting collectors' inclination to be active in a region. Inventory completeness is a zero-inflated response variable (i.e., there are many cells without records; Zuur et al. 2010) and we used $\log_{10}(x+1)$-transformation to reduce extreme deviations from normality. We first carried out AIC-based model selection of ordinary least square (OLS) models to identify important effects of these variables on inventory completeness. For the best model (lowest AIC), we found significant

positive autocorrelation in residuals for lag-distances up to ca. 870 km. Using the variables in the best OLS-model, we re-analysed effects in a spatially explicit GLS model, and we repeated this analysis without the zero-cells to avoid spurious conclusions due to zero-inflation.

4.3. Results

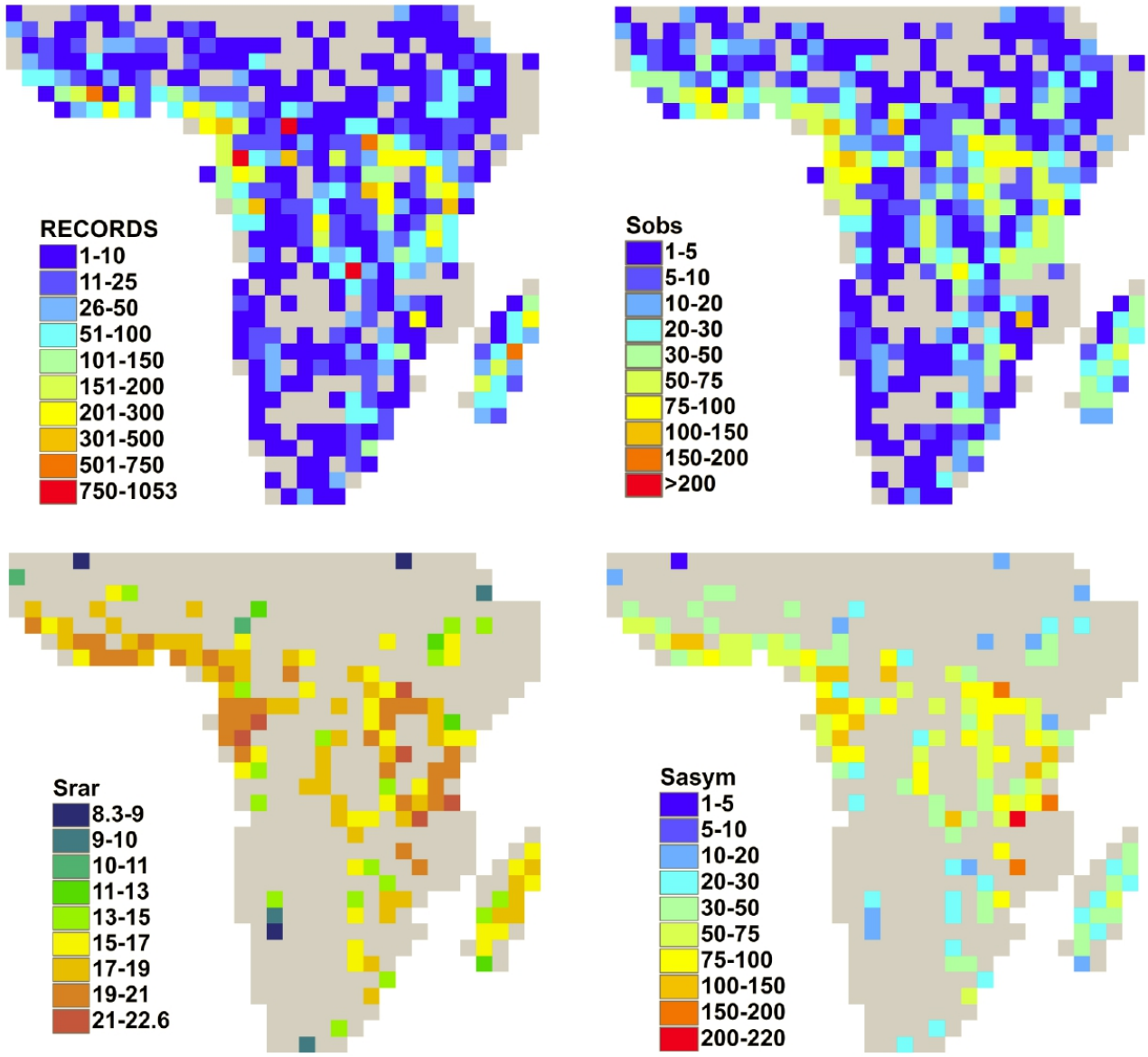
4.3.1. Observed and estimated species richness

A total of 21,194 records provide occurrence data for 322 species over 405 grid cells (of 200 x 200 km size) covering all of Sub-Saharan Africa (145 additional grid cells had no data available; Figure 4.1). After applying 'pruning' rules of data inclusion for estimating species richness (see Methods), 146 cells were left for analyses.

Most occurrence samples (N) come from the coastal parts of western and central Africa, from the Great Lakes regions of eastern Africa and from Madagascar. Few records are available for the drier parts of southern Africa. This geographically highly uneven availability of occurrence data was strongly reflected in the patterns of observed species richness (S_{obs}). N and S_{obs} are strongly positively correlated (linear correlation of $logN \sim logS_{obs}$: $r = 0.95$, n = 405 grid cells), suggesting pervasive effects of sampling effort even at this coarse spatial resolution. Restricting the test to the 146 cells with ≥25 records confirms this relationship ($r = 0.90$).

To overcome these sampling effects of richness we calculated rarified species richness (S_{rar}) and estimated expected full species richness given sufficient sampling using parametric asymptotic (S_{asym}) and non-parametric (S_{Chao}) methods. S_{asym} showed similar geographic patterns (Figure 4.1C). These measures accordingly exhibited much weaker relationships with N (i.e., $logN \sim logS_{asym}$, $r = 0.65$; $logN \sim logS_{Chao}$, $r = 0.64$) and there was barely an association with rarefied species richness ($logN \sim logS_{rar}$, $r=0.45$). Relationships with observed species richness were also weak ($S_{obs} \sim S_{asym}$, $r = 0.78$; $S_{obs} \sim S_{Chao}$, $r= 0.70$; $log_{10}S_{obs} \sim S_{rar}$, $r= 0.67$). Estimates of full species richness agree with each other (i.e., $S_{asym} \sim S_{Chao}$, $r= 0.94$) while showing some deviation from rarefied data ($S_{rar} \sim logS_{asym}$, $r = 0.88$; $S_{rar} \sim logS_{Chao}$, $r= 0.84$).

Figure 4.1 (A) Number of records (N), (B) observed species richness (S_{obs}), (C) rarefied species richness at 25 records (S_{rar}), (D) asymptotic estimate of total species richness (S_{asym}). Note that S_{obs} and S_{asym} are shown on the same colour scale. Grid cells without data are shown in grey.



4.3.2. Environmental models and interpolation

For both rarefied species richness (S_{rar}), and asymptotic estimators (S_{asym}), the strongest environmental models included all predictors according to AIC (see Table 4.1). Notably, environment explains considerably more of the variation in S_{rar} (pseudo-R^2 = 0.41) than S_{asym} (pseudo-R^2 = 0.14), and the models agree only partly in the importance of variables. For both models, positive coefficients of similar magnitude were found for tree and herb cover. The model for S_{rar} is additionally driven by PET and, more weakly, topographic heterogeneity, whereas that for S_{asym} is mainly affected by AET. Figure 4.2 illustrates these differences by extrapolating the models

(note, e.g., different prediction for the Congo Basin, a region of very high AET). The model for S_{Chao} shows performance similar to that of S_{asym} (pseudo-R^2 = 0.15; Appendix 4.2).

Environmental models based on different species richness estimates lead to broadly similar predicted patterns of diversity (Figure 4.2; $S_{rar} \sim \log_{10}S_{asym}$, r = 0.97; N=550), and so did co-kriging interpolations ($S_{rar} \sim \log_{10}S_{asym}$, r = 0.95). Estimates of S_{asym} and S_{Chao} (see Appendix 4.3) are correlated for the environmental model (r = 0.92) and even stronger for co-kriging (r = 0.97). Surprisingly, predictions from environmental models and co-kriging interpolations within the same metrics deviate considerably from each other (S_{rar}: r = 0.83; S_{asym}: r = 0.75; S_{Chao}: r = 0.68). Deviations are particularly strong in the Ethiopian Highlands and the Horn of Africa (the environmental model predicts more species in the former, fewer in the latter, than co-kriging). For both estimates of total species richness (S_{asym} and S_{Chao}), environmental models predict more species in the Congo Basin and fewer in Tanzania/Mozambique than co-kriging (residual data not shown).

Table 4.1 Generalised least squares (GLS) model details for rarefied species richness (S_{rar}) and asymptotic estimates of species richness (S_{asym}). Pseudo-R^2 = 0.405 for S_{rar}, 0.138 for S_{asym} (n = 146 grid cells).

$\log_{10}S_{rar}$				$\log_{10}S_{asym}$		
Variable	*Coefficient*	*t*	*P*	*Coefficient*	*t*	*p*
(Intercept)	0.62548	7.307	0.000	0.52334	1.258	0.211
Topo. Het.	0.00001	1.839	0.068	0.00003	1.348	0.180
AET	0.00004	1.255	0.212	0.00032	2.273	0.025
PET	0.00010	2.253	0.026	0.00016	0.760	0.448
Tree	0.00430	6.074	0.000	0.00544	1.893	0.061
Herb	0.00404	6.817	0.000	0.00562	2.286	0.024

4.3.3. Inventory completeness

We used the predictions of the best-performing model of total species richness, co-kriging of S_{asym}, to estimate the geographic variation in inventory completeness (S_{obs}/S_{asym}) and undetected species richness (S_{asym}-S_{obs}; Figure 4.3). Model selection based on AIC (see Methods) led to a model including population density, railway lines, airports, touristic hotspots, protected areas and colonial history as the most important variables for predicting inventory completeness (explaining ca. 21% of data variability). Coefficients (Table 4.2) reveal the expected positive effects of infrastructure (traffic access, tourism) and of protected areas, but also effects of colonial history (although the large majority of data stemmed from the post-colonial era). In particular, the formerly Portuguese regions (i.e., Mozambique), but surprisingly also former British regions, were less well-sampled or -

mobilized than formerly French and Belgian regions. A univariate model (not shown) that does not account for differences in infrastructure (which may itself be an outcome of colonial history) confirms the effect of past Portuguese occupation but no other effects of colonial history. Models with slightly higher AIC (ΔAIC <2; data not shown) contain additional positive effects of road density and negative effects of pristine regions. A map of residuals from the OLS model (not shown) indicates only weak spatial structure, with particularly positive residuals (i.e., better sampling than predicted) in Madagascar and Cameroon, and negative residuals in the Sahel, western Congo Basin and Zimbabwe.

Inventory completeness (Figure 4.3, left) is related to ($\log_{10}(x+1)$-transformed) number of records (Figure 4.1; $r^2 = 0.85$), and repeating the OLS analysis with records as a response variable (a proxy of sampling effort) lead to identical conclusions (not shown).

When looking at absolute numbers of yet-to-be-recorded species, Mozambique and southern Tanzania, as well as the Congo Basin, stand out as containing much unrecorded (at grid cell level, not necessarily undescribed) biodiversity (Figure 4.3, right).

Table 4.2 OLS and GLS models explaining estimated inventory completeness (Figure 4.3; log10(x+1)-transformed) of cells by human geographic factors. Country names refer to colonial powers in 1919 (see Appendix 4.1 for details on predictor variables; n = 502 grid cells). Data are zero-inflated, but a GLS model without the 145 zero-cells recovered all results except the marginal effect of protected areas (Appendix 4.5).

	OLS; $\boldsymbol{R^2_{adj}}$ **= 0.21**			GLS**; pseudo-**$\boldsymbol{R^2}$ **= 0.22**		
	Coefficient	*t*	*p*	*Coefficient*	*t*	*p*
(Intercept)	0.03285	2.356	0.019	0.03286	2.340	0.020
Britain	-0.02825	-2.837	0.005	-0.02828	-2.815	0.005
Belgium	0.01243	0.838	0.402	0.01190	0.794	0.428
Portugal	-0.06021	-4.015	0.000	-0.06045	-3.997	0.000
France	0*			0*		
$\log_{10}$(Popul+1)	0.02954	4.427	0.000	0.02980	4.433	0.000
Airports	0.04446	3.378	0.001	0.04387	3.350	0.001
Railways	0.00015	3.064	0.002	0.00015	3.050	0.002
Tourism	0.04163	3.968	0.000	0.04115	3.936	0.000
Protected	0.01691	1.982	0.048	0.01652	1.942	0.053

*) zero by default

Figure 4.2 (A) Estimates of species richness based on the environmental model of rarefied species richness at 25 records ($S_{rar\ (environ)}$, upper left), (B) Co-kriging interpolation of rarefied species richness ($S_{rar\ (co\text{-}krig)}$, upper right), (C) Environmental model of asymptotic estimate of total species richness ($S_{asym\ (environ)}$, lower left), (D) Co-kriging interpolation of asymptotic estimate of total species richness ($S_{asym\ (co\text{-}krig)}$, lower right). See Table 4.1 for details on environmental models. RMSEs for co-kriging interpolations are 1.95 for S_{rar} and 29.05 for S_{asym}.

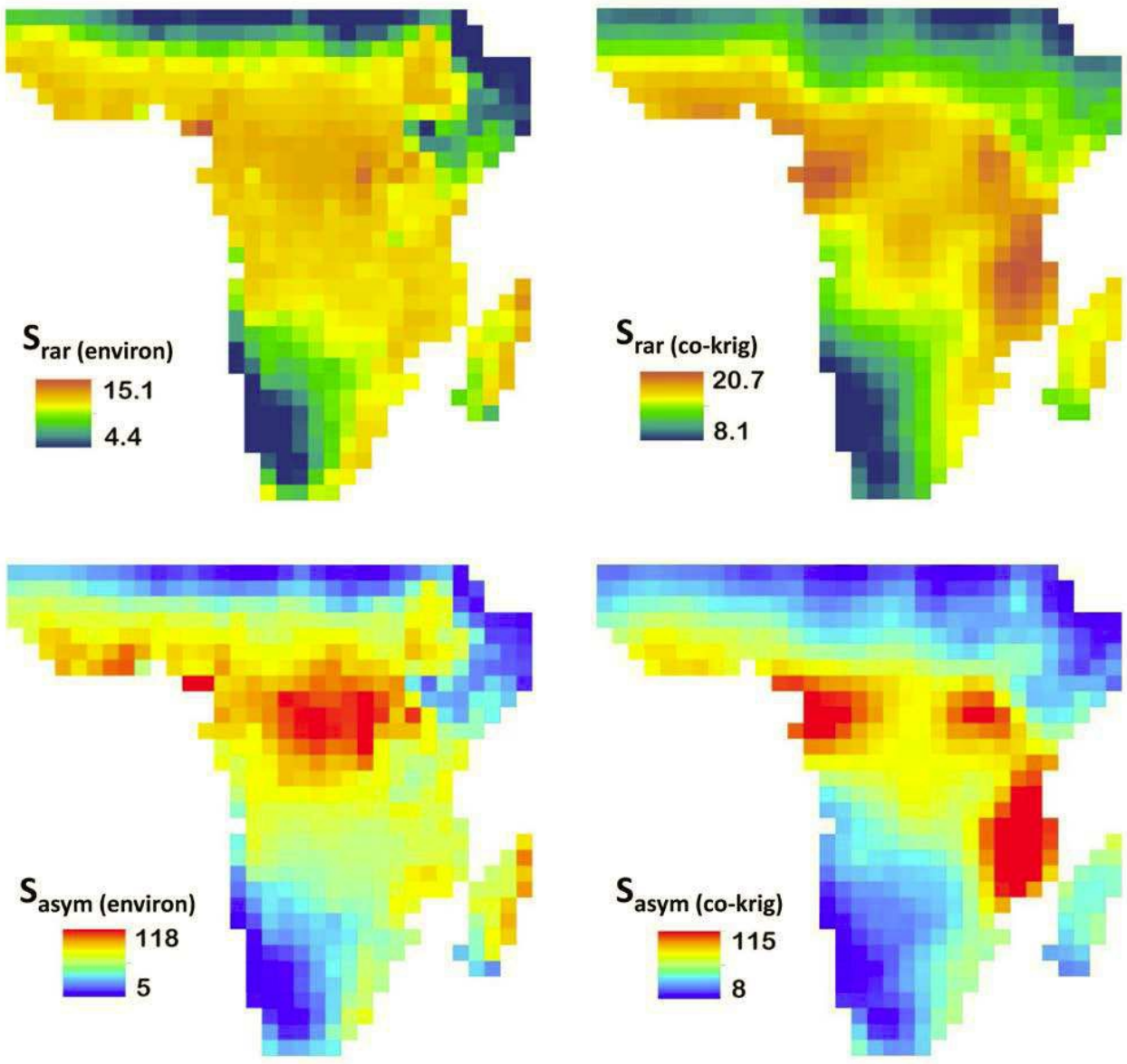
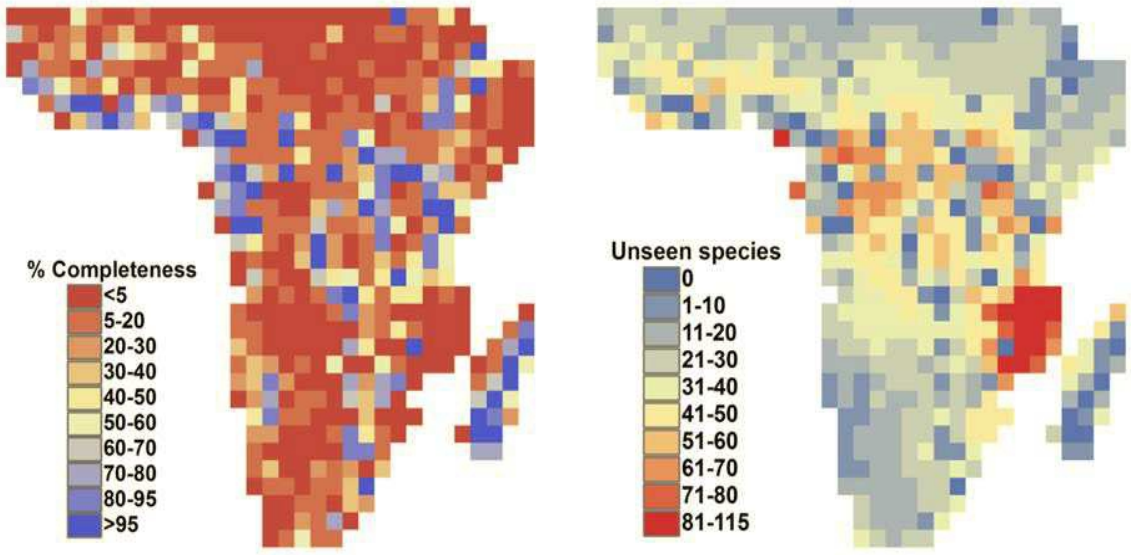


Figure 4.3 (*Left*) Cell-wide inventory completeness (based on co-kriging estimate of S_{asym}; these data were used to investigate effects of human geographic factors, Table 4.2). (*Right*) Estimated number of unrecorded species in each grid cell measured as the difference between S_{obs} and S_{asym}.



4.4. Discussion

4.4.1. Controlling species richness for sampling effort

Our analyses demonstrate that for incompletely sampled taxa (i.e., the great majority of taxa in most regions), numerical estimates of cell-wide species richness based on the relative distributions of records and species can provide data that enables first large-scale mapping and analysis of diversity. These sorts of assessments are urgently needed to put global biodiversity research on a broader taxonomic basis. As observed data are often heavily affected by sampling effort (e.g., Palmer *et al*., 2002; Boakes *et al*., 2010), such estimates may currently be the only alternative to overlaying estimated range maps of individual species (based on expert knowledge or distribution modelling). Expert range maps for individual species are, for tropical regions, currently only available for vertebrates, and they can have spatial characteristics different from cell-based estimates, with consequences for further analysis and inference (McPherson & Jetz, 2007).

Although estimates of total species richness are easiest to understand and interpret, our data suggest that rarefaction may currently be the more reliable method of controlling for sampling effort in diversity patterns. We found S_{rar} to be less dependent on record numbers than S_{asym} or S_{Chao}, and the environmental model based on S_{rar} explained considerably more variability. This indicates that rarefaction introduces less random error than extrapolation. Based on similar arguments, Fiedler & Truxa (2012) recently came to the same conclusions for finer-scale data.

4.4.2. Environmental effects and spatial interpolation

We found positive effects of energy-related variables in environmental models, but there was inconsistency between models whether links with AET (a proxy of primary productivity) or PET (a proxy of solar energy input) are more important. Plausible mechanisms have been postulated for both variables (see Evans *et al*., 2005 for review), and published analyses leave uncertainty similar to that revealed here (e.g., Mittelbach *et al*., 2001; Currie *et al*., 2004; Buckley & Jetz, 2007). Energy availability was found to have a large effect on regional and local richness (Jetz & Fine, 2012). Additionally temperature was found to be positively associated with ectotherm richness whereas primary productivity is correlated with endotherm richness (Buckley *et al.*, 2012). At a much smaller scale temperature was found to be negatively associated with butterfly richness (Stefanescu *et al.*, 2004). Possibly the coarse-scale, imprecise measurement of currently available AET data prevents clear distinction between these effects. Interesting is the high similarity of coefficients for tree and herb cover (i.e., forest vs. savannah) after controlling for energy and productivity, suggesting that other differences between those habitat types (such as 3-D structure) are not very important at this spatial scale of analysis.

We found relatively weak correspondence of patterns recovered from the environmental models and from co-kriging. Spatial interpolation can yield equal or better estimates than environmental models (e.g., Bahn & McGill, 2007; Lin *et al*., 2008), although they are less informative with regard to the causes of patterns. By being closer to observed data patterns, interpolation can also map historical effects undetectable by correlation with the current environment. Our co-kriging estimates (Figure 4.2) clearly identify three areas of high diversity, i.e. the coastal mountains of western Central Africa, and the northern and southern mountain ranges of East Africa. Notably, this pattern is not explained by topographic heterogeneity (which was included in environmental models). All three regions were identified as regions of complex biogeographical history and high endemism in other taxa (e.g., Jetz *et al*., 2004; Linder *et al*., 2012), suggesting potential effects of geographical history. Also, co-kriging interpolations yield patterns of species richness broadly similar to those published for birds (Jetz & Rahbek 2002; Lin *et al.*, 2008), amphibians (Buckley & Jetz 2007) and plants (Kreft & Jetz 2007). Scale-dependency of species richness hinders quantitative comparison across studies (Rahbek 2005, Beever *et al*. 2006). However, the prediction of high diversity not only in the montane areas of southern Tanzania and Mozambique but also in their coastal lowlands appears novel; it is unwarranted by actual data (Figure 4.1) and requires further data collection for confirmation.

4.4.3. Sampling effort and the large-scale evaluation of biodiversity

Our data showed clearly that cell-wide inventory completeness was not equally distributed in space (Figure 4.3). However, the causality of the relationship between sampling effort (i.e., number of records) and observed species richness is not entirely clear. Collectors may be drawn particularly to places known or presumed to be high in species diversity (which often also feature high human population density; Balmford *et al.*, 2001). Alternatively, more comprehensive sampling may lead to finding more species.

Inventory completeness was substantially related to accessibility and infrastructure. Modelling cannot infer causality directly, but it is plausible to conclude that collectors make conscious decisions to visit those places that are easy to access. Protected areas had a positive effect on inventory completeness, although it cannot be concluded from our data whether this is caused by specific conservation interest in surveys or by the infrastructure allowing access to, e.g., National Parks. “Pristine” landscapes, on the other hand, had a (weak) negative effect, which is most likely due to lack of access. This (non-significant) effect is somewhat in contradiction to the assumption that such places have often a more complete inventory, and also to Guernard *et al.* (2012), who estimated many unrecorded ant genera in regions of high anthropogenic habitat destruction.

Even though some patterns of model residuals match most Africa-researchers’ preconceived expectations on collection intensity (e.g., poor knowledge of the Congo Basin, well-sampled Madagascar), we mostly observed only idiosyncratic deviations from model expectations of inventory completeness. Some large positive residuals (more complete data than modelled) seemed to be associated with single places with large quantities of data collected over a few years, suggesting intense activity by a single collector or a particular survey program. Additionally, georeferencing issues could also cause such effects. Records saying nothing but “Kivu”, for example, were referenced to the same coordinates whereas they could sometimes relate to a much wider geographic interpretation, i.e. the Kivu provinces or the entire region around Lake Kivu. Furthermore, some well-sampled places did not stand out in the population or traffic network, but they may nevertheless have drawn collectors (such as expatriate workers, missionaries) because of their administrative (e.g., Yaounde, Cameroon’s capital) or economic importance (e.g., Lubumbashi, a mining town in the Congo). We did not include locations of universities (cf. Moerman & Estabrook, 2006) in our analyses as very little of our data stemmed from African collectors or collections (e.g., databasing the collection of the Natural History Museum of Addis Ababa yielded <10% of available records for Ethiopia). Exploratory analysis of inventory completeness and human geographic variables, using geographically weighted regression (not shown), suggested some spatial patterns that deserve further study, such as increasing predictability of inventory completeness from West to East.

4.5. Conclusions

Sampling effort is a crucial variable when assessing large-scale species richness patterns and ignoring this would probably lead to flawed perceptions of patterns. Numerical estimates based on the accumulation of species with records, in combination with environmental models and spatial interpolation can help estimating broad-scale richness patterns. However, they necessarily contain estimation error, and important patterns should be backed up by future field surveys. Similar to what has been found for the much better studied vertebrates, vegetation cover, energy-related variables and topographic heterogeneity are important environmental correlates also for sphingid moth species richness, while leaving considerable variation unexplained possibly due to historical component in the patterns of species richness. Inventory completeness can be predicted to a certain degree from human population density, infrastructure and colonial history. Our approach and results expose areas of extensive and poor sampling given expected discoverable species richness, thus highlighting regions where future efforts of sampling should be directed.

4.6. Acknowledgements

We thank all the professional and amateur collectors (too numerous to mention here) who made their data available for our project. S.P. Loader, W. Schwanghart and two anonymous reviewers provided critical comments on an earlier version of the manuscript. M. Curran, M. Kopp, R. Hagmann, S. Widler and S. Lang helped processing the distributional data. The study received financial support from the Swiss National Science Foundation (SNF, project 3100AO_119879) and the Synthesys program of the EU.

4.7. References

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733-742.

Balmford, A., Moore, J.L., Brooks, T., Burgess, N., Hansen, L. A, Williams, P. & Rahbek, C (2001) Conservation conflicts across Africa. *Science*, **291**, 2616-9.

Beale, C., Lennon, J., Yearsley, J. & Brewer, M. (2010) Regression analysis of spatial data. *Ecology Letters*, **13**, 246:264.

Beck, J. & Kitching, I.J. (2007) Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. *Journal of Applied Ecology*, **44**, 672-681.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S. a., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M. & Dormann, C.F. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673-683.

Beck, J., Kitching, I.J. & Eduard Linsenmair, K. (2006a) Determinants of regional species richness: an empirical analysis of the number of hawkmoth species (Lepidoptera: Sphingidae) on the Malesian archipelago. *Journal of Biogeography*, **33**, 694-706.

Beck, J., Kitching, I.J. & Linsenmair, K.E. (2006b) Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. *Ecotropica*, **12**, 1–13.

Beck, J., Schwanghart, W., Chey, V.K. & Holloway, J.D. (2011) Predicting geometrid moth diversity in the Heart of Borneo. *Insect Conservation and Diversity*, **4**, 173-183.

Beever, E.A., Swihart, R.K. & Bestelmeyer, B.T. (2006) Linking the concept of scale to studies of biological diversity: evolving approaches and tools. *Diversity and Distributions*, **12**, 229–235.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Ding, C.Q., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS biology*, **8**, e1000385.

Buckley, L.B., Hurlbert, A.H. & Jetz, W. (2012) Broad-scale ecological implications of ectothermy and endothermy in changing environments. *Global Ecology and Biogeography*, **21**, 873–885.

Buckley, L.B. & Jetz, W. (2007) Environmental and historical constraints on global patterns of amphibian richness. *Proceedings of the Royal Society (B)*, **274**, 1167-1173.

Chao A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11,** 265-270.

Colwell, R.K. & Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society (B)*, **345**, 101-18.

Currie, D.J., Mittelbach, G.G., Cornell, H.V., Field, R., Guegan, J.-F., Hawkins, B. a., Kaufman, D.M., Kerr, J.T., Oberdorff, T., O'Brien, E. & Turner, J. R. G. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters*, **7**, 1121-1134.

Evans, K.L., Warren, P.H. & Gaston, K.J. (2005) Species–energy relationships at the macroecological scale: a review of the mechanisms. *Biological Reviews*, **80**, 1–25.

Fiedler, K. & Truxa, C. (2012) Species richness measures fail in resolving diversity patterns of speciose forest moth assemblages. *Biodiversity and Conservation*, **21**, 2499-2508.

Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guegan, J.-F., Kaufman, D. M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M. & Turner, J.R.G. (2009) Spatial species-richness gradients across scales: a meta-analysis. *Journal of Biogeography*, **36**, 132-147.

Godfray, H.C.J., Lewis, O.T. & Memmot, J. (1999) Studying insect diversity in the tropics. *Philosophical Transaction of the Royal Society (London) B*, **354**, 1811-1824.

Gotelli, N. & Colwell, R. (2001) Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Guénard, B., Weiser, M.D. & Dunn, R.R. (2012) Global models of ant diversity suggest regions where new discoveries are most likely are under disproportionate deforestation threat. *Proceedings of the National Academy of Science*, **109,** 7368-7373.

Hawkins, B. a. & DeVries, P.J. (2009) Tropical niche conservatism and the species richness gradient of North American butterflies. *Journal of Biogeography*, **36**, 1698-1711.

Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guegan, J.-F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E., Porter, E.E. & Turner, John R. G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105-3117.

Hortal, J., Diniz-Filho, J.A.F., Bini, L.M., Rodriguez, M.A., Baselga, A., Nogues-Bravo, D., Rangel, T.F. Hawkins, B.A. & Lobo, J.M. (2011) Ice age climate, evolutionary constraints and diversity patterns of European dung beetles. *Ecology Letters*, **14**, 741–748.

Jenkins, C.N., Sanders, N.J., Andersen, A.N., Arnan, X., Brühl, C. A., Cerda, X., Ellison, A.M., Fisher, B.L., Fitzpatrick, M.C., Gotelli, N.J., Gove, A.D., Guénard, B., Lattke, J.E., Lessard, J.-P., McGlynn, T.P., Menke, S.B., Parr, C.L., Philpott, S.M., Vasconcelos, H.L., Weiser, M.D. & Dunn, R.R. (2011) Global diversity in light of climate change: the case of ants. *Diversity and Distributions*, **17**, 652-662.

Jetz, W. & Fine, P.V. (2012) Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. *PloS Biology*, **10,** e1001292, doi: 10.1271/journal.pbio.1001292.

Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548-51.

Jetz, W., Rahbek, C. & Colwell, R.K. (2004) The coincidence of rarity and richness and the potential signature of history in centres of endemism. *Ecology Letters*, **7**, 1180–1191.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution*, **27**, 151-159.

Kitching I.J. & Cadiou J.-M. (2000) *Hawkmoths of the world*. The Natural History Museum, London. Cornell University Press, London.

Kreft, H. & Jetz, W. (2007) Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences*, **104**, 5925-30.

Kudrna, O., Harpke, A., Lux, K., Pennersdorfer, J., Schweiger, O., Settele, J. & Wiemers, M. (2011) *Distribution Atlas of Butterflies in Europe*. Gesellschaft für Schmetterlingsschutz. Halle, Germany.

Kumschick, S., Schmidt-Entling, M.H., Bacher, S., Hickler, T., Espadaler, X., & Nentwig, W. (2009) Determinants of local ant (Hymenoptera: Formicidae) species richness and activity density across Europe. *Ecological Entomology*, **34**, 748–754.

Lin, Y.-P., Yeh, M.-S., Deng, D.-P. & Wang, Y.-C. (2007) Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. *Global Ecology and Biogeography*, **17**, 175–188.

Linder, H.P., de Klerk, H.M., Born, J, Burgess, N.D., Fjeldsa, J. & Rahbek, C. (2012) The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. *Journal of Biogeography*, **39**, 1189–1205.

Martin, L.J., Blossey, B. & Ellis, E. (2012) Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, **10**, 195–201.

McPherson, J.M. & Jetz, W. (2007) Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. *Global Ecology and Biogeography*, **16**, 657–667.

Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I. & Gough, L. (2001) What is the observed relationship between species richness and productivity? *Ecology*, **82**, 2381–2396.

Moerman, D.E. & Estabrook, G.F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*, **33**, 1969–1974.

Mora, C., Tittensor, D.P. & Myers, R.A. (2008) The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society (B)*, **275**, 149-155.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S. & Wohlgemuth, T. (2002) Quantitative tools for perfecting species lists. *Environmetrics*, **13**, 121-137.

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Reddy, S. & Dávalos, L. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719-1727.

Ruggiero, A. & Hawkins, B.A. (2008) Why do mountains support so many species of birds? *Ecography*, **31**, 306-315.

Stefanescu, C., Herrando, S. & Paramo, F. (2004) Butterfly species richness in the north-west Mediterranean Basin: the role of natural and human-induced factors. *Journal of Biogeography*, **31**, 905-912.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V. & Worm, B. (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature*, **466**, 1098-1101.

Zagmajster, M., Culver, D., Christman, M. & Sket, B. (2010) Evaluating the sampling bias in pattern of subterranean species richness: combining approaches. *Biodiversity and Conservation*, **19**, 3035-3048.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1,** 3-14.

Appendix

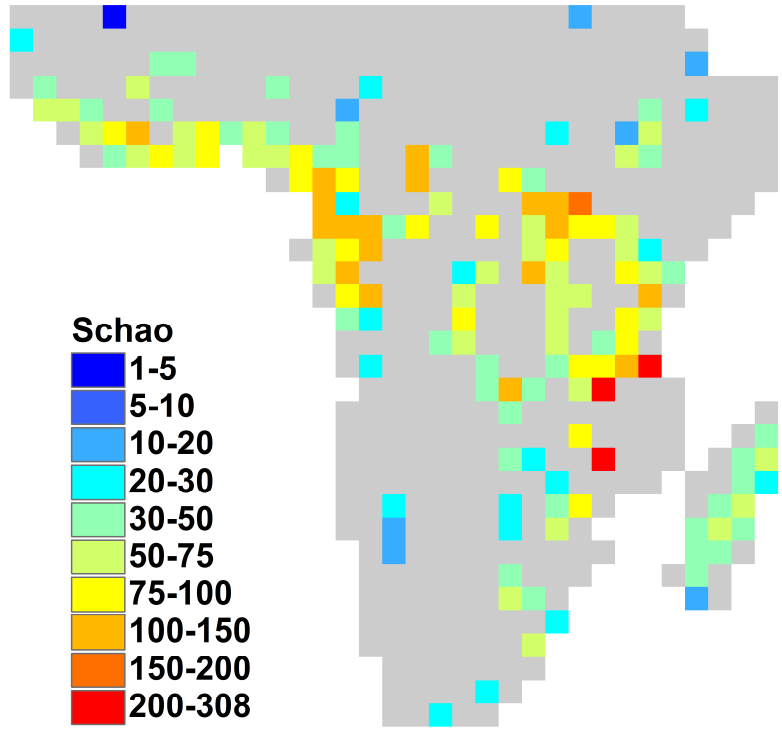
Appendix 4.1. Human geographic factors (continuous [cont.], categorical [cat.] or presence-absence [P/A]), data sources, and modelled effects on inventory completeness of 200 x 200 km cells. All online sources were accessed in May 2010. Tourism hotspots were identified as the "top ten places" of each country as listed in Lonely Planet guide book series (considered the most authorities source of information by most Africa travellers). We coded 'colonial history' as presence of absence of Great Britain, France, Belgium and Portugal in the main part of each grid cell in 1919 (diplomatic refinements, e.g. colony vs. protectorate, were ignored); we excluded the few grid cells with other colonial history or no data for other variables, leaving 502 grid cells in analysis.

Variable	**Data source**
Road density [area of 2 km buffer, cont.]	http://www.diva-gis.org/gData
Railway density [area of 2 km buffer, cont.]	http://www.diva-gis.org/gData
Airports [P/A]	http://goafrica.about.com/
Tourism hotpots [P/A]	http://www.lonelyplanet.com/africa
Protected areas [P/A]	http://www.wdpa.org/
Pristine nature areas [P/A]	http://www.ciesin.columbia.edu/wild_areas/
Colonial history, in 1919 [cat.]	
Human population, 2005 [cont.]	http://gcmd.nasa.gov/records/GCMD_Landscan.html
Armed conflict since 1945 [P/A]	http://www.prio.no/

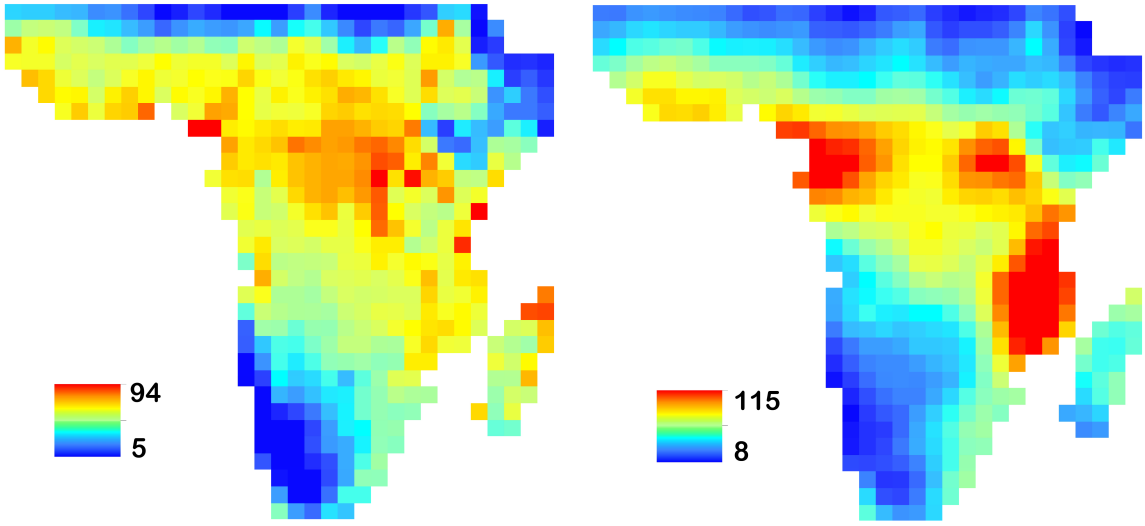
Appendix 4.2 GLS model details for Chao1-estimates of species richness (S_{Chao}). Note that for S_{Chao} the best model (lowest AIC) was not the full model, but one without AET. Pseudo-R^2 = 0.145 (N = 146 gird cells).

$\log_{10}S_{Chao}$: **AIC = -36.0**				
Variable	*Coefficient*	*SE*	*t*	*p*
(Intercept)	0.352494	0.442870	0.796	0.426
Topo. Het.	0.000044	0.000023	1.902	0.059
PET	0.000366	0.000224	1.629	0.106
Tree	0.008514	0.002187	3.894	0.000
Herb	0.007185	0.002379	3.020	0.003

Appendix 4.3 *Chao1*-estimated species richness. Grey cells denote no data.



Appendix 4.4 Species richness estimates based on S_{Chao}. *(Left)* Extrapolation of environmental model (Table A2); *Right* Co-kriging extrapolation of S_{Chao} (RMSE = 35).



Appendix 4.5 Spatially explicit model explaining estimated inventory completeness (Figure 4.3; log10(x+1)-transformed) by human geographic factors, using only cells with at least one species recorded (i.e., no zero-inventory completeness).

N = 367	GLS**; pseudo-R^2 = 0.18**			
	Coefficient	*SE*	*t*	*p*
(Intercept)	0.079839	0.016615	4.805	0.000
Britain	-0.040175	0.011798	-3.405	0.001
Belgium	-0.020264	0.016157	-1.254	0.211
Portugal	-0.074995	0.018758	-3.998	0.000
France	0*			
$\log_{10}$(Popul+1)	0.025628	0.007868	3.257	0.001
Airports	0.037299	0.014017	2.661	0.008
Railways	0.000121	0.000055	2.221	0.027
Tourism	0.032594	0.011408	2.857	0.005
Protected	0.011792	0.009702	1.215	0.225

*) zero by default

CHAPTER 5

Addressing the Wallacean shortfall: Distribution and biodiversity of the hawkmoths of the Old World

Liliana Ballesteros-Mejia[1*], Ian J. Kitching[2], Peter Nagel[1], Jan Beck[1]

[1] University of Basel, Department of Environmental Science (Biogeography), St. Johanns-Vorstadt 10, 4056 Basel, Switzerland

[2] The Natural History Museum, Department of Entomology, Cromwell Road, London SW7 5BD, UK.

*Author for correspondence: Tel.: +41-2670803, E-mail: liliana.ballesteros@unibas.ch

ABSTRACT

For the vast majority of the species described today, there is very little knowledge about their distribution and ecology, a phenomenon called Wallacean Shortfall. Available distribution data are biased towards very few charismatic taxa. For invertebrates, the figures are dramatic and even though they account for the biggest part of the species richness in the planet, knowledge is both scarce and scattered. New advances in technology (i.e., statistical methods and remote sensing) can contribute to improve this knowledge. Here we report our endeavours of assembling a multi-source database of distributional records for all 982 no-American taxa of the Sphingidae family of Lepidoptera, provide algorithm-based distribution maps, and study resulting patterns of biodiversity. We used Maxent, a popular technique of species distribution modelling (SDM), in combination with climatic and vegetation data, to estimate the distribution of the species at 5 x 5 km resolution across. We then superimposed resulting grids to study patterns of biodiversity at two different spatial scales (α-diversity: 5 x 5 km grid cells, and γ-diversity: 200 x 200 km grid cells). We also used these data to map β-diversity. We could model the distribution for 789 taxa, whereas we provided expert-based range estimates for the remaining 193 taxa. Annual temperature range emerged as the variable that contributes most to shape the distribution of species in models, closely followed by variables related to precipitation. Vegetation data did not contribute highly to our models. Our maps of α and γ diversity reveal the expected gradient towards the tropics of species richness. In contrast, beta diversity did not show a latitudinal gradient but a rather altitudinal one, with high β in mountainous regions and along main biogeographic boundaries. To the best of our knowledge this is the first distributional data set of a complete family of invertebrates at large (i.e., almost global) extent and fine resolution. There were many challenges inherent to assembling data, and we discuss which steps of work required particular attention: taxonomic and georeferencing errors and spatial bias in data. Our results will contribute to understand and move forward the study of insect biodiversity patterns at a macro-scale. We also hope that this study help and encourage others to embark on similar tasks.

Introduction

Current research questions in biogeography, macroecology and biodiversity research refer to topics such as the mechanisms that shape global biodiversity patterns, how species ranges will be affected by changing climate and landuse patterns, or how phylogeny, species traits and the environment interact to determine what species we find at a given site (Morrone 2009). However, a closer look at the published literature reveals that the majority of large-scale analyses have been carried out on a limited set of taxa (i.e., vertebrates and vascular plants) that do not represent the actual phylogenetic distribution of biodiversity (Beck et al. 2012a). Invertebrates represent the vast majority of taxa, and among the known species richness herbivorous insects make a sizable contribution (Godfray et al. 1999, Hamilton et al. 2010). Unfortunately, for most species of those groups we only have a vague idea of where they occur, while even for well-studied taxa (i.e., vertebrates) geographical distribution data is usually at coarse grain compared to other environmental variables (Jetz et al. 2012).

The lack of information on species' geographic ranges (the 'Wallacean shortfall'; Lomolino 2004) is only part of a larger data limitation problem. We are in the midst of a proclaimed "biodiversity crisis" (Wilson et al. 2003), yet we have seen and described only a limited part of the species diversity on the planet, i.e. an estimated of 23-33% of the total of multicellular species (Hamilton et al. 2010; the 'Linnean shortfall'). Only for a minority of these, usable knowledge is available on ecological traits, phylogeny, or biological interactions with other species and their relevance for ecosystem function (Wilson et al. 2003). Such data would be urgently needed to appreciate and understand the full spectrum of taxonomic and functional diversity in an ecological context, which in turn would be needed to conserve or manage the functioning of ecosystems. These topics are of utmost global societal importance (Diamond 2006), and the fact that we still know so little of life on earth (for many reasons that are beyond the scope of this paper) must be considered a severe limitation of current science.

The shortcomings listed above have been recognized and begun to be addressed by increasing attempts to utilize information technology approaches to make scatted distribution data more accessible and to facilitate synergistic collaborations to close these gaps in knowledge. Here, we focus on the Wallacean shortfall, but we recognize close linkage with the other data deficiencies and their prospective solutions (e.g., technological advances in taxonomy; Joppa et al. 2011, Bik et al. 2012, Deans et al. 2012, Maddison et al. 2012). Technology also plays a major role in addressing the knowledge gap for distributions, which is only partly a problem of limited incentive and finance for 'boots-on-the-ground' field research. Rather, to a large degree it is about storing, processing and

distributing information. Much data on species occurrence is already available in the form of natural history collections, however only digitizing and making them broadly available will fully facilitate their use in biodiversity research (see Jetz et al. 2012, Beck et al. 2012 for overview to current initiatives). At the same time, Ecological Niche Modelling or Species Distribution Modelling (SDM; Elith and Leathwick 2009) is a fast growing methodology that builds correlative models combining distributional records with (mostly) climatic and remotely sensed landcover data to infer suitable habitat, i.e. likely regions of occurrence for a taxon. Even though these methods make many unwarranted assumptions, have clear limitations and manifold options for wrong implementation and erroneous interpretation (Beale and Lennon 2012, Warren 2012) they nevertheless appear the best current approach to attain accurate, high-resolution and reproducible distribution estimates for many taxa.

We acknowledge and appreciate the opportunities for efficient workflow and intercontinental collaboration provided by the advances in online tools (e.g., Jetz et al. 2012), which seems a promising direction of addressing these issues. However, under current habits of scientific accreditation in biology and ecology, we are sceptic that this will be achieved by 'quantum contributions' (*sensu* Maddison et al. 2012) of a multitude of data contributors that are driven by a selfless urge to provide good data. Quality problems in current online databases (e.g., Yesson et al. 2007, Beck et al. subm Chapter 3 in this thesis) are probably to a large degree due to the fact that data providers are not data users.

In recent decades, there have been many endeavours on collecting, compiling and making available distributional data on different taxa for various purposes. Typically these data stem from highly non-random observations and surveys in space and time. A common output of analyzing such data is a set of distribution maps, ideally also available a digital form. Data vary enormously in three aspects: (1) Data type (i.e., presence-absence data per grid cells, based on surveys; model predictions of occurrence; or expert-drawn range maps), (2) resolution (grain size) and (3) extent.

Currently, reasonably reliable presence-absence data exist only for conspicuous, charismatic and well-studied taxa in well-searched regions (Gibbons et al 1993). Additionally, for some taxonomic groups (i.e., birds, mammals, amphibians and some selected plant groups [add references or URLs]) expert drawn maps are available. Implicitly, such expert-drawn maps typically have spatial resolutions between 100-200 km (Jetz et al. 2012), with a tendency for more accuracy in the temperate zones than in the tropics (Hurlbert and Jetz 2007)

Among the datasets available, there is for plants based on local inventories resulting in a map at 1° resolution and is published in (Kreft and Jetz 2007), and a global bird distribution at 1° resolution

was used in Jetz and Rahbek (2002). At the same resolution but at the continental scale (i.e. Africa) there are databases of distributional data on mammals, amphibians and reptiles held by the Zoological Museum at the University of Copenhagen (Galster et al. 2007, Hansen et al. 2007, Rasmussen et al. 2007). Additionally, at continental and country extents, but a much coarser resolution (10 km and 50 km UTM grid respectively), maps of distribution for amphibians and reptiles are provided in the atlases by (Godinho et al. 1999) and (Gasc et al. 1997).

For invertebrates there is even less comprehensive distribution data available in terms of taxonomic scope, extent and resolution. Asher et al. (2001) assembled distribution maps at 10 x 10 km grid cells for all butterflies of Britain and Ireland, Settele et al. 2008, compiled the data from the project "Mapping European Butterflies Project" (MEB: www.european-butterflies.eu) providing maps at a resolution of 50 x 50 km resolution, Similarly, Scott (1986) provides expert-range maps all for butterflies (and skippers) in North America (North of Mexico, Canada, Alaska Greenland, Iceland, Bermuda and Hawaii).

In this paper, we first report in detail of our endeavour to provide detailed distribution and biodiversity data for all non-American (i.e., Old World + Australia/Pacific) members of the Sphingidae, a family of the Lepidoptera. Sphingid moths are a suitable model taxon for macroecological studies on herbivorous insects (see below), and the dataset that we document here is the first of its kind (i.e., high resolution, almost global extent distribution data for a higher insect taxon). We describe the source data compilation (i.e., specimen records) and the generation of distribution estimates, and some baseline properties of input and output data. Our experiences gained throughout this work, including errors, hindsight and quantitative data descriptions, may be useful for those who attempt similar tasks in other taxonomic groups, and we hope this will encourage more researchers to follow through such a task. We will, among other topics, discuss the delicate balance between 'objective', algorithm-based SDM and "expert knowledge", between speed and accuracy in processing data, and the problems of obtaining new specimen distribution data from places where most diversity is to be found, i.e. tropical countries. In the second part of this paper, we present maps for sphingid diversity based upon stacking species-specific distribution predictions. In particular, we present species richness at two different scales (5 x 5 km and 200 x 200 km grid cells) and define these as α-, respectively γ-diversity. From those data, we also calculate and map patterns of β-diversity.

Lepidoptera, family Sphingidae

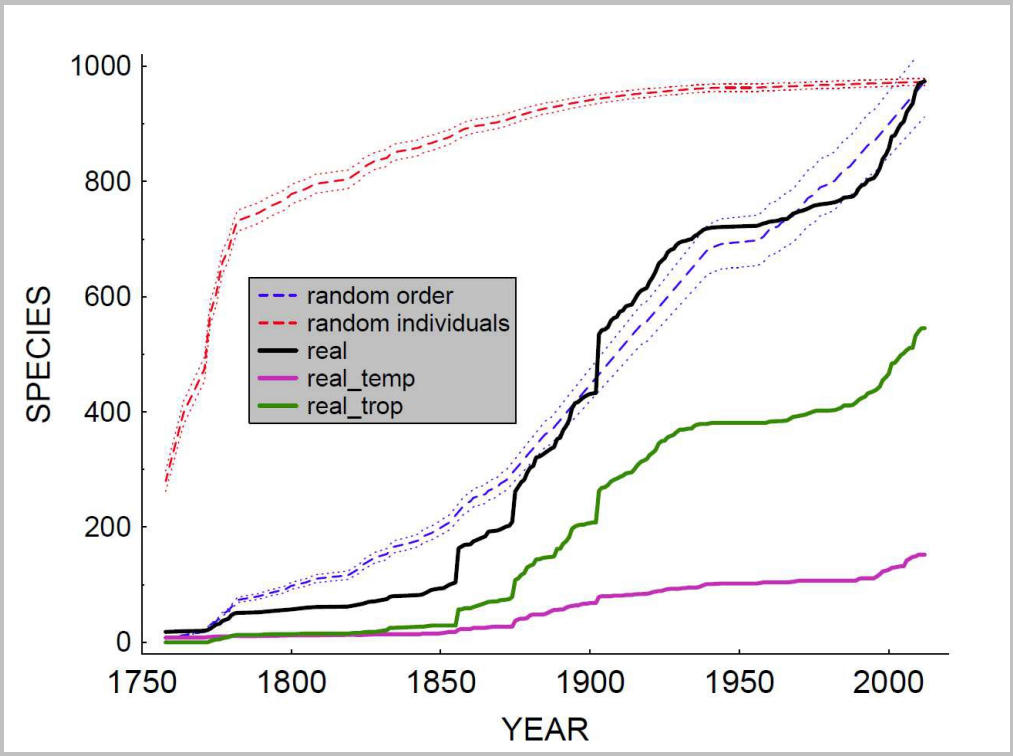
Sphingidae or 'hawkmoths' are relatively large, fast-flying and sometimes extremely dispersive members in the bombicoid clade of Macrolepidoptera (Kitching and Cadiou 2000, Regier et al.

2009, Mutanen et al. 2010). Most adults are nocturnal, but some day-flying genera occur (*Macroglossum, Hemaris, Sataspes*). Caterpillars ('hornworms') are folivorous with a moderate to low degree of hostplant specialization (i.e., specialization below plant family is rare; Kitching and Cadiou 2000, Robinson et al. 2001, Beck et al. 2006a). Adults are mostly nectarivorous (with some exceptions, such as honey-feeding *Acherontia*) or do not feed at all (an ancestral trait in the bombicoid clade), and this distinction has been shown to be related to phylogeny, manifold life history traits (e.g. morphology, sexual dimorphism), habitat preference and distribution (Janzen 1984, Holloway 1987, Beck et al. 2006b, Beck and Kitching 2007, Kawahara et al. 2009). Globally ca. 1470 species are known (Kitching and Cadiou 2000 and recent updates), whereas outside the Americas (i.e., our study area) 982 autochtonous taxa are recognized in this study (see below for taxonomic treatment). There is almost no overlap with the Americas (only one species, *Hyles galli*, is autochtonous to both regions), which made this geographic split feasible.

Sphingids are an attractive group to both amateur collectors and taxonomists, and in consequence more is known about their distribution, biology (e.g., hostplant associations) and taxonomy than for most other invertebrate groups, with the exception of diurnal butterflies (Papilionoidea) for some regions. Despite this, however, large gaps in knowledge exist particularly for tropical taxa, and the accumulation pattern of known species (Box 5.1) indicates that substantial numbers of yet unknown taxa may exist.

Box 5.1 Accumulation of taxonomic richness with time

Solid lines in the figure show the accumulation of known species of Sphingidae, derived from the year of description, for 973 described species in our study region (Old World; see main text for taxonomic treatment, candidate species were not considered). Note the steep increase in known numbers around the year 1900 due to the efforts of Rothschild and Jordan (1903). The increase in recent years is probably caused by a combination of exploration efforts in poorly-sampled regions (tropics, China) as well as the utilization of integrative taxonomy (i.e. considering molecular data such DNA barcoding), which lead to the recognition of cryptic species diversity.



Broken lines represent averaged randomizations of species accumulation (± 95% confidence intervals) based on the frequencies of species encountered in our database (ca. 109'000 occurrence records, carried out with EstimateS 8.0; Colwell, 2005).

The blue line describes what would be expected, on average, if the order of yearly additions of species due to new descriptions was random, i.e. if there were no era-specific highs and lows in taxonomic activity. One would expect a linear increase, but because there were years without any new species we plot the resulting, slightly irregular pattern to allow direct comparison with the real accumulation (black solid line). The relevant information from this particular randomization stems from comparing confidence intervals with the real data curve (black line). This allows judging if and where real species description rates were outside of the expected random variability derived from our data set, and hence require further historical explanation (e.g., low rate of new description from ca. 1780 to 1880).

The broken red line is a species accumulation curve where all specimens had the same chance to be picked and described at any given year (*individual shuffling* in EstimateS). This mimics, i.e., that the entire collection of specimens was available, and taxonomists would pick random individuals each year and describe them if they were new, without any biases due to geography (i.e., occurrence in inaccessible regions) or systematic preference. At the beginning there will be a steep increase of species description because all the species would be relatively new, but with years passing by the curve levels off because the chances to pick up a specimen of a common species already described would be higher. Obviously, the expected pattern is very different from the real one, indicating that *(a)* strong biases prevailed in real species accumulation (i.e., taxonomic description patterns), and *(b)* efforts of estimating numerically the expected total of species from description rates (Scoble et al. 1995, Costello and Wilson 2011) is complicated by violation of the common assumption of random draws in species accumulation in such extrapolation methods.

Methods

Raw data compilation and processing

Taxonomy and nomenclature

For the majority of species, we followed the nomenclature given of Kitching and Cadiou (2000) and more recent taxonomic publications. However, taxonomic findings can have a considerable time-lag until reaching official status according to the *International Code of Zoological Nomenclature* (ICZN 1999), so we allowed deviations for the purposes of this data compilation. In particular, we did not consider some recent descriptions where we were quite sure that they were erroneous (although they are not yet refuted in publication), whereas we accepted some recent splits and revisions based on compelling evidence even if not yet published (including some 'in litteris species'). We adjusted all nomenclature to this system, but in some cases (e.g. *Hippotion boerhaviae*-complex) we did not consider distribution records that could not clearly be associated with a currently valid taxon (i.e., specimens inaccessible, no pictures or other backup data). For higher-taxon associations we followed the molecular phylogeny of Kawahara et al. (2009), which confirmed the monophyly of major traditional systematics units (i.e., subfamilies, tribes) although rearranging their topology.

Distribution records

We extracted distribution records by screening all the published literature and checklists (ca. 1664 references) (see EA1 submitted with the thesis), and we carried out own field sampling in parts of Europe, Africa and Southeast-Asia. Earlier projects have compiled data on geographic distribution records of sphingids for larger regions and presented them online (i.e., Pittaway 1997-2012, Pittaway and Kitching 2000-2012, Beck and Kitching 2004-2008); these data have been fully integrated here. We also downloaded distribution records from online data bases such as the Global Biodiversity Information Facility (GBIF; www.gbif.org, Nov. 2009) and Barcode of Life Database (BOLD; www.barcodinglife.org, Aug. 2010).

Furthermore, we visited natural history collections (private and public) on four continents and extracted specimen-label data (among them, e.g., complete collections of the Natural History Museum, London; Museum National d'Histoire naturelle, Paris; Royal Museum for Central Africa, Tervuren; see Appendix 5.1). Additionally, we received data from colleagues (amateur collectors, by-catches in other entomological projects, etc.) and from picture-based identification request (mostly to IJK). We also actively searched the internet (Lepidoptera blogs, specimen trading sites)

for data of interest. Generally, to avoid errors due to misidentifications, difficult taxa were checked by IJK (either on the specimen or by photograph) unless data stemmed from a renowned expert on sphingid identification.

We ignored records that could, with high likelihood, be considered erroneous (locality or taxon). Furthermore, we ignored records that were known to be single vagrant specimens or where we could reasonably infer that single specimens were transported by human traffic far out of their autochthonous range. We also excluded two New World species that recently have established populations in our study region (*Darapsa myron* in the Bangkok area, and *Agrius cingulata* in West-Africa; cf. Ballesteros-Mejia et al. 2011; chapter 6 in this thesis).

A number of relatively common and highly dispersive taxa are known to establish summer populations but not surviving the cold season (see Beck et al. subm., Chapter 3 in this thesis). However, details on boundaries between permanent (i.e., overwintering) and non-permanent populations are only known for Europe. Therefore, we used these data to assign coldest-month isotherms as northern boundaries for these species and applied these thresholds across their entire range (e.g., East Asia) in order to model permanent ranges. This approach fitted with migrant ranges provided in Beck and Kitching (2004-2008). Thus, all models refer to permanent ranges. Local migrations may occur in further, non-European taxa as well as in some arid regions, but insufficient data prevented us from considering these.

Georeferencing

Based on locality data associated to specimen labels, we assigned geographic coordinates to distribution records. If not given from GPS measurement, we found coordinates of localities in online gazettes (e.g., http://www.fallingrain.com/world/), GoogleEarth, the Times Atlas of the World (2010) and local maps. For difficult localities (e.g., old records with changing names, small places, transliterations from non-Latin spellings), we also made use of historical atlases (including http://worldmap.harvard.edu/africamap/) and travel itinerates of the collectors in questions. Hints towards localities were often found from broad internet searches, such as, traveller blogs or sites on Christian missions or military history (e.g. US navy battles in the West-Pacific). Uncertainties arose particularly from creative transliterations of Chinese localities, from incomplete data (e.g. place name but no information on province or country) and from very common place names. In the latter case, we tried to guess the most likely locality based on other records for the species and the 'home range' or travel route of the collector.

We generally aimed at georeferencing at a resolution of 0.01° (ca. 1 km) or higher. However, sometimes this was not feasible either because we could not find sites precisely enough, or because

no detailed locality was given in original sources. We coded estimated spatial precision of records in four classes (i.e., '0.01° or better'; '0.1° or better'; '1° or better'; 'unspecific' (e.g., "Southern India"); Wieczorek et al. 2004) to facilitate filtering our data depending on the resolution required for specific analyses. If data could not be localized precisely but altitude information was given, we set coordinates to a region of similar altitude (using GoogleEarth, which incorporates a 90 m-resolution digital elevation model) to minimize environmental deviation from the true site. We excluded *Hippotion leucocephalus* from our dataset (known only from the holotype, locality data: "*Africa, ?*").

Georeferencing was the most work-intensive and error-prone step of work, and we spent great lengths to identify errors by mapping data, checking consistency between records processed by different people (see Acknowledgements), etc. We also subjected data already containing coordinates to this procedure (e.g., collectors' GPS-data or downloads from GBIF), and found quite a lot of errors (often probably due to mistyping). We prioritized efforts of databasing and georeferencing towards regions and taxa with relatively few records, hereby adding more information per man-hour (Beck et al. subm., Chapter 3 in this thesis).

Distribution modelling

We based all SDMs on climatic data provided by WorldClim (www.worldclim.org; 30-50 year averages) as well as information on vegetation cover from remote sensing (MODIS continuous fields, based on 2000-2001 data: percentage of tree, herb and bare ground per pixels; http://modis.gsfc.nasa.gov/). We excluded climatic variables that seemed highly redundant or insignificant in the light of our knowledge of sphingid ecology. Specifically, we used the following, continuous-data variables: *Altitude, annual temperature range, annual precipitation, annual temperature, mean temperature of the coldest quarter, mean temperature of the driest quarter, mean temperature of the warmest quarter, mean temperature of the wettest quarter, precipitation of the coldest quarter, precipitation of the driest quarter, precipitation of the warmest quarter, precipitation of the wettest quarter, precipitation seasonality, tree cover, herb cover* and *bare ground cover*. All modelling was carried out on a spatial resolution of 2.5 arcminutes (≈ 5 km).

For species with a large number of available records we used only data that were georeferenced with high precision (e.g. when more than 50 spatially independent records (at 5 x 5 km grid cell resolution) where available for the species, all records with a precision ≥0.1° were excluded), whereas we included records up to a resolution of 1° for data-deficient species. Unspecific records were not used for modelling, but we considered them when editing for dispersal barriers (see below).

For species known from less than five spatially independent localities we supplemented SDM data (if models could be run at all) with tentative, expert-drawn range estimates. Similarly, we supplemented final range areas with expert assessments of likely occurrence if models did not converge or in regions without environmental data (i.e., small Pacific islands, see EA2).

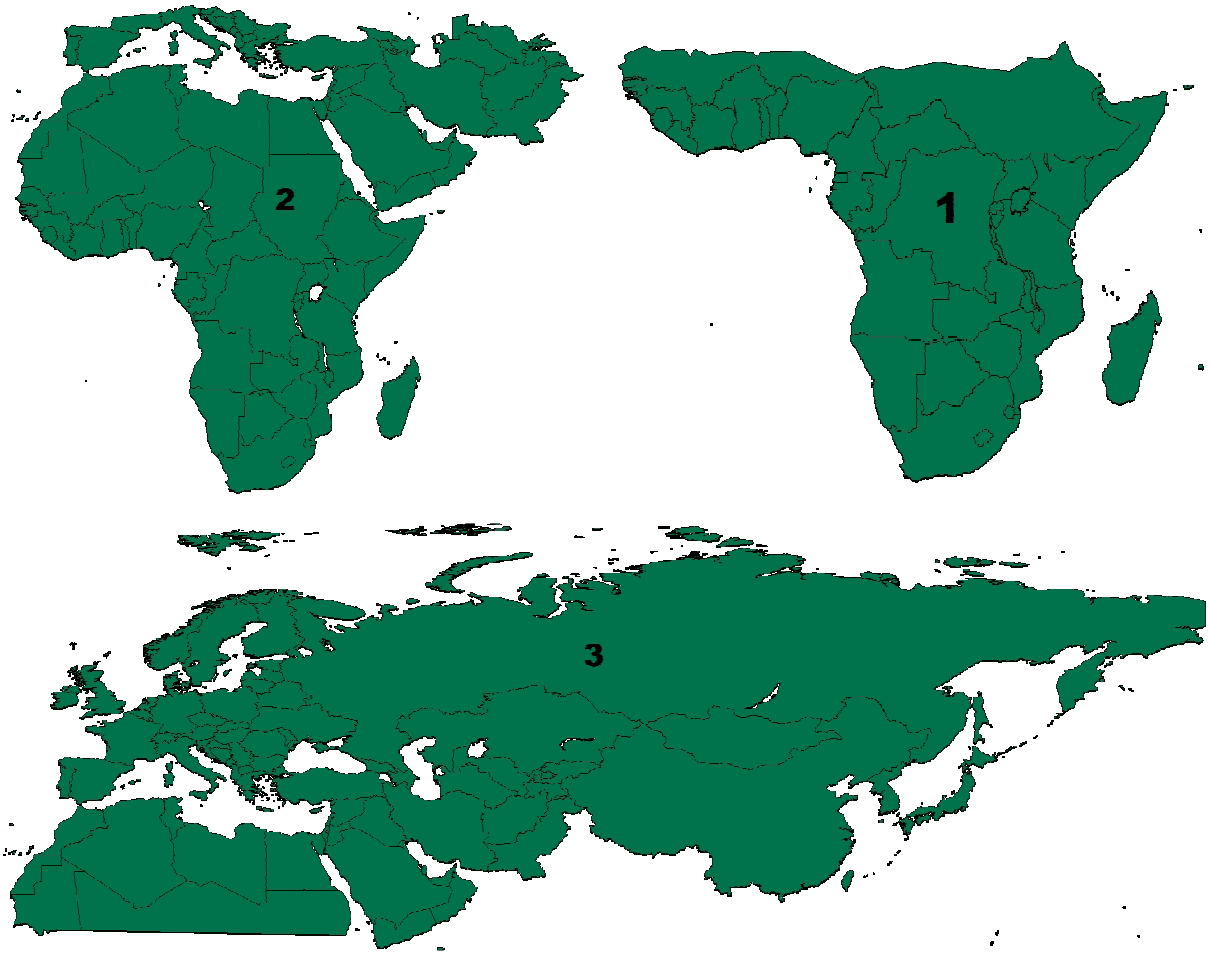
There is a large number of SDM methods available. We used a stratified-random selection of test species to evaluate their performance with our data according to three different evaluation criteria (Ballesteros–Mejia et al. subm.; Chapter 2 in this thesis, for details). We concluded that maximum entropy modelling (Maxent; Phillips et al. 2006) and random forest (RF; Breiman 2001) were the best-performing methods (see also Elith et al. 2006), and we used only those for final modelling of all species. Our study advised against model averaging approaches (Araujo and New 2004, Thuillier et al. 2009), and we did not find obvious links of model performance with species- or data-set characteristics. As a consequence, we separately modelled and processed data for these two methods. However, due to time constraints results presented here are only based on the Maxent modelled maps. Raw outputs of RF models are deposited with this thesis for later processing.

All modelling of 'presence-only' specimen records require the creation of data to compare with, i.e. pseudo-absences (assumed absence; Ferrier 2002) or background sample (environment sample across the landscape (Phillips et al. 2009)). Different methods exist to create these data, and we followed recent advice in the literature on best practices (Mateo et al. 2010). For Maxent, we used 10'000 background points chosen according to a bias file produced from kernel densities of raw records (100 km search radius; Phillips et al. 2009), otherwise default software settings and logistic output. For RF, we chose 10'000 pseudo-absences from outside a 40 km-buffer around known records (VanDerWal et al. 2009). For both methods, we a priori restricted the modelling region according to the known biogeography of species (e.g., excluding Europe and Asia when modelling a species restricted to sub-saharan Africa). For that purpose, we divided the study region into seven sub-regions (trying to follow established biogeographical regions with some slight variations; see Figure. 5.1 for a map): 1) Sub-saharan Africa, 2) Africa +Arab peninsula, 3) Palaearctic, 4) Eastern Palaearctic + Oriental region, 5) Oriental region + Australia, 6) Australia + Western Pacific islands and 7) Australia. However, if in doubt on the potential spread of species we rather modelled larger regions.

We followed standard procedures of model evaluation by randomly splitting the available data and using 75% for model fitting or "training" and the remaining 25% for testing. We used a crossvalidation procedure to retrieve the area under the receiver-operating characteristic (AUC; Hanley and McNeil 1982) based on five replicate model runs, and we used averages from five runs as model predictions. Additionally, all models were evaluated by us for plausibility (see below). We tried to find and fix sources of error for obviously bad models (input data problems, among them

niche misspecifications due to large quantities of spatially biased records from GBIF; Böller, 2012). We excluded some species from modelling and provide expert-based range estimates where this did not lead to improvement (see EA3, submitted with this thesis for list).

Figure 5.1. Different regions used as a basis to restrict the areas for SDM. They are based on recognized biogeographical regions incorporating some slight modifications. Region 1) Subsaharan Africa, 2) Africa plus Arab peninsula , 3) Paleartic , 4)Oriental Region , 5) Oriental Region plus Australia , 6) Australia plus Pacific Islands ,7)Australia.



Range estimates: Post-editing, thresholding, and expert estimates

Current standard SDM methods cannot, by design, account for dispersal limitation when estimating distributions (although progress has been made, see Glor and Warren 2011). Rather, output is a measure of habitat suitability, irrespectively of whether the species has reached a region or not. We applied expert-opinion post-editing of modelled distributions based on known biogeographic barriers, such as separations of zoogeographical regions known for sphingids (Beck et al. 2006d) or for other taxa (e.g., Wallace 1869, Kreft and Jetz 2010, Linder et al. 2012), sea, deserts or mountain ranges, as well as large gaps in suitable habitat inferred from niche model output. We assumed that species did not cross such potential dispersal barriers unless we found positive

evidence (i.e., records). These edits were reviewed by LBM, JB and IJK, adjusted were necessary, and then implemented by clipping the extent of model predictions to the required area.

For transforming continuous suitability into binary presence-absence predictions we used the minimum predicted area rule (Engler et al. 2004), i.e. setting a threshold so that at least 90% of recorded presences are predicted correctly. We re-projected all binary predictions into Mollweide equal area projection (5 x 5 km resolution) to facilitate measurement of range area and further analysis.

Thus, for all species with SDM-based range maps we have three stages of model outputs for further analyses: raw SDM output (i.e., continuous mapping of suitability from 0 to 1), raw SDM output expert-edited for dispersal limitation, edited SDM thresholded output (providing binary presence-absence prediction). Apart from raw output which we have for Maxent and RF, edited outputs are currently only available for Maxent.

For some species we could not provide SDMs (too few data, poor models, regions without environmental data). In these cases we created 'expert-opinion' range estimates by plotting records on maps of altitude, temperature, precipitation and tree cover and drawing estimated extents of occurrence (as provided e.g. in Beck and Kitching 2004-2008 for Southeast-Asian taxa). Following suggestions by Hurlbert and Jetz (2007) we intersected these with our assessment of habitat restrictions and converted them to a grid of the same resolution as SDM-based maps (e.g., for a montante species all lowland cells were cut out; specific rules for each species were documented).

A database allows selecting and comparing these outputs according to species properties, dataset and SDM criteria, among them AUC (as tentative measure of model quality), sample size of available records, higher-taxon association, region, and comments added during editing. Depositing these various stages will facilitate further amendments to the data, e.g. if relevant new records for a species have been found.

Mapping and analysing biodiversity

To provide a first appraisal of biodiversity patterns in a scale-dependent manner, we map species richness at local scale (α, defined as 5 x 5 km grid cells) and at regional scale (γ, defined as 200 x 200 km grid cells; cf. Ballesteros-Mejia et al *in press*; Chapter 4 in this thesis). We used these data to map the regional heterogeneity of communities (i.e., β), applying the multiplicative concept of β-diversity ($\beta = \gamma/\text{average } \alpha$; Whittaker 1960, Tuomisto 2010). We carried out calculations and analyses of β-diversity only for large islands and the continental part of our region, (i.e. excluding small Pacific islands) and also excluding from the alpha diversity map those cells with zero-values.

Software

We stored and processed distribution records and taxonomic data in MS Access. Mapping and other geoprocessing, as well as editing and further analysis of model output was carried out in ArcGIS 9.3 and 10, Geospatial Modelling Environment (http://www.spatialecology.com/gme/) and R. SDM was carried out in Maxent software version 3.3.3e (Phillips et al. 2006), while we used BIOMOD (a platform for R; Thuiller et al. 2009) for RF models. Computing speed is an issue when working with many species in high resolution over large extents. We run trial-models in R on a Linux system, whereas we use a computer cluster at the University Computing Centre for final BIOMOD runs for all species. Java-based Maxent software cannot easily be sped up, instead we resorted to nightly parallel runs on many different computers (i.e., our colleagues' machines), controlled through *remote desktop* function. Other statistical analyses were carried out in R. We used a data backup system of external hard drives as well as memory storage at the University Computing Centre to store data. We also used the online system *Google Drive* for collaborative editing.

Results

Raw data properties

In total, we had 109'880 records available for the analysis of 982 species. A record indicates a unique combination of locality, year and collector or source, but it may contain one or many specimens. Not included in this figure are records that we excluded a priori: 233 records for being considered 'confirmed errors', vagrants or human-transported specimens; 571 records where insufficient data on species or locality was given, and ca. 4'000 records considered as 'low priority' for our georeferencing efforts (e.g., common taxa, well-sampled regions, unspecific localities).

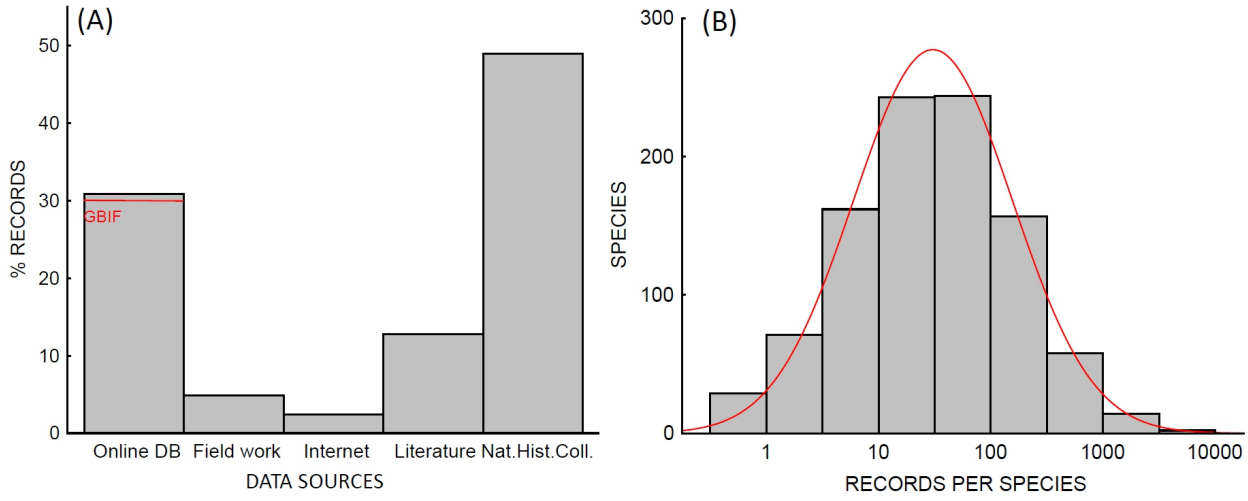
Of the records used, 79.6 % were georeferenced with a precision ≤0.1° latitude/longitude (ca. 11 km), whereas 2.6 % were considered highly unspecific. 69.4 % of records contained information on the year of collection, and most of these (i.e., 83.5 %) stemmed from after 1950 (despite spanning a 200 year range, from 1811 to 2011). Only 2.6 % of records were from pre-1900, and we observed a clear decline in collecting activity during 2nd world war, and strong increase afterwards. Interestingly, there was a peak in collecting during the 1990's (48.0% of records are from 1989 or younger), and a decline in the 2000's. 31.8 % or records had associated altitude data, which suggested that sampling activity was quite proportional to available land area up to 5400 m a.s.l. (i.e., approximately linear decline of records with altitude on a log-scale plot; not shown). Almost 50% of data stemmed from data-basing specimen labels in natural history collections, whereas online databases (of which GBIF made up by far the largest portion) made the second-largest

contribution with >30%. However, Beck et al. (2013; Chapter 3 in this thesis) showed that GBIF data were often less informative with regard to species' niches and their geographic distributions than collections data (even for European species). The number of records per species followed closely a lognormal distribution, indicating that in the dataset, despite its substantial size, many rare and poorly-known taxa persist (Figure 5.2).

Data availability was highly uneven geographically. Ballesteros-Mejia et al. (2013; Chapter 4 in this thesis) have analysed this in detail for sub-Saharan Africa, reporting a very close correlation of available records and observed species richness and identifying historical and human-geographic determinants of these patterns.

Notably, 39.2 % of records represent spatial replicates (i.e., the species was already known from the same locality yet from a different year or a different collector or source). When considered on the grid cell resolution as used for modelling (i.e., ca. 5 x 5 km), the number of spatially unique records shrank to 49'418 (i.e., ca. 45% of original records). Replicates can be valuable for confirmation or for numerical techniques of species richness estimation (Ballesteros-Mejia et al. 2013; Chapter 4 in this thesis) but they do not contribute to SDM.

Figure 5.2. (**A**) Contribution of data sources to available records (in red contribution of the Global Biodiversity Information Facility, GBIF), "Literature" refers to scientific publications (paper and books), whether in papers or online while "internet" denotes informal online sources such as blogs, reports, etc. (**B**) Distribution of records over species. A log normal distribution was fitted.



SDM output: Model quality and predictor contributions

We could compute and post-edit SDM range estimates for 789 species, whereas we provide expert range estimates for the remaining 193 species. AUC values of test data for Maxent SDMs were median = 0.94 (25-75 percentiles = 0.88-0.973; minimum= 0.3598; maximum = 0.997). From

models, 90% retrieved AUC values > 0.8 (Figure 5.3), which represent good or excellent models (Swets 1988), however AUC is also affected by features such as modelling extent, making cross-species comparisons of model quality difficult (Beale and Lennon 2010). Models that we considered highly unrealistic were not included in this data set (see Methods).

We found that across all species the variable that contributed most to models was *annual temperature range*, although variability between species was large (Figure 5.4). It is followed by *precipitation of the warmest, driest* and *coldest quarters*. Vegetation cover data came to contribute only in 6th, 7th and 11th place.

A closer look on the contribution of variables, however, reveals some differences between tribes. *Annual temperature range* is retained as the most prevalent contributor to the models for most tribes (Acherontini, Ambulycini, Macroglossini, Sphingulini), while for Dilophonotini and Sphinguini the variable that contributes most was *Precipitation in the warmest quarter* (10% and 6.17%, respectively). The importance of vegetation cover also varies from tribe to tribe. *Bare ground cover* appeared to be important in the models of Acherontini and Ambulycini, whereas *tree cover* appeared to be important for the species-rich tribes Smerinthini and Macroglossini (3rd and 4th place of importance with 3.48% and 2.41%, respectively; Figure 5.5(E) and 5.5(F)). Field data from tropical Southeast Asia suggested that Smerinthini tend to be forest adapted whereas members of the Macroglossini tribe are more adapted to open, disturbed landscapes (Beck et al. 2006a, 2006b, 2006c; Beck and Nässig 2007). Supporting this, response curves for several Smerinthini show a positive link between suitability and percentage of tree cover. In contrast, response curves for Macroglossini models show negative links with tree cover (see Appendix 5.2 for some exemplary curves).

Figure 5.3. Histogram of AUC values (Area under the receiver-operating characteristic) for the test data (i.e. distributional records not used for model fitting)

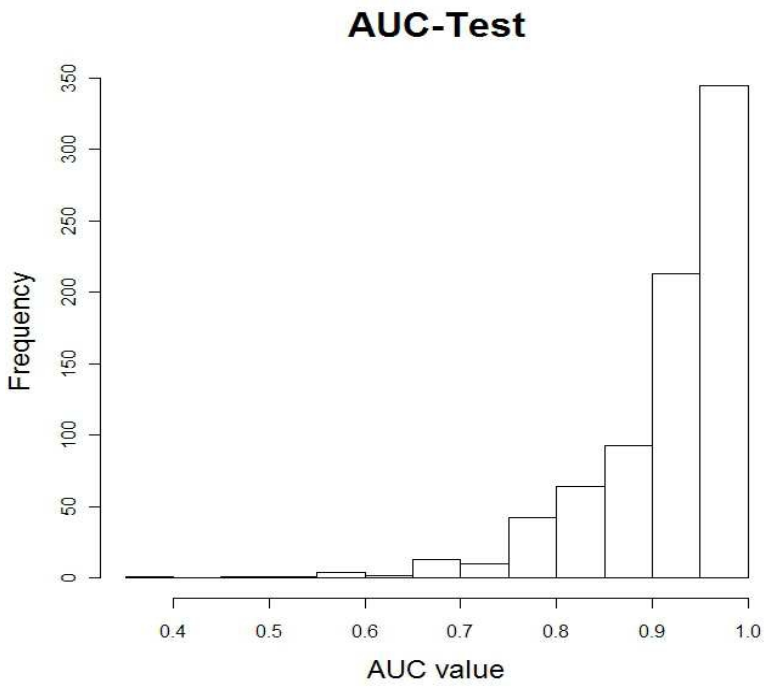


Figure 5.4. Boxplot (median, quartiles, range) of the variation in variable contribution to the model across all the species (N=789)

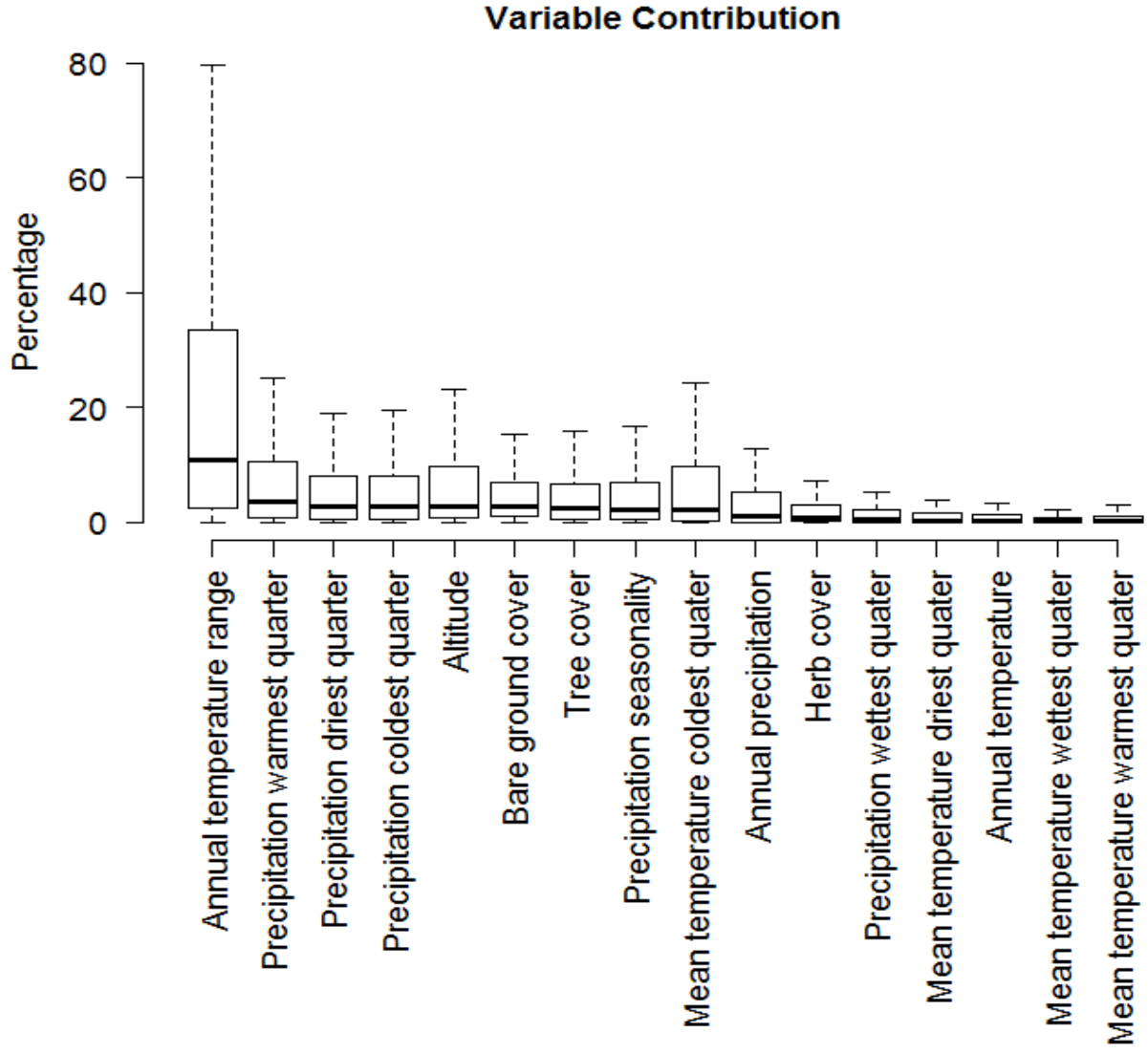
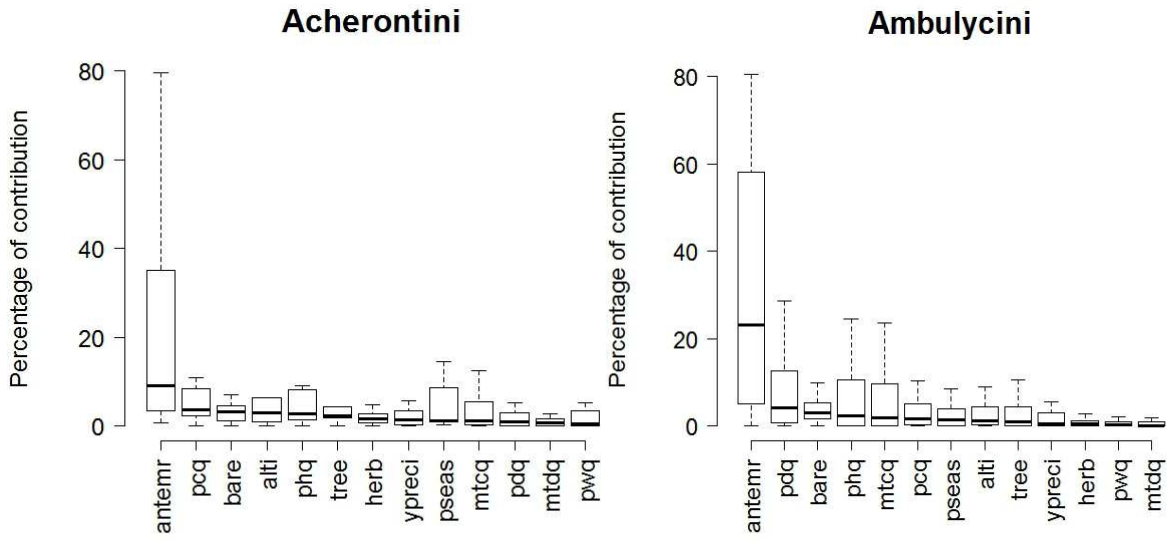
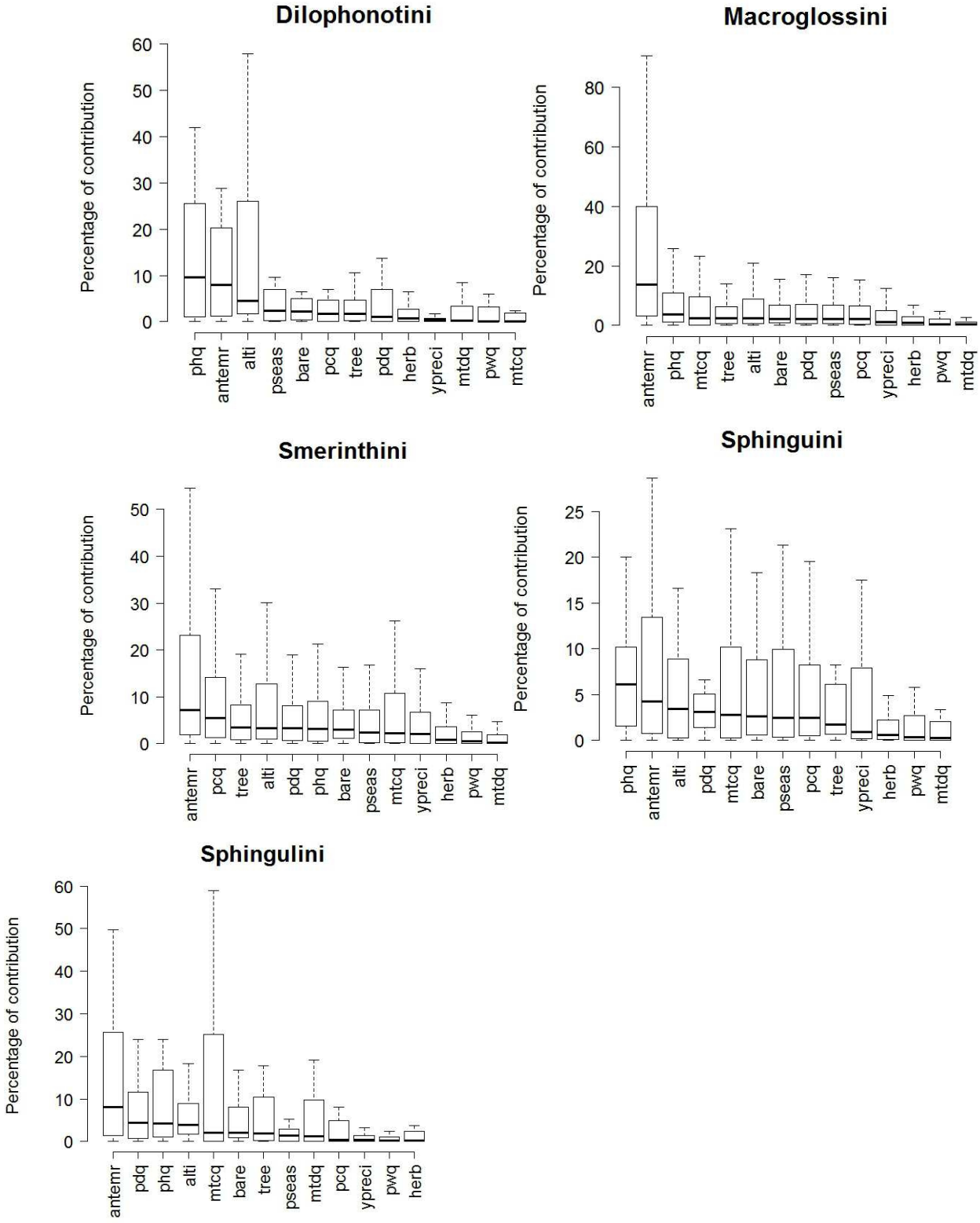


Figure 5.5. Boxplots of the variable contribution to the models showed by tribe.





Alpha, Beta and Gamma diversity

The map of estimated alpha diversity sphingid moths (5 x 5 km cells; Figure 5.6A) reveals areas with more that 155 species per cell. There is a strong latitudinal gradient, with species richness increasing towards the tropics. These data highlight areas known for harbouring high species richness also in other taxa, such as Western African forest, Eastern Arc Mountains in

Tanzania/Kenya, the Indo-Burmese region, South-Central China and Sundaland, with estimated species richness exceeding 100 species per 5 x 5 km grid cell. Additionally, it warns about the areas with high species richness not so conspicuous: Other areas of high species richness are the Albertine Rift in Eastern Africa, the north-western part of Tanzania and Congo Basin. The mountain region at the border between Mozambique and Tanzania, and the most western part of the Himalaya belt stretching up until almost to the Hindu Kush still feature species richness greater than 80 species per 5 x 5 km.

Large-scale diversity (i.e. gamma) in the study region is presented in Figure 5.7A, predicting species richness of up to 186 species per 200 x 200 km cell. In broad pattern, it highlights the same areas of high species richness as the previous map, yet extends them to include Malawi, the northern part of Zambia, Mozambique, all of Thailand, and the Philippines with species richness exceeding 90 species per 200 x 200 km grid cell. At both resolutions maps reveal areas of intermediate species richness (40-80 species per grid cell) in Angola, South Africa, the Horn of Africa, Madagascar, India, Sri Lanka, the Caucasus, and all Southeast Asia east to New Caledonia, and the humid-tropical part of Australia. In general, we see a very similar pattern across the region in both Alpha and Gamma diversities, which is confirmed by the positive correlation between them (linear correlation, 3003 cells, **α** ave_200km and **γ** r= 0.9524795)

Patterns of alpha and gamma diversity vary from tribe to tribe. Macroglosssini and Smerithini are the two tribes contributing the most to the species richness pattern of the family, with places with species richness up to 87 and 47 for α-diversity) and 103 and 60 (for γ-diversity) respectively (Figures 5.6E and 5.6F) and 5.7E and 5.7F)). Four out of seven tribes (Acherontini, Macroglossini, Smerinthini and Sphingini; Figures 5.6 and 5.7) exhibit centres of high species richness in both tropics (i.e. African tropics and Oriental tropics), Ambulycini's richness pattern is concentrated in the Oriental tropics, whereas Dilophonotini's pattern is wider across the region (i.e. places with high species richness all over Paleartic region, Philippines, and Northern part of Australia).

Variation of species composition within 200 x 200 km grid cells (i.e., beta diversity) is shown in Figure 5.8. It is measured independently of species richness (see Methods). Hardly any study allows comparing values of beta diversity, following the same methodology and definitions, across a larger, e.g. latitudinal, gradient, hence complicating an assessment of what areas have "very high" beta diversity (Beck et al. 2012b). A detailed analysis of the data presented here is pending, but tentative observations suggest that many of the high beta-diversity areas are located in mountainous landscapes. This has been noticed in other studies (McKnight et al. 2007, Ruggiero and Hawkins 2008, Jankowski et al. 2009) and for insects in particular (Davis et al. 1999, Brehm et al. 2003). It

also highlights regions that may be biogeographic transition zones due to historical or ecological effects, i.e. (1) between sub-saharan Africa and Europe, (2) between the western and the eastern Palaeartic, along the Reignie line (De Latin 1967) through Tibet to the Indus valley. What seems conspicuously absent is a clear latitudinal gradient.

Discussion

Addressing the shortfall

Species' distribution is often one of the key variables in macroecological studies (Gaston, 2003) and therefore the Wallacean shortfall is a drawback when analysing broad-scale patterns of diversity. This shortfall is present at almost all groups of organisms, but it is even stronger in insects despite of their abundance, richness and ecological importance (i.e. as functional group within the ecosystem; pollinators; Diniz-Filho et al. 2010). Their study is critical to understand ecological and evolutionary processes that drive diversity in terrestrial ecosystems around the world (Thomas et al. 2008).

Here we undertook the task of contributing to its closure. To the best of our knowledge we have compiled the most complete database on occurrence localities for a higher invertebrate taxon at almost-global scale, which we combined with SDM techniques to produce distributional maps for the all species in our research region. Despite the fact that many species are known from single or few localities or even single individuals (Figure. 5.2), for the majority of the species we had enough data to apply SDMs. We retrieved an AUC value >0.8 (Figure 5.3), which indicates good performance of the model (Swets 1988), for 90% of the models.

Environmental effects on species distribution

Across all species, annual temperature range was usually the most important environmental factor to predict the distribution of sphingid moths, followed by precipitation in the warmest, driest and coldest quarters (Figure 5.4). Thus, temperature-related variables seem to play a major role in shaping distributions, although causality can not be inferred from correlative models. Nevertheless, similar patterns have been noted before (Turner et al. 1987, Ballesteros-Mejia et al. 2011, Chapter 6 in this thesis) and can be attributed to direct effects of temperature on moth physiology, or to indirect effects of climate on vegetation structure and plant diversity. Notably, data on vegetation structure (MODIS layers) did not contribute a lot to explain distributions, possibly because much of their variation is explained by climate themselves. We did not have data on taxonomic composition of plants, which may play a role at least for the more host-specific feeders. Many other factors

could theoretically play a role in determining the distribution of the species (e.g., habitat disturbance, competition, predations, meta-population dynamics; Beale and Lennon 2012) yet data on such variables are not available at the scale and extent treated here. However, many other studies found that species distributions of large scales and large extents are mainly shaped by climatic variables (Buckley and Jetz 2007, Hawkins et al. 2008, Field et al. 2009, Keil and Hawkins 2009). Other studies show the same predictors found here as important also drive species richness in other taxa (i.e., annual temperature, precipitation and altitude; Buckley and Jetz 2007, Terribile et al. 2009).

This discussion is elucidated by considering that species distribution models (SDMs) are based on assessments of environmental niches, which have been defined in at least two different concepts. Grinnellian niches define a set of environmental conditions within which a species can survive and reproduce (Grinell, 1917). Eltonian niches define the place or role of the species within the ecological community (Elton, 1927). SDM's are based on Grinnellian niches (Soberon, 2010). Within the Grinellian niche concept, two aspects can be recognized. The *fundamental niche* defines the set of conditions where a species can survive and reproduce physiologically, while the *realized niche* defines the conditions where species can survive and reproduce in the presence of other members of a community, most importantly of their competitors (Hutchinson, 1957). In recent years there has been a debate about whether SDMs model fundamental or realized niches. Several authors argued that SDMs aim to identify the realize niche of the species even without including biotic interactions variables, because they use actual (i.e., realized) distributional data to build the model (Guisan and Zimmermann 2000, Austin 2002, Pearson and Dawson 2003). However, Soberon and Peterson (2005) and Soberón (2010) argue that what is modelled here is the fundamental part of the niche because it does not explicitly include (unless stated otherwise) any variable concerning dispersal limitation of the species or biotic interactions. We tend to think that we are modelling realized niches here by producing a model that closely resemble realized distributions of species based on observations where the species were actually found.

Despite this debate, SDM seems to be able to capture a significant amount of the ecological signature even when biotic data is often lacking in the models (Elith and Leathwick 2009). The environmental factors commonly used in these models determine the size and shape of the species distributions at continental or regional scales (Hortal et al. 2010).

Difference between tribes

Between the tribes explanatory variable contribution is roughly the same, with some differences on the strength of the contribution from one to another. Patterns of alpha and gamma diversity differ

from tribe to tribe, probably as consequence of different phylogeographic histories of these lineagues.

It is interesting to point out the two possible phylogenetic differences that might that might exist in the tribe Sphingini, which exhibits two geographically strongly differentiated, groups (one in Africa and one in Oriental region: Figure 5.6G). It would be necessary to have a detailed analysis of the phylogenetic effects on range sizes across the tribes (Beck et al 2006e. Probably there is a positive relation between geographic range and extinction resistance (Jablonski 2008), depending on the selective pressures that act upon them.

Alpha, gamma and beta diversity

Our maps for alpha and gamma diversity show the expected increment of species richness towards the tropics. However, there are some details in the alpha diversity pattern that requires further attention. There are some places that seem to have quite high estimates of species richness, Borneo for example. Previous studies have reported no more than 60 species even for the most species rich places in South-East Asia (Beck et al. 2007), and at Borneo light trapping sites the number of species did not exceed 40 even where more than 900 individuals were caught (Beck, unpublished data). It can be due to the fact that maps of species richness generated by overlaying continuous ranges of species might lead to a systematic over-prediction than those based on local inventories (Lennon et al. 2003, Jetz and Fine 2012). In contrast, there are some other places where a much lower species richness was predicted (i.e. Temburong District of Brunei) that might be due three factors: 1) a technical problem when adding up the maps or 2) a technical problem with the MODIS vegetation data (i.e. lack of some cells in certain parts of the map), or 3) to a systematic error in the WorldClim data, (i.e. weather stations recording differently in Malaysia and Brunei), something that have been observed for the German-French border along the Rhine river due to French weather station recording slightly different (Pers. Comm. E. Parlow, University of Basel), .

Moreover, it is often assumed that beta diversity is higher in the tropics (Novotny and Weiblen 2005) Tropics offers a set of conditions to promote coexistence of many species; Here (i.e. in the tropics), species exhibit narrower physiological tolerances (therefore their range size are smaller) plus a lack of overlap in the thermal regimes over tropical altitudinal gradients, sets conditions for reduced dispersal and overlap in species distributions across elevation, consequently, it would lead to high rates of allopatric speciation (Ghalambor et al. 2006, Ruggiero and Hawkins 2008). Interestingly, our results could not confirm a higher β-diversity in the tropics, but it does show that places with higher values correspond to mountainous landscapes which provide conditions of habitat heterogeneity and dispersal limitation.

In addition to habitat heterogeneity and dispersal limitation, the velocity and magnitude of past climatic changes may also have affected species range sizes and therefore probably also β-diversity. Species with smaller range (common in the tropics) might be prone to extinction due to climatic changes, which as stronger outside the tropics. Species with larger ranges are more resilient and would persist (Dynesius and Jansson 2000).

Challenges and limitations

There were many challenges to overcome during this project. These might provide valuable experience for those that set out to undertake similar studies like this one. In the following we list some of our main practical insights on how to do things better than we did.

1) Going to Museum collection is worth it. There are vast amounts of otherwise inaccessible data stored (Figure 5.2A), and these data may be more valuable than some other data sources (Beck et al. subm, Chapter 3 of this thesis). We generally experienced a lot of support and a positive attitude towards our study from museum staff, both regarding administrative access as well as within-collection support.

2) When assembling a database of occurrences of this magnitude, errors in taxonomy (i.e., misidentification, nomenclature issues) are one of the most common, yet often hidden problems. Without the participation of a taxonomic expert in the inspection and determination of original specimens when visiting collections or screening published data, it is almost impossible to provide high-quality data. Poorly determined collections or databases should be used with extreme care. Progress had been made on this issue for plants, where efforts to standardized nomenclature resulted in a software-tool designed to solve name-related problems in vegetation databases (Jansen and Dengler 2010).

3) Georeferencing is the bottleneck in working with distribution record data. It is a very time consuming process (we stent an estimated >1460 man-hours on this during the first part of the project), and it can be a high source of error. It is extremely important to check also those records that already have geographical coordinates. Experience but also motivation is crucial factor to provide good georeferencing data.

4) Obtaining field records from the tropics proved very important to fill gaps in data for these often undersampled yet species-rich reasons. Furthermore, there in particular, molecular data can help to solve taxonomic problems. However, regulations in tropical as well as in developed countries make it increasingly difficult to obtain such data (Renner et al. 2012), without apparent benefit to anyone. Guidelines for scientific research in tropical countries are often unpractical and naïve to large-extent

projects and local conditions, and sometimes entirely absurd (e.g., ban to visit places where DNA is found). Currently, data from tropical regions are often obtained from private collectors who in turn employ local collectors – a situation that is neither scientifically nor ethically desirable.

Conclusions

Our analyses have shown that the estimation of species distribution using SDM can be a valuable source of information for advancing our understanding of patterns of biodiversity. Although challenging and not without its problems it is clear that such they can provide a good approach, particularly in areas where sampling is lacking. SDM is not a certainly not replacement of real field work but provide an impetus to the discovery and description of new and rare species and improve our knowledge from those already known.

Geographical biases calls for great caution in the interpretation of results in some areas and/or approaches that can account for such. The gathering of as much distribution information as possible remains a high priority. Museum collections therefore have an important data source that remains to be exploited for many taxa and geographical areas, although much progress is been done. In undertaking studies outlined in this paper it is important to carefully choose the method that suits better your purposes. We have demonstrated that Maxent is a valuable one. Tropics and mountainous areas remain places where we can find the higher amount of species of Sphingids and analysis of their main drivers are still pending for the whole area.

For the successful completion of the project after four years was a great advantage to have an interdisciplinary team of collaborators (taxonomist, GIS & modeling skilled person, macroecologist, biogeographer). Certainly there are countless research topics that can be addressed by having this kind of data, but still we need more, so we hope that this project stimulates others to carry out also similar projects for other taxa specially insects.

It is our future plan to make this database publicly available through the website facility The map of life (http://www.mappinglife.org/).

Acknowledgements

We are very grateful with all those amateurs and professional collectors (to many to mention) who have made available their collections for us to use within this project. To those who help us georeference the data (R. Hagmann, M. Curran, M. Koop, S. Lang, S. Widler). To Patrick Vogt who help with setting up the use of the cluster. The study received financial support from the Swiss

National Science Foundation (SNF, project 3100AO_119879), the Synthesys program of the EU and the Freiwillige Akademische Gesellschaft (FAG) Basel.

Electronic Appendix

Together with this thesis, it is submitted the following electronic appendix, they are placed at the network drive of the University Computing Centre (URZ).

(Intranet: \\nlu-jumbo.nlu.p.unibas.ch\nlu-gis$\GIS)

EA1: List of published data sources

EA2: List of the species whose ranges expands into the Pacific Islands

EA3: List of the species with expert-drawn range maps

References

Araujo MB and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47

Asher, J., M. Warren, R. Fox, P. Harding, G. Jeffcoate, and S. Jeffcoat (Eds.). 2001. The Millennium Atlas of Butterflies in Britain and Ireland. . Oxford University Press, Oxford.

Austin, M. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157:101–118.

Ballesteros-Mejia, L., I. J. Kitching, and J. Beck. 2011. Projecting the potential invasion of the Pink Spotted Hawkmoth (Agrius cingulata) across Africa. International Journal of Pest Management 57 :153 – 159.

Ballesteros-Mejia, L., I. J. Kitching, W. Jetz, P. Nagel, and J. Beck. 2013. Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography 22: 586-595

Beale, C. M., and J. J. Lennon. 2012. Incorporating uncertainty in predictive species distribution modelling. Philosophical transactions of the Royal Society of London. Series (B) 367:247–258.

Beck, J., L. Ballesteros-Mejia, P. Nagel, I,J. Kitching, 2013. Online solutions and the "Wallacean shortfall": what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions 1-8. DOI: 10.1111/ddi.12083

Beck, J., L. Ballesteros-Mejia, C. M. Buchmann, J. Dengler, S. a. Fritz, B. Gruber, C. Hof, F. Jansen, S. Knapp, H. Kreft, A.-K. Schneider, M. Winter, and C. F. Dormann. 2012a. What's on the horizon for macroecology? Ecography 35:1–11.

Beck, J., J. D. Holloway, C. V. Khen, and I. J. Kitching. 2012b. Diversity partitioning confirms the importance of beta components in tropical rainforest Lepidoptera. The American naturalist 180:E64–E74.

Beck, J., and I. Kitching. 2004-2008. The Sphingidae of Southeast-Asia (incl. New Guinea, Bismarck and Solomon Islands) version 1.5.

Beck, J., and I. J. Kitching. 2007. Correlates of range size and dispersal ability: a comparative analysis of sphingid moths from the Indo‐Australian tropics. Global Ecology and Biogeography 16:341–349.

Beck, J., I. J. Kitching, and J. Haxaire. 2007. The latitudinal distribution of sphingid species richness in continental Southeast Asia: What causes the "biodiversity hotspot" in northern Thailand. Raffles Bulletin of Zoology 55:179–185.

Beck, J., I. J. Kitching, and K. E. Linsenmair. 2006a. Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. Ecotropica 12:1–13.

Beck, J., I. J. Kitching, and K. E. Linsenmair. 2006b. Effects of habitat disturbance can be subtle yet significant: biodiversity of hawkmoth-assemblages (Lepidoptera: Sphingidae) in Southeast-Asia. Biodiversity and Conservation 15:465–486.

Beck, J., I. J. Kitching, and K. E. Linsenmair. 2006c. Extending the study of range – abundance relations to tropical insects: sphingid moths in Southeast Asia. Evolutionary Ecology Research 8:677–690.

Beck, J., I. J. Kitching, and K. E. Linsenmair. 2006d. Wallace's line revisited: has vicariance or dispersal shaped the distribution of Malesian hawkmoths (Lepidoptera: Sphingidae)? Biological Journal of the Linnean Society 89:455–468.

Beck, J., I. J. Kitching, and K. E. Linsenmair. 2006e. Measuring range sizes of Southeast-Asian hawkmoths (Lepidoptera: Sphingidae): effects of scale, resolution and phylogeny. Global Ecology and Biogeography 15, 339–348.

Beck J., W. Nässig, 2007. Diversity and abundance patterns, and revised checklist, of saturniid moths (Lepidoptera: Saturniidae) from Borneo. Nachrichten des Entomologischen Vereins Apollo 28: 155-164.

Bik, H. M., D. L. Porazinska, S. Creer, J. G. Caporaso, R. Knight, and W. K. Thomas. 2012. Sequencing our way towards understanding global eukaryotic biodiversity. Trends in ecology and evolution 27:233–43.

Böller, M. 2012. Modellierung von Verbreitungsgebieten mit MaxEnt: Der Effekt von verzerrten presence-only Datensätzen der Global Biodiversity Information Facility (GBIF). BSc Thesis. University of Basel.

Brehm, G., J. Homeier, and K. Fiedler. 2003. Beta diversity of geometrid moths (Lepidoptera: Geometridae) in an Andean montane rainforest. Diversity and Distributions Distributions 9:351–366.

Breiman, L. 2001. Random forests. Machine learning 45:5–32.

Buckley, L. B., and W. Jetz. 2007. Environmental and historical constraints on global patterns of amphibian richness. Proceedings of the Royal Society (B) 274:1167–1173.

Colwell, R. K. 2005. EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application published at: http://purl.oclc.org/estimates.

Costello, M. J., and S. P. Wilson. 2011. Predicting the number of known and unknown species in European seas using rates of description. Global Ecology and Biogeography 20:319–330.

Davis, A. L. V., C. H. Scholtz, and S. L. Chown. 1999. boundaries and community Species turnover , in gradient assemblages across an altitudinal South Africa of dung beetle biogeographical composition. Journal of Biogeography 26:1039–1055.

Deans, A. R., M. J. Yoder, and J. P. Balhoff. 2012. Time to change how we describe biodiversity. Trends in ecology and evolution 27:78–84.

Diamond, J. 2006. Collapse: How Societies Choose to Fail or Succeed. Page 573. Viking, Penguin Group, New York.

Diniz-Filho, J. A. F., P. De Marco Jr, and B. a. Hawkins. 2010. Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. Insect Conservation and Diversity 3:172–179.

Dynesius, M., and Jansson, R. 2000. Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. Proceedings of the National Academy of Sciences (B) 97:9115–9120.

Elith, J., and J. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Elith J, C. Graham, R. Anderson, M. Dudık, S. Ferrier, A. Guisan, R. Hijmans, F. Huettmann, J. Leathwick, A.Lehmann, J. Li, L. Lohmann , B. Loiselle, G. Manion, C. Moritz, M. Nakamura Y. Nakazawa, Jm. Overton, A. Peterson, S. Phillips, K. Richardson, R. Scachetti-Pereira, R. Shapire, J. Soberon, S. Williams, M. Wisz, N. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

Elton, C. 1927. Animal Ecology. Sidgwick & Jackson, London.

Engler R, A. Guisan, L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263-274

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? Systematic Biology 51:331–363.

Field, R., B. A. Hawkins, H. V. Cornell, D. J. Currie, J. A. F. Diniz-Filho, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. M. O'Brien, and J. R. G. Turner. 2009. Spatial species-richness gradients across scales: a meta-analysis. Journal of Biogeography 36:132–147.

Ghalambor, C. K., R.B. Huey, P.R. Martin, J.J. Tewksbury and G. Wang. 2006. Are mountain passes higher in the tropics? Janzen's hypothesis revisited. Integrative and Comparative Biology 46: 5-17

Galster, S., N. D. Burgess, J. Fjeldsa°, L. A. Hansen, and C. Rahbek. 2007. One degree resolution databases of the distribution of 1085 mammals in Sub-Saharan Africa.

Gasc, J. P., A. Cabela, J. Crnobrnja-Isailovic, D. Dolmen, K. Grossenbacher, P. Haffner, J. Lescure, H., T.S. Martens, M. Veith, and A. Zuiderwijk. 1997. Atlas of amphibians and reptiles in Europe. Societas Europaea Herpetologica and Museum National d'Histoire Naturelle, Paris.

Gaston, K.J. 2003. The structure and dynamics of geographic ranges. Oxford University Press, Oxford.

Gibbons, D.W., J.B. Reid and R.A.Chapman. 1993 The new atlas of breeding birds in Britain and Ireland: 1988-1991. T. & A.D. Poyser

Glor, R. E., and D. Warren. 2011. Testing ecological explanations for biogeographic boundaries. Evolution; international journal of organic evolution 65:673–83.

Godfray, H. C., T. Lewis, and J. Memmott. 1999. Studying insect diversity in the tropics. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 354:1811–24.

Godinho, R., J. Teixeira, R. Rebelo, P. Segurado, and A. Loureiro. 1999. Atlas of the continental Portuguese herpetofauna: an assemblage of published and new data. Revista Espanola de Herpetologia 13:61–82.

Grinnell, J. 1917. The Niche-Relationships of the California Thrasher. The Auk 34:427–433.

Guisan, a, and N. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Hamilton, A. J., Y. Basset, K. K. Benke, P. S. Grimbacher, S. E. Miller, V. Novotný, G. A. Samuelson, N. E. Stork, G. D. Weiblen, and J. D. L. Yen. 2010. Quantifying uncertainty in estimation of tropical arthropod species richness. The American naturalist 176:90–5.

Hanley, J. A., and B. J. McNeil. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143:29–36.

Hansen, L. A., N. D. Burgess, J. Fjeldsa°, and C. Rahbek. 2007. One degree resolution databases of the distribution of 739 amphibians in Sub-Saharan Africa.

Hawkins, B. A., M. Rueda, and M. Á. Rodríguez. 2008. What Do Range Maps and Surveys Tell Us About Diversity Patterns? Folia Geobotanica 43:345–355.

Holloway J.D. (1987). The moths of Borneo, part 3: Lasiocampidae, Eupterotidae, Bombicidae, Brahmaeidae, Saturniidae, Sphingidae. The Malay Nature Society and Southdene Sdn. Bhd., Kuala Lumpur.

Hortal, J., N. Roura-Pascual, N. Sanders, and C. Rahbek. 2010. Understanding (insect) species distributions across spatial scales. Ecography 33:51–53.

Hurlbert, A. H., and W. Jetz. 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences (B) 104:13384–13389.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposium On Quantitative Biology 22:415–427.

ICZN. 1999. International Code of Zoological Nomenclature, 4th edition. International Trust for Zoological Nomenclature., London.

Jablonski, D. 2008. Species selection: theory and data. Annual Review of Ecology, Evolution, and Systematics 39:501–524.

Jankowski, J. E., A. L. Ciecka, N. Y. Meyer, and K. N. Rabenold. 2009. Beta diversity along environmental gradients: implications of habitat specialization in tropical montane landscapes. The Journal of animal ecology 78:315–327.

Jansen, F., and J. Dengler. 2010. Plant names in vegetation databases - a neglected source of bias. Journal of Vegetation Science 21:1179–1186.

Janzen, D. H. 1984. Two ways to be a tropical big moth: Santa Rosa saturniids and sphingids. Pages 85–140 *in* R. Dawkins and M. Ridley, editors. Oxford surveys in Evolutionary Biology, 1st edition. Oxford University Press, Oxford.

Jetz, W., and P. V. a Fine. 2012. Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. PLoS biology 10:e1001292.

Jetz, W., J. M. McPherson, and R. P. Guralnick. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends in ecology and evolution 27:151–159.

Jetz, W., and C. Rahbek. 2002. Geographic range size and determinants of avian species richness. Science. 297:1548–51.

Joppa, L. N., D. L. Roberts, N. Myers, and S. L. Pimm. 2011. Biodiversity hotspots house most undiscovered plant species. Proceedings of the National Academy of Sciences (B) 108:13171–13176.

Kawahara, A. Y., A. a Mignault, J. C. Regier, I. J. Kitching, and C. Mitter. 2009. Phylogeny and biogeography of hawkmoths (Lepidoptera: Sphingidae): evidence from five nuclear genes. PloS one 4:e5719.

Keil, P., and B. A. Hawkins. 2009. Grids versus regional species lists: are broad-scale patterns of species richness robust to the violation of constant grain size? Biodiversity and Conservation 18:3127–3137.

Kitching, I., J., and J. M. Cadiou. 2000. Hawkmoths of the world. The Natural History Museum and. Cornell University Press, London.

Kreft, H., and W. Jetz. 2007. Global patterns and determinants of vascular plant diversity. Proceedings of the National Academy of Sciences (B) 104:5925–5930.

Kreft, H., and W. Jetz. 2010. A framework for delineating biogeographical regions based on species distributions. Journal of Biogeography 37:2029–2053.

De Latin, G. 1967. Grundriss der Zoogeographie. G. Fisher-Verlag, Jena. Page 602.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. 2003. Contribution of rarity and commonness to patterns of species richness. Ecology Letters 7:81–87.

Linder, H. P., H. M. de Klerk, J. Born, N. D. Burgess, J. Fjeldså, and C. Rahbek. 2012. The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. Journal of Biogeography

Lomolino, M.V. 2004 Conservation biogeography. Frontiers of Biogeography: new directions in the geography of nature (eds. Lomolino MV & Heaney LR), Sinauer Associates, Sunderland, Massachusetts. pp 293–296.

Maddison, D. R., R. Guralnick, A. Hill, A.-L. Reysenbach, and L. a McDade. 2012. Ramping up biodiversity discovery via online quantum contributions. Trends in ecology and evolution 27:72–77.

Mateo, R. G., T. B. Croat, Á. M. Felicísimo, and J. Muñoz. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Diversity and Distributions 16:84–94.

McKnight, M. W., P. S. White, R. I. McDonald, J. F. Lamoreux, W. Sechrest, R. S. Ridgely, and S. N. Stuart. 2007. Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. PLoS biology 5:e272.

Morrone J. J. 2009. Evolutionary Biogeography: An integrative approach with case studies. Columbia University Press, New York.

Mutanen, M., N. Wahlberg, and L. Kaila. 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proceedings of the Royal Society (B) 277:2839–2848.

Novotny, V., and G. D. Weiblen. 2005. From communities to continents: beta diversity of herbivorous insects. Annales Zoologici Fennici 42:463–475.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12:361–371.

Phillips, S., R. Anderson, and R. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

Pittaway, A. R. 1997-2012. Sphingidae of the Western Palaearctic.

Pittaway, A. R., and I. J. Kitching. 2000-2012. Sphingidae of the Eastern Palaearctic (including Siberia, the Russian Far East, Mongolia, China, Taiwan, the Korean Peninsula and Japan).

Rasmussen, J. B., L. A. Hansen, N. D. Burgess, J. Fjeldsa°, and C. Rahbek. 2007. One degree resolution databases of the distribution of 467 snakes in Sub-Saharan Africa.

Regier, J. C., A. Zwick, M. P. Cummings, A. Y. Kawahara, S. Cho, S. Weller, A. Roe, J. Baixeras, J. W. Brown, C. Parr, D. R. Davis, M. Epstein, W. Hallwachs, A. Hausmann, D. H. Janzen, I. J. Kitching, M. A. Solis, S.-H. Yen, A. L. Bazinet, and C. Mitter. 2009. Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. BMC evolutionary biology 9:280.

Renner, S. C., D. Neumann, M. Burkart, U. Feit, P. Giere, A. Gröger, A. Paulsch, C. Paulsch, M. Sterz, and K. Vohland. 2012. Import and export of biological samples from tropical countries–considerations and guidelines for research teams. Organisms Diversity & Evolution 12:81–98.

Robinson, G. ., P. R. Ackery, I. J. Kitching, G. W. Beccaloni, and L. M. Hernández. 2001. Hostplants of the moth and butterfly caterpillars of the Oriental Region. The Natural History Museum and Southdene Sdn Bhd,Kuala Lumpur.

Rothschild, L.W. and K. Jordan. 1903. A revision of the lepidopterous family Sphingidae. Novitates Zoologicae, 9 (suppl.): 1–972.

Ruggiero, A., B. Hawkins. 2008. Why mountains support that many species of birds?. Ecography. 31:306-315

Scoble, M. J., K. Gaston J., and A. Crook. 1995. Using taxonomic data to estimate species richness in geometridae. Journal of the Lepidopterists' Society 49:136–147.

Scott, J.A. 1986. The Butterflies of North America: A Natural History and Field Guide. Standford University Press, Standford.

Settele, J., O. Kudrna, A. Harpke, I. Kuehn, C. van Swaay, R. Verovnik, M. Warren, M. Wiemers, J. Hanspach, T. Hickler, E. Kühn, I. van Halder, K. Veling, A. Vliegenthart, I. Wynhoff, and O. Schweiger. 2008. Climatic Risk Atlas of European Butterflies. BIORISK – Biodiversity and Ecosystem Risk Assessment. . Pensoft, Sofia.

Soberon, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species distributional areas. Biodiversity Informatics 2:1–10.

Soberón, J. M. 2010. Niche and area of distribution modeling: a population ecology perspective. Ecography 33:159–167.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science 240:1285–1293.

Terribile, L. C., M. Á. Olalla-Tárraga, J. A. F. Diniz-Filho, and M. Á. Rodríguez. 2009. Ecological and evolutionary components of body size: geographic variation of venomous snakes at the global scale. Biological Journal of the Linnean Society 98:94–109.

Thomas, C. D., C. R. Bulman, and R. J. Wilson. 2008. Where within a geographical range do species survive best? A matter of scale. Insect Conservation and Ecology 1:2–8.

The Times Atlas of the World: Comprehensive Edition. 2010. Times Books, HarpenCollins Publishers. Edition 10th. London

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. BIOMOD–A platform for ensemble forecasting of species distributions. Ecography 32:369–373.

Tuomisto, H. 2010. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33:2–22.

Turner, J. R. G., C. M. Gatehouse, and C. A. Corey. 1987. Does solar energy control organic diversity? Butterflies, moths and the British climate. Oikos 48:195–205.

VanDerWal, J., L. P. Shoo, C. Graham, and S. E. Williams. 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220:589–594.

Wallace, A. R. 1869. The Malay Archipelago. Oxford in Asia Hardback Reprint (1986). Oxford University Press, Oxford.

Warren, D. L. 2012. In defense of "niche modeling". Trends in ecology and evolution 27:497–500.

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279–338.

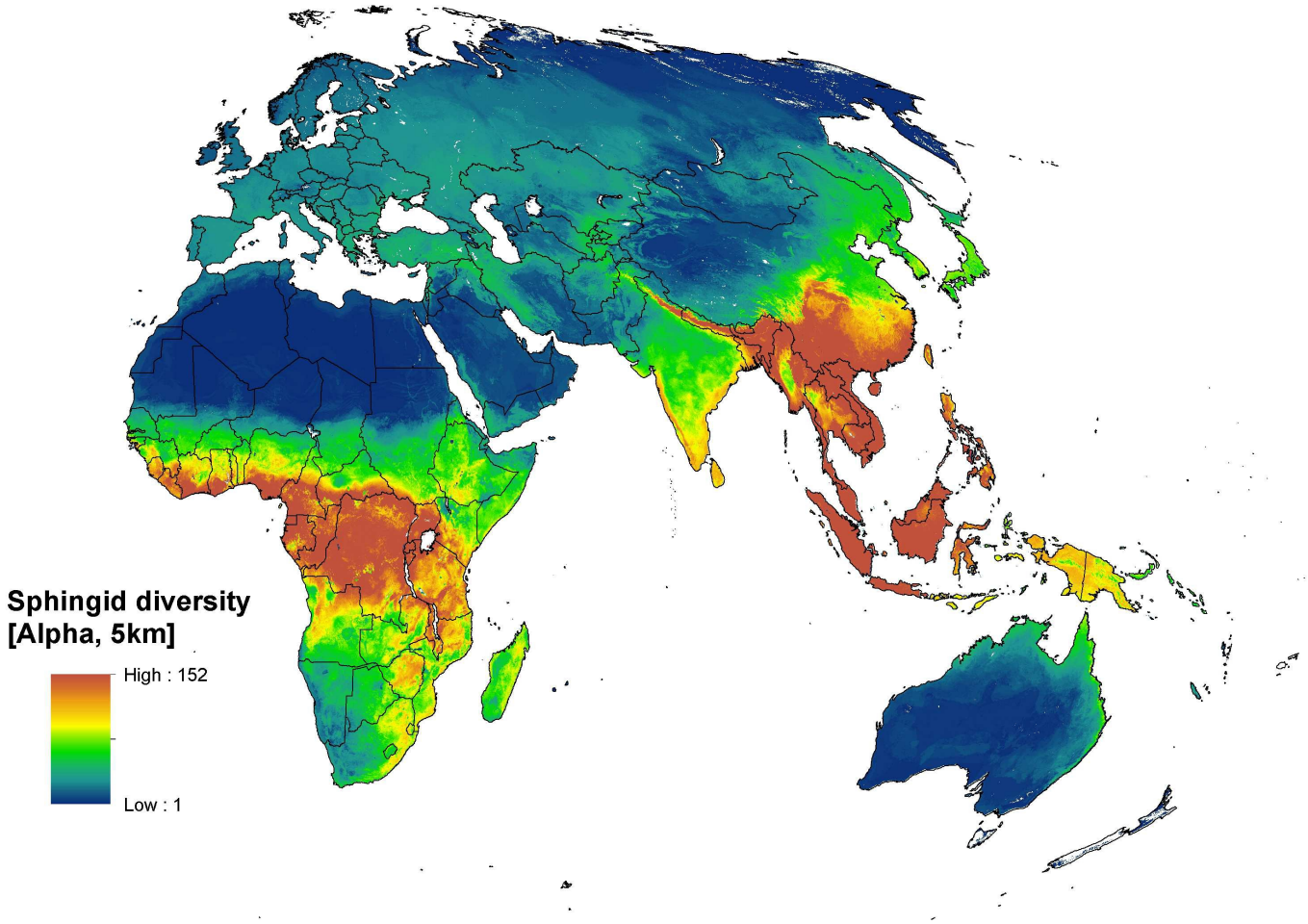
Wieczorek, J., Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18:745–761.

Wilson, E.O. 2003. The encyclopedia of life. Trends in Ecology and Evolution 18:77:80

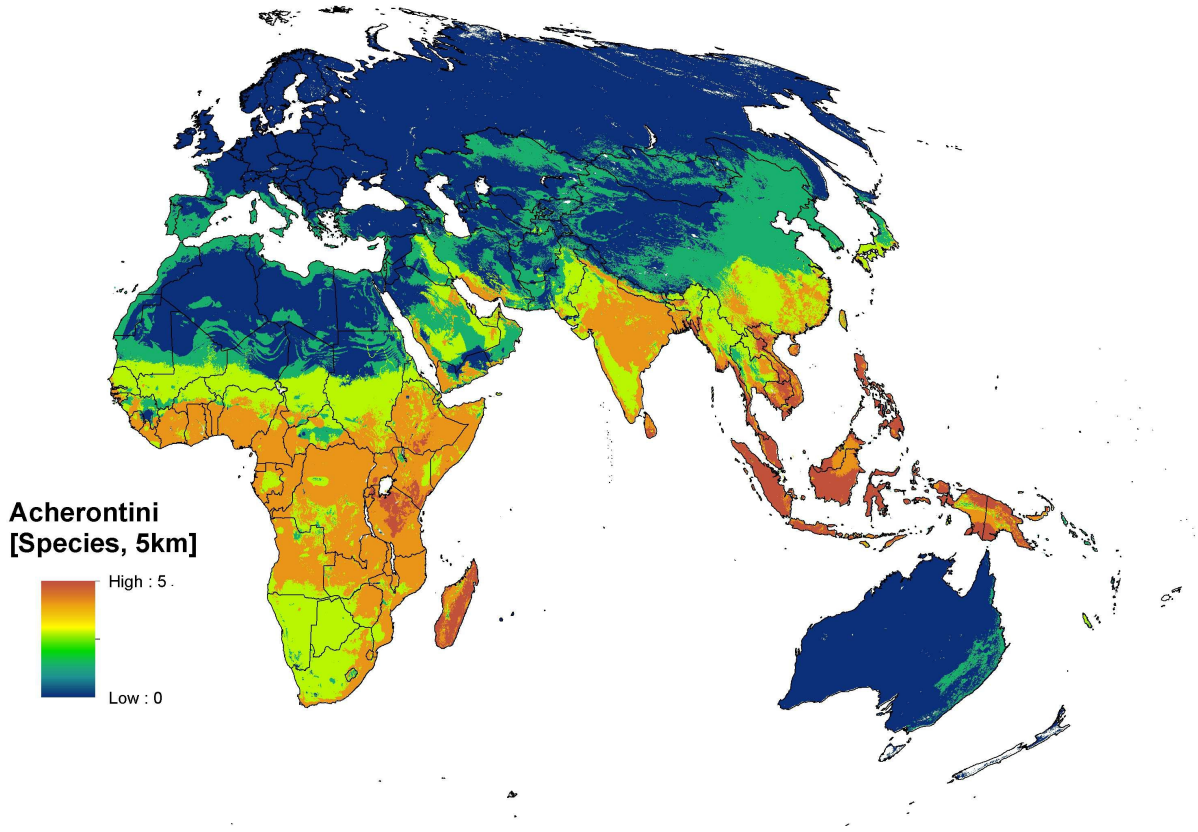
Yesson, C., P.W.Brewer, , T. Sutton, N. Caithness,J.S. Pahwa,,M.Burgess, W.A., Gray, R.J. White, A.C. Jones, F.A. Bisby, and A. Culham,. 2007: How global is the Global Biodiversity Information Facility? PloS One 2, e1124.

Figure 5.6. Sphingid diversity across the Old World total and by tribes. (α-diversity, 5 x 5km cellsize). (A) Total sphingid species diversity. (B) Species diversity of the Acherontini tribe.(C) Species diversity of the Ambulycini tribe. (D) Species diversity of the Dilophonotini tribe. (E) Species diversity of the Macroglossini tribe. (F) Species diversity of the Smerinthini tribe (G) Species diversity of the Sphingini tribe. (H) Species diversity of the Sphingulini tribe.

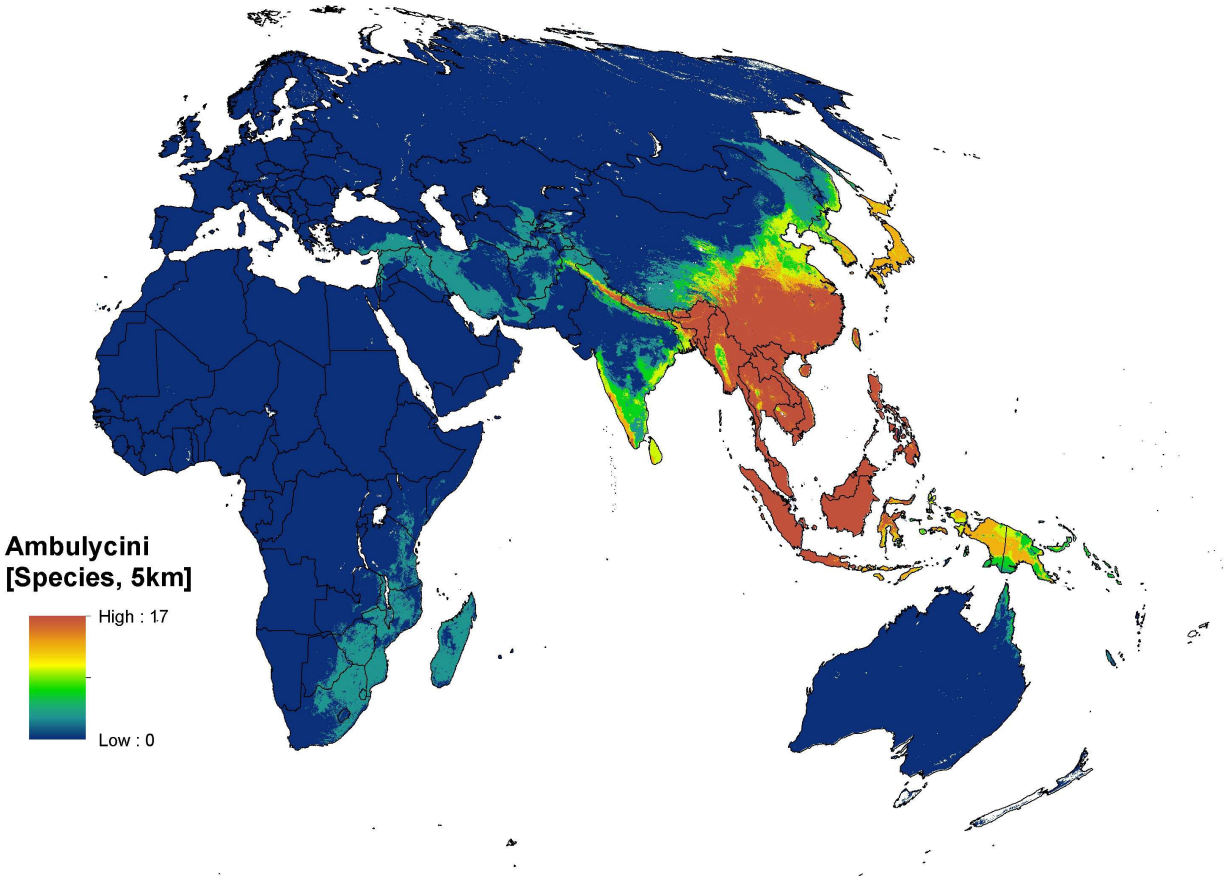
(A)



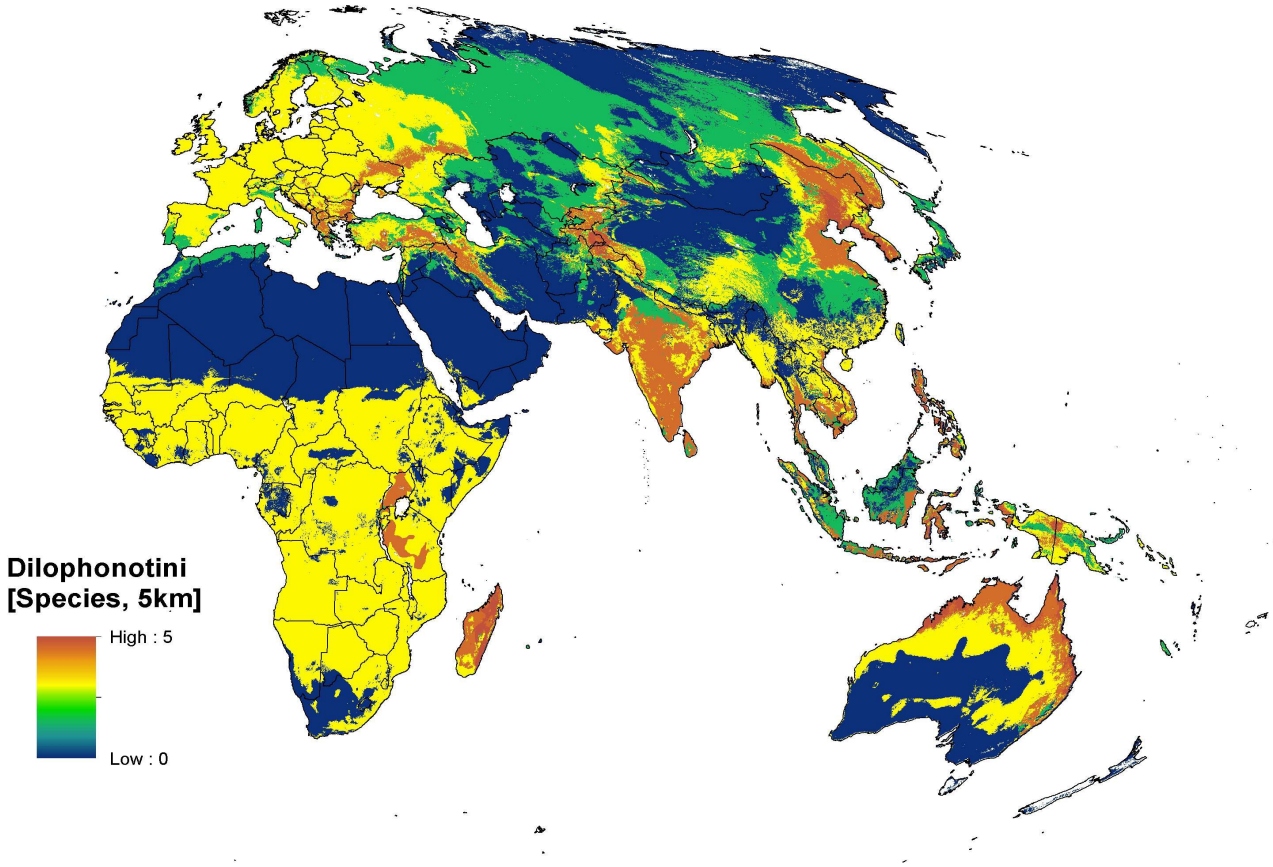
(B)


Acherontini
[Species, 5km]
High : 5
Low : 0

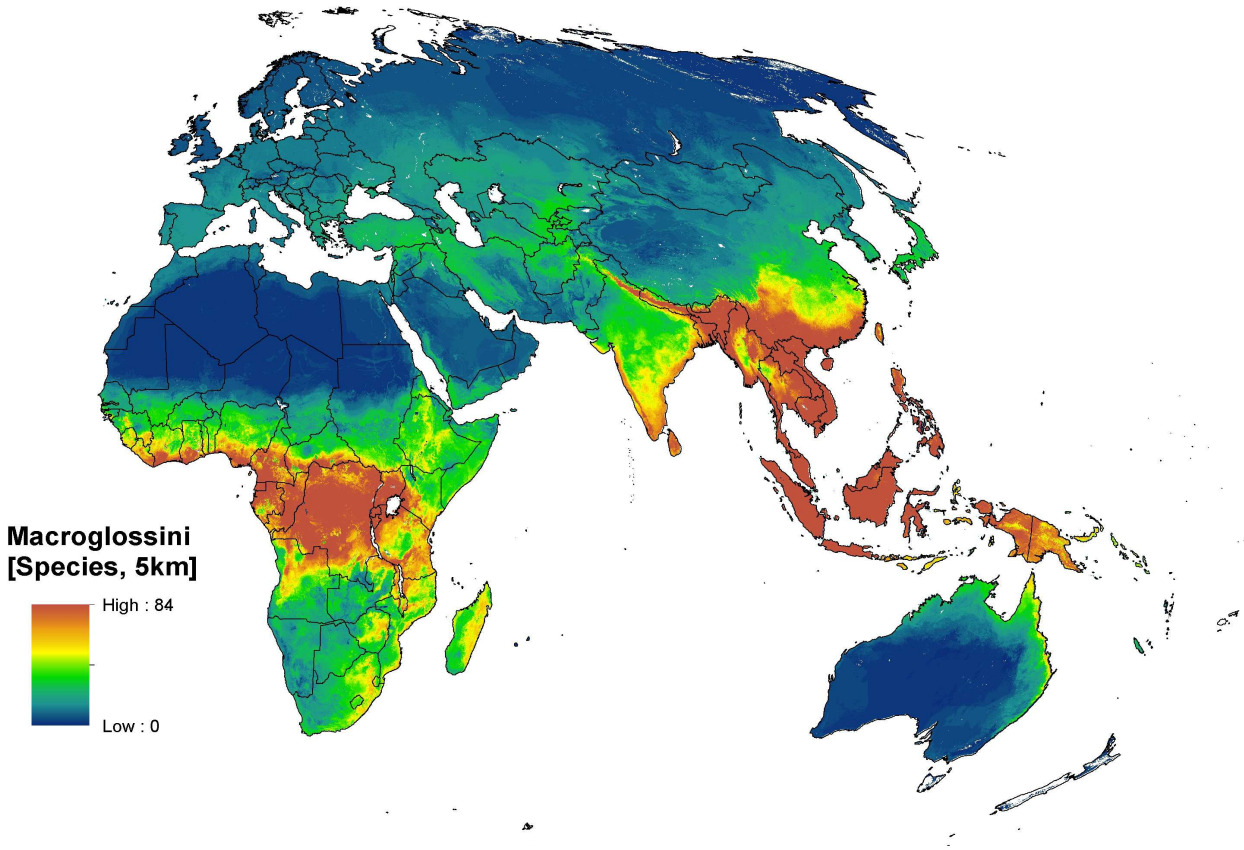
(C)


Ambulycini
[Species, 5km]
High : 17
Low : 0

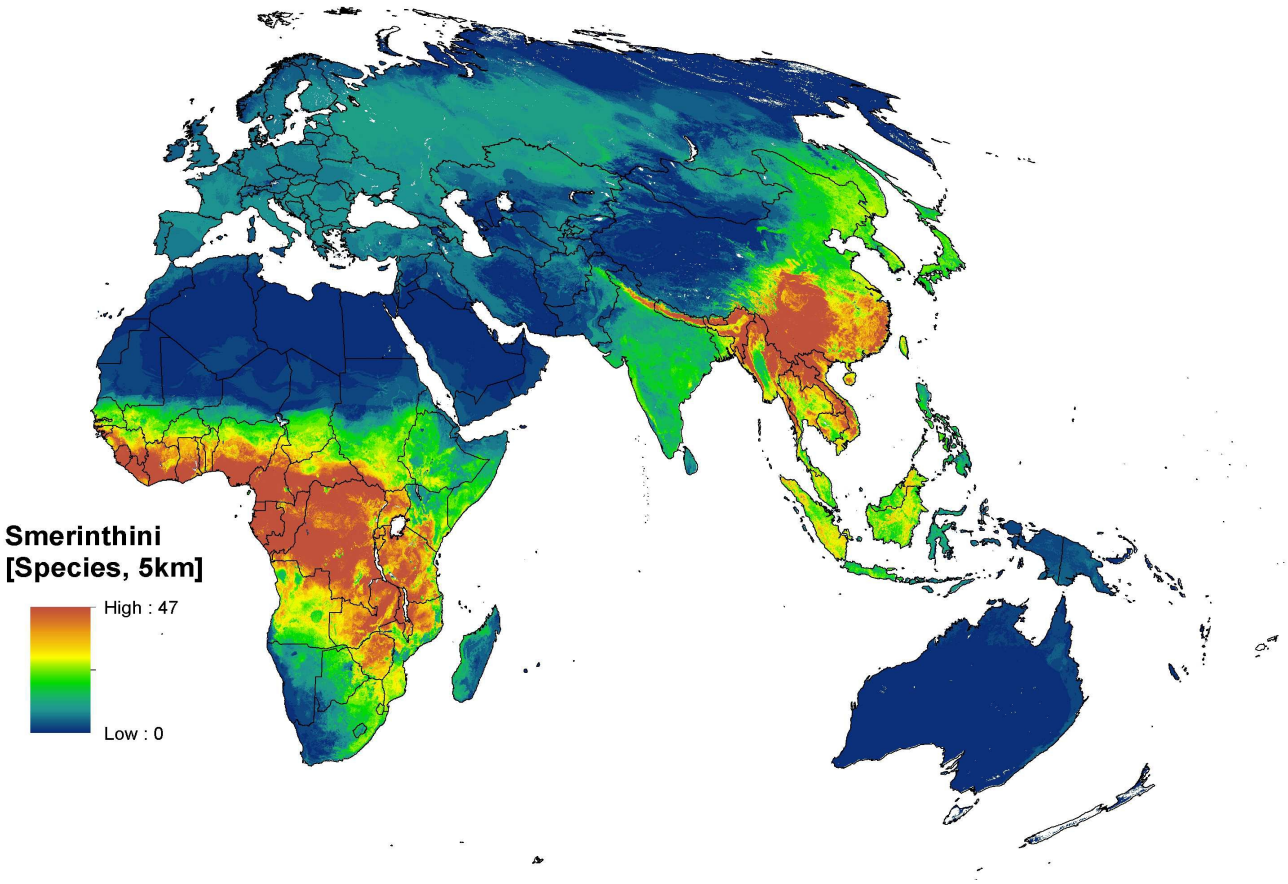
(D)


Dilophonotini
[Species, 5km]
High : 5
Low : 0

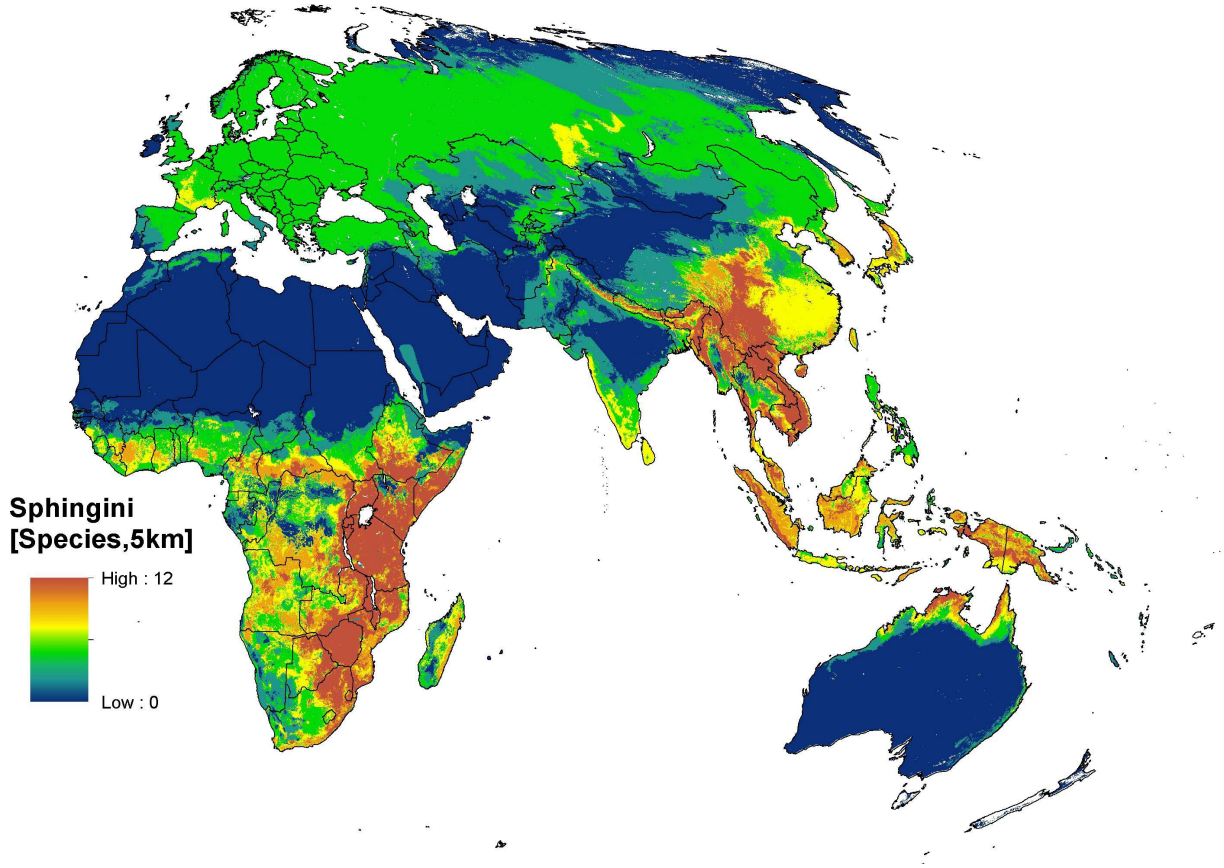
(E)


Macroglossini
[Species, 5km]
High : 84
Low : 0

(F)



(G)



(H)

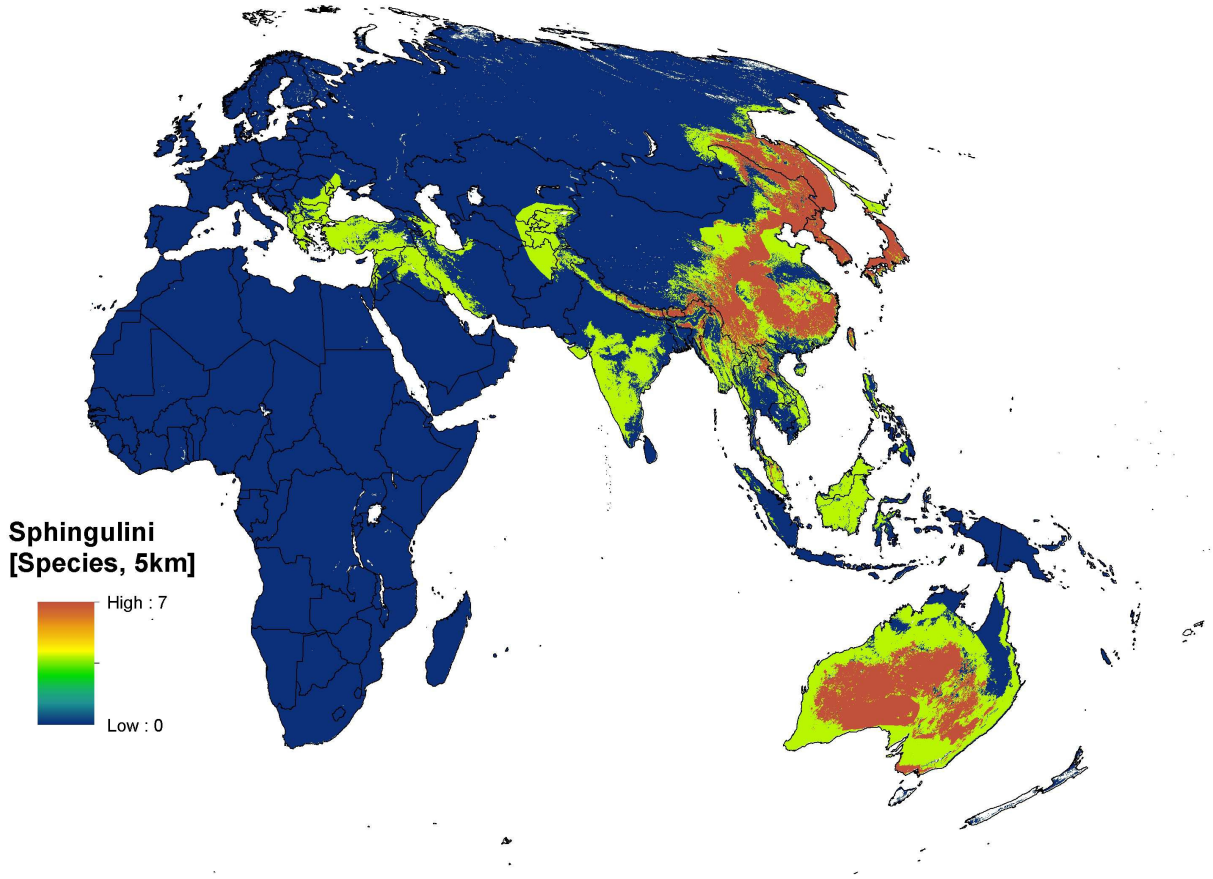
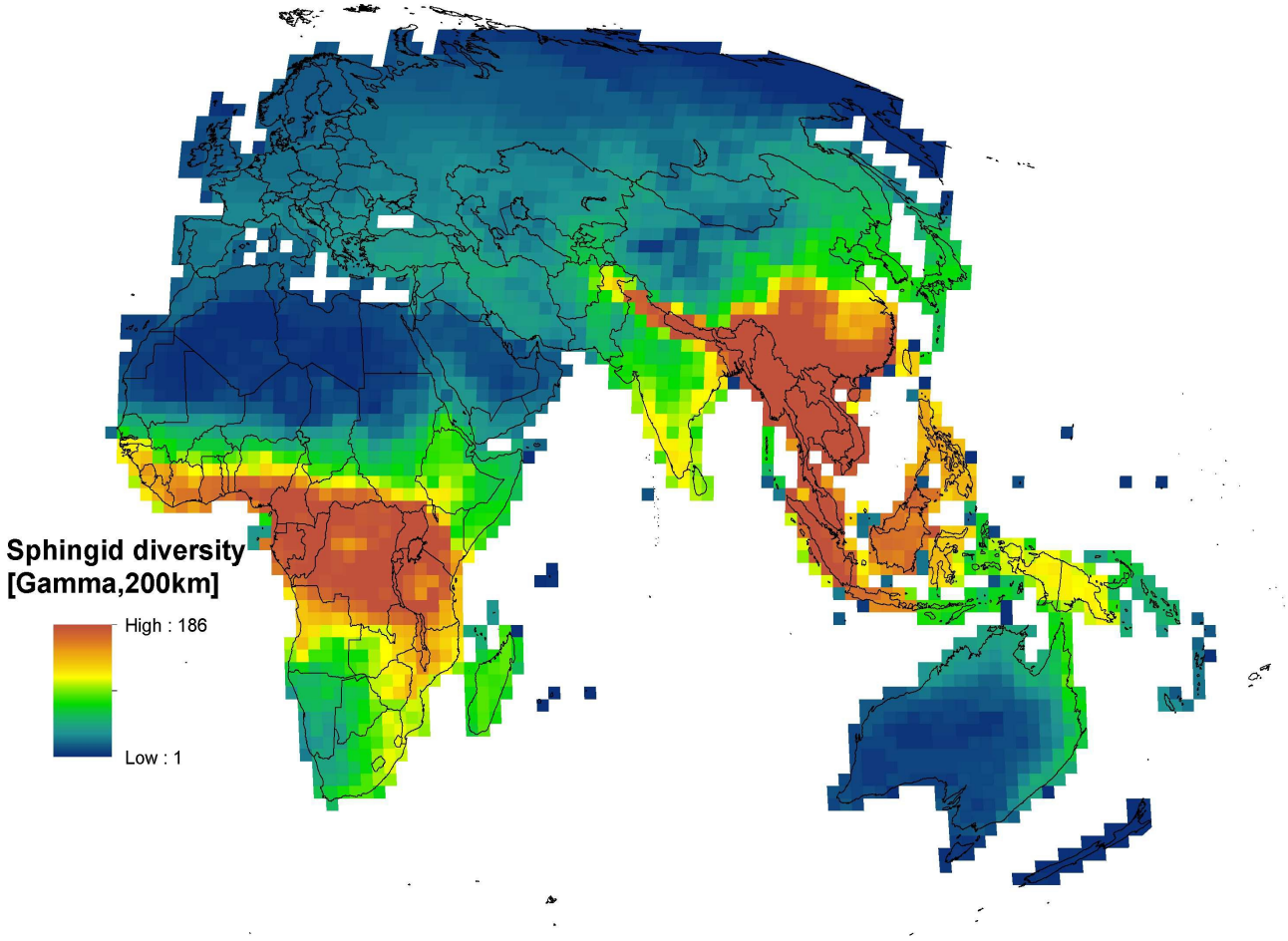
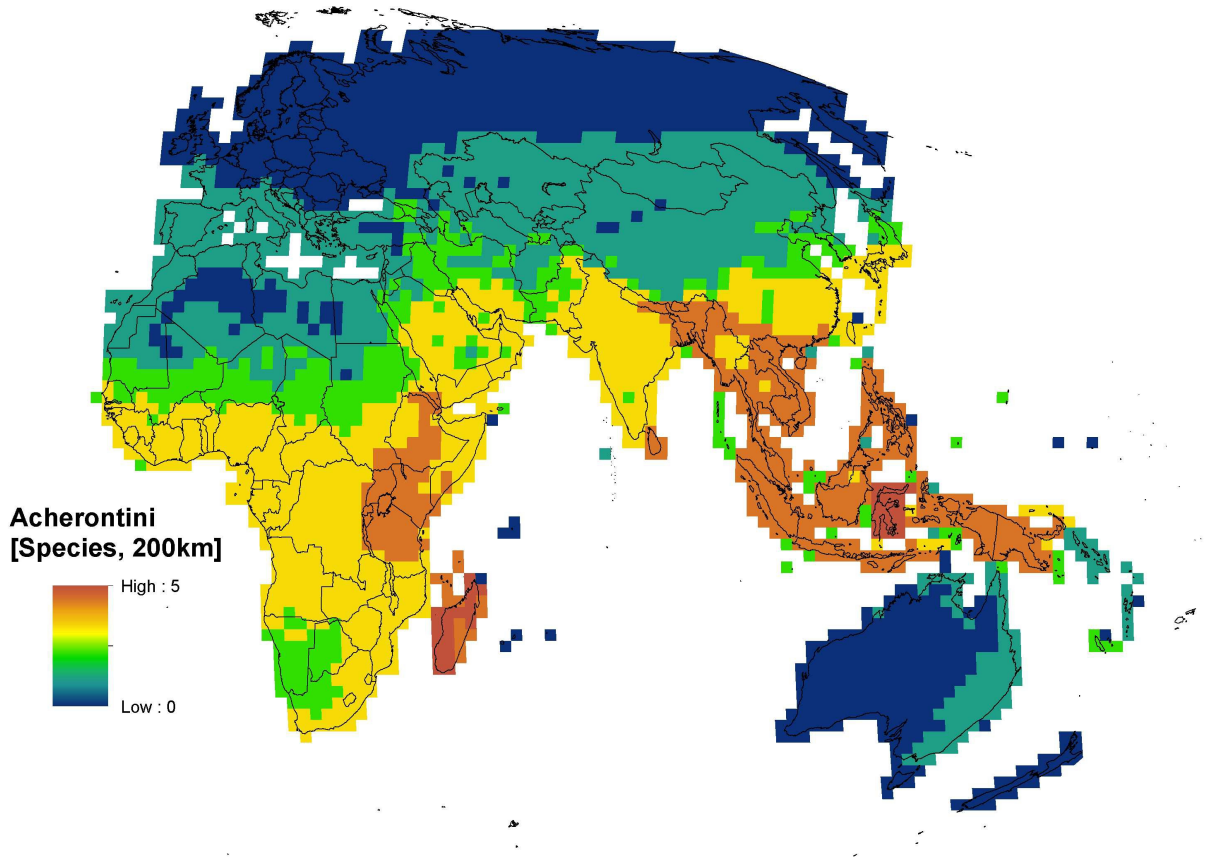

Sphingulini
[Species, 5km]
High : 7
Low : 0

Figure 5.6. Sphingid diversity across the Old World total and by tribes. (γ-diversity, 200 x 200 km cellsize). (A) Total sphingid species diversity. (B) Species diversity of the Acherontini tribe.(C) Species diversity of the Ambulycini tribe. (D) Species diversity of the Dilophonotini tribe. (E) Species diversity of the Macroglossini tribe. (F) Species diversitz of Smerinthini tribe. (G) Species diversity of the Sphingini tribe. (H) Species diversity of the Sphingulini tribe.

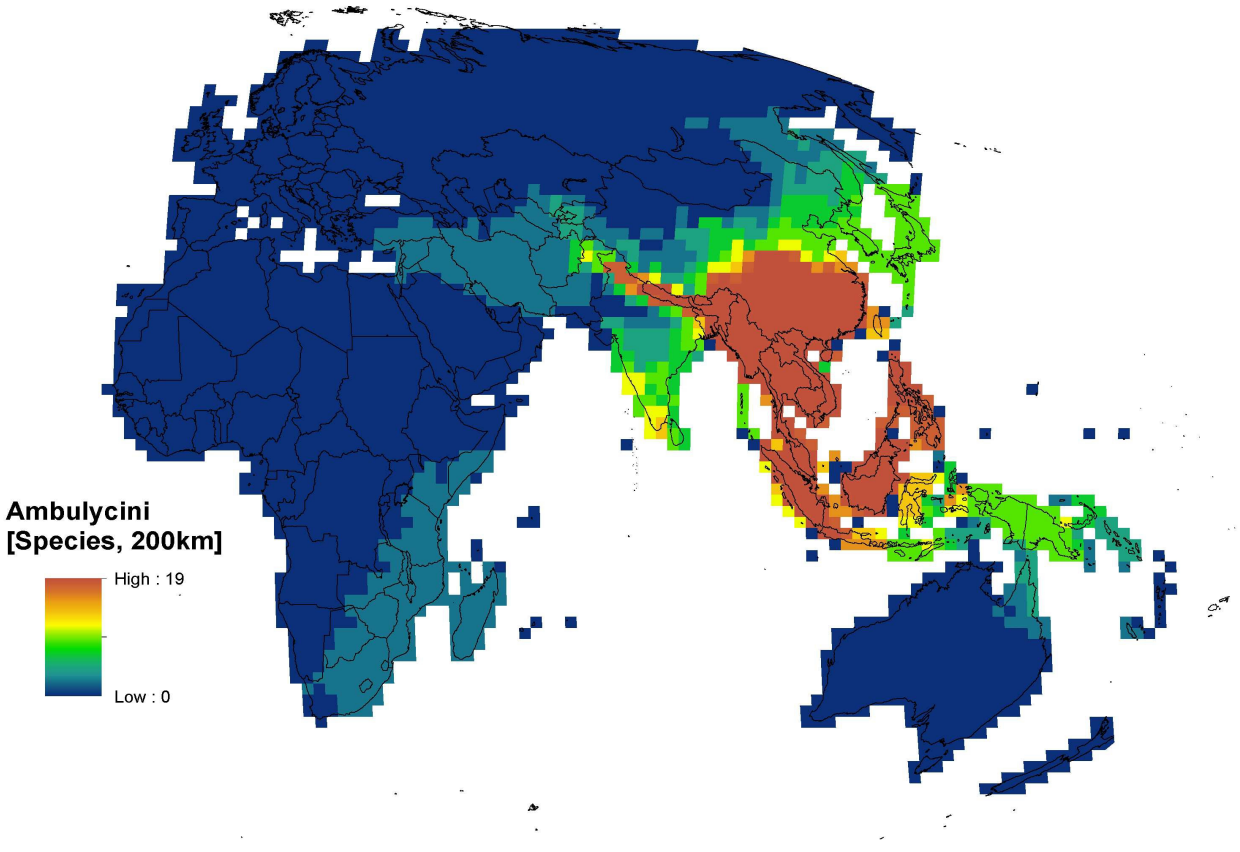
(A)



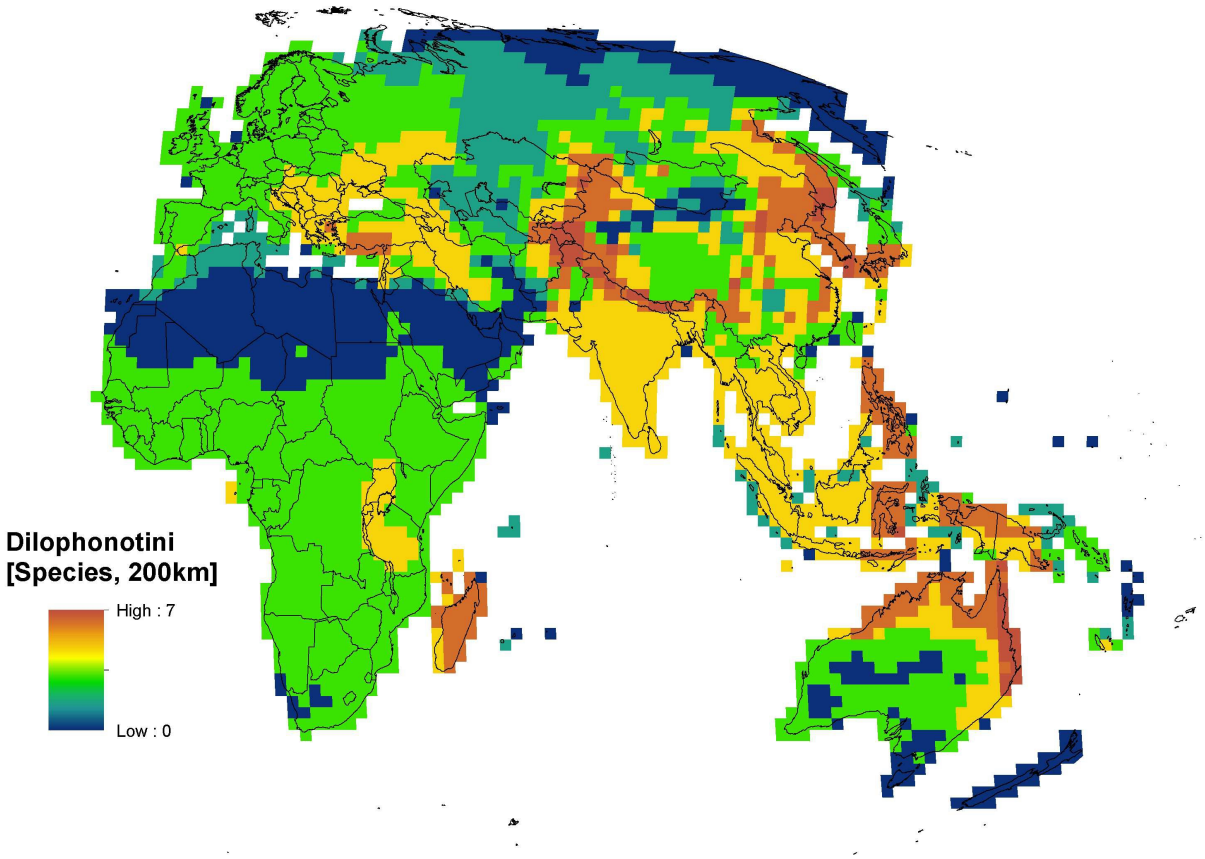
(B)


Acherontini
[Species, 200km]
High : 5
Low : 0

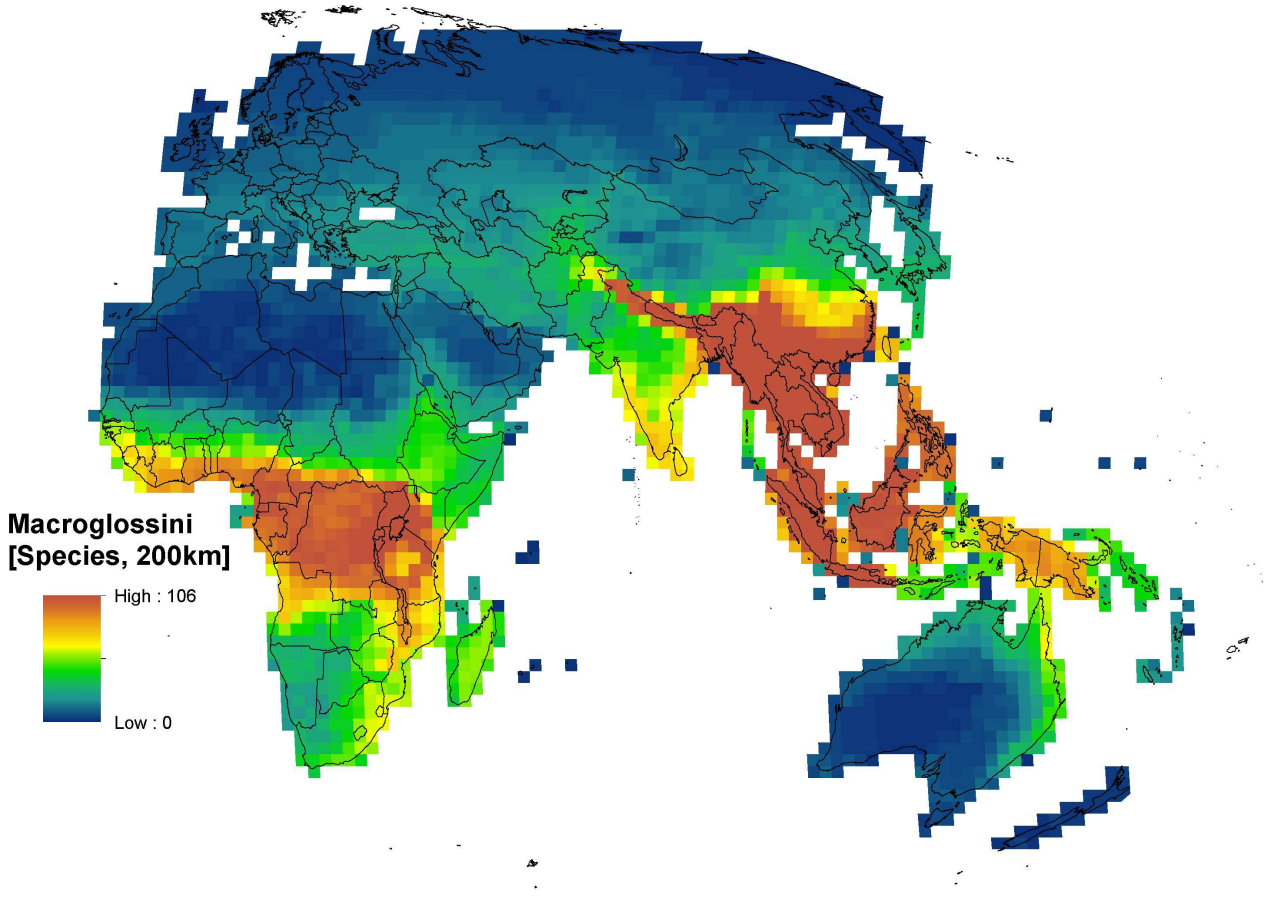
(C)


Ambulycini
[Species, 200km]
High : 19
Low : 0

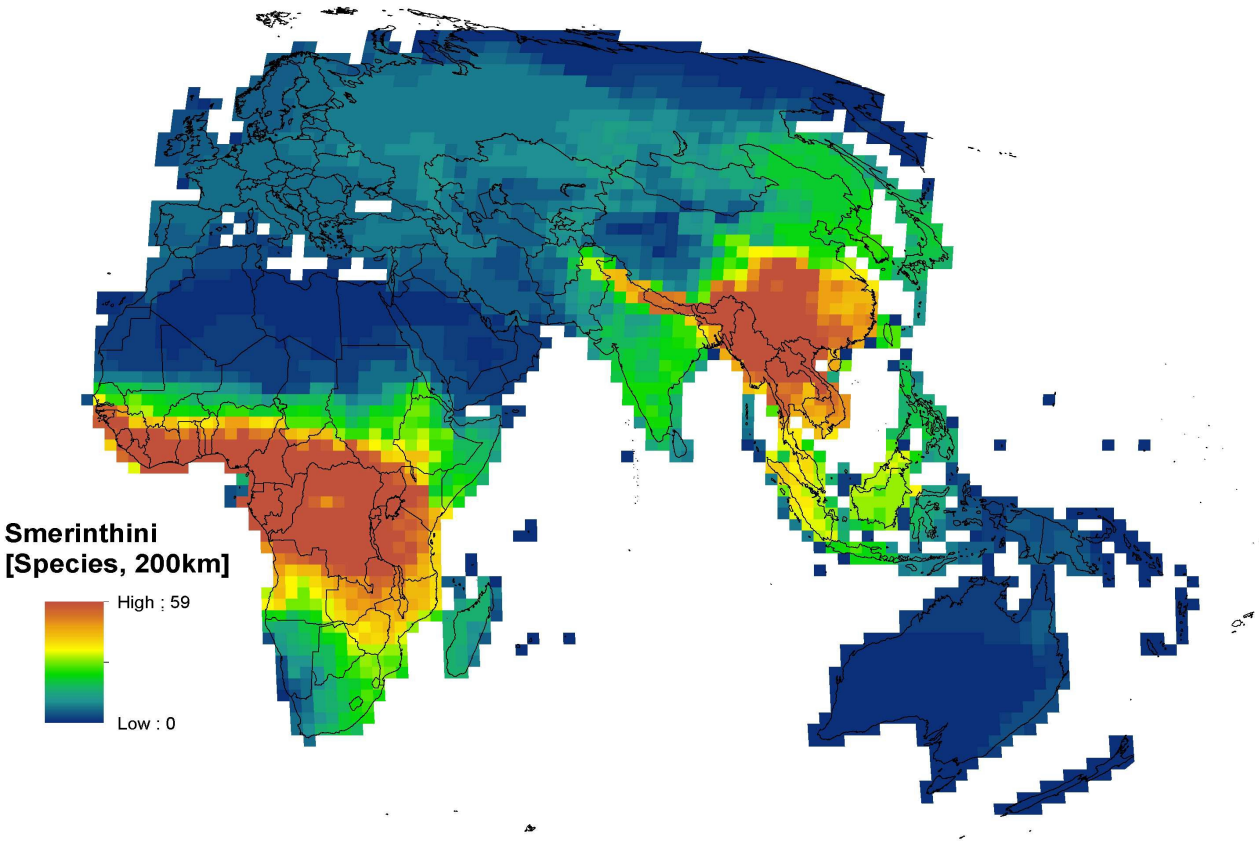
(D)


Dilophonotini
[Species, 200km]
High : 7
Low : 0

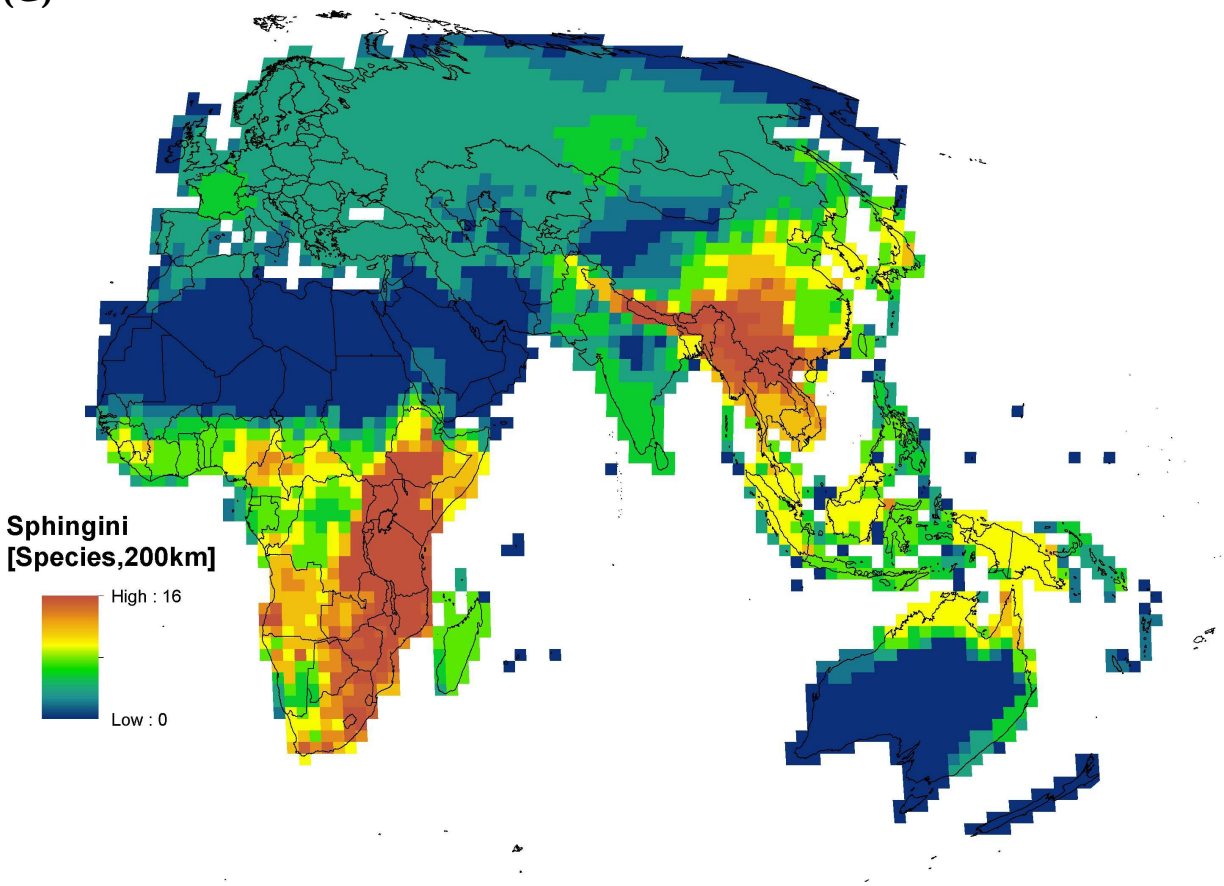
(E)


Macroglossini
[Species, 200km]
High : 106
Low : 0

(F)


Smerinthini
[Species, 200km]
High : 59
Low : 0

(G)


Sphingini
[Species,200km]
High : 16
Low : 0

(H)

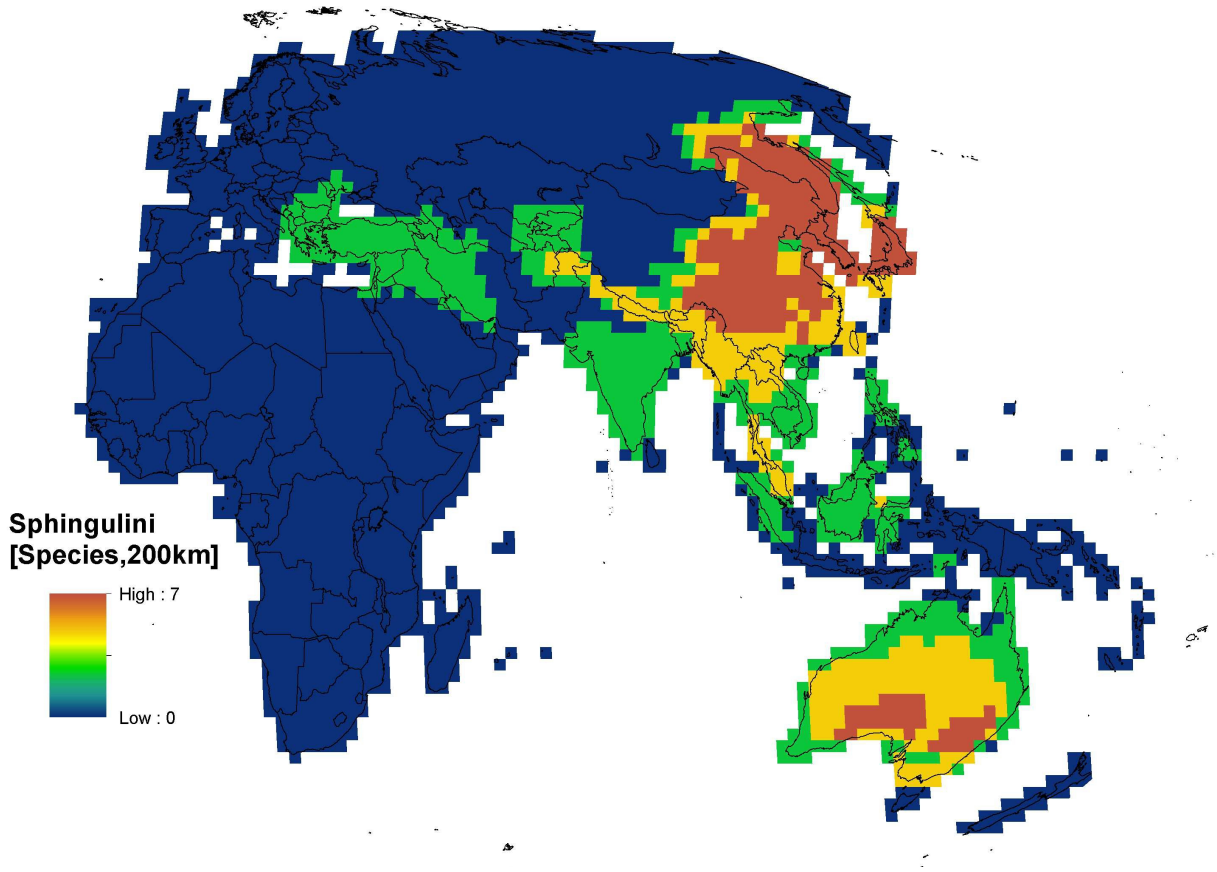
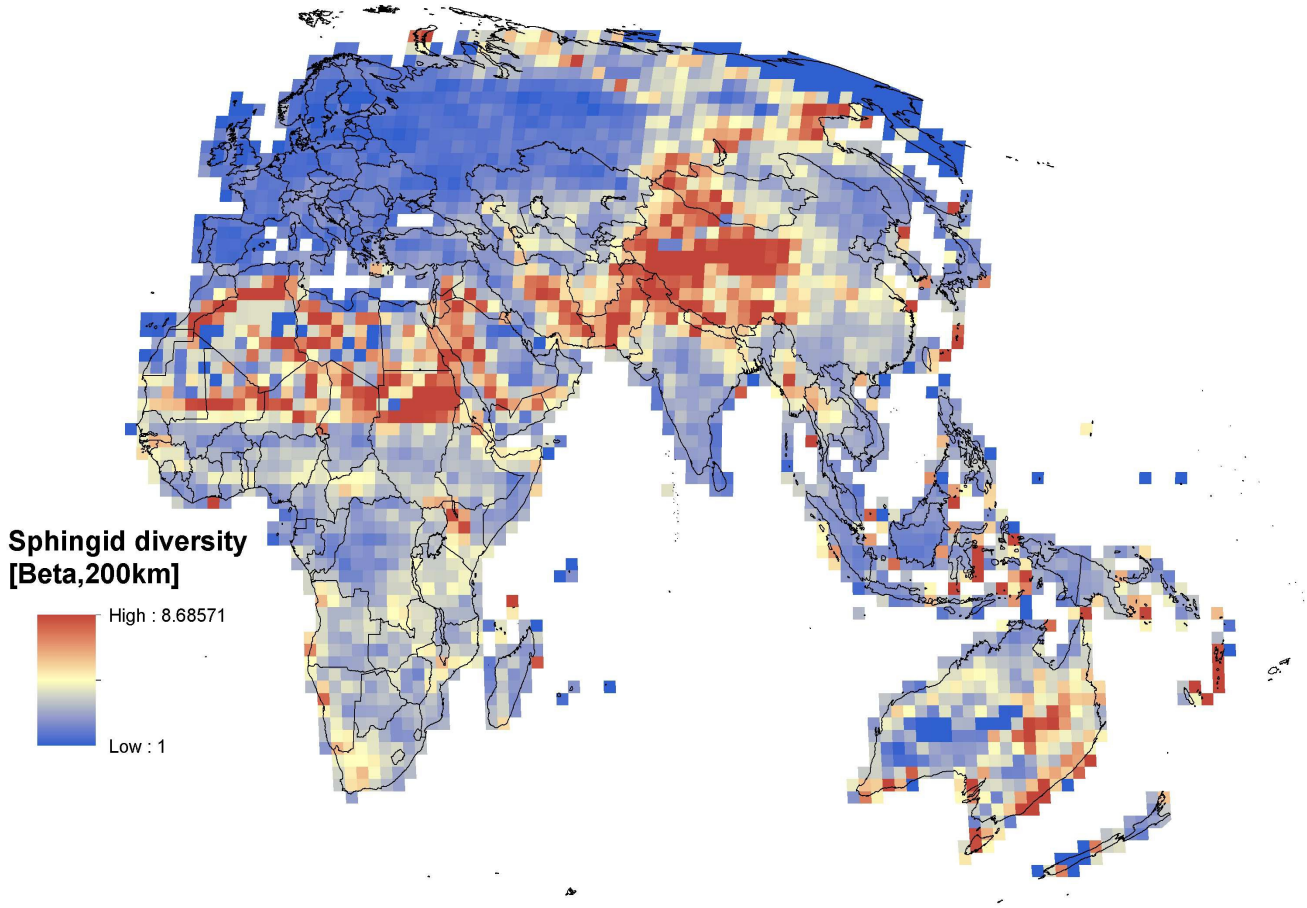


Figure 5.8. Pattern of sphingid β-diversity (Estimated as as α average / γ; 200 x 200 km cellsize).



Appendix 5.1. List of unpublished data sources (i.e., museum and private collections).

ABBREVIATION	COLLECTION
ACMU	Faculty of Agriculture, Chiang Mai University, Chiang Mai, Thailand
AddisMus	Museum Addis Abeba (Arat Kilo), Ethiopia
AES Exhibition	Amateur Entomologists' Society
AMES	Allyn Museum of Entomology, Sarasota, Florida, USA
AMNH	American Museum of Natural History, New York, New York, USA
AMSA	Australian Museum, Sydney, Australia
ANIC	Australian National Insect Collection, CSIRO, Canberra, A.C.T., Australia
ANSP	Academy of Natural Science of Philadelphia, Philadelphia, Pennsylvania, USA
ASCB	Czech Academy of Sciences, Ceske Budejovice, Czech Republic
BAUB	Beijing Agricultural University, Beijing, China
BCMU	Department of Biology Insect Collection, Chiang Mai University, Chiang Mai, Thailand
BDAF	Bermuda Department of Agriculture and Fisheries, Bermuda
BIOG	Biodiversity Institute of Ontario, Univeristy of Guelph, Guelph, Ontario, Canada
BMED	Bohart Museum of Entomology, University of California, Davis, California, USA
BMNH_suppl	Natural History Museum, London, UK, supplementary collection
BMNH	Natural History Museum, London, UK
BPBM	B.P. Bishop Museum, Honolulu, Hawaii, USA
BUMD	Bet Ushishkin Museum, Qibbutz Dan, Israel
CABF	A. Bergmann collection, Forst, Germany
CAIB	A. Iorio collection, Bologna, Italy
CAKM	A. Koslov collection, Moscow, Russia
CAKP	A. Knorke collection, Preslau, Germany
CAKV	A. Koutroumpas collection, Volos, Greece
CALO	A. Lévêque collection, Orléans, France
CAMC	A.M. Cotton collection, Chiang Mai, Thailand
CAMF	A. Martínez Fernandez collection, Ares, A Coruña, Spain
CANR	A. Napolov collection, Riga, Latvia
CAPS	A. Pessoa collection, Sobral, Ceará, Brazil
CARP	A.R. Pittaway collection, Cholsey, UK
CASD	A. Schintlmeister collection, Dresden, Germany
CASF	California Academy of Sciences, San Francisco, California, USA
CASM	A. Sochivko collection, Moscow, Russia
CAVC	A.V. Chuvilin collection, Moscow, Russia
CAZS	A. Zwick collection, Schlitz, Germany
CBDS	B. De Sousa collection, Lisbon, Portugal
CBGA	B. Guerrero Aguado collection, Gerona, Spain
CBSC	B.C. Schmidt collection, Canada
CCCH	C.C. Hoffmann collection, Mexico
CCCM	C. Congdon collection, Mufindi, Tanzania
CCEM	C.E. Meyer collection, Canberra, Australia
CCGT	C.G. Treadaway collection, Frankfurt am Main, Germany
CCLV	C. López Vaamonde collection, Spain
CCMC	C.G.C. Mielke collection, Curitiba, Paraná, Brazil
CCST	C.-S. Tzen collection, Taipei, Taiwan
CDAL	D.A. Lane collection, Atherton, Australia
CDEB	D.E. Bowman collection, Golden, Colorado, USA
CDGS	D.G. Sevastopulo collection, Mombasa, Kenya
CDHJ	D.H. Janzen ACG voucher collection, Philadelphia, Pennsylvania, USA
CDRN	D. Rolfe collection, Northfleet, UK
CdlM	J. de la Maza collection, Mexico
CEIB	Coleçâo Entomológica do Insituto Butantan, Sâo Paulo, Brazil
CENB	E.O. Núñez-Bustos collection, Martinez, Buenos Aires, Argentina

CESB	E.S. Brown collection, Muguga, Kenya (ex Carcasson, 1968)
CEvS	E. van Schayck collection, Bochum, Germany
CEWD	E.W. Diehl collection, Pematang Siantar, Sumatra, Indonesia
CFBG	F. Bénéluz collection, Bélizon, French Guiana
CFGT	F. Gil-T. collection, Granada, Spain
CFKZ	F. Karrer collection, Zofingen, Switzerland
CFSW	F.S. Schmit collection, Warmenhuizen, Netherlands
CGEK	G. Ebert collection, Karlsruhe, Germany [to C_UE?]
CGRM	G. Riedel collection, Munich, Germany
CHBy	H. Byrne collection, USA
CHFK	H. Falkner collection, Karlsruhe[?], Germany
CHHM	H.H. Hacker collection, Munich, Germany
CHNM	Croatian Natural History Mueum, Zagreb, Croatia
CHHO	H. Hjelde collection, Oslo, Norway
CHSB	H.S. Barlow collection, Genting, Malaysia
CHvM	H. van Mastrigt collection, Jayapura, Indonesia
CICI	W. Clark collection, College of Idaho, Caldwell, Idaho, USA
CIFL	Centro de Investigación Forestal de Lourizán, Pontevedra, Spain
CJAC	J.R. Alvarez Corral collection, Pointe-à-Pitre, France
CJBP	J. Bury collection, Poland
CJBT	J.B. Walsh collection, Tucson, Arizona, USA
CJBB	J. Beck collection, University of Basel, Switzerland (many specimens not collected)
CJdF	J. de Freina collection, Munich, Germany
CJFL	J.F. LeCrom collection, Bogotá, Colombia
CJFW	J. de Freina collection in T. Witt Museum, Munich, Germany
CJH?	J. Hyatt collection, USA
CJHL	J.H. Lourens collection, Lucena City, Luzon, Philippines
CJMC	J.-M. Cadiou collection, Saint-Cloud, France
CJPT	J.P. Tuttle collection, Australia
CJTM	J.T. Moss collection, Brisbane, Australia
CKJK	K.-J. Kleiner collection, Idar-Oberstein, Germany
CKKB	K. Kernbach collection, Berlin, Germany
CKMI	K. Martini collection, Ingolstadt, Germany
CKWB	K.W. Brown collection, Uganda Forestry Department, Uganda (ex Carcasson, 1968)
CLTU	Li collection, Tianjin University, Tianjin, China
CLWC	L. Willan collection, Australia
CMBB	M. Barnes collection, Maya Beach, Belize
CMcC	C. McCleery collection, Lindi, Tanzania (ex Carcasson, 1968)
CMGA	M.G. Allen collection, UK
CMNH	Carnegie Museum of Natural History, Pittsburgh, Pennsylvania, USA
CMRV	M.R. Vincent collection, Southampton, UK
CMSA	M.S. Adams collection, USA
CMSB	M. Singer collection, Bariloche, Argentina
CMSM	M.S. Moulds collection, Sydney, Australia
CMSW	M. Ströhle collection, Weiden, Germany
CNBF	Centro Nazionale per lo Studio e la Conservazione della Biodiversità Forestale "Bosco Fontana", Verona, Italy
CNCO	Canadian National Collection of Insects, Arachnids and Nematodes, Ottawa, Canada
CNUB	Institute of Natural Sciences, Colombian National University, Bogotá, Colombia
COMC	O.H.H. Mielke collection, Curitiba, Paraná, Brazil
CPAC	Brazil
CPBo	P. Boireau collection, France
CPBT	P. Basson collection, Tsumeb, Namibia
CPEA	P. Ek-Amnuay collection, Bangkok, Thailand

CPSB	P. Smetacek collection, Bhimtal, India
CPSV	P. Schmit collection, Videlles, France
CQNU	College of Life Science, Chongqing Normal University, Chongqing, China
CRBL	R.B. Lachlan collection, Brisbane, Australia
CRCK	R.C. Kendrick collection, Hong Kong, China
CRDK	R.D. Kennett collection, Bangkok, Thailand
CREL	R. Lampe collection, Nürnberg, Germany
CREW	R.E. Wells collection, Jackson, California, USA
CRJM	R.J. Murphy collection, Malawi
CRLM	R. Lichy collection, Maracay, Venezuela
CRSP	R.S. Peigler collection, San Antonio, Texas, USA
CRVP	R. Vinciguerra collection, Palermo, Italy
CRVY	R.V. Yakovlev collection, Barnaul, Russia
CSAR	S.A. Ryabov collection, Tula Exotarium, Oktyabrskaya, Russia
CSBS	S.V. Beschkow collection, Sofia, Bulgaria
CSGC	S. Gorgeev collection, Chita, Russia
CSHY	S.-H. Yen collection, Taipei, Taiwan
CSKM	S. Kovalenko collection, Moscow, Russia
CSKN	S. Kager, Nürnberg, Germany
CSLB	California State University, Long Beach, California, USA
CSNB	S. Naumann collection in MNHU, Berlin, Germany
CSNF	S. Naumann collection in FSFM, Frankfurt am Main, Germany
CSSB	S. Sáfián collection, Budapest, Hungary
CTKI	T. Klemetti collection, Imatra, Finland
CTMJ	T. Mano collection, Japan
CTWH	T.W. Harman collection, Turville Heath, UK
CUAT	University of Arizona collection, Tucson?, Arizona, USA
CUBB	Université of Brazzaville collection, Brazzaville, Congo Republic
CUDW	R. Perissinotto, University of Durban-Westville, Durban, South Africa [?]
CUIC	Cornell University, Ithaca, New York, USA
CUPP	University of Pennsylvania collection, Philadelphia, Pennsylvania, USA
CUVG	Colección de Artrópodos de la Universidad del Valle de Guatemala, Guatemala City, Guatemala
CVGM	V.A. Ganson collection, Moscow, Russia
CVOB	V.O. Becker collection, Brasília, Brazil
CVZU	V. Zolotuhin collection, Uljanovsk, Russia
CWAN	W.A. Nässig collection, Mühlheim am Main, Germany
CYBM	Y. Bezverkhov collection, Moscow, Russia
CYHC	Y.-H. Chen collection, Taipei, Taiwan
CZAP	Z. Ahmed collection, Pakistan
C_?B	?. Belyaev collection, Russia
C_?R	?. Roberts collection, California, USA
C_?S	?. Ströhle, Weiden, Germany
C_AA	A. Amarillo collection, Bogotá, Colombia
C_AB	A. BjÝrnstad collection, Norway
C_AC	A. Chaminade collection, France
C_AF	A. Floriani collection, Milan, Italy
C_AG	A. Geyer collection, Germany
C_AH	A. Hauenstein collection, Untermunkheim-Schonenberg, Germany
C_AK	A. Kingston collection
C_AL	A. Legrain collection
C_AM	A. Miyata collection, Japan
C_AP	A. Pinratana collection, Bangkok, Thailand
C_AR	A. Russell collection, El Roble de Heredia, Costa Rica
C_AS	A. Saldaitis collection, Vilnius, Lithuania

C_BS	B. Surholt collection,Germany
C_BT	B. Turlin collection, France
C_BW	B. Wenczel collection, Kloten, Switzerland
C_CC	C. Conlan collection, San Diego, California, USA
C_CD	C. Descoins collection, Bailly, France
C_CH	C. Howard collection, Zimbabwe
C_CL	C. Lemaire collection, Gorde, France
C_CS	C. Schultze collection, Germany
C_DB	D. Benyamini collection, Israel
C_DC	D. Camiade collection, Sallespisse, France
C_DH	D. Herbin, Péchabou, France
C_EF	E. Furtado collection, Diamantino, Mato Grosso, Brazil
C_EH	E. Haig collection, (ex Boorman, 1960)
C_FB	F. Brandt collection, Germany
C_FK	F. Katoh collection,Japan
C_FM	F. Meister collection, Prenslau, Germany
C_FS	F. Salvador collection, El Salvador
C_GK	G. Köhl, Trier, Germany
C_GM	G. Muller collection, Freising, Germany
C_GP	G. Ping collection, Brunei
C_GT	G. Terral collection, Rosny-sur-Seine, France
C_HB	H. Bänziger collection, Chiang Mai, Thailand
C_HF	H. Fukuda collection, Yotsukaido, Japan
C_HK	H. Käch collection, Tumbaco, Pichincha, Ecuador
C_HL	H. Lehmann collection.
C_HP	H. Politzar collection, France
C_HS	H. Schnitzler collection, Frechen, Germany
C_IR	I. Robertson collection, Ilonga, Tanzania (ex Carcasson, 1968)
C_JB	J. Boorman collection.
C_JH	J. Haxaire collection, Laplume, France
C_JJ	J. Jensen collection, Chile
C_JK	J. Kielland collection, Boroy, Norway
C_JN	J. Noble collection, Annaheim Hills, California, USA
C_JP	J. Poulard collection, Lyon, France
C_JW	J. White collection, Mexico
C_KH	K. Hories collection, Japan
C_KK	K. Kudo collection, Japan
C_KN	K. Nakao collection, Japan
C_KW	K. Wolfe collection, Escondido, California, USA
C_LA	L. Aarvik collection, Norway
C_LB	L. Beaudoin collection, Aulnay-sous-Bois, France
C_LC	B. Lalanne-Cassou collection, Paris, France
C_LK	L. Kühne collection, Potsdam-Babelsberg, Germany
C_LR	L. Racheli collection, Rome, Italy
C_LS	L. Schwartz collection.
C_MB	M. Beeke collection, Stemwede, Germany
C_MD	M. Desfontaine collection.
C_MM	M. Moosburg collection, Munich, Germany
C_MN	M. Newport collection
C_MO	M. Ochse collection, Weisenheim am Berg, Germany
C_MY	M. Yamamoto collection, Japan
C_NI	N. Ivshin collection, Moscow, Russia
C_OI	O. di Iorio collection, Buenos Aires, Argentina
C_OM	O. Mooser collection, Mexico D.F., Mexico
C_PA	P. Annoyer collection,

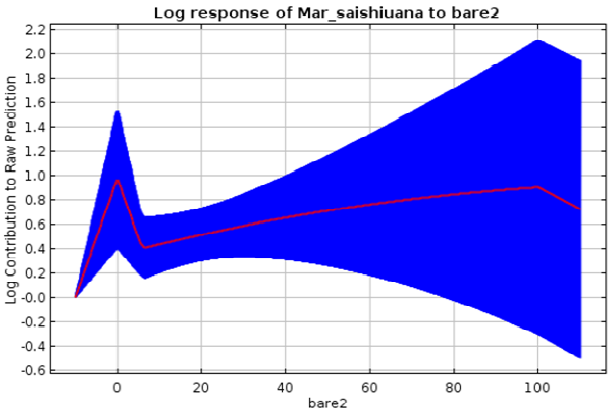
C_PB	P. Basquin collection, Yvetot-Bocage, France
C_PD	P. Darge collection, Clénay, France
C_PE	P. Eikenboom collection,, The Netherlands
C_PM	P. Moretto collection,
C_PR	P. Régnier collection, France
C_PS	P. Schütz collection,
C_PU	P. Ustjuzhanin collection, Primorskiy Kray, Russia
C_RB	R. Brechlin collection, Pasewalk, Germany
C_RG	R. Galley collection,
C_RL	Collection Rob Lachlan
C_RM	R. Minetti collection, La Ciotat, France
C_RP	R. Paul collection, Romania
C_RR	R. Rougerie collection, France
C_RW	R. Wemcken collection, Bannewitz, Germany
C_SH	S. Haapala collection, Imatra, Finland
C_SJ	S. Jakl collection,
C_SK	S. Kohll, Kayl, Luxembourg
C_SL	S. Löffler collection, Lichtenstein
C_SN	S. Naumann collection, Berlin, Germany
C_TB	T. Bouyer collection, Chênée, Belgium
C_TD	T. Decaëns collection, Rouen, France
C_TF	T. Frankenbach collection, Lindenburg, Germany
C_TK	T. Klemetti collection, Imatra, Finland
C_TM	T. Melichar collection, Pribram, Czech Republic
C_TV	T. Vaglia collection, Quebec, Canada
C_UB	U. Brosch collection, Hille, Germany
C_UE	U. Eitschberger collection, Marktleuthen, Germany
CULP	U. & L.H. Paukstadt collection, Wilhelmshaven, Germany
C_UW	U. Weritz collection, Braunschweig, Germany
C_VS	V. Sinjaev collection, Moscow, Russia
C_VV	V. Visinskas collection, Vilnius, Lithuania
C_WH	W. Harding collection, USA
C_WM	W. Mooney collection,USA
C_WS	W. Sieker collection, Madison, WI, USA
C_YD	Yu. Derzhavets collection, Saint Petersburg, Russia
C_YE	Y. Estradel collection
C_YK	Y. Kishida collection, Japan
DCRS	Dodo Creek Research Station, Honiara, Guadalcanal, Solomon Islands
DEIE	Deutsche Entomologisches Institüt, Eberswalde, Germany
DMNH	Denver Museum of Natural History, Denver, Colorado, USA
DNHP	Ditsong National Museum of Natural History, Pretoria, South Africa
DPIM	Department of Primary Industries, Mareeba, Australia
DPPA	Department of Plant Protection, Anhui Agricultural College, Hefei, Anhui, China
DPPD	Department of Primary Production, Darwin, Australia
EIHU	Entomological Institute, Hokkaido University, Sapporo, Japan
ELUR	Entomology Laboratory, College of Agriculture, University of the Ryukyus
EMEB	Essig Museum of Entomology, University of California, Berkeley, California, USA
EMJU	Entomological Museum of the Jilin Agricultural University, China
EMNH	Estonian Museum of Natural History, Tallinn, Estonia
EMPW	EcoMusée du Parc W, Diapaga, Burkina Faso
ETHZ	Eidgenössische Technische Hochschule collection, Zurich, Switzerland
FAKI	Faculty of Agronomy of Karadj, Karadj, Iran
F_Altermatt	Collection Florian Altermatt, Zurich
FIML	Fundación e Instituto Miguel Lillo, Tucumán, Argentina
FRC	Forest Research Centre of Sabah, Sepilok, Malaysia

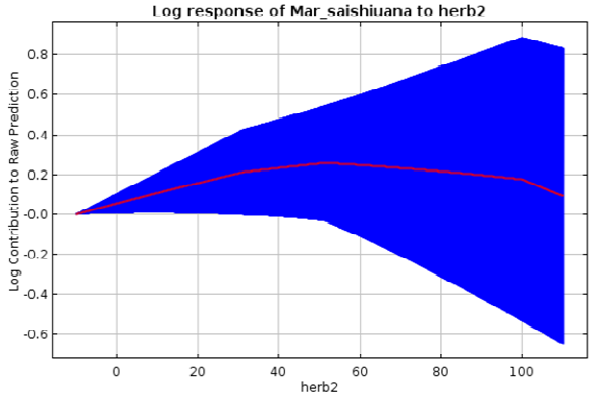
FRIM	Forest Research Institute of Malaysia, Kepong
FSCA	Florida State Collection of Arthropods, Gainesville, Florida, USA
FUUB	Entomological Collection, Federal University of Uberlândia, Minas Gerais, Brazil
GNMT	Georgian National Simon Janashia Museum, Tbilisi, Georgia
HMCM	Harvard Museum of Natural History, Cambridge, Massachusetts, USA
HMHT	Houston Museum of Natural History, Houston, Texas, USA
HMIM	Hyke Mirzayans Insect Museum, PPDRI, Tehran, Iran
HNHM	Hungarian Natural History Museum, Budapest, Hungary
HAUC	Institute of Entomology, Hunan Agricultural University, Changsha, Hunan, China
IBAJ	Inst. de Biología de la Altura, Univ. Nac. de Jujuy, San Salvador de Jujuy, Argentina
ICNM	Instituto Colombiano Nacional Museo de Historia Natural, Bogotá, Colombia
IEPE	Institute of Entomology and Plant Pathology of Evin, Tehran, Iran
IFAN	Institut Fondamental d'Afrique Noire, Dakar, Senegal
IMBV	Instituto Multidisciplinario de Biología Vegetal, Córdoba, Argentina
IMMC	Insectarium de Montréal, Montréal, Canada
INBC	Instituto Nacional de Biodiversidad (INBio), San José, Costa Rica
INCA	INCA Life Science Ltd, Chongqing, China
Inoue	Inoue-collection, housed at Natural History Museum London
INPA	Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, Brazil
IOCR	Instituto Ozwaldo Cruz, Rio de Janeiro, Brazil
IPMB	Institut für Pharmazie und Molekulare Biotechnologie, Heidelberg, Germany
IRSN	Institut Royal des Sciences Naturelles de Belgique, Brussels, Belgium
ITZA	Instituut voor Taxonomische Zoölogie, Amsterdam, The Netherlands
JB	Collection Jan Beck
JB_Basel Field Trip	Collection Jan Beck
JB_obs	Observation Jan Beck
Kailash Chandra, pers. comm.	Collection Kailash Chandra, Jabalpur, India
KMKK	Kitale Museum, Kitale, Kenya
KRSU	Kawanda Research Station collection, Kawanda, Uganda (ex Carcasson, 1968)
KSFC	Kunming Southwest Forestry College, Kunming, Yunnan, China
KSSK	Kasulu Secondary School collection, Kasulu, Tanzania
KUMB	Kasetsart University Main Collection, Bangkok, Thailand
KUSB	Kasetsart University Student Collection, Bangkok, Thailand
LACM	Los Angeles County Museum of Natural History, Los Angeles, California, USA
LCBM	La Ceiba Butterfly/Insect Museum, La Ceiba, Honduras
LEUR	Laboratoire ECODIV, Université de Rouen, Rouen, France
LI Hou Hun (Tianjin Univ.), pers. comm.	collection Li Hou Hun (Tianjin Univ., China)
LSNK	Landessammlungen für Naturkunde, Karlsruhe, Germany
MACN	Museo Argentino de Ciencias Naturales Bernardino Rivadavia, Buenos Aires, Argentina
MBGP	Mission Biologique au Gabon, Paris, France
MBUL	Museu Bocage, Museu Nacional de Historia Natural, Universidade de Lisboa, Lisbon, Portugal
MCEB	Museu Entomológico Ceslau Biezanko da Universidade Federal de Pelotas, Pelotas, Rio Grande do Sul, Brazil
MCLB	McGuire Center for Lepidoptera and Biodiversity, Gainesville, USA
M. Curran	Collection of Julian and Ray from Malawi, mediated by Micheal Curran, Zurich
MCZR	Museo Civico di Zoologia, Rome, Italy
MDBB	M. De Baar collection, Brisbane, Queensland, Australia
MECN	Museo Ecuatoriano de Ciencias naturales, Quito, Ecuador
MHLY	Muséum d'Histoire de naturelle de Lyon, Lyon, France
MHND	Muséum d'Histoire naturelle de Dijon, Dijon, France
MNHG	Muséum d'Histoire naturelle de Genève, Geneva, Switzerland
MHNL	Museo de Historia Natural "Javier Prado", Lima, Peru

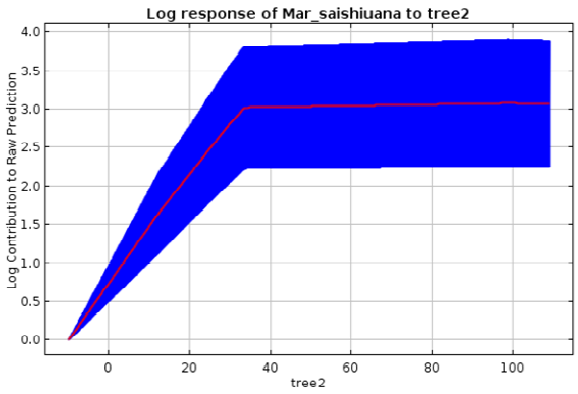
MHNM	Museo de Historia Natural de la Ciudad de México, Mexico City, Mexico
MHNT	Muséum d'Histoire naturelle de Toulouse, Toulouse, France
MichaelGeiser	Collection by Michael Geiser, Basel
MIZA	Museo del Instituto de Zoología Agrícola Francisco Fernández Yépez, Maracay, Venezuela
MNFB	Museum für Naturkunde, Freiburg im Breisgau, Germany
MNHC	Museo Nacional de Historia Natural, Santiago, Chile
MNHN	Muséum national d'Histoire naturelle, Paris, France
MNHU	Museum für Naturkunde, Leibnitz-Institut für Evolutions- und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Germany
MNRJ	Museu Nacional do Rio de Janeiro, Rio de Janeiro, Brazil
MNSD	Museo Nacional de Historia Natural, Santo Domingo, Dominican Republic
MPEG	Museu Paraense Emílio Goeldi, Pará, Brazil
MPMM	Milwaukee Public Museum, Milwaukee, Wisconsin, USA
MRAC	Musée Royal de l'Afrique Centrale, Tervuren, Belgium
MSNG	Museo Civico di Storia Naturale, Genoa, Italy
MSSK	Mlole Secondary School collection, Kigoma, Tanzania
MSUE	Michigan State University Museum, East Lansing, USA
MTWM	Museum Thomas Witt, Munich, Germany
MWTA	Makasuto Wildlife Trust, Abuko, The Gambia
MZBC	Museum Zoologicum Bogoriense, Cibinong, Indonesia
MZHF	Zoological Museum, Helsinki, Finland
NAUY	Northwestern Agricultural University, Yangling, Shaanxi, China
NBIB	National Bureau of Agriculturally Important Insects, Bangalore, India
NHLA	Natural History Museum of Los Angeles County, Los Angeles, USA
NHMC	Natural History Museum and Institute, Chiba, Japan
NHMP	Natural History Museum, Prague, Czech Republic
NHMS	Natural History Museum, Santa Cruz, Bolivia
NHMV	Naturhistorisches Museum, Vienna, Austria
NHRS	Naturhistoriska Riksmuseet, Stockholm, Sweden
NMKN	National Museums of Kenya, Nairobi, Kenya
NMNS	National Museum of Natural Sciences, Taichung, Taiwan
NMZB	National Museum of Zimbabwe, Bulawayo, Zimbabwe
NCBN	Netherlands Centre for Biodiversity Naturalis, Leiden, The Netherlands
NSMT	National Science Museum, Tokyo, Japan
NTMD	Northern Territory Museum, Darwin, Australia
NTUT	National Taiwan University, Taiwan
NYSM	New York State Museum, Albany, New York, USA
OMNZ	Otago Museum, Dunedin, New Zealand
OSUC	Oregon State University collection, Corvallis, Oregon, USA
OUMO	Oxford University Museum of Natural History, Oxford, UK
PDBC	Project Directorate of Biological Control, Bangalore, India
PLAU	Insect Coll., PLA University of Agricultural & Animal Sciences, Changchun, Jilin, China
PMNH	Yale Peabody Museum of Natural History, New Haven, Connecticut, USA
PSUH	Prince Of Songkhla University, Hat Yai, Thailand
QENP	Lock collection, Queen Elizabeth National Park, Uganda
QMBA	Queensland Museum, Brisbane, Australia
RAWR	Riyadh Agricultural and Water Research Centre, Riyadh, Saudi Arabia
RCAH	A.K. Hundsdörfer research collection, Dresden, Germany
Roger Kitching, pers. comm.	Collection Roger Kitching, Brisbane, Australia
RMBR	Raffles Museum of Biodiversity Research, National University of Singapore, Singapore
ROMT	Royal Ontario Museum, Toronto, Canada
SabahParksColl	Collection of Sabah Parks at Kinabalu Park Headquarter, Sabah, Malaysia
SAMZ	South Africa Museum, Cape Town, South Africa

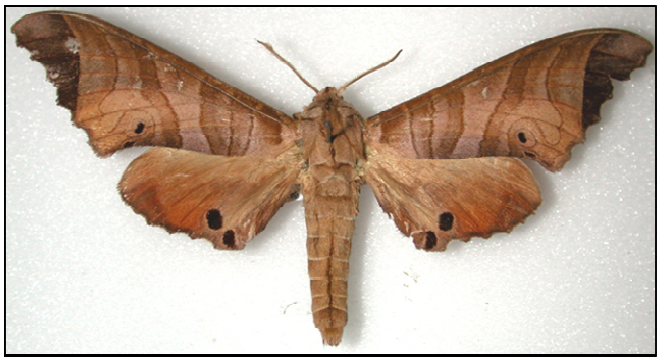
SCAU	South China Agricultural University, Guangzhou, China
SDNH	San Diego Natural History Museum, San Diego, California, USA
SIES	Shanghai Institute of Entomology collection, Shanghai, China
SMFL	Forschungsinstitüt Senckenberg, Frankfurt am Main, Germany
SMNK	Staatliches Museum für Naturkunde in Karlsruhe, Karlsruhe, Germany
SMNS	Staatliches Museum für Naturkunde in Stuttgart, Stuttgart, Germany
SMTD	Staatliches Museum für Tierkunde in Dresden, Dresden, Germany
SMUA	E.H. Strickland Entomological Museum, University of Alberta, Edmonton, Alberta, Canada
SSUS	Laboratory of Animal Systematics and Faunistics, Samara State University, Samara, Russia
SZMN	Siberian Zoological Museum, Novosibirsk, Russia
TAMU	Dept of Entomology collection, Texas A&M University, College Station, Texas, USA
TAUI	Tel Aviv University, Tel Aviv, Israel
TFRI	Taiwan Forestry Research Institute, Taipei, Taiwan
TMET	Texas Museum of Entomology, Pipe Creek, Texas, USA
UABC	Universidad Autonoma de Baja California Norte, Ensenada, Mexico
UCIN	University College Ibadan collection, Ibadan, Nigeria
UFPC	Padre Jesus Moure collection, Universidade Federal do Paraná, Curitiba, Brazil
UGAG	University of Georgia collection, Athens, Georgia, United States
UGIC	University of Guam Insect Collection, Mangilao, Guam
UHIM	University of Hawaii Insect Museum, Honolulu, Hawaii, USA
UKMB	Universiti Kebangsaan Malaysia, Bangi, Malaysia
UMCE	Inst. de Ent. de la Univ. Metropolitana de Ciencias de la Educación, Santiago, Chile
UMCP	University of Maryland, College Park, Maryland, USA
UMZC	University Museum of Zoology, Cambridge, Cambridge, UK
UNAM	Universidad Nacional Autónoma de México, Mexico D.F., Mexico
UNSM	Museo Historia Natural, Universidad Nacional Mayor de San Marcos, San Marcos, Peru
UOPO	University of Osaka Prefecture, Osaka, Japan
UOSC	Universidad de Oriente collection, Santiago de Cuba, Cuba
UPJP	Dept of Syst. & Ecol., Universidade Federal da Paraíba, Joao Pessoa, Paraíba, Brazil
USAO	Museum of Natural History, University of Science and Arts of Oklahoma
USCC	Zoological Museum, University of San Carlos, Cebu, Philippines
USNM	United States National Museum, Washington, D.C., USA
UWIT	University of the West Indies, Trinidad, Trinidad & Tobago
WAMP	Western Australia Museum, Perth, Australia
ZFMK	Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany
ZIMH	Zoologisches Institut und Zoologisches Museum, Hamburg, Germany
ZISP	Zoological Institute, St Petersburg, Russia
ZMAN	Zoölogisch Museum Amsterdam, Amsterdam, The Netherlands
ZMKU	Zoological Museum of Kiev University, Kiev, Ukraine
ZMRI	Zoological Museum, Rome, Italy
ZMUC	Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
ZSBS	Zoologische Staatssammlung des Bayerischen Staates, München, Germany

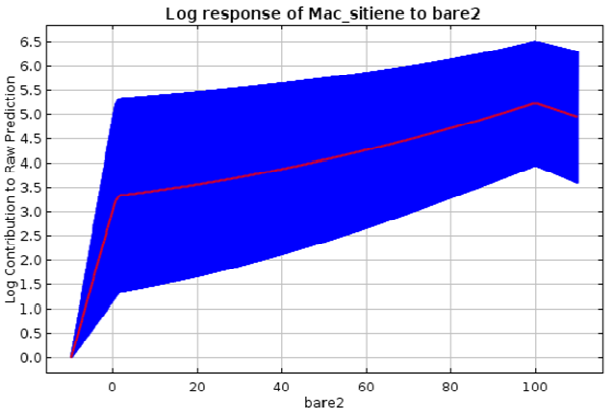
Appendix 5.2. Examplary response curves for a representative of the tribe macroglossini and smerinthini. Studies have shown that macroglossini species tend to be

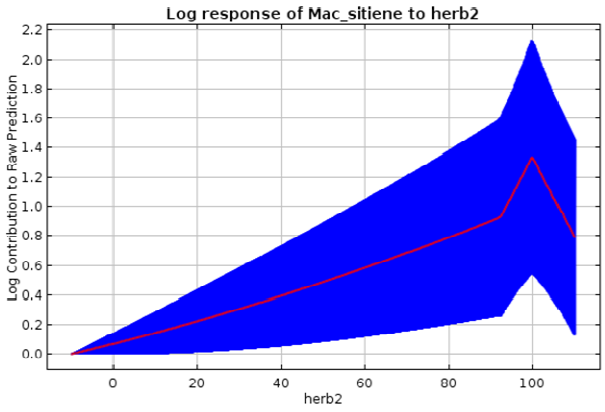


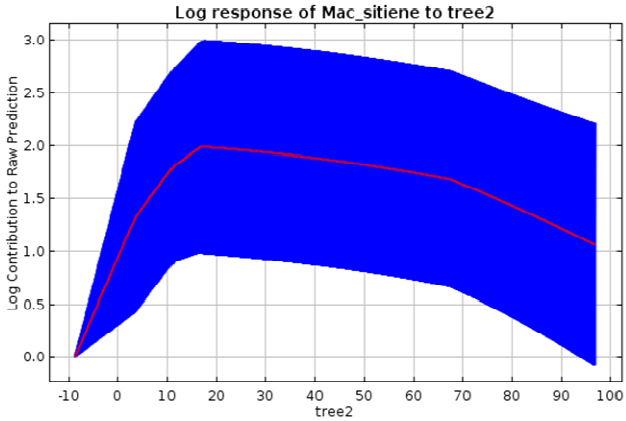


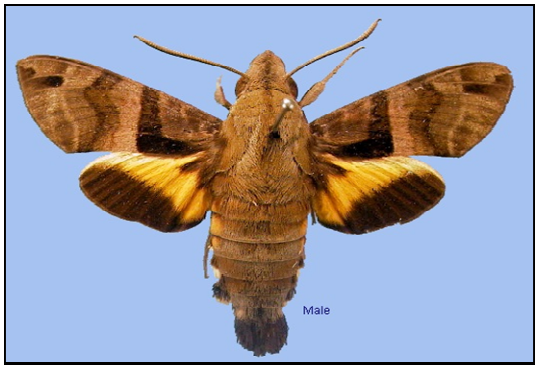












CHAPTER 6

Projecting the potential invasion of the Pink Spotted Hawkmoth (*Agrius cingulata*)

Liliana Ballesteros-Mejia[1*], Ian J. Kitching[2], Jan Beck[1*]

[1] University of Basel, Department of Environmental Science (Biogeography), St. Johanns-Vorstadt 10, 4056 Basel, Switzerland

[2] The Natural History Museum, Department of Entomology, Cromwell Road, London SW7 5BD, UK.

*Author for correspondence: Tel.: +41-2670810, E-mail:jan.beck@unibas.ch

Published in: *International Journal of Pest Management. (2011) 57:2,153-159.*

Abstract

Agrius cingulata (Lepidoptera, Sphingidae) is widespread in the Americas, but has recently begun to spread into Africa. In parts of its native range, the species is a pest on sweet potato, which is also an important crop plant in Africa. We used two types of ecological niche models, based on native distribution records and climate and vegetation structure data, to estimate which regions of Africa are potentially suitable for the species to become established. The results show that, under the simplifying assumption that the species will occupy the same ecological niche in Africa as in its native range, *A. cingulata* may find suitable habitat across wide stretches of sub-Saharan Africa. We conclude that early monitoring programs of the spread and actual status of the species in Africa may be worthwhile.

Keywords: BIOMOD, Climate niche, Ecological niche model, *Ipomoea*, Maxent, Species distribution, Sphingidae, sweet potato.

6.1. Introduction

During the past two centuries, and as a result of human traffic or other activities, many organisms have succeeded in crossing biogeographical dispersal barriers. Some of these non-native taxa successfully established and became invasive (Didham et al. 2005). Such invasive species have become an issue of concern and interest in fields such as evolutionary biology (Cody & Overton 1996), conservation (Ricciardi 2003; Gurevitch & Padilla 2004), economics (Pimentel 2002) and agronomy (Mullin et al. 2000; Jordan et al. 2008). Invaders can impact the ecological community to which they have been introduced by altering ecosystem functioning and threatening native biodiversity (Strayer et al. 2006) and they can become agricultural pests (Chalfant et al. 1990; for recent examples of invading pest insects see Desneux et al. 2010; Haack et al. 2010; Ragsdale et al. 2011). Becoming an invader or pest is a multi-step process (Colautti & MacIsaac 2004) that requires a number of key factors, among them colonization opportunities and the ecological suitability of the new habitat. Successful invaders have also been shown to share certain traits (Sutherland 2004), such as high dispersal ability and an opportunistic lifestyle.

Ecological Niche Models (hereafter ENMs) are a set of very powerful tools that have been used to predict the potential spread of species into new areas (Elith & Leathwick 2009). Combining geographical and environmental information, ENMs assess one step of the complex process of invasion, i.e. the environmental suitability of the to-be invaded landscape (Peterson & Vieglais 2001; Thuiller et al. 2005). ENMs cannot, in their current standard applications, consider potential evolutionary change (i.e., niche shifts; Broennimann et al. 2007; Hortal et al. 2010) or dispersal limitations (Bomford et al. 2009), although progress is being made to integrate these aspects into ENMs (Dirnböck & Dullinger 2004, Elith et al. 2010).

The Pink Spotted Hawkmoth, *Agrius cingulata* [Fabricius, 1775], is a large (9.5-12 cm wingspan) member of the lepidopteran family Sphingidae. It is widespread with a native range across the Americas (Figure 6.1). It is also known from the Falklands, Galapagos and Hawaii, although clear evidence is lacking as to whether these distant archipelagos were colonized naturally or with human assistance (D. Rubinoff, pers. comm.). The species performs regular summer migrations to the North (e.g., to Canada) and possibly also to the South (Figure 6.1).

However, thirty years ago, *A. cingulata* was reported from the Cape Verde Islands, off the west coast of Africa (Bauer & Traub 1980), and the species is now firmly established on several of the islands, most notably Sao Filipe and Santo Antao. The species might well have been transported there by transatlantic trading of its larval host plant. The caterpillars of *A. cingulata* feed on sweet potatoes (*Ipomoea batatas*), as well as other Convolvulaceae, and in its native range, *A. cingulata* is

regarded a pest of sweet potato (Talekar 1987). Then, in April 2002, a female *A. cingulata* was captured on mainland Africa, in the vicinity of Man, Ivory Coast, and is now in the collection of Tomas Melichar (Příbram, Czech Republic). This record may well indicate the start of a colonization of the African continent. A single specimen of *A. cingulata* has also been recorded in Portugal (Marabuto 2006), but this is considered to be a non-breeding vagrant and is not considered further here.

Sweet potatoes are an economically important, widely grown crop in Africa (Horton 1988; Woolfe 1992). The spread of a new herbivore known to be an agricultural pest on the same crop in its native range is therefore of major concern, and it will be useful to observe closely its spread and pest status in Africa. Towards this aim, we here present ENM analyses of suitable regions for *A. cingulata* in Africa, so as to predict to where the moth may spread, as well as defining the sets of environmental conditions that might enhance such an invasion. These predictions may be useful for identifying endangered regions and for targeting early field surveys or monitoring programs.

6.2. Methods

To identify suitable habitats for *A. cingulata* in Africa, we fitted models that successfully predicted the native New World range of the species. These models were then projected onto environmental data for Africa, assuming the colonizing populations will occupy the same ecological niche as the native populations (see Discussion).

6.2.1. Species records

We had available 361 presence records (Figure 6.1) of *A. cingulata* from the collections of the Natural History Museum (London), published literature, the Global Biodiversity Information Facility database (www.gbif.org; accessed May 2010) and the Barcode of Life Database (www.boldsystems.org; accessed May 2010). We only included in the analysis records from confirmed or very likely permanent and breeding populations, and excluded summer migrant populations and vagrant specimens (Figure 6.1). We georeferenced records to a precision of 0.01° latitude/longitude wherever this was possible. Because modelling (see below) was carried out with a raster grid resolution of 2.5 arc minutes (≈ 5 x 5 km), only 235 of these records were situated in unique grid cells within the native range. Only these entered the models as independent records.

6.2.2. Environmental variable selection

We used climate data from the WorldClim database (version 1.4; www.worldclim.org; accessed Feb. 2009). These data are based on average monthly weather conditions recorded from 1950-2000. Furthermore, we used MODIS vegetation data (http://glcf.umiacs.umd.edu/data/vcf; accessed Feb. 2009; three layers, indicating percent coverage of herbs, trees, and bare ground, respectively; see Buermann et al. 2008 for use of such data in ENM). In the native, breeding range, *A. cingulata* occurs across a broad environmental gradient (e.g., latitudinal extent 36°S – 42°N).

For initial selection of relevant environmental variables, we used 14 variables that we considered to be of potential importance (see Appendix). To select the best set of environmental variables for niche modelling, we ran in total 15 models with different combinations of variables (see Appendix). We assessed their quality based on the area under their receiver-operating characteristics of a cross-validation from repeated runs of the model (AUC, a standard metric of model fit; Marzban 2004) and chose the best combination of variables (Table 6.1) for further modelling.

We carried out this initial modelling with a maximum entropy model (software Maxent 3.3.2, Phillips & Dudík 2008). In comparison with other methods, this method has proved to be a very effective algorithm for modelling species distribution with presence–only data (Elith & Leathwick 2009).

We used several methods to predict the potential distribution of *Agrius cingulata*. We applied the Maxent model (see above), but additionally we also used an ensemble forecasting technique (Araújo & New 2007), i.e. BIOMOD (http://r-forge.r-project.org/projects/biomod/; accessed May 2010; Thuiller et al. 2009). Ensemble forecasting is based on the idea that by using several modelling techniques and calculating a measure of central tendency (mean or median) from the whole spectrum, the range of projections can be evaluated and a more reliable prediction can be made. BIOMOD applied seven different algorithms as a model ensemble (Thuiller et al. 2009; see also caption of Table 6.2), but it did not include Maxent. A consensus map of these methods was produced using the median as measure of central tendency. The median is less influenced than the mean by extreme output values of the different algorithms, and has therefore been suggested to be more reliable (Araújo & New 2007). As in initial variable selection, AUC values were used for evaluating model quality. While there are some known problems with this measure (Lobo et al. 2008), it carried the advantage that it is independent of the choice of a threshold converting continuous "probability of occurrence" model output into a categorical presence-absence prediction (Pearce & Ferrier 2000). It is currently still unclear how to make best (i.e., objective and informed) choices on such thresholds (e.g. Liu et al. 2005; Jiménez-Valverde & Lobo 2007). We ran two repetitions of the BIOMOD models to assess the consistency of results.

Figure 6.1. Distribution of presence records for *A. cingulata* in both its native and invasive range. Only native breeding records (black crosses) were used for modelling (see Methods). The record from Portugal was considered a non-breeding vagrant.

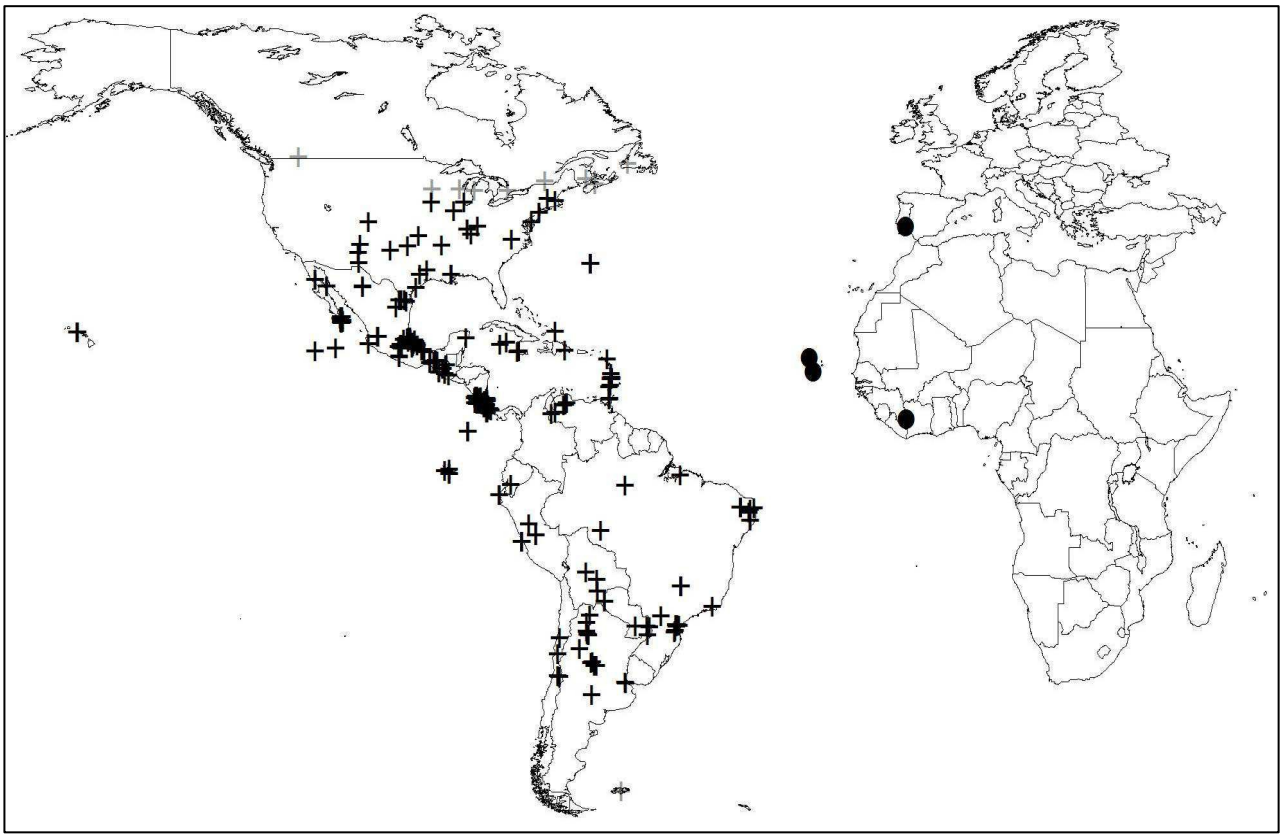


Table 6. 1 Relative contribution of the environmental variables to the best maximum entropy (Maxent) model.

Variable	% Contribution
Annual temperature (antemp)	42.9
Mean temperature coldest quarter (mtcq)	15.1
Herb cover (herb)	9.0
Mean diurnal range (mdr)	8.2
Annual precipitation (anprec)	7.9
Precipitation driest quarter (pdq)	4.3
Precipitation coldest quarter (pcq)	3.7
Mean temperature wettest quarter (mtwq)	2.8
Tree cover (tree)	2.3
Mean temperature warmer quarter (mthq)	2.2
Bare ground cover (bare)	0.8
Mean temperature driest quarter (mtdq)	0.7

6.3. Results

The best Maxent model (Figure 6.2) was of very good predictive quality (AUC = 0.930). It contained 12 (out of 14 tested) environmental variables (Table 6.1). We used this set of variables for further building of the ensemble forecasting models.

Different variables had widely different relative importance in the different models of the ensemble (Table 6.2). Figure 6.3 shows predictions from BIOMOD ensemble forecasting. Some modelling methods of the ensemble performed better than others (Table 6.3). The method that performed best among the seven model algorithms was Random Forest (RF). Its predictive performance was very good in both repetitions (AUC = 0.978 and 0.982, respectively).

Comparisons of variable contributions to the Maxent-model (Table 6.1) and BIOMOD models (Table 6.2) revealed that temperature (either annual or coldest quarter) was the most influential variable in most models. Inspection of response curves in the Maxent model revealed a steep rise in probability of occurrence once annual mean temperature was higher than ca. 8°C, whereas a coldest-quarter temperature of ca. 15°C led to a unimodal peak in modelled probability. However, inconsistencies across models with regard to the importance of other variables (Table 6.1 and 6.2) complicated conclusions on the biological relevance of other variables.

In the native range, the models agree with the wide distribution that has been reported for this species (Figures 6.2 and 6.3). Occurring throughout the Neotropics and almost all the adjacent subtropical regions, it is particularly widespread within Central America and the northern part of South America.

Output of both model approaches (Figures 6.2 and 6.3) consistently predicted that *A. cingulata* can be expected to find suitable. areas (probability of occurrence >0.5) across large parts of southern and eastern part of Africa, Madagascar, and along the Mediterranean coast in Northern Africa. The rainforested Congo Basin, on the other hand, does not seem to provide prime habitat for the species. Disagreement among model approaches, however, was found particularly in the very arid zones (i.e., Sahara & Namib deserts), where BIOMOD predicted considerable higher (Δ >0.2) probability of occurrence than Maxent, whereas Maxent placed higher probabilities, in comparison to BIOMOD, on parts of the Sahel (data not shown in detail). GIS-compatible model outputs are available at www.biogeography.unibas.ch/beck.

Figure 6.2. Potential distribution of A. cingulata in its Native American range and in Africa according to the maximum entropy model (Maxent)**.** All Ecological Niche Models were based on native (i.e., American) distribution records only (see Methods).

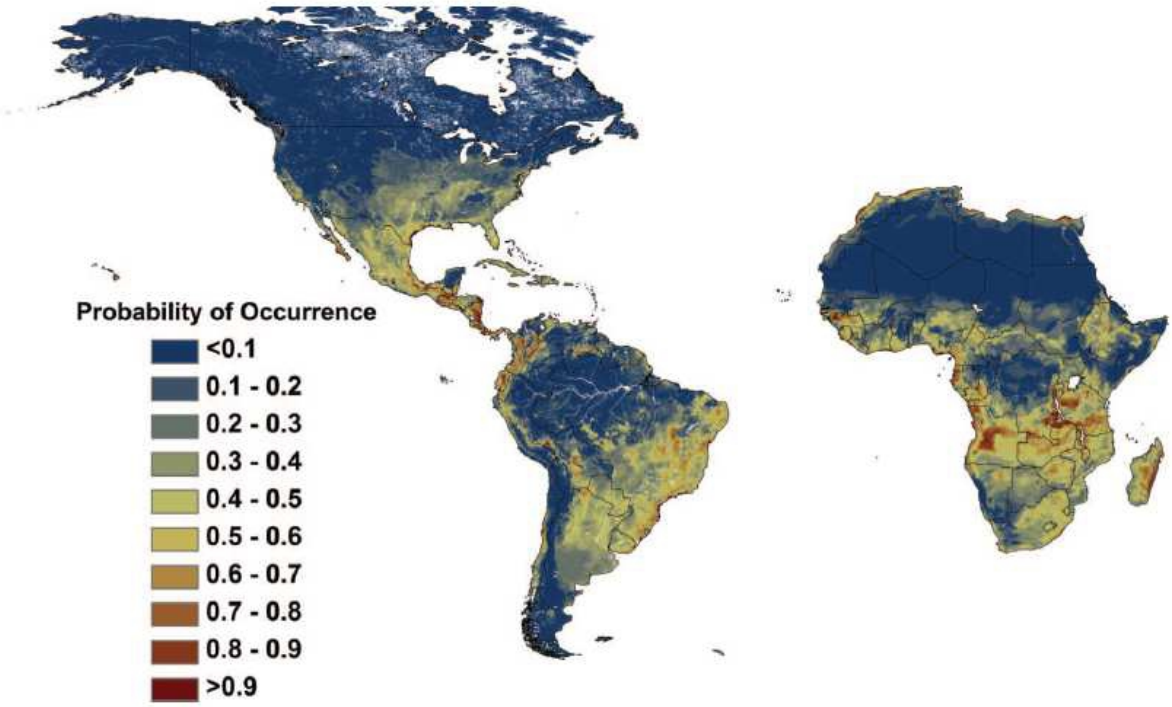


Figure 6.3. Median probabilities of occurrence from an ensemble of the seven model used in BIOMOD. The colour scheme is identical to that in Figure 6.2. All Ecological Niche Models were based on native (i.e. American) distribution records only (see section 6.2).

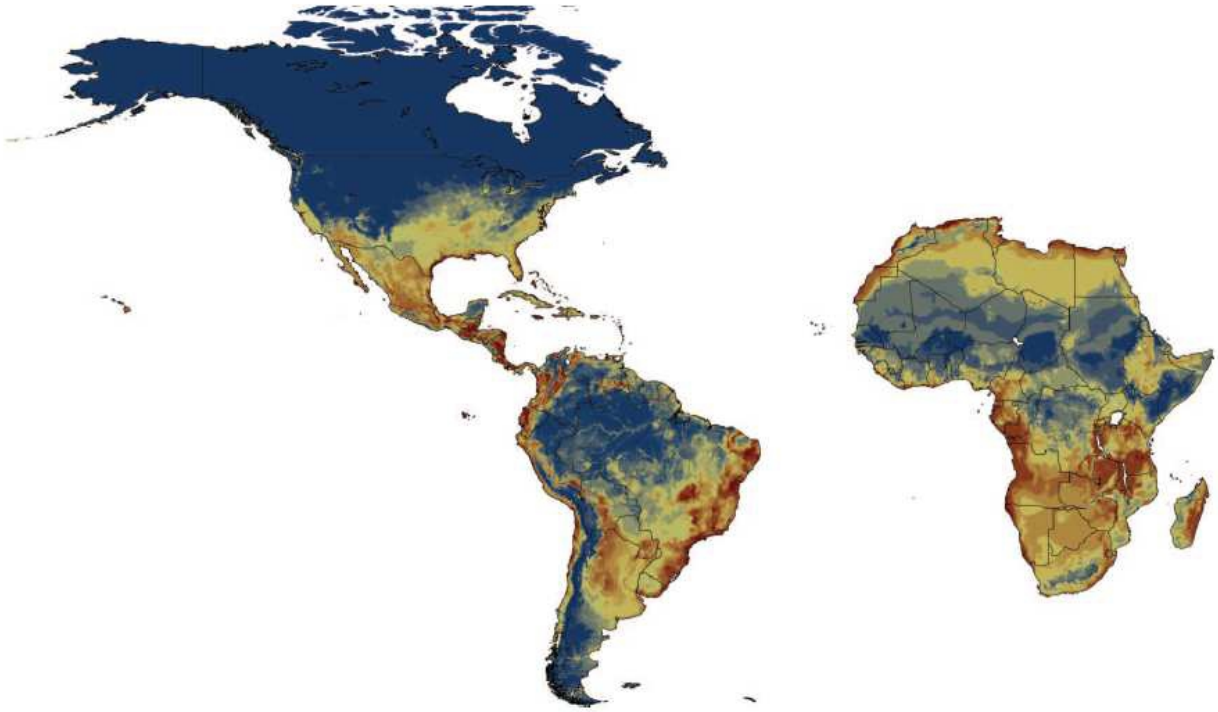


Table 6. 2. Relative importance of the variables within the different methods used by BIOMOD. See Table 6.1 for acronyms of variables (first row). Acronyms of model types (first column) are: CTA = Classification Tree Analysis, GAM = Generalized Additive Model, GBM = Generalized Boosting Model, GLM = Generalized Linear Model, MARS = Multiple Adaptive Regression Splines, RF = Random Forest, SRE = Surface Range Envelope.

Method	**mtcq**	**anprec**	**pdq**	**pcq**	**antem**	**mdr**	**mtwq**	**mtdq**	**herb**	**tree**	**mthq**	**bare**
CTA	0.374	0.120	0.011	0.103	0.456	0.232	0.097	0.142	0.058	0.094	0.319	0.041
GAM	0.750	0.079	0.039	0.034	0.000	0.110	0.13	0.000	0.000	0.050	0.000	0.025
GBM	0.317	0.012	0.010	0.005	0.117	0.040	0.010	0.010	0.010	0.040	0.000	0.005
GLM	0.209	0.110	0.037	0.039	0.133	0.145	0.000	0.000	0.000	0.071	0.000	0.000
MARS	0.694	0.463	0.180	0.146	0.569	0.179	0.000	1.014	0.000	0.024	0.653	0.000
RF	**0.508**	**0.040**	**0.015**	**0.023**	**0.128**	**0.128**	**0.048**	**0.017**	**0.019**	**0.026**	**0.042**	**0.003**
SRE	0.013	0.047	0.017	0.010	0.006	0.023	0.055	0.008	0.010	0.030	0.018	0.019

Table 6.3. Performance measure AUC (see Methods) of the modelling methods used in BIOMOD (see Table 6.2 for acronyms). The column *Cross Validation* shows the predictive accuracy according to the criteria for evaluation. *Sensitivity* shows the true positive fraction of the model (correct predicted presences) and *Specificity* shows the true negative fraction of the model (correct predicted absences). Two repetitions (Rep. 1, 2) of BIOMOD runs were carried out (see Methods). There is no AUC-evaluation available for SRE as it does not provide probability values but only the presence absence of the species (Busby 1991).

Model	***Cross Validation***		***Sensitivity***		***Specificity***	
	Rep1	**Rep2**	**Rep1**	**Rep2**	**Rep1**	**Rep2**
CTA	0.885	0.916	90.00	91.43	91.6	91.6
GAM	0.932	0.957	88.86	88.86	88.8	88.9
GBM	0.934	0.944	88.00	88.86	87.8	89.1
GLM	0.938	0.958	88.57	87.14	88.6	87.0
MARS	0.939	0.960	88.00	88.86	88.0	88.9
RF	0.978	0.982	98.29	100.0	98.3	100.0

6.4. Discussion

Our data (Figures 6.2 and 6.3) show that there is a considerable potential for *A. cingulata* to spread through large parts of Africa due to the suitability of environmental conditions for the species. One of its host plants, sweet potato, is a widely grown crop across Africa, which may additionally increase the species' chances for a rapid and wide spread. Sweet potato is the main source of household income (>50%) in many rural communities in East African countries (Kenya, Uganda, Tanzania and Rwanda; Low 1997). These regions have been highlighted by our results (Figures 6.2 and 6.3) as highly suitable climatically for *A. cingulata*. Hence, we can predict the risk of an emergent pest on an important economic crop.

It may well be possible that the current invasive status of *A. cingulata* is underestimated by recorded data as the species could easily be confused in Africa with congeneric *A. convolvuli*, a widespread paleotropical taxon (Beck & Kitching 2004-2008). The taxa look similar and may often have been identified based solely on the sampling locality. *A. convolvuli* is also regarded as a pest of sweet potato foliage, causing great damage especially during the crop season in summer, when three generations can occur (Talekar 1987).

Our range predictions, however, are based entirely on matching native environmental requirements with conditions in Africa. Ideally, one would test the fit of the projected invasive range using record data from the newly colonized region, and investigate niche evolution or other confounding factors by comparing models fitted to native vs. invasive presence records (Randin et al. 2006). Unfortunately, in the case of *A. cingulata*, there are too few African records yet to make this more rigorous approach feasible.

Furthermore, climatic input data were based on 50 year averages from the recent past, while climatic conditions can be assumed to change. Range predictions can be derived from applying ENM to future climatic scenarios (e.g., Settele et al. 2008), but this would further increase uncertainty in predictions, e.g., with regard to the correctness of future climate scenarios, or due to novel combination of climatic variables (see Elith et al. 2010 for recent methodological advance).

Temperature turned out to be the most important predictor of the distribution of *A. cingulata*, and temperature minima, in particular, could be crucial in determining the places where eggs are laid or where larval and pupal development could successfully occur. Coldest quarter temperatures have been shown to affect overwintering survival in butterflies (Hill et al. 2003). In another sphingid, *Hyles lineata*, caterpillar activity can be temperature–dependent, and investigated specimens ceased to move and feed below some threshold (Casey 1976).

While the two different modelling approaches employed in this study (Maxent and BIOMOD ensemble forecasting) broadly agreed in predicted patterns, we also noticed disagreements in some details (see Results). Methods such as generalized additive models or generalized boosting models, calculated within BIOMOD, are known to be close-fitting to the data, making them more sensitive to the sample peculiarities and therefore more prone to overfitting.

Discrepancies between realized (affected by biotic interactions) and fundamental niches may be another source of error when projecting invasive ranges (Soberón & Nakamura 2009), as biotic interactions may change in new biogeographic regions. Unfortunately, such discrepancies can only be tested experimentally and not with an ENM approach. Nevertheless, our results on the native range (America) reflected the recorded distribution of the species well, indicating a sufficiently good choice of fundamental niche dimensions as predictor variables. In the absence of further information, our projections of these models onto Africa are the best estimate for highlighting regions at risk.

6.5. Conclusions

Based on climatic match, we conclude that *Agrius cingulata* could spread widely across the African continent, creating the potential for causing major damage on a widely grown crop (i.e., sweet potato). Early, careful surveys and monitoring in regions predicted as suitable would help to assess the actual occurrence and status of populations (i.e., migrant/vagrant or breeding). This may be advisable as a first step to recognize an emerging pest early enough to take suitable measures against its potential agronomic effects. Certainly, records of *Agrius* hawkmoths from Africa can no longer be simply assumed to be the native *A. convolvuli*.

6.6. Acknowledgements

We thank Tomas Melichar for alerting us to the presence of *A. cingulata* on the African mainland and thus providing the initial impetus for this study. The study benefited from the AMNH workshop in species distribution modelling (i.e., R.G. Pearson and S.J. Philips). Two anonymous reviewers provided valuable comments on an earlier draft of the manuscript. We received financial support from the Swiss National Science Foundation (SNF, grant no. 31003A_119879).

6.7. References

Araújo, M. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in Ecology and Evolution*, **22**, 42-47.

Bauer, E. & Traub, B. (1980) Zur Macrolepidopterenfauna der Kapverdischen Inseln. Teil 1. Sphingidae und Arctiidae. *Entomologische Zeitschrift*, **90**, 244-248.

Beck, J. & Kitching, I. (2004-2008) *The Sphingidae of Southeast-Asia (incl. New Guinea, Bismarck & Solomon Islands), version 1.5*. Website at http://www.sphin-sea.unibas.ch/ (accessed August 2010).

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A. & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 1-9.

Buermann, W., Saatchi, S., Smith, T.B. Zutta, B.R., Chaves, J.A.,Milá, B. & Graham, C.H (2008) Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography*, **35**, 1160–1176.

Busby, J. (1991) BIOCLIM-A bioclimate analysis and prediction system. *Plant Protection Quarterly*, **6**, 8–9.

Casey, T. (1976) Activity patterns, body temperature and thermal ecology in two desert caterpillars (Lepidoptera: Sphingidae). *Ecology*, **57**, 485-497.

Chalfant, R., Jansson, R. & Seal, D. (1990) Ecology and management of sweet potato insects. *Annual Review of Entomology*, **35**, 157-180.

Cody, M. & Overton, J. (1996) Short-term evolution of reduced dispersal in island plant populations. *Journal of Ecology*, **84**, 53-61.

Colautti, R.I. & MacIsaac, H.J. (2004) A neutral terminology to define 'invasive' species. *Diversity and Distributions*, **10**, 135-141.

Desneux, N., Wajnberg, E., Wyckhuys, K.A.G., Burgio, G., Arpaia, S., Narváez-Vasquez, C.A., González-Cabrera, J., Catalán Ruescas, D., Tabone, E., Frandon, J., Pizzol, J., Poncet, C., Cabello, T. & Urbaneja, A. (2010). Biological invasion of European tomato crops by *Tuta absoluta*: Ecology, history of invasion and prospects for biological control. *Journal of Pest Science*. **83**, 197-215.

Didham, R., Tylianakis, J. & Hutchison, M. (2005) Are invasive species the drivers of ecological change? *Trends in Ecology and Evolution*, **20**, 470-474.

Dirnböck, T. & Dullinger, S. (2004) Habitat distribution models, spatial autocorrelation, functional traits and dispersal capacity of alpine plant species. *Journal of Vegetation Science*, **15**, 77-84.

Elith, J., Kearney, M. & Phillips S. (2010) The art of modelling range-shifting species. *Methods in Ecology & Evolution*, **1**, 330–342.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677-697.

Gurevitch, J. & Padilla, D.K. (2004) Are invasive species a major cause of extinctions? *Trends in Ecology and Evolution*, **19**, 470-474.

Haack, R.A., Herard, F., Sun, J.H. & Turgeon, J.J. (2010). Managing invasive populations of Asian longhorned beetle and citrus longhorned beetle: A worldwide perspective. *Annual Review of Entomology*, **55**, 521-546.

Hill, J., Thomas, C. & Huntley, B. (2003) Modelling present and potential future ranges of European butterflies using climate response surfaces. In: *Butterflies: Ecology and Evolution Taking Flight*. Boggs, C., Watt, W. & Ehrlich, P., p. 149-167. The University of Chicago Press, Chicago.

Hortal, J., Roura-Pascual, N., Sanders, N. & Rahbek, C. (2010) Understanding (insect) species distributions across spatial scales. *Ecography*, **33**, 51-53.

Horton, D. (1988) World patterns and trends in sweet potato production and use. In: *Exploration, maintenance and utilization of sweet potato genetic resources: report of the First Sweet Potato planning conference 1987*, pp. 17-27. International Potato Center, Lima.

Jiménez-Valverde, A. & Lobo, J. (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica*, **31**, 361-369.

Jordan, N., Larson, D. & Huerd, S. (2008) Soil modification by invasive plants: effects on native and invasive species of mixed-grass prairies. *Biological Invasions*, **10**, 177-190.

Liu, C., Berry, P., Dawson, T. & Pearson, R. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385-393.

Lobo, J., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145-151.

Low, J. (1997) *Prospects for sustaining potato and sweet potato cropping systems in Southwest Uganda*. CIP, Lima, Peru.

Marabuto, E. (2006) The occurrence of a neotropical hawkmoth in southern Portugal: *Agrius cingulatus* (Fabricius, 1775) (Lepidoptera: Sphingidae). *Boletín Sociedad Entomológica Aragonesa*, **38**, 163-166.

Marzban, C. (2004) The ROC curve and the area under it as performance measures. *Weather and Forecasting*, **19**, 1106-1114.

Mullin, B., Anderson, L. & DiTomaso, J. (2000) Invasive plant species. *The Council for Agricultural Sciences and Technology*, **13**, 1-18.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225-245.

Peterson, A. & Vieglais, D. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience*, **51**, 363-371.

Peterson, A. (2003) Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology*, **78**, 419-433.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161-175.

Pimentel, D. (2002) *Biological invasions: economic and environmental costs of alien plant, animal, and microbe species*. CRC Press, Boca Raton (Florida, USA).

Ragsdale, D.W., Landis, D.A., Brodeur, J., Heimpel, G.E. & Desneux, N. (2011) Ecology and management of the soybean aphid in North America. *Annual Review of Entomology*, **56**, 375-399.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689-1703.

Ricciardi, A. (2003) Predicting the impacts of an introduced species from its invasion history: an empirical approach applied to zebra mussel invasions. *Freshwater Biology*, **48**, 972-981.

Settele, J., Kudrna, O., Harpke, A., Kühn, I., van Swaay, C., Verovnik, R., Warren, M., Wiemers, M., Hanspach, J., Hickler, T., Kühn, E., van Halder, I., Veling, K., Vliegenthart, A., Wynhoff, I. & Schweiger, O. (2008) *Climatic risk atlas of European butterflies*. Biorisk 1 (Special Issue), Pensoft (Sofia), 719 pp.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences (USA)*, **106**, 19644-19650.

Strayer, D., Eviner, V., Jeschke, J. & Pace, M. (2006) Understanding the long-term effects of species invasions. *Trends in Ecology and Evolution*, **21**, 645-651.

Sutherland, S. (2004) What makes a weed a weed: life history traits of native and exotic plants in the USA. *Oecologia*, **141**, 24–39.

Talekar, N. (1987) Insect pests of sweet potato in the tropics. In: *Proceeding of Symposium on Crop Protection in the Tropics*. 11th International Congress of Plant Protection. Manila (Philippines).

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M. (2009) BIOMOD – A platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369-373.

Thuiller, W., Richardson, D., Pysek, P., Midgley, G., Hughes, G. & Rouget, M. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234-2250.

Woolfe, J.A. (1992) *Sweet potato: an untapped food resource*, *Volume 1991*. Cambridge University Press, Cambridge (MA).

Appendix

Appendix 6.1. 15 preliminary Maxent-Models (M1-M15) with varying predictors analyzed for initial selection of input variables. Cell values give percent contribution of each variable to the respective models. Models were compared according to the area under the receiver-operating curve (AUC). The model with the highest AUC (M11, in bold) was used for final prediction, and the same variable combination was utilized for BIOMOD models. See Methods for details.

Variable	M1 (AUC 0.67)	M2 (AUC 0.85)	M3 (AUC 0.87)	M4 (AUC 0.67)	M5 (AUC 0.87)
Annual Temperature	-	90.6	66.8	-	62.1
Annual Precipitation	-	9.4	12.3	-	7.7
Mean diurnal range	-	-	-	13.6	11.5
Herb cover	56.2	-	15.2	43.7	14.1
Tree cover	12.6	-	3.5	11.3	3.4
Bare ground cover	31.1	-	2.2	31.4	1.1
Precipitation driest quarter	-	-	-	-	-
Precipitation coldest quarter	-	-	-	-	-
Precipitation warmest quarter	-	-	-	-	-
Precipitation wettest quarter	-	-	-	-	-
Mean Temperarature wettest quarter	-	-	-	-	-
Mean Temperature warmer quarter	-	-	-	-	-
Mean Temperature driest quarter	-	-	-	-	-
Mean Temperature coldest quarter	-	-	-	-	-

Variable	M6 (AUC 0.89)	M7 (AUC 0.90)	M8 (AUC 0.89)	M9 (AUC 0.88)	M10 (AUC 0.88)
Annual Temperature	60.6	60.3	45.8	42.9	39.2
Annual Precipitation	7.5	6.9	6.0	7.0	7.6
Mean diurnal range	11.2	9.4	8.6	9.8	9.2
Herb cover	13.4	15.1	15.1	12.5	12.6
Tree cover	3.0	1.7	2.0	3.3	1.4
Bare ground cover	0.8	1.6	-	-	0.6
Precipitation driest quarter	-	-	-	-	-
Precipitation coldest quarter	-	-	-	-	3.5
Precipitation warmest quarter	-	-	-	-	-
Precipitation wettest quarter	-	-	-	-	-
Mean Temperarature wettest quarter	3.6	-	1.8	2.6	-
Mean Temperature warmer quarter	-	2.8	1.5	1.6	1.8
Mean Temperature driest quarter	-	-	-	1.4	2.5
Mean Temperature coldest quarter	-	-	17.7	17.9	19.2

Variable	M11 (AUC 0.93)	M12 (AUC 0.90)	M13 (AUC 0.91)	M14 (AUC 0.90)	M15 (AUC 0.88)
Annual Temperature	**42.9**	38.0	41.1	-	-
Annual Precipitation	**8.2**	6.4	2.0	-	-
Mean diurnal range	**7.9**	7.9	9.7	-	-
Herb cover	**9.0**	12.2	10.3	-	11.4
Tree cover	**2.8**	2.5	2.6	-	3.6
Bare ground cover	**0.8**	1.1	0.9	-	1.3
Precipitation driest quarter	**4.3**	5.2	3.9	4.0	4.1
Precipitation coldest quarter	**3.7**	2.6	2.3	5.3	3.3
Precipitation warmest quarter	**-**	2.5	2.6	4.2	8.8
Precipitation wettest quarter	**-**	-	5.3	-	-
Mean Temperarature wettest quarter	**2.3**	1.8	2.3	6.0	4.5
Mean Temperature warmer quarter	**2.2**	1.5	1.4	1.5	1.6
Mean Temperature driest quarter	**0.7**	1.3	1.4	1.0	2.8
Mean Temperature coldest quarter	**15.1**	17.0	14.1	69.7	54.1

CHAPTER 7

Synthesis & Conclusions

7.1. Synthesis

The work presented in this thesis, represents the first documented almost global database of high-resolution maps of distributions for a complete family of herbivore insects: The hawkmoths of the Old World (Lepidoptera: Sphingidae). Along the process of producing it, various analyses regarding methodology, the worth of the data, biodiversity patterns and particular uses of the methodologies were carried out.

Choosing the right method:

There is a large body of literature that have used species distribution models for successfully inferring the current ranges of the species. Despite the wide range of existing algorithms to estimate species distributions there is still disagreement on what method to use under what circumstance and how accurate they are. The results presented here, support the statement that SDMs can successfully predict species distributions. The methods used to build distribution maps can vary greatly from one another and in absence (most of the time) of truly independent ways to assess their accuracy we should resort to other methods such: AUC, MPA or expert opinion. Our results consistently show that relatively new methods like Maxent outperform others, followed by Random Forest, but also call for caution in the case of model averaging (see Chapter 2 for details).

Data set used here includes a relevant extension of properties in terms of geographical scale, distribution phylogenetic variability and wide range of sample size compared to other studies (Elith et al. 2006 among others), in addition to represent data properties inherent to the majority of distribution data available, where the only ecological data known for a vast number of species (invertebrates in particular) is a name and a location.

Worth of the data: There is an increased awareness of the value of data from museum collections. They represent an important source of information that so far has not been fully exploited. At the same time, platforms such as Global Biodiversity Information Facility (GBIF) which aims to make available biodiversity data is not used as often as they should. An important assumption for SDM's is that the occurrence data included in the models represent the range size and the environmental tolerance of the species as complete as possible. Our results suggest that quantity does not necessarily imply the best quality (Chapter 3 for details). Collection data provided the most

complete information the range of species nevertheless GBIF was a good addition of information, but not a replacement.

Species richness, climatic drivers and inventory completeness:

As rightfully pointed out by various authors, climate is expected to be an important determinant of species distribution, and therefore might also determine patterns of species richness. However those patterns (i.e. species richness) might be blurred by its strong correlation with sampling effort (see Chapter 4). Numerical estimators in combination with models of environmental effects proved to set out the baseline to disentangle potential causes of differential sampling from those variables driving species richness. It seems to be the avenue to advance the knowledge of understudied groups like insects.

Furthermore, patterns of species richness are also possible to map by overlaying the resultant grids distributions of individual species (Chapter 5) providing knowledge about richness in areas that might be particularly poorly sampled.

Inventory completeness could be partially predicted from human geographical factors and it might be helpful to prioritize collecting in the future.

Specific uses of SDMs

Species distribution models have been utilized for many different purposes; however, conservation is one of the biggest. Predicting potential distribution of species that could be regarded as pests in their own ranges threatening the local biodiversity is of major concern in conservation. Our results show that SDMs can help with delimiting zones of potential spread, to initiate early monitoring and efforts for control (Chapter 6).

Certainly the potential of data produced here and the methods used throughout the analysis is not yet exhausted. There is still pending numerous questions to address and analysis to make.

Phylogenetic relationships and different life histories between the tribes in the family suggest they affect several aspects of their distribution (i.e. habitat and food preferences, range sizes, niche breath etc, Beck et al 2006a, 2006b, 2006c). So might be interesting to make a detailed analysis about effects of phylogeny on range sizes to test if there is phylogenetic conservatism among them. In the same line, we also have the opportunity to study whether relevant traits or regional environmental conditions would explain the variation in strength of the phylogeny effect on range size. It would be interesting to study range shifts under changing climates, and test whether there is a phylogenetic compound in the resilience to those changes.

As outlined above, climate is expected to play an important role as driver for α, β, and γ diversity patterns. Analysis for testing different hypothesis is still pending;

1. Whether water energy availability provides good explanations and quantifies how much variation different parameters explain.
2. Whether non-environmental factors provide more explanation for the variation of these patterns or cover residual variation from the environmental parameters.

7.2. Conclusion

SDM is a powerful tool to provide information about the occurrence of the species. However it is important not to forget, that these models are correlative, so no causality can be directly inferred from them; if a variable shows good association with the occurrence does not mean necessarily that directly determine their distribution.

Nevertheless, for most of the species (particularly invertebrates), this might be the best (and perhaps the only) approach given the limited information that is generally held upon them. This is a first step to know more about them. Amidst the recent biodiversity crisis that we are in, SDMs despite of their limitations, provide invaluable information to fill the gaps in our knowledge about spatial distribution patterns of biodiversity in the world.

SUMMARY

The geographic distribution of the species is one of the basic units of information in ecology and biogeography, yet for the vast majority of the species in the world is quite unknown. It is a phenomenon called the Wallacean shortfall. The knowledge available about the distribution is biased to few certain taxa and regions. For insects, in particular, this information is even sparser despite being very specious taxa and play an important ecological role. Species Distribution Modelling (SDMs) offers a potentially powerful tool that might help to fill those gaps in our knowledge especially for those species from which there is very little known about their ecology and places where collecting has been very scarce. In this thesis, I used a database compiled from museum and private collections, publications (including online databases) and fieldwork data already assembled by my co-authors.
Such database contains over 109,880 distributional records of the global distribution for all the 982 non-American taxa of the Sphingidae family of Lepidoptera which then I combined with SDMs algorithms to provide high-resolution distribution maps for all the taxa in the family and study patterns of biodiversity. Since the purpose of this document is to provide stand alone manuscripts that are at the point of submission are either submitted, in review or published, I will refer to "we" throughout much of the text.

As a first step, we compared the performance of 8 commonly used SDM's algorithms while considering some intrinsic properties of the species and data with a representative sample of the species in the family (*Chapter 2*). The algorithm that performed the best was Maxent followed by Random Forest, however we could not confirm effects of species traits or data properties influencing the modeling performance.

Subsequently, in *Chapter 3* we assessed the value of different data sources, by comparing an independent compilation of occurrence data vs GBIF database, and its contribution to different aspects of the range of the species (i.e. range filling, range extent and climatic niche space). GBIF provided more records than other sources though contributed with less information about the range, so it is not yet an alternative to manual compilation of distributional data.

Species diversity patterns based on numerical estimators are studied in *Chapter 4*, in relation with their main environmental correlates for a fraction of the study region. We also provided assessment of inventory completeness in the same region. Variables describing vegetation emerged as important predictors of species richness. Variables capturing heat, energy availability and topographic heterogeneity were identified as further parameters influencing species richness. Inventory completeness is positively associated with densely populated areas, accessibility, protected areas and colonial history. We discussed how this approach sets the baseline to estimate diversity patterns in under-studied taxa.

A detailed documentation of data acquisition, processing and modeling is compiled in *Chapter 5*. We applied the modeling technique chosen in *Chapter 2* in combination with environmental data and vegetation cover data to the whole dataset. We could retrieve models for 789 taxa whereas we provided expert drawn range maps for the remaining 193. In general, annual temperature range was the factor contributing the most to shape species' distributions followed by variables related to precipitation. Variables related to vegetation did not highly contribute. In a next step, we superimposed the resultant grids to study patterns of biodiversity at two spatial scales (α = 5 x 5 km and γ = 200 x 200 km) and then used them to calculate β-diversity. The α and γ diversity maps exhibited a latitudinal gradient of species richness towards the tropics whereas β-diversity patterns revealed rather a altitudinal gradient, higher in mountainous regions and along biogeographical boundaries. This set of maps is the result of a collaborative project that to the best of our knowledge

compiles the first distributional data set for a complete family of invertebrates at an almost global scale. Achievements, challenges and limitations of the project are also reported and discussed.

A specific application of SDM is shown in *Chapter 6.* We predicted the potential the range of an invasive species (*Agrius cingulata*), native to the American continent which have recently spread and established populations in Africa. We used two types of SDM based on native range records and environmental data. Our results showed that *Agrius cingulata* could find suitable habitat across wide stretches across Sub-saharan Africa. Early monitoring programs might be valuable to evaluate the status of the invasion.

In C*hapter 7*, a general discussion of the results plus an outlook to further research with this data is presented. All the maps (i.e. from raw data, intermediate steps until the final map) together with appendixes containing details of the models, list of species, literature and museum collection sources are deposited on the network drive at the University Computing Centre of Basel. It is our plan for the future to make this database available throughout the website facility: The map of life (http://www.mappinglife.org/).

RESUMEN

La distribución geográfica de las especies es una de las unidades básicas de información en estudios de ecología y biogeografía, sin embargo desafortunadamente dicha información es desconocida para la mayoría de las especies del planeta. Este fenómeno es mejor conocido como el déficit de Wallace (o el termino en ingles "Wallacean shortfall"). El conocimiento que existe disponible acerca de la distribución de las especies está sesgado hacia ciertos taxa y regiones. Para insectos, en particular, esta información es aún más escasa a pesar de ser un taxón muy rico en especies y desempeñar un papel ecológico muy importante. El modelamiento de distribución de especies (MDS) ofrece una herramienta poderosa que puede ayudar a llenar esas lagunas, especialmente para esas especies de las que muy poco se sabe acerca de su ecología y en aquellos lugares donde el muestreo ha sido escaso. En esta tesis, he usado una base de datos recopilada de museos y colecciones privadas, publicaciones (incluyendo bases de datos en internet) y datos de trabajo de campo llevado a cabo por mis coautores. Dicha base de datos contiene más de 109.880 registros de la distribución global para todas las 982 especies no Americanas de la familia Sphingidae de lepidópteros. Seguidamente los combine con algoritmos de MDSs para proporcionar mapas de distribución con alta resolución de todas las especies en la familia y luego estudie patrones de biodiversidad. Ya que este es un proyecto colaborativo y el propósito de este documento como tesis es proveer capítulos que puedan leerse sin necesidad de hacer referencia a los otros, me referiré a "nosotros" a lo largo de una gran parte del texto.

Como primer paso, comparamos el rendimiento de 8 algoritmos de (MDS) comúnmente utilizados, teniendo en cuenta al mismo tiempo algunas de las propiedades intrínsecas de las especies y de los datos. Elegimos un grupo de 64 especies como muestra representativa de las todas las especies de la familia (*Capítulo 2*). El algoritmo que proporciono mejores resultados fue Maxent seguido por "Random forest", sin embargo no fue posible confirmar si ciertos atributos particulares de las especies o propiedades de los datos influyen en desempeño de los algoritmos.

Posteriormente, en el *Capítulo 3* evaluamos el valor de diferentes fuentes de datos. En particular comparamos los datos de la compilación independiente de datos de ocurrencias vs la base de datos online de GBIF, y su contribución a los diferentes aspectos de los rangos de la especie (es decir: la ocupación, la extensión, y el nicho climático observado). GBIF ofrece más registros que otras fuentes, sin embargo estos registros contribuyeron con menos información acerca de los diferentes aspectos del rango, por lo que todavía no ofrece una alternativa a la recopilación manual de datos de distribución.

Patrones de diversidad y riqueza de especies basados en estimadores numéricos se estudiaron en el *Capítulo 4*, en relación con sus principales determinantes ambientales en una fracción de nuestra área de estudio. También proporcionamos una evaluación de que tan completo es el inventario de especies en la misma región. Variables que describen la estructura de la vegetación surgen como importantes predictoras de la riqueza de especies. Asimismo variables relacionadas con captura de calor, la energía y heterogeneidad topográfica se identificaron como otros parámetros que influyen en la riqueza de especies. Que tan completo es el inventario está asociado positivamente con áreas densamente pobladas, accesibilidad de la zona, áreas protegidas y la historia colonial. También discutimos cómo este enfoque establece una línea base para estimar patrones de diversidad en aquellos taxones menos estudiados.

Una documentación detallada de la adquisición de datos, procesamiento y modelado de especies se describió en el *Capítulo 5*. Se aplicó el algoritmo de modelación elegido en el *Capítulo 2*, en combinación con los datos ambientales y de estructura de vegetación como predictoras. Pudimos recuperar modelos para 789 especies, mientras proveemos mapas basados en opinión de expertos

para las 193 especies restantes. En general, el rango de temperatura anual se identifico como el factor que más contribuyó a la distribución de las especies, seguida por variables relacionadas con la precipitación mientras que variables relacionadas con estructura de la vegetación no aportaron mucho. Seguidamente, se superponen los mapas producidos en el paso anterior para estudiar patrones de biodiversidad en 2 escalas espaciales distintas (α = 5 x 5 km (biodiversidad local) y γ = 200 x 200 km (biodiversidad regional)) adicionalmente estos datos se utilizaron para calcular β-diversidad. Los mapas de diversidad α y γ exhiben un gradiente latitudinal de incremento de la riqueza de especies hacia los trópicos, mientras que el mapa de diversidad β revela mas bien un patrón de gradiente altitudinal siendo mayor en regiones montañosas y a lo largo de las fronteras biogeográficas. Este conjunto de mapas es el resultado de un proyecto colaborativo que según nuestro conocimiento compila los primeros datos de distribución establecidos para una familia completa de invertebrados en una escala casi global. También reportamos y discutimos logros, desafíos y limitaciones del proyecto.

Una aplicación específica de MSD se muestra en el *Capítulo 6*. La predicción del rango de una especies invasiva (*Agrius cingulata*), cuyo rango nativo es el continente americano sin embargo recientemente ha extendido dicho rango y ha establecido poblaciones en África. Se utilizaron dos tipos de MDS basados en los registros de ocurrencia en el rango nativo en combinación con datos ambientales. Nuestros análisis reportan que *Agrius cingulata* podría encontrar un hábitat adecuado a través de amplias extensiones de todo el subsahara africano. Programas tempranos de monitoreo podrían ser valiosos para evaluar el estado de la invasión.

En el *Capítulo 7*, se presenta una discusión general de los resultados, además de planes futuros de investigación. Todos los mapas (es decir, datos sin procesar, pasos intermedios y el mapa final) junto con los anexos que contienen detalles de los modelos, listas de especies, fuentes de literatura y listas de los museos de donde recopilamos los datos de ocurrencia, están depositados en unidad de red en el Centro de Computación de la Universidad de Basilea. Nuestro plan es hacer disponibles estos datos através del proyecto de Internet: The map of life (http://www.mappinglife.org/).

ACKNOWLEDGEMENTS

There is a superb bunch of people that I need to thank for being by my side throughout all these years. But undoubtedly my family deserves to be right on top of the list, my parents: Jaime and Elssy, all their teachings, love and unconditional support with all my “genius” ideas have made possible all the things I have achieved. My brothers and sisters: Jaime, Randy, Melissa and Isabella have supported me in every step no matter how far away I decided to go and no matter what it was. You are the biggest motor in my life. ☺

Then, a huge thanks, all my respect and admiration to my supervisor, Jan Beck for his guidance, patience in dealing with my latin moods. His corrections, comments and challenges made this experience unforgettable.

I should also thank, Dr. Ian Kitching for all his help to deal with the taxonomy, even if it meant duplicate my work when decided to split one or more species.

Simon Loader deserves a very special thanks, his friendship was one of the best part of my life in Basel and at the Institute, his trust, patience and support kept me going, especially the difficult moments, million thanks!!!!.

Many thanks go to Prof. Dr. Peter Nagel for his support in every respect.

I thank Dr. Carsten Bruehl for taking over the co-report.

Very special thanks also to all the people in the Institute of Biogeography. The pleasant working atmosphere made the work easier in every respect. Reto Hagmann, not only shared the office with me but also made “the light” in there every time we needed it. To Ruth Kimser for all her support. Definitively the “frog people”: Simon, Chris, Dominik, Sara, Maiti-Matteo made the visits to Kberg a worthy walk.

People around Basel, Sebastian, Lina, Camila, always there, willing to help, to support and have fun even with my workaholic style of life. Claudia Avila, my (not anymore)-flatmate/almost sister for listening, helping, celebrating and even lecturing, made my life at home very pleasant.
My friends (las brujas) in Colombia for made me feel every time I went back like if I would not have ever left thanks!

Thanks to the SNF for funding the initial 3 years of the project and the Freiwillige Akademische Gesellschaft Basel (FAG) for funding the last 6 months.

CURRICULUM VITAE

Name	Liliana Ballesteros Mejia
Date of birth	July 5th 1978
Place of birth	Popayan, Colombia
1989 – 1994	Colegio Rosario Campestre, Colombian qualification for university entrance with commerce as main subject.
1996 – 2002	BSc in Biology from Universidad de los Andes (Bogota Colombia)
	Voluntary semestre for Fundacion Yubarta Fieldwork for course "Genetics II" Cali – Colombia Fieldwork for course "Entomology" (Boyaca- Colombia) Fieldwork for course "Behavioural ecology" (Villavicencio - Colombia) **BSc Thesis** "Effect of whale-watching boats in breathing and diving behaviour of mother and calf groups' of humpback whales (*Megaptera novaeangliae*) Malaga and surroundings, Colombian Pacific. 2002 " Nine moths fieldwork in Malaga-Colombian Pacific.
2002 - 2003	Scientific assistant in Fundacion Yubarta (Cali - Colombia)
2004 - 2006	Work experience as science teacher for primary and middle school at Colegio Nuevo Reino de Granada (Cota - Colombia)
2006 - 2008	TopMaster in Evolutionary biology and Ecology, Rijksuniversiteit Groningen, (Groningen, The Netherlands).
	Master project 1: "Life history approach to allometric scalling and growth" Theoretical Biology Group – University of Groningen
	Master project 2: "Effect of spatial patterns of resource distribution on the predation of three fruit species in tropical forest" Community and Conservation Ecology Group – University of Groningen Barro Colorado Island (BCI) - Smithsonian Tropical Research Institute Panama. Three months fieldwork.
2008 – Present	PhD thesis: "Examining and addressing the Wallacean shortfall: Species distribution models and biodiversity patterns of Hawkmoths in the Old World" Supervisor: PD Dr Jan Beck.
Publications	Projecting the potential invasion of the Pink Spotted Hawkmoth (*Agrius cingulata*) across Africa. Ballesteros-Mejia, L., Kitching I.J., Beck, J. (2011). International Journal of Pest Management 57 p. 153-159
	What's on the horizon of Macroecology? Present status and future perspectives. Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S., Gruber, B., Hof, C. Jansen, F., Knapp, S., Kreft, H., Schneider A-K. Winter, M., Dormann, C.F. (2012). Ecography 35 p. 1-11.

Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Ballesteros-Mejia, L., Kitching, I.J., Nagel, P., Jetz, W., Beck, J. Global Ecology and Biogeography 2013. 22, 586-595

Revisiting the indicator problem: can three epigean arthropod taxa inform about each other's biodiversity? Beck, J., Pfiffner, L., Ballesteros-Mejia, L., Blick, T. Luka, H. Diversity and Distributions 2012. 1-12. DOI: 10.1111/ddi.12021

REFERENCES

Araujo MB and New M (2007) Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47

Araújo, M. B., and M. Luoto. (2007). The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16:743–753.

Asher, J., M. Warren, R. Fox, P. Harding, G. Jeffcoate, and S. Jeffcoat (Eds.). (2001). The Millennium Atlas of Butterflies in Britain and Ireland. . Oxford University Press, Oxford.

Austin M (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157: 101-118

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200:1–19.

Bahn, V. and McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography, 16, 733-742.

Baillie J, Hilton-Taylor C, Stuart SN (2004) IUCN Red List of Threatened Species: a Global Species Assessment. World Conservation Union, IUCN Glad, Switzerland and Cambridge

Ballesteros-Mejia L, Kitching IJ, Beck J (2011) Projecting the potential invasion of the Pink Spotted Hawkmoth (Agrius cingulata) across Africa. International Journal of Pest Management 57:153-159

Ballesteros-Mejia L, Kitching IJ, Jetz W, Nagel P, Beck J (in press) Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Global Ecology and Biogeography.

Balmford, A., Moore, J.L., Brooks, T., Burgess, N., Hansen, L. A, Williams, P. and Rahbek, C (2001) Conservation conflicts across Africa. Science, 291, 2616-9.

Barve N, Barve V, Jiménez-Valverde A, Lira-Noriega A, Maher SP, Peterson AT, Soberón J, Villalobos F (2011) The crucial role of the accessible area in ecological niche modelling and species distribution modelling. Ecological Modelling 222: 1810-1819

Beale CM, Lennon JJ (2012) Incorporating uncertainty in predictive species distribution modeling. Philosophical Transactions of the Royal Society B 367:247-258

Beck J, Kitching IJ (2007) Correlates of range size and dispersal ability: a comparative analysis of sphingid moths from the Indo-Australian tropics. Global Ecology and Biogeography 16: 341-349

Beck J, Kitching IJ, Linsenmair KE (2006a) Measuring range sizes of South-East Asian hawkmoths (Lepidoptera: Sphingidae): effects of scale, resolution and phylogeny. Global Ecology and Biogeography 15: 339–348

Beck, J., Kitching, I.J. Linsenmair, KE. (2006b) Determinants of regional species richness: an empirical analysis of the number of hawkmoth species (Lepidoptera: Sphingidae) on the Malesian archipelago. Journal of Biogeography, 33, 694-706.

Beck J, Kitching IJ, Linsenmair KE (2006c) Diet breadth and host plant relationships of Southeast-Asian sphingid caterpillars. Ecotropica 12: 1–13

Beck J, Kitching IJ, Linsenmair KE (2006d) Effects of habitat disturbance can be subtle yet significant: biodiversity of hawkmoth-assemblages (Lepidoptera: Sphingidae) in Southeast-Asia. Biodiversity and Conservation 15: 465–486

Beck J, Kitching IJ, Linsenmair KE (2006e) Extending the study of range – abundance relations to tropical insects: sphingid moths in Southeast Asia. Evolutionary Ecology Research 8: 677-690

Beck, J. and Kitching, I.J. (2007) Estimating regional species richness of tropical insects from museum data: a comparison of a geography-based and sample-based methods. Journal of Applied Ecology, 44, 672-681.

Beck, J., I. J. Kitching, and J. Haxaire. (2007). The latitudinal distribution of sphingid species richness in continental Southeast Asia: What causes the “biodiversity hotspot” in northern Thailand. Raffles Bulletin of Zoology 55:179–185.

Beck J., W. Nässig. (2007). Diversity and abundance patterns, and revised checklist, of saturniid moths (Lepidoptera: Saturniidae) from Borneo. Nachrichten des Entomologischen Vereins Apollo 28: 155-164.

Beck, J., and I. Kitching. (2004-2008). The Sphingidae of Southeast-Asia (incl. New Guinea, Bismarck and Solomon Islands) version 1.5.

Beck, J., Schwanghart, W., Chey, V.K. and Holloway, J.D. (2011) Predicting geometrid moth diversity in the Heart of Borneo. Insect Conservation and Diversity, 4, 173-183.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S. a., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M. and Dormann, C.F. (2012a) What's on the horizon for macroecology? Ecography, 35, 673-683.

Beck, J., Holloway, J. D., Khen, C.V., and Kitching, I. J. (2012b). Diversity partitioning confirms the importance of beta components in tropical rainforest Lepidoptera. The American Naturalist 180:E64–E74.

Beever, E.A., Swihart, R.K. and Bestelmeyer, B.T. (2006) Linking the concept of scale to studies of biological diversity: evolving approaches and tools. Diversity and Distributions, 12, 229–235.

Bik, H. M., D. L. Porazinska, S. Creer, J. G. Caporaso, R. Knight, and W. K. Thomas. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology and Evolution 27:233–43.

Bini, L. M., J. A. F. Diniz-Filho, T. F. L. V. B. Rangel, R. P. Bastos, and M. P. Pinto.(2006). Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. Diversity and Distributions 12:475–482.

BirdLife International (2000) Threatened Birds of the World. BirdLife International and Lynx Editions, Cambridge and Barcelona

Boakes EH, McGowan PJK, Fuller RA, Chang-qing D, Clark NE, O'Connor K, Mace GM (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8: e1000385

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution 24: 127–135

Böller, M. (2012). Modellierung von Verbreitungsgebieten mit MaxEnt: Der Effekt von verzerrten presence-only Datensätzen der Global Biodiversity Information Facility (GBIF). BSc Thesis. University of Basel. Unpublished.

Brehm, G., Homeier, J. and Fiedler, K.. (2003). Beta diversity of geometrid moths (Lepidoptera: Geometridae) in an Andean montane rainforest. Diversity and Distributions 9:351–366.

Breiman, L. (2001). Random forests. Machine learning 45:5–32.

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W, Fortin ,M.-J., Randin, C., Zimmermann, N.E., Graham, C.H. and Guisan, A. (2011) Measuring ecological niche overlap from occurrence and spatial environmental data. Global Ecology and Biogeography 21, 481-497.

Brown, J. H., G. C. Stevens, and D. M. Kaufman. (1996). The geographic range: Size, Shape, Boundaries, and Internal Structure. Annual Review of Ecology and Systematics 27:597–623.

Brown, J., and M. Lomolino. (1998). Biogeography. Sinauer Press, Massachusetts.

Buckley, L.B. and Jetz, W. (2007) Environmental and historical constraints on global patterns of amphibian richness. Proceedings of the Royal Society (B), 274, 1167-1173.

Buckley, L.B., Hurlbert, A.H. and Jetz, W. (2012) Broad-scale ecological implications of ectothermy and endothermy in changing environments. Global Ecology and Biogeography, 21, 873–885.

Carpenter, G. A.N. Gillison, J.Winter. (1993). DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation 2: 667-680

Chase, J. M., and M. A. Leibold. (2003). Ecological Niches: Linking Classical and Contemporary Approaches. Universtity of Chicago Press, Chicago.

Chao A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11,** 265-270.

Chessel, D., Dufour, A.B. and Thioulouse, J. (2004) The ade4 package –I: One-table methods. R News. 4, 5-10.

Clark, B.R., Godfray, H.C.J., Kitching, I.J., Mayo, S.J. and Scoble, M.J. (2009) Taxonomy as an e-Science. Philosophical Transactions of the Royal Society(A) 367, 953-966.

Colwell, R. K. (2005). EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application published at: http://purl.oclc.org/estimates.

Colwell, R.K. and Coddington, J.A. (1994) Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society (B), 345, 101-18.

Costello, M. J., and S. P. Wilson. (2011). Predicting the number of known and unknown species in European seas using rates of description. Global Ecology and Biogeography 20:319–330.

Currie, D.J., Mittelbach, G.G., Cornell, H.V., Field, R., Guegan, J.-F., Hawkins, B. a., Kaufman, D.M., Kerr, J.T., Oberdorff, T., O'Brien, E. and Turner, J. R. G. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. Ecology Letters, 7, 1121-1134.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88: 2783-92

Danner, F., Eitschberger, U., and Surholt, B. (1998) Die Schwärmer der westlichen Palaearktis. Bausteine zu einer Revision (Lepidoptera: Sphingidae). Herbipoliana 4, 1-368, 1-720.

Davis, A. L. V., C. H. Scholtz, and S. L. Chown.(1999). boundaries and community Species turnover , in gradient assemblages across an altitudinal South Africa of dung beetle biogeographical composition. Journal of Biogeography 26:1039–1055.

De Latin, G. (1967). Grundriss der Zoogeographie. G. Fisher-Verlag, Jena. Page 602.

Deans, A. R., M. J. Yoder, and J. P. Balhoff. (2012). Time to change how we describe biodiversity. Trends in Ecology and Evolution 27:78–84.

Diamond, J. (2006). Collapse: How Societies Choose to Fail or Succeed. Page 573. Viking, Penguin Group, New York.

Diniz-Filho, J. A. F., P. De Marco Jr, and B. a. Hawkins. (2010). Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. Insect Conservation and Diversity 3:172–179.

Dynesius, M., and Jansson, R. (2000). Evolutionary consequences of changes in species' geographical distributions driven by Milankovitch climate oscillations. Proceedings of the National Academy of Sciences (B) 97:9115–9120.

Elith J, Graham C, Anderson R, Dudık M, Ferrier S, Guisan A, Hijmans R, Huettmann F, Leathwick J, Lehmann A, Li J, Lohmann, L, Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton Jm, Peterson A, Phillips S, Richardson K, Scachetti-Pereira R, Shapire R, Soberon J, Williams S, Wisz M, Zimmermann N (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods in Ecology and Evolution 1: 330-342

Elith J, Leathwick J (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 265-275

Elith J, Leathwick J (2009) Species Distribution Models: Ecological Explanation and Prediction across Space and Time. Annual Reviews of Ecology, Evolution and Systematics 40: 677-697

Elton, C. (1927). Animal Ecology. Sidgwick & Jackson, London.

Engler R, A. Guisan, L. Rechsteiner. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263-274

Evans, K.L., Warren, P.H. and Gaston, K.J. (2005) Species–energy relationships at the macroecological scale: a review of the mechanisms. Biological Reviews, 80, 1–25.

Ferrier S (2002) Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? Systematic Biology 51: 331-363

Ferrier S, Watson G, Pearce J, Drielsma M (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling Biodiversity and Conservation 11: 2275-2307

Fiedler, K. and Truxa, C. (2012) Species richness measures fail in resolving diversity patterns of speciose forest moth assemblages. Biodiversity and Conservation, 21, 2499-2508.

Field, R., B. A. Hawkins, H. V. Cornell, D. J. Currie, J. A. F. Diniz-Filho, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. M. O'Brien, and J. R. G. Turner. (2009). Spatial species-richness gradients across scales: a meta-analysis. Journal of Biogeography 36:132–147.

Field, R., Hawkins, B.A., Cornell, H.V., Currie, D.J., Diniz-Filho, J.A.F., Guegan, J.-F., Kaufman, D. M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M. and Turner, J.R.G. (2009) Spatial species-richness gradients across scales: a meta-analysis. Journal of Biogeography, 36, 132-147.

Franklin, J. (2009). Mapping species distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge.

Galster, S., N. D. Burgess, J. Fjeldsa°, L. A. Hansen, and C. Rahbek. 2007. One degree resolution databases of the distribution of 1085 mammals in Sub-Saharan Africa.

Gasc, J. P., A. Cabela, J. Crnobrnja-Isailovic, D. Dolmen, K. Grossenbacher, P. Haffner, J. Lescure, H., T.S. Martens, M. Veith, and A. Zuiderwijk. (1997). Atlas of amphibians and reptiles in Europe. Societas Europaea Herpetologica & Museum National d'Histoire Naturelle, Paris.

Gaston, K.J. 2003. The structure and dynamics of geographic ranges. Oxford University Press, Oxford.

Gaston, K. J. & Blackburn, T. M. (2000). Pattern and process in macroecology. Oxford, UK: Blackwell Science.

Ghalambor, C. K., R.B. Huey, P.R. Martin, J.J. Tewksbury and G. Wang (2006). Are mountain passes higher in the tropics? Janzen's hypothesis revisited. Integrative and Comparative Biology 46: 5-17

Gibbons, D.W., J.B. Reid and R.A.Chapman. (1993) The new atlas of breeding birds in Britain and Ireland: 1988-1991. T. & A.D. Poyser

Giovanelli JGR, de Siqueira MF, Haddad CFB, Alexandrino J (2010) Modeling a spatially restricted distribution in the Neotropics: How the size of calibration area affects the performance of five presence-only methods. Ecological Modelling 221: 215-224

Glor, R. E., and D. Warren. (2011). Testing ecological explanations for biogeographic boundaries. Evolution; international journal of organic evolution 65:673–83.

Godfray, H.C.J., Lewis, O.T. and Memmot, J. (1999) Studying insect diversity in the tropics. Philosophical Transaction of the Royal Society (London) B, 354, 1811-1824.

Godfray, H.J.C., Clark, B.R., Kitching, I.J., Mayo, S.J. and Scoble, M.J. (2007) The web and the structure of taxonomy. Systematic Biology 56, 943-955.

Godinho, R., J. Teixeira, R. Rebelo, P. Segurado, and A. Loureiro. (1999). Atlas of the continental Portuguese herpetofauna : an assemblage of published and new data. Revista Espanola de Herpetologia 13:61–82.

Gotelli, N. and Colwell, R. (2001) Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, 4, 379-391.

Graham CH, Elith J, Hijmans RJ, Guisan A, Townsend PA, Loiselle BA (2008) The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45: 239-247

Graham, C., Ferrier, S., Huettman, F. and Moritz, C. (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.

Graham, C. H., Ron, S. R., Santos, J. C., Schneider, C. J., and Moritz, C. (2004). Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58:1781–93.

Grinnell, J. (1917). The Niche-Relationships of the California Thrasher. The Auk 34:427–433.

Guénard, B., Weiser, M.D. and Dunn, R.R. (2012) Global models of ant diversity suggest regions where new discoveries are most likely are under disproportionate deforestation threat. Proceedings of the National Academy of Science, 109, 7368-7373.

Guisan, a, and N. Zimmermann. (2000). Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Guisan, A., and C. Rahbek. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography 38:1433–1444.

Guisan, A., and W. Thuiller. (2005). Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guralnick R, Hill A (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25: 421-8.

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22

Hamilton, A. J., Y. Basset, K. K. Benke, P. S. Grimbacher, S. E. Miller, V. Novotný, G. A. Samuelson, N. E. Stork, G. D. Weiblen, and J. D. L. Yen. (2010). Quantifying uncertainty in estimation of tropical arthropod species richness. The American Naturalist 176:90–5.

Hanley, J. A., and B. J. McNeil. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143:29–36.

Hansen, L. A., J. Fjeldsa°, N. D. Burgess, and C. Rahbek. (2007). One degree resolution databases of the distribution of 1789 birds in Sub-Saharan Africa.

Hanski, I. (1999). Metapopulation Ecology. Oxford Universtiy Press, Oxford UK.

Hastie T, Tibshirani R, Friedman J (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, New York.

Hawkins BA (2001) Ecology's oldest pattern? Trends in Ecology and Evolution 16, 470.

Hawkins, B. a. and DeVries, P.J. (2009) Tropical niche conservatism and the species richness gradient of North American butterflies. Journal of Biogeography, 36, 1698-1711.

Hawkins, B. A., M. Rueda, and M. Á. Rodríguez. (2008). What Do Range Maps and Surveys Tell Us About Diversity Patterns? Folia Geobotanica 43:345–355.

Hawkins, B. A., R. Field, H. V. Cornell, D. J. Currie, J.-F. Guegan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, E. O'Brien, E. E. Porter, and J. R. G. Turner.(2003). Energy,water, and broad-scale geographic patterns of species richness. Ecology 84:3105–3117.

Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber.(2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography 16:754–763.

Heisey, D.M., Osnas, E.E., Ross, P.C., Oly, D.O., Langenberg, J.A. & Miller, M.W. (2010) Rejoinder : sifting through model space. Ecology, 91, 3503–3514.

Hepinstall JA, Krohn WB, Sader SA (2002) Effects of niche width on the performance and agreement of avian habitat models. In: Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, Samson FB (eds), Predicting species occurrences,. Island Press, Covelo (CA), pp. 593-606

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modelling methods. Ecography 29: 773-785.

Hof C, Rahbek C, Araújo MB (2010) Phylogenetic signals in the climatic niches of the world's amphibians. Ecography 33: 242-250.

Holloway J.D. (1987). The moths of Borneo, part 3: Lasiocampidae, Eupterotidae, Bombicidae, Brahmaeidae, Saturniidae, Sphingidae. The Malay Nature Society and Southdene Sdn. Bhd., Kuala Lumpur.

Hortal, J., Diniz-Filho, J.A.F., Bini, L.M., Rodriguez, M.A., Baselga, A., Nogues-Bravo, D., Rangel, T.F. Hawkins, B.A. and Lobo, J.M. (2011) Ice age climate, evolutionary constraints and diversity patterns of European dung beetles. Ecology Letters, 14, 741–748.

Hortal, J., N. Roura-Pascual, N. Sanders, and C. Rahbek. (2010). Understanding (insect) species distributions across spatial scales. Ecography 33:51–53.

Hundsdörfer, A.K., Mende, M.B., Kitching, I.J., and Cordellier, M. (2011) Taxonomy, phylogeography and climate relations of the Western Palaearctic spurge hawkmoth (Lepidoptera, Sphingidae, Macroglossinae). Zoologica Scripta 40, 403–417.

Hundsdörfer, A.K., Rubinoff, D., Attié, M., Wink, M., and Kitching, I.J. (2009) A revised molecular phylogeny of the globally distributed hawkmoth genus Hyles (Lepidoptera: Sphingidae), based on mitochondrial and nuclear DNA sequences. Molecular Phylogenetics and Evolution 52, 852–865.

Hurlbert, A. H., and W. Jetz. (2007). Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Sciences (B) 104:13384–13389.

Hutchinson, G. E. (1957). Concluding remarks. Cold Spring Harbor Symposium On Quantitative Biology 22:415–427.

ICZN. (1999). International Code of Zoological Nomenclature, 4th edition. International Trust for Zoological Nomenclature., London.

Jablonski D (2008) Species selection: theory and data. Annual Reviews of Ecology, Evolution and Systematics 39: 501-524.

Jankowski, J. E., A. L. Ciecka, N. Y. Meyer, and K. N. Rabenold. (2009). Beta diversity along environmental gradients: implications of habitat specialization in tropical montane landscapes. The Journal of animal ecology 78:315–327.

Jansen, F., and J. Dengler. (2010). Plant names in vegetation databases - a neglected source of bias. Journal of Vegetation Science 21:1179–1186.

Janzen, D. H. (1984). Two ways to be a tropical big moth: Santa Rosa saturniids and sphingids. Pages 85–140 in R. Dawkins and M. Ridley, editors. Oxford surveys in Evolutionary Biology, 1st edition. Oxford University Press, Oxford.

Jenkins, C.N., Sanders, N.J., Andersen, A.N., Arnan, X., Brühl, C. A., Cerda, X., Ellison, A.M., Fisher, B.L., Fitzpatrick, M.C., Gotelli, N.J., Gove, A.D., Guénard, B., Lattke, J.E., Lessard, J.-P., McGlynn, T.P., Menke, S.B., Parr, C.L., Philpott, S.M., Vasconcelos, H.L., Weiser, M.D. and Dunn, R.R. (2011) Global diversity in light of climate change: the case of ants. Diversity and Distributions, 17, 652-662.

Jetz W, McPherson JM, Guralnick RP (2012) Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology and Evolution 23:151-159

Jetz, W. and Fine, P.V. (2012) Global gradients in vertebrate diversity predicted by historical area-productivity dynamics and contemporary environment. PloS Biology, 10, e1001292

Jetz, W. and Rahbek, C. (2002) Geographic range size and determinants of avian species richness. Science, 297, 1548-51.

Jetz, W., Rahbek, C. and Colwell, R.K. (2004) The coincidence of rarity and richness and the potential signature of history in centres of endemism. Ecology Letters, 7, 1180–1191.

Jiménez-Valverde A (2011) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography 21:498–507.

Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan D., Emmott, S., 2013. Troubling trends in scientific software use. Science 340, 814-815.

Joppa, L. N., D. L. Roberts, N. Myers, and S. L. Pimm. (2011). Biodiversity hotspots house most undiscovered plant species. Proceedings of the National Academy of Sciences (B) 108:13171–13176.

Kawahara AY, Mignault AA, Regier JC, Kitching IJ, Mitter C (2009) Phylogeny and biogeography of hawkmoths (Lepidoptera: Sphingidae): evidence from five nuclear genes. PloS One 4: e5719

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? Oikos 115:186–191.

Keil, P., and B. A. Hawkins. (2009). Grids versus regional species lists: are broad-scale patterns of species richness robust to the violation of constant grain size? Biodiversity and Conservation 18:3127–3137.

Kitching, I., J., and J. M. Cadiou.(2000). Hawkmoths of the world. The Natural History Museum & Cornell University Press, London.

Kreft, H. and Jetz, W. (2007) Global patterns and determinants of vascular plant diversity. Proceedings of the National Academy of Sciences (B), 104, 5925-30.

Kreft, H., and W. Jetz. (2010). A framework for delineating biogeographical regions based on species distributions. Journal of Biogeography 37:2029–2053.

Kremen, C., A. Cameron, A. Moilanen, S. J. Phillips, C. D. Thomas, H. Beentje, J. Dransfield, B. L. Fisher, F. Glaw, T. C. Good, G. J. Harper, R. J. Hijmans, D. C. Lees, E. Louis, R. a Nussbaum, C. J. Raxworthy, A. Razafimpahanana, G. E. Schatz, M. Vences, D. R. Vieites, P. C. Wright, and M. L. Zjhra. (2008). Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. Science 320:222–226.

Kudrna, O., Harpke, A., Lux, K., Pennersdorfer, J., Schweiger, O., Settele, J. and Wiemers, M. (2011) Distribution Atlas of Butterflies in Europe. Gesellschaft für Schmetterlingsschutz. Halle, Germany.

Kumschick, S., Schmidt-Entling, M.H., Bacher, S., Hickler, T., Espadaler, X., and Nentwig, W. (2009) Determinants of local ant (Hymenoptera: Formicidae) species richness and activity density across Europe. Ecological Entomology, 34, 748–754.

Leathwick, J.R., and M. Austin.(2001). Competitive interactions between tree species in New Zealand's old-growth indigenous forest. Ecology 82:2560–2573.

Lennon, J. J., P. Koleff, J. J. D. Greenwood, and K. J. Gaston. (2003). Contribution of rarity and commonness to patterns of species richness. Ecology Letters 7:81–87.

Levin, S. 1992. The Problem of Pattern and Scale in Ecology. Ecology 73:1943–1967.

Lin, Y.-P., Yeh, M.-S., Deng, D.-P. and Wang, Y.-C. (2007) Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. Global Ecology and Biogeography, 17, 175–188.

Linder, H.P., de Klerk, H.M., Born, J, Burgess, N.D., Fjeldsa, J. and Rahbek, C. (2012) The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. Journal of Biogeography, 39, 1189–1205.

Lobo J, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17:145-151

Loiselle BA, Jørgensen PM, Consiglio T, Jiménez I, Blake JG, Lohmann LG, Montiel OM (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? Journal of Biogeography 35: 105-116

Lomolino MV (2004) Conservation Biogeography. In: Lomolino MV, Heaney LR (eds) Frontiers of Biogeography: New directions in the geography of Nature. Sinauer Associates, Inc. Publishers, Sunderland, MA, pp 293-296

MacArthur, R.H. and Wilson, E. (1967). The Theory of Island Biogeography. Princeton University Press, New Jersey.

Mackey, B. G., and D. B. Lindenmayer. (2001). Towards a hierarchical framework for modelling the spatial distribution of animals. Journal of Biogeography 28:1147–1166.

Maddison, D. R., Guralnick, R., Hill, A., Reysenbach, A.-L., and McDade, L.A.(2012). Ramping up biodiversity discovery via online quantum contributions. Trends in Ecology and Evolution 27:72–77.

Marmion M, Hjort J, Thuiller W, Luoto M (2009) Statistical consensus methods for improving predictive geomorphology maps. Computers & Geosciences 35: 615–625

Martin, L.J., Blossey, B. and Ellis, E. (2012) Mapping where ecologists work: biases in the global distribution of terrestrial ecological observations. Frontiers in Ecology and the Environment, 10, 195–201.

Mateo RG, Croat TB, Felicísimo ÁM, Muñoz J (2010) Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Diversity and Distributions 16:84-94

Mateo RG, Felicísimo AM, Muñoz J (2010) Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. Journal of Vegetation Science 21:908-922

Matthew, W.D. (1911) Climate and evolution. Annals of the New York Academy of science 24: 171:318.

Mazzei, P., D. Morel,R. Panfili., I. Pimpinelli, D.Reggianti. (1999-2012). Moths and butterflies of Europe and North Africa. http://www.leps.eu.

McKnight, M. W., P. S. White, R. I. McDonald, J. F. Lamoreux, W. Sechrest, R. S. Ridgely, and S. N. Stuart.(2007). Putting beta-diversity on the map: broad-scale congruence and coincidence in the extremes. PLoS Biology 5:e272.

McPherson, J.M. and Jetz, W. (2007) Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. Global Ecology and Biogeography, 16, 657–667.

Mittelbach, G.G., Steiner, C.F., Scheiner, S.M., Gross, K.L., Reynolds, H.L., Waide, R.B., Willig, M.R., Dodson, S.I. and Gough, L. (2001) What is the observed relationship between species richness and productivity? Ecology, 82, 2381–2396.

Moerman, D.E. and Estabrook, G.F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. Journal of Biogeography, 33, 1969–1974.

Mora, C., Tittensor, D.P. and Myers, R.A. (2008) The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proceedings of the Royal Society (B), 275, 149-155.

Morrone J. J. (2009). Evolutionary Biogeography: An integrative approach with case studies. Columbia University Press, New York.

Muñoz J, Felicísimo ÁM (2004) Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15:285–292

Murray J, Goldizen A (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies Petrogale penicillata. Journal of Applied Ecology 46:842-851

Mutanen, M., Wahlberg, N., and Kaila, L. (2010). Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proceedings of the Royal Society (B) 277:2839–2848.

Newbold T (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography 34:3-22

Newbold T, Gilbert F, Zalat S, El-Gabbas A, Reader T (2009) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. Journal of Biogeography 36:2085-2095

Newbold T, Reader T, Zalat S, El-Gabbas A, Gilbert F (2009) Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. Biodiversity and Conservation 18:3629-3641

Novotny, V., and Weiblen. G. D.. (2005). From communities to continents: beta diversity of herbivorous insects. Annales Zoologici Fennici 42:463–475.

O'Connell, A. F. J., Gilbert, A. T., and Hatfield, J. S.. (2004). Contribution of Natural History Collection Data to Biodiversity Assessment in National Parks. Conservation biology 18:1254–1261.

Opler, P. A., K. Lotts, T. Naberhaus. Coordinators.(2012). Butterflies and Moths of North America.

Palmer, M.W., Earls, P.G., Hoagland, B.W., White, P.S. and Wohlgemuth, T. (2002) Quantitative tools for perfecting species lists. Environmetrics, 13, 121-137.

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225-245

Pearce J, Ferrier S, Scotts D (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. Journal of Environmental Management 62:171–184

Pearson R, Raxworthy C, Nakamura M, Peterson A (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34:102-117

Pearson, R. G. (2007). Species' Distribution Modeling for conservation Educators and Practitioners. Synthesis. American Museum of Natural History: 50.

Pearson, R. G., and T. P. Dawson. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12:361–371.

Pearson, R. G., Dawson, T. P., and Berry, P.M., Harrison, P.A. (2002). SPECIES: A spatial evaluation of climate impact on the envelope of species. Ecological Modelling 154:289-300

Peters, D. P. C., J. E. Herrick, D. L. Urban, R. H. Gardner, and D. D. Breshears. (2004). Strategies for ecological extrapolation. Oikos 106:627–636.

Peterson, A. T., E. Martínez-meyer, and C. González-salazar.(2004). Reconstructing the Pleistocene geography of the Aphelocoma jays (Corvidae). Diversity and Distributions 10:237–246.

Peterson, A. T., V. Sánchez-Cordero, E. Martínez-Meyer, and A. G. Navarro-Sigüenza.(2006). Tracking population extirpations via melding ecological niche modeling with land-cover information. Ecological Modelling 195:229–236.

Phillips SJ, Anderson R, Schapire R (2006) Maximum entropy modelling of species geographic distributions. Ecological Modelling 190: 231-259

Phillips SJ, Dudík M (2008) Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161-175

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19: 181-197

Pittaway, A. R. (1997-2012). Sphingidae of the Western Palaearctic.

Pittaway, A. R., and I. J. Kitching. (2000-2012). Sphingidae of the Eastern Palaearctic (including Siberia, the Russian Far East, Mongolia, China, Taiwan, the Korean Peninsula and Japan).

Pöyry J, Luoto M, Heikkinen RK, Saarinen K (2008) Species traits are associated with the quality of bioclimatic models. Global Ecology and Biogeography 17:403-414

Pulliam, H. R. (2000). On the relationship between niche and distribution. Ecology Letters 3:349–361.

R Development Core Team (2009), 'R: A language and environment for statistical computing. R foundation for statistical computing'

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. Ecology Letters, 8, 224–239.

Rahbek, C., and G. R. Graves. (2001). Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Sciences (B) 98:4534–9.

Rasmussen, J. B., L. A. Hansen, N. D. Burgess, J. Fjeldsa°, and C. Rahbek. 2007. One degree resolution databases of the distribution of 467 snakes in Sub-Saharan Africa.

Reddy, S. and Dávalos, L. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography, 30, 1719-1727.

Regier, J. C., A. Zwick, M. P. Cummings, A. Y. Kawahara, S. Cho, S. Weller, A. Roe, J. Baixeras, J. W. Brown, C. Parr, D. R. Davis, M. Epstein, W. Hallwachs, A. Hausmann, D. H. Janzen, I. J. Kitching, M. A. Solis, S.-H. Yen, A. L. Bazinet, and C. Mitter. (2009). Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. BMC evolutionary biology 9:280.

Renner, S. C., D. Neumann, M. Burkart, U. Feit, P. Giere, A. Gröger, A. Paulsch, C. Paulsch, M. Sterz, and K. Vohland. (2012). Import and export of biological samples from tropical countries–considerations and guidelines for research teams. Organisms Diversity and Evolution 12:81–98.

Robinson, G. ., P. R. Ackery, I. J. Kitching, G. W. Beccaloni, and L. M. Hernández. (2001). Hostplants of the moth and butterfly caterpillars of the Oriental Region. The Natural History Museum and Southdene Sdn Bhd,Kuala Lumpur.

Rothschild, L.W. and K. Jordan. (1903). A revision of the lepidopterous family Sphingidae. Novitates Zoologicae, 9 (suppl.): 1–972.

Ruggiero, A. and Hawkins, B.A. (2008) Why do mountains support so many species of birds? Ecography, 31, 306-315.

Rupprecht F, Oldeland J, Finckh M (2011) Modelling potential distribution of the threatened tree species Juniperus oxycedrus: how to evaluate the predictions of different modelling approaches? Journal of Vegetation Science 22:647-659

Scoble, M. J., Gaston, K. J., and Crook, A. (1995). Using taxonomic data to estimate species richness in geometridae. Journal of the Lepidopterists' Society 49:136–147.

Scoble, M.J., Clark, B., Godfray, H.C.J., Kitching, I.J. and Mayo, S. (2007) Revisionary taxonomy in a changing e-landscape. Tijdschrift voor Entomologie 150, 305-317.

Scott, J.A. (1986). The Butterflies of North America: A Natural History and Field Guide. Standford University Press, Standford.

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. Journal of Biogeography 31:1555-1568

Seoane J, Bustamante J, Diaz-Delgado R (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology 19:512–522

Settele, J., O. Kudrna, A. Harpke, I. Kuehn, C. van Swaay, R. Verovnik, M. Warren, M. Wiemers, J. Hanspach, T. Hickler, E. Kühn, I. van Halder, K. Veling, A. Vliegenthart, I. Wynhoff, and O. Schweiger. (2008). Climatic Risk Atlas of European Butterflies. BIORISK – Biodiversity and Ecosystem Risk Assessment. Pensoft, Sofia.

Soberón, J. and Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society, London (B) 359, 689–698.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. Ecology letters 10:1115–1123.

Soberón, J. M. (2010). Niche and area of distribution modeling: a population ecology perspective. Ecography 33:159–167.

Soberon, J., and Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species distributional areas. Biodiversity Informatics 2:1–10.

Soberon, J., Arriaga, L. and Lara, L. (2002) Issues of quality control in large, mixed-origin entomological databases. Towards a global biological information infrastructure (eds. Saarenmaa H, Nielsen ES), pp. 1–72. European Environment Agency, Copenhagen.

Stefanescu, C., Herrando, S. and Paramo, F. (2004) Butterfly species richness in the north-west Mediterranean Basin: the role of natural and human-induced factors. Journal of Biogeography, 31, 905-912.

Stockwell, D. R. B., and Noble, I. R. (1992). Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Mathematics and Computers in Simulations 33:385–390.

Svenning JC, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand,S (2011) Applications of species distribution modelling to paleobiology. Quaternary Science Reviews 30:2930-2947

Swets, J. A.(1988). Measuring the accuracy of diagnostic systems. Science 240:1285–1293.
Terribile, L. C., Olalla-Tárraga, M. Á., Diniz-Filho, J. A. F. and Rodríguez, M. Á. (2009). Ecological and evolutionary components of body size: geographic variation of venomous snakes at the global scale. Biological Journal of the Linnean Society 98:94–109.

The Times Atlas of the World: Comprehensive Edition. (2010). Times Books, HarpenCollins Publishers. Edition 10th. London
Thomas, C. D., Bulman, C. R., and Wilson, R. J. (2008). Where within a geographical range do species survive best? A matter of scale. Insect Conservation and Ecology 1:2–8.
Thorn, J. S., Nijman, V., Smith, D. and Nekaris, K. a. I.(2009). Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). Diversity and Distributions 15:289–298.
Thuiller W (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9: 1353-1362
Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD–A platform for ensemble forecasting of species distributions. Ecography 32 369-373
Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V. and Worm, B. (2010) Global patterns and predictors of marine biodiversity across taxa. Nature, 466, 1098-1101.
Tuomisto, H.(2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography 33:2–22.
Turner, J. R. G., Gatehouse, C. M. and Corey, C. A. (1987). Does solar energy control organic diversity? Butterflies, moths and the British climate. Oikos 48:195–205.
VanDerWal J, Shoo LP, Graham C, Williams SE (2009) Selecting pseudo-absence data for presence-only distribution modelling: How far should you stray from what you know? Ecological Modelling 220 589-594
Wallace, A. R. 1869. The Malay Archipelago. Oxford in Asia Hardback Reprint (1986). Oxford University Press, Oxford.
Wallace, A.R. (1876). The Geographical Distribution of Animals,Macmillan
Warren, D. L. (2012). In defense of "niche modeling". Trends in Ecology and Evolution 27:497–500.
Warren, M., Robertson, M.P., and Greeff, J.M. (2010). A comparative approach to understanding factors limiting abundance patterns and distributions in a fig tree-fig wasp mutualism. Ecography 33:148–158.
Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279–338.
Whittaker, R.H., Levin, S.A., and Root, R.B. (1972). Niche, Habitat and Ecotope. The American Naturalist 107:321–338.
Wieczorek, J., Guo, Q. and Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18:745–761.
Wilson, E.O. (2003). The encyclopedia of life. Trends in Ecology and Evolution 18:77:80
Wilson, R.J., Z.G. Davies, and C.D. Thomas. (2010). Linking habitat use to range expansion rates in fragmented landscapes: a metapopulation approach. Ecography 33:73–82.
Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008) Effects of sample size on the performance of species distribution models. Diversity and Distributions 14:763-773
Yates CJ, Elith J, Latimer AM, Le Maitre D, Midgley GF, Schurr FM, West AG (2010) Projecting climate change impacts on species distributions in megadiverse South African Cape and Southwest Australian Floristic Regions: opportunities and challenges. Austral Ecology 35: 374–391
Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A. and Culham, A. (2007): How global is the Global Biodiversity Information Facility? PloS One 2, e1124.

Zagmajster, M., Culver, D., Christman, M. and Sket, B. (2010) Evaluating the sampling bias in pattern of subterranean species richness: combining approaches. Biodiversity and Conservation, 19, 3035-3048.

Zuur, A.F., Ieno, E.N. & Elphick, C.S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, **1,** 3-14.