Notes

Cyril Matthey-Doret

August 18, 2016

Choices and methods:

The wasp species used in the analysis were chosen on the basis of: 1. Reproductive mode (see below): Only genera containing asexual species were considered. 2. Information available: Species for which no or little information was known were avoided, as they could not be used in any analysis later. 3. Species of the Chalcidoidea were used as this superfamily has a lot of asexual species, and this taxon is well documented.

Manual dataset:

Data from this set was gathered by hand, in the literature. It covers 138 species, 51 of which are parthenogenetic.

Variables studied:

- **Reproductive mode:** Wether a species reproduce via sexual reproduction, or parthenogenesis. Those are respectively encoded as “sex”" and “asex”. This information was extracted from an earlier list of species compiled by Casper Van Der Kooi and for sexual species, confirmed by searcching the litterature. Species for which the inference of reproductive mode was uncertain were not considered in the analysis to prevent mistakes. Species for which only females were known (based on more than ~ 1-5 specimens) were considered to be asexual.
- **Body length:** Gathered from the litteratude, using various taxonomic keys and scientific books or articles. The size is measured in milimeters, as a single value, or as a range, if available. This measure does not include the ovipositor length.
- **Distribution:**
  - *Number of locations:* Obtained by hand, after removing redundant location names (such as obsolete locations like USSR, if there already is an overlapping location such as Russia, or large countries when more accurate information is present. For example, if USA and New York are present, USA would be dropped in favor of New York.) The filtered list of locations was then entered in a batch geocoding service (ref: http://www.findlatitudeandlongitude.com/batch-geocode/) working with the google maps API and the output coordinates representing the geographical center of each location were stored in a file. Locations are generally countries, but for large countries (e.g. Brazil, Russia, India, USA, China, Canada) states or regions are used, if available. Some locations are also islands.
  - *Biogeographical regions:* Deduced from the list of countries and states used for number of locations. The biogeographical regions used are encoded in the following way: 1-Nearctic, 2-Neotropical, 3-West Palearctic, 4-Ethiopian, 5-East Palearctic, 6-Oriental, 7-Australasian. In the case where a species was found in all 7 biogeographical regions, the notation “WW” was used, thus considering the species is distributed worldwide.
  - *Latitude:* The lowest and highest latitude values at which a species was found were extracted from the geocoding output list, using a small python script. Those values represent the extreme northern and southern coordinate that will be used to compute latitude range (i.e. *max - min*) and maximum distance from equator (i.e. *max(abs(max),abs(min))*), in further analyses.

  - *Longitude:* longitude coordinates were extracted following the same steps as latitude. The extremes were extracted, representing the most western and eastern coordinates, considering greenwich as the starting point. These values will be used in combination with latitude in order to compute the area of distribution for the different species.

- **Host range:**
  - *Host order:* Host orders were encoded in a binary way. If a wasp species has host in a given order, those will be denoted with a "1", whereas the orders which do not host the species are encoded with a 0. Throughout the dataset, there are 8 orders of host insects, and also 3 taxons of host plants.
  - *Host family:* Additionaly to the 8 orders, 3 families of hemiptera have been stored: Aphididae (aphids), Aleyrodidae (whiteflies) and Coccidae (scale insects), since they are especially relevant in this biological system.
  - *Host species:* The number of host species was retrieved from the nhm database, by visually counting them for every wasp species.

- **Pair:** Pairs or comparison groups of asexual and sexual species have been formed, based on molecular phylogenies if available. If molecular data was not available, morphologic data was used.

- **Ecology:** Encoded as "phyt", standing for phytophagous, if species had plant hosts, or as "par", for parasitoid, if they did not. This allows to distinguish between phytophagous and parastoid wasp species, as they have different biologie. This variable was inferred from the "Host order" information.

Issues:

- Some variables had to be dropped:
  - Originally, wing dimorphism was used, but since most of the time, sexual dimorphism happen in antennae or segment length and not in the wings, the variable was dropped.
  - Development time was originally used, but since it depends a lot on other factors that are not necessarily mentioned in studies, such as temperature, temperature fluctuations, host instar, day/light cycles or genotype, the variable was dropped.
  - Host reproduction: Aphelinidae species have many host with different reproduction modes, almost never exclusively sex/asex.

- Distribution is biased by two factors:
  - Number of studies done on species; Species that are interesting for biocontrole are more studied, and sampled in more countries. That is eeven more true for the automated dataset, described further.
  - Insertions in new locations for biocontrole, followed by establishment in the region. This results in messy distributions. Many well studied species are all over the world.

Automated dataset:

Data from this set was exclusively retrieved from the Universal Chalcidoidea Database, on the Natural History Museum website. This dataset covers 8357 species, 51 of which are parthenogenetic. Only genera for which at least one asexual species was known to us was used in the set. The data has been scraped from the website, for each of the species URL, for both distribution and host species. This was done with a python script, using the package "urllib2" to open the URL's and store them as python objects, and the package "BeautifulSoup4" to parse the html code and extract informations.

Distribution : A python script was used to scrape location lists of the nhm database for each species. This was done by opening each species' URL using the urllib2 package and then parsing these URLs with the

BeautifulSoup 4 package. Since the database is made to be human readable, location names sometimes caused issues (e.g. chinese provinces names are followed by a non-official phonetic name) and some old location have become obsolete (this is especially true for regions in the former USSR). Those locations names had to be transformed into correct names using a hand-made list of exceptions. This allowed the locations to be passed to a geocoding service (Nominatim, website:https://www.openstreetmap.org) via the python package geopy and transformed into coordinates automatically. The script made a similar transformation as what was done by hand, by removing all redundant (i.e. countries when state is already present), obsolete (e.g. USSR) and irrelevant (very broad regions such as Continents or biogeographic regions) locations.

Host range : A python script was used to scrape host species, families and orders directly of the nhm website, using the same procedure as for the locations (i.e. urllib2 and BeautifulSoup packages).

Variables studied:

- **Reproductive mode:** The reproductive mode was inferred from the list provided by Casper. All species retrieved from the database that were not present in the list were considered to be sexual.
- **Distribution:**
  - *Latitude:* Latitude was obtained using the open source geocoding service Nominatim (website: https://www.openstreetmap.org). The location names extracted from the database were passed to the service via the python package "geopy", and the output coordinates were stored.
  - *Longitude:* Longitude was obtained along with latitude.
  - *Number of locations:* The number of locations mentioned in the database was modified in the same way as for the manual dataset; if locations belong to a given country and that country is also present as a value in the list, the more accurate location is kept and the country is removed. The redundant locations such as "Europe"", or"Palearctic"" are also removed, they do not contribute to the number of locations and their coordinate are not stored. In total, 477 different locations are used for this dataset.
- **Host range:**
  - *Host order:* The script identified the number of host orders present in the list of hosts using html parsing and regular expressions.
  - *Host family:* The same method is used for families.
  - *Host species:* The number of host species was identified based on the number of times a specific html tag used for species was encountered in the list of primary hosts.
- **Ecology:** The ecology was only estimated in each genera and all species from genera which were estimated to have >50% phytophagous species were considered to be phytophagous species, otherwise they were considered parasitoids. These values have also been obtained by detecting a specific tag, but they should only be used for indicative purpose as they will probably underestimate of the number phytophagous species (see "Issues"" below).

Issues:

- Since the database was made to be human-readable, some of the locations names were not fit for analysis via the geocoding service. For instance, chinese provinces had a phonetic name appended to their real name. Those modified names cause error when sent to the program and had to be taken care of by using an exception list which transformed invalid names into corresponding valid names and sent them to the program once transformed. A second exception list was used for the broader, less informative locations such as continents and biogeographical areas, simply removing these.
- The ecology was not tested species-wise because the database was not reliable for this information. Indeed, the script used a specific section header to detect the presence of host plants on a URL, but sometimes there are inconsistencies and host plant are listed in the "primary host section" instead of the "host plant section", which would result in the absence of the header and cause the script to

count the species as a parasitoid. This is the reason the ecology was only estimated as a proportion of phytophagous species in each genus and these values will be underestimated.

- The exception lists should normally include all wrong location names and invalid locations for the set of species used, except if the database is modified. The only type of error that could normally happen is when 2 locations share the same name and the wrong name has been used. This kind of error depends on the geocoding service and could be checked by passing the list of 477 locations through other geocoding services and detecting only major differences, since they will always be small discrepancies probably caused by the way center is calculated or small changes in border shape.

Manual vs Automated:

Comparing data obtained by hand and automatically reveals some differences, however thos are rather minor and should not be too problematic since they are consistent over all species.

Distribution : Differences in distribution are caused by exception list which may be caused by counting mistakes in manual version.

Host range : Difference in number of host species are caused by «unspecified» species; while those were considered as a species on their own when counted by hand, the script does not take them into account since they are given a different structure (i.e. empty field in an html tag) than other species in the html code. None of the 2 method is wrong since those records could either be member of an existing species already in the list, or member of a new species, or a species not in the list. There are a few cases where large differences can be noticed. This is when manual data was gathered from litterature in cases where the database seemed unreliable or not up to date. The manual and automated versions will be used in 2 different analyses. While the manual version uses pairs to compare asexual versus sexual species, the automated version will be used to perform broader comparisons on taxonomic levels (genera and families), comparing asexual species based on Casper's list, with all other species on the database being considered sexual.

Variables transformations:

Extreme values of latitude will be used to calculate the range of distribution (i.e. max-min) and the maximum absolute distance from the equator. Longitude and latitude will be used together to compute the area of distribution by multiplying both ranges.

Statistical analysis:

A generalized linear model was used, for both the automated dataset and the manual dataset. The manual model counts pairs as a random effect nested in genus, reproductive mode as a fixed effect and takes a quantitative variable such as the number of host species or latitude range as response variable. The model for the automated data counts genus as a fixed effect nested in family, and reproductive mode as another fixed effect. It also takes one of the quantitative variable as response variable.