**What makes a great speciator?**

Andrea Estandia1, Nilo Merino-Recalde1, Bruce C. Robertson2, Sonya M. Clegg1

1Edward Grey Institute of Field Ornithology, Department of Zoology, University of Oxford,, OX1 3SZ, Oxford, United Kingdom

2Department of Zoology, University of Otago, Dunedin 9054, New Zealand

**Abstract**

XXXXX

**Introduction**

Great speciators have puzzled evolutionary biologists for more than half a century due to their paradoxical ability to colonise entire archipelagoes while diversifying into multiple subspecies despite potential gene flow. Mayr and Diamond (2001) identified a set of 31 Northern Melanesian bird species as great speciators based on their number of subspecies and allospecies and the number of islands were they were present. The original list was produced by applying arbitrary cut-offs and it included representatives of several groups of birds from raptors to passerines, but they had some things in common: they were short-to-intermediate colonisers, they were abundant, and many of them were montane. Diamond et al. (1976) proposed that the recognition of a class comprising short-distance colonists partly resolved the paradox. If a species rarely colonised nearby islands but the new established populations survived and readily altered their habitat preference as a consequence of competition, this could lead to phenotypic divergence on each of the islands. This idea was also central in Wilson (1961) evolutionary framework “the taxon cycle”, which aims to explain the speciation process in archipelagos. This framework encapsulates the idea that island taxa go through phases of range expansions and contractions. The cycle starts with the colonisation of many islands by a species that will eventually differentiate on each of those islands, giving rise to phenotypically different populations (great speciators). However, there are some great speciators that are found thousands of miles away from Northern Melanesia. These birds should have shown an excellent long-distance dispersal ability that allowed them to colonise remote islands, but such ability is costly to maintain (Bonte et al. 2012) so it would be then strongly selected against after colonisation leading to sedentary populations in a matter of generations (Diamond et al. 1976). This idea gained popularity, despite the fact that most great speciators do not show long-distance dispersal, and it has been widely explored yielding examples of rapid evolutionary shifts in dispersal ability (Pedersen et al. 2018). However, it is still unclear what shapes the unique diversity and distribution of great speciators.

Language taxonomy follows a power law-distribution (Wichmann 2005), and this is likely the case for any hierarchical system, including biological taxonomy. The ubiquity of power-law distributions is a result of processes of a stochastic nature (Buraczewski et al. 2016). Great speciators could easily be the tail of this distribution, with only a few species that contain many subspecies. However, no study has considered that the number of subspecies that we observe in great speciators might emerge from stochastic processes alone and we might not need to invoke changes in dispersal ability.

The term ‘great speciator’ has been applied in an inconsistent way in the literature. In its original incarnation, the term applied to species with a more than 7 subspecies across their range (restricted to Northern Melanesia) or if the ratio between number of subspecies and number of occupied islands was greater than 0.33. This concept was never used to describe a specious family or genus, but a species that showed high levels of phenotypic diversity (Mayr and Diamond 2001). However, the term has been appropriated for higher taxonomic levels: for example, *Zosterops griseotinctus* was classified as one of the ‘original’ great speciators, but subsequent studies have conferred that status to the entire species-rich Zosteropidae family (Moyle et al. 2009)—even though some members do not have any subspecies or exclusively live on mainlands. Another common misperception is that great speciators have rapid diversification rates, whether applied at subspecies, species or high levels (Moyle et al. 2009; Pedersen et al. 2018). However divergence rate was not a consideration in the original definition, and as species diversification rates are not correlated with subspecies richness…something else needed here (Phillimore et al. 2006).

Here we revisit the great speciator concept to identify potential species that fit the great speciator concept using a quantitative approach. We apply the original definition by Mayr and Diamond (2001) and a custom definition to two datasets at different spatial scales, i) all the islands of the world (global island dataset) and ii) Northern Melanesia. We build phylogenetic regression models to identify ecological variables that may explain the number of subspecies given the opportunities they have to phenotypically diverge. Finally, we build spatially explicit models under a neutral scenario to explore whether geography alone can explain the distribution and diversity observed in great speciators, or whether is necessary to invoke another mechanism (e.g. low extinction rates or post-colonisation changes in dispersal ability).

**Methods**

*Species Distribution Data*

We collected resident and breeding distribution data for 10,079 bird species from BirdLife International and matched the shapefiles to the Clements Checklist v2019.

Using vectorised maps downloaded from Natural Earth (ref) we created two landmass shapefiles i) a ‘global island’ shapefile where each closed polygon was considered an independent landmass, retaining only those that had a smaller area than Madagascar. This size selection was chosen as the great speciator concept is an inherently allopatric one, and sympatric differentiation fundamentally depends on area and isolation of the island (McArthur and Wilson 1963) , and on dispersal and population density (Diamond 1977). We found that sympatric subspeciation is frequent on islands the size of Madagascar and larger. However, there were still some species that contained more than two subspecies that inhabited the same island. Those rare cases were excluded from the dataset; ii) the second dataset was based on the regional definition of Mayr and Diamond (2001) for Northern Melanesia. This dataset was produced for comparative purposes with the original work and for updates on taxonomy.

We intersected each of the species distribution shapefiles with each of the landmass shapefiles (the global island and Northern Melanesia) using the R package *sp,* permitting calculation of the number of islands where each species was present. We mapped subspecies range descriptions for each species available in the Clements Checklist, and counted the number of subspecies per species across the islands included in the ‘global island’ dataset and the Northern Melanesia dataset.

*Great Speciator Index*

We used the original index proposed by Mayr and Diamond (2001) for the Northern Melanesia dataset to compare the original great speciators and the ones emerging with the current taxonomy. Their classification has two main rules: i) If the species in question has seven or more subspecies across its range, it is automatically classified as a great speciator; and ii) if the ratio of number of subspecies/number of islands occupied by the species is higher than 0.33 and the species has from four to six subspecies, it also belongs to the great speciators. This classification has its limits. The arbitrary cut-offs were probably chosen because they fit their observations at the regional scale, but once this set of rules are applied to larger geographical scales these thresholds stop making sense. First, with increasing distribution range, the number of species also increases, but not in a linear fashion (Phillimore et al. 2007). While in Northern Melanesia we can have species distributed in a whole archipelago that have a substantial number of subspecies yielding a ratio greater than 0.33, if we increase range size to a global dataset, this does not hold. Many species have more than seven subspecies because their distribution range is very wide and even if there were very mobile, they probably could not maintain an unstructured population (except for seabirds). For example, the peregrine falcon (*Falco peregrinus*) is present in 3574 islands in our ‘global island’ dataset and has 8 subspecies being present in most continents. The original definition would classify it as a great speciator. However, the peregrine falcon is likely to generate subspecies due to isolation by distance rather than by the mechanisms thought to give rise to great speciators. Further, the second rule proposed by Mayr and Diamond (2001) would classify any species with four subspecies in four different islands as a great speciator. At the regional scale this might represent species that live on nearby islands, which indicates that despite having high chances of gene flow they have differentiated. However, at a global scale even if we double the number of subspecies to eight species distributed on eight different islands, if these islands are separated enough, it would not represent the essence of the great speciators which is a dispersal paradox.

We develop a 'Great Speciator Index' (GSI) that tries to solve the problems described above. This index only has a slight modification of the original index, it is also based on the species distribution (number of islands) and number of subspecies across these islands, but the number of subspecies is divided by the logarithm of the sum of the number of landmasses where the species is distributed (Equation 1). Another key difference is that we do not apply any cut-off.

GSI is higher if a species contains many subspecies and it is distributed across several islands, but it will be smaller if with the same distribution it contains fewer subspecies. The following examples aim to illustrate the behaviour of the index. The island thrush (*Turdus poliocephalus*) is the species with the highest number of subspecies (43 subspecies). This species lives on 82 islands in a moderately restricted geographical area (Oriental and Oceanian regions according to the Holt et al. (2013) framework). On the same number of islands, the pheasant coucal (*Centropus phasianinus*) only has 4 subspecies. The GSI for the island thrush will naturally be higher than for the pheasant coucal. Species that have a different subspecies on each of the inhabited islands, will have a higher GSI as more subspecies have. The same GSI would be obtained for a species distributed across 20 islands with 7 subspecies, as one distributed across 270 islands with 13 subspecies.

*Ecological variables*

We calculated range size as the distance between the two most distant polygons within the distribution of a species. Relative island density for each species was calculated by creating a distance matrix among all islands and calculating the median distance. To capture the geographic distribution of occupied islands, we created a distance matrix among all occupied islands for each species, and extracted the median island distance.

We calculated the intersection between species distribution, average annual temperature, average annual precipitation and altitude obtained from WorldClim v.1 with a resolution of 2.5 minutes of a degree.

The following species-level morphological and ecological variables were used: hand-wing index (HWI) and body size from (Sheard et al. 2020), migratory behaviour from Eyres et al. (2017), diet from Wilman et al. (2014), and habitat from Tobias et al. (2016).

Diversification rates

We used BAMM (Bayesian Analysis of Macroevolutionary Mixtures) to calculate tip-specific diversification rates on the Hackett topology (Hackett et al. 2008). Even though many studies after Jetz et al. (2012) use the two backbone topologies used in the original study (Hackett et al. (2008) and Ericson et al. (2006)), we decided to only use the Hackett topology based on the assumption that tip-specific rates are robust to topology uncertainty (Rabosky et al. 2015). We computed the maximum clade credibility tree for the Hackett topology only including those species for which genetic data was available (6663 species). We controlled for incomplete taxon sampling by calculating the percentage of species per family included in the analysis, according to the taxonomy used in Jetz et al. (2012).

We used MCMC with four independent chains and 200 million iterations each, discarding 10% as burn-in. We set a mean of 0.02 on the rate parameter of the Poisson distribution as it is recommended for trees containing more than 5000 tips. Default priors were applied for the rate parameters. To test for convergence we calculated the effective sample size (ESS).

*Regression models*

We fit Bayesian phylogenetic generalised linear mixed models in *brms* (Burke 2017). We tested the effect of ecological variables on the number of subspecies per species relative to the number of islands where the whole species is present. We ran models for the two subsets (world and Northern Melanesia) separately.

The binomial response variable was ‘number of subspecies’ versus ‘number of islands minus number of subspecies’. This captures subspeciation ‘opportunity’ such that 10 subspecies distributed across 10 islands would have a response variable (10, 0), having taken all opportunities to subspeciate, whereas for five subspecies distributed across 10 islands it would be (5, 5), with five opportunities ‘missed’. We applied binomial error distribution with logit link to determine the effect of the following explanatory variables: hand-wing index, body mass, range size, relative isolation of islands, temperature range, precipitation range, altitude range, diversification rate, diet, realm, habitat and migratory status. We modelled the interaction between HWI and body mass, and among temperature, precipitation and altitude ranges. Body mass, range size, and relative isolation of islands were log-transformed because…. All quantitative variables were scaled from 0 to 1. To control for phylogenetic relatedness, we included a phylogenetic variance-covariance matrix (P-matrix) calculated from the Jetz et al. (2012) phylogeny as a random variable.

We ran four independent chains with 4000 iterations, discarding 1000 as burn-in.

[delete – captured belowWe also ran an intercept-only model, that only included the response variable and the phylogeny. ]

To check for convergence, we inspected the caterpillar plots, analysed the posterior predictive checks, and checked that the ESS was greater than 500 (following ref).

To explore the strength of phylogenetic signal (the tendency of related species to resemble each other more in the response variable (?) than species drawn at random from the same tree), we calculated an alternative to the Intraclass Correlation Coefficient (ICC), that accommodates non-Gaussian models by decomposing the variance based on the posterior predictive distribution.

We re-ran the models excluding those explanatory variables for which we had high levels of missing data to increase the number of observations and maximise the degrees of freedom. In these models we restricted the explanatory variables to the following set: range size, relative isolation of islands, temperature range, precipitation range, altitude range and diversification rate.

To evaluate model performance we compared our fitted models with an intercept-only model (including only the response variable and the P matrix), using leave-one-out cross-validation (LOO), a robust, fully Bayesian model selection approach (Vehtari et al., 2017).

*Simulations*

The model consists of a square grid where every cell has a different probability of being colonised, capturing geographic distance. Each simulation is individually based. Species start at a random cell, colonising other cells and establishing populations. New metapopulations can remain connected to other populations if they are re-colonised in a period of X steps, but they will subspeciate if re-colonisation does not occur. Metapopulations can go extinct too. Area is the same for all cells and population size varies depending on the number of individuals colonising a new island.

Each time step in a model run proceeds through a simulation of extinction, dispersal, reproduction. We also implement carrying capacity. If the number of individuals in a population reaches X, 70% of the individuals will randomly die. However, reproduction continues to happen at the same rate.

1) Extinction phase: each occupied cell independently faces a Bernoulli trial against the probability of extinction (binomial distribution where there are two outcomes: extinction or survival as a population). Extinction probability is invariant across cells and time.

2) Dispersal phase: each occupied cell is examined for the possibility that the species in a cell is able to colonise each unoccupied cell throughout the grid. It will be more likely to colonise closer cells, but not all adjacent or close cells have the same probability of being colonised. The probability of colonisation of each cell is extracted from real distances among islands in an archipelago.

3) Speciation phase: the speciation rate is fixed. A subspeciation event will only depend on the number of steps that the population has spent isolated. If immigrants arrive to the island at any point (before or after) the subspeciation event, the steps for subspeciation are reset and the metapopulation will become part of the original species. For obtaining subspecies status again, it will need to wait the same iterations as when it first colonised it.

We extracted the probability of colonisation as the distance among islands in Northern Melanesia, this led to a 73x73 matrix. We ran the simulation in 321 fictional species, as this is the number of island birds seen in Northern Melanesia. Iterations?

We compared the distribution of subspecies resulting from the simulations with those observed from the same region.

Birds comprise the most well-studied group of vertebrates that occurs across archipelagos (Lerner et al., 2011; Lamichhaney et al., 2015). The active colonising ability of volant birds is unmatched by members of any other terrestrial vertebrate group, and this is reflected in the distribution of bird species across the worlds’ islands, particular isolated oceanic islands. Nevertheless, bird species vary in their likelihood of successful island colonisation, in large part due to variation in dispersal propensity and capability (Diamond et al. 1976).

Results