Original Research Article

A simple method for assessing the completeness of a geographic range size estimate

Garima Gupta a,*, Jonathon Dunn a, Roy Sanderson a, Richard Fuller b, P.J.K. McGowan a

a Newcastle University, UK
b University of Queensland, Australia

ARTICLE INFO

Article history:
Received 2 May 2019
Received in revised form 13 September 2019
Accepted 13 September 2019

Keywords:
Geographic range size
Galliformes
Locality database
Conservation

ABSTRACT

Measuring geographic range size is a fundamental part of ecology and conservation. Geographic range size is used as a criterion by the IUCN Red List of Threatened Species in estimating species extinction risk. Yet the geographic distributions of many threatened species are poorly documented, and it is often unclear whether a geographic range size estimate is complete. Here we use a large and near-exhaustive database of species occurrences to (i) estimate extent of occurrence (a measure of geographic range size routinely used in Red List assessments), and (ii) develop a method to assess whether our estimate for each species is complete. We use an extensive database of point locality records for 24 Himalayan Galliformes, a group of highly threatened bird species. We examine the chronological pattern of increase of geographic range size estimates and compare this accumulation curve with a null model generated by performing 1000 iterations for each species using the point locality information in random order. Using Generalised Estimation Equations (GEE) and Generalised Least Square (GLS), we show that estimates of geographic range size for most species has now asymptoted, and that the range size estimates have improved more rapidly over time than expected by chance, suggesting relatively efficient sampling over time. The approach used in this study can be used as a simple method for assessing the completeness of a geographic range size estimates for any taxon.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The geographic distribution of a species is fundamental to understanding its ecology and conservation needs, and there has been much research analysing the spatial occurrence of biodiversity (Gaston, 2000; Myers et al., 2000; Hawkins et al., 2003; Koleff et al., 2003; Orme et al., 2005; Naidoo et al., 2008). The size of geographic range of a species plays a prominent role in categorizing species according to their short-term likelihood of extinction, including listing on the IUCN Red List of Threatened Species (Gaston and Fuller, 2009), as well as how their distributions may change in response to anthropogenic perturbations such as habitat loss (Channell and Lomolino, 2000; Ceballos and Ehrlich, 2002) and climate change (Parmesan and Yohe, 2003; Thomas et al., 2004). Small absolute range size, or rapid declines in range size can indicate a high risk of
imminent extinction, because species with small geographic range are more vulnerable to stochastic threats than species with widespread distributions, and declining geographic range can lead to population reductions (Bland et al., 2016).

It therefore follows that knowledge of species distributions influences conservation efforts at all scales (Margules and Pressey, 2000; Whittaker et al., 2005). Knowing where a species occurs is important as it allows conservationists to make an accurate assessment of threats for individual species. It also allows us to understand global patterns of biodiversity in relation to threats (Joppa et al., 2016), which enables conservationists to identify how best to ameliorate threats and to target conservation actions. The distributions of species are also commonly used to evaluate the coverage of protected areas, and to inform the placement of new protected areas (Venter et al., 2014; Watson et al., 2014; Butchart et al., 2015).

There are several ways of describing geographic range size (Gaston and Fuller, 2009), but all methods rely ultimately on information about species occurrences. Our knowledge of these occurrences is generated from field records of individual taxa often collected for reasons far removed from those for which they might be used in macro-ecological or applied conservation analyses. Such data collection is labour intensive, requires a high level of expertise, and is expensive. Consequently, only a very small proportion of the planet has so far been covered by systematic spatial surveys (Price et al., 1995; Hagemeijer and Blair, 1997), and the comprehensiveness of distributional data varies spatially and temporally with factors such as observer effort, taxon detectability and ease of identification (Bibby et al., 2000; Boakes et al., 2010).

There is a potential for much of the information used in large-scale spatial analyses to be biased, particularly for tropical species, where species richness is very high and taxonomy poorly known. For example, no tree species has been accurately mapped in the Amazon basin, and there are significant known taxonomic biases in estimates of size of species' geographic range (Pitman et al., 1999; Ruokolainen et al., 2002). If such biases are widespread across taxa and regions, spurious patterns may arise in large-scale analyses such as those described above, and substantial errors in extinction risk estimation could be made. Despite the improving knowledge and availability of data sets on a wide range of species, our understanding of species' geographic distribution remains inadequate (Whittaker et al., 2005; Rondinini et al., 2006; Jetz et al., 2012).

Here, we develop a framework for testing the efficiency of our sampling of species' geographic range that could in principle be applied to any spatial dataset prior to conducting extinction risk assessments or large-scale ecological analyses. The underlying principle of the modelling framework is that we gain more information about the distribution of individual species the more effort we spend surveying, but that all else being equal, the information gained will eventually asymptote as we move towards a position of perfect knowledge of a species' distribution. In this case, the more records we have of an individual species, the more likely we are to get a more complete picture of the distribution. In the absence of systematic sampling, we assume that knowledge about the distribution will accrue with time as records are made opportunistically. In effect, we expect that the overall estimate of geographic range size will be asymptotically related to the number of records.

Distribution data for Himalayan Galliformes are analysed here, but we emphasise that the method is generally applicable. Specifically, we explore the size of the geographic range of 24 species using a large and near-exhaustively collected dataset of localities to assess the completeness of our geographic range size estimate. Largely restricted to forested habitats, most Himalayan Galliformes are severely affected by hunting and habitat loss, and many are declining (Fuller and Garson, 2000).

Ultimately, we will never know the “true” size of the geographic range of any species. Instead, we suggest examining the pattern of accumulation of information on a species' geographic range over time and comparing this with a null model. Here we test: a) the completeness of our knowledge of size of species’ geographic range; b) whether our knowledge of the geographic range of this group of birds has improved more rapidly than expected by chance; and c) whether this improvement has accelerated toward the present.

2. Methods

2.1. Bird records

The point locality data were extracted from GALLIFORM: Eurasian DatabaseV.10 (Boakes et al., 2010), which contains data accurate to 0.62—30 miles. This database contains records on 131 Galliformes species from a wide range of sources including museum specimens, references, and trip reports (Fig. 1). The data were opportunistically collected, hence is a ‘presence-only’ dataset with no absence points. In addition, we cannot exclude the possibility that the species could be found in areas where surveys have not been undertaken or survey effort was incomplete.

The 24 species occupying the Greater Himalayas are studied here. The study area covers approximately seven million square kilometres of north-west and north-east Indian states, northern Pakistan, Nepal and Bhutan, representing the major parts of the Greater Hindu-Kush Himalayan mountain system, delineated by 11 WWF ecoregions (Eastern Himalayan alpine shrub and meadows; Eastern Himalayan broadleaf forests; Eastern Himalayan subalpine conifer forests; Himalayan subtropical broadleaf forests; Himalayan subtropical pine forests; Northeastern Himalayan subalpine conifer forests; Northern Triangle temperate forests; Northwestern Himalayan alpine shrub and meadows; Western Himalayan alpine shrub and meadows; Western Himalayan broadleaf forests; and Western Himalayan subalpine conifer forests; see Dunn (2015)) in the Greater Himalaya (Wikramanayake et al., 2002). For each species, all records from the date of their first occurrence up to 2007, when the last records were entered into the database, were used to create shape files. Records without a year or geographical co-ordinates were omitted.
2.2. Area accumulation curve: modelling historical sampling of geographic range

Point locality records were arranged in chronological order (henceforth ‘historical records’). At the addition of each record for a species, we constructed the minimum convex polygon (MCP) around the occurrence locations. The MCP measures the extent of occurrence of a species, defined as “the area contained within the shortest continuous imaginary boundary which can be drawn to encompass all the known, inferred or projected sites of present occurrence of a taxon, excluding cases of vagrancy” (IUCN, 2012). Extent of occurrence is a measure of spatial risk-spreading that is not inferior to methods that attempt to map occupied areas, but simply has a different purpose (Gaston and Fuller, 2009). Extent of occurrence is routinely used for in IUCN Red List assessments of extinction risk (Santini et al., 2019).

Records were added in chronological order, and a new MCP and extent of occurrence derived at each step. This process was iterated until the final locality record from the most recent year was added. This resulted in a series of MCPs for each species based on cumulative year (e.g. if the earliest year was 1950, the first MCP was constructed based on all 1950 records. Therefore, the MCP for 1951 was based on all records from 1950 plus those from 1951, and so on for all subsequent MCPs).

For each species, the MCP area was plotted as a function of year and as a function of count for number of records. The resulting accumulation curves were then compared to the simulated curve derived by randomising the addition sequence as described below.

Fig. 1. Geographic range accumulation curves for satyr tragopan (Tragopan satyra). A, Comparison of historical (red curves) and simulated area accumulation curves (green curves) for year. B, Comparison of historical and simulated area accumulation curves for number of records. When the area of the range based of historical records exceeds that of the range based on simulated records, it indicates that our knowledge of that species’ range is better than random. Plots based on N = 1000 iterations and show mean range areas ± 1SE. Note: Standard errors estimated at each point are so small that they cannot be represented on the plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)
2.3. Generating the random simulation model

The random accumulation curves were generated by performing 1000 iterations for each species in which the point locality records used in MCP construction were added in a random and not chronological order (henceforth ‘simulated records’).

For each iteration, we created a new column called ‘year random’, which was based on the actual year column but the order of the year was shuffled. Therefore, the number of record counts for each unique year was the same, but the locality information associated with the years was different. Data were summarised to obtain mean MCP area ± 1SE for each year and for each record count across the 1000 iterations. The mean MCP area was plotted as a function of year for year records and as a function of count for count records, both with the standard error bars. The simulated curves thus represented the predicted range size estimate after the addition of each record if all parts of the geographic range were sampled with equal probability. This is a Monte Carlo approach.

2.4. Analysis of asymptote

To assess the completeness of our estimate of size of species geographic range, we tested whether the historical and simulated accumulation curves had reached an asymptote. An asymptote indicated that the addition of more records did not change the MCP area estimate, indicating that new knowledge has not increased our estimate of the species’ geographic range. Failure to reach an asymptote indicated that our knowledge of that species’ geographic range was incomplete and still expanding spatially. To assess whether an asymptote had been reached, we undertook the following procedure: 1) identify total number of records and the total area of the geographic range; 2) identify the year or number of records that corresponded to 80% of the total number of records; 3) calculate the difference between total MCP area and the MCP area that corresponded to 80% of the total number of records; and then 4) area accumulation curves were considered asymptotic when the final 20% of the records added less than 10% to the geographic range size area estimate. There is no standard threshold by which an asymptote is identified in this context: the 20% and 10% figures used here are arbitrary, but reasonable approximations.

2.5. Statistics

To examine the completeness of our knowledge of species’ geographic range size, we used McNemar’s test to determine whether the number of species with an asymptote was similar for the random accumulation curve vs. historical accumulation curve. Data were paired for each species and coded 1 where the area accumulation curve reached an asymptote and 0 where the curve did not reach asymptote.

To test whether our knowledge of species’ geographic range has improved more rapidly than expected by chance, we used logistic regression models to compare historical and simulated area accumulation curves separately for each individual species. We hypothesised that the historical and simulated curves for each species would be temporally autocorrelated in that the value in any one year would be dependent to some extent on the values in previous year. We created a binomial variable (1/0) indicating that the simulated area in one year was greater (1) or less than (0) than observed. We assessed the trend (1/0) in accordance with time using Generalised Estimating Equations (GEE) with an autoregressive error structure.

In this way, MCP area was implicitly assumed to reflect range knowledge, with larger MCP areas indicating better range knowledge. Thus, for both the logistic regression and GEE models, the predictor variable was year and the response variable was whether historical range area exceeded or was less than simulated range area. The only difference was that each single-species logistic regression model used a single row of data for each year, whereas the multi-species GEE used multiple rows of data for each year for each species. A significant positive logistic regression model would indicate that our knowledge has improved more rapidly than expected by chance.

Finally, we used generalised least squares models to test whether our improvement in knowledge has accelerated towards the present. To do this, we calculated the difference in area actually observed in each year and that predicted from the simulated range we then investigated if there was any trend in this difference with year. The 1970s marks the time when there was a change in the forest policies and new legislation was enacted in the Himalayan region to protect the forests after the demonstrations by “hill tribes” against ongoing deforestation in the Greater Himalaya (Shah, 2008). The dependent variable was whether historical range area exceeded or was less than simulated range area (coded as 1/0 as before) and the independent variable was whether the time period was before/after 1970. If the improvement in knowledge has accelerated towards the present, we predicted the probability of obtaining a 1 to be greater post-1970 than pre-1970.

3. Results

The random simulation models showed that sampling all areas of a geographic range with equal probability should lead to an asymptotic area accumulation curve (see example given in Fig. 1), with the probability of each new record falling within the known MCP range increasing as each record is added. The actual historical patterns of geographic range size estimates generally produced sigmoidal area accumulation curves i.e. knowledge initially increased slowly, with a rapid phase of improvement before finally reaching an asymptote. Using the number of records added or year of record as the independent variable produced graphs of similar pattern.
Sixteen out of 24 (66.6%) species’ historical range accumulation curves reached asymptote whereas 20 out of 24 (83.3%) species’ random range accumulation curves reached asymptote (see Table 1).

Table 1 shows that for 16 of 24 Himalayan species, the curve for historical records has reached an asymptote suggesting that sampling effort for those species is good and hence, our knowledge of those species’ geographic range size is complete. For eight species the curve did not reach an asymptote for the historical records, although the simulated curve reached asymptote for six of the species, suggesting that sampling effort is not adequate to determine geographic range size robustly in those cases.

Common hill partridge (*Arborophila torquata*) and koklass pheasant (*Pucrasia macrolopha*) did not reach an asymptote for either historical or simulated records suggesting that either sampling is not sufficiently broadly distributed in space and time, or that more survey effort is needed. The curve for Blyth’s tragopan (*Tragopan blythii*) and Sclater’s monal (*Lophophorus impejanus*) reached an asymptote for historical records but not for the simulated records, suggesting that survey effort for these two species has been better than random.

McNemar’s test showed that there was no difference in the number of species with historical accumulation curves that reached an asymptote and those with simulated accumulation curves that reached an asymptote (*McNemar*’s $\chi^2 = 1.13$, df = 1, p-value = 0.29). This suggests that sampling, and thus our knowledge of all Himalayan Galliformes species’ ranges reflect reality and that overall we know the range of our species rather well.

Our estimates of size of geographic range have improved more rapidly than predicted by chance for 16 of the 24 (66.6%) species while the opposite was true in the remaining eight cases (Table 1). When all species were pooled together, the GEE model suggested that for the majority of species, geographic range knowledge has improved more rapidly than at random through time ($\beta = 0.00529$, SE = 0.002, p < 0.028).

The difference between actual area and the random area increased over time, and after 1970 the difference became positive, which indicates that after 1970 the actual (historical) area was greater than that predicted (simulated) using the simulated model. However, the estimates of area were not independent of each other, in that records in any one year would also contribute to the observed area and the simulated areas in subsequent years, indicating that the data were serially correlated. This means that the estimates of significance are likely to have been biased. Therefore, to avoid this problem an autocorrelation component was included in the GLS, which meant that the contribution of time to increased area was not significant (which is actually implausible).
4. Discussion

We found that our knowledge of the geographic range sizes of Himalayan Galliformes is generally rather complete, has improved rapidly over time, and has accelerated since 1970. An intensive research and survey programme was developed after 1970 (McGowan et al., 1999; Fuller et al., 2000), and similar efforts have occurred for many other highly threatened groups. We now have evidence that this effort laid solid knowledge foundations of species distribution in this group.

Despite this, knowledge of the geographic range size of two of the 24 species (koklass pheasant and common hill partridge) appears to remain incomplete, suggesting that sampling efforts are still insufficient to describe the complete geographic range of these species. If such knowledge gaps are representative of birds globally, then many hundreds of species might still have incomplete estimates of geographic range size.

In spite of having reasonably large numbers of records, the range accumulation curve for koklass pheasant (*Pucrasia macrolopha*) did not reach an asymptote for either historical or simulated records. This might be due to the fact that koklass pheasant has an extremely large suspected geographic range (BirdLife International, 2019) and we need more survey effort to accurately quantify its geographic range size. However, it could also be that early survey efforts were focussed in only a few areas of the suspected total range before 1970, as a positive linear regression model for koklass pheasant suggests that knowledge of geographic range size of koklass pheasant has accelerated towards the present.

We found that of the six threatened Himalayan Galliformes species, geographic range size knowledge of five species, namely cheer pheasant (*Catreus wallichii*), Blyth’s tragopan (*Tragopan blytii*), Himalayan quail (*Ophrysia superciliosa*), chestnut-breasted partridge (*Arborophila mandellii*) and Selater’s monal (*Lophophorus sclateri*) has not accelerated towards the present. This could be because the small, fragmented population of cheer pheasant has a patchy distribution (BirdLife International, 2019) that was previously understudied, or that species such as Blyth’s tragopan and Selater’s monal occur at least partly in areas that were difficult to access historically. For the Critically Endangered Himalayan quail, the species has not been reliably recorded since 1876 (BirdLife International, 2019) suggesting that thorough surveys are required, as there is a possibility that the species may be rediscovered (Dunn et al., 2015).

To our knowledge, this is the first attempt to evaluate sampling effort across species’ geographic ranges. Once this method has been used to identify whether geographic ranges have been described adequately, other techniques may be used to extend our understanding and help focus conservation research efforts further. For example, Grainger et al. (2018) developed a Bayesian belief network for a highly threatened bird species, Edwards’s pheasant (*Lophura edwardsi*) to assess the probability of its persistence and where surveys or other conservation action should be targeted in light of suspected uncertainty in its distribution.

The biggest constraint in identifying the complete geographic range of a species is the paucity of documented species occurrence records. Often survey effort is heavily biased in space and time (Tingley and Beissinger, 2009) and surveyors tend to focus on areas rich in biodiversity for documenting localities (Boakes et al., 2010). Also, habitats where species of interest have been recorded in the past may be more likely to be surveyed subsequently. Consequently, some habitats may be undersurveyed where other similar habitats have few records, possibly because of unrelated environmental conditions. This makes it difficult to identify the true geographic range of a species, as areas with other biodiversity values are often understudied.

Our results show that examining data chronologically may enable the identification of taxa for which further geographic data are required, and provide a way of prioritising taxa and areas for further survey work. The methods we outline here can also help identify biases in survey efforts since it is crucial to resolve the current spatial biases in biodiversity monitoring to correctly estimate extinction risk (Boakes et al., 2016).

The main limitation of our approach is that the accumulation curves are unable to distinguish between where a species’ geographic range has expanded or contracted and where survey effort has been better targeted. Geographic range expansion in this case, however, seems unlikely to any meaningful extent because these species are largely, if not entirely, sedentary (except for the common quail *Coturnix coturnix*), and have quite specific habitat requirements. There may be a similar issue surrounding species detectability with the diverse methods that have been used over time (collecting specimens, targeted surveys, and birder trip reports). Whilst it is not possible, therefore, to distinguish between a geographic range expansion and an increasing ability to detect and record the species with time, the long-standing keen interest in collecting these species, hunting them, and now recording them seems likely to have ensured that detectability has remained fairly constant despite the use of different detection methods. Further simulation modelling could discover the effect of different geographic range change trajectories and changes in detectability on accumulation curves. For example, if a species’ geographic range has declined, it is unlikely this will be reflected in the historical accumulation curve.

This study provides a novel means to examine the quality of a locality dataset and assess the robustness of geographic range size estimates. The importance of using geographic information appropriately in global conservation priority-setting cannot be overstated. However, MCP is a first step to assess a geographic distribution, and estimates of the area occupied within the extent of occurrence will be important for planning conservation actions.

Declaration of competing interest

I confirm that there are no conflict of interests with this manuscript.
