CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

DOCTORAL THESIS

MAPPING SPECIES DISTRIBUTION RANGES BY MEANS OF HETEROGENEOUS DATA

Laura Ríos Pena

ESCOLA DE DOUTORAMENTO INTERNACIONAL

PROGRAMA DE DOUTORAMENTO EN INGENIERÍA DO MEDIO RURAL E CIVIL

LUGO

2018

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DO AUTOR/A DA TESE

Mapping species distribution ranges by means of heterogeneous data

Dna. Laura Ríos Pena

Presento a miña tese, seguindo o procedemento axeitado ao Regulamento, e declaro que:

1) A tese abarca os resultados da elaboración do meu traballo.
2) De selo caso, na tese faise referencia ás colaboracións que tivo este traballo.
3) A tese é a versión definitiva presentada para a súa defensa e coincide coa versión enviada en formato electrónico.
4) Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En Lugo, 29 de junio de 2018

Asdo. Laura Ríos Pena

Esta tesis se llevó a cabo en la Estación Biológica de Doñana (EBD-CSIC) con financiación procedente del programa de ayudas Severo Ochoa para la formación de doctores en Centros de Excelencia Severo Ochoa, Subprograma Estatal de Formación, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016 del Ministerio de Economía y Competitividad (MINECO) (SVP−2013−067977) y de la Agencia Estatal de Investigación del Ministerio de Economía, Industria y Competitividad, España con los proyectos CGL2012-35931 y CGL2017-83045-R AEI/FEDER EU, cofinanciado con FEDER a E.R., R.B.M., M.G.S.

AUTORIZACIÓN DOS DIRECTORES

Mapping species distribution ranges by means of heterogeneous data

Dr. Eloy Revilla Sánchez
Estación Biológica de Doñana – CSIC, Sevilla (España)

Dr. Miguel Clavero Pineda
Estación Biológica de Doñana – CSIC, Sevilla (España)

INFORMAN:

Que a presente tese, correspóndese co traballo realizado por Dna. Laura Ríos Pena, baixo a nosa dirección, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como director desta non incorre nas causas de abstención establecidas na Lei 40/2015.

En Lugo, 29 de junio de 2018

Directores:

Asdo. Eloy Revilla Sánchez

Asdo. Miguel Clavero Pineda

Tutor:

Asdo. Manuel Fco. Marey Pérez
Universidade de Santiago de Compostela

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AGRADECIMIENTOS

La elaboración de una tesis doctoral es una tarea compleja, prolongada en el tiempo, y en la que se entremezclan periodos de optimismo con otros de desánimo. Quiero dejar aquí reflejados los nombres de algunas de las personas que contribuyeron a los primeros o mitigaron los segundos.

En primer lugar, quiero dar mis más sinceras gracias a mis supervisores, Dr. Eloy Revilla y Dr. Miguel Clavero, por brindarme la oportunidad para conocer este "mundo de la biología de la conservación" que para mí hasta ese mismo instante era totalmente desconocido. Por todo el apoyo y el asesoramiento recibido y por toda la paciencia y familiaridad con la que me trataron. Ha sido un viaje nuevo, interesante y por su puesto de gran provecho.

Otra persona a la que le estoy muy agradecida y al que me gustaría dedicarle unas palabras es al tutor de esta tesis, Dr. Manuel Marey. Conocí a Marey en 2011 debido a que era el profesor de una asignatura de la carrera que estaba cursando. Posteriormente, Marey se ofreció a supervisar mi proyecto fin de carrera, una oferta que fue el comienzo de mi interés por la investigación. También me animó a estudiar un Máster en Técnicas Estadísticas y así obtener una sólida base en estadística, que, desde las primeras etapas de mi doctorado, me di cuenta que había sido un gran consejo. Por todos estos aspectos han hecho que para mí, Marey, sea un excelente mentor.

En lo que respecta a mi estancia en el departamento de estadística de la Universidad de Gotinga (Alemania), debo reconocer al Profesor Thomas Kneib y Dr. Nadja Klein por su ayuda incondicional en cualquier proceso de computación y utilización de herramientas estadísticas que, tanto en el periodo de estancia como a posteriori, he ido necesitado. Ha sido un placer trabajar y colaborar con vosotros.

Más recientemente, y por un periodo de tres meses, tuve la oportunidad de trabajar en Research Centre in Biodiversity and Genetic Resources (CiBiO/InBIO) en la Universidad de Lisboa. Gracias al Prof. Henrique Pereira por brindarme la oportunidad para desempeñar parte del trabajo del tercer capítulo de esta tesis y al Dr. César Capinha con quién trabajé más de cerca que, sin ningún tipo de obligación, estuvo ahí siempre que lo necesité. Agradezco enormemente su buena disposición para discutir mi trabajo. En este punto

también quiero dar las gracias a toda la gente del grupo, pues el ambiente era muy familiar y desde el primer día me hicieron sentir como uno más de la casa.

Gracias a Sara Varela, que aunque no fue directora de esta tesis, así lo sentí. Por todos sus buenos consejos, apoyo y comprensión recibida, que no ha sido poca, porque desde el primer momento en que nos pusimos en contacto, el trabajo en conjunto, los e-mails y Skype me ayudaron a trepar por una montaña que cada vez se me iba haciendo más cuesta arriba. Parte de todo esto es gracias a ti.

Gracias al personal del LAST e informáticos de la EBD, por el indispensable soporte técnico.

A mis compañeros de la EBD, en especial a la gente del despacho y del café de las once, los que estaban cuando llegué, los que se han ido y los que se han incorporado con el paso del tiempo, con quienes he compartido (y continúo compartiendo) muchas vivencias y donde las risas y los buenos momentos fueron siempre en común denominador de cada día.

Un agradecimiento especial a mis amigos de A Coruña y Sevilla, con quienes he pasado buenos momentos tanto antes (A Coruña) como durante mi doctorado (A Coruña y Sevilla) y que espero que continúe para hacerlo.

A toda mi familia, en especial a mis padres que nunca dejaron de apoyarme y de creer en mí. Por todos sus consejos, su cariño y comprensión a lo largo de toda la vida. Soy muy afortunada de teneros y siempre seréis mi guía y ejemplo a seguir.

Finalmente, a Raúl, gracias por enseñarme que el mejor regalo de la vida es compartirla y aprovecharla con la persona que siempre me saca una sonrisa, con la que me apoya y confía en mí. Gracias por ser paciente, por escucharme y ayudarme. Gracias por ser tú.

A mis padres

"If we knew what it was we were doing, it would not be called research, would it?"

Albert Einstein (1879 – 1955)

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Index

ABSTRACT

The mapping of species ranges is one of the most relevant and widely used pieces of information in the study of biodiversity. Knowing the distribution range of species is a fundamental first step us to understand the factors that determine those distributions, as well as the patterns in the richness and abundance of species in a biogeographical context, all this being necessary information to establish conservation strategies. The distribution range is a conceptual construction that describes the area where a taxon occurs. The basic units of information for constructing these ranges are spatially and temporally referenced observations of species (i.e. records). Direct field sampling on very large spatial scales is rarely feasible, as it requires significant resources and time. Therefore, large-scale biodiversity analyses tend to be based on a variety of data reporting information on species observations or distributions, ranging from point location data obtained from databases or wildlife atlases to species distribution maps based on expert knowledge. In spite of been essential, our knowledge on the distribution of species is far from complete, even for the best studied taxa. Given the great relevance of species distribution maps, it is surprising to note that very little attention has been paid to analyse how these maps are affected by the quality of the baseline data and the diversity of methods used to construct them. This is the central axis of the thesis, which structured in four main chapters.

In **Chapter I** we conducted a bibliographic review in order to obtain information from scientific publications that use species distribution ranges in their studies. We noted how distribution ranges have been generated and identified which are the most commonly used methods to generate distribution ranges from georeferenced data, along with the advantages and disadvantages provided by each of them. Most often researchers do not provide information on how ranges have been constructed. The lack of explicit information on the data and methods used in the construction of distribution ranges severely affect the interpretation of results. Finally, the methods commonly used to delineate the areas have been insufficiently evaluated. We urge researchers to be explicit both in what they consider the ranges of distribution of species and in the methods they use to generate them. This will allow

for more robust comparisons between the ranges of distribution of species generated by different methods.

In **Chapter II** we assessed the accuracy of five geographic algorithms commonly used to delineate species ranges with the aim of providing guidelines to minimize Type I error and maximize sensitivity of the resulting species ranges. To this aim, we generated hypothetical range areas with the same total surface but varying in shape, number of fragments, heterogeneity in fragment size and simulated sets of species records varying in numbers, spatial distribution and presence of errors and biases. The recommended algorithms have been Adaptive Local Convex Hull (a-LoCoH) and Kernel Density Estimation (KDE). KDE algorithm has the highest sensitivity and a-LoCoH algorithm has the lowest type I error rate. Both behaved similarly well when describing range fragmentation. We provide recommendations to minimize the effects of data quantity and quality, and provide guidance to choose an algorithm when defining species distribution ranges based on species observations.

Chapter III of this thesis explores options for a systematic and replicable generation of range maps that take into account the different sources of variability and the exponential increase in the availability of species records. We offer a unified and repeatable methodology for building species range maps, which we compare with the existing maps of the International Union for Conservation of Nature (IUCN). The combination of IUCN distribution maps with georeferenced species data available from the Global Biodiversity Information Facility (GBIF) is a promising route to providing information on where mapped distributions are reliable and where they are uncertain. Lack of information or availability of information in certain areas makes it difficult to implement systematic approaches to the construction of distribution maps. So we also reveal priority sites for lack of information or sampling effort on a global scale.

Chapter IV assesses the variability in the description of species distribution ranges based on non-systematic data gathering (e.g. using records from available databases) or on systematic and specific surveys. As a case study, we used the southern water vole (*Arvicola sapidus*) in peninsular Spain, using the results of a citizen science initiative specifically focussed on this species and comparing them with those of a previous atlas. The resulting distribution maps had notable differences, which were related to identification errors and

heterogeneous sampling effort in the non-systematic dataset as well as to actual changes in range due to predation by invasive American mink. The likelihood of commission errors increases in areas where there are species that may be confused with the water vole and by mink predation. The probability of omission errors increases in areas with low sampling effort and the existence of rodents easily confused with the study species. We emphasize the need to be cautious in using available information sources to generate range maps, particularly in areas with little data or signs of heterogeneous spatial coverage.

In conclusion, this thesis explores the different dimensions of species distribution maps and offers a necessary perspective to deal with problems posed by sciences such as ecology or conservation biology. We also try to understand the nature of the uncertainty involved in distribution maps to help interpret existing results and guide future research. The information metrics developed throughout this thesis could be incorporated into online tools that allow researchers and funding agencies to identify priority species and areas to improve information sources along with their associated distribution maps.

RESUMEN

El mapeo de áreas de distribución de especies es una de las piezas de información más relevantes y ampliamente utilizadas en el estudio de la biodiversidad. Conocer el área de distribución de las especies es un primer paso fundamental para entender los factores que determinan esas distribuciones, así como los patrones de riqueza y abundancia de las especies en un contexto biogeográfico, siendo toda esta información necesaria para establecer estrategias de conservación. El área de distribución es una construcción conceptual que describe el área donde ocurre un taxón. Las unidades básicas de información para la construcción de estas áreas son las observaciones de referencia espacial y temporal de las especies (es decir, los registros). El muestreo directo sobre el terreno a escalas espaciales muy grandes rara vez es factible, ya que requiere recursos y tiempo considerables. Por lo tanto, los análisis de biodiversidad a gran escala tienden a basarse en una variedad de datos que reportan información sobre observaciones o distribuciones de especies, que van desde datos de localización de puntos obtenidos de bases de datos o atlas de vida silvestre hasta mapas de distribución de especies basados en el conocimiento de expertos. A pesar de ser esencial, nuestro conocimiento sobre la distribución de las especies está lejos de ser completo, incluso para los taxones mejor estudiados. Dada la gran relevancia de los mapas de distribución de especies, es sorprendente observar que se ha prestado muy poca atención al análisis de cómo estos mapas se ven afectados por la calidad de los datos de línea de base y la diversidad de los métodos utilizados para construirlos. Este es el eje central de la tesis, que se estructura en cuatro capítulos principales.

En el **Capítulo I** realizamos una revisión bibliográfica para obtener información de publicaciones científicas que utilizan áreas de distribución de las especies en sus estudios. Observamos cómo se han generado e identificado las áreas de distribución que son los métodos más comúnmente utilizados para generar áreas de distribución a partir de datos georreferenciados, junto con las ventajas y desventajas proporcionadas por cada uno de ellos. En la mayoría de los casos, los investigadores no proporcionan información sobre cómo se han construido las áreas. La falta de información explícita sobre los datos y métodos utilizados en la construcción de las áreas de distribución afecta severamente a la

interpretación de los resultados. Por último, los métodos utilizados habitualmente para delimitar las zonas no se han evaluado suficientemente. Instamos a los investigadores a ser explícitos tanto en lo que consideran áreas de distribución de las especies como en los métodos que utilizan para generarlas. Esto permitirá realizar comparaciones más sólidas entre las áreas de distribución de las especies generados por diferentes métodos.

En el **Capítulo II** evaluamos la exactitud de cinco algoritmos geográficos comúnmente utilizados para delinear las áreas de distribución de las especies con el objetivo de proporcionar directrices para minimizar el error de Tipo I y maximizar la sensibilidad de las áreas de distribución de las especies resultantes. Con este objetivo, generamos áreas de distribución hipotéticas con la misma superficie total pero variando en forma, número de fragmentos, heterogeneidad en el tamaño de los fragmentos y conjuntos simulados de registros de especies variando en número, distribución espacial y presencia de errores y sesgos. Los algoritmos recomendados han sido Adaptive Local Convex Hull (a-LoCoH) y Kernel Density Estimation (KDE). El algoritmo KDE tiene la sensibilidad más alta y el algoritmo a-LoCoH tiene la tasa de error tipo I más baja. Ambos se comportaron similarmente bien al describir la fragmentación del área. Proporcionamos recomendaciones para minimizar los efectos de la cantidad y calidad de los datos, y proporcionamos orientación para elegir un algoritmo a la hora de definir las áreas de distribución de las especies en base a las observaciones de las especies.

El **Capítulo III** de esta tesis explora las opciones para una generación sistemática y replicable de mapas de áreas de distribución que tengan en cuenta las diferentes fuentes de variabilidad y el aumento exponencial en la disponibilidad de registros de especies. Ofrecemos una metodología unificada y repetible para construir mapas de áreas de distribución de especies, que comparamos con los mapas existentes de la Unión Internacional para la Conservación de la Naturaleza (UICN). La combinación de los mapas de distribución de la UICN con los datos de especies georreferenciados disponibles del Fondo Mundial para la Información sobre la Biodiversidad (GBIF) es una vía prometedora para proporcionar información sobre dónde son fiables los mapas de distribución de especies y dónde son inciertos. La falta de información o la disponibilidad de información en determinadas zonas dificultan la aplicación de enfoques sistemáticos para la elaboración de mapas de distribución.

Así que también revelamos sitios prioritarios por falta de información o esfuerzo de muestreo a escala global.

El **Capítulo IV** evalúa la variabilidad en la descripción de las áreas de distribución de las especies basándose en la recolección de datos no sistemáticos (por ejemplo, usando registros de bases de datos disponibles) o en encuestas sistemáticas y específicas. Como caso de estudio se utilizó el topillo de las agua (*Arvicola sapidus*) en la España peninsular, utilizando los resultados de una iniciativa de ciencia ciudadana centrada específicamente en esta especie y comparándolos con los de un atlas anterior. Los mapas de distribución resultantes presentaban diferencias notables, relacionadas con errores de identificación y esfuerzos heterogéneos de muestreo en el conjunto de datos no sistemáticos, así como con cambios reales en el área de distribución debido a la depredación por el visón americano invasor. La probabilidad de errores de comisión aumenta en áreas donde hay especies que pueden ser confundidas con el topillo de agua y por la depredación del visón. La probabilidad de errores por omisión aumenta en áreas con bajo esfuerzo de muestreo y la existencia de roedores fácilmente confundibles con la especie estudiada. Hacemos hincapié en la necesidad de ser cautelosos al utilizar las fuentes de información disponibles para generar mapas de área de distribución, en particular en zonas con pocos datos o signos de cobertura espacial heterogénea.

En conclusión, esta tesis explora las diferentes dimensiones de los mapas de distribución de especies y ofrece una perspectiva necesaria para abordar problemas planteados por ciencias como la ecología o la biología de la conservación. También tratamos de entender la naturaleza de la incertidumbre involucrada en los mapas de distribución para ayudar a interpretar los resultados existentes y guiar la investigación futura. Las métricas de información desarrolladas a lo largo de esta tesis podrían ser incorporadas en herramientas en línea que permitan a los investigadores y agencias de financiamiento identificar especies y áreas prioritarias para mejorar las fuentes de información junto con sus mapas de distribución asociados.

RESUMO

O mapeo das áreas de distribución de especies é unha das pezas de información máis relevantes que se usan ampliamente no estudo da biodiversidade. Coñecer a área de distribución da especie é un primeiro paso fundamental para comprender os factores que determinan esas distribucións, así como os patróns de riqueza e abundancia da especie nun contexto bioxeográfico, toda esta información é necesaria para establecer estratexias de conservación. A área de distribución é unha construción conceptual que describe a zona onde ocorre un taxón. As unidades básicas de información para a construción destas áreas son as observacións espaciais e temporais de referencia da especie (é dicir, os rexistros). A mostraxe directa no chan a escalas espaciais moi grandes raramente é factible, xa que require un tempo e recursos considerables. Polo tanto, as análises de biodiversidade a gran escala tenden a basearse nunha variedade de datos que reportan información sobre observacións ou distribucións de especies, que van dende datos de localización de puntos obtidos a partir de bases de datos ou atlas de vida salvaxe ata mapas de distribución de especies baseados no coñecemento de expertos. A pesar de ser esencial, o noso coñecemento sobre a distribución das especies está lonxe de ser completo, incluso para os taxons mellor estudados. Dada a gran relevancia dos mapas de distribución de especies, é sorprendente observar que se prestou moi pouca atención á análise de como estes mapas están afectados pola calidade dos datos base e a diversidade dos métodos utilizados para construílos. Este é o eixe central desta tese, que está estruturada en catro capítulos principais.

No **Capítulo I** realizamos unha revisión bibliográfica para obter información de publicacións científicas que usan áreas de distribución de especies nos seus estudos. Observamos como se xeraron e identificaron as áreas de distribución, que son os métodos máis utilizados para xerar áreas de distribución a partir de datos xeorreferenciados, xunto coas vantaxes e desvantaxes proporcionadas por cada un deles. Na maioría dos casos, os investigadores non proporcionan información sobre como se construíron as áreas. A falta de información explícita sobre os datos e métodos utilizados na construción das áreas de distribución afecta gravemente á interpretación dos resultados. Por último, os métodos xeralmente utilizados para delinear áreas non foron suficientemente valorados. Instamos aos

investigadores a que sexan explícitos tanto no que consideran áreas de distribución de especies como nos métodos que utilizan para xeralas. Isto permitirá facer comparacións máis sólidas entre as áreas de distribución das especies xeradas por diferentes métodos.

No **Capítulo II** foi valorada a precisión de cinco algoritmos xeográficos comunmente utilizados para delinear as áreas de distribución das especies, a fin de proporcionar directrices para minimizar o erro de Tipo I e maximizar a sensibilidade das áreas de distribución das especies resultantes. Para este fin, foron xeradas áreas distribución hipotéticas coa mesma superficie total, pero variando en forma, número de fragmentos, heteroxeneidade no tamaño dos fragmentos e conxuntos simulados de rexistros de especies variando en número, distribución espacial e presencia de erros e sesgo. Os algoritmos recomendados foron Adaptive local Convex Hull (a-Locoh) e Kernel Density Estimation (KDE). O algoritmo KDE ten a maior sensibilidade e o algoritmo de a-Locoh ten a menor taxa de erro de tipo I. Ambos se comportaron de forma similar ao describir a fragmentación da área. Proporcionamos recomendacións para minimizar os efectos da cantidade e calidade dos datos e proporcionar orientación para elixir un algoritmo ao definir as áreas de distribución da especie con base nas observacións da especie.

O **Capítulo III** desta tese explora as opcións para unha xeración sistemática e replicable de mapas de distribución das áreas que teñan en conta as distintas fontes de variabilidade e o aumento exponencial na dispoñibilidade de rexistros de especies. Ofrecemos un sistema unificado e repetible para construír mapas de distribución das especies, en comparación cos mapas existentes da Unión Internacional para a Conservación da Natureza (UICN). A combinación da distribución de mapas de especies da UICN e datos xeorreferenciados dispoñibles a partir do Global Biodiversity Information Facility (GBIF) é un camiño prometedor para proporcionar información sobre onde son fiables os mapas de distribución das especies e onde son incertas. A falta de información ou dispoñibilidade da información en certas áreas dificultan a aplicación de enfoques sistemáticos para a elaboración de mapas de distribución de especies. Así que tamén revelamos áreas prioritarias por falta de información ou esforzo de mostraxe a escala global.

O **Capítulo IV** avalía a variabilidade na descrición das áreas de distribución da especie baseándose na recollida de datos non sistemáticos (por exemplo, utilizando rexistros de bases de datos dispoñibles) ou en enquisas sistemáticas e específicas. Como caso de estudo

utilizouse o topillo das auga (Arvicola sapidus) na España peninsular, utilizando os resultados dunha iniciativa científica cidadá centrada especificamente nesta especie e comparándoos cos dun atlas anterior. Os mapas de distribución resultantes presentaron diferenzas notables, relacionadas con erros de identificación e esforzos de mostraxe heteroxéneos no conxunto de datos non sistemáticos, así como cambios reais na área de distribución debido á depredación do visón americano invasor. A probabilidade de erros de comisión aumenta en áreas onde hai especies que poden confundirse co topillo de auga e pola depredación do visón. A probabilidade de erros por omisión aumenta en áreas con baixo esforzo de mostraxe e a existencia de roedores fácilmente confundidos coas especies estudadas. Destacamos a necesidade de ser cautelosos ao utilizar as fontes de información dispoñibles para xerar mapas de área de distribución, especialmente en áreas con poucos datos ou signos de cobertura espacial heteroxénea.

En conclusión, esta tese explora as distintas dimensións dos mapas de distribución de especies e ofrece unha perspectiva necesaria para abordar problemas derivados de ciencias como a ecoloxía ou a bioloxía de conservación. Tamén intentamos comprender a natureza da incerteza involucrada nos mapas de distribución para axudar a interpretar os resultados existentes e orientar a futura investigación. As métricas de información desenvolvidas ao longo desta tese poderían incorporarse a ferramentas en liña que permitan aos investigadores e axencias de financiamento identificar áreas de especies e prioridades para mellorar as fontes de información e os seus mapas de distribución asociados.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

RESUMEN EN LENGUA CASTELLANA DE MÁS DE 3000 PALABRAS

La biodiversidad se distribuye de forma heterogénea por toda la Tierra. Conocer los lugares en los que están presentes las diferentes especies es uno de los principales objetivos de las ciencias naturales, especialmente en disciplinas como la biogeografía, la macroecología y la biología de la conservación. Un conocimiento preciso de la distribución de las especies permite describir los patrones geográficos de la biodiversidad, informar el manejo y conservación de los recursos naturales, identificar áreas prioritarias para la conservación o investigar las relaciones evolutivas a través del espacio (Margules et al., 2002; Rondinini et al., 2011). El área de distribución de las especies (u otro nivel taxonómico) es una construcción conceptual que describe el área donde está presente un taxón o especie. Las unidades básicas de información para la construcción de estas áreas son las observaciones de referencia espacial y temporal de las especies (es decir, los registros). El muestreo directo sobre el terreno a escalas espaciales muy grandes rara vez es factible, ya que requiere recursos y tiempo considerables. Por lo tanto, los análisis de biodiversidad a gran escala tienden a basarse en una variedad de datos que reportan información sobre observaciones o distribuciones de especies, que van desde datos de localización de puntos obtenidos de bases de datos o atlas de vida silvestre hasta mapas de distribución de especies basados en el conocimiento de expertos. A pesar de ser esencial, nuestro conocimiento sobre la distribución de las especies está lejos de ser completo, incluso para los taxones mejor estudiados.

El área de distribución puede caracterizarse en términos de su tamaño, forma y otros descriptores de sus límites, fragmentación o estructura interna (Brown et al., 1996, Lucas et al., 2016). Como herramienta conceptual, el área de distribución proporciona una descripción resumida de la compleja dinámica espacio-temporal de las poblaciones. La caracterización de las áreas de distribución depende de cómo se definen, la calidad y cantidad de los datos de línea de base disponibles y el enfoque metodológico elegido para construirlos; temas que a menudo se pasan por alto en la literatura científica.

Definición de las áreas de distribución de las especies

Tal y como se ha definido anteriormente, y dado que este concepto será tratado a lo largo de la presente tesis doctoral, el área de distribución es una construcción conceptual que define un espacio topológico en el que se supone que la especie o taxón está presente dadas las observaciones y la resolución espacial y temporal impuestas. Sin embargo, este concepto a veces se contextualiza en la literatura científica de diferentes maneras, con el potencial de confusión cuando se usa el concepto de una manera no transparente.

La UICN, en la evaluación más influyente del estado de conservación de las especies (UICN, 1994, 2001), define la extensión de la ocurrencia (EOO) como el área contenida dentro del límite continuo más corto que abarca todos los sitios de ocurrencia actual de un taxón. El EOO puede incluir discontinuidades o disyunciones dentro de la distribución general de los taxones, tales como grandes áreas de hábitat obviamente inadecuado. El área de ocupación (AOO) es un subconjunto del EOO y describe el área donde una especie está realmente presente (Gaston, 1991; 2003). Estos dos parámetros se utilizan en los protocolos de la UICN para evaluar el estado de conservación de las áreas de distribución (Gaston and Fuller, 2009; UICN Standards and Petitions Subcommittee, 2010). De la misma manera, otras definiciones de áreas de distribución también se utilizan en la literatura científica actual para generar mapas de áreas de distribución de especies. Algunos se basan exclusivamente en registros georreferenciados, y otros utilizan estimaciones de idoneidad ambiental junto con registros georreferenciados, que pueden traducirse en áreas en las que supuestamente se cubren los requisitos ambientales de la especie. Sin embargo, la distribución de una especie no sólo está determinada por el nicho ecológico, sino también por las barreras de dispersión, las interacciones bióticas y los factores históricos (Oswald et al., 2016, Husáková y Münzbergová, 2016, Schloss et al., 2012). Los modelos ecológicos de nicho, más frecuentemente conocidos como modelos de distribución de especies (MDF), son herramientas metodológicas utilizadas para delinear las áreas donde se cumplen las condiciones para la existencia de una especie, basándose en los datos de ocurrencia conocidos y las condiciones ambientales en esos lugares. Por lo tanto, los MDFs por definición no identifican las áreas de distribución de las especies. Sin embargo, este salto de área de distribución a área de distribución potencial ocurre frecuentemente en la literatura.

Datos de biodiversidad

Bajo el explosivo aumento de los datos globales, el término "big data" se utiliza para describir enormes conjuntos de datos. Estos grandes datos generan nuevas oportunidades para descubrir nuevos valores y también incurren en nuevos desafíos al tratar de organizar y manejar estos conjuntos de datos de manera efectiva (Maldonado et al., 2015; Stephenson et al., 2017). En ciencias como la ecología o la biología de la conservación, las bases de datos de la ciencia ciudadana se están convirtiendo en una forma importante de recopilar información sobre la distribución de las especies (Dickinson et al., 2012; Tiago et al., 2017). Las observaciones recogidas por un gran número de voluntarios, en grandes extensiones espaciales y períodos temporales, a menudo proporcionan un gran número de registros (Chandler et al., 2012), lo que permite realizar estudios que de otro modo serían inviables. El incremento de los registros de especies a partir de las iniciativas de ciencia ciudadana en los últimos años es particularmente importante para grupos taxonómicos visibles y fáciles de identificar. La posibilidad de recoger, a través de aplicaciones móviles con conexión a Internet, observaciones georeferenciadas del mundo natural (por ejemplo, avistamientos de fauna) a través de interfaces interactivas de geovisualización (por ejemplo, Google Maps, Google Earth y Microsoft Virtual Earth) o el uso de sensores en dispositivos móviles nos permite recoger una gran cantidad de datos del entorno. Además de la gran oportunidad que ofrecen las plataformas de ciencia ciudadana, las bases de datos de biodiversidad también agregan información publicada (libros, monografías, artículos o actas de congresos), colecciones de historia natural, información recogida en encuestas, encuestas específicas o repositorios en línea (Soberón y Peterson, 2004; Guralnick et al. 2007). Por lo tanto, las bases de datos sobre biodiversidad proporcionan una gran cantidad de información heterogénea y las iniciativas para generar, almacenar y conectar estas bases de datos también han proliferado en las últimas décadas.

Ambiciosas infraestructuras internacionales como el Fondo Mundial de Información sobre la Biodiversidad (GBIF, http://www.gbif.org/) tratan de vincular todas estas colecciones de bases de datos sobre biodiversidad entre países y continentes. GBIF es en la actualidad la base de datos de biodiversidad más grande y más ampliamente utilizada (Beck et al., 2012,

2014; Jetz et al., 2012). El objetivo de GBIF es "hacer que los datos primarios del mundo sobre biodiversidad estén libre y universalmente disponibles a través de Internet" (Yesson et al., 2007; GBIF, 2008). Actualmente, GBIF proporciona un portal único para acceder a más de 975 millones de registros (en abril de 2018). Esta disponibilidad masiva de datos sobre la biodiversidad, junto con la rápida aparición de nuevas técnicas e instrumentos para analizar dicha información, ha facilitado el análisis y la interpretación a gran escala de los datos sobre la biodiversidad y la distribución de las especies. Por lo tanto, estos datos proporcionan un recurso inestimable para documentar la biodiversidad y su distribución a través del tiempo y el espacio para la investigación, la educación y la formulación de políticas (Williams et al., 1996; Winker, 2004). Sin embargo, estas fuentes de datos incurren en sesgos potenciales relacionados con ambigüedades taxonómicas, cobertura territorial desigual, errores tipográficos y de georeferenciación o incertidumbre geográfica (Soberón y Peterson, 2004; Newbold, 2010) que ahora son reconocidos por la comunidad científica. Estas limitaciones han puesto en duda la utilidad de las bases de datos públicas, incluso si todos los datos disponibles pudieran recopilarse exhaustivamente (Hortal et al., 2008; Stropp et al., 2016).

Existen tres limitaciones principales para caracterizar la distribución de las especies, que van desde la información contenida en las bases de datos de biodiversidad: i) esfuerzo de estudio desconocido, ii) ausencias desconocidas, y iii) recurrencia desconocida. Estas limitaciones están interrelacionadas entre sí, por lo que sólo cuando se compilan exhaustivamente todos los sucesos conocidos es posible estimar el esfuerzo de muestreo con cierta fiabilidad, ayudando así a diferenciar la ausencia de pruebas de la evidencia de ausencia. Por lo tanto, una base de datos de biodiversidad que recopile exhaustivamente toda la información disponible sobre la identidad y distribución de un grupo de especies permitiría tanto identificar áreas bien encuestadas (por ejemplo, Hortal y Lobo, 2005) como obtener estimaciones de la ocurrencia repetida y/o la probabilidad de ausencia de especies particulares (por ejemplo, Guillera-Arroita et al., 2010). A pesar de la importancia ampliamente reconocida de evaluar la calidad y la integridad de los datos como paso preliminar en cualquier estudio de biodiversidad, este proceso a menudo se descuida. Podría decirse que esto se debe en parte a que este proceso de evaluación lleva mucho tiempo, requiere el uso de varias aplicaciones informáticas y la repetición del mismo proceso para cada una de las unidades territoriales o emplazamientos considerados (o, en general, para cualquier tipo de unidad espacial).

El sesgo espacial en los datos de distribución de especies es un fenómeno general con el potencial de distorsionar fuertemente nuestra visión sobre los patrones de biodiversidad a gran escala (Ballesteros-Mejia et al., 2013; Boakes et al., 2010; Yang et al., 2013). Una multitud de factores, tales como dónde se llevaron a cabo las encuestas y a qué escala espacial, qué datos o especímenes fueron recolectados, y cuáles de estos datos fueron almacenados y archivados.

Enfoque metodológico

Se han desarrollado muchos métodos diferentes para generar áreas de distribución a partir de los registros de observación, pero se ha prestado poca atención a comprender cómo las variaciones en la cantidad y calidad de los datos de línea de base y la implementación de diferentes metodologías afectan la precisión de los mapas de distribución de las especies (Graham e Hijmans, 2006; Maldonado et al., 2015). Hasta ahora, podemos diferenciar dos técnicas principales para construir áreas de distribución de las especies: algoritmos geográficos y mapas de áreas dibujadas por expertos.

Los algoritmos geográficos son métodos matemáticos que utilizan únicamente observaciones espacio-temporales (Burgman y Fox, 2003; Getz y Willmers, 2004; Getz et al., 2007) para definir un espacio topológico como la extensión de la ocurrencia o el área donde se supone que la especie está presente dadas las observaciones y la resolución espacial impuestas por las observaciones vecinas (Bronstein et al., 2007). Estos métodos sólo utilizan registros para definir el espacio geográfico que representa el área en la que se supone que una especie está presente (Burgman y Fox, 2003; Bronstein et al., 2007). Al requerir sólo registros de especies para la construcción de mapas de áreas de distribución, estos métodos conectados a las bases de datos de biodiversidad en línea nos permitirían mantener las áreas de distribución siempre actualizadas. Además, estos métodos son fácilmente repetibles siempre y cuando el procedimiento esté debidamente anotado. Los mapas de área de distribución dibujados por expertos derivan de un dibujo manual de un polígono simplificado en torno a registros conocidos, utilizando el conocimiento experto de las preferencias de hábitat de las especies y la información ambiental auxiliar, como los tipos de hábitats o los accidentes geográficos (Maréchaux et al., 2017; Herkt et al., 2017). Se trata de un método que presenta un alto nivel de abstracción, además de ser difícil de replicar por no haber sido realizado con

una metodología clara y repetible. En la actualidad, la Unión Internacional para la Conservación de la Naturaleza (UICN; http://www.iucnredlist.org/technical-documents/spatial-data) es el depositario más completo de esta gama de especies dibujadas por expertos. Los mapas de expertos de la UICN están disponibles para todas las especies de aves (Butchart et al., 2004), anfibios (Stuart et al., 2004), mamíferos (Schipper et al., 2008) y reptiles (Böhm et al., 2013), así como para varias especies de otros taxones en muchas regiones del mundo (UICN, 2017), y se están utilizando cada vez más para la investigación.

Los MDFs también se utilizan a menudo en la literatura científica para generar áreas de distribución de las especies (Boitani et al., 2011; Elith et al., 2006), un enfoque que es inapropiado siguiendo nuestra definición del concepto de área de distribución de las especies. Mientras que los algoritmos geográficos generan áreas de distribución fenomenológicas, basados sólo en los registros de especies, los MDF generan estimaciones de idoneidad ambiental, que pueden traducirse en áreas donde los requerimientos ambientales de las especies están supuestamente cubiertos. A pesar de ello, no tenemos ninguna garantía de que la especie esté realmente presente.

Conclusión

Dada la gran relevancia de los mapas de distribución de especies, es sorprendente observar que se ha prestado muy poca atención al análisis de cómo estos mapas se ven afectados por la calidad de los datos de línea de base y la diversidad de los métodos utilizados para construirlos. Este es el eje central de la tesis, que se estructura en cuatro capítulos principales.

En el **Capítulo I** realizamos una revisión bibliográfica para obtener información de publicaciones científicas que utilizan áreas de distribución de las especies en sus estudios. Observamos cómo se han generado e identificado las áreas de distribución que son los métodos más comúnmente utilizados para generar áreas de distribución a partir de datos georreferenciados, junto con las ventajas y desventajas proporcionadas por cada uno de ellos. En la mayoría de los casos, los investigadores no proporcionan información sobre cómo se han construido las áreas. La falta de información explícita sobre los datos y métodos utilizados en la construcción de las áreas de distribución afecta severamente a la

interpretación de los resultados. Por último, los métodos utilizados habitualmente para delimitar las zonas no se han evaluado suficientemente. Instamos a los investigadores a ser explícitos tanto en lo que consideran áreas de distribución de las especies como en los métodos que utilizan para generarlas. Esto permitirá realizar comparaciones más sólidas entre las áreas de distribución de las especies generados por diferentes métodos.

En el **Capítulo II** evaluamos la exactitud de cinco algoritmos geográficos comúnmente utilizados para delinear las áreas de distribución de las especies con el objetivo de proporcionar directrices para minimizar el error de Tipo I y maximizar la sensibilidad de las áreas de distribución de las especies resultantes. Con este objetivo, generamos áreas de distribución hipotéticas con la misma superficie total pero variando en forma, número de fragmentos, heterogeneidad en el tamaño de los fragmentos y conjuntos simulados de registros de especies variando en número, distribución espacial y presencia de errores y sesgos. Los algoritmos recomendados han sido Adaptive Local Convex Hull (a-LoCoH) y Kernel Density Estimation (KDE). El algoritmo KDE tiene la sensibilidad más alta y el algoritmo a-LoCoH tiene la tasa de error tipo I más baja. Ambos se comportaron similarmente bien al describir la fragmentación del área. Proporcionamos recomendaciones para minimizar los efectos de la cantidad y calidad de los datos, y proporcionamos orientación para elegir un algoritmo a la hora de definir las áreas de distribución de las especies en base a las observaciones de las especies.

El **Capítulo III** de esta tesis explora las opciones para una generación sistemática y replicable de mapas de áreas de distribución que tengan en cuenta las diferentes fuentes de variabilidad y el aumento exponencial en la disponibilidad de registros de especies. Ofrecemos una metodología unificada y repetible para construir mapas de áreas de distribución de especies, que comparamos con los mapas existentes de la Unión Internacional para la Conservación de la Naturaleza (UICN). La combinación de los mapas de distribución de la UICN con los datos de especies georreferenciados disponibles del Fondo Mundial para la Información sobre la Biodiversidad (GBIF) es una vía prometedora para proporcionar información sobre dónde son fiables los mapas de distribución de especies y dónde son inciertos. La falta de información o la disponibilidad de información en determinadas zonas dificultan la aplicación de enfoques sistemáticos para la elaboración de mapas de distribución.

Así que también revelamos sitios prioritarios por falta de información o esfuerzo de muestreo a escala global.

El **Capítulo IV** evalúa la variabilidad en la descripción de las áreas de distribución de las especies basándose en la recolección de datos no sistemáticos (por ejemplo, usando registros de bases de datos disponibles) o en encuestas sistemáticas y específicas. Como caso de estudio se utilizó el topillo de las agua (*Arvicola sapidus*) en la España peninsular, utilizando los resultados de una iniciativa de ciencia ciudadana centrada específicamente en esta especie y comparándolos con los de un atlas anterior. Los mapas de distribución resultantes presentaban diferencias notables, relacionadas con errores de identificación y esfuerzos heterogéneos de muestreo en el conjunto de datos no sistemáticos, así como con cambios reales en el área de distribución debido a la depredación por el visón americano invasor. La probabilidad de errores de comisión aumenta en áreas donde hay especies que pueden ser confundidas con el topillo de agua y por la depredación del visón. La probabilidad de errores por omisión aumenta en áreas con bajo esfuerzo de muestreo y la existencia de roedores fácilmente confundibles con la especie estudiada. Hacemos hincapié en la necesidad de ser cautelosos al utilizar las fuentes de información disponibles para generar mapas de área de distribución, en particular en zonas con pocos datos o signos de cobertura espacial heterogénea.

En conclusión, esta tesis explora las diferentes dimensiones de los mapas de distribución de especies y ofrece una perspectiva necesaria para abordar problemas planteados por ciencias como la ecología o la biología de la conservación. También tratamos de entender la naturaleza de la incertidumbre involucrada en los mapas de distribución para ayudar a interpretar los resultados existentes y guiar la investigación futura. Las métricas de información desarrolladas a lo largo de esta tesis podrían ser incorporadas en herramientas en línea que permitan a los investigadores y agencias de financiamiento identificar especies y áreas prioritarias para mejorar las fuentes de información junto con sus mapas de distribución asociados.

GENERAL INTRODUCTION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Introduction

Biodiversity is distributed heterogeneously across the Earth. Knowing the places in which the different species are present is a main objective of natural sciences, especially in disciplines such as biogeography, macroecology and conservation biology. An accurate knowledge of species distributions allows describing the geographical patterns of biodiversity, informing the management and conservation of natural resources, identifying priority areas for conservation or investigating evolutionary relationships through space (Margules et al., 2002; Rondinini et al., 2011). The distribution range of species (or other taxonomic level) is a conceptual construction that describes the area where it is present. The range can be characterized in terms of its size, shape and other descriptors of its limits, fragmentation or internal structure (Brown et al., 1996, Lucas et al., 2016). As a conceptual tool, the distribution range provides a summarized description of the complex spatio-temporal dynamics of populations. The characterization of distribution ranges depends on how they are defined, the quality and quantity of the available baseline data and the methodological approach chosen to build it; issues that are often overlooked in the scientific literature.

Defining species distribution ranges

As defined above, and as this concept will be treated throughout the PhD thesis, the distribution range is a conceptual construction that defines a topological space where the species or taxon is assumed to be present given the observations and the spatial and temporal resolution imposed. However, this concept is sometimes contextualized in the scientific literature in different ways, with potential for confusion when using the concept in a non-transparent way.

The IUCN, in the most influential assessment of the conservation status of species (IUCN, 1994, 2001), defines the extent of occurrence (EOO) as the area contained within the shortest continuous boundary that encompasses all sites of present occurrence of a taxon. EOO may include discontinuities or disjunctions within the overall distribution of taxa, such as large areas of obviously unsuitable habitat. The area of occupancy (AOO) is a subset of the EOO and describes the area where a species is actually present (Gaston, 1991; 2003). These two parameters are used in the IUCN protocols to asses conservation status from distribution

ranges (Gaston and Fuller, 2009; IUCN Standards and Petitions Subcommittee, 2010). Likewise, other distribution ranges definitions are also used in the current scientific literature to generate species range maps. Some are based exclusively on georeferenced records, and others use estimates of environmental suitability together with georeferenced records, which can be translated into areas where the environmental requirements of the species are supposedly covered. However, the distribution of a species is not only determined by the ecological niche, but also by dispersal barriers, biotic interactions and historical factors (Oswald et al., 2016, Husáková and Münzbergová, 2016, Schloss et al., 2012). Ecological niche models, most frequently known as species distribution models (SDMs), are methodological tools used to delineate the areas where the conditions for the existence of a species are met, based on known occurrence data and the environmental conditions in those locations. Therefore, SDMs by definition do not identify species distribution ranges. Nevertheless, this leap from distribution range to potential distribution range occurs frequently in the literature.

Glossary I: Relevant concepts used throughout this study

Species records are the geographic coordinate's data, often available in online biodiversity databases.

Distribution range is the area where a taxon is present, which can be characterized in terms of its size, shape and descriptors of the limits, fragmentation or internal structure.

Species range maps are visuals representations of distribution ranges.

Ecological niche is a multidimensional space in which each dimension (component of the niche) corresponds to a resource or requirement of a species. This fundamental (or potential) niche is limited by the interaction with other species, resulting in the real niche (observed).

Species distribution models (SDMs) are tools used for modelling species geographic distributions based on correlations between known occurrence records

and the environmental conditions at occurrence localities

Extent of occurrence (EOO) is the area contained within the shortest continuous boundary that encompasses all sites of present occurrence of a taxon.

Area of occupancy (AOO) is defined as the area within its 'extent of occurrence' which is occupied by a taxon, excluding cases of vagrancy

Biodiversity data

Under the explosive increase in global data, the term "big data" is used to describe huge data sets. These big data generates new opportunities to discover new values and also incur new challenges when trying to organize and manage these datasets effectively (Maldonado et al., 2015; Stephenson et al., 2017). In sciences such as ecology or conservation biology, citizen science databases are becoming an important way to collect information on species distributions (Dickinson et al., 2012; Tiago et al., 2017). Observations gathered by a large number of volunteers, over broad spatial extents and temporal periods often provide a large number of records (Chandler et al., 2012), allowing studies that would otherwise be unfeasible. The increment of species records from citizen science initiatives in recent years is particularly important for conspicuous and easy to identify taxonomic groups. The possibility of collecting, through mobile applications with internet connections, georeferenced observations of the natural world (e.g., wildlife sightings) via interactive geovisualization interfaces (e.g., Google Maps, Google Earth, and Microsoft Virtual Earth) or the use of sensors in mobile devices allow us to collect a large amount of data from the environment. In addition to the great opportunity offered by citizen science platforms, biodiversity databases also aggregate published information (books, monographs, papers or conference proceedings), collections of natural history, information collected in surveys, specific surveys or online repositories (Soberón and Peterson, 2004; Guralnick et al. 2007). Therefore, biodiversity databases provide a large amount of heterogeneous information and initiatives to generate, store and connect these databases have also proliferated in recent decades.

Ambitious international infrastructures such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) seek to link all these collections of biodiversity databases between countries and continents. GBIF is at the moment the largest and most

widely used biodiversity database (Beck et al., 2012, 2014; Jetz et al., 2012). The objective of GBIF is to‘make the world's primary data on biodiversity freely and universally available via the Internet’ (Yesson et al., 2007; GBIF, 2008). Currently, GBIF provides a single portal to access more than 975 million records (as for April 2018). This massive availability of biodiversity data, together with the rapid emergence of new techniques and tools to analyze such information, has facilitated large-scale analyses and interpretation of biodiversity and species distribution data. Such data thus provide an invaluable resource to document biodiversity and its distribution through time and space for research, education and policy making (Williams et al., 1996; Winker, 2004). However, these data sources incur potential biases related to taxonomic ambiguities, unequal territorial coverage, typographical and georeferencing errors or geographical uncertainty (Soberón and Peterson, 2004; Newbold, 2010) that are now recognized by the scientific community. These limitations have called into question the usefulness of public databases, even if all available data could be gathered exhaustively (Hortal et al., 2008; Stropp et al., 2016).

There are three main limitations for characterizing species distributions ranges from the information contained in biodiversity databases: i) unknown survey effort, ii) unknown absences, and iii) unknown recurrence. These limitations are mutually interrelated, so only when all known occurrences are comprehensively compiled it is possible to estimate sampling effort with some reliability, thereby helping to differentiate the absence of evidence from the evidence of absence. Therefore, a biodiversity database that compiles exhaustively all available information on the identity and distribution of a group of species would enable both identifying well-surveyed areas (e.g. Hortal and Lobo, 2005) and obtaining estimates of the repeated occurrence and/or the probability of absence of particular species (e.g. Guillera-Arroita et al., 2010). Despite the widely recognized importance of evaluating data quality and completeness as a preliminary step in any biodiversity study, this process is often neglected. Arguably, this is in part because such evaluation process is highly time-consuming, it requires the use of several software applications, and repeating the same process for each one of the territorial units or sites considered (or, in general, for any type of spatial unit).

Spatial bias in species distribution data is a general phenomenon with the potential of strongly distorting our view on large-scale biodiversity patterns (Ballesteros-Mejia et al., 2013; Boakes et al., 2010; Yang et al., 2013). A multitude of factors, such as where surveys

were carried out and at what spatial scale, what data or specimens were collected, and which of these data were stored and finally mobilized, can cause such biases. Data provided by GBIF are no exception to these problems. The distribution of museums and their funding, data digitalization policies of sharing data with GBIF may weigh particularly high as factors leading to spatial bias in the data made available. The identification of spatial biases in biodiversity databases is essential to interpret the results obtained in the generation of species distribution maps (Tiago et al., 2017). Only by taking into account these biases, such as the existence of under-sampled regions, can we support and improve the adoption of conservation measures by decision-makers (Tulloch et al., 2013).

In addition to spatial bias, another disadvantage of these databases lies in errors, mostly spatial and taxonomic errors. Errors in occurrence data are caused by a variety of factors, including mistakes in transfer of data from field sheets to electronic databases, rounding errors, failure to specify the geographical datum used to measure geographical location and retrospective georeferencing of imprecise locality descriptions (Graham et al., 2007; Varela et al., 2011). In this sense, a great effort is being made worldwide to reduce these errors and biases in these databases (Soberón and Peterson, 2004; Guralnick et al., 2007).

Methodological approach

Many different methods have been developed to generate distribution ranges from observation records but little attention has been paid to understand how variations in the quantity and quality of baseline data and the implementation of different methodologies affect the accuracy of species range maps (Graham and Hijmans, 2006; Maldonado et al., 2015). So far, we can differentiate two main techniques to build species distribution ranges: geographic algorithms and expert-drawn range maps.

Geographic algorithms are mathematical methods that use only spatio-temporal observations (Burgman and Fox, 2003; Getz and Willmers, 2004; Getz et al., 2007) to define a topological space as the extent of occurrence or the area where the species is assumed to be present given the observations and spatial resolution imposed by neighboring observations (Bronstein et al., 2007). These methods only use records to define the geographic space that

represents the area in which a species is assumed to be present (Burgman and Fox, 2003; Bronstein et al., 2007). By requiring only species records for the construction of range maps, these methods connected to the online biodiversity databases would allow us to keep the ranges always up to date. In addition, these methods are easily repeatable as long as the procedure is properly annotated. Expert-drawn range maps derive from a manual drawing of a simplified polygon around known records using expert knowledge of habitat preferences of species and auxiliary environmental information, such as the types of habitats or geographical features (Maréchaux et al., 2017; Herkt et al., 2017). It is a method that presents a high level of abstraction as well as being difficult to replicate because they were not made with a clear or repeatable methodology. Actually, the most comprehensive repository of such expert-drawn range is provided by the International Union for the Conservation of Nature (IUCN; http://www.iucnredlist.org/technical-documents/spatial-data). IUCN expert maps are available for all species of birds (Butchart et al., 2004), amphibians (Stuart et al., 2004), mammals (Schipper et al., 2008) and reptiles (Böhm et al., 2013) as well as for several species in other taxa in many regions across the world (IUCN, 2017), and they are being used increasingly for research.

SDMs are also often used in the scientific literature to generate species distribution ranges (Boitani et al., 201;, Elith et al., 2006), an approach that is inappropriate following our definition of the concept of species distribution range. While geographic algorithms generate phenomenological distribution ranges, based only on species records, SDMs generate estimates of environmental suitability, which can be translated into areas where the environmental requirements of the species are supposedly covered. In spite of that, we have no guaranty that the species is actually present.

Objectives

The main objective of this PhD thesis is the evaluation of methodologies for obtaining species distribution maps from the different approaches that characterize the distribution ranges in relation to their definition, the existing sources of biodiversity information and the different methods used to build distribution ranges. Specifically, the following questions are addressed:

- ✓ To know which are the most used methods in the description of species distribution ranges and if the authors adequately describe the implementation of these methods. To this end, a bibliographic review is made of the current approaches that describe the ranges of species based on georeferenced data, the methodologies used in the process of constructing species distribution areas and the advantages and disadvantages of each method (Chapter 1).

- ✓ To evaluate the accuracy of commonly used geographic algorithms in reproducing reference areas from geographic records that vary in the quantity and quality of biodiversity databases (number, spatial bias and errors) to provide guidelines on how to delineate distribution ranges while minimizing the Type I error rate and maximizing the sensitivity. To do this, we construct reference ranges with limitations in terms of: the shape, number and size of the range fragments and baseline data (amount of information, spatial distribution and errors) (Chapter 2).

- ✓ Develop unified techniques, systematic and replicable to generate species range maps using geographic algorithms and online biodiversity databases. To do this, we build ranges of species using geographic algorithms and online georeferenced records. We compared the ranges generated by species with those provided by IUCN and identified concordant and discordant areas. Finally, we generated a spatially explicit estimate of the sampling effort, with the aim of discerning between commission and omission errors in discordant areas to identify areas around the world that require a record-collection effort to enable proper functioning of systematic approaches to the generation of species distribution maps (Chapter 3).

- ✓ To compare the distribution ranges generated from non-systematic and systematic data collection strategies, using as a case study the southern water vole in peninsular Spain. We used two different sources of information on the species distribution, and built distribution range maps using geographic algorithms. Finally, we investigated factors that might be associated with omissions and commission errors to reduce the risk of omitting unmapped areas from the range maps. (Chapter 4).

CHAPTER I - Defining species distribution ranges: current approaches, methodologies and limitations

ABSTRACT. Currently there is not a consensus when it comes to defining the distribution range of a species. This is because distribution ranges are generally constructed using heterogeneous data and without providing information on how they have been generated. This lack of explicit information often means that the interpretation of the results obtained using them is challenging. Here, we conducted a literature review in order to identify publications that used species distribution ranges in their studies. For each paper, we looked for all the information related to how the distribution ranges have been generated, whether these papers had explicit information on how the distribution ranges have been constructed and, which are the most common methods used to generate distribution areas from georeferenced data. The results obtained indicate that (1) species distribution ranges are rarely defined in papers, (2), those that do offer little or no information on how species distribution ranges have been generated, (3) there is a long list of different methods used to generate species distribution ranges, and (4) the methods employed varied considerably even when dealing with the same information source. Additionally, we describe the methods most commonly used and their advantages and disadvantages and provide recommendations to help in selecting the best method that allows mapping distribution ranges. We urge researchers to be explicit in both what they consider species distribution ranges and the methods they use to generate them. Our recommendations will increase the reproducibility of studies and allow for more solid comparisons between species distribution ranges generated with different methods.

Key words: bias and error, biodiversity databases, GBIF, geographic algorithms, polygons, range maps.

Ríos-Pena, L., Clavero, M., & Revilla, E. Defining species distribution ranges: current approaches, methodologies and limitations (*In prep).*

RESUMEN

Actualmente no existe consenso a la hora de definir el área de distribución de una especie. Esto se debe a que las áreas de distribución se construyen generalmente utilizando datos heterogéneos y sin proporcionar información sobre cómo se han generado. Esta falta de información explícita a menudo significa que la interpretación de los resultados obtenidos al utilizarlos es desafiante. Aquí, realizamos una revisión de la literatura con el fin de identificar las publicaciones que utilizaron áreas de distribución de las especies en sus estudios. Para cada trabajo, se buscó toda la información relacionada con cómo se han generado las áreas de distribución, si estos documentos tenían información explícita sobre cómo se han construido las áreas de distribución y cuáles son los métodos más comunes utilizados para generar áreas de distribución a partir de datos georreferenciados. Los resultados obtenidos indican que (1) las áreas de distribución de las especies rara vez se definen en los documentos, (2) los que ofrecen poca o ninguna información sobre cómo se han generado las áreas de distribución de las especies, (3) existe una larga lista de métodos diferentes utilizados para generar áreas de distribución de las especies, y (4) los métodos empleados variaron considerablemente incluso cuando se trataba de la misma fuente de información. Además, describimos los métodos más comúnmente utilizados y sus ventajas y desventajas, y ofrecemos recomendaciones para ayudar a seleccionar el mejor método que permite mapear las áreas de distribución. Instamos a los investigadores a ser explícitos tanto en lo que consideran las áreas de distribución de las especies como en los métodos que utilizan para generarlas. Nuestras recomendaciones aumentarán la reproducibilidad de los estudios y permitirán realizar comparaciones más sólidas entre áreas de distribución generados con diferentes métodos.

Palabras clave: sesgo y error, bases de datos de biodiversidad, algoritmos geográficos, GBIF, polígonos, mapas de áreas de distribución.

INTRODUCTION

Species distribution ranges can be defined as the areas that enclose all the localities where a species (or whichever the taxon) has been recorded. This concept is key in biogeography, macroecology, conservation biology and large-scale community ecology (Brown et al., 1996). It is used to describe spatial patterns of biodiversity and identify the processes shaping these patterns, to inform the management and conservation of natural resources, to identify priority areas for conservation or to investigate evolutionary relationships across space (Rondinini et al., 2011; Meyer et al., 2016).

The study of distribution ranges emerged in the 18th and 19th centuries, when naturalists such as Candole, Wallace, Hooker or Darwin documented the patterns in the distribution of the variety of plants and animals around the world and speculated on the drivers generating them (Egerton 2012). The first works that dealt explicitly with the characteristics of distribution ranges came from the hand of Willis (1922), who quantified the areas of distribution ranges in several taxonomic groups, and Arrhenius (1921), who worked on species/area relationships. For most of the twentieth century, research on distribution ranges was directed primarily towards identifying the ecological factors determining the boundaries of species ranges (Billings, 1952; Andrewartha and Birch, 1954; MacArthur, 1972). In 1977, Sydney Anderson published the first of several papers focused on measuring the areas of the mapped ranges of vertebrates in North America and Australia (Anderson, 1977). However, it was not until the publication of Rapoport's monograph Aerography (Rapoport, 1982) when the interest of the scientific community on studying species distribution ranges began to rise (Anderson and Marcus, 1992). Rapoport provided evidence of a decrease in species range size from high to low latitudes, which was to be known as the Rapoport's rule, using mammal subspecies data from North America. Since the last decade of the 20th century the study of distribution ranges has grown noticeably, boosted by the development of extensive online databases that compile occurrences of species (García-Roselló et al., 2015).

In the 21st century, under the explosive rise of global data, a new concept of "big data" is emerging, which is mainly used to describe large datasets and the methods associated with them. These large datasets have in turn led to a new conceptual development, the so-called Ecological Niche Models (ENMs). Those models use occurrence records in order to establish a model of the suitability of the local environmental conditions for the appearance of the

target species. These estimates translate into areas where the species' environmental requirements are assumed to be met, but which, accordingly, do not define the area of actual presence. Consequently, ENMs do not identify species distribution ranges, although that leap from “possible distribution range” to “distribution range” is often made, especially when there is no alternative.

The choice of methodology is often critical in determining the characteristics of the distribution ranges generated (Muñoz and Felicisimo 2004; Fotin et al., 2005; Tsoar et al. 2007; Mota-Vargas and Rojas-Soto, 2011). Different methods have been developed to generate distribution ranges (Burgman and Fox, 2003; Getz and Willmers, 2004; Graham and Hijmans, 2006; Getz et al., 2007). These can be grouped into two broad categories: 1) geographical methods that define the area of presence based exclusively on the geographic coordinates of the occurrence data (Burgman and Fox, 2003; Bronshtein et al 2007; Jaryan et al., 2013; Sharifi et al., 2012; Getz et al., 2007; Asaeedi et al., 2013; Kondoh et al., 2013); and 2) expert knowledge approaches that use the original records together with personal knowledge and/or intuitions as source of information to establish the boundaries, shape and size of a species' distribution (Gaston, 1996; Brown and Lomolino, 1998; Orme et al., 2005). Expert knowledge approaches essentially involve the implementation of an informal distribution niche modelling but that is non-repeatable (Graham and Hijmans, 2006).

The key role of species distribution ranges in many ecology-related disciplines and the fact that they can be generated with methods that may produce disparate results for the same source of information, call for the need of unifying criteria on how to generate species ranges. Here, we review current practice in describing species ranges from georeferenced data. We developed a literature review in order to know how frequently authors explicitly report how distribution ranges are generated, and identify the most commonly used methodologies. We then describe the advantages and disadvantages of each of the methodologies identified in the literature review. Our crosscutting aim is to highlight how different methodologies to generate species distribution ranges may produce different outcomes, thus emphasizing the need to standardize approaches.

USES AND OPERATIONAL DEFINITIONS

The general definition of the area enclosing all the localities where a species is present is translated in the literature into different operational definitions. Krebs (2001) defined the distribution range in terms of the variability of abundance, stating that the abundance of an organism within its range must always be greater than zero and the boundary of a distribution range is equal to the contour line where the abundance is equal to zero. Espinosa and Llorente (1993) made a distinction between ecological and geographical distributions, and define the former as the behaviour of a population parameter along an environmental gradient, be it a gradient of conditions (temperature, pH, salinity, etc.) or of resources (availability of food, shelter, breeding sites, etc.). Zunino and Zullini (2003) defined the species distribution range as the fraction of the geographical area where that species is present and can interact in a non-ephemeral manner with the ecosystem, while Soberón (2007) defined the distribution areas in terms of the actual or potential spatial locations that individuals comprising a species can occupy and one particular type of niche in terms of the parameters of population equations, thus mixing the concept with that of ecological niche models. Therefore, there is a conceptual discussion about the distribution range in the scientific literature that leads to large differences in the way distribution ranges are generated (Kreft et al 2006, McPherson and Jetz, 2007). Species distribution ranges are highly dynamic and can expand and contract over time, although it is rarely considered (Davis and Shaw, 2001; Gaston, 2003). Acknowledging such dynamism has important implications for the understanding of biodiversity patterns and the conservation of biological diversity (Lamoureux et al., 2006; Myers et al., 2000) as, for example, for assessing the conservation status of species (e.g. IUCN, 1994).

We reviewed scientific publications working with species distribution ranges and analysed whether they were explicit in defining ranges and describing how those ranges had been generated. We searched for papers focused on species distribution ranges in the Web of Science (WOS), using different keywords (Table 1), as well as filters by language (English and Spanish), research domains (science technology) and research areas (Zoology, Environmental Sciences, Ecology and Biodiversity Conservation). We obtained 2034 articles of which only 127 articles contained explicit information about what is and how species distribution ranges have been generated. Out of the 127 selected papers, 100 worked directly with species distribution ranges and 27 with species distribution models. Of the 100 papers

containing information of how ranges were constructed, only 17 offered sufficient information to replicate accurately the distribution ranges, i.e., they gave information about the chosen methods and specified the parameters selected in the methods for the construction of the distribution range. Therefore, 83% of the documents focusing on the construction of species ranges did not provide information on how the areas on which results were subsequently obtained had been delimited. In addition, 21.3% of papers that define the concept of species distribution apply species distribution models to generate their own maps of areas of presence, thus using potential areas to make inference.

Table 1: Keywords used in the literature review to search for papers that use species ranges in their analyses. Total refers to the total number of papers obtained in each keyword search. These papers have been reviewed and the total number of articles working with species distribution ranges has been selected (Information). The total number of papers valid as source of information has been designated as SDR (Species Distribution Range). TOTAL is the total sum of Information and SDR papers that show unique values, without repetition of papers.

Keywords	Total	Information	SDR
"species distribution range"	186	10	8
"distribution range" AND area AND species	1145	35	20
"distribution area" AND map* AND species*	135	3	0
"distribution area" AND map* AND species NOT model*	103	23	23
"geographic distribution" AND species* AND map* NOT model* AND area* AND range*	44	10	8
"distribution" AND species* AND range* AND area* size* AND map* AND geographic* NOT model* NOT predict*	62	13	11
"distribution area" AND species* AND geographic* NOT	108	10	9

model* NOT predict* NOT home ranges*			
"geographic distribution*" AND "geographic range*" OR "geographic boundary*" AND specie* NOT "species distribution model*"	251	24	21
TOTAL		**127**	**100**

METHODS AND THEIR ADVANTAGES AND DISADVANTAGES

The development of systematic methodologies for representing species distributions maps began in the decade of the 1950s and involved identifying organism's distributions on maps and connecting the disjoint distribution ranges by lines calls strokes (Croizat L. 1958, 1964). Since then, a large number of methodologies have been developed to solve the problem of how to draw species distribution ranges on maps. Distribution ranges are normally constructed using georeferenced records (Hirsch and Chiarello, 2012; Desender et al., 2010; Laplana et al., 2013). However, different methods can provide substantially different results (Fig 1; Appendix 1.S2, Table 1.S2). We conducted a second literature review to identify the most commonly used methodologies for generating species distribution ranges from georeferenced data. This new database is composed only of papers that use geographic methods to generate distribution ranges. A total of 100 publications form this database (Appendix 1.S1, Table 1.S1). The information extracted from each paper was mainly focused on the methodology used to generate species ranges from data records and the detailed description of each method.

The cartographic method turned out to be the most commonly used, appearing in more than 50% of the publications. It was followed by the minimum convex polygon (MCP, with 20%), the expert delineation (15%), kernel density estimation (10%), and the indicator kriging, hull (concave and convex) and local convex hull (k-LoCoH and r-LoCoH) methods (each with 5%). We run a third search for each selected method in order to find at least 20 articles were it is used (Table 2). In addition, for each described method, we present its advantages and disadvantages in order to know which method is most suitable for the sources of information obtained and the objectives defined:

(1) Cartographic method

This approach consists in superposing a grid to a map containing recorded localities of a given taxon; the distribution range encompasses all the grid cells containing at least one record. This is the typical approach used in atlases. The results of this simple procedure are extremely sensitive to the scale (grid cell size) used in the calculation. Consequently, when a fine scale is used the resulting distribution range will be small and unrecorded occurrences derived from a heterogeneous sampling effort will be overlooked. In contrast, using coarser resolutions may result in mapping large unoccupied areas, resulting in range overestimations. Therefore, the choice of a scale is not a simple matter, and could be a source of inconsistencies and biases (IUCN 2001). A reasonable solution to the problem of assigning a suitable scale was provided by Willis et al. (2003), who suggested that grid cell size could be defined as 10% of the distance between the most distant pair of points. This criterion allows calculating a specific scale to each particular species depending on its range configuration. This method is currently being used in Red List assessments to calculate the area of occupancy (AOO) defined as the area within its extent of occurrence (EOO, area contained within the shortest continuous imaginary boundary that can be drawn to encompass all the known, inferred or projected sites of present occurrence of a taxon), which is occupied by a taxon. Nevertheless, the challenge remains when dealing with species with very large ranges and with a biased spatial sampling effort.

(2) Expert-drawn range maps

This method manually draws a simplified polygon around known occurrence locations using expert knowledge of a species' habitat preference and auxiliary environmental information, such as the presence of specific land uses or geographic barriers (Hawkins et al., 2008). This method is not repeatable and, in addition, given the rather high level of abstraction involved, deduced range boundaries typically ignore most of the internal structure as well as spatial outliers (Brown et al., 1996). The most comprehensive repository of such expert-drawn range maps is provided by the International Union for the Conservation of Nature (IUCN) and its partner, BirdLife International (often built by experts modifying the outcome of other methods such as MCP, see below). At present, IUCN expert maps are available for birds (Butchart et al., 2004), amphibians (Stuart et al., 2004), mammals (Schipper et al., 2008) and

reptiles (Böhm et al., 2013), and they are being used increasingly for macroecological research, often because they are the only source of occurrence information readily available. IUCN expert maps were created with a specific purpose in mind, namely to guide conservation efforts. IUCN maps tend to be conservative, underestimating the geographic range (sensu extent of occurrence; see Gaston and Fuller, 2009) of many species, especially in poorly surveyed regions such as the species-rich tropics—even in case of well-studied taxa (Ficetola et al., 2014; Pineda and Lobo, 2012).

(3) Minimum Convex Polygon

The minimum convex polygon (MCP; Mohr, 1947) (also called a convex hull) is the smallest polygon in which no internal angle exceeds 180 degrees and which contains all the presence records. This method is simple and easy to compute. Its main problem is that it tends to overestimate ranges because it includes large areas in which the focal species is not (or may not be) present, at least as long as the point clouds move away from the rounded or elliptical shapes (Mota-Vargas and Rojas-Soto, 2012; Burgman and Fox, 2003). The MCP approach is sensitive to outliers and to sample size, precluding comparisons of polygons generated with different sample sizes. Despite these caveats, it is the most used method in the assessment of the conservation status of species (IUCN, 2014; 2017; Burgman and Fox, 2003).

(4) Alpha convex and concave hulls

Alpha convex and concave hull are defined as a generalization of convex hull (Edelsbrunner et al., 1983; Burgman and Fox, 2003). These methods differ in the estimation of the internal angles. While the angles can be convex or concave, in the concave hull method, the angles are exclusively convex for the convex hull method. The alpha hull methods have been shown to be more efficient when species ranges have a concave shape, while the convex hull method tends to overestimate them. However, both methods are similarly good when the shape of the range is convex (Asaeedi et al, 2013).

(5) Kernel Density Estimation (KDE)

Kernel density estimation (KDE; Silverman, 1986) is frequently used to estimate distribution ranges (Burgman and Fox, 2003; Fortin et al., 2005). This methodology requires to select a bandwidth *h* (Seaman and Powell 1996), a free parameter that has a strong influence on the resulting range estimate. The bandwidth determines the relationship between the distance of a location from a point and the contribution of the location to the density estimate at that point. There are two types of kernel density estimations, fixed and adaptive. In the former the bandwidth is a fixed value over the plane, whereas in the latter there is a smoothing parameter that varies over the plane so that areas with a low concentration of records have higher h values than areas with a high concentration of points. The density estimation will be high in areas with many observations, and low in areas with few data. However, the choice of bandwidth will dramatically change the KDE, as a bandwidth that is too high or low will result in over- or under smoothing, respectively (Willians et al., 2014; Quintero et al., 2015). Otherwise, a poorly selected bandwidth is likely to produce an unrealistic structure in the density estimate (Spencer and Ghaznavi, 2017). Simulations have shown that estimates using core density work well because they faithfully reproduce the areas taken as "true" (Getz et al., 2007; Fleming et al., 2017; Cross et al., 2016).

(6) Local convex hull (LoCoH)

Local Convex hull (LoCoH) is both a generalization of the minimum convex polygon (MCP) method and a non-parametric kernel method (Getz et al., 2007). The distribution ranges are constructed by associating a local distribution function with each species record and then adding and normalizing these local distribution functions to obtain a function of distribution that belongs to the data as a whole (Getz and Wilmers, 2004). If the local distribution function is a parametric distribution, such as a symmetric bivariate normal distribution then the method is referred to as a kernel method (a parametric kernel method). On the other hand, if the local kernel element associated with each point is a local convex polygon constructed from the point and its k-1 nearest neighbours, then the method is nonparametric and referred to as a k-LoCoH (or fixed point LoCoH. There are two modifications of the k-LoCoH method. The first modification is a “fixed radius” r, or r-LoCoH, in which all the points in a fixed “sphere

of influence" of radius r are used to construct the local hulls. The second modification is the adaptive method, or *a*-LoCoH, in which all points within a variable sphere are used to construct the local hulls such that the sum of the distances between nearby points is less than or equal to *a*. The LoCoH methods require selecting the values of *k, r* and *a* parameters that have a strong influence over the resulting range estimate. Getz et al., 2007 provides a guide for selecting the values of these parameters. LoCoH methods are advantageous for precisely mapping the species distributions for which the absence of records indicates true gaps in occurrence (Chirima and Owen-Smith, 2017; Getz et al., 2007; Doherty and Witt, 2017).

(7) Indicator kriging

Indicator kriging is a non-linear geostatistical technique that interpolates site-specific point data over surfaces. It was introduced by Journel (1983) and it is mainly used to predict species occurrence probabilities, later transforming the probabilities into area. This method has been pointed out as a suitable approach for both frequent and rare species with highly biased records (Stelzenmüller et al., 2004). Indicator kriging estimates the probability of exceeding specific threshold values, z_k, at a given location. In indicator kriging, the data, $z(x)$, are transformed into an indicator variable as follows

$$i(x, z_k) = \begin{cases} 1, & \text{if } z(x) \leq z_k \\ 0, & \text{otherwise} \end{cases}$$

At an unsampled location, x_0, the probability that $z(x) \leq z_k$ can be estimated using a linear combination of neighbouring indicator variables. This ordinary indicator kriging estimator is,

$$Prob\left[z(x_0) \leq \frac{z_k}{n}\right] = \sum_{\alpha=1}^{n} \lambda_\alpha i(\mathrm{x}_\alpha; z_k)$$

where $i(\mathrm{x}_\alpha; z_k)$ represents indicator values at x_α, α=1,…,n, and λ_α, determined by solving the following kriging system, is the kriging weight of $i(\mathrm{x}_\alpha; z_k)$used in estimating $Prob\left[z(x_0) \leq \frac{z_k}{n}\right]$.

An ordinary indicator kriging system can be solved using,

$$\sum_{\beta=1}^{n} \lambda_\beta \gamma_i(x_\alpha - x_\beta; z_k) + \mu = \gamma_i(x_\alpha - x_0; z_k)$$

$$and \sum_{\beta=1}^{n} \lambda_\beta = 1$$

where μ is the Largrange multiplier; $\gamma_i(x_\alpha - x_\beta; z_k)$ is the indicator variogram between indicator variables at the αth and βth sampling points; $\gamma_i(x_\alpha - x_0; z_k)$ is the variogram between indicator variables the ath sampling point and x_0, and α= 1,..., n. This technique provides reliable interpolation results when there are gaps in sampling effort, reducing time and money to achieve this collection of data in the field. Its main disadvantage lies in the computational complexity of the method that tries to determine what should be the sampling density.

Table 2: Description of the most common methods used to generate species distribution ranges.

Method	Definition	Expression	Parameter values	Uses	References for description of methods
Cartographic	It uses a regular grid to summarize the position of species records.	Geometric unit = side * side	Grid size	Atlas (Gasc et al., 1997); suggested by IUCN for measuring the species area of occupancy (IUCN, 2010).	Hernández and Navarro, 2007; Mota-Vargas and Rojas-Soto, 2012
Minimum Convex Polygon (MCP)	Convex polygon the lower perimeter that contains the set of points in the plane and no exceeds 180 degrees.	$\hat{A} = \left(x_1(y_n - y_2) + \sum_{i=2}^{n-1} x_i(y_{i-1} - y_{i+1}) + x_n(y_{n-1} - y_1) \right) /2$	where$(x_i, y_i), i = 1,2, \dots n$ are the coordinates of the locations	Home range of any animal; to build species distribution maps; to generate extent of occurrence by IUCN.	Morh, 1947.

Kernel Density Estimation (KDE)	It is a probability density function for a sample. If placed a kernel in each of the referenced data, the weighted sum of these functions is also a probability density function data around the data used.	$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$	$(x_1, x_2, \ldots, x_n)$ is an independent and identically distributed sample with an unknown density f, k(·) is the kernel and h > 0 is a smoothing parameter called the bandwidth.	To estimate the home ranges of animals from radio-tracking data and to constructed distribution range maps.	Worton, 1989; Diggle et al., 2005.
Indicator Kriging	Geostatistics method of estimation points. It is based on data transformed from continuous values to binary values obtaining a new set of binary data for each quantile.	$I = (Z \leq q) := \begin{cases} 1, if\ Z \leq q \\ 0, if\ Z > q \end{cases}$	Quantile $Z_q \in \{q_{0.25}, q_{0.50}, q_{0.75}\}$	To predict probabilities of species occurrence; and then transform those probabilities into distribution ranges.	Kondoh et al 2013; Stelzenmüller et al 2010.
Alpha-Convex Hull	Family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points dependent on the value of the parameter α.	$\left\{\sum_{i=1}^{\lvert S\rvert} \alpha_i x_i \lvert (\forall_i: \alpha_i \geq 0) \wedge \sum_{i=1}^{\lvert S\rvert} \alpha_i = 1\right\}$	X = set of geographical coordinates, α can take values between zero and infinite.	To estimate species range maps; to evaluate species distribution patterns and path planning.	Edelsbrunner et al., 1983.

Alpha-Concave Hull	For a set P of points is the enclosing α – polygon (all interior angles are less than or equal to 180 + α degrees) with smallest area that contains P.	-	$0 \geq \alpha \leq 180$	The most common applications are in computational geometry, shape approximation, roof design and geometry modeling.	Moreira and Santos, 2007; Meyer et al., 2017.
Local Convex Hull (LoCoH)	***k*-LoCoH**	Set of convex hulls where each convex hull is built from a k-point and its nearest neighbours k.	$k = \sqrt{n}$, n is number of points in the total set.	To estimate the range size of a specie or taxon; to construct a probability distribution that represents the probability of finding a species within its range.	Getz and Wilmers, 2004; Getz et al., 2007
	***r*-LoCoH**	Set of convex hulls where each convex hull is built from a record within a "sphere of influence" of radius r around each record.	r is half of the maximum nearest neighbour distance between points (i.e. the radius of a sphere that will allow all points to be joined)		
	***a*-LoCoH**	all points within a variable sphere around a root point are used to construct the local hulls such that the sum of the	a is maximum distance between any two points in the data set.		

		distances between nearby points and the root point is less than or equal to a**.**			
Expert	The records are placed on a map and a polygon is drawn free hand. The lines are not drawn from one point to another, but instead, they pass either close to or distant from, the locality records at the discretion of the author.	-	They take into account specific criteria, such as vegetation or habitat, omission or exclusion of areas, knowledge of the species studied or environmental suitability	To draw species distribution range maps.	Graham et al., 2006; Hurlbert et al., 2007; Jetz et al., 2012

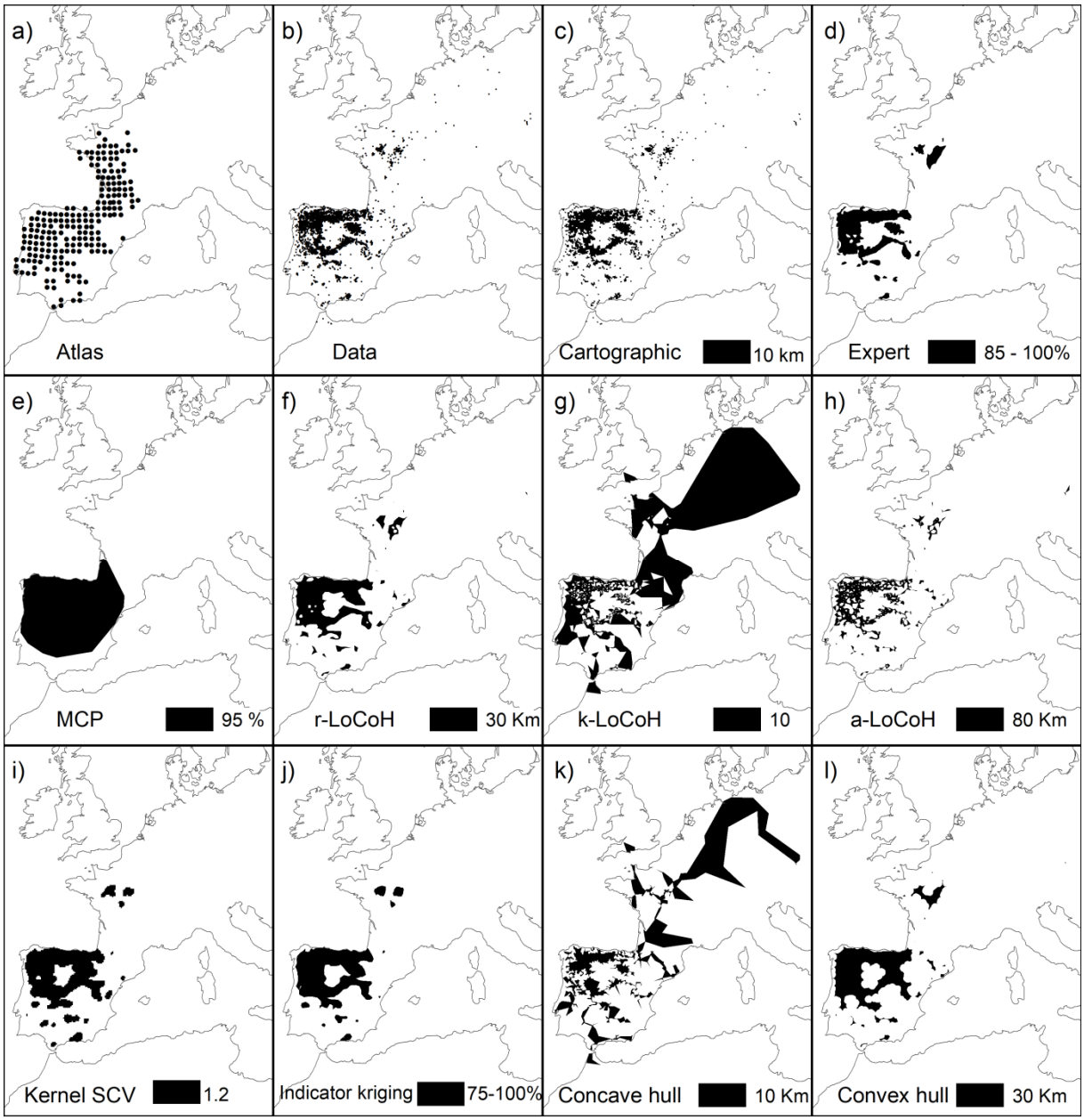


Figure 1. Example of distribution ranges generated using different algorithms for the Pyrenean oak (*Quercus pyrenaica* Willd.) using different geographical methods from GBIF records. (a) Distribution range from the Atlas Florae Europaeae (AFE, Jalas and Suominen 1988). (b) Filtered GBIF records used to build the range maps. (c) Presence grids with a size of 10 Km. (d) Representation of 85-100% overlapping area by expert method. (e) Range map not including 5% of the records furthest from the total records density. (f, g h) Range maps with the LoCoH methods where the legend value is the input parameter selected in each

method. (i) KDE range map where legend value is the bandwidth selected. (j) Range map with probability > 75%. (k, l) Range map of concave and convex hull methods where legend value is the input parameter selected in each method.

GENERAL CONSIDERATIONS AND RECOMMENDATIONS

The distribution of life on Earth shows complex spatial and temporal patterns (Brown et al., 1996; Graham and Hijmans, 2006). Any characterization of species distribution ranges is necessarily a simplification of such complex patterns. Therefore, we need to be aware that in that simplification we make many important decisions that need to be explicit in order to properly communicate our work. First, it is necessary to define beforehand and depending on the research question, the explicit definition of distribution range to be used, the data available and its quality, and the method that best fits the purpose of the research. The literature review shows that many methods are currently used and that they can generate different distribution range maps from the same spatial data. Given the wide choice available, we consider that it is convenient to establish some criteria to standardize the delineation of distribution ranges. A potential approach could be to overlap all the individual distribution maps obtained with each geographic method to obtain an ensemble range. In this way, we can establish comparative measures based on the percentage of concordance and discordance between methods. As a general recommendation, we should always provide standardized distribution maps.

Given a definition of the distribution range, and before selecting the geographic method to generate the distribution range, we must decide if we are going to work with all the observations recorded for the species or not. Applying filters to debug the available records is a fundamental process that should not be overlooked. This is because debugging data according to some previously established criteria and described in a clear and transparent manner would help to reduce the possible spatial biases and errors of the data while still allowing for replicated analyses by other researchers. Nevertheless, after filtering for errors the database can still be spatially biased. Therefore, we must consider the fact of selecting a method with the potential to generate a single or multiple polygons. If we work with spatially biased data and select methods that do not fragment the areas, the resulting distribution range may tend to overestimate the total area, including areas where the species is not present

(commission error, Hurlbert and Jetz, 2007). In short, commission errors could be more problematic than omission errors if they lead to the false belief that species are present when they are not (Rodrigues et al., 2004). If we work with all the records available, with data that contain biases and we select methods that tend to fragment, the distribution range tends to be reduced and consequently, it can give rise to omission errors (when a species is considered absent in a place where it occurs, Burgman and Fox, 2003), underestimating the distribution range (Cantú-Salazar and Gaston, 2013). Overall, the possibility of producing commission and omission errors reinforce calls for caution in the using uncritically range maps as sources of data on the presence or absence of species (Hurlbert and Jetz, 2007). Omission errors can affect the efficiency of conservation planning by biasing results towards known species occurrences, potentially missing important areas where a species also exists.

CONCLUSIONS

The accumulation of large quantitative databases, the development of computer software for statistical and spatially explicit analyses, and advances in mathematical and computer simulation modelling are helping in providing a more synthetic view of the distribution range concept. The discovery of quantitative patterns in the characteristics of ranges has led inevitably to the search for the causal processes and the development and testing of hypotheses on the mechanisms. Although there is much to learn about the patterns and the processes that generate them, it is necessary to establish explicit criteria when using these methodologies that allow us to compare ranges to, for example, identify priority areas for conservation or to investigate evolutionary relationships through space (Margules et al., 2002; Rondinini et al., 2011).

Finally, the choice of a methodology should be guided by the amount and quality of data available to delineate ranges, the quality of the data available, and the questions we want to answer. In addition, a number of technical challenges related to spatial and temporal data resolution and the management of uncertainty and biases associated with input data should be highlighted. At this point we can state that, depending on the quantity and quality of the data and their spatial distribution, it is necessary to make a rigorous selection of the geographical method to build the distribution range, including its parameterization. We hope that following

these guidelines will help in obtaining a more accurate representation of the current distribution of species on Earth.

ACKNOWLEDGEMENTS

We are very grateful to all those people from all over the world who have helped us to understand the methodological development of geographic algorithms. We also thank the expert staff at Doñana Biological Station for helping us build our own expert range. Finally, we are thankful to GBIF for making and maintaining their databases freely available online. This work was supported by the following grants and projects: Spanish Ministry of Economy, Industry and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R+D+I (SEV-2012-0262) and Agencia Estatal de Investigación from Ministry of Economy, Industry and Competitiveness, Spain with projects CGL2012-35931 and CGL2017-83045-R AEI/FEDER EU, co-financed with FEDER to E.R., R.B.M., M.G.S.

Supporting Information

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix 1.S1. Bibliographic review

Table 1.S1. Papers selected after the 1st bibliographic review which contained explicit information about what is and how species distribution ranges have been generated. We add two fields to the table, taxon and methodology, both obtained from each revised paper.

Methodology	TAXON	References
Cartographic	invertebrate	Desender, K., et al. 2010. Changes in the distribution of carabid beetles in Belgium revisited: have we halted the diversity loss? Biological Conservation, 143, 1549–1557.
	mammal	Ramos P.L., et al. 2009. Distribución actual del meloncillo, *Herpestes ichneumon* (Linnaeus, 1758), en el sur de la provincia de Salamanca y en el norte de la provincia de Cáceres. Galemys, 21 (NE): 133-142.
	mammal	Barros, P., et al. 2016. Confirmation of European snow vole Chionomys nivalis (Mammalia: Rodentia: Cricetidae) occurrence in Portugal, Italian. Journal of Zoology, 83(1): 139-145.
	plant	Idžojtić, M., et al. 2006. Hosts and distribution of Viscum L. ssp. album in Croatia and Slovenia. Plant Biosyst 140:50–55.
	mammal	van Moll, G. 2005. Distribution of the badger (Meles meles L.) in the Netherlands; changes between 1995 and 2001. Lutra 48 (1): 3-34.
	mammal	Llaneza, L., et al. 2004. Distribución y aspectos poblacionales del Lobo Ibérico en la provincia de Ourense. [Distribution and populational aspects of The Iberian Wolf in the Ourense province]. Ecología 18: 227-238.
	invertebrate	Monroy, F., et al. 2003. Distribution of earthworms in the north-west of the Iberian Peninsula. European Journal of Soil Biology 39(1): 13–18.
	mammal	Leranoz, I., Castien, E. 1996. Evolution of wild boar Sus scrofa L, 1758 in Navarra N Iberian peninsula. Journal of the Natural History Museum and Institute Chiba 6(2): 201-207.
	invertebrate	Cuppen, J.G.M. 1986. On the habitats, distribution and life-cycles of the Western European species of the genus *Helochares Mulsant* (Coleoptera: Hydrophilidae). Hydrobiologia, 132 (1986), pp. 169-183

plant	Hernández, H., et al. 2010. ¿Es la rareza geográfica frecuente entre las cactáceas del Desierto Chihuahuense? Revista mexicana de Biodiversidad. 81:183-175.
anphibians and reptiles	Cogălniceanu, D., et al. 2008. The current distribution of herpetofauna in the Maramureş county and the Maramureş Mountains Nature Park, (Maramureş, Romania), Transylvanian Review of Systematical and Ecological Research, Vol. 5, The Maramureş Mountains Nature Park, pp. 189 – 200.
plant	Banuelos, M. J., et al. 2004. Modelling the distribution of *Ilex aquifolium* at the north-eastern edge of its geographical range. Nordic Journal of Botany 23: 129–142.
plant	Miguel-Talonia, C., Téllez-Valdés, O., Murguía-Romero, M. (2014). Las cactáceas del Valle de Tehuacán-Cuicatlán, México: estimación de la calidad del muestreo. Revista Mexicana de Biodiversidad 85: 436–444.
mammal	Scheick, B.K., McCown, W. 2014. Geographic distribution of American black bears in North America. Ursus 25: 24–33.
mammal	Laplana, C., Sevilla, P. 2013. Documenting the biogeographic history of Microtus cabrerae Thomas, 1906 (Arvicolinae, Rodentia, Mammalia) through its fossil record. Mammal Review, 43: 309–332.
invertebrate	Krištín, A., et al. 2007. Distribution and ecology of Ruspolia nitidula (Scopoli 1786) and Aiolopus thalassinus (Fabricius 1781) (Orthoptera) in Slovakia. Linzer Biologische Beiträge, 39, 451–461.
mammal	Queirolo, D., et al. 2011. Historical and current range of the Near Threatened maned wolf Chrysocyon brachyurus in South America. Oryx 45: 296–303.
bird	Tizzani, P., et al. 2013. Recent distribution of red-legged partridge *Alectoris rufa* in Piedmont (North Western Italy): Signs of recent spreading. Avocetta 37:83-86.
invertebrate	Cianferoni, F. 2013. Distribution of Cymatia rogenhoferi (Fieber, 1864) (Hemiptera, Heteroptera, Corixidae) in the West-Palaearctic Region, with the first record for the Italian mainland. North-Western Journal of Zoology 9(2): 245-249.
plant	Costion, C.M., et al. 2013. Palau's rare and threatened palm Ponapea palauensis (Arecaceae): Population density, distribution, and threat assessment. Pac. Sci. 67: 599–608.
invertebrate	Vicente-Arranz, J.C., et al. 2013. Distribution and phenology of the species of family Zygaenidae Latreille, 1809 in the province of Avila (Spain) (Lepidoptera: Zygaenidae). SHILAP 41(161): 113-127.

invertebrate	Popović, M., et al. 2014. Distribution and threats of *Phengaris teleius* (Lepidoptera: Lycaenidae) in Northern Serbia. Acta Zool Hungarica 60:173–183.
anphibian	Cogălniceanu D., et al. 2013. Diversity and distribution of amphibians in Romania. ZooKeys 296: 35- 57.
anphibian	Soorae, P., et al. 2013. Distribution and ecology of the Arabian and Dhofar Toads (Duttaphrynus arabicus and D. dhufarensis) in the United Arab Emirates and adjacent areas of northern Oman. Zool. Middle East., 59, pp. 229-234.
invertebrate	MacGown, J. A., Wetterer, J. K. 2013. Distribution and biological notes of Strumigenys margaritae(Hymenoptera: Formicidae: Dacetini). Terrestrial Arthropod Reviews 6: 247–255.
bird	Andreotti, A., et al. 2008. Landscape-scale spatial distribution of the lanner falcon (Falco biarmicus feldeggii) breeding population in Italy.Ambio 37: 443–447.
bird	Michelsen-Heath, S., Gaze, P. 2007. Changes in abundance and distribution of the rock wren (Xenicus gilviventris) in the South Island, New Zealand. Notornis, 54, pp. 71-78.
mammal	Gaubert, P., et al. 2008. Has the common genet (Genetta genetta) spread into south-eastern France and Italy? Ital. J. Zool. 75, 43–57.
mammal	Uhrin M., et al. 2008. Lesser Mouseeared Bat (*Myotis blythii*) in Slovakia: distributional status with notes on its biology and ecology (Chiroptera: Vespertilionidae). Lynx n. s. 39: 153–190.
reptiles	Pupins M., Pupina A. 2017. Updated distribution of the European Pond Turtle, Emys orbicularis (L., 1758) (Emydidae) on the extreme northern border of its European range in Latvia. Acta Zoologica Bulgarica, Supplement 10: 133–137.
mammal	Encabo, I., et al. 2007. Área de campeo de quirópteros en el término municipal de Carcaixent (Valencia): Nuevas citas para el Atlas de los Mamíferos Terrestres. Galemys 19:37–44.
anphibian	Sillero, N., et al. 2014. Updated distribution and biogeography of amphibians and reptiles of Europe. Amphibia-Reptilia, 35(1):1-31
mammal	Martínez, A.I., et al. 2005. Sondeo y evolución de la distribución de la nutria paleártica (Lutra lutra Linnaeus,

	1758) en el País Vasco (N España). Galemys, 1, pp. 25-46
mammal	Vandel, J.M., Stahl, P. 2005. Distribution trend of the Eurasian *Lynx lynx* populations in France. Mammalia 69,145–158.
mammal	González-Prat, F., Puig, D., Folch, A. 2001. Distribución de la marmota alpina *Marmota marmota* (Linnaeus, 1758) en el extremo suroriental del Pirineo. Galemys 13:139–148.
bird	Brown, A.F., et al. 1995. The distribution, numbers and breeding ecology of Twite Acanthis flavirostris in the south Pennines of England. Bird Study 42: 107–121.
mammal	Kryštufek, B., Vohralík, V. 1994. Distribution of the Forest dormouse *Dryomys nitedula* (Pallas, 1779) (Rodentia, Myoxidae) in Europe. Mammal Review 24: 161–177.
bird	Bruderer, B., Bruderer, H. 1993. Distribution and habitat preference of Red-backed Shrike (*Lanius collurio*) in southern Africa. Ostrich 64:141–147.
invertebrate	Johansson, F. 1993. The distribution of Odonata in Västerbotten and South Lapland, northern Sweden. Entomol Fenn, 4:165–168.
invertebrate	Utrio, P. (1979). Geographic distribution of mosquitoes (Diptera, Culicidae) in eastern Fennoscandia. Notulae Entomologicae, 59: 105–123.
invertebrate	Panelius, S. 1978. The detailed geographical distribution of Tettigonia cantans in Finland (Orthoptera, Tettigoniidae). Notulae Entomologicae, 584: 151-157.
bird	Onmuş, O., et al. 2009. Distribution of breeding birds in the Gediz Delta, Western Turkey. Zoology in the Middle East 47: 39-48.
mammal	Kelly, J. R., et al. 2009. Records of recovering American marten, Martes americana, in New Hampshire. Canadian Field-Naturalist, 123: 1- 6.
mammal	Pavlinić, I., et al. 2010. The atlas of Croatian bats (Chiroptera), part I. Nature Croatia, 19, pp. 295-337.
mammal	Arlt, M.L., Manseau, M. 2011. Historical changes in caribou distribution and land cover in and around Prince Albert National Park: land management implications. Rangifer, Special Issue No. 19, 17–32.
bird	Aguiar, A., et al. 2011. Owls (Strigiformes) in Parque Nacional Peneda-Gerês (PNPG) – Portugal. Nova Acta

	Científica Compostelana (Bioloxía), 19: 83-92.
bird	Kylin, H., et al. 2011. Distribution of the subspecies of Lesser Black-backed Gulls Larus fuscus in sub-Saharan Africa. Bird Study, 58: 186–192.
invertebrate	Harvey, D.J., Gange, A.C., Hawes, C.J., Rink, M. 2011. Bionomics and distribution of the stag beetle, Lucanus cervus (L.) across Europe. Insect Conservation and Diversity, 4, 23–38.
mammal	Palacios, D.M., et al. 2008. Distribution and relative abundance of oceanic cetaceans in Colombia's Pacific EEZ from survey cruises and platforms of opportunity. Journal of Cetacean Research and Management 12(1): 45-50.
mammal	Cano P.D., et al. 2012. Aportes al conocimiento de la distribución del ciervo de los pantanos (Blastocerus dichotomus) en la provincia de Corrientes, Argentina. Mastozoologia Neotropical, 19(1): 35-45.
mammal	Lunney, D., Crowther, M.S., Shannon, I., Bryant, J.V. 2009. Combining a map-based public survey with an estimation of site occupancy to determine the recent and changing distribution of the koala in New South Wales. Wildl. Res. 36:262–273.
mammal	Parsons, B.T., Middleton, A.D. 1937. The distribution of the Grey squirrel (Sciurus carolinensis) in Great Britain in 1937. Journal of Animal Ecology, 6, 386–390.
bird	Karakaş, R. 2012. Does black-winged kite Elanus caeruleus (Desfontaines, 1789) have an expansion in its range in Turkey? Acta Zoologica Bulgarica 64(2): 209-214.
mammal	Jerina, K., et al. 2013. Range and local population densities of brown bear Ursus arctos in Slovenia. — Eur. J. Wildl. Res. 59: 459–467.
mammal	Lloyd, H.G. 1983. Past and present distribution of red and grey squirrels. Mammal Rev., 13: 69–80.
plant	Kalwij, J. M. et al. 2014. Spatially-explicit estimation of geographical representation in large-scale species distribution datasets. – PLoS One 9: e85306.
plant	Van Landuyt, W., et al. 2008. Changes in the distribution area of vascular plants in Flanders (northern Belgium): eutrophication as a major driving force. Biodiversity and Conservation, 17, 3045–3060.
bird	Roselaar, C.S., et al. 2007. Geographic patterns in the distribution of Palaearctic songbirds. Journal of Ornithology, 148: 271–280.
bird	Herrando, S., et al. 2010. Assessing regional variation in conservation value using fine-grained bird

		atlases. Biodivers. Conserv. 19, 867–881.
MPC	mammal	Niemi, M., et al. 2012. Movement data and their application for assessing the current distribution and conservation needs of the endangered Saimaa ringed seal. Endangered Species Research 19: 99–108.
	bird	Mota-Vargas, C., Rojas-Soto, O.R. 2012. The importance of defining the geographic distribution of species for conservation: the case of the Bearded Wood-Partridge. Journal for Nature Conservation 20: 10–17.
	invertebrate	Jiménez-Valverde, A., et al. 2007. Exploring the distribution of Sterocorax Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. Journal of Biogeography, 34: 1426-1438.
	bird	Kleman, L. Jr., Vieira, J. S. 2013. Assessing the extent of occurrence, area of occupancy, territory size, and population size of Marsh Tapaculo (Scytalopus iraiensis). Anim. Biodivers. Conserv. 36: 47–57.
	mammal	Printes, R.C., et al. 2011. Distribution and status of the Critically Endangered blond titi monkey Callicebus barbarabrownae of north-east Brazil. Oryx, 45(3):439-443.
	bird	Mattos, J.C.F., et al. 2009. Abundance, distribution and conservation of the Restinga Antwren Formicivora littoralis (Aves: Thamnophilidae). Bird Conserv Intern 19: 392-400.
	bird	García, J.T., et al. 2008. Assessing the distribution, habitat, and population size of the threatened Dupont's lark Chersophilus duponti in Morocco: lessons for conservation. Oryx 42, 592–599.
	mammal	El Alqamy, H., Baha El Din, S. 2006. Contemporary status and distribution of gazelle species (Gazella dorcas and Gazella leptoceros) in Egypt. Zoology in the Middle East 39, 5-16.
	reptiles	Vasconcelos R, et al. 2013. Review of the distribution and conservation status of the terrestrial reptiles of the Cape Verde Islands. Oryx 47: 77–87.
	mammal	Kowalchuk, K.A., Kuhn, R.G. 2012. Mammal distribution in Nunavut: Inuit harvest data and COSEWIC's species at risk assessment process. Ecol Soc 17:4.
	bird	Tobias, J. A., Brightsmith, D.J. 2007. Distribution, ecology and conservation status of the Blue-headed Macaw Primolius couloni. Biological Conservation 139: 126–138.
	plant	Randrianasolo, A., et al. 2002. Application of IUCN criteria and Red List categories to species of five Anacardiaceae genera in Madagascar. Biodiversity and Conservation, 11, 1289–1300.
	mammal	Thorn, M., et al. 2011. Large-scale distribution patterns of carnivores in northern South Africa: implications for

		conservation and monitoring. Oryx 45(4): 579–586.
	mammal	Martinez, J., Wallace, R.B. 2007. Further notes on the distribution of endemic Bolivian titi monkeys, *Callicebus modestus* and *Callicebus olallae*. Neotropical Primates, *14*(2): 47–54.
	plant	Jaryan, V., et al. 2012. Extent of Occurrence and Area of Occupancy of Tallow Tree (*Sapium sebiferum*): Using the Red list Criteria for Documenting Invasive Species Expanse. *National Academy Science Letters*, Springer.
	reptiles	González-Maya J. F., et al. 2014. "Distribution, range extension, and conservation of the endemic black-headed bushmaster (Lachesis melanocephala) in Costa Rica and Panama," Herpetol. Conserv. Biol., 9(2), 369 – 377.
KDE	-	Steiniger, S., Hunter, A.J.S. 2013. A scaled line-based kernel density estimator for the retrieval of utilization distributions and home ranges from gps movement tracks. Ecol Inform., 13:1–8.
	mammal	Long, J.A., Nelson, T.A. 2012. Time geography and wildlife home range delineation. J. Wildl. Manag. 76 (2), pp. 407-413.
	invertebrate	Irwin, M. D., Coelho, J. R. 2000. Distribution of the Iowan Brood of periodical cicadas (Homoptera: Cicadidae: Magicicada spp.) in Illinois. Ann. Entomol. Soc. Am. 93: 82–89.
	mammal	Charles, C., Schwartz, U.S. 2002. Distribution of grizzly bears in the Greater Yellowstone Ecosystem: 1990–2000. Ursus 13: 203–213.
	mammal	Bader, M. 2000. Distribution of grizzly bears in the U.S. northern Rockies. Northwest Science 74:325–334.
Convex and concave Hull	mammal	Hirsch, A., Chiarello, A.G. 2012. The endangered maned sloth *Bradypus torquatus* of the Brazilian Atlantic forest: a review and update of geographical distribution and habitat preferences. Mammal Review 42: 35–54.
	reptiles	Bombi, P., et al. 2011. When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. Diversity and Distributions, 17, 581–592.
	plant and reptiles	Attorre, F., et al. 2012. The use of spatial ecological modelling as a tool for improving the assessment of geographic range size of threatened species. Nat. Conserv., 21(1), pp. 48-55.
LoCoH	-	Getz, W. and Wilmers, C. 2004. A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. Ecography 27: 489–505.
	-	Getz, W.M., et al. 2007. LoCoH: non-parameteric kernel methods for constructing home ranges and utilization distributions. PLoS ONE, 2, e207.

Indicator kriging		Hengl, T., Sierdsema, H., Radović, A., Dilod, A. 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. Ecological Modeling 24: 3499–3511.
Expert (IUCN) and SDM	mammal	Geissman T, Lwin N, Aung SS, Aung TN, Aung ZM, et al. (2011) A new species of snub-nosed monkey, Genus Rhinopithecus Milne-Edwards, 1872 (Primates, Colobinae), from northern Kachin state, northeastern Myanmar. Am J Primatol 73: 96 –107.
	mammal	Long, Y.C., et al. 1994. Report on the distribution, population, and ecology of the Yunnan snub-nosed monkey (*Rhinopithecus bieti*). *Primates* 35: 241–250.
	mammal	Stone, O.M., et al. 2012. Distribution and population estimate for the chacma baboon (Papio ursinus) in KwaZulu-Natal, South Africa. Primates, 53: 337–344.
	bird	Vanderwerf, E.A., et al. 2013. Current distribution and abundance of the O‘ahu ‘Elepaio (Chasiempis ibidis). Wilson J Ornithol, 125(3): 600–8.
	reptiles	Pike, D.A., Roznik, E.A. 2009. Drowning in a sea of development: distribution and conservation status of a Sand-Swimming Lizard, Plestiodon reynoldsi. Herpetological Conservation and Biology 4:96–105
	mammal	Aquino, R., et al. 2009. Geographic distribution and demography of Pithecia aequatorialis (Pitheciidae) in Peruvian Amazonia. Am. J. Primatol. 71: 964–968.
	mammal	Lizcano, D.J., et al. 2002. Geographic distribution and population size of the mountain tapir (Tapirus pinchaque) in Colombia. Journal of Biogeography, 29, 7–15.
	reptiles	Butler, J.A., Heinrich, G.L. 2013. Distribution of the ornate diamondback terrapin (Malaclemys terrapin macrospilota) in the Big Bend region of Florida. Southeastern Naturalist, 12: 552–567.
	mammal	Balciauskas, L. 2008. Wolf numbers and distribution in Lithuania and problems of species conservation. Ann. Zool. Fenn. 45, 329–334.
	bird	Schroeder, M.A., et al. 2004. Distribution of Sage-Grouse in North America. Condor 106: 363–376.
	mammal	Lortkipanidze, B. 2010. Brown bear distribution and status in the South Caucasus. Ursus 21: 97-103.
	mammal	Wibisono, H.T., Pusparini, W. 2010. Sumatran tiger (*Panthera Tigris Sumatrae*): A review of conservation status. Integrative Zoology 5: 309–318.

mammal	Bergl, R.A., et al. 2010. Remote sensing analysis reveals habitat, dispersal corridors and expanded distribution for the critically endangered Cross River gorilla Gorilla gorilla diehli. Oryx, 46, 278–289.
mammal	Ma, C., et al. 2014. Distribution and conservation status of *Rhinopithecus strykeri* in China. *Primates, 55*(3), 377–382.
mammal	Sampaio, R., et al. 2010. New distribution limits of *Bassaricyon alleni* Thomas 1880 and insights on an overlooked species in the Western Brazilian Amazon. Mammalia 74: 323-327.
mammal	Brugière D., et al. 2009. Distribution of chimpanzees and interactions with humans in Guinea-Bissau and Western Guinea, West Africa. Folia Primatol. 80: 353–358.
mammal	Pinto, L.P.S., Rylands, A.B. (1997). Geographic distribution of the golden-headed lion tamarin, Leontopithecus chysomelas: implications for its management and conservation. Folia Primatologica68:161 – 180.
mammal	Gautier, J.P., et al. 1992. The distribution of Cercopithecus (lhoesti) solatus: An endemic guenon of Gabon. Revue de Ecologie (La Terre et la Vie) 47: 367-381.
mammal	Kierulff, M.C.M., Rylands, A.B. 2003. Census and distribution of the golden lion tamarin (Leontopithecus rosalia). American Journal of Primatology, Washington, 59 (1): 29-44.

Appendix 1.S2: An example application with *Quercus pyrenaica* Wild.

Information source

We downloaded all records for Pyrenean oak (*Quercus pyranaica* Wild.*)* available via GBIF in March 2015. There were 12590 georeferenced records, which were contributed by 103 data publishers from 9 countries. We excluded all records located at sea because they were considered location errors and, eliminated the duplicate rows (records with the same geographic coordinate values) from the database set. These steps led to a reduction of 12355 raw records.

An approximation of the distribution range of the Pyrenean oak using the cartographic method was obtained from the Atlas Florae Europaeae (AFE; Jalas and Suominen 1988), which provides current and historical ranges of native and naturalized European tree species using 50 × 50 km grid. According to AFE, the Pyrenean oak has an Atlantic-Mediterranean distribution, covering mainly from western and south-western France, to the Iberian Peninsula and northern Morocco, with about 95% of the range being included in Spain and Portugal. We build distribution ranges from the previously obtained GBIF data and applied the geographic methods: cartographic, expert-drawn range, MCP, k, r and a-LoCoH, KDE, indicator kriging and convex and concave hull methods (see Figure 1). To build expert-drawn range maps, we randomly selected 20 participants from a group of scientists from the Doñana Biological Station (EBD-CSIC) who were either knowledgeable about the tools to build species distribution ranges or were knowledgeable about the biology and distribution of *Quercus pyrenaica*, or both. The 20 participants were given a sheet where the Spain map and the geo-referenced points of the filtered GBIF database were drawn. They were asked to draw a free hand the species distribution range. A priori, we indicated that the points that appeared represented on the map corresponded with the geographic coordinates obtained from the GBIF database for the species of study. Once we obtained the 20 expert-drawn range maps, we proceeded to digitize them and then superimposed the ranges. Our final distribution range corresponded to the distribution range where we assume an overlap equal to or greater than 85%. For the rest of the geographic methods we followed the methodological procedure indicated in the previous section.

The results showed that, as expected, the distribution ranges reported by each of the methods were different in relation to the total area size, number of fragments and distribution sites (Table 3; Figure 1). For the available records of the species, the cartographic method reported a total of 1,643 presence grids using a grid size of 10 Km (Figure 1c). The average total size of the distribution ranges was 333,312.2 Km^2, being the k-LoCoH method that generated the largest size of distribution range followed by MPC and concave hull. The largest total range showed an increase of 79% over the size of the smallest range, which was built with the cartographic method. The average number of fragments obtained in the construction of the distribution ranges was 15.7 fragments, being the methods that obtained the highest number of a- and r-LoCoH fragments. The MPC and concave hull methods are methods that do not fragment ranges, a necessary characteristic to take into account when choosing geographic methods as this makes it possible to include large areas without information.

Table 1.S2: Results obtained from the total area and number of fragments of the distribution range maps with each geographic method and their corresponding input parameter.

Methods	**Input parameter**	**Total size (Km^2)**	**Number fragments**
Cartographic	10 km grid	164,300.0	-
Expert	≥ 85% concordant range	251,781.4	18
MCP	95% records	517,813.4	1
r-LoCoH	r= 30 Km	282,250.0	29
k-LoCoH	k= 10	778,133.1	1
a-LoCoH	a= 80 Km	198,225.0	42
KDE	h=1.2	298,315.1	15
Indicator Kriging	≥ 25%	271,294.7	13
Concave hull	10 Km	305,294.1	1
Convex hull	30 Km	265,714.7	21

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

CHAPTER II - Outlining distribution ranges with geographic algorithms when data quality is heterogeneous

ABSTRACT. Accurate mapping of the areas where a species is present is fundamental in sciences such as biogeography, macroecology and conservation biology, both for basic and applied purposes. The method used to delineate distribution ranges influences the results and until now, these methods have been insufficiently evaluated in relation with the amount and quality of the information used. The accuracy of the geographic algorithms most commonly used to generate species ranges depends, to a large extent, on the quality of the data, and this dependence is complex. Here, we evaluate by simulation how precise are five geographical algorithms in the estimation of reference ranges with the same total area but varying in shape, number of fragments and heterogeneity in the size of the fragments, and with sets of observations that vary in sample size, spatial distribution, and presence of errors and biases. Adaptive Local Convex Hull (a-LoCoH) and Kernel Density Estimation (KDE) algorithms are the recommended algorithms, with KDE algorithm having the highest sensitivity and a-LoCoH the lowest Type I error rate. Both behaved similarly when describing range fragmentation. Finally, we offer recommendations to minimize the effects of data amount and quality, and provide a guide to help in choosing algorithms when we have to define species distribution ranges based on species observations.

Key words: bias and errors in datasets, geographic algorithms, range maps, reference range, sensitivity, type I error rate.

Ríos-Pena, L., Varela, S., Clavero, M., & Revilla, E. Outlining distribution ranges with geographic algorithms when data quality is heterogeneous (*In prep).*

RESUMEN

El mapeo preciso de las áreas donde está presente una especie es fundamental en ciencias como son la biogeografía, macroecología y biología de la conservación, tanto para fines básicos como aplicados. El método utilizado para delinear los rangos de distribución influye en los resultados y hasta ahora, estos métodos han sido evaluados de manera insuficiente. La precisión de los algoritmos geográficos más comúnmente utilizados para generar áreas de distribución de especies depende en gran medida de la calidad de los datos, y esta dependencia es compleja. Aquí, evaluamos por simulación qué tan precisos son cinco algoritmos geográficos en la estimación de áreas de referencia con el mismo tamaño de área total pero variando en forma, número de fragmentos y heterogeneidad en el tamaño de los fragmentos y, con conjuntos de observaciones que varían en tamaño de muestra, distribución espacial y presencia de errores y sesgos. Los algoritmos recomendados son Adaptive Local Convex Hull (a-LoCoH) y Kernel Density Estimation (KDE), el algoritmo KDE tiene la sensibilidad más alta y a-LoCoH contempla la tasa de error de tipo I más baja. Ambos se comportaron de manera similar cuando describieron la fragmentación de rango. Finalmente, ofrecemos recomendaciones para minimizar los efectos de la cantidad y calidad de datos, y proporcionamos una guía para ayudarnos a elegir un algoritmo cuando tenemos que definir áreas de distribución de especies en función de las observaciones de las especies.

Palabras clave: sesgo y errores en bases de datos, algoritmos geográficos, mapas de áreas, áreas de referencia, sensibilidad, tasa de error tipo I.

INTRODUCTION

A distribution range is a conceptual construct describing the area where a taxon occurs. The distribution range is a central concept in biogeography, macroecology and conservation biology that is used to describe biodiversity patterns, to inform the management and conservation of natural resources, to identify priority areas for conservation or to investigate evolutionary relationships across space (Margules *et al*., 2002, Myers *et al*., 2000, Rondinini *et al*., 2011). Reliable descriptions of species distribution ranges at different spatial and temporal scales are fundamental for conservation (e.g, replicability is critical to define trends) and other research purposes (e.g, a range based on true presence data) (Cox and Moore, 2004; Dormann, 2007). As a conceptual tool, the distribution range is so successful because it provides an upscaled description of the complex spatiotemporal dynamics of populations. Characterizing distribution ranges is fundamental to answer questions dealing with the patterns and processes determining the location of species in space and time. This characterization is usually done through variables such as area, shape and descriptors of boundaries, fragmentation or internal structure (Brown *et al*., 1996, Beselga et al. 2012). However, these properties depend on how distribution ranges are defined, which is in turn influenced not only by the definition and methodological approach chosen, but also by the quality and quantity of data available, issues that are frequently overlooked in the scientific literature.

Many different methods have been developed to generate distribution ranges from observation records. Geographic algorithms use those records to define a topological space representing the area where a species is assumed to be present, given the spatial structure and resolution of the baseline data (Burgman and Fox, 2003; Bronstein *et al*., 2007). These geometric algorithms are substantially different from methods that require additional environmental predictors, such as niche modelling approaches (species distribution models SDM; Boitani *et al*., 2011, Elith *et al*., 2006). While the former generate phenomenological distribution ranges based only on where a species has been observed, the latter generate estimates of environmental suitability. However, the results of SDMs are often translated into areas of probable presence (where the environmental requirements are assumedly covered) and used as if distribution ranges. This usage may be problematic because distributions are also determined by factors not necessarily related to habitat suitability,

such as dispersal barriers, biotic interactions, local population dynamics, human impacts and historical processes (Sexton et al., 2009; Peterson et al., 2011). Expert maps can be considered a type of informal SDMs, since experts tend to use environmental proxies, such as habitat types or geographic accidents, to delineate range areas (Maréchaux et al., 2017; Herkt et al., 2017).

Geo-referenced observations are the basic data used to construct distribution ranges (Hirsch and Chiarello, 2012). The availability of such records is improving due to national and international networks of data mobilization, including citizen science initiatives, and storage (Garcillan *et al*., 2003). The combination of these sources of information have the potential of offering large amounts of data with a broad temporal and geographic coverage (Sousa-Baena *et al*., 2014). However, their heterogeneity may also induce biases (taxonomic, spatial and temporal) and uncertainty (e.g., errors) that may hamper the usefulness of data repositories (Rocchini *et al*., 2011). The dynamic nature of the distribution of species, the quality and quantity of available observations, the process of mapping them and the methods used to delineate range areas will affect the outcome in arguably non-trivial ways (Graham and Hijmans, 2006). There is thus a need to understand how different methods respond to changes in the quantity and quality of the baseline data.

Here, we evaluate the accuracy of five commonly used geographic algorithms when generating distribution areas from records varying in number, spatial distribution and presence of errors and biases in order to provide guidelines on how to delineate distribution ranges while minimizing Type I error rate and maximizing sensitivity. To that aim, we generate hypothetical reference ranges of equal total area, but varying in shape, number of fragments and heterogeneity in fragment size, and simulated sets of records varying in sample size, spatial distribution and presence of errors. These sets of records emulate the variability of available data on real species, based on the patterns observed in the information provided by the Global Biodiversity Information Facility (GBIF) and the International Union for Conservation of Nature (IUCN). We evaluate how accurate the different geographic algorithms are in reproducing reference ranges under several constraints, which reproduce the limitations found in commonly used data sources.

MATERIAL AND METHODS

Reference areas

We constructed nine hypothetical distribution ranges with the same total area, but varying in shape (circle, half bagel and star), number of fragments (one or three) and heterogeneity in the size of the fragments (equal or different, Figure 1). We introduced this variability in order to represent the heterogeneity of shapes that distribution ranges may have (Burgman and Fox, 2003). These hypothetical ranges were latter used as reference ranges to test the accuracy of geographic algorithms in reproducing them using different sets of simulated records.

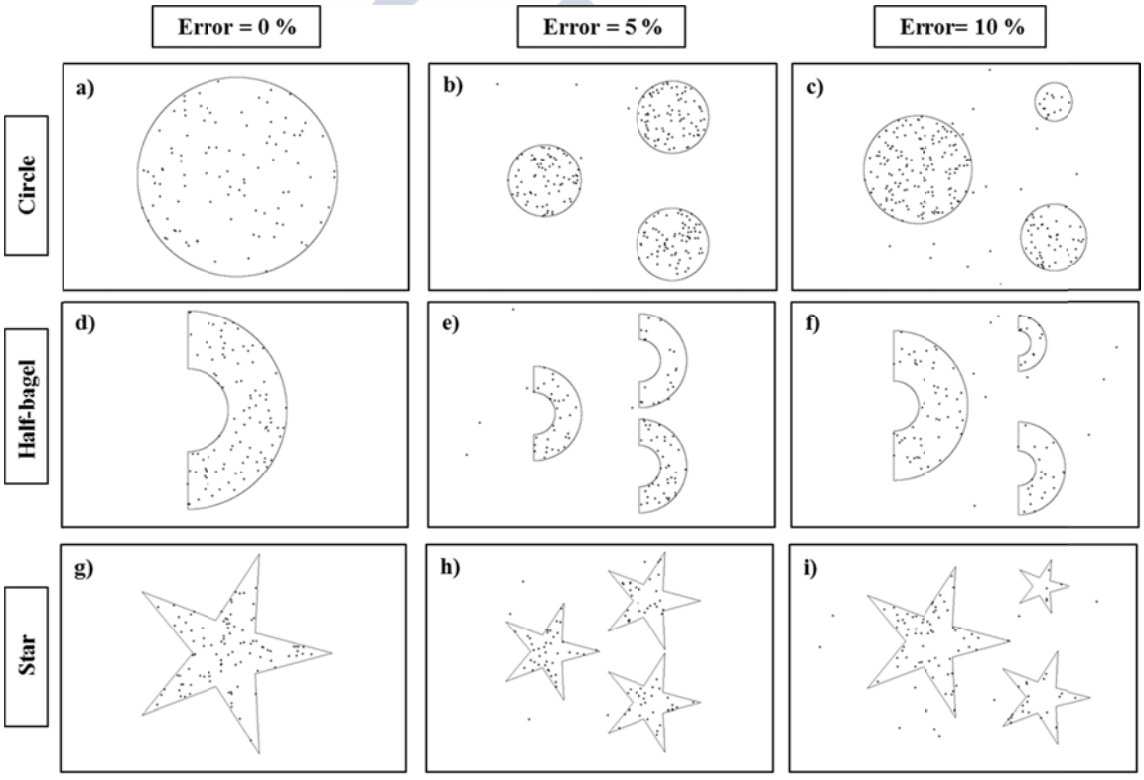


Figure 1. Representation of the reference ranges used to generate the simulated datasets: circular shapes (a, b, c), half bagels (d, e, f) and stars (g, h, i), ranging from one to three fragments (a, d, g), of equal (b, e, h) and different size (c, f, i). The total area is constant in all reference ranges. In this figure simulated records are randomly distributed with a sample size of 100 with no spatial errors (a, d, g) and with the effect of spatial error at 5% (b, e, h) and 10% (c, f, i) of the total sample size.

Simulated species records

We simulated species records within the reference ranges, introducing variability in: i) the number of records; ii) their spatial distribution; and iii) presence of errors (locations outside the reference range). In order to approximate the variations in real species data we described the patterns observed in GBIF data and the IUCN Red List of Threatened Species (The IUCN Red List; IUCN 2014; 2017). GBIF (www.gbif.org) is an international initiative that compiles and distributes data gathered from diverse sources, including museum collections, standardized biological surveys, national and regional databases, citizen science initiatives and direct inputs from individual scientists (Graham *et al*., 2004). The IUCN Red List provides complete and updated distribution ranges of several species (http://www.iucnredlist.org/technical-documents/spatial-data). From these two sources we collected data involving mammals (Class Mammalia). Mammals are a species-rich, globally distributed and thoughtfully studied group that is well represented in the currently available online databases (Meyer *et al*. 2016; Ceballos and Ehrlich, 2006). We used the taxonomy provided by Schipper *et al*., (2008).

We downloaded mammal geo-referenced records collected between 1980 and 2016 from GBIF (accessed December 2016) to obtain a reference of the amount of data recently collected in Mammals. We used that information to set the range of values to be used in the generation of datasets. From the IUCN Red List we obtained range maps of 4,440 mammal species, selecting only parts of the species' ranges coded as "extant". Only 3,392 species were taxonomically coincident between GBIF and IUCN, and they accumulated 3,012,333 geo-referenced records in GBIF during the time of reference.

To test how the spatial distribution of records affects the characterization of species ranges, we simulated records using three types of spatial distributions: 1) random, with randomly distributed records; 2) uniform, with records distributed at regular distance intervals; and 3) clustered, with records distributed in groups of heterogeneous size. To simulate the clustering structure of real species records, we focused on the 2,224,505 geo-referenced records provided by GBIF for terrestrial, continental mammals (i.e. eliminating all records in the sea, in islands with a surface area equal to or less than 50,000 km^2 and the Antarctica, Figure 2a). We assigned each record to one of the following five main landmasses: North America, South America (both separated by the Isthmus of Panama),

Africa, Eurasia and Australia. We used a density-based clustering algorithm (DBSCAN, Ester *et al*., 1996) to characterize the spatial aggregation of species records and identify the main clusters, defined as dense regions in the data space separated by areas with a lower density of records. DBSCAN searches for an optimal number of clusters on the basis of two parameters: a distance threshold (ε) that defines the neighborhood of a record and a minimum number of records (m) required to define a dense region. The DBSCAN algorithm starts by randomly selecting a record, then it takes its ε-neighborhood and, if it contains at least m elements, it aggregates the records into the same cluster. The process goes across all records, creating density-connected clusters. The optimal ε value is estimated with the average of the average distances of every point to its k-nearest neighbors, where k value is specified by the user. Next, these k-distances are plotted in an ascending order. The aim is to determine the knee in the distribution, which corresponds to the optimal ε (the threshold where a sharp change occurs along the k-distance curve) (Ester *et al*., 1996). We calculated the optimal ε for each continent, obtaining the following values: Africa 2.5 km, Australia 1.0 km, Eurasia 2.2 km, North America 2.0 km, and South America 1.8 km, along with an m of 100 in Africa and South America and 150 for the remaining 3 continents. We run this procedure with each of the continents using dbscan function of the fpc R package (Hennig, 2015) (R Core Team, 2016; version 3.2.5). We calculated the center of gravity of each cluster to obtain the frequency distribution of the number of records as a function of the distance to it (Figure 2b). Then, we calculated the number of clusters and centers of gravity that fell within the distribution range of each mammal species (based on the distribution polygons provided by the IUCN Red List) and used the mean number of clusters per species and the probability distribution of the distance of records to the center of gravity of clusters to generate simulated spatially biased datasets.

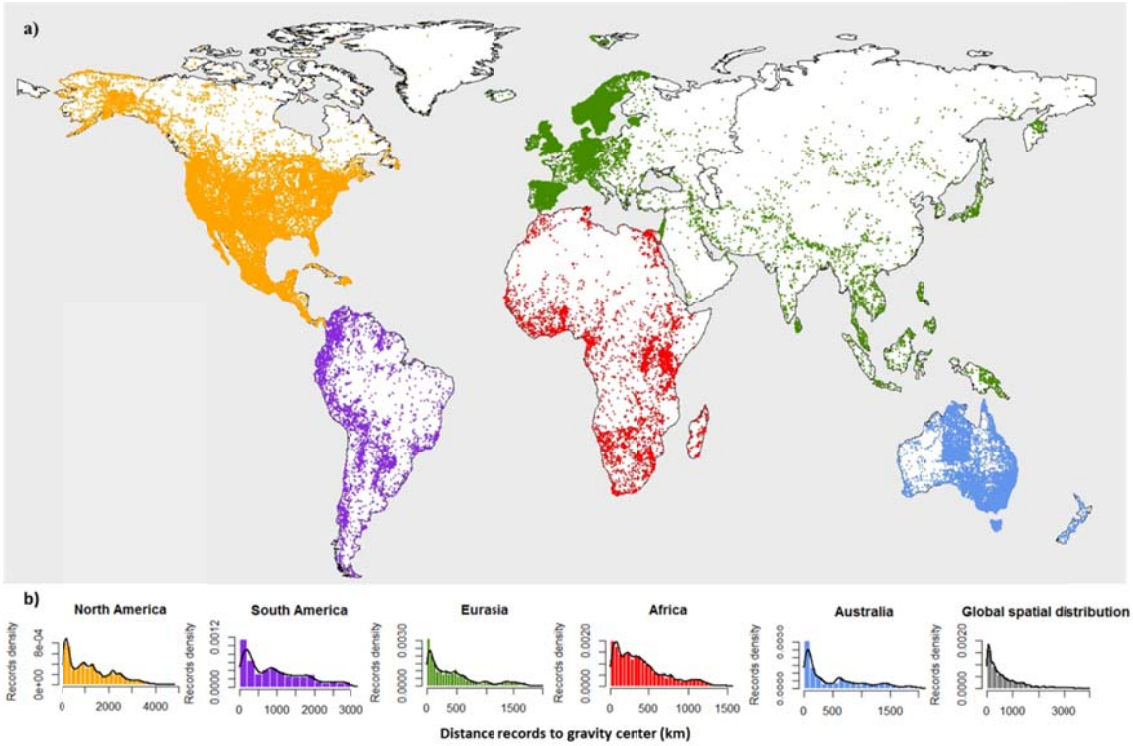


Figure 2. Spatial distribution of the 2,387,155 geo-referenced records in GBIF for terrestrial mammals between 1980 and 2017 (geo-referenced records in the ocean, islands with a size equal to or less than 50,000 km^2 and Antarctica were not considered) matching the binomial or trinomial scientific name of species with IUCN distribution maps. Shown are spatial distributions of the geo-referenced records in clusters as obtained using the DBSCAN algorithm per continents and to global scale (b): North America (orange), South America (purple), Eurasia (green), Africa (red), Australia (blue) and, global spatial distribution (grey). Plots with bars ordered proportionally to the number of records and the distances in Km to the centers of gravity of clusters for each continent and at global scale.

From the several potential sources of error of species records, here we refer only to spatial misallocations, which result in records in areas where the focal species is absent (Hurlbert and Jetz, 2007). In order to approximate a lower bound for the error rates that might be present in the available data, we estimated the proportion of records corresponding to terrestrial mammal species that lay in the ocean or the Antarctica, which we consider a conservative estimate of the error rate (i.e. only part misallocations are detected). For the upper bound of the error rate we assumed the IUCN range maps as ground truth and calculated for each species the proportion of GBIF records that fell

outside the range map. Also, we calculated the measures of central dispersion of records falling outside IUCN range maps. We consider this proportion as a high potential error rate (an overestimate of the real errors), since an unknown proportion of records outside IUCN maps do indeed correspond to true records. Within this two extreme bounds we simulated four error rates in our dataset: 0.05, 0.1 and 0.2 (supplementary Table 1.S1). Finally, on the data set that contained 10% errors, we eliminated 5% of the most extreme values and recalculated scenarios of errors.

We used *spsample* function of the *sp* R package (Pebesma and Bivand, 2005) to produce the different sets of simulated species records, introducing the variability in numbers, spatial distribution and error rate.

Geographic algorithms

We selected five widely used algorithms to generate species range maps: Minimum Convex Polygon (MCP), Kernel Density Estimation (KDE) and k, r and a Local Convex Hulls (LoCoH).

MCP generates the smallest convex polygon that contains all records and has no internal angles exceeding 180° (Rapoport, 1982; O'Rourke, 1998). This method generates a single polygon (i.e. it does not consider the possible fragmentation of a range) and does not require input parameters. Polygons can be generated for any given percentage of the available records by excluding the most extreme observations. We generated species ranges with 100%, 95%, 90%, 85% and 80% of our simulated records using "mcp" function of the *adehabitatHR* package (Calenge, 2015) in R.

KDE requires the selection of a bandwidth parameter (h), a free parameter that has a strong influence over the resulting range estimate. The bandwidth determines the relationship between the distance of a given observation from an evaluation point and the contribution of the location to the density estimate at that point. We selected the fixed kernel method (method where h remains constant for all records), estimating the bandwidth through least-squares Cross-Validation method because it uses a resampling, cross-validation approach that minimizes error between true and estimated distributions (LSCV, Li and Racine, 2003; Gitzen *et al*., 2006). We used "npudens" function of the *np* R

package, which uses the method of Li and Racine (2003) to obtain the kernel density function. We obtained range maps applying a Thin Plate Splines (TPS) model (Donato and Belongie, 2002) to the weighted density of records. TPS are a spline-based technique for data interpolation and smoothing. We used the 0%, 5%, 10%, 15% and 20% isolines to plot the density map.

LoCoH methods (k-LoCoH, r-LoCoH and a-LoCoH; Getz *et al.*, 2007) have in common the construction of small convex hulls for each observation and its neighbors. Convex hulls are merged together starting from the smallest to the largest until all records are included. k-LoCoH constructs convex hulls associated with each observation and its $(k-1)$ nearest neighbors. r-LoCoH constructs convex hulls from all records at distance r, being r half of the maximum nearest neighbor distance between records. Finally, a-LoCoH generates convex hulls from all records within a radius a such that the sum of the distances between the records is less than or equal to a parameter. The three LoCoH methods require the estimation of their respective input parameters (k, r and a). This is a key issue because relatively low values of the parameters can generate a high level of fragmentation in the resulting ranges, which disappear with higher values (Getz *et al.*, 2007). We used the Minimum Spurious Hole Covering (MSHC) rule (Getz and Wilmers, 2004) to select k, r and a values of the parameters. The values obtained with the MSHC rule were a first approximation for the selection of input parameters. We chose two values above and two values below the value obtained with MSHC (supplementary Table 1.S1), resulting in five values of parameter studied for each LoCoH method.

In summary, we generated 9 reference ranges, and for each of them we considered 7 values of sample size (number of records), with three types of spatial distribution and four levels of error. For each of these scenarios we used the five geographic algorithms with five values of input parameters in each algorithm, to generate species ranges aimed at reproducing the reference ranges. This approach generated 18,900 combinations of factors to study ($9 \times 7 \times 3 \times 4 \times 5 \times 5$), each one of which was replicated 50 times, so we generated a total of 945,000 distribution ranges that provide us with acceptable estimates of the mean and 95% confidence intervals (Manly, 1997) for the accuracy metrics.

Accuracy of distribution ranges

For each of the estimated ranges we calculated the total area (A_E) and the total number of fragments (F_R) and compared these values with its reference range. We based this comparison on three metrics: i) sensitivity; ii) Type I error rate; and iii) Observed to predicted fragments ratio (henceforth fragment ratio) (see definitions in Table 1). Sensitivity and Type I error rate have values between 0 and 1. A sensitivity equal to 1 implies that the range generated by the geographic algorithm encompasses all the reference range, while a sensitivity of 0 would indicate no coincidence. Likewise, Type I error rate would be 1 if the range generated by the geographic algorithm did not overlap with the reference range, and 0 when all the range generated by the algorithm was included within the reference range. Values of the ratio of predicted fragments range between 0 and ∞. A value of 1 would indicate that the range generated by the geographic algorithms had the same number of fragments than the reference range, while smaller and larger values would indicate less and more fragments, respectively (Table 1).

Table 1. Descriptors of the distribution ranges estimated with the geographic algorithms and for their comparison with the reference ranges.

Measure	Description	Symbol/Formula
Area	Area of the reference ranges.	A_R
Estimated area	Area of the ranges generated with the geographic algorithms.	A_E
Number of fragments	Number of fragments of the reference ranges	F_R
Estimated number of fragments	Number of fragments generated by the geographic algorithms.	F_E
True positive area	Overlapping area between the reference ranges and the estimated ranges.	a
False positive area	Area included in the estimated ranges but not in the reference ranges.	A_E - a
False negative	Area included in the reference range but not	A_R - a

area	in the estimated ranges.	
Sensitivity	True positive area in relation to the area of the reference range: proportion of the reference range correctly predicted by the estimated range.	a / A_R
Type I error rate	False positive area in relation to the area of the estimated range: proportion of the estimated range that is not included in the reference range.	$(A_E - a) / A_E$
Type II error rate	False negative area in relation to the area of the reference range: proportion of the reference range that is not included in the estimated range.	$(A_R - a) / A_R$
Ratio of predicted fragments	Ratio of the number of estimated fragments to the number of fragments of the reference ranges.	F_E / F_R

We analyzed the variations in sensitivity and Type I error rate obtained in each simulated scenario using generalized linear models (GLMs, McCullagh and Nelder, 1989) with beta distribution and a logit link, using shape, number of fragments, number of records, spatial distribution, error rate, algorithms and parametrization as explanatory variables. Beta regression is suitable for modelling continuous variables restricted to the standard unit interval, as it incorporates the natural asymmetry and heteroscedasticity of these data (Ferrari and Cribari-Neto, 2004). These models were constructed using the "betareg" function in the *betareg* R package (Cribari-Neto and Zeileis, 2010) and were fitted via Maximum Likelihood estimation of regression parameters. The variation in the ratio of predicted fragments was analyzed using log-normal GLMs. Model fitting via Maximum Likelihood estimation of regression parameters were done using the "glm" function in the stats R package.

To construct the regression models, we first set the 3 levels of spatial distribution (uniform, random and clustered) and construct the models for each algorithm with its 5

input parameters. We then analyze the effect of a biased spatial distribution and the proportion of errors (5, 10 and 20%) on the predictors for the three dependent variables, that is, we establish two factors and calculate the estimates per algorithm. Finally, we calculated the regression models incorporating the effect of the shape of the reference range as a categorical variable for the three spatial distribution levels fixed and added the effect of spatial errors as a continuous explanatory variable for each algorithm. A total of 154 regression models were constructed.

RESULTS

Effect of the amount of information available

A total of 4,403 species had at least one record (i.e. around 20% of the species lacked recent records). In the group of species with information, 370 (8%) had a single record and 1432 (33%) had less than 10. The mean number of records per species (considering only species with at least one record) was 928.4, median 29, denoting a strongly right-skewed distribution (Table 2). These values do not change much if we consider all data available in GBIF (Table 2). We chose the number of records for our simulated datasets as a function of the observed frequency distribution of mammal records, selecting seven different levels of sample size, ranging between 10 and 1000 records, corresponding approximately with quantiles 0.30, 0.60, 0.70, 0.80, and 0.85 (8, 56, 114, 267 and 473, respectively), the median value (29) and the mean value (929).

Table 2: Number of records available and measures of central dispersion during different periods and total across Mammalia class registered in GBIF. Only species with at least one record are considered. Data downloaded in December 2016.

Time window	N records	N Species	N Genus	N Families	N Orders	Range (Min-Max) /sp	Avg (SD) /sp	Median (Q1;Q3) /sp
Before 1980	1,785,955	4,300	1,136	209	29	1-114,849	415.3 (2683.2)	30 (6;156)
1980 - 1989	618,963	2,945	1,102	192	28	1-39,057	210.2 (1004.7)	16 (4;84)
1990 - 1999	943,627	3,064	1,084	195	29	1-103,028	308.0 (2343.8)	15 (4;87)
2000 - 2009	1,132,152	3,094	1,134	216	29	1-64,368	365.9 (2330.2)	14 (3;81)
2010 - 2016	1,392,915	2,440	1,000	203	28	1-495,474	570.9 (10696.9)	9 (2;51)
Total	5,873,612	4,404	1,456	263	29	1-496,872	938.4 (9417.0)	30 (6;170)

Sensitivity improves with increasing number of records for all algorithms when records are distributed uniformly or randomly, reaching values above 0.75 with as few as 100 records (Figure 3a, 3b and supplementary Fig 1.S2, a and d). The highest sensitivities were obtained with k-LoCoH and kernel algorithms at small sample sizes, while at high sample sizes the three LoCoH algorithms performed better (Figure 3a, 3b). Type I error rates increased with sample size for the r-LoCoH and were quite stable for the MCP at around 0.5 (Figure 3f, 3g). The remaining algorithms show decreasing Type I error rates with increasing sample size. The a-LoCoH had low Type I error rates (bellow 0.25) even for low sample sizes (Figure 3f, 3g). At low sample sizes all methods underestimated the number of fragments (Figure 3k, 3l). Kernels severely overestimated the number of fragments with increasing sample size, while k-LoCoH and, especially, a-LoCoH tended to provide accurate estimations of the number of fragments (Figure 3k, 3l). The shape, number of fragments and the heterogeneity of the reference ranges affect the quality of the range estimates obtained with the different algorithms, especially when their shape is

irregular and concave, i.e., half bagel and star shapes, but their relative performance is coarsely maintained (Table 3, supplementary Fig 3.S2).

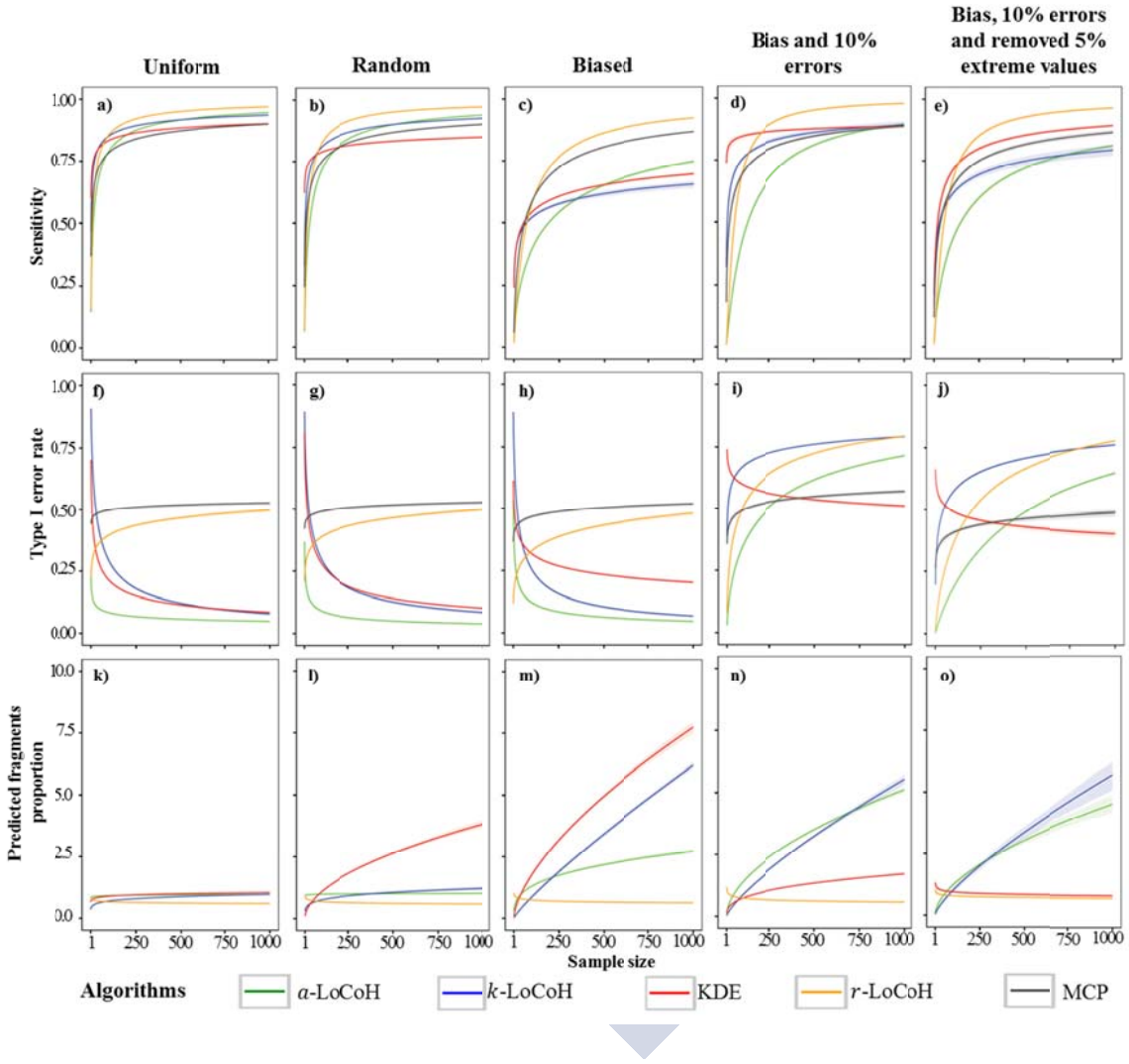


Figure 3. Example plots of the prediction of generalized linear models with beta distribution describing the sensitivity and type I error rate with records following uniform spatial distribution (a, f), random spatial distribution (b, g) and biased spatial distribution of data (c, h) and generalized linear models with log-normal distribution for explaining the proportion fragments predicted respect to size sample with uniform spatial distribution (k), random spatial distribution (l) and clustered spatial distribution (m). We also show the sensitivity, type I error rate and proportion of predicted fragments when the data present bias and 10% errors (d, i, n) and bias and 10% errors but extract 5% of the most extreme errors (e, j, o). The solid lines depict the prediction and the shaded areas depict 95%

confidence intervals for a reference range where the sample size responds to a continuous explanatory variable.

Table 3. Qualitative summary of the results obtained when evaluating the accuracy of the five geographical algorithms when the quality and quantity of information available varies in sample size and presence of spatial bias and presence of errors. X: adequate behavior with increasing sample size; XX: good performance; XXX: best performing algorithm; ( ): classification when a fraction of the observations (extreme) is left unused to control for errors.

Methods	**Data quality**				**Aims**
	uniform	random	biased	biased + errors	
MPC	XX	X	XX	X	sensitivity
					Type I error rate
					spatial structure
KDE	XX	X	X	X	sensitivity
	X	X	X	X	Type I error rate
	XXX			X	spatial structure
a-LoCoH	XX	X	X	X	sensitivity
	XXX	XXX	XXX		Type I error rate
	XXX	XXX			spatial structure
k-LoCoH	XX	X	X	X	sensitivity
	X	X	X		Type I error rate
	XXX	X			spatial structure
r-LoCoH	XXX	XXX	XXX	XXX	sensitivity

				Type I error rate
	X		(X)	spatial structure

These results indicate that if our sample size is small and we need to maximize the true area included in the estimated range we should favor k-LoCoH and kernel methods. At larger sample sizes, the three LoCoH methods behave reasonably well. However, if our objective is to minimize the true area included in the estimated range we should favor the use of the a-LoCoH method.

Effect of biases in the spatial distribution of records

The density of records available per continent shows a strong spatial bias in sampling effort. Of a total of 2,224,505 records, Eurasia had the greatest data density with 103,987.5 records/10^6 km^2 (47.6% the available records), followed by Australia with 77,050.9, North America with 14,434.3, South America 4,232.6 and Africa with 1,682.1 records/10^6 km^2 (29.5%, 17.1%, 3.4% and 2.3% of the available records, respectively). Within continents, the spatial distribution of records was far from homogeneous, reaching maxima in parts of Europe, North America and Australia and minima in large parts of Asia, Africa and South America (Figure 2a). We identified a total of 26, 23, 13, 8 and 6 clusters in Eurasia, Africa, Australia, North America and South America, respectively. The number of records available decreased very fast as we move away from the center of gravity of those clusters, both at continental and global scales (Figure 2b). On average, IUCN species ranges overlapped with 3.2 clusters (SD= 7.85, median = 1.0, 3rd quartile = 3.0). The average number of centers of gravity within those species ranges was 1.1 (SD = 2.54, median = 0.5, 3rd quartile = 1.6). For 84 out of the 3,005 terrestrial species evaluated here, the range provided by the IUCN Red List did not overlap with any cluster, while 1462 species ranges did not contain any center of gravity. Based on these results we selected the mean value of the number of clusters overlapping IUCN ranges on a global scale to simulate spatial bias in the distribution of records.

As expected, spatially biased data strongly influences range estimation (Table 3, Figure 3c, h, m). Sample sizes higher than 200 records are required to obtain sensitivities above 0.5, and, even with large sample sizes (≥ 500 records), only r-LoCoH and MPC were able to reach sensitivities above 0.75 (Figure 3c). Most of the reduced sensitivity due to spatially-biased data occurs in irregular or fragmented reference ranges, while the sensitivity of the circular reference ranges was barely affected by this bias (supplementary Figure 3.S2). Type I error rate increased with sample size for the r-LoCoH and was high for the MCP at any sample size (Figure 3h). For the remaining algorithms Type I error rates behaved properly, decreasing with increasing sample size. The a-LoCoH had low error rates (below 0.25) even at low sample size (Figure 3c, h, m). The algorithms that estimated correctly the number of fragments when data were unbiased (a-LoCoH and k-LoCoH), overestimate the number of fragments with increasing sample size when using spatially-biased baseline data (Figure 3m).

Additional effect of spatial errors

Of the 3,005 species of coincident terrestrial mammals between GBIF and IUCN ranges, 1,305 species had at least one record in the ocean, water bodies or the Antarctica. Assuming all these records as errors, the average lower bound error rate was 7.6% errors per species (SD = 18.99, median = 0.00, 1ºquantile = 0.00). A total of 307 species (10.2%) did not contain any GBIF record within their IUCN distribution range, while 536 (17.8%) had all records within it. Of the remaining 3,005 species, 51.1% had less than 25% of their records outside their IUCN range, 32.8% of the species had between 25% and 75% of their records outside and 16.1% of the species had more than 75% of their records outside. The median percentage of records outside the IUCN range was 23.3% (average 33.7%, SD = 32.84, 1ºquantile = 6.1 and 3ºquantile = 50.0) which we took as an upper estimate of the error rate. Based on this information, we selected three levels of error, 5, 10 and 20%, and investigated the combined effect of spatially biased data with errors.

In the presence of spatial errors, the algorithms tend to overestimate ranges (i.e. higher Type I error rates) a trend that is more evident as the percentage of error increases (Table 3, supplementary Figure 1.S2j and Figure 2.S2). Increasing sample size improves sensitivity, at the cost of reaching high Type I error rates (Figure 3d, 3i). KDE was the only

algorithm showing a slightly decreasing Type I error rate with increasing sample size in the presence of errors and was performing reasonably well in describing the number of fragments (Figure 3n). Finally, when errors are present, eliminating extreme records may potentially help in reducing the impact of errors. The elimination of the 5% most spatially extreme records from the simulated datasets (note that they are not necessarily errors) reduced type I error rate and improved the estimation of the number of fragments, but sensitivity decreased for all algorithms except r and a-LoCoH that barely noticeable (Figure 3e, j, o, supplementary Figure 1.S2, m-o).

DISCUSSION

The algorithms we have explored have been evaluated mostly in relation with the delineation of home ranges, normally with data of high quality, with virtually no errors or biases and from the perspective of defining a good spatial representation (spatial structure of the home range) as a function of sample size (Gaston and Fuller *et al*., 2009). Our work shows that the accuracy of geographic algorithms used to generate species ranges largely depends on the interaction between the quality of data and the method used, and that this relationship is complex. This problem has long been acknowledged, but there are no clear recommendations on how to build a distribution range using data on species presence of heterogeneous quality and it is still too often the case that ranges are delineated without offering information on the quality of the data used or even the method used.

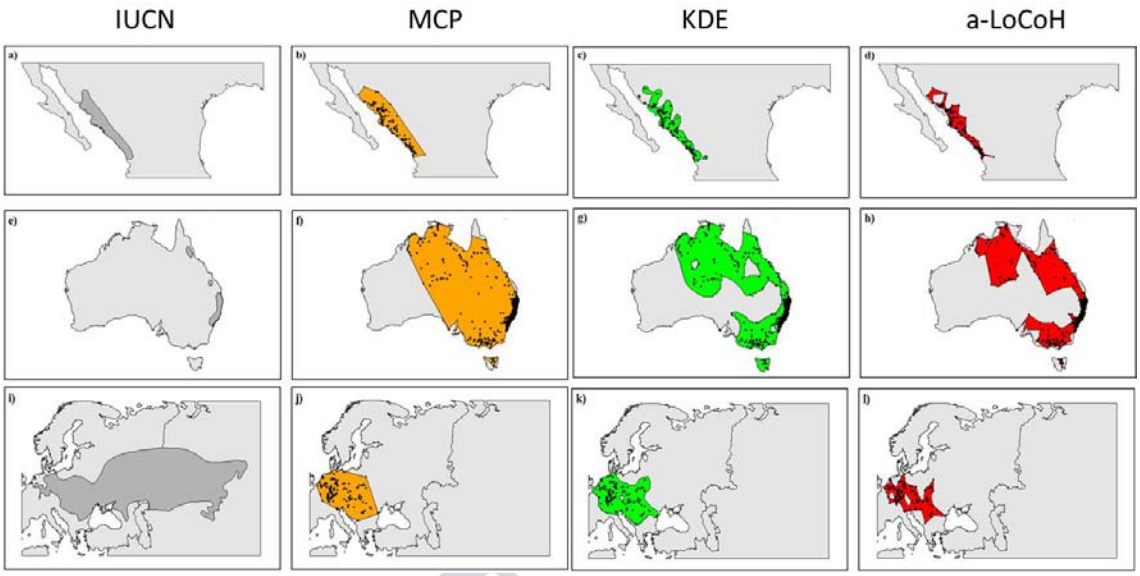


Figure 4: Examples of distribution ranges for three species, *Chaetodipus pernix* (a-d), *Vespadelus pumilus* (e-h), and *Cricetus cricetus* (i-l), as provided by the IUCN (extant) and estimated using georeferenced records available in the Global Biodiversity Information Facility (GBIF) using MCP, KDE and a-LoCoH algorithms. The three species represent cases with a good coverage (*Chaetodipus*) and biased sampling effort (*Vespadelus* and *Cricetus*).

The impact of data quality

Most often, the data available to define ranges is of heterogeneous quality. We explored three components of data quality, the quantity of data, their spatial bias and the presence of errors, of which sample size was the easiest to evaluate (Boitani *et al*., 2011; Burgman and Fox, 2003). In general, when baseline data were randomly or uniformly distributed the accuracy of all geographic algorithms improved with sample size, levelling with as few as 200 to 250 records. Algorithms are robust even at small sample sizes, but the tradeoff between sensitivity and type I error rate that occurs in MCP and r-LoCoH make these two methods a non-preferred choice. This is good news since most of the species have few data available, with median data availability as low as 30 records. If the distribution range of the focal species is irregular or fragmented, a-LoCoH and KDE are the methods of choice, with the first yielding the lowest Type I error rates and a more accurate representation of the fragmentation of the range.

Data originated from a uniform or an unbiased random sampling are rare or lacking for most regions and species (Gaston and Rodrigues, 2003; Rocchini *et al*., 2011). This type of data is normally produced in specific surveys, as may occur with species subject to monitoring, species with a restricted distribution range in areas with a high density of records or data generated in systematic national inventories. In the rest of the cases, the heterogeneity of sampling effort induces a background bias that may affect the estimation of ranges (Meyer *et al*., 2016; Pimm *et al*., 2014). A simple look at the distribution of all records across species shows that data density varies in space and that we can easily identify clusters of data overlapping with the distribution ranges defined by the IUCN for most species. This type of bias strongly decreased the sensitivity of all methods, especially when the reference distribution range is irregular or fragmented. This means that the distribution ranges generated with the data currently available will leave undetected areas where the focal species are present. KDE is the method with the best simultaneous behavior of both sensitivity and type I error rate, but at the cost of requiring relatively large data sets, and, even then, leaving as much as 25% to 35% of the reference range undetected. The existence of spatial biases in the data precludes the detection of complete ranges, making necessary to increase sample size to improve estimates. Spatial biases in species records are relevant in GBIF and other global data sources because heterogeneous factors such as human population density, access to technology, presence of a well-developed transport system or funding availability may affect their collection, storage and mobilization (Beck *et al*., 2013). At this point, it is important to work to characterize and, if possible, reduce the presence of spatial biases in data repositories (Cantú-Salazar and Gaston, 2013; Beck *et al*., 2013; 2014).

Spatial errors are another widespread problem present in biodiversity databases (Maldonado *et al*., 2015). They can be generated in many ways and at any moment of the data lifecycle. Nevertheless, it is very difficult to obtain accurate overall estimates of how important this problem is. We used a conservative approach to define lower and upper bounds within which the actual error rates may be located. Our lower bound estimate shows that errors are indeed a problem, with more than 40% of the terrestrial species having records in the ocean, water bodies or the Antarctica. The upper bound is much more difficult to estimate. We used the IUCN distribution ranges, which depict areas of potential distribution created using a combination of a geographic algorithm and expert opinion (and

are therefore more akin to SDM), as ground truth to compare how often the available observations lay outside. As we have seen, geographic algorithms used with biased data will leave outside an important fraction of the true range, and therefore the median 23% of observations outside IUCN ranges is an upper overestimate of the actual error rate. The presence of errors affect the performance of geographic algorithms, being its main drawback the overestimation of the distribution range (Getz and Wilmers, 2004; Burgman and Fox, 2003). The reliability of the ranges obtained depends largely on the quantification and control of spatial errors in the sources of information. If species occupy very small spatial scales in relation with the errors, the results will be sensitive to the actual location of the errors. If species have large geographic ranges and the proportion of location errors is small, the results will be less affected. When data contain errors and sampling effort is spatially biased, there is a substantial deterioration of Type I error rates, which increase with sample size in all methods except for KDE. Even at large sample sizes and for KDE as much as 50% of the area delineated as part of the range may be incorrectly included within the range. Algorithms, such as a-LoCoH, with a good simultaneous behavior in sensitivity and Type I error rates are strongly affected by the presence of errors. One possible way to reduce the impact of spatial errors is to exclude extreme values from the dataset. The exclusion of extreme records before constructing the ranges helps reasonably to improve the accuracy of the algorithms to reproduce the reference ranges mostly by reducing the Type I error rate, but it does not affect qualitatively the overall performance of the different algorithms.

How to define a distribution range using geographic algorithms

The first recommendation is that when defining a distribution range we must be explicit with our aims, the data used and its quality, and the method chosen (Ríos-Pena *et al*., in prep-chapter 1). Geographic algorithms are often used when we need to maximize sensitivity while minimizing Type I error rates to identify areas actually occupied, while species distribution models offer a good approximation to the areas with enough quality to be potentially occupied and therefore tend to overestimate the area occupied (high Type I error rates). It is too often the case that authors do not provide information on how ranges are built (Ríos-Pena *et al*., in prep-chapter 1). Some characteristics of the data available

may have important consequences in the definition of ranges, such as biases in sampling effort that affect the spatial distribution of records or the existence of errors (Figure 4). Data quantity and quality should also be explicit to acknowledge the limitations of the method of choice. The only thing we can do when there is no or few data available is to collect information, while we can minimize the impact of bias, by resampling data in oversampled areas, and that of errors by carefully crosschecking the data and by removing a fraction of extreme observations.

Our results show that based on actual data there is no single best method simultaneously for sensitivity, Type I error rate and range fragmentation. Depending on our aims and the quantity and quality of data available, some methods should be preferred over others (Guillera-Arroita *et al*., 2015; Qiao *et al*., 2015; Diniz-Filho *et al*., 2015). All methods show a good behavior for sensitivity, as expected since they have been designed to maximize it, even at low sample sizes. In most cases, this is so at a high cost in Type I error rate, including large areas where the species might not be present. More importantly, not all methods behave properly in their Type I error rate with increasing sample size, as is the case with MCP and r-LoCoH, which increase their error rate with growing sample sizes and therefore, should be avoided. Range fragmentation is the most difficult property to reproduce. KDE and a-LoCoH have the best behavior when data is unbiased and errors free (Figure 4).

In case we have a good quality dataset with a not too complex spatial configuration, it is straightforward to define the range (Figure 4). KDE offers a good compromise if data is biased and there is a possibility of errors in the dataset. Again, the estimation of the kernel value out of the data should be explicit. In case we need to be sure that the area is occupied by the focal species, a-LoCoH performs well in avoiding the inclusion of false positive areas, maintain a low error rate when there are no errors in the dataset. At low sample sizes, a combination of both methods provides a core area defined by the a-LoCoH with a low Type I error rate plus a larger area defined by the KDE where the error rate may be higher. Nevertheless, we should differentiate the areas generated by each of the methods when generating ranges using a combination of approaches since the uncertainty of the different areas may be substantially different. This recommendation also applies when combining geographic algorithms with SDM or expert knowledge, as is the case of UICN

distribution ranges. Doing so would help in controlling the uncertainty when using those ranges in theoretical and applied research.

CONCLUSIONS

Our results have strong implications for the construction of species distribution maps based only actual observations and the mobilization of data. Georeferenced records available in global databases suffer from gaps in data coverage and spatial, taxonomic and species-level biases (Ficetola *et al*., 2014; Meyer *et al*., 2016). The quality of available information is not homogeneous across species, nor are species lacking information randomly distributed across families and regions. In short, the heterogeneity in data availability and quality is a serious limitation to generate unbiased distribution ranges. The lack of standardized criteria to accept minimum levels in the quantity and quality of information and in the methods used hinder the potential use of distribution ranges in applications such as the prioritization of conservation, interspecific comparative studies and other basic and applied uses in research.

Our study demonstrates that a correct estimate of distribution ranges requires data of good quality. To this end, we should apply substantial amounts of taxonomic knowledge, time and funding to collect, verify and clean up public databases. Additionally, users should carefully clean up the datasets before use by conservatively removing poorly annotated records and those that may have an erroneous location. We have to be aware of the requirements and limitations of the different geographic algorithms to estimate distribution ranges depending on the type of data and research question that we want to address and accordingly select the one that most suits our needs. Finally, in all cases we must be transparent with the data and the method used.

ACKNOWLEDGEMENTS

We are grateful to the GBIF and IUCN Red List team for making and maintaining their databases freely available online. We also acknowledge the Conservation Biology group of Estación Biológica de Doñana, CSIC for helpful suggestions. This work was

supported by the following grants and projects: Spanish Ministry of Economy, Industry and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R+D+I (SEV-2012-0262) and Agencia Estatal de Investigación from Ministry of Economy, Industry and Competitiveness, Spain with projects CGL2012-35931 and CGL2017-83045-R AEI/FEDER EU, co-financed with FEDER to E.R., R.B.M., M.G.S.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

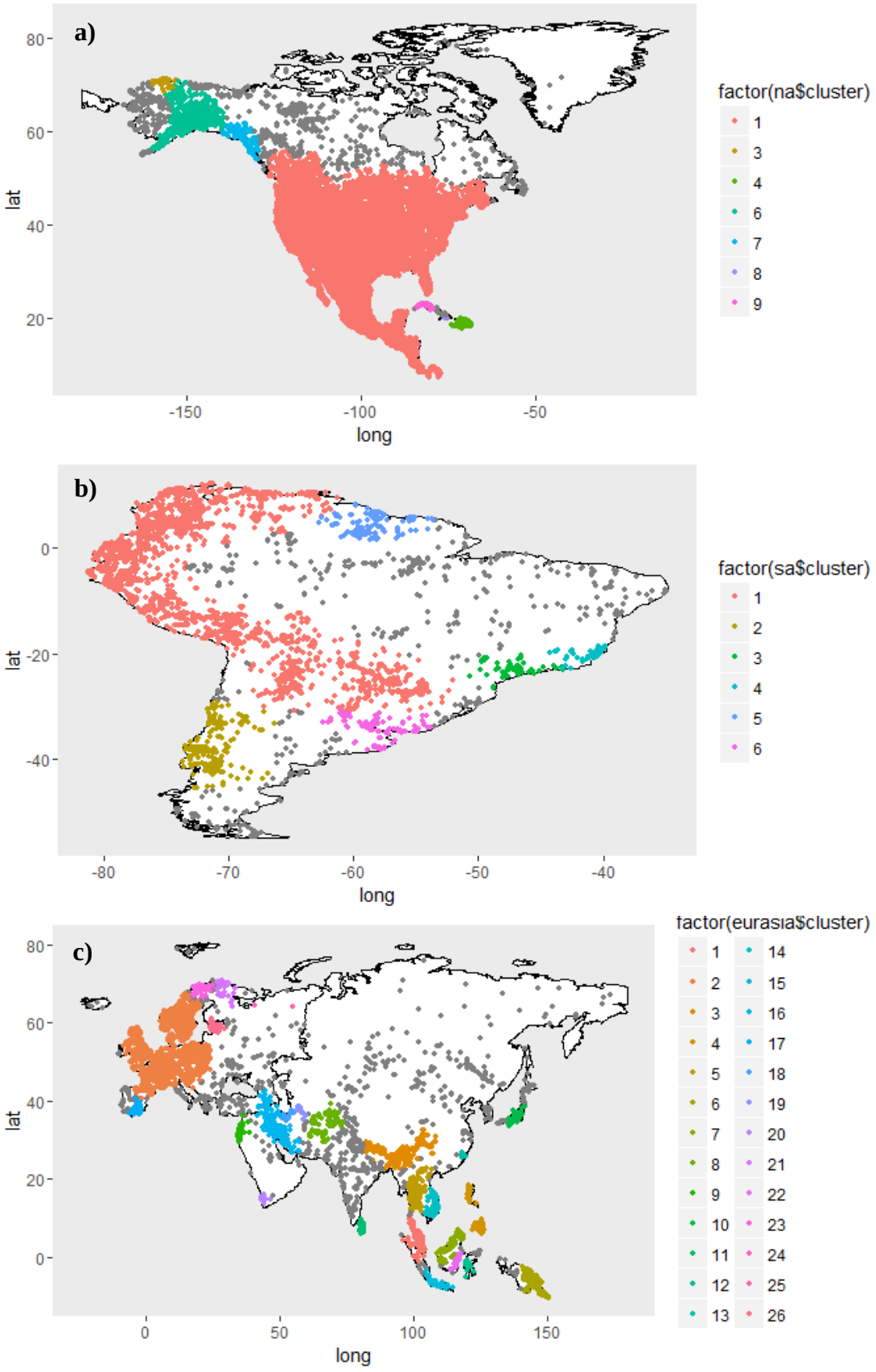
Supporting Information

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix 2.S1. Supplementary methods

Table 2.S1. Summary table with the different levels for reference ranges, records and algorithms used in the simulations

Reference ranges	**Symbol**	**Levels**
Shape	Shape	Circle, half bagel, star
Fragments	N_frag	1, 3 equal, 3 different
Recods		**Levels**
Size	N_sample	10, 25, 50, 100, 250, 500, 1000
Distribution	Sampling_method	Random, uniform, clustered
Errors	errors	0, 5, 10, 20%
Algorithm		**Parameters**
MCP	perc	100, 95, 90, 85, 80
k-Local Convex Hull	k	10, 15, 20, 25, 30
r-Local Convex Hull	r	2, 2.2, 2.4, 2.8, 3
a-Local Convex Hull	a	5, 5.5, 6, 6.5, 7
Kernel	thresh	1, 5, 10, 15, 20


a)
80
60
40
20
lat
-150
-100
-50
long
factor(na$cluster)
1
3
4
6
7
8
9
b)
0
-20
-40
lat
-80
-70
-60
-50
-40
long
factor(sa$cluster)
1
2
3
4
5
6
c)
80
60
40
20
0
lat
0
50
100
150
long
factor(eurasia$cluster)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

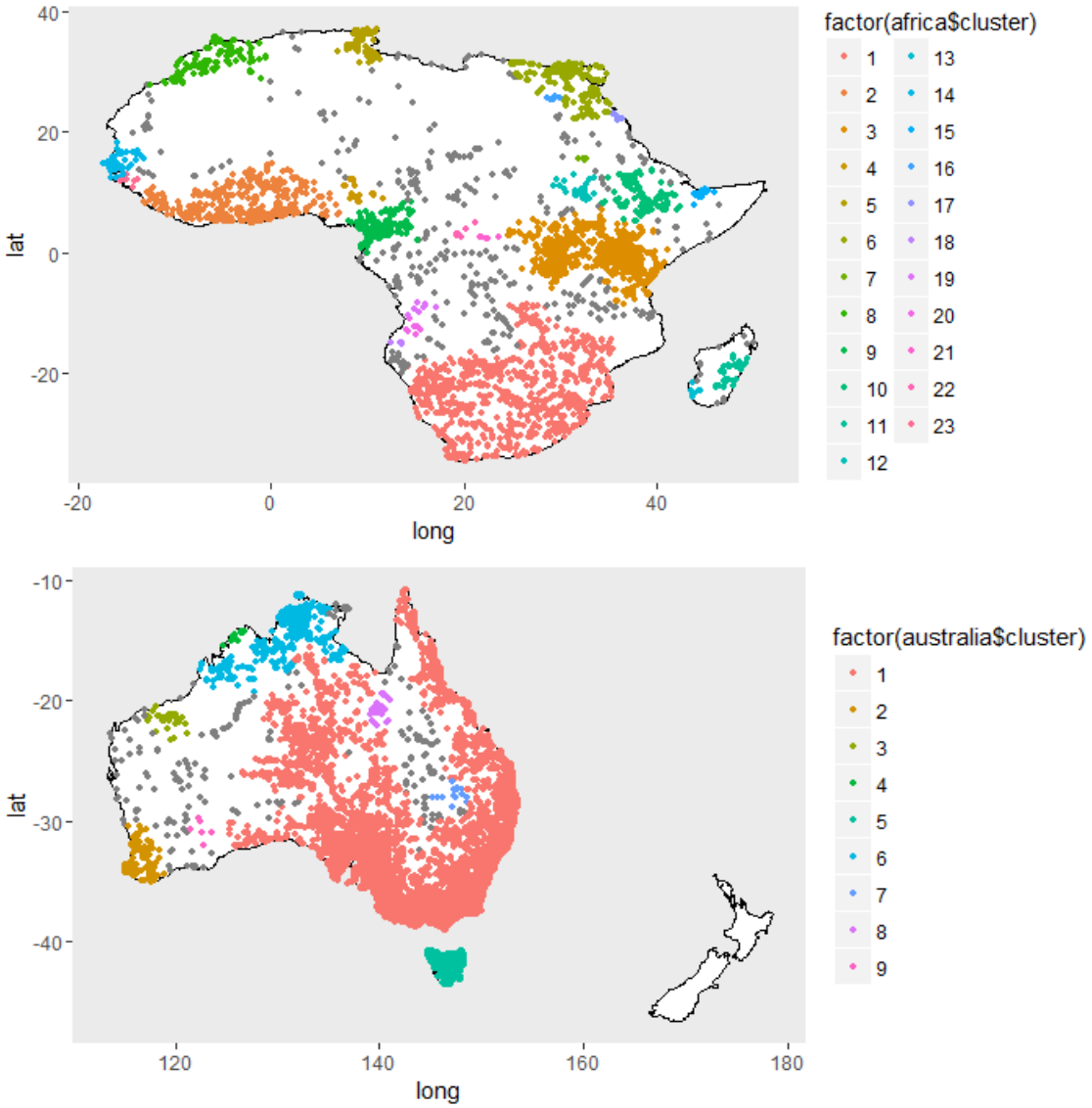


Figure 2.S1. Graphical representation of the spatial layout, clusters and number of clusters per continent: a) North America, b) South America, c) Eurasia, d) Africa and, e) Australia.

Appendix 2.S2. Supplementary results

Table 2.S2. Results of the generalized linear models with beta distribution for explaining the sensibility and type I rate error and with log-normal distribution for explaining the predicted fragments proportion respect to sample size with three spatial distributions (uniform, random and biased spatial distribution) and of the spatial errors when the distribution of the records is biased. We report the coefficient estimated and standard error (SE) for intercept, log (N_sample), log (Error) and types of shape included in the models.

Model	Intercept (Circle)	Circle dif	Circle equal	Half bagel	Half bagel dif	Half bagel equal	Star	Star dif	Star equal	Log (N sample)	Log (Error)
Sensitivity (uniform spatial distribution)											
MCP	-0.71 (±0.02)	0.14 (±0.02)	0.10 (±0.02)	-0.01 (±0.02)	0.12 (±0.02)	0.08 (±0.02)	0.47 (±0.02)	0.34 (±0.02)	0.27 (±0.02)	0.04 (±0.02)	-
KDE	0.34 (±0.02)	0.04 (±0.02)	-0.14 (±0.02)	-0.12 (±0.02)	-0.12 (±0.02)	-0.19 (±0.02)	0.38 (±0.02)	0.28 (±0.02)	0.25 (±0.02)	0.27 (±0.01)	-
k-LoCoH	-0.14 (±0.02)	-0.16 (±0.02)	-0.39 (±0.02)	-0.08 (±0.02)	-0.20 (±0.02)	-0.39 (±0.01)	0.54 (±0.02)	0.37 (±0.02)	0.23 (±0.02)	0.42 (±0.01)	-
r-LoCoH	-1.83 (±0.02)	-0.36 (±0.02)	-0.43 (±0.02)	-0.06 (±0.02)	-0.16 (±0.02)	-0.53 (±0.02)	1.01 (±0.03)	0.37 (±0.02)	0.24 (±0.02)	0.80 (±0.01)	-
a-LoCoH	-1.40 (±0.01)	-0.69 (±0.02)	-0.86 (±0.01)	-0.21 (±0.02)	-0.79 (±0.01)	-0.88 (±0.01)	-0.11 (±0.02)	-0.46 (±0.01)	-0.53 (±0.01)	0.71 (±0.01)	-
Type I error rate (uniform spatial distribution)											
MCP	-0.18 (±0.01)	0.20 (±0.01)	0.22 (±0.01)	-1.57 (±0.01)	0.22 (±0.01)	0.54 (±0.01)	-0.94 (±0.01)	0.39 (±0.01)	0.49 (±0.01)	0.05 (±0.01)	-
KDE	-0.50 (±0.07)	1.40 (±0.07)	1.24 (±0.07)	0.97 (±0.07)	2.40 (±0.06)	2.68 (±0.06)	2.21 (±0.06)	2.97 (±0.06)	3.02 (±0.06)	-0.66 (±0.01)	-
k-LoCoH	3.36 (±0.04)	-0.28 (±0.03)	-0.26 (±0.05)	-2.22 (±0.05)	-0.34 (±0.04)	-0.62 (±0.04)	-1.18 (±0.04)	0.16 (±0.04)	0.15 (±0.04)	-0.83 (±0.01)	-
r-LoCoH	-1.79 (±0.03)	0.19 (±0.03)	0.18 (±0.03)	-0.83 (±0.03)	0.57 (±0.02)	0.23 (±0.03)	0.51 (±0.02)	0.98 (±0.02)	1.01 (±0.02)	0.21 (±0.01)	-
a-LoCoH	1.02 (±0.25)	-0.29 (±0.12)	-0.32 (±0.12)	0.21 (±0.13)	-0.45 (±0.12)	-0.70 (±0.12)	0.25 (±0.13)	-0.07 (±0.13)	-0.26 (±0.12)	0.01 (±0.01)	-
Predicted fragments proportion (uniform spatial distribution)											
KDE	-0.60 (±0.02)	-0.51 (±0.02)	-0.18 (±0.01)	-0.01 (±0.01)	-0.48 (±0.01)	-0.45 (±0.01)	-0.02 (±0.01)	-0.52 (±0.01)	-0.64 (±0.02)	0.12 (±0.01)	-
k-LoCoH	0.15 (±0.01)	-0.55 (±0.01)	-0.52 (±0.01)	0.00 (±0.01)	-0.92 (±0.01)	-0.60 (±0.01)	0.11 (±0.01)	-0.94 (±0.01)	-0.94 (±0.01)	-0.03 (±0.01)	-

r-LoCoH	-0.12 (±0.01)	-0.08 (±0.01)	-0.01 (±0.01)	-0.01 (±0.01)	-0.07 (±0.01)	-0.30 (±0.01)	-0.01 (±0.01)	-0.11 (±0.01)	-0.16 (±0.01)	0.02 (±0.01)	
a-LoCoH	-0.60 (±0.02)	-0.51 (±0.02)	-0.18 (±0.01)	-0.01 (±0.01)	-0.48 (±0.01)	-0.45 (±0.01)	-0.02 (±0.01)	-0.52 (±0.01)	-0.64 (±0.02)	0.12 (±0.01)	-
Sensitivity (random spatial distribution)											
MCP	-1.34 (±0.02)	0.14 (±0.02)	0.11 (±0.02)	0.11 (±0.02)	0.29 (±0.02)	0.13 (±0.02)	0.44 (±0.02)	0.34 (±0.02)	0.25 (±0.02)	0.49 (±0.01)	-
KDE	0.47 (±0.04)	-0.04 (±0.03)	-0.05 (±0.03)	0.02 (±0.03)	0.01 (±0.03)	-0.05 (±0.03)	0.11 (±0.03)	0.10 (±0.03)	0.07 (±0.03)	0.18 (±0.01)	-
k-LoCoH	-0.72 (±0.03)	-0.17 (±0.02)	-0.35 (±0.02)	-0.02 (±0.02)	-0.12 (±0.02)	-0.30 (±0.02)	0.40 (±0.02)	0.21 (±0.02)	0.09 (±0.02)	0.48 (±0.01)	-
r-LoCoH	-2.73 (±0.02)	-0.46 (±0.02)	-0.54 (±0.01)	0.13 (±0.02)	-0.10 (±0.02)	-0.37 (±0.02)	0.86 (±0.02)	0.16 (±0.02)	0.02 (±0.02)	0.96 (±0.01)	-
a-LoCoH	-2.28 (±0.02)	-0.59 (±0.02)	-0.78 (±0.02)	-0.22 (±0.02)	-0.69 (±0.02)	-0.82 (±0.02)	-0.20 (±0.02)	-0.50 (±0.02)	-0.60 (±0.02)	0.81 (±0.01)	-
Type I error rate (random spatial distribution)											
MCP	-0.28 (±0.01)	-1.51 (±0.01)	0.28 (±0.01)	-1.49 (±0.01)	0.20 (±0.01)	0.57 (±0.01)	-0.99 (±0.01)	0.40 (±0.01)	0.53 (±0.01)	0.07 (±0.01)	-
KDE	0.16 (±0.03)	1.80 (±0.03)	1.78 (±0.03)	0.96 (±0.03)	2.08 (±0.03)	2.39 (±0.03)	2.06 (±0.03)	2.81 (±0.03)	2.94 (±0.03)	-0.68 (±0.01)	-
k-LoCoH	3.17 (±0.05)	-2.86 (±0.04)	-0.21 (±0.05)	-2.35 (±0.05)	-0.33 (±0.04)	-0.52 (±0.04)	-1.31 (±0.04)	0.17 (±0.04)	0.13 (±0.04)	-0.78 (±0.01)	-
r-LoCoH	-1.89 (±0.03)	-1.03 (±0.02)	0.13 (±0.03)	-0.83 (±0.03)	0.52 (±0.02)	0.37 (±0.02)	0.42 (±0.02)	0.90 (±0.02)	0.99 (±0.02)	0.23 (±0.01)	-
a-LoCoH	-0.36 (±0.05)	-1.42 (±0.05)	0.11 (±0.10)	-1.69 (±0.06)	-0.30 (±0.04)	-0.17 (±0.04)	-0.83 (±0.05)	0.23 (±0.04)	0.36 (±0.04)	-0.39 (±0.01)	-
Predicted fragments proportion (random spatial distribution)											
KDE	-2.01 (±0.04)	-0.73 (±0.02)	-0.86 (±0.03)	-0.05 (±0.01)	-0.80 (±0.02)	-0.96 (±0.03)	-0.03 (±0.01)	-0.96 (±0.03)	-0.95 (±0.03)	0.56 (±0.01)	-
k-LoCoH	-0.75 (±0.04)	-0.61 (±0.04)	-0.36 (±0.03)	-0.06 (±0.03)	-0.61 (±0.04)	-0.68 (±0.04)	-0.06 (±0.03)	-0.66 (±0.04)	-0.85 (±0.04)	0.19 (±0.01)	-
r-LoCoH	0.21 (±0.01)	-0.48 (±0.01)	-0.52 (±0.01)	0.00 (±0.01)	-0.88 (±0.01)	-0.65 (±0.01)	0.01 (±0.01)	-0.85 (±0.01)	-0.88 (±0.01)	-0.42 (±0.01)	-
a-LoCoH	0.14 (±0.01)	-0.25 (±0.01)	-0.21 (±0.01)	-0.10 (±0.01)	-0.28 (±0.01)	-0.45 (±0.01)	-0.10 (±0.01)	-0.33 (±0.01)	-0.40 (±0.01)	0.02 (±0.01)	-
Sensitivity (biased spatial distribution)											
MCP	-2.47 (±0.03)	-0.31 (±0.03)	-0.47 (±0.02)	-0.04 (±0.03)	-0.24 (±0.02)	-0.61 (±0.02)	-0.16 (±0.02)	-0.41 (±0.02)	-0.66 (±0.02)	0.68 (±0.01)	-
KDE	-0.50 (±0.04)	-0.74 (±0.03)	-0.90 (±0.03)	-0.15 (±0.03)	-0.64 (±0.03)	-0.99 (±0.04)	-0.77 (±0.03)	-0.90 (±0.03)	-1.10 (±0.03)	0.30 (±0.01)	-
k-LoCoH	-0.14 (±0.05)	-1.05 (±0.04)	-1.24 (±0.04)	-0.25 (±0.04)	-1.03 (±0.04)	-1.47 (±0.04)	-1.01 (±0.04)	-1.42 (±0.04)	-1.73 (±0.04)	0.27 (±0.01)	-
r-LoCoH	-3.61 (±0.04)	-1.29 (±0.03)	-1.50 (±0.03)	-0.18 (±0.03)	-0.10 (±0.03)	-1.51 (±0.03)	-0.43 (±0.03)	-1.25 (±0.03)	-1.56 (±0.03)	1.08 (±0.01)	-
a-LoCoH	-2.52 (±0.03)	-1.51 (±0.02)	-1.67 (±0.02)	-0.48 (±0.02)	-1.57 (±0.02)	-2.02 (±0.02)	-1.55 (±0.02)	-2.11 (±0.02)	-2.34 (±0.02)	0.77 (±0.01)	-

Type I error rate (biased spatial distribution)

MCP	-0.47 (±0.02)	0.15 (±0.02)	0.32 (±0.02)	-1.48 (±0.02)	0.12 (±0.02)	0.52 (±0.029	-1.23 (±0.02)	0.25 (±0.02)	0.49 (±0.02)	0.10 (±0.01)	-
KDE	-0.94 (±0.03)	1.56 (±0.03)	1.69 (±0.03)	0.76 (±0.03)	1.81 (±0.03)	2.03 (±0.03)	1.56 (±0.03)	2.36 (±0.03)	2.50 (±0.03)	-0.32 (±0.019	-
k-LoCoH	3.03 (±0.06)	-0.07 (±0.05)	-0.08 (±0.06)	-2.12 (±0.06)	-0.54 (±0.05)	-0.62 (±0.04)	-1.58 (±0.05)	-0.14 (±0.05)	-0.19 (±0.05)	-0.74 (±0.01)	-
r-LoCoH	-2.33 (±0.05)	0.27 (±0.04)	0.12 (±0.04)	-0.77 (±0.04)	0.36 (±0.04)	0.20 (±0.04)	0.10 (±0.04)	0.51 (±0.04)	0.63 (±0.03)	0.31 (±0.01)	-
a-LoCoH	0.48 (±0.06)	-0.12 (±0.07)	0.11 (±0.08)	-1.41 (±0.07)	-0.48 (±0.06)	-0.39 (±0.06)	-0.92 (±0.06)	0.09 (±0.06)	0.08 (±0.06)	-0.47 (±0.01)	-

Predicted fragments proportion (biased spatial distribution)

KDE	-2.57 (±0.06)	-0.97 (±0.03)	-1.06 (±0.03)	-0.11 (±0.01)	-1.14 (±0.03)	-1.14 (±0.03)	-0.15 (±0.01)	-1.20 (±0.03)	-1.07 (±0.03)	0.79 (±0.01)	-
k-LoCoH	-3.57 (±0.11)	-0.36 (±0.05)	-0.34 (±0.05)	0.11 (±0.04)	-0.46 (±0.06)	-0.46 (±0.06)	0.49 (±0.04)	-0.18 (±0.05)	0.07 (±0.04)	0.78 (±0.02)	-
r-LoCoH	0.28 (±0.01)	-0.41 (±0.01)	-0.44 (±0.01)	0.01 (±0.01)	-0.71 (±0.02)	-0.55 (±0.01)	0.05 (±0.01)	-0.64 (±0.02)	-0.65 (±0.02)	-0.06 (±0.01)	-
a-LoCoH	-1.20 (±0.03)	-0.19 (±0.02)	-0.19 (±0.02)	0.15 (±0.02)	-0.32 (±0.02)	-0.36 (±0.03)	0.46 (±0.02)	-0.37 (±0.03)	-0.45 (±0.03)	0.34 (±0.01)	-

Sensitivity (spatial bias and errors)

MCP	-2.38 (±0.04)	-0.28 (±0.03)	-0.51 (±0.03)	-0.12 (±0.03)	-0.37 (±0.03)	-0.61 (±0.02)	-0.14 (±0.03)	-0.34 (±0.02)	-0.56 (±0.03)	0.54 (±0.01)	0.53 (±0.01)
KDE	-0.92 (±0.05)	-0.37 (±0.04)	-0.50 (±0.04)	-0.08 (±0.04)	-0.41 (±0.04)	-0.60 (±0.04)	-0.23 (±0.04)	-0.49 (±0.04)	-0.53 (±0.04)	0.32 (±0.01)	0.65 (±0.01)
k-LoCoH	-1.87 (±0.06)	-0.96 (±0.04)	-1.01 (±0.04)	-0.35 (±0.04)	-0.96 (±0.04)	-1.27 (±0.04)	-0.78 (±0.04)	-1.26 (±0.04)	-1.34 (±0.04)	0.47 (±0.01)	0.80 (±0.01)
r-LoCoH	-5.81 (±0.06)	-1.48 (±0.03)	-1.79 (±0.03)	-0.30 (±0.04)	-1.35 (±0.03)	-1.78 (±0.03)	-0.88 (±0.03)	-1.66 (±0.03)	-1.93 (±0.03)	1.55 (±0.02)	0.50 (±0.01)
a-LoCoH	-4.82 (±0.03)	-1.47 (±0.02)	-1.65 (±0.02)	-0.54 (±0.02)	-1.55 (±0.02)	-1.90 (±0.02)	-1.48 (±0.02)	-1.95 (±0.02)	-2.05 (±0.02)	1.09 (±0.01)	0.47 (±0.01)

Type I error rate (spatial bias and errors)

MCP	-3.95 (±0.04)	2.46 (±0.03)	2.62 (±0.03)	1.85 (±0.03)	2.53 (±0.03)	2.75 (±0.03)	1.85 (±0.03)	2.61 (±0.03)	2.73 (±0.03)	0.16 (±0.01)	0.47 (±0.01)
KDE	-1.05 (±0.04)	0.82 (±0.02)	0.81 (±0.02)	0.50 (±0.02)	0.99 (±0.02)	1.06 (±0.02)	0.93 (±0.02)	1.17 (±0.02)	1.23 (±0.02)	-0.15 (±0.01)	0.56 (±0.01)
k-LoCoH	-1.98 (±0.02)	0.41 (±0.01)	0.41 (±0.01)	0.13 (±0.01)	0.35 (±0.01)	0.49 (±0.01)	0.28 (±0.01)	0.58 (±0.01)	0.59 (±0.01)	0.28 (±0.01)	0.45 (±0.01)
r-LoCoH	-4.51 (±0.03)	0.24 (±0.02)	0.39 (±0.02)	0.26 (±0.02)	0.55 (±0.02)	0.58 (±0.02)	0.41 (±0.02)	0.66 (±0.02)	0.69 (±0.02)	0.60 (±0.01)	0.61 (±0.01)
a-LoCoH	-6.85 (±0.05)	0.45 (±0.03)	0.50 (±0.03)	0.16 (±0.03)	0.58 (±0.03)	0.74 (±0.03)	0.55 (±0.03)	0.94 (±0.03)	0.93 (±0.03)	0.73 (±0.01)	1.02 (±0.01)

Predicted fragments proportion (spatial bias and errors)

KDE	-3.40 (±0.05)	-0.88 (±0.02)	-0.82 (±0.02)	-0.02 (±0.01)	-0.94 (±0.02)	-0.95 (±0.03)	-0.03 (±0.01)	-0.91 (±0.02)	-0.93 (±0.03)	0.43 (±0.01)	0.61 (±0.01)

k-LoCoH	-3.55 (±0.07)	-0.25 (±0.03)	-0.26 (±0.03)	0.44 (±0.03)	-0.26 (±0.03)	-0.28 (±0.03)	0.83 (±0.02)	-0.20 (±0.03)	-0.09 (±0.03)	0.78 (±0.01)	-0.07 (±0.01)
r-LoCoH	0.49 (±0.02)	-0.84 (±0.01)	-0.80 (±0.01)	0.01 (±0.01)	-0.96 (±0.01)	-0.83 (±0.01)	0.03 (±0.01)	-0.90 (±0.01)	-0.86 (±0.01)	-0.07 (±0.01)	-0.01 (±0.01)
a-LoCoH	-2.80 (±0.03)	-0.57 (±0.01)	-0.54 (±0.01)	0.18 (±0.01)	-0.62 (±0.02)	-0.60 (±0.02)	0.45 (±0.01)	-0.68 (±0.01)	-0.72 (±0.02)	0.60 (±0.01)	0.25 (±0.01)

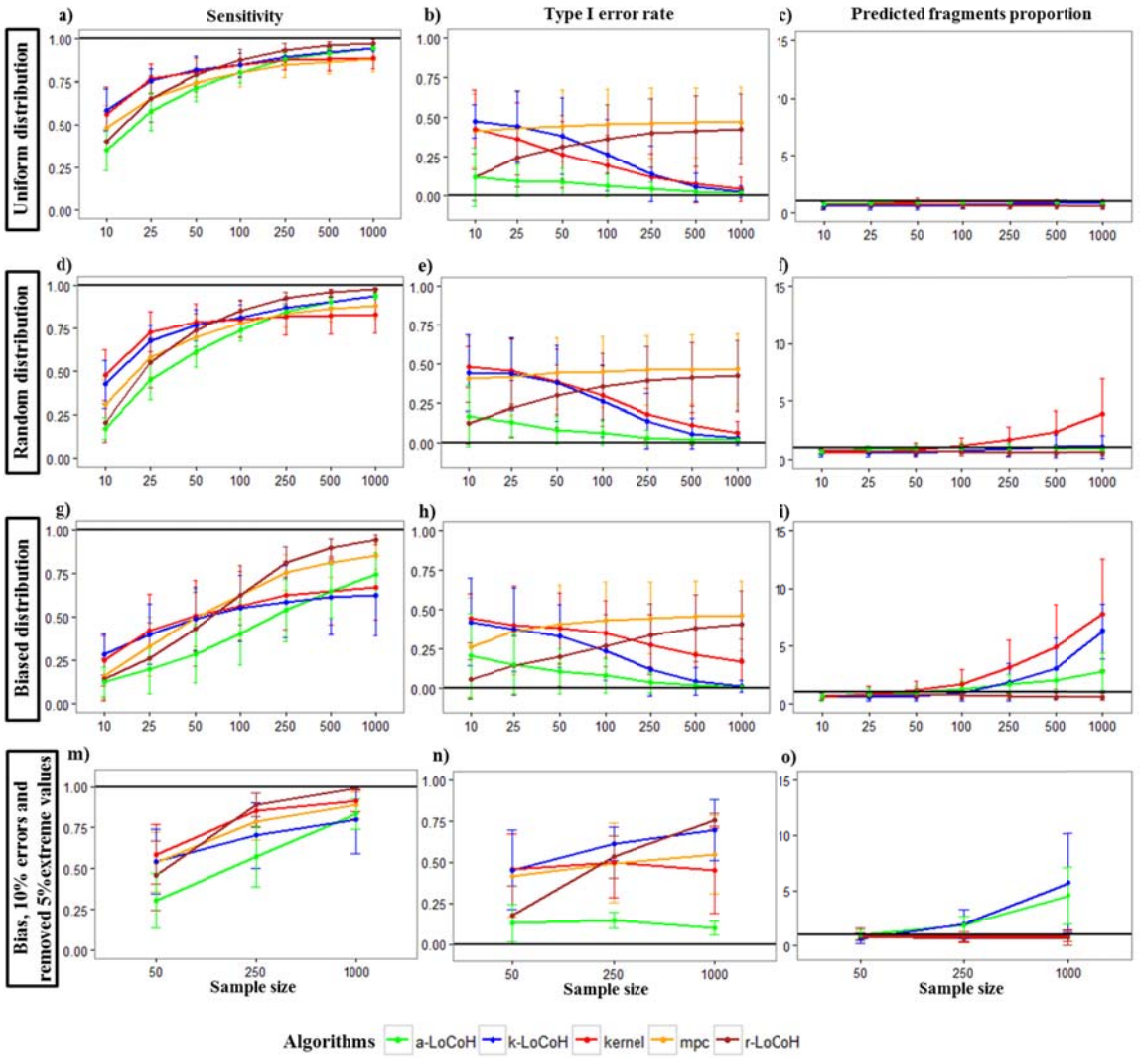


Figure 2.S2. Performance of the geographic algorithms when estimating sensitivity (a, d, g, j and m), type I error rate (b, e, h, k and n) and rate of predicted fragments (c, f, i, l and o) respect to the number of records available (sample size) when these data are uniformly (a, b, c), randomly (d, e, f), bias distributed (g, h, i), bias distributed and 10% errors (j, k, l) and bias distributed, 10% errors and removed 5% the most extreme values (m, n, o) . The black line represents the correct value describing the reference ranges.

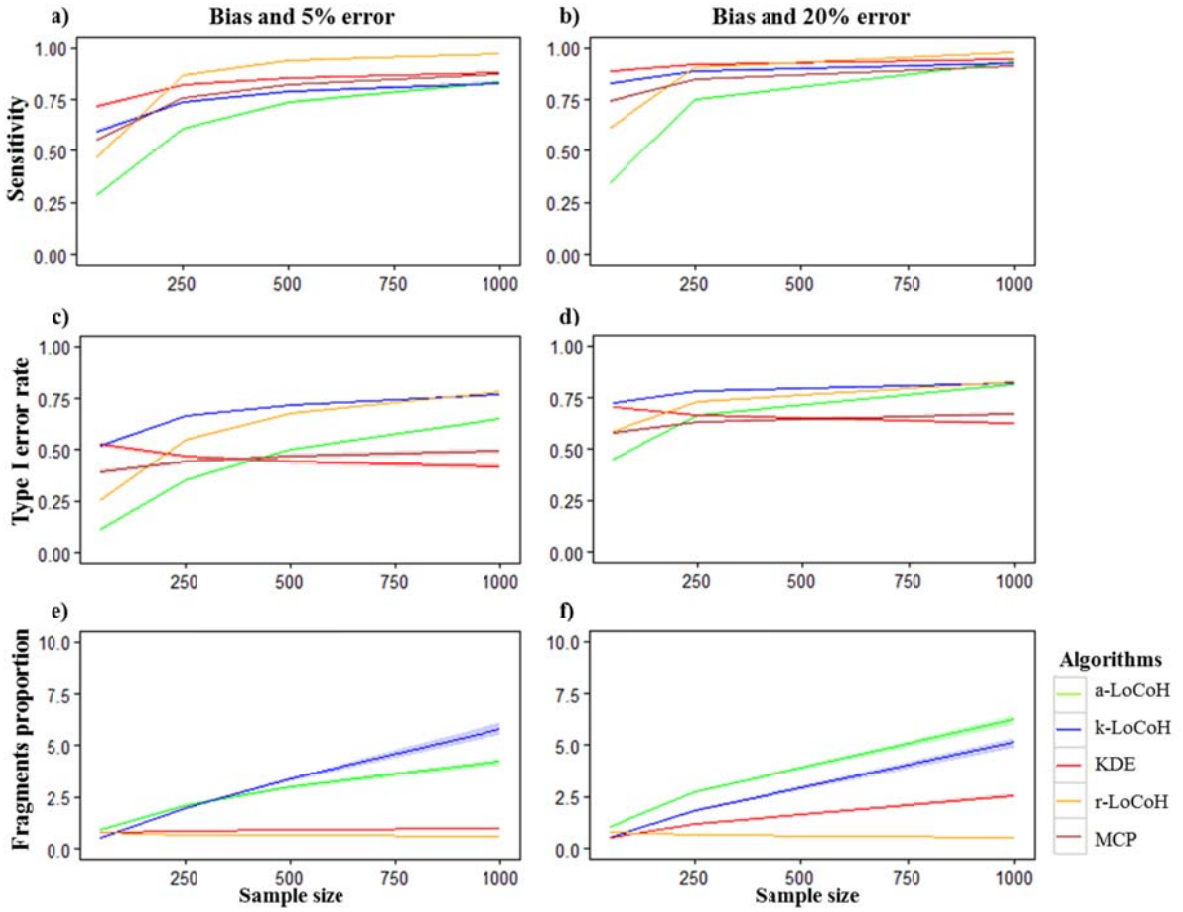


Figure 2.S3. Generalized linear models with beta distribution for explaining the sensitivity (a-b) and type I error rate (c-d) and, generalized linear models with log-normal distribution for explaining the predicted fragments proportion (e-f) respect to size sample with biased spatial distribution and error (5 and 20%) in database. The solid lines to color depict the mean relationship and the shaded areas depict 95% confidence intervals.

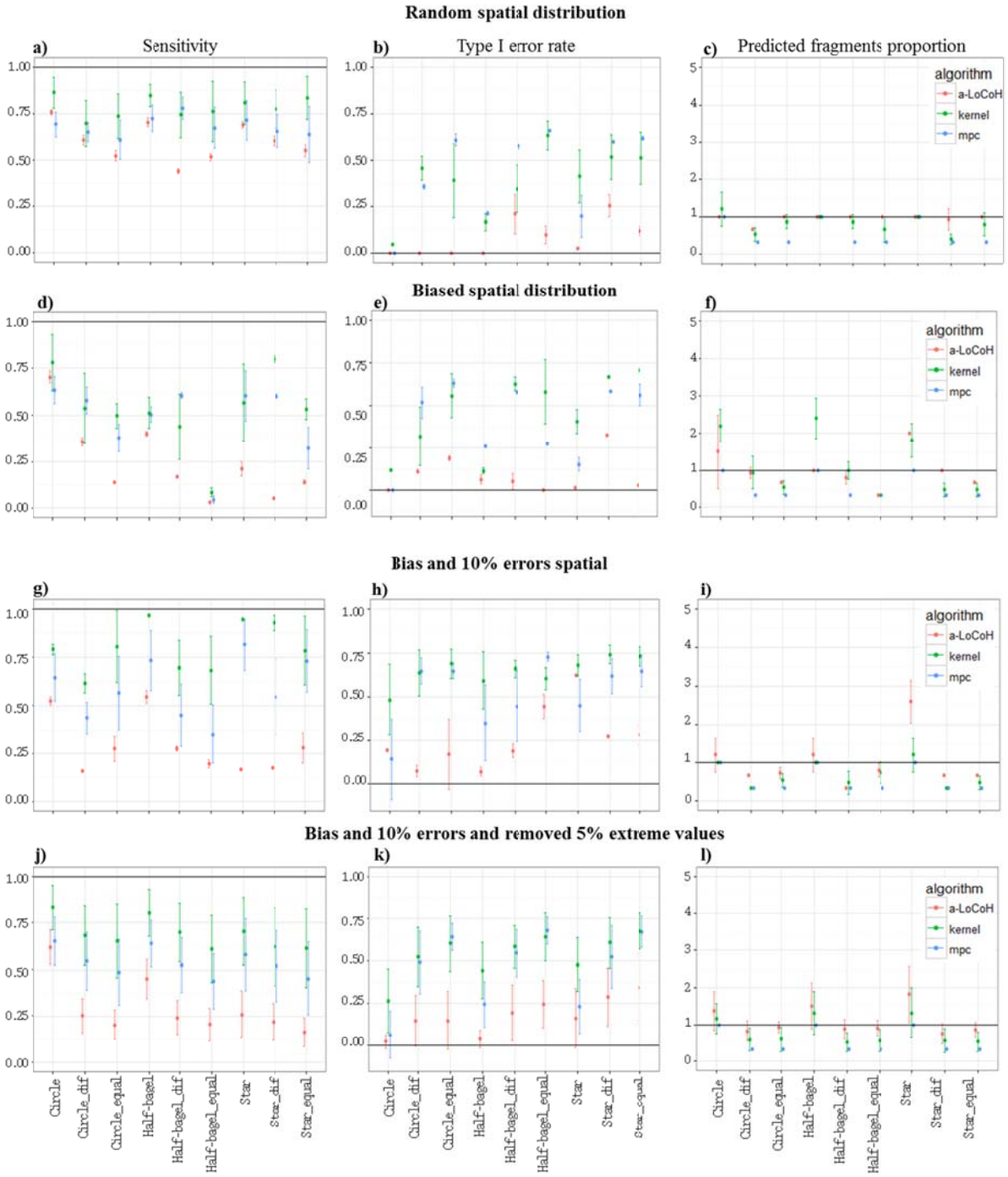


Figure 2.S4. Performance of the geographic algorithms when estimating sensitivity (a, d, g, j), type I error rate (b, e, h, k) and predicted fragments proportion (c, f, i, m). The samples were random (a, b, c) and bias (d-l) distributed with a sample size of 50 records and without errors spatial (a-f), 10% errors spatial (g-i) and 10% errors spatial and removed 5% the most extreme values (j-l). The black line represents the correct value describing the reference ranges.

CHAPTER III - Towards systematic species range maps

ABSTRACT. Range maps of thousands of species, compiled and made freely available by the International Union for Conservation of Nature (IUCN), are being increasingly used to support spatial conservation planning and in basic ecological research. However, the methodology used for building these maps is non-replicable, and the coarse nature of these maps makes them prone to commission and omission errors, calling into question their informative value. Here, we offer a systematic and easily replicable methodology to build species distribution ranges, which we compare with the already existing IUCN maps. Our results suggest that combining IUCN range maps with available georeferenced data in Global Biodiversity Information Facility (GBIF data) is a promising route to provide information on where the mapped distributions are reliable and where they are uncertain, to obtain a unified and easily repeatable methodology. The lack of information or availability of information in certain areas makes it difficult to implement systematic approaches to the construction of distribution range maps. We also disclose the priority sites where sampling effort should be increased. This is all the more urgent in the little-known hyper-diverse regions where decisions relevant to conservation must continue despite the scarcity of biodiversity data.

Key words: Conservation priority areas, Distribution ranges, GBIF data, geographic algorithms, IUCN range maps, omission and commission errors.

Ríos-Pena, L., Clavero, M., & Revilla, E. Towards systematic species range maps (*In prep).*

RESUMEN

Los mapas de áreas de distribución de miles de especies, compilados y distribuidos gratuitamente por la Unión Internacional para la Conservación de la Naturaleza (UICN), se utilizan cada vez más para apoyar la planificación de la conservación espacial y en la investigación ecológica básica. Sin embargo, la metodología utilizada para la construcción de estos mapas no es replicable, y la naturaleza gruesa de estos mapas los hace propensos a errores de comisión y omisión, poniendo en duda su valor informativo. Aquí, ofrecemos una metodología sistemática y fácilmente replicable para construir áreas de distribución de especies, que comparamos con los mapas ya existentes de la UICN. Nuestros resultados sugieren que la combinación de los mapas de áreas de distribución de la UICN con los datos georreferenciados disponibles en el Fondo Mundial para la Información sobre la Biodiversidad (GBIF) es una ruta prometedora para proporcionar información sobre dónde son confiables las distribuciones cartografiadas y dónde son inciertas, a fin de obtener una metodología unificada y fácilmente repetible. La falta de información o de disponibilidad de información en ciertas áreas dificulta la implementación de enfoques sistemáticos para la construcción de mapas de distribución. También revelamos los sitios prioritarios donde el esfuerzo de muestreo debe ser incrementado. Esto es aún más urgente en las regiones hiper-diversas poco conocidas donde las decisiones relevantes para la conservación deben continuar a pesar de la escasez de datos sobre biodiversidad.

Palabras clave: Áreas prioritarias de conservación, áreas de distribución, datos de GBIF, algoritmos geográficos, mapas de áreas de distribución de la UICN, omisiones y errores de comisión.

INTRODUCTION

Accurate mapping of species distribution ranges is a fundamental goal of modern biogeography, both for basic and applied purposes. Range maps provide information about the places where species occur, through a simplified abstraction of the complex spatio-temporal dynamics of the species' populations (Sexton et al., 2009). These maps constitute the baseline information for multiple proposes in fundamental and applied ecology and biogeography, including the identification of priority conservation actions (Wilson et al., 2007; Carwardine et al., 2008) and areas (Venter et al., 2014; Watson et al., 2014), or the description of biodiversity patterns (Orme et al., 2006; Di Marco and Santini, 2015; Faurby and Svenning, 2015), impacts of climate change (Lawler et al., 2010), or patterns in species distribution changes (Rodrigues et al., 2017).

The basic information units to build range maps are species records, which inform about the presence of a species in a given place and time. There are multiple methods that allow transforming sets of georeferenced records into species ranges (Burgman and Fox, 2003; Getz and Willmers, 2004; Getz et al., 2007) and in recent years there has been an important debate on the accuracy of these methodologies (Rondinini et al., 2006; Gaston and Fuller, 2009; Guisan et al., 2013). However, authors and institutions often overlook the availability of specific methodologies to generate range maps and build those maps on the base of expert knowledge or outputs of species distribution models (Herkt et al., 2017).

Species ranges provided by the Red List of the International Union for Conservation of Nature (IUCN) are the most comprehensive (taxonomically and geographically) global dataset on the distribution of species (IUCN, 2017). These maps were constructed following expert-knowledge approaches, through which experts apparently combined species records and their own knowledge to establish the boundaries, fragmentation, shape and size of the distribution area of each species (Rodrigues et al., 2006; Rodriguez et al., 2011). The expert knowledge approach is essentially an informal species distribution modeling, the procedures and outputs of which are non-repeatable (Johnson et al., 2012). The non-repeatability of IUCN maps is a common feature of other repositories of species distribution ranges and might introduce uncertainties to the uses and applications of those distribution maps. There is thus a need to construct distribution maps using systematic, repeatable approaches in which the exclusive source of information are temporally and spatially explicit species records. Such

approaches should take advantage of the exponential increase in the availability of species records, generated by both specialists and citizen scientists, and their storage and distribution through web-platforms, such as the Global Biodiversity Information Facility (GBIF) (Boitani et al., 2011; Jetz et al., 2012). These data repositories have been, however, shown to have several problems, including unequal and non-transparent quality of the original data or spatial and temporal biases due to the unequal sampling efforts (Graham et al., 2008; Yesson et al., 2007). Therefore, the election and implementation of methods for a systematic production of range maps should take into account the variability in the availability and quality of the original species records.

This work explores the options for a systematic and replicable generation of range maps that take into account different sources of variability in the quality of species records. We constructed species ranges applying two geographic algorithms, the adaptive Local Convex Hull (a-LoCoH) and the kernel Density Estimation (KDE), to the available georeferenced records in GBIF for mammalian carnivores (order Carnivora), a diverse, relatively well-known and widely distributed group of species. We compared the ranges generated for each carnivore species with those provided by IUCN and identified the concordant and discordant areas. We further used the whole dataset extracted from GBIF to generate a spatially-explicit estimate of sampling effort, aiming at discerning between commission and omission errors in discordant areas. Finally, we used the crossing of estimated species ranges and the distribution of sampling effort to identify areas across the world that require an effort of record gathering to allow an appropriate functioning of the systematic approaches for the generation of species distribution maps.

MATERIAL AND METHODS

Species distribution data

Two databases with geographic information on species distributions have been used in this study: i) geo-referenced species occurrence records (i.e. points) collected from GBIF; and ii) species distribution areas (i.e. polygons) obtained from IUCN. Occurrence records provide direct evidence that a particular species was present at a specific geographical point at a

certain point in time (Soberón and Peterson, 2004), while range maps intend to delimit the whole area where a species is known or assumed to occur (Graham and Hijmans, 2006).

We downloaded the 1642820 records of mammalian carnivores stored by GBIF (http://www.gbif.org) on October 6th 2017, selecting the following metadata: 1) scientific name, 2) year of registration, 3) geographical coordinates, 4) institution to which the data belong, and 5) basis of record (observation, literature, preserved specimen, fossil specimen, living specimen, human observation, machine observation, material sample or unknown). From this dataset, we excluded records when they: i) did not have geographic coordinates or they were (0, 0) (reduced to 1341335 records); ii) did not have a scientific name or were based on a fossil specimen (reduced to 1105494 records); iii) were placed in the sea (reduced to 728861 records); and iv) were coastal records referred to marine species and records located in Antarctica (reduced to 669914 records). Finally, we reduced all rows (i.e. records) with identical values in all fields to a single record (392845 records retained) and kept only the information for species that had at least 10 valid records. The final dataset retained after this filtering process contained 338770 records (Figure 1), which referred to 175 species, included in 86 genera in 14 families.

We downloaded the IUCN polygons where origin is coded as extant (resident and introduced) distribution ranges of mammalian carnivores for the 175 species selected in GBIF.

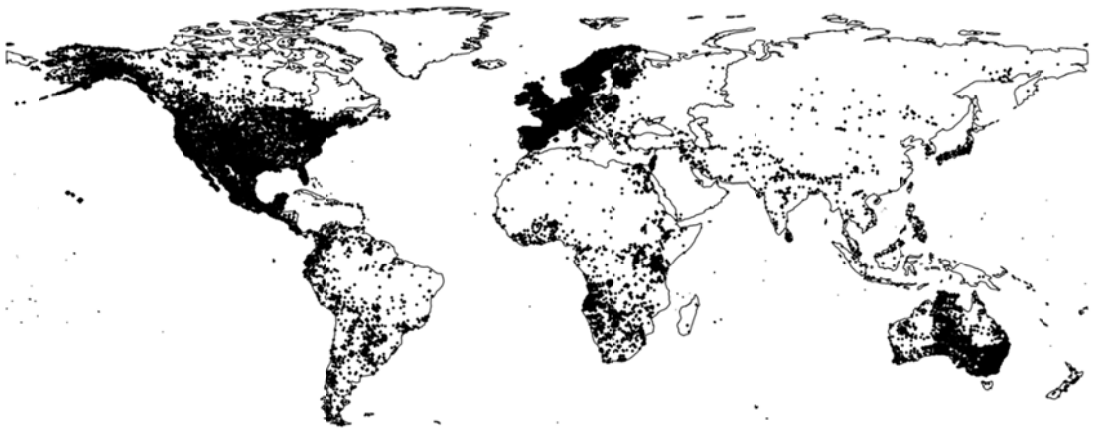


Figure 1: Global map of the distribution of point-occurrence records mobilized via GBIF after the filtering of database for the terrestrial Carnivorous Order (October, 2017). Antarctica has not been considered.

Species range maps by geographical algorithms

We used two geographic algorithms to build the distribution ranges: 1) Kernel Density Estimation (KDE); and 2) adaptive Local Convex Hull (a-LoCoH).

KDE algorithm is based on a kernel density function and is frequently used to estimate distribution ranges (Worton, 1989; Gitzen et al. 2006). It requires selecting a bandwidth parameter (h), which controls the degree of smoothing applied to the data and has a strong influence over the resulting estimates of range area. We applied the fixed kernel method, selecting h through the Maximum Likelihood Cross-Validation method (CVh, Habbema et al., 1974; Duin 1976). We used "npudens" function of the "np" package in R (Hayfield and Racine, 2007; R Core Team, 2017), which uses the method of Li and Racine (2003) to obtain the kernel density function. Range maps were constructed through a Thin Plate Splines (TPS) model, a technique based on providing a smooth interpolation of the data given in two or more dimensions (Donato and Belongie, 2002).

Adaptive (a)-LoCoH algorithm uses all records to generate the range within a variable circle around a root record such that the sum of the distances between the records and the root record is less than or equal to the parameter a (input parameter), which has to be specified. This method adjusts the radius of the circle that circumscribes each local convex hull, such that smaller convex hulls are placed where there is more concentration of records and larger convex hulls where the records are more distant from each other (Getz et al., 2007). We selected the value of a as the value of the maximum distance between occurrence points for each species and constructed species range with a-LoCoH using adehabitatHR package in R (Calenge, 2006) (Figure 2, a-d).

Methodology for the build of range maps

Out of the final set of records obtained from GBIF after filtering, and to obtain the species distribution ranges, we excluded 5% of the records farther from the total density of points by species. This exclusion was intended at avoiding in a systematic way the occurrence of geographically anomalous locations, such as erroneous locations or those of records for which the coordinates reported are those of the museum where the specimen is stored.

Georrefencing errors of this kind are difficult to identify systematically (i.e. not through an expert-dependent species per species filtering) from the original database.

For each species, we obtained 3 ranges: 1) Distribution range built with a-LoCoH (a-LoCoH range) algorithm, 2) distribution range built with KDE (KDE range) algorithm and, 3) distribution range obtained through IUCN Red List (IUCN range) (Table 1). We defined GEOGAL range as the union of both geographic algorithms (a-LoCoH and KDE ranges) and TOTAL RANGE as the union of the ranges of both geographical algorithm and IUCN range (GEOGAL and IUCN ranges). We calculated the concordant and discordant areas between both geographic algorithms and defined two levels in the overlap of ranges: (i) confident range and (ii) possible range. Confident range was designated as the concordant area between the ranges described through the a-LoCoH and KDE algorithms. Possible range, in contrast, describes the discordant area between the ranges resulting from the two geographical algorithms. In addition, we calculated the concordant and discordant areas between GEOGAL and IUCN ranges and defined three levels based on the overlap of ranges: (i) presence, (ii) possible presence and (iii) possible absence or lack of information. Presence was the concordant area between IUCN, KDE and a-LoCoH ranges. Possible presence described the concordant area between KDE or a-LoCoH range and IUCN range and, possible absence or lack of information was not concordant area between GEOGAL range and IUCN range. Here, there was not information of records of the focal species (Table 1, Figure 2 a-e).

Table 1. Variables used to quantify the proportion of concordant and discordant ranges between the geographical ranges constructed from GBIF geo-referenced records and the ranges obtained directly from IUCN.

Metrics	**Description**	**Symbol**
a-LoCoH range	Range generated with a-LoCoH algorithm	$A_{a-LoCoH}$
KDE range	Range generated with KDE algorithm	A_{KDE}
IUCN range	IUCN polygons where presence is coded as extant	A_{IUCN}
GEOGAL range	$A_{KDE} \cup A_{a-LoCoH}$	A_{GEOGAL}
TOTAL range	$A_{KDE} \cup A_{a-LoCoH} \cup A_{IUCN}$	A_T
Confident range	$A_{KDE} \cap A_{a-LoCoH}$	$A_{conGEOGAL}$
Possible range	$(A_{KDE} \cup A_{a-LoCoH}) - (A_{KDE} \cap A_{a-LoCoH})$	$A_{disGEOGAL}$

Presence	$A_{KDE} \cap A_{a-LoCoH} \cap A_{IUCN}$	$A_{conTOTAL}$
Possible presence	$(A_{KDE} \cap A_{IUCN})$ *or* $(A_{a-LoCoH} \cap A_{IUCN})$	$A_{ppTOTAL}$
Possible absence or lack of information	$(A_{KDE} \cup A_{a-LoCoH} \cup A_{IUCN}) - (A_{KDE} \cup A_{a-LoCoH}) \cap (A_{IUCN})$	$A_{paTOTAL}$
% concordant GEOGAL	$(A_{conGEOGAL}/A_{GEOGAL}) * 100$	$P_{conGEOGAL}$
% discordant KDE range	$(A_{KDE}/A_{GEOGAL}) * 100$	$P_{disGEOGAL_{KDE}}$
% discordant a-LoCoH range	$(A_{a-LoCoH}/A_{GEOGAL}) * 100$	$P_{disGEOGAL_{a-LoCoH}}$
% concordant TOTAL range	$((A_{GEOGAL} \cap A_{IUCN})/ A_T) * 100$	$P_{T_{conc}}$
% discordant GEOGAL range	$((A_{GEOGAL} \setminus A_{IUCN})/ A_T) * 100$	$P_{T_{disc-GEOGAL}}$
% discordant IUCN range	$((A_{IUCN} \setminus A_{GEOGAL})/ A_T) * 100$	$P_{T_{disc-IUCN}}$

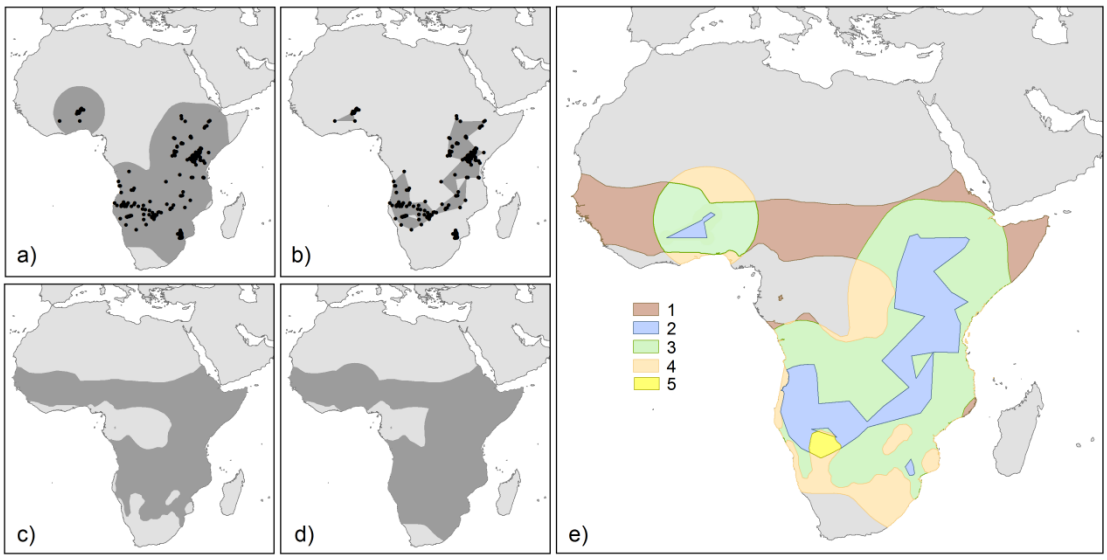


Figure 2: Distribution range maps of the spotted hyena (*Crocuta crocuta*). (a, b) Range maps generated with the KDE (a) and a-LoCoH (b) algorithms using 95% of the GBIF records for the species, which are also plotted in both panels. (c) IUCN range map for the spotted hyena. (d) TOTAL range resulting from the union of IUCN, KDE and a-LoCoH ranges. (f) Different

polygons representing the concordant and discordant areas resulting from the overlap of the IUCN, KDE and a-LoCoH ranges: 1, discordant area between A_{IUCN} and A_{GEOGAL} ranges; 2, concordant area between A_{IUCN}, A_{KDE} and $A_{a-LoCoH}$ ranges; 3, concordant area between A_{IUCN} and A_{KDE} ranges but discordant between them and the $A_{a-LoCoH}$ range; 4 discordant area between the A_{KDE} range and both A_{IUCN} and $A_{a-LoCoH}$ ranges and finally; and 5, discordant area between the A_{GEOGAL} range and the A_{IUCN}.

We tried to discern whether the discordances between the ranges defined through geographic algorithms and those contributed by IUCN may be due to geographic biases in the amount of information generating areas that lack species records (i.e. areas not identified through geographic algorithms) or due to inaccuracies of IUCN range maps (e.g. areas with records not included in IUCN ranges). To do so, for each focal species we selected all carnivore records, except those of the focal species, which fall within the TOTAL range (A_T) polygon, made a KDE with those records and cut it in three isopleths, generating areas that we interpreted as different sampling intensities (Figure 3): i) In the area defined by the 75% isopleth within A_T we assumed that there was enough accumulated information on other carnivore species to think that records of the focal species would had been generated if present; ii) in the area limited by the 75% and 90% isopleths (i.e. including between 10 and 25% of the records) we assumed a low density of information of the focal species, but the species could be present; iii) in the Area limited by the 90% isopleth (i.e. less than 10% of the records), we assumed that the scarcity of information did not allow a solid assessment of absence of the focal species. We highlight this as a priority area of information search.

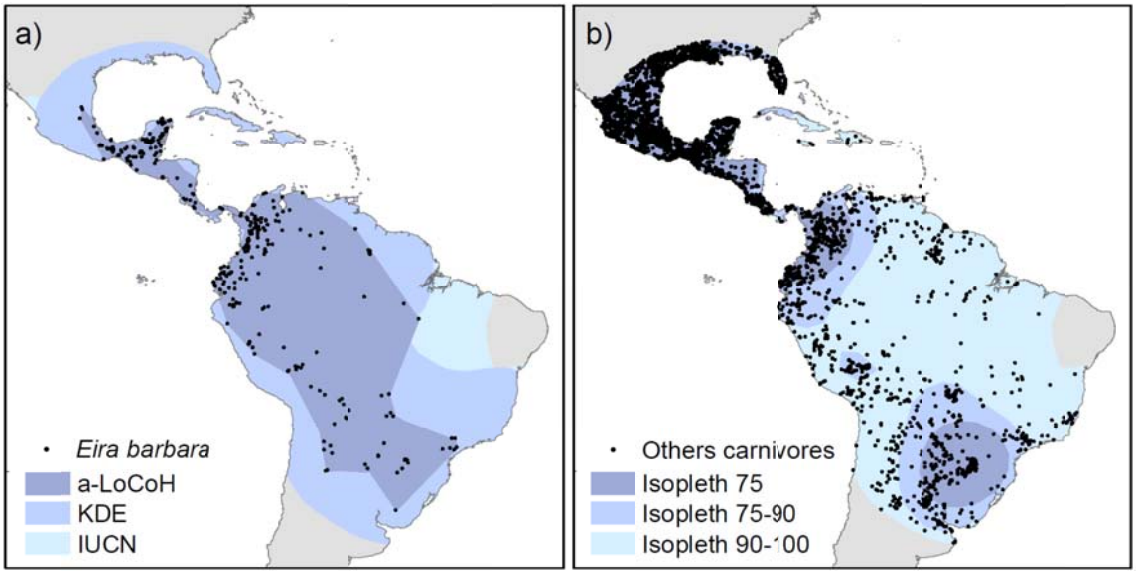


Figure 3: Graphic representation of distribution range of the tayra (*Eira barbara*) and of the information on other carnivores that is available within it. (a) Tayra range as resulting from the use of the two geographical algorithms and that provided by IUCN. (b) Representation of the TOTAL range of the tayra and all records of other carnivores reported by GBIF within it. The TOTAL range is divided in three polygons delimited by the 75% and 90% isopleths of the KDE applied to carnivore records except those of the tayra itself.

Global priorities for information gathering

After identifying areas poorly-sampled that hinder the systematic construction of species ranges, we summarized this information for all carnivores at the global level. Our aim was to highlighting regions that concentrate discrepancies in range descriptions due to an apparent lack of information, and that should consequently be considered as priority for gathering biodiversity records.). We used a 2°-square grid to count the number of KDE polygons that while being included in the TOTAL range of particular species contain less than 10% of the total density of records of carnivore species, excluding the focal one (isopleths 90-100%). This count reflects the number of species in need of reducing uncertainties in the definition of distribution ranges through the collection of biodiversity information. This metric already reflects the spatial variation of the information needs to build solid, systematic species distribution ranges, but can however be strongly influenced by the spatial variation in the species richness of carnivores. To take this possible relationship into account, we estimated the number of carnivore species present each 2°-square by counting the number of polygons

of IUCN ranges that intersected each square. We then calculated the ratio between the count of KDE polygons and polygons of UICN range, a metric that should inform on the proportion of carnivore species in need of better biodiversity information in each given square. However, the value of this metric does not have 1 as its upper limit because the numerator and denominator derived from different data sources (a combination of GBIF and IUCN and exclusively IUCN, respectively). The grids containing information on species distribution were grouped into 5 quantiles (0.10, 0.25, 0.50, 0.75, 0.90) (Figure 4).

RESULTS

Analyses of species distribution data

In our dataset, the mean record count per species with GBIF records was 1935.8 (SD= 5487.9, 1st quartile = 27.5, median=104.0, 3rd quartile = 642.5). North/Central America had the greatest number of species in our data set but the largest number of presence records was in Europe. Oceania had the fewest species and South America the lowest number of presence records (Table 2). After superimposing the GBIF records on the IUCN range by species, on average, 24.5% of GBIF records were outside their UICN ranges (SD = 27.97, 1st quartile = 4.98, median = 12.70, 3rd quartile = 33.60). Only 6 carnivore species had all GBIF records were within IUCN ranges, while all GBIF records were outside IUCN ranges for 3 species (*Genetta pardina*, *Genetta tigrina*, and *Herpestes javanicus*). The number of records falling outside the UICN ranges was strongly correlated with the total number of records of the species (Pearson's correlation on log-transformed data: r= 0. 83, df= 173, $p < 0.0001$, Figure 3.S1). Europe and North America had the highest proportion of records within the IUCN range, followed by Africa and Oceania. Range fit was lowest in Asia and South America.

Table 2: Summary of number of records of terrestrial carnivore species used for the analyses records by Continents and globally.

Record count	N species	N records	Min	Max	Mean	SD	Median
Global	175	338770	10	34970	1936.0	5487.9	104
North/Central	90	62955	1	6844	699.5	1352.4	54

America South	70	5618	1	970	80.3	155.4	15
America Europe	45	220249	1	33620	4894.0	7825.0	53
Africa	65	6214	1	1074	95.6	151.5	46
Asia	72	13364	1	3250	185.6	494.4	20
Oceania	9	30369	2	14310	3374.0	6028.0	36

IUCN and GBIF-based ranges

The comparison of the area of the ranges generated with a-LoCoH and KDE and IUCN showed that, on average, the smallest species ranges corresponded to those constructed with the a-LoCoH algorithm, followed by the IUCN ranges (59.6% larger than a-LoCoH) and the larger ranges were those generated with KDE algorithm (21.6% larger than IUCN range). The proportion of $A_{conGEOGAL}$ respect to a-LoCoH range was high (mean= 92.6%, median= 100.0, 3rd quartile = 100.0). In fact, the KDE range included totally the a-LoCoH range in 78.1% of cases. Contrastingly, the proportion of concordance range between a-LoCoH and KDE respect to KDE total range was low (mean= 33.5%, median= 30.0, 3rd quartile = 50.0). These differences in concordance seem to be generated by an overestimation of the range through KDE. The concordant range between a-LoCoH and KDE ranges is designated as the "confident range" of the focal species studied. On average, we had a 21.4% discordant range between methods ($A_{disGEOGAL}$) that was designated as a "possible presence".

There was a rather high agreement between the species ranges defined from GBIF records and those provided by IUCN. On average, 75.1% of the IUCN range was included in the (A_{GEOGAL}) (1st quartile= 27.3, median= 58.1, 3rd quartile= 76.3) and 70.4% of A_{GEOGAL} was included in IUCN range (1st quartile= 53.2, median= 80.1, 3rd quartile= 0.95). We found a positive association between the size of the A_{GEOGAL} and A_{IUCN} concordance range (Pearson's correlation: r= 0.55, df= 173, p < 0.0001, Figure 3.S2) and a strong positive association between the size A_{IUCN} range and A_{GEOGAL} concordance range (Pearson's correlation: r= 0.70, df= 173, p < 0.0001, Figure 3.S3). In the discordant areas ($A_{disTOTAL}$), in which we did not have presence information of the focal species ("range of absence or lack of information") we

obtained on average a 24.6% of discordant range (1st quartile = 4.5, median = 19.9, 3rd quartile = 46.8).

Global priority areas for information gathering

The results obtained after estimating the number of distribution ranges per grid with the IUCN ranges on a global scale showed on average 8.48% ranges (SD= 5.80, median= 8.0, 3rd quartile= 12.0) while for the areas that correspond to the isopleths 90-100 of KDE range per grid on average was 5.29% (SD= 5.82, median= 4.0, 3rd quartile= 8.0). On a continental scale, the highest average values of number of IUCN ranges per grid were Africa, followed by South America and Asia, and the least average value corresponded to Oceania. This circumstance also occurred when we calculate the number of polygons of isopleths 90-100 of the KDE range, although with lower average values in relation to those obtained in IUCN range (Table 3, Figure 4 a-b). The results obtained from the ratio between the isopleth 90-100 and the UICN ranges showed that of the 5,441 study grid cells, 5.9% could not be evaluated because they represent: 1) places where there were no carnivores naturally (as was the case in Australia or New Guinea), but where some species had been introduced, or 2) cells without information of both IUCN and KDE ranges. In 94.1% of the remaining grid cells, we obtained information on species distribution (Table 3, Figure 4c). The 0.10 quantile, which incorporated the zero value, represented 27.9% of well-sampled places. The grids with value 1 or higher were integrated into the same group and represented 24.1% of places with lack of information. At a continental level, North/Central America followed by Europe contained the highest percentages of well-sampled places. The continent with the highest percentage of grids lacking information was Africa, followed by Asia and South America (Figure 4c).

Table 3: Summary of the polygon count of IUCN and isopleths 90-100 corresponding to the KDE range by grids shown by continents and on a global scale.

Continent	Polygon	Mean	Median	3rd quartile	SD
Global	IUCN	8.48	8.0	12.0	5.80
	KDE 90-100	5.29	4.0	8.0	5.82
North/Central	IUCN	7.53	8.0	12.0	5.34

America	KDE 90-100	2.00	1.0	3.0	3.02
South America	IUCN	12.02	14.0	16.0	5.47
	KDE 90-100	10.87	12.0	15.0	5.49
Europe	IUCN	6.86	8.0	9.0	3.53
	KDE 90-100	4.13	4.0	7.0	4.08
Africa	IUCN	13.4	14.0	19.0	6.98
	KDE 90-100	12.6	11.0	19.0	7.06
Asia	IUCN	8.1	8.0	10.0	4.57
	KDE 90-100	4.7	4.0	7.0	3.97
Oceania	IUCN	0.82	1.0	1.0	0.38
	KDE 90-100	0.13	0.0	0.0	0.52

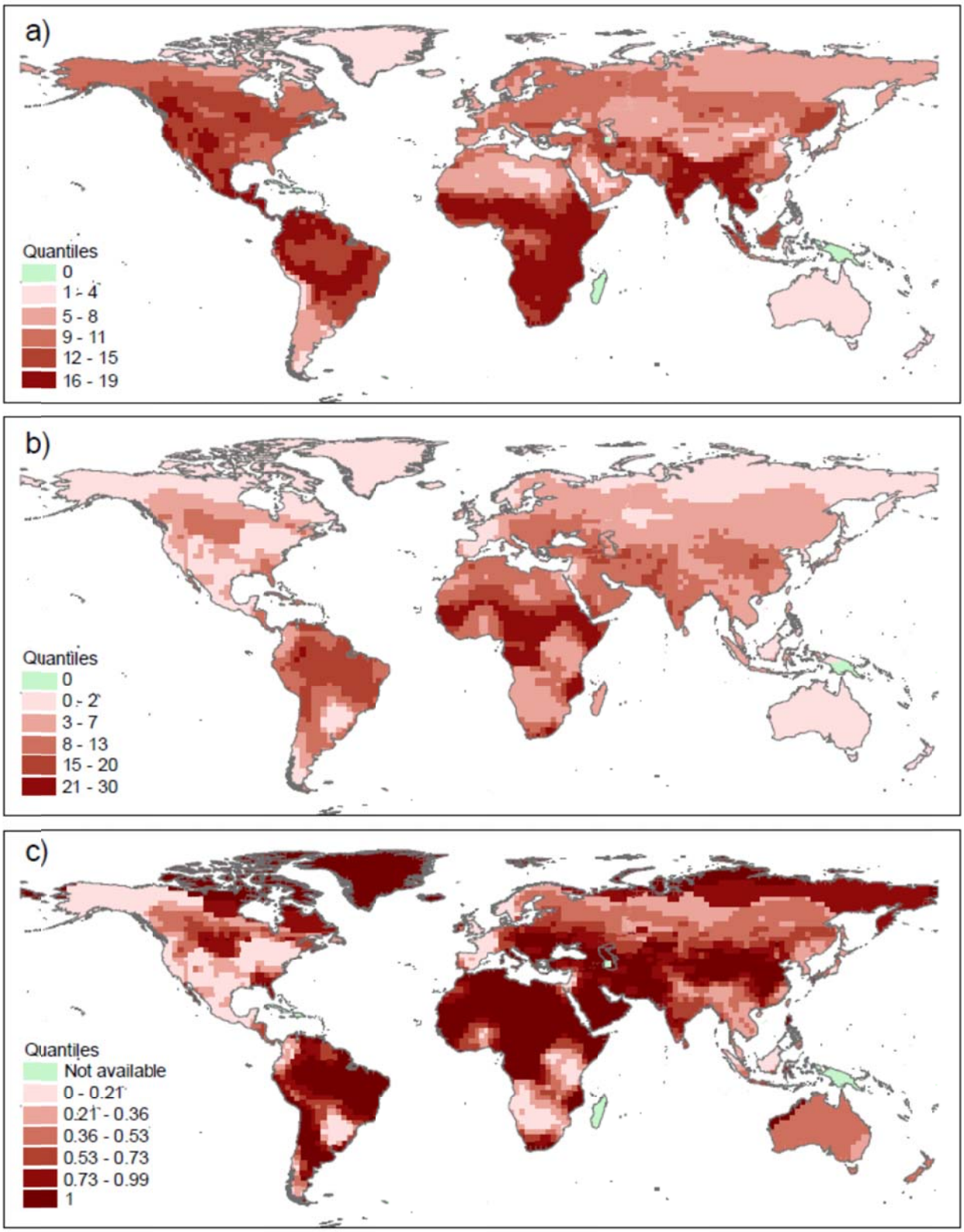


Figure 4: Global inequalities and gaps in the information on the distribution of terrestrial carnivores. (a) Count by cells of the number of IUCN polygons. (b) Count by cells of the number of polygons the isopleth 90-100 of KDE. (c) Quotient between the counts obtained in

(b) and in (a). Plots (a) and (b) have a legend that represents the real values in 5 categories (quantiles) obtained after the count and, the zero category, which represents the areas where there were no polygons. The graph (c) represents with its legend the real values in 5 levels (quantiles) obtained from the quotient and includes two more categories, not available, which represents the places without polygons and the one that represents the places with the highest information priority.

DISCUSSION

Our evaluation of range maps generated systematically for the carnivore order provides important information about possibilities of implementing such approaches for other biological groups. On the one hand, there was a rather high correspondence of the ranges obtained through geographic algorithms (KDE and a-LoCoH) and those provided by IUCN, suggesting that the systematically generated range maps provide a reasonable and useful description of the species distribution. On the other hand, there was a large geographic variation in the adjustment between maps generated with geographic algorithms and those of IUCN, which is parallel to the global bias observed for the research efforts and the ecological observations (Martin et al., 2012). The current study demonstrates the urgent need for increased investment to update, improve and complete the information sources, particularly in the especially rich areas of Asia, Africa and South America (Meyer et al., 2015). This improvement in the sources of information would allow us to obtain distribution ranges with a systematic approach. Its main advantages would lie in transparent analysis of more consistent data; being explicitly target-driven; and combining two forms of flexibility, namely opportunities to change data and targets, and opportunities to assess the options for achieving targets (Pressey and Cowling, 2001; Cowling et al., 2003). This is possible thanks to the expert knowledge based on approach on biodiversity persistence and pragmatic management and implementation issues not normally included in biodiversity feature-site data matrices (Dinerstein et al., 2000; Maddock and Samways, 2000).

This study demonstrates the existence of significant gaps in Global data. We have shown that there are geographical, temporal and taxonomic gaps in the quantity and quality of information mobilized through GBIF, which is very variable among species, even when, as

we do here with mammalians carnivores, a small and taxonomically coherent group of species is analyzed (Boakes et al., 2010; Amano et al., 2016). A substantial proportion of carnivorous mammals (18%) did not have geo-referenced records. The usefulness of GBIF as a repository of species records might be hindered by the lack of appropriate error-detecting filters, causing a significant occurrence of spatial errors (inaccurate geographical coordinates) in datasets (Yesson et al., 2007). Precision problems can be particularly problematic in old records, due to both positional errors and taxonomic changes (Boitani et al., 2011). For example, we found that 8% of carnivore georeferenced records were located in the sea, almost half of the records (41.4%) were exact repetitions of other records (i.e. all metadata fields were identical). Many of these errors can be identified and deleted through a species-per-species expert knowledge inspection of the available records, but such an inspection may be feasible only when the number of species is relatively low, becoming unworkable for larger datasets. Thus, in order to maintain the systematic nature of our approach and avoid the expert knowledge intervention, we excluded for each mammalian carnivore species the most geographically extreme 5% of the records (Ficetola et al., 2013; Hurlbert and Jetz, 2007).This procedure can reduce considerably positioning errors of the GBIF records and thus increase the quality the range maps constructed using them (Boitani et al., 2011).

In fact, true errors of omission (caused by ignorance of the presence of species or imprecise mappings) probably correspond to regions where the species are relatively rare, and therefore sites of low specific richness, not very different from the zero value obtained from the presence maps. The precision between species ranges reflected a global bias towards regions of the world better sampled with easier access to the information (Martin et al., 2012), and possibly certain taxonomic groups as well, that is, the coincidence between the range maps generated from GBIF and those from the IUCN is greater where there is more basic information about species records. This variability in the adjustment can also have a taxonomic reflection, so that the species or groups of species that accumulate more information will be those for which systematic distribution maps can be constructed more robustly. For example, the continent with the highest adjustment range (i.e., Europe) was also the continent with the most observations (Figure 1). For a more complete evaluation of the unoccupied areas (areas without information), additional analysis based on grid squares covering all continents was needed. Obvious data gaps in parts of Asia, Africa and South America highlight the need to identify the causes of the lack of information in these areas

(Meyer et al., 2015). Comprehension of the key factors that drive the biases in the availability of species records is important to prioritize the activities in data mobilization. In addition, bias drivers could be explicitly incorporated into the construction tools of species distribution ranges.

Identifying the best way or a consensual way to build distribution ranges maps is still a matter of debate. The horizon pursued in this study is to provide a transparent and easy to implement method for the construction of standardized and temporarily dynamic distribution maps that allow generating a more robust knowledge of the distribution patterns of biodiversity on Earth. However, gaps in accessible digital information on the distribution of species block the prospects of safeguarding biodiversity and ecosystem services. Therefore, we identify the regions with areas with lack of information (Figure 4) in order to make known the places where it is necessary to invest resources to improve the accuracy of the range presence maps. Finally, filling the current data gaps on a global scale will allow us to know all the places where the species are distributed, improve and update the species distribution maps and also help us when establishing strategies for biodiversity conservation little known of the Earth, so it is urgently needed a more effective use and mobilization of data and, a cultural change on the exchange of data.

ACKNOWLEDGEMENTS

We thank the IUCN Red List and Global Biodiversity Information Facility for making their database freely available online. This work was funded by the following grants and projects: Spanish Ministry of Economy, Industry and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R+D+I (SEV-2012-0262) and Agencia Estatal de Investigación from Ministry of Economy, Industry and Competitiveness, Spain with projects CGL2012-35931 and CGL2017-83045-R AEI/FEDER EU, co-financed with FEDER to E.R., R.B.M., M.G.S.

CHAPTER IV - The southern water vole as a case study: systematic vs. non-systematic data sources to build range maps

ABSTRACT. Range maps are among the most frequently used distribution data in biodiversity studies. As with any biological data, range maps can have some level of measurement error, but this error is rarely quantified. We assessed the error by comparing range maps obtained with systematic surveys and non-systematically by accumulating the available biodiversity information for the Southern water vole, *Arvicola sapidus*, in Spain. We built distribution maps using two geographic algorithms and provided explicit measures of spatial accuracy of ranges (omission and commission errors) that can allow us to reduce the risk of omitting undetected areas from range maps. The results of our study provided information on the nature of the distribution maps and the errors associated with the range maps that were explained through taxonomic errors, heterogeneous sampling effort and actual changes in the range due to the different sampling periods of the two datasets. This study provides precision measures that can be useful to understand the distributional changes over time and for future research using range maps as reference data. Finally, we emphasize the need to be cautious when using the available information sources to generate these range maps, particularly in areas with few data or with signs of heterogeneous spatial coverage.

Key words: Commission and omission errors, sampling effort, GBIF data, geographic algorithms, distribution ranges, taxonomic errors.

Ríos-Pena, L., Román, J., Clavero, M., & Revilla, E. The southern water vole as a case study: systematic vs. non-systematic data sources to build range maps (*In prep).*

RESUMEN

Los mapas de áreas de distribución se encuentran entre los datos de distribución más utilizados en los estudios de biodiversidad. Al igual que con cualquier dato biológico, los mapas de rango pueden tener algún nivel de error de medición, pero este error rara vez se cuantifica. Evaluamos el error asociado con *Arvicola sapidus* al comparar los mapas de áreas de distribución obtenidos con los enfoques de recolección de datos sistemáticos (encuesta) y no sistemáticos (acumulación de información de biodiversidad disponible). Construimos mapas de áreas de distribución utilizando dos algoritmos geográficos y proporcionamos medidas explícitas de precisión espacial de rangos (errores de omisión y comisión) que pueden permitirnos reducir el riesgo de omitir áreas no asignadas de los mapas de rango. Los resultados de nuestro estudio proporcionaron información sobre la naturaleza de los mapas de distribución y los errores asociados con los mapas de distribución que se explicaron a través de errores taxonómicos, esfuerzo de muestreo heterogéneo y cambios reales en el área debido a la depredación. Este estudio proporciona medidas de precisión que pueden ser útiles para comprender los cambios de distribución a los que están expuestas las especies a lo largo del tiempo y para futuras investigaciones utilizando mapas de rango como datos de referencia. Finalmente, enfatizamos la necesidad de tener cuidado al usar fuentes de información disponibles para generar estos mapas de rango, particularmente en áreas con pocos datos o con signos de cobertura espacial heterogénea.

Palabras clave: errores de comisión y omisión, esfuerzo de muestreo, algoritmos geográficos, áreas de distribución, errores taxonómicos

INTRODUCTION

In recent years, a growing number of studies have investigated patterns of biodiversity at broad spatial scales. These studies have helped us to understand the factors determining species distributions, richness and abundance, thus providing the information needed to set up conservation strategies (Lawler et al., 2010; Rondinini et al., 2011; Sandel et al., 2011; Hof et al., 2012). Since direct field sampling over large spatial scales is rarely feasible, as it requires significant resources and time, these broad-scale analyses must rely on compilations of data obtained from databases, faunistic atlases and geographical range maps. Unfortunately, our knowledge of biodiversity distribution is far from complete, and we have a limited knowledge of actual species distribution even for the best-studied taxa (Lomolino, 2004; Mokany and Ferrier, 2011; Ficetola et al., 2013). The strength of broad-scale biodiversity analyses and their usefulness for conservation purposes is directly related to the quality of the baseline data. Among species distribution data, errors are routinely quantified for some data types (point localities) but not for others (geographical range maps) (Rondinini et al., 2006; Rocchini et al., 2011). Geographical range maps encompass the areas where a species is thought to be found, and assume the species' presence inside the range and its absence outside. Even with this assumption, tests are needed to estimate the reliability of the range edge (Gaston, 2003; Rocchini et al., 2011).

Species range maps may be affected by multiple sources of error, such as incomplete information on some species or in some areas, limited spatial resolution, or errors when digitizing the distribution ranges, which may influence the output of analyses based on these maps (Hurlbert and Jetz, 2007; Foody, 2011; Cantú-Salazar and Gaston, 2013). Determining the level of accuracy of range maps can improve their usefulness in ecology, conservation and evolutionary biology, allowing for a better understanding of the strengths and limitations of analyses that use maps as baseline data (Hurlbert and Jetz, 2007; Rocchini et al., 2011).

The availability of information on species occurrences is currently growing at an exponential rate, but this information is most often collected in a non-systematic way and have several biases, mainly geographical, due to wrong spatial information, and taxonomic, when the species is incorrectly classified (Meyer et al., 2016a; Troudet et al., 2017). Several of these biases are overcome when the information on species occurrences is originated through systematic surveys, but this approach is much more costly in terms of effort and

money. It is thus necessary to compare the descriptions of the species distributions that can result from the use of systematic versus non-systematic information sources, in order to evaluate whether the effort required for systematic surveys is worthwhile.

Here we compare the range generated based on non-systematic and systematic data-collection strategies, using the southern water vole (Arvicola sapidus; henceforth water vole) in Spain as a study case. We used two sources of information on the distribution of the water vole: i) records contained in the Global Biodiversity Information Facility (GBIF) (i.e. non-systematic source); and ii) information from a stratified systematic survey, specifically designed to describe the status of the species in Spain (http://elrateador.blogspot.com.es/; Román, 2010). We constructed range maps based on the two information sources and using two different geographic algorithms. Finally, we investigated the factors that could be associated with omission and commission errors and that may allow us to reduce the risk of omitting unmapped areas of range maps. The results of our study provide insights into the nature of the maps and presence records, but also identify priorities for future research.

MATERIAL AND METHODS

Water vole data

We compared water vole range maps in Spain generated from two sets of records: i) a dataset extracted from the GBIF ("GBIF data"), which includes information from a wide, often unknown temporal window and is subjected to different biases and sources of error (Boakes et al., 2010; Troudet et al., 2017); and ii) a dataset based on a systematic survey designed specifically to detect the presence of the water vole ("survey data"), in which spatial biases are not present and some sources of error (e.g. taxonomic identification) can be assumed to be reduced (Román, 2010; Peralta et al., 2016). Records from both data sources were summarized using a grid of UTM 10×10 km cells, considering that a cell was positive for the water vole when it contained at least one record.

GBIF is a data portal established in 2001 to allow free and open access to global biodiversity data. It currently (April 2018) holds more than 977,000,000 species records, approximately half of the records localities originated from museum records and the rest from field studies (Edwards, 2004; Boitani et al., 2011). We downloaded all records stored by

GBIF for the water vole on January 12th 2018, resulting in 6,676 records from three countries (Spain, Portugal and France). We selected only the 5,567 records from peninsular Spain, in order to compare them with the data obtained from the systematic Spanish survey. We further excluded records that did not have geographical coordinates (x, y), had location errors (i.e. records at sea) and repeated geographical coordinates. We also excluded the observations of the systematic survey (see below) which are also available in GBIF. In total 1,134 records were eliminated. After these different filtering, the GBIF data contained 4,433 records that represented 1,968 positive UTM grid cells (Figure 1a, c).

The survey data were generated through a citizen-science based systematic survey, designed specifically to identify the presence of the species (Román, 2010). The survey used UTM 10×10 km cells as spatial sampling units and selected for sampling 1000 cells regularly distributed across peninsular Spain. Cells were assigned to local teams that were previously trained in the identification of water vole signs and tracks through different one-day workshops carried out across Spain. Surveyors visited 3 points within each cell, chosen due to the existence of *a priori* favorable habitat (i.e. rivers, wetlands or any type of aquatic system with abundant herbaceous vegetation, flood-prone meadows). Any given cell was considered positive if the water vole was detected in any of the 3 visited points, and negative otherwise. The survey involved visits to 2914 points, the water vole being detected in 1018 of them (Figure 1b), resulting in 587 positive UTM 10×10 km cells (Figure 1d).

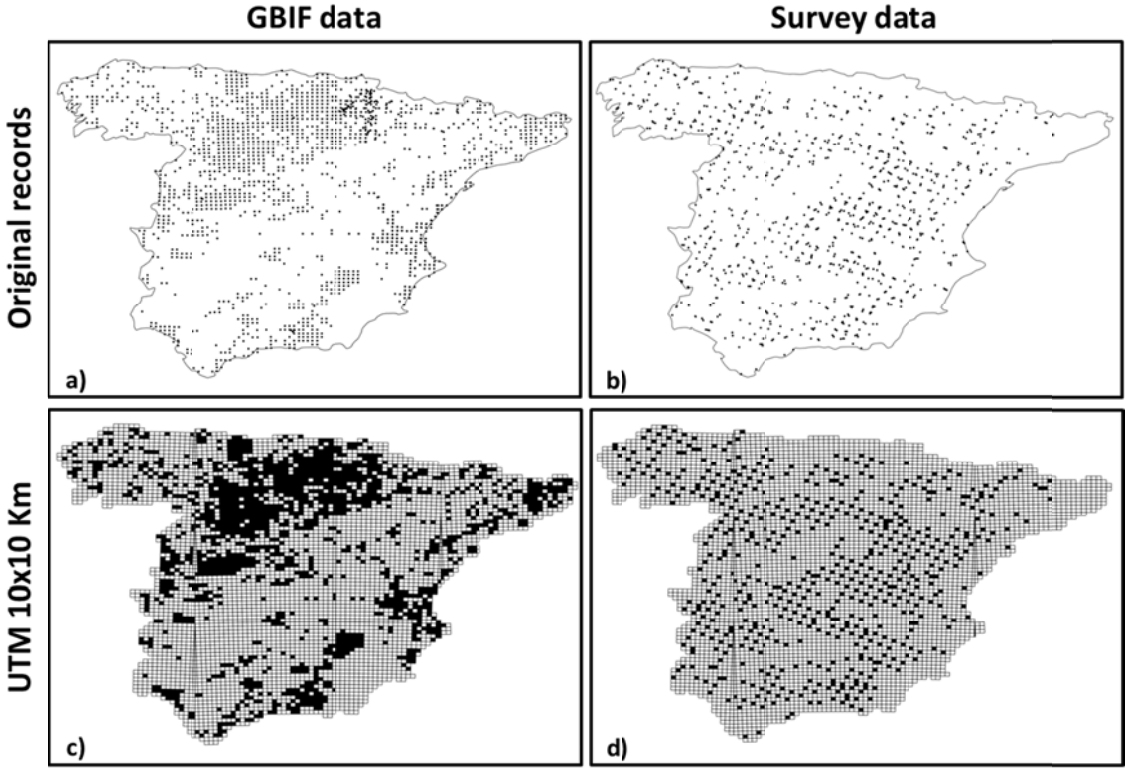


Figure 1: Records (above) and UTM 10×10 km cells (bellow) with presence of the water vole (*Arvicola sapidus*) using data obtained from GBIF (left) and the Spanish water vole survey (right).

Range maps

Range construction for both the GBIF and Survey datasets was done by using the centroid of the UTM 10×10 km cells containing at least one water vole record as the geographical point of reference. We used two methods for range building, the adaptive Local Convex Hull (a-LoCoH; Getz, et al., 2007) and Kernel Density Estimation (KDE; Worton, 1989) (Figure 2).

KDE algorithm requires the selection of a bandwidth parameter (h), a free parameter that affects the resulting range estimate. The bandwidth determines the relationship between the distance of a used location from an evaluation point and the contribution of the location to the density estimate at that point. We estimated the bandwidth through Maximum Likelihood Cross-Validation (CVh, Habbema et al., 1974; Horne and Garton, 2006). We applied "npudens" function of the *np* R package, which uses the method of Li and Racine (2003) to

obtain the kernel density function. We obtained range maps applying a thin plate splines (TPS) model (Donato and Belongie, 2002) to the weighted density of observations. TPS are a spline-based technique for data interpolation and smoothing. We created one threshold that allows cutting the records density map. The value assigned to TPS was 0%, i.e., the zero value included all observations within the estimated distribution range.

The a-LoCoH algorithm was developed by Getz et al. (2007) and is based on the construction of small convex hulls for each record and its neighbors. Convex hulls are merged together from smallest to largest and are ordered from the smallest to the largest. The a-LoCoH uses all points within a variable sphere around a root point such that the sum of the distances between these points and the root point is less than or equal to the a parameter, the value of which must be selected. We obtained the parameter using the Minimum Spurious Hole Covering (MSHC) rule (Getz and Wilmers, 2004; Getz et al., 2007).

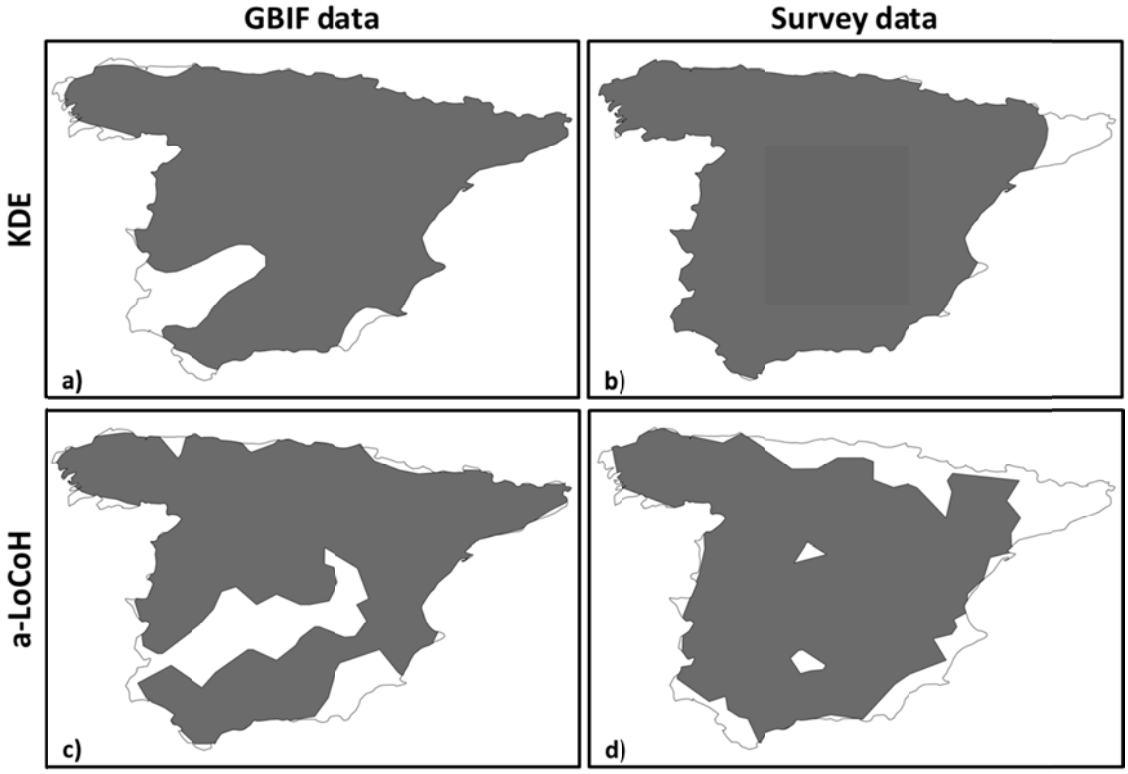


Figure 2: Range maps of *Arvicola sapidus* obtained using the KDE (above) and a-LoCoH (bellow) algorithms from GBIF (left) and survey (right) databases.

Commission and omission errors

We evaluated commission and omission errors in the representation of the water vole range using the maps resulting from the water vole survey as ground truth (Table 1). The survey data might contain errors, resulting in both false presences and false absences, but we assumed that the probability of these errors would be low due to the narrow focus of the water vole survey and the specific formation of the surveyors (Román, 2010; Peralta et al., 2016). We thus considered commission errors as those included in the water vole range when analyzing GBIF data but not when using survey data and omission errors as areas not included in the water vole range when using GBIF data but included when using the survey data (Table 1). Both commission and omission errors were quantified in terms of UTM cells within each one of these error categories. We hypothesized that commission errors could occur due to taxonomic errors and/or due to actual temporal changes in the presence of the species, while omission errors could result from spatial heterogeneity in sampling effort, which in turn, can be described using the total amount of information available for any species (i.e. sampling effort) and/or from taxonomic errors.

We analyzed the occurrence of commission and omission errors using binary logistic regression models and a link logit (McCullagh and Nelder, 1989). For each geographic method (KDE and a-LoCoH), we had the 10 Km UTM squares classified according to Table 1. In order to evaluate the commission errors, the response variable took value 1 when the grid cells belonged to commission errors and value 0 when they belonged to concordant area; and to evaluate the omission errors, the response variable took value 1 when the grids belonged to omission errors and value 0 when they belonged to concordant area. The explanatory variables used in the regression models to explain the discordance between distribution ranges are associated to multiple processes among which we have considered:

1) *Taxonomic errors*, which cause the distribution range to consider the presence of the species in an area where it is really absent and vice versa. We considered that rodent species that might be involved in the misidentification of the water vole, either through signs or direct observations were *Rattus norvegicus*, *Microtus cabrerae*, *Microtus agrestis* and *Arvicola terrestris/shermann*. We downloaded GBIF records for these species and counted the number of species present in each UTM cell, resulting in a count ranging from 0 to 4.

2) *Sampling effort*, we downloaded all the records of mammals in peninsular Spain and counted the number of records in 10 x 10 UTM squares, as an indicator of total sampling effort. The explanatory variable that represents the sampling effort is designated COUNT.

3) *Predation* by invasive exotic species, which could produce changes in the distribution range of the water vole as has been shown for *Arvicola terrestris* in other areas. Here we evaluate the specific case of *Neovison vison* (Aars et al., 2001), considering that the presence of at least one GBIF record in a grid corresponded to a presence grid and took value 1, when it did not record data, the grid was assigned the value 0.*vison*

The models were constructed and adjusted by estimation of maximum likelihood of the regression parameters using the "glm" function of R version 3.3.2 (R Development Core Team, 2017).

Table 1: contingency matrix to record commission and omission errors through the comparison of the water vole ranges constructed through information sources from GBIF and survey data.

		Survey data	
		Presence	**Absence**
GBIF data	**Presence**	Concordant	Commission errors
	Absence	Omission errors	Empty

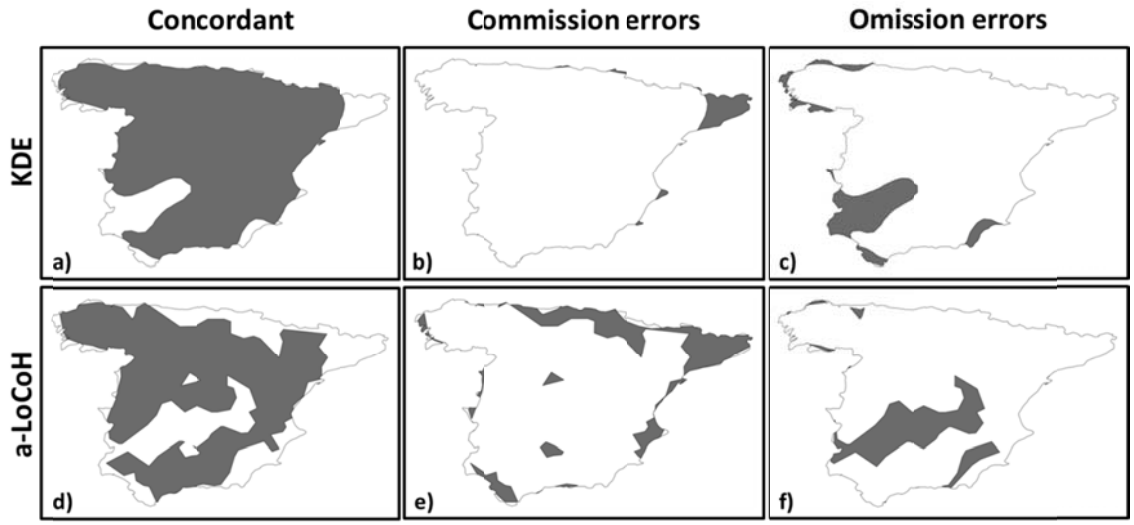


Figure 3: Concordant and discordant areas of water vole using the KDE (above) and a-LoCoH (bellow) algorithms with the GBIF and survey data. (a, d) Concordant ranges between GBIF and survey data, (b, e) Commission errors and, (c, f) omission error (see Table 1).

RESULTS

Range maps

The KDE method generated larger water vole range areas with respect to the a-LoCoH method for both data sources. Range areas constructed using the survey data were consistently larger than those obtained based on GBIF data (Table 2). The larger size of the range based on the survey data is surprising due to the smaller number of positive cells in this dataset (587) than in the GBIF data (1968), indicating a more heterogeneous spatial distribution of observations in the GBIF dataset. Both methods generated unfragmented ranges.

Table 2: Description of the distribution ranges resulting from the two geographic methods (KDE and a-LoCoH) by the two data source (GBIF and survey)

	Range area (Km^2)		Concordant area (Km^2)
	GBIF	**Survey**	
KDE	439,548.5	477,049.5	423,976.1
a-LoCoH	398,629.8	402,067.9	325,427.4

The percentage of concordant area between the ranges generated from the two databases and for both geographic methods was high. The percentage of coincident area when using the KDE algorithm was 96.4% and 88.9% for GBIF and survey data, respectively, while with the a-LoCoH algorithm these figures were 81.6% and 80.9% (Table 2).

Commission and omission errors

We evaluated the discordant area for both geographical methods (KDE and a-LoCoH respectively) and we obtained a size of 15,572.4 and 73,202.4 km^2 (3.6 and 18.4 % commission errors) for the GBIF data and, 53,073.4 and 76,640.5 km2 (18.4 and 19.1% omission errors) with the survey data (Figure 3). The number of discordant cells was much higher when using the a-LoCoH method than that obtained with the KDE method (Table 3). While omission errors were around three times more common than commission ones when ranges had been constructed using KDE, both types of errors were approximately as common when using the a-LoCoH (Table 3). Cells not included in any of the distribution ranges were excluded from the regression analyses (Table 3).

Table 3: Number of cells and percentage associated with the concordant and discordant areas between the two species information sources (GBIF and survey) for a-LoCoH and KDE algorithms. Percentages were calculated in reference to the 5314 UTM 10km cells in peninsular Spain.

	a-LoCoH	KDE
Concordant range	3687 (69.4%)	4573 (86.1%)
Commission	738 (13.9%)	184 (3.5%)
Omission	659 (12.4%)	548 (10.3%)
Empty	230 (4.3%)	9 (0.02%)

The models that best explained the probability of incurring in errors where those that included all the explanatory variables, bith for commission (i.e. taxonomic errors and predation) and omission (i.e. taxonomic errors and sampling effort) errors. When we evaluated the commission errors obtained with the KDE method, the models showed that the probability of incurring in commission errors increased in areas where there are many species of rodents that can be confused with the southern water vole. When we used the a-LoCoH method, the probability of incurring commission errors was related to the existence of species of rodents that can be confused with the southern water vole and predation by American mink, which is driving the decline and range contraction of the water vole. When we evaluated omission errors with the KDE method, the probability of incurring in omission errors increased in areas with little GBIF information (i.e. small sampling effort) and the existence of rodents easily confused with the study species. The omission errors generated with the a-LoCoH method increased mainly due to the lack of sampling effort and, to a lesser extent, by the existence of species easily confused with the southern water vole (Table 4).

Table 4: Results of regression analyses based on binomial logistic regression models (GLM) to explore the effect of taxonomic errors and predation on the discordant distribution ranges with GBIF data respect to survey data and the effect of taxonomic errors and sampling effort on the discordant distribution ranges of the survey data with respect to GBIF data. We report the coefficient estimate and its standard error [β (SE)] for all variables and the Akaike Information Criterion (AIC) of each model for comparison.

Model	β (SE)			**Model comparison**
	Predation	Sampling effort	Taxonomic error	AIC
COMMISSION – KDE (N= 4757)				
Predation	-15.47 (292.286)	-	-	1520.1
Taxonomic error	-	-	0.80 (0.097)***	1495.7
Pred + Tax	**-16.64 (471.094)**	**-**	**0.85 (0.096)*** **	**1447.6**
COMMISSION – a-LoCoH (N= 4425)				
Predation	-1.28 (0.196)***	-	-	3933.4
Taxonomic error	-	-	0.92 (0.058)***	3724.9
Pred + Tax	**-1.55 (0.200)*** **	**-**	**0.98 (0.058)*** **	**3638.8**
OMISSION – KDE (N= 5121)				
Sampling effort	-	-0.04 (0.003)***	-	3146.2
Taxonomic error	-	-	-0.31 (0.069)***	3467.5
Sam + Tax	**-**	**-0.05 (0.003)*** **	**0.47 (0.087)*** **	**3119.5**
OMISSION – a-LoCoH (N= 4346)				
Sampling effort	**-**	**-0.03 (0.002)*** **	**-**	**3442.9**
Taxonomic error	-	-	-0.58 (0.070)***	3630.3
Sam + Tax	-	-0.03 (0.003)***	-0.14 (0.080) ·	3442.1

***p<0.001; **p<0.01; *p<0.05; ·p<0.1

DISCUSSION

Range maps provide important information on the properties of critical elements of biogeographical, large-scale ecology and biodiversity conservation studies. In our case, range maps represent the actual and most current distribution range of the studied species, and omission errors have relatively small, although significant, effects on the estimated range, particularly in certain geographic areas. On the other hand, the fit between the range maps suggests a strong variation, and the geographical variation is parallel to the bias observed for the research efforts and the ecological observations (Martin et al., 2012). This measures of range accuracy that may be useful for future research using the southern water vole range maps as baseline data, and demonstrates a need for greater investment in the continuous update and improvement of the data necessary to generate the distribution ranges.

A general characterization of the data showed that the GBIF data provide more amount of information along with a more heterogeneous distribution, causing a strong spatial bias, with areas with lots of coverage while others have no observations. Nevertheless, GBIF is configured within the framework of a platform where a large amount of spatial information accumulates over time (Boitani et al., 2011). The survey data includes only recent data (data of the year in which the survey was conducted) and a more homogeneous distribution throughout the peninsular Spain. Direct field sampling provides full coverage of the species distribution reducing the spatial bias. Nevertheless, this type of sampling is rarely feasible over large spatial scales, as it would require significant resources and time (Meyer et al., 2016), but the use of this data and particularly those from specific samplings in areas of low data coverage, even on a small scale, could be used to improve or refine ranges on a global scale.

The analysis of range maps involved several steps that could explain some of the observed errors (Chanson et al., 2008). In agreement with previous studies (Hurlbert and Jetz, 2007; Cantú-Salazar and Gaston, 2013), we found that range maps contained errors, being the omission error rates higher than the commission error rates. This was unexpected since range maps are considered much more prone to range overestimation than to underestimation (Rondinini et al., 2006), even though other studies had also found non-trivial levels of omission errors in range maps (Beresford et al., 2011, Cantú-Salazar and Gaston, 2013). Omission errors may result from an underestimation of species' range extension (Cantú-Salazar and Gaston, 2013). In our analysis, omission errors are due to areas where there is little GBIF data (unequal sampling effort). The accuracy of the ranges reflects the bias towards better sampled regions with easier access to information (Martin et al., 2012). The accumulation of information over time (GBIF data) often does not necessarily serve to overcome the biases (in this case, spatial) (Meyer et al., 2016) of the sources of information, but it can make lose the temporal dynamics of the distributions, as happens with the disappearance of the southern water vole in the areas where the American mink has arrived. The rapid expansion of this invasive exotic species is threatening the survival of the southern water vole and the patterns found throughout the distribution areas they show this fact by producing a decrease in their distribution area that is starting in the NE of peninsular Spain. The loss of temporal dynamics happens not only because of how GBIF accumulates the data temporarily, but because some data providers themselves give the data in blocks

without temporal information, giving only the year in which this information is added. As occurs with some museum data, sometimes the data provided covers long periods of information that is not properly metadated. Although these platforms are making a great effort to encompass all biodiversity information on a global scale, both users and providers should be cautious and careful with the detail of the information used and provided if we want it to be useful for future research.

The concordant areas can be interpreted as a good approximation of the real area presence of the species, given the high percentage of overlap of the areas generated with different sources of primary information and methods. In spite of obtaining large concordant areas, the identification of the causes that lead to omission and commission errors, make clear that we need data workflows with integrated feedback loops and analysis of the places lacking information or with low sampling effort to determine where the efforts are sufficient and where additional data should be collected (Kissling et al., 2017).

The lack of sampling effort arises not only because the data do not exist or are not accessible, but because of low detectability (Ruete, 2015). This is of particular relevance in the case of time-explicit data, which are fundamental to understanding the trends of biodiversity, and where the only way to fill in the gaps is to make data recognizable (Mihoub et al., 2017). The availability of standardized and complete metadata when providing information on the basic characteristics of the data, including taxonomic, spatio-temporal information, as well as methodology, is very important too. Such metadata allow a for a rapid assessment of the quality of the data and the aptitude for its intended use. The improvement of metadata can act as an interim solution in cases where there are difficulties in avoiding full access to data. Conservation actions and strategies require data of sufficient quality. These data must have a minimal geographical, taxonomic and temporal resolution (Westgate et al., 2013). Without such data, inappropriate actions in conservation management become more likely. In the absence of these data, at least the interpretation of such data should be made considering the influence of possible errors on the distribution maps. We emphasize that efforts are required to increase the spatial uniformity of the sampling effort and, that data providers must put all available detail and possible care to allow for making the most out of their data.

ACKNOWLEDGEMENTS

We thank to Global Biodiversity Information Facility for making their database freely available online. This work was funded by the following grants and projects: Spanish Ministry of Economy, Industry and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R+D+I (SEV-2012-0262) and Agencia Estatal de Investigación from Ministry of Economy, Industry and Competitiveness, Spain with projects CGL2012-35931 and CGL2017-83045-R AEI/FEDER EU, co-financed with FEDER to E.R., R.B.M., M.G.S.

GENERAL DISCUSSION

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

This thesis documents the knowledge, construction, analysis and evaluation of the complex spatial and temporal patterns in which individual organisms are distributed on the Earth's surface and provides important information on the properties of these data. The update and continuous improvement of distribution range maps deepens our knowledge on the distribution of species and this is desirable when seeking to guide decisions on priority conservation actions based on threat levels. (Ficetola et al., 2013; Di Marco et al., 2013). Each of the four chapters of this thesis provides an exhaustive discussion of the relevant issues, so this final section represents a synthetic effort to integrate its main findings into a macro-ecological framework.

Defining species distribution ranges

The process of constructing a species distribution range begins with the definition of the range. This definition must be clear and concise and is essential for the subsequent selection of the geographic algorithm that best suits the purpose of the research (Ríos-Pena et al., Chapter 1). The different methodologies used in the construction of distribution ranges produce different distribution range maps for the same pattern of spatial data, highlighting the importance of considering the benefits and shortcomings of the method used to create the maps. Nowadays, the efforts to do comparative, and especially quantitative, research are complicated by problems of defining and mapping distribution ranges (Brown et al., 1996; Graham and Hijmans, 2006). This is therefore a priority objective to be pursued by the scientific community in future research.

To construct species' ranges, researchers often use georeferenced records available in global databases. These databases suffer from gaps in data coverage and spatial, taxonomic and species-level biases (Ficetola et al., 2014; Meyer et al., 2016). The quality of the available information is not homogeneous among species, nor is species that lack information randomly distributed among families and regions. Heterogeneity in data availability and quality is a serious constraint to generating unbiased distribution ranges. The quantity and quality of data must also be explicit in order to recognize the limitations of the chosen method (Ríos-Pena et al., in prep-chapter 2). Our study shows that a correct estimation of distribution ranges requires good quality data. To this end, we must apply substantial amounts of taxonomic

knowledge, time and funding to collect, verify and clean up public databases. Users should carefully clean the datasets before using them by conservatively eliminating poorly annotated records and records that may be misplaced (Ficetola et al., 2015). It is necessary to establish standardized criteria with minimum levels in the quantity and quality of information to facilitate the use of the distribution ranges.

Characterization of the information used to construct range maps

Three components have been explored to understand how the quality of data in biodiversity databases affects the construction of distribution ranges using geographic algorithms: data quantity, spatial bias and the presence of errors. The amount of information is a very limiting factor and therefore it is necessary to set a minimum number of observations from which we believe it is possible to generate a distribution range. Throughout this thesis, the minimum value for the analyses was 10 records per species, but a much higher number is required to improve the estimates. The only thing we can do when there is no or few data available is to go out and collect it. We can minimize the impact of spatial biases in sampling effort by resampling data in oversampled areas, and that of errors by carefully crosschecking the data and by removing a fraction of extreme observations (Ríos-Pena et al., Chapter 2).

In general, when data are distributed randomly or uniformly, the accuracy of all geographic algorithms improves with sample size. Data from uniform or unbiased random sampling are rare or non-existent for most regions and species (Gaston and Rodrigues, 2003; Rocchini et al., 2011). Heterogeneity of sampling effort induces a bias that may affect the estimation of ranges (Meyer et al., 2016; Pimm et al., 2014). This type of bias significantly decreases the sensitivity of all methods, especially when the range is irregular or fragmented. This means that distribution ranges generated with currently available data leave undetected areas where focal species are present. The existence of spatial biases in the data prevents the detection of complete ranges, making it necessary to substantially increase the sample size to improve estimates (Boitani et al., 2011; Burgman and Fox, 2003). Spatial biases in species records are relevant in GBIF and other global data sources because heterogeneous factors such as human population density, access to technology, the presence of a well-developed transport system or the availability of funds can affect their collection, storage and

mobilization. (Beck et al., 2013). At this point, it is important to work to characterize and, if possible, reduce the presence of spatial biases in data repositories (Cantú-Salazar y Gaston, 2013; Beck et al., 2013; 2014).

Spatial errors are another widespread problem in biodiversity databases (Maldonado et al., 2015). They can be generated in many ways and at any point in the data lifecycle. However, it is very difficult to obtain precise overall estimates of the importance of this problem. The presence of errors affects the performance of geographic algorithms, with the main disadvantage being the overestimation of the distribution ranges (Getz y Wilmers, 2004; Burgman y Fox, 2003). The reliability of the intervals obtained depends to a large extent on the quantification and control of spatial errors in the information sources. When the data contain errors and the sampling effort is spatially biased, there is a substantial deterioration in Type I error rates that increase with sample size. A possible way to reduce the impact of spatial errors is to exclude extreme values from the data set. The exclusion of extreme records before building ranges reasonably helps to improve the accuracy of algorithms for reproducing reference ranges, especially by reducing the Type I error rate, but does not qualitatively affect the overall performance of the different geographic algorithms (Ríos-Pena et al., Chapter 2).

Based on actual data, there is no single best method for sensitivity, type I error rate and range fragmentation (Guillera-Arroita et al., 2015; Qiao et al., 2015; Diniz-Filho et al., 2015). Depending on our objectives and the quantity and quality of the available data, some geographic algorithms should be preferred to others. All geographic algorithms show good behaviour in terms of sensitivity, even with low sample sizes. In most cases, this is at a high cost in the Type I error rate, including large areas where the species may not be present. More importantly, not all methods behave adequately in their Type I error rate with increasing sample size and should therefore be avoided (Ríos-Pena et al., Chapter 2). Range fragmentation is the most difficult property to reproduce. We must be aware of the requirements and limitations of the different geographical algorithms to estimate distribution ranges according to the type of data and the research question we want to address and, consequently, select the one that best suits our needs. Finally, in all cases we must be transparent with the data and the method used.

Systematic mapping of distribution ranges

Our study and evaluation for the production of systematically generated range maps for the order of carnivores provides important information on the possibilities of implementing such approaches for other biological groups (Ríos-Pena et al., Chapter 3). On the one hand, there is a fairly high correspondence between the ranges obtained through geographic algorithms (study conducted with KDE and a-LoCoH methods) and those provided by IUCN, suggesting that systematically generated range maps provide a reasonable and useful description of the species distribution. On the other hand, there was a large geographical variation in the fit between maps generated with geographical algorithms and those of IUCN, which is parallel to the overall bias observed for research efforts and ecological observations (Martin et al., 2012). This thesis demonstrated the urgent need for increased investment to update, improve and supplement information sources, particularly in the particularly rich areas of Asia, Africa and South America (Ríos-Pena et al., Chapter 3). This improvement in the sources of information would allow us to obtain distribution ranges in a more systematic way. Its main advantages would be more transparent analysis of more coherent data; be explicitly goal-oriented; and combine two forms of flexibility, namely, opportunities to change data and objectives, and opportunities to assess options for achieving the objectives (Pressey and Cowling, 2001; Cowling et al., 2003; Meyer et al., 2015).

Systematic vs. non-systematic sampling

For the specific case of *Arvicola sapidus* in peninsular Spain, the information for the focal species was obtained through GBIF and a homogeneous systematic sampling throughout peninsular Spain. GBIF provided more information, but more heterogeneously distributed in space, causing a strong spatial bias, with areas of high coverage while others had no observations. In addition, GBIF is configured within the framework of a platform on which a large amount of spatial information is accumulated over time (Boitani et al., 2011). The accumulation of information over time often does not necessarily serve to overcome biases, in this case spatial, of information sources (Meyer et al., 2016), but it can cause the temporal dynamics of distributions to be lost. Although these platforms are making a great effort to cover all biodiversity information on a global scale, the citizen scientists and administrations

that are data providers of the platforms need to be cautious and careful with the detail of the information used and provided if it is to be useful for future research. Survey data included only recent data (data from the year the survey was conducted) and had a more homogeneous distribution throughout the peninsula, reducing spatial bias in the estimation of *Arvicola* distribution However, this type of sampling is rarely feasible at large spatial scales, as it would require significant resources and time (Meyer et al., 2016), but the use of these data, and particularly those from specific sampling in areas of low data coverage, even at small scales, should be used to improve or refine ranges at global scales.

The analysis of overlapping distribution ranges built with KDE and a-LoCoH algorithms through systematic and non-systematic sampling showed that the concordant areas can be interpreted as a good approximation of the real presence of the species' area (Ríos-Pena et al., in prep-chapter 4). The identification of the causes leading to errors of omission and commission, in turn, makes it clear once again that we need data workflows with integrated feedback loops and analyses of missing or under-sampled sites to determine where efforts are sufficient and where additional data should be collected (Ríos-Pena et al., Chapter 4). The lack of sampling effort is due not only to the fact that the data do not exist or are not accessible, but also to the low detectability (Ruete, 2015). This is particularly relevant in the case of time-sensitive data, which are fundamental to understanding biodiversity trends, and where the only way to fill the gaps is to make the data recognizable (Mihoub et al., 2017). Conservation actions and strategies require data of sufficient quality. These data must have a minimum geographic, taxonomic and temporal resolution (Westgate et al., 2013). Without this type of data, inappropriate actions in conservation management are more likely. In the absence of these data, at least the interpretation of these data should take into account the influence of possible errors in the distribution maps (Kissling et al., 2017).

Synthesis

In this thesis we have shown that, the scientific literature often overlook to make an explicit interpretation of the distribution range concept, causing serious difficulties in the description of the patterns related to species distributions. There are many geographical algorithms to generate distribution ranges from occurrence data but they are rarely used for this purpose, and it is more common to use the distribution maps provided by the IUCN Red

List. But reported macro-ecological patterns may depend on the characteristics of the distribution ranges used as baseline data, which are, in turn, dependent of the methodologies used, even when working with well-studied taxonomic groups such as terrestrial mammals. We attribute the differences observed in size, shape and geographical location derived primarily from the precautionary principle that underpins the IUCN expert maps and makes them relatively sensitive to geographical variation in the sampling effort, which is common in most parts of the world. Finally, we urge caution in the process of defining, using data and building range maps. We provide a systematic tool for the construction of species distribution ranges in order to allow for comparisons between species distribution maps, since there is no geographic algorithm that works best, but everything will depend on the research question we want to answer.

BIBLIOGRAPHY

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Aars, J., Lambin, X., Denny, R., Griffin, A. (2001). Water vole in the Scottish uplands: distribution patterns of disturbed and pristine populations ahead and behind the American mink invasion front. Anim. Conserv., 4: 187–194.

Aguiar, A., Lopes, A. L., Pimenta, M., Luís, A. (2011). Owls (Strigiformes) in Parque Nacional Peneda-Gerês (PNPG) – Portugal. Nova Acta Científica Compostelana (Bioloxía), 19: 83–92.

Akay, A.E., Inac S., Yildrim, I.C. (2011). Monitoring the local distribution of striped hyenas (Hyaena hyaena L.) in the Eastern Mediterranean Region of Turkey (Hatay) by using GIS and remote sensing technologies. Environ. Monit. Assess. 181: 445–455.

Andreotti, A., G. Leonardi, M. Sarà, M. Brunelli, L. De Lisio, A. De Sanctis, M. Magrini, R. Nardi, et al. (2008). Landscape-scale spatial distribution of the lanner falcon (Falco biarmicus feldeggii) breeding population in Italy.Ambio 37: 443–447.

Angulo, A., Icochea, J. (2010). Cryptic species complexes, widespread species and conservation: lessons from Amazonian frogs of the Leptodactylus marmoratus group (Anura: Leptodactylidae). Systematics and Biodiversity, 8: 357–370.

Aquino, R., Cornejo, F.M., Lozano, E.P., Heymann, E.W. (2009). Geographic distribution and demography of *Pithecia aequatorialis* (Pitheciidae) in Peruvian Amazonia. Am. J. Primatol. 71: 964–968.

Arlt, M.L., Manseau, M. (2011). Historical changes in caribou distribution and land cover in and around Prince Albert National Park: land management implications. Rangifer, Special Issue No. 19, 17–32.

Attorre, F., et al. (2012). The use of spatial ecological modelling as a tool for improving the assessment of geographic range size of threatened species. Nat. Conserv., 21(1), pp. 48–55

Bader, M. (2000). Distribution of grizzly bears in the U.S. northern Rockies. Northwest Science 74: 325–334.

Balciauskas, L. (2008). Wolf numbers and distribution in Lithuania and problems of species conservation. Ann. Zool. Fenn. 45, 329–334.

Banuelos, M.J., Kollmann, J., Hartvig, P., Quevedo, M. (2004). Modelling the distribution of Ilex aquifolium at the north-eastern edge of its geographical range. Nordic Journal of Botany 23: 129–142.

Barros, P., Vale-Gonçalves, H.M., Paupério, J., Cabral, J.A., Rosa, G. (2016). Confirmation of European snow vole Chionomys nivalis (Mammalia: Rodentia: Cricetidae) occurrence in Portugal, Italian. Journal of Zoology, 83(1): 139–145.

Baselga, A., Lobo, J.M., Svenning, J.-C., Araújo, M.B. (2012). Global patterns in the shape of species geographical ranges reveal range determinants. Journal of Biogeography, 39, 760–771.

Beck, J., Ballesteros-Mejia, L., Nagel, P., Kitching, I.J. (2013). Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions, 19: 1043–1050.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics, 19: 10–15.

Beresford, A., Buchanan, G., Donald, P., Butchart, S.H.M., Fishpool, L.D.C., Rondinini, C. (2011). Poor overlap between the distribution of Protected Areas and globally threatened birds in Africa. Anim. Conserv. 14: 99–107.

Bergl, R.A., Warren, Y., Nicholas, A., Dunn, A., Imong, I., Sunderland-Groves, J., Oates, J.F. (2010). Remote sensing analysis reveals habitat, dispersal corridors and expanded distribution for the critically endangered Cross River gorilla Gorilla gorilla diehli. Oryx, 46, 278–289.

Boakes, E.H., McGowan, P.J., Fuller, R.A., Chang-qing, D., 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(6): e1000385.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P., Rondinini, C. (2011). What spatial data do we need to develop global mammal conservation strategies? Philosophical Transactions of the Royal Society B: Biological Sciences 366: 2623–2632.

Bombi, P., Luiselli, L., D'Amen, M. (2011). When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. Diversity and Distributions, 17, 581–592.

Bronstein, A. M., Bronstein, M. M., & Kimmel, R. (2007). Calculus of non-rigid surfaces for geometry and texture manipulation. IEEE Transactions on Visualization and Computer Graphics, 13(5), 902–913.

Brown, A.F., Crick, H.Q.P., Stillman, R.A. (1995). The distribution, numbers and breeding ecology of Twite Acanthis flavirostris in the south Pennines of England. Bird Study 42: 107–121.

Brown, J.H., Stevens, G.C., Kaufman, D.M. (1996). The geographic range: size, shape, boundaries, and internal structure. Annu. Rev. Ecol. Syst., **27**, 597–623.

Bruderer, B., Bruderer, H. (1993). Distribution and habitat preference of Red-backed Shrike (*Lanius collurio*) in southern Africa. Ostrich 64:141–147.

Brugière D., Badjinca, I., Silva, C., Serra, A. (2009). Distribution of chimpanzees and interactions with humans in Guinea-Bissau and Western Guinea, West Africa. Folia Primatol. 80: 353–358.

Burgman, M.A., Fox, J.C. (2003). Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. Anim. Conserv. 6: 19–28.

Butler, J.A., Heinrich, G.L. (2013). Distribution of the ornate diamondback terrapin (Malaclemys terrapin macrospilota) in the Big Bend region of Florida. Southeastern Naturalist, 12: 552–567.

Calenge, C. (2006). The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. Ecol. Modell. 197: 516–519.

Cano P.D., Cardozo H.G., Ball H.A., D'Alessio S., Herrera P., Lartigau B. (2012). Aportes al conocimiento de la distribución del ciervo de los pantanos (Blastocerus dichotomus) en la provincia de Corrientes, Argentina. Mastozoologia Neotropical, 19(1): 35–45.

Carwardine, J., Wilson K.A., Watts M., Etter A., Klein C.J., Possingham H.P. (2008). Avoiding costly conservation mistakes: the importance of defining actions and costs in spatial priority setting. PLoS ONE 3, e2586.

Cantú-Salazar, L., Gaston, K.J. (2013). Species richness and representation in protected areas of the Western hemisphere: discrepancies between checklists and range maps. Diversity and Distributions, 19: 782–793.

Ceballos, G., Ehrlich, P. R. (2006). Global mammal distributions, biodiversity hotspots, and conservation. Proceedings of the National Academy of Sciences of the United States of America, 103(51), 19374–19379.

Chanson, J.S., Stuart, S.N., Cox, N., Young, B.E., Hoffman, M. (2008). The Global Amphibian Assessment (GAA): history, objectives and methodology. Threatened amphibians of the world (ed. by S.N. Stuart, M. Hoffman, J.S. Chanson, N.A. Cox, R.J. Berridge, P. Ramani and B.E. Young), pp. 30–32. Lynx Editions, IUCN and Conservation International, Barcelona.

Charles, C., Schwartz, U.S. (2002). Distribution of grizzly bears in the Greater Yellowstone Ecosystem: 1990–2000. Ursus 13: 203–213.

Cianferoni, F. (2013). Distribution of Cymatia rogenhoferi (Fieber, 1864) (Hemiptera, Heteroptera, Corixidae) in the West-Palaearctic Region, with the first record for the Italian mainland. North-Western Journal of Zoology 9(2): 245–249.

Cogălniceanu, D., et al. (2008). The current distribution of herpetofauna in the Maramureş county and the Maramureş Mountains Nature Park, (Maramureş, Romania), Transylvanian Review of Systematical and Ecological Research, Vol. 5, The Maramureş Mountains Nature Park, pp. 189–200.

Cogălniceanu D., Szekely P., Samoilă C., Iosif R., Tudor M., Plăiaşu R., Stănescu F., Rozylowicz, L. (2013): Diversity and distribution of amphibians in Romania. ZooKeys, 296: 35–57.

Costion, C.M., Kitalong, A.H., Perlman, S., Edwards, W. (2013). Palau's rare and threatened palm Ponapea palauensis (Arecaceae): Population density, distribution, and threat assessment. Pac. Sci. 67: 599–608.

Cox, C.B., Moore, P.D. (2004). Biogeography: An Ecological and Evolutionary Approach. Oxford: Blackwell.

Cribari-Neto, F., Zeileis, A. (2010). "Beta Regression in R." Journal of Statistical Software, 34(2): 1–24.

Cuppen, J.G.M. (1986). On the habitats, distribution and life-cycles of the Western European species of the genus *Helochares Mulsant* (Coleoptera: Hydrophilidae). Hydrobiologia, 132, pp. 169-183.

de Tores, P.J., Hayward, M.W., Dillon, M.J., Brazell, R. (2007). Review of the distribution, causes for the decline and recommendations for management of the quokka, *Setonix brachyurus*(Macropodidae: Marsupialia), an endemic macropod marsupial from south-west Western Australia. Conservation Science West Aust 6: 13–73.

Dennis, R.L.H. (2001). Progressive bias in species status is symptomatic of fine-grained mapping units subject to repeated sampling. Biodiversity and Conservation 10: 483–494.

Desender, K., Dekoninck, W., Dufrêne, M., Maes, D. (2010). Changes in the distribution of carabid beetles in Belgium revisited: have we halted the diversity loss? Biological Conservation, 143, 1549–1557.

Di Marco, M., Santini, L. (2015). Human pressures predict species' geographic range size better than biological traits. Global Change Biology, 21, 2169–2178.

Diniz-Filho, J.A.F., Rodrigues, H., Telles, M.P.D.C., Oliveira, G.D., Terribile, L.C., Soares, T.N., Nabout, J.C. (2015). Correlation between genetic diversity and environmental suitability: taking uncertainty from ecological niche models into account. Molecular Ecology Resources, 15: 1059–1066.

Domisch, S., Wilson, A. M. and Jetz, W. (2016). Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. Ecography, 39: 1078–1088.

Donato, G., Belongie, S. (2002). Approximate thin plate spline mappings. In: ECCV 2002, Part III. pp. 21–31.

Dormann, C.F. (2007). Promising the future? Global change projections of species distributions. Basic and Applied Ecology 8: 387–397.

Downs, J.A., Horner, M.W.,Tucker, A.D. (2012). Time-geographic density estimation for home range analysis. Annals of GIS, 17(3): 163–171.

Duin, R.P.W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions in Computing C-25: 1175–1179.

Encabo, I., Barba, E., Belda, E.J., Monrós, J.S. (2007). Área de campeo de quirópteros en el término municipal de Carcaixent (Valencia): Nuevas citas para el Atlas de los Mamíferos Terrestres. Galemys 19:37–44.

Edwards, J.L. (2004). Research and societal benefits of the Global Biodiversity Information Facility. BioScience, 54: 485–486.

El Alqamy, H., Baha El Din, S. (2006). Contemporary status and distribution of gazelle species (Gazella dorcas and Gazella leptoceros) in Egypt. Zoology in the Middle East 39, 5–16.

Elith, J., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129–151.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226–231.

Faurby, S., Svenning, J.C. (2015). Historic and prehistoric human□driven extinctions have reshaped global mammal diversity patterns. Diversity and Distributions, 21: 1155–1166.

Ferrari, S.L.P., Cribari-Neto, F. (2004). "Beta Regression for Modelling Rates and Proportions." Journal of Applied Statistics, 31: 799–815.

Ficetola GF, Pennati R, Manenti R (2013). Spatial segregation among age classes in cave salamanders: habitat selection or social interactions? Population Ecology, 55: 217–226.

Ficetola, G.F., Rondinini, C., Bonardi, A., Katariya, V., Padoa-Schioppa, E., Angulo, A. (2014). An evaluation of the robustness of global amphibian range maps. Journal of Biogeography, 41, 211–221.

Flueck, W.T., Smith-Flueck, J.A.M., Naumann, C.M. (2003). The current distribution of red deer (*Cervus elaphus*) in southern Latin America. Eur J Wild Res 49:112–119.

Foody, G.M. (2011). Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. Global Ecology and Biogeography, 20: 498–508.

Fortin, M.J., Keitt, T.H., Maurer, B.A., Taper, M.L., Kaufman, D.M., Blackburn, T.M. (2005). Species geographic ranges and distributional limits: pattern analysis and statistical issues. Oikos 108: 7–17.

García, J.T., et al. (2008). Assessing the distribution, habitat, and population size of the threatened Dupont's lark Chersophilus duponti in Morocco: lessons for conservation. Oryx 42, 592–599.

Garcillan, P.P., Ezcurra, E., Riemann, H. (2003). Distribution and species richness of woody dryland legumes in Baja California, Mexico. Journal of Vegetation Science, 14(4): 475–486.

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

Gaston, K.J., Fuller, R.A. (2009). The sizes of species' geographic ranges. Journal of Applied Ecology, 46: 1–9.

Gaston, K.J., Rodrigues, A.S.L. (2003). Reserve selection in regions with poor biological data. Cons. Biol. 17: 188–95.

Gaubert, P., Jiguet, F., Bayle, P., Angelici, F.M. (2008). Has the common genet (Genetta genetta) spread into south-eastern France and Italy? Ital. J. Zool. 75, 43–57.

Gautier, J.P., Moysan, F., Feistner, A.T.C., Loireau, J.N., Cooper, R. (1992). The distribution of Cercopithecus (lhoesti) solatus: An endemic guenon of Gabon. Revue de Ecologie (La Terre et la Vie) 47: 367–381.

Getz, W.M., Fortmann-Roe, S., Cross, P.C., Lyons, A.J., Ryan, S.J., Wilmers, C.C. (2007). LoCoH: non-parameteric kernel methods for constructing home ranges and utilization distributions. PLoS ONE, 2, e207.

Getz, W.M., Wilmers, C.C. (2004). A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. Ecography, 27: 489–505.

Gitzen, R.A., Millspaugh, J.J., Kernohan, B.J. (2006). Bandwidth selection for fixed kernel analysis of animal range use. Journal of Wildlife Management: 70: 1334–1344.

González-Prat, F., Puig, D., Folch, A. (2001). Distribución de la marmota alpina *Marmota marmota* (Linnaeus, 1758) en el extremo suroriental del Pirineo. Galemys 13:139–148.

Goodwin Z.A., Harris D.J., Filer D., Wood J.R.I., Scotland R.W. 2015. Widespread mistaken identity in tropical plant collection. Curr. Biol. 25: R1066-R1067.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol. Evol., 19: 497–503.

Graham, C.H., Hijmans, R.J. (2006). A comparison of methods for mapping species ranges and species richness. Global Ecology and Biogeography, 15: 578–587.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend, P.A., Loiselle, B.A. (2008). The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45: 239–247.

Geissman, T., Lwin, N., Aung, S.S., Aung, T.N., Aung, Z.M., et al. (2011). A new species of snub-nosed monkey, Genus *Rhinopithecus* Milne-Edwards, 1872 (Primates, Colobinae), from northern Kachin state, northeastern Myanmar. Am J Primatol 73: 96–107.

González-Maya J.F., Castañeda F., González R., Pacheco J., Ceballos G. (2014). "Distribution, range extension, and conservation of the endemic black-headed bushmaster (Lachesis melanocephala) in Costa Rica and Panama," Herpetol. Conserv. Biol., 9(2), 369–377.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography, 24: 276–292.

Guisan, A., Tingley, R., Baumgartner, J.B. et al. (2013). Predicting species distributions for conservation decisions. Ecology Letters, 16: 1424–1435.

Habbema, J.D.F., Hermans, J., van den Brock, K. (1974). A stepwise discriminant analysis program using density estimation. In: Proceedings of COMPSTAT 1974, Physica Verlag, Heidelberg, pp. 101–110.

Harvey, D.J., Gange, A.C., Hawes, C.J., Rink, M. (2011). Bionomics and distribution of the stag beetle, Lucanus cervus (L.) across Europe. Insect Conservation and Diversity, 4, 23–38.

Hayfield, T., Racine, J.S. (2007). The np package, Manual, McMaster University.

Hengl, T., Sierdsema, H., Radović, A., Dilod, A. 2009. Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. Ecological Modeling 24: 3499–3511.

Hennig, C. (2015). Fcp: Flexible procedures for clustering. R package version 3.2.5. https://CRAN.R-project.org/package=fpc.

Herkt K.M.B., Skidmore A.K., Fahr J., (2017). Macroecological conclusions based on IUCN expert maps: A call for caution. Global Ecol Biogeogr. 26: 930–941.

Hernández, H., Gómez-Hinojosa C., Hoffman G. (2010). ¿Es la rareza geográfica frecuente entre las cactáceas del Desierto Chihuahuense? Revista mexicana de Biodiversidad. 81: 183–175.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29: 773–785.

Herrando, S., Brotons, L., Guallar, S., Quesada, J. (2010). Assessing regional variation in conservation value using fine-grained bird atlases. Biodivers. Conserv. 19, 867–881.

Hirsch, A., Chiarello, A.G. (2012). The endangered maned sloth Bradypus torquatus of the Brazilian Atlantic forest: a review and update of geographic distribution and habitat preferences. Mammal Review 42: 35–54.

Hof, C., Araújo, M.B., Jetz, W., Rahbek, C. (2012). Additive threats from pathogens, climate and land-use change for global amphibian diversity. Nature, 480, 516–519.

Horne, J.S., Garton, E.O. (2006). Likelihood cross-validation versus least squares cross-validation for choosing the smoothing parameter in kernel home-range analysis. Journal of Wildlife Management 70: 641–648.

Horne, J.S., Garton, E.O., Krone, S.M., Lewis, J.S. (2007). Analyzing animal movements using Brownian bridges. Ecology 88: 2354–2363.

Hortal, J., (2008). Uncertainty and the measurement of terrestrial biodiversity gradients. Journal of Biogeography 35: 1335–1336.

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

Hwang, W.-H., He, F. (2011). Estimating abundance from presence/absence maps. Methods Ecol. Evol. 2, 550–559.

Idžojtić, M., Kogelnik, M., Franjić, J., Škvorc, Ž. (2006). Hosts and distribution of *Viscum* L. ssp. *album* in Croatia and Slovenia. Plant Biosyst 140: 50–55.

Irwin, M. D., Coelho, J. R. (2000). Distribution of the Iowan Brood of periodical cicadas (Homoptera: Cicadidae: Magicicada spp.) in Illinois. Ann. Entomol. Soc. Am. 93: 82–89.

IUCN (2012). IUCN Red List Categories and Criteria: Version 3.1. Second edition. Gland, Switzerland and Cambridge, UK: IUCN. Available at www.iucnredlist.org/technical-documents/categories-and-criteria

IUCN (2017). Guidelines for Using the IUCN Red List Categories and Criteria. Version 13. Prepared by the Standards and Petitions Subcommittee. Available at http://www.iucnredlist.org/documents/RedListGuidelines.pdf

Jaryan, V., Uniyal, S. K., Kumar, A., Gupta, R. C., Singh, R. D. (2012). Extent of Occurrence and Area of Occupancy of Tallow Tree (*Sapium sebiferum*): Using the Red list Criteria for Documenting Invasive Species Expanse. *National Academy Science Letters*, Springer.

Jerina, K., et al. (2013). Range and local population densities of brown bear Ursus arctos in Slovenia. — Eur. J. Wildl. Res. 59: 459–467.

Jetz, W., McPherson, J.M., Guralnick, R.P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol. Evol. 23: 151–159.

Jiménez-Valverde, A., Ortuño, V.M., Lobo, J.M. (2007). Exploring the distribution of Sterocorax Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. Journal of Biogeography, 34: 1426–1438.

Johansson, F. (1993). The distribution of Odonata in Västerbotten and South Lapland, northern Sweden. Entomol Fenn, 4:165–168.

Johnson, C.J., Hurley, M., Rapaport, E., Pullinger, M. (2012). Using expert knowledge effectively: lessons from species distribution models for wildlife conservation and management. In: Expert Knowledge and its Application in Landscape Ecology (eds A.H. Perera, C.A. Drew & C.J. Johnson), pp. 153–171. Springer, New York.

Kalwij, J. M. et al. (2014). Spatially-explicit estimation of geographical representation in large-scale species distribution datasets. – PLoS One 9: e85306.

Karakaş, R. (2012). Does black-winged kite Elanus caeruleus (Desfontaines, 1789) have an expansion in its range in Turkey? Acta Zoologica Bulgarica 64(2): 209–214.

Keith, D.A. (2000). Sampling designs, field techniques and analytical methods for systematic plant population surveys. Ecol. Mgmt. Restor. 1: 125–139.

Kelly, J. R., Fuller, T. K., Kanter, J. J. 2009. Records of recovering American marten, Martes americana, in New Hampshire. Canadian Field-Naturalist, 123: 1–6.

Kierulff, M.C.M., Rylands, A.B. (2003). Census and distribution of the golden lion tamarin (Leontopithecus rosalia). American Journal of Primatology, Washington, 59 (1): 29–44.

King, J.L., Sue, M.C., Muchlinski, A.E. (2010). Distribution of the Eastern Fox Squirrel (Sciurus niger) in Southern California. The Southwestern Naturalist, 55, pp. 42–49.

Kissling, W.D., et al. (2017). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological Reviews. 93, pp. 600 – 625.

Kleman, L.Jr., Vieira, J.S. (2013). Assessing the extent of occurrence, area of occupancy, territory size, and population size of Marsh Tapaculo (Scytalopus iraiensis). Anim. Biodivers. Conserv. 36: 47–57.

Kowalchuk, K.A., Kuhn, R.G. (2012). Mammal distribution in Nunavut: Inuit harvest data and COSEWIC's species at risk assessment process. Ecol Soc 17:4.

Krištín, A., Kaňuch, P., Sarossy, M. (2007). Distribution and ecology of Ruspolia nitidula (Scopoli 1786) and Aiolopus thalassinus (Fabricius 1781) (Orthoptera) in Slovakia. Linzer Biologische Beiträge, 39, 451–461.

Kryštufek, B., Vohralík, V. (1994). Distribution of the Forest dormouse *Dryomys nitedula* (Pallas, 1779) (Rodentia, Myoxidae) in Europe. Mammal Review 24: 161–177.

Kugler, K.A., Geluso, K. (2009). Distribution of the eastern woodrat (Neotoma floridana campestris) in southern Nebraska. Western North American Naturalist 69: 175–179.

Kylin, H., Bouwman, H., Louette, M. (2011). Distribution of the subspecies of Lesser Black-backed Gulls Larus fuscus in sub-Saharan Africa. Bird Study, 58: 186–192.

Laplana, C., Sevilla, P. (2013). Documenting the biogeographic history of Microtus cabrerae Thomas, 1906 (Arvicolinae, Rodentia, Mammalia) through its fossil record. Mammal Review, 43: 309–332.

Lawler, J.J., Shafer, S.L., Bancroft, B.A., Blaustein, A.R. (2010). Projected climate impacts for the amphibians of the western hemisphere. Conserv. Biol, 24: 38–50.

Leranoz, I., Castien, E. (1996). Evolution of wild boar *Sus scrofa L*, 1758 in Navarra N Iberian Peninsula. Journal of the Natural History Museum and Institute Chiba 6(2): 201–207.

Li, Q., Racine, J.S. (2003). "Nonparametric estimation of distributions with categorical and continuous data," Journal of Multivariate Analysis, 86: 266–292.

Lizcano, D.J., Pizarro, V., Cavelier, J., Carmona, J. (2002). Geographic distribution and population size of the mountain tapir (Tapirus pinchaque) in Colombia. Journal of Biogeography, 29, 7–15.

Llaneza, L., Álvares, F., Ordiz, A., Sierra, P., Uzal, A. (2004). Distribución y aspectos poblacionales del Lobo Ibérico en la provincia de Ourense. [Distribution and populational aspects of The Iberian Wolf in the Ourense province]. Ecología 18: 227–238.

Lloyd, H.G. (1983). Past and present distribution of red and grey squirrels. Mammal Rev., 13: 69–80.

Lomolino, M.V. (2004). Conservation biogeography. Frontiers of biogeography: new directions in the geography of nature (ed. by M.V. Lomolino and L.R. Heaney), pp. 293–296. Sinauer, Sunderland, MA.

Long, Y.C., Kirkpatrick, R.C., Zhong, T., Xiao, L. (1994). Report on the distribution, population, and ecology of the Yunnan snub-nosed monkey (*Rhinopithecus bieti*). *Primates* 35: 241–250.

Long, J.A., Nelson, T.A. (2012). Time geography and wildlife home range delineation. J. Wildl. Manag., 76 (2), pp. 407–413.

Lortkipanidze, B. (2010). Brown bear distribution and status in the South Caucasus. Ursus 21: 97–103.

Lunney, D., Crowther, M.S., Shannon, I., Bryant, J.V. (2009). Combining a map-based public survey with an estimation of site occupancy to determine the recent and changing distribution of the koala in New South Wales. Wildl. Res. 36: 262–273.

Ma, C., Huang, Z. P., Zhao, X. F., Zhang, L. X., Sun, W. M., Scott, M. B., et al. (2014). Distribution and conservation status of *Rhinopithecus strykeri* in China. *Primates, 55*(3): 377–382.

MacGown, J.A., Wetterer, J.K. (2013). Distribution and biological notes of Strumigenys margaritae (Hymenoptera: Formicidae: Dacetini). Terrestrial Arthropod Reviews 6: 247–255.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J. et al. (2015). Estimating species diversity and distribution in the era of big data: to what extent can we trust public databases? Glob. Ecol. Biogeogr., 24: 973–984.

Manly, B.F.J. (1997). Randomization, bootstrap and Monte Carlo methods in biology, 2nd ed. Chapman and Hall, London.

Marboutin, E., Pruszek, M., Calenge, C., Duchamp, C. (2011). On the effects of grid size and shape when mapping the distribution range of a recolonising wolf (Canis lupus) population. European Journal of Wildlife Research, 57, 457–465.

Maréchaux, I., Rodrigues, A.S.L., Charpentier, A. (2017). The value of coarse species range maps to inform local biodiversity conservation in a global context. Ecography, 40: 1166–1176.

Margules, C., Pressey, R.L., Williams, P.H. (2002). Representing biodiversity: data and procedures for identifying priority areas for conservation. J. Biosci. 27, pp. 309–326.

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

Martin E., Hans-Peter K., Joerg S., Xiaowei X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Institute for Computer Science, University of Munich. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).

Martínez, A.I., et al. (2005). Sondeo y evolución de la distribución de la nutria paleártica (Lutra lutra Linnaeus, 1758) en el País Vasco (N España). Galemys, 1, pp. 25–46.

Martinez, J., Wallace, R.B. (2007). Further notes on the distribution of endemic Bolivian titi monkeys, *Callicebus modestus* and *Callicebus olallae*. Neotropical Primates, *14*(2): 47–54.

Mattos, J.C.F., Vale, M.M., Vecchi, M.B., Alves, M.A.S. (2009). Abundance, distribution and conservation of the Restinga Antwren Formicivora littoralis (Aves: Thamnophilidae). Bird Conserv Intern 19: 392–400.

McCullagh, P., Nelder, J.A. (1989). Generalized Linear Models, volume 37 of Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2 edition.

Merow, C., Wilson, A. M., Jetz, W. (2016). Integrating occurrence data and expert maps for improved species range predictions. Global Ecol. Biogeogr, 26: 243–258.

Meyer, C., Kreft, H., Guralnick, R.P., Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. Nature Communications, 6, 8221.

Meyer, C., Jetz, W., Guralnick, R. P., Fritz, S. A., Kreft, H. (2016). Range geometry and socio-economics dominate species-level biases in occurrence information. Global Ecol. Biogeogr., 25: 1181–1193.

Meyer, C., Weigelt, P., Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecol Lett 19: 992–1006.

Michelsen-Heath, S., Gaze, P. (2007). Changes in abundance and distribution of the rock wren (Xenicus gilviventris) in the South Island, New Zealand. Notornis, 54, pp. 71–78.

Miguel-Talonia, C., Téllez-Valdés, O., Murguía-Romero, M. (2014). Las cactáceas del Valle de Tehuacán-Cuicatlán, México: estimación de la calidad del muestreo. Revista Mexicana de Biodiversidad 85: 436–444.

Miguelez, D., Valladares, L.F. (2010). Hábitat y distribución de *Aphelocheirus murcius* Nieser & Millán, 1989 (Hemiptera: Aphelocheiridae) en el norte de la Península Ibérica. Limnetica 29: 387–392.

Mihoub, J.B., Henle, K., Titeux, N., Brotons, L., Brummitt, N., Schmeller D.S. (2017). Setting temporal baselines for biodiversity: the limits of available monitoring data for capturing the full impact of anthropogenic pressures. Sci. Rep., 7, 41591.

Mokany, K., Ferrier, S. (2011). Predicting impacts of climate change on biodiversity: a role for semi-mechanistic community-level modelling. Diversity and Distributions, 17: 374–380.

Monroy, F., Aira, M., Dominguez, J., Morino, F. (2003). Distribution of earthworms in the north-west of the Iberian Peninsula. European Journal of Soil Biology 39(1): 13–18.

Mota-Vargas, C., Rojas-Soto, O.R. (2012). The importance of defining the geographic distribution of species for conservation: the case of the Bearded Wood-Partridge. Journal for Nature Conservation, 20: 10–17.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J. (2002). Biodiversity hotspots for conservation priorities. Nature, 403: 853−858.

Niemi, M., Auttila, M., Viljanen, M., Kunnasranta, M. (2012). Movement data and their application for assessing the current distribution and conservation needs of the endangered Saimaa ringed seal. Endangered Species Research 19: 99–108.

Ó Teangana, D., Reilly, R., Montgomery, W.I., Rochford, J. (2000). Distribution and status of the red squirrel (*Sciurus vulgaris*) and grey squirrel (*Sciurus carolinensis*) in Ireland. Mammal Rev 30: 45–56.

Onmuş, O., Durusoy, R., Eken, G. (2009). Distribution of breeding birds in the Gediz Delta, Western Turkey. Zoology in the Middle East 47: 39–48.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A., Miller, P., Louzao, M. (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation, 156, 94–104.

Orme, C.D.L., Davies, R.G., Olson, V.A., Thomas, G.H., Ding, T.-S., Rasmussen, P.C. et al. (2006). Global patterns of geographic range size in birds. PLoS Biology, 4, e208.

O'Rourke, J. (1998). Computational geometry in C. Cambridge: Cambridge University Press.

Palacios, D.M., Gerrodette, T., García, C., Avila, I.C., Soler, G.A., Bessudo, S., Trujillo, F. (2008). Distribution and relative abundance of oceanic cetaceans in

Colombia's Pacific EEZ from survey cruises and platforms of opportunity. Journal of Cetacean Research and Management 12(1): 45–50.

Panelius, S. (1978). The detailed geographical distribution of Tettigonia cantans in Finland (Orthoptera, Tettigoniidae). Notulae Entomologicae, 584: 151–157.

Parsons, B.T., Middleton, A.D. (1937). The distribution of the Grey squirrel (Sciurus carolinensis) in Great Britain in 1937. Journal of Animal Ecology, 6, 386–390.

Pavlinić, I., Đaković, M., Tvrtković, N. (2010). The atlas of Croatian bats (Chiroptera), part I. Nature Croatia, 19, pp. 295–337.

Pebesma, E.J., Bivand, R.S. (2005). Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/

Peralta, D., Leitão, I., Ferreira, A., Mira, A., Beja, P., Pita, R. (2016). Factors affecting southern water vole (Arvicola sapidus) detection and occupancy probabilities in Mediterranean farmland. Mammalian Biology-Zeitschrift für Säugetierkunde, 81(2): 123–129.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Nakamura, M., Martinez-Meyer, E., Araújo, M.B. (2011). Ecological niches and geographical distributions. Princeton University Press, Princeton, New Jersey, USA.

Pike, D.A., Roznik, E.A. (2009). Drowning in a sea of development: distribution and conservation status of a Sand-Swimming Lizard, Plestiodon reynoldsi. Herpetological Conservation and Biology 4: 96–105.

Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., Raven, P.H., Roberts, C.M., Sexton J.O. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. Science 344: 1246752.

Pinto, L.P.S., Rylands, A.B. (1997). Geographic distribution of the golden-headed lion tamarin, Leontopithecus chysomelas: implications for its management and conservation. Folia Primatologica68:161–180.

Printes, R.C., Rylands, A.B., Bicca-Marques, J.C. (2011). Distribution and status of the Critically Endangered blond titi monkey Callicebus barbarabrownae of north-east Brazil. Oryx, 45(3): 439–443.

Popović, M., Radaković, M., Đurđević, A., Franeta, F., Verovnik, R. (2014) Distribution and threats of *Phengaris teleius* (Lepidoptera: Lycaenidae) in Northern Serbia. Acta Zool Hungarica 60:173–183.

Pupins M., Pupina A. (2017). Updated distribution of the European Pond Turtle, Emys orbicularis (L., 1758) (Emydidae) on the extreme northern border of its European range in Latvia. Acta Zoologica Bulgarica, Supplement 10: 133–137.

Qiao, H., Soberón, J., Peterson, T.A. (2015). No silver bullets in correlative ecological niche modeling: insights from testing among many potential algorithms for niche estimation. Methods in Ecology and Evolution, 6: 1126–1136.

Queirolo, D., Moreira, J.R., Soler, L., Emmons, L.H., Rodrigues, F.H.G., Pautasso, A.A., Cartes, J.L., Salvatori, V. (2011). Historical and current range of the Near Threatened maned wolf *Chrysocyon brachyurus* in South America. Oryx 45: 296–303.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Ramos P.L., Merchán T., Rocha G., Hidalgo de Trucios S.J. (2009). Distribución actual del meloncillo, Herpestes ichneumon (Linnaeus, 1758), en el sur de la provincia de Salamanca y en el norte de la provincia de Cáceres. Galemys, 21 (NE): 133–142.

Randrianasolo, A., Miller, J.S., Consiglio, T.K. (2002). Application of IUCN criteria and Red List categories to species of five Anacardiaceae genera in Madagascar. Biodiversity and Conservation, 11: 1289–1300.

Rapoport, E.H. (1982). Areography. Geographic Strategies of Species. Trad. B. Drausal, Pergamon Press, Oxford. ISBN 978-0-08-028914-4.

Rocchini, D., Hortal, J., Lengyel, S., Chiarucci A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography 35, 211–226.

Rodrigues, J.F.M., Olalla-Tárraga, M.Á., Iverson, J.B., Akre, T.S.B., Diniz-Filho, J.A.F. (2017). Time and environment explain the current richness distribution of non-marine turtles worldwide. Ecography, 40: 1402–1411.

Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M. (2006). The value of the IUCN Red List for conservation. Trends in Ecology & Evolution 21: 71–76.

Rodriguez, J.P., Rodriguez-Clark, K., Baillie, J.E., Ash, N., Benson, J., et al. (2011). Establishing IUCN Red List Criteria for Threatened Ecosystems. Conservation Biology 25: 21–29.

Román, J. (2010). Manual de campo para un sondeo de rata de agua ("Arvicola sapidus"). SECEM.

Rondinini, C., Di Marco, M., Chiozza, F., Santulli, G., Baisero, D., Visconti, P., Hoffmann, M., Schipper, J., Stuart, S.N., Tognelli, M.F., Amori, G., Falcucci, A., Maiorano, L., Boitani L. (2011). Global habitat suitability models of terrestrial mammals. Philos Trans R Soc Lond B Biol Sci 366(1578): 2633–2641.

Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H., Possingham, H.P. (2006). Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology Letters 9: 1136–1145.

Roselaar, C.S., Sluys, R., Aliabadian, M., Mekenkamp, P.G.M. (2007). Geographic patterns in the distribution of Palaearctic songbirds. Journal of Ornithology, 148: 271–280.

Sampaio, R., Munaril, D.P., Röhe, F., Ravetta, A.L., Rubim, P., Farias, I.P., da Silva, M.N.F., Cohn-Haft, M. (2010). New distribution limits of *Bassaricyon alleni* Thomas 1880 and insights on an overlooked species in the Western Brazilian Amazon. Mammalia 74: 323–327.

Sandel, B., Arge, L., Dalsgaard, B., Davies, R.G., Gaston, K.J., Sutherland, W.J., Svenning, J.-C. (2011). The influence of Late Quaternary climate-change velocity on species endemism. Science, 334: 660–664.

Scheick, B.K., McCown, W. (2014). Geographic distribution of American black bears in North America. Ursus 25: 24–33.

Schipper, J. et al. (2008).The status of the world's land and marine mammals: diversity, threat, and knowledge. Science 322: 225–230.

Schroeder, M.A., et al. (2004). Distribution of Sage-Grouse in North America. Condor 106: 363–376.

Sexton, J.P., McIntyre, P.J., Angert, A.L., Rice, K.J. (2009). Evolution and ecology of species range limits. Annual Review of Ecology, Evolution, and Systematics, 40: 415–436.

Sillero, N., et al. 2014. Updated distribution and biogeography of amphibians and reptiles of Europe. Amphibia-Reptilia, 35(1):1–31

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

Soorae, P., et al. (2013). Distribution and ecology of the Arabian and Dhofar Toads (Duttaphrynus arabicus and D. dhufarensis) in the United Arab Emirates and adjacent areas of northern Oman. Zool. Middle East., 59, pp. 229–234.

Sousa-Baena, M.S., Garcia, L.C., Peterson, A.T. (2014). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory, in: Lluís Brotons (Ed.), Diversity and Distributions 20 (14), pp. 369–381.

Steiniger, S., Hunter, A.J.S. (2013). A scaled line-based kernel density estimator for the retrieval of utilization distributions and home ranges from gps movement tracks. Ecol Inform., 13:1–8.

Stephenson, P.J., Brooks, T.M., Butchart, S.H.M., Fegraus, E., Geller, G.N., Hoft, R., Hutton, J., Kingston, N., Long, B., McRae, L. (2017). Priorities for big biodiversity data. Frontiers in Ecology and the Environment 15, 124–125.

Stone, O.M., Laffan, S.W., Curnoe, D., Rushworth, I., Herries, A.I. (2012). Distribution and population estimate for the chacma baboon (Papio ursinus) in KwaZulu-Natal, South Africa. Primates, 53: 337–344.

Taulman, J.F., Robbins, L.W. (1996). Recent range expansion and distributional limits of the nine-banded armadillo (Dasypus novemncinctus) in the United States. Journal of Biogeography, 23: 635–648.

Taulman, J.F., Robbins, L.W. (2014). Range expansion and distributional limits of the nine-banded armadillo in the United States: an update of Taulman & Robbins (1996). Journal of Biogeography, 41: 1626–1630.

Tiago, P., Pereira, H.M., Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. Basic and Applied Ecology.

Tizzani, P., et al. (2013). Recent distribution of red-legged partridge Alectoris rufa in Piedmont (North Western Italy): Signs of recent spreading. Avocetta 37:83-86.

Thorn, M., Green, M., Keith, M., Marnewick, K., Bateman, P.W., Cameron, E.Z., Scott, D.M. (2011). Large-scale distribution patterns of carnivores in northern South Africa: implications for conservation and monitoring. Oryx 45(4): 579–586.

Tobias, J.A., Brightsmith, D.J. (2007). Distribution, ecology and conservation status of the Blue-headed Macaw Primolius couloni. Biological Conservation 139: 126–138.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F.(2017). Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7, 9132.

Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K., Wilson, K.A. (2013). To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. Diversity and Distributions 19, 465–80.

Uhrin M., Benda P., Obuch J., Danko Š. (2008). Lesser Mouseeared Bat (*Myotis blythii*) in Slovakia: distributional status with notes on its biology and ecology (Chiroptera: Vespertilionidae). Lynx n. s. 39: 153–190.

Utrio, P. (1979). Geographic distribution of mosquitoes (Diptera, Culicidae) in eastern Fennoscandia. Notulae Entomologicae, 59: 105–123.

Van Landuyt, W., Vanhecke, L., Hoste, I., Hendrickx, F., Bauwens, D. (2008). Changes in the distribution area of vascular plants in Flanders (northern Belgium): eutrophication as a major driving force. Biodiversity and Conservation, 17, 3045–3060.

van Moll, G. (2005). Distribution of the badger (Meles meles L.) in the Netherlands; changes between 1995 and 2001. Lutra 48 (1): 3-34.

Vandel, J.M., Stahl, P. (2005). Distribution trend of the Eurasian *Lynx lynx* populations in France. Mammalia 69: 145–158.

Vanderwerf, E.A., Lohr, M.T., Titmus, A.J., Taylor, P.E., Burt, M.D. (2013). Current distribution and abundance of the O‘ahu ‘Elepaio (Chasiempis ibidis). Wilson J Ornithol, 125(3): 600–8.

Vasconcelos, R., Brito, J.C., Carranza, S., Harris, D.J. (2013). Review of the distribution and conservation status of the terrestrial reptiles of the Cape Verde Islands. Oryx 47: 77–87.

Venter, O., Fuller, R.A., Segan, D.B., Carwardine, J., Brooks, T., Butchart S.H.M., Di Marco, M., Iwamura, T., Joseph, L., O’Grady D. (2014). Targeting Global Protected Area Expansion for Imperiled Biodiversity. PLoS Biology 12, e1001891.

Vicente-Arranz, J.C., et al. (2013). Distribution and phenology of the species of family Zygaenidae Latreille, 1809 in the province of Avila (Spain) (Lepidoptera: Zygaenidae). SHILAP 41(161): 113-127.

Visconti, P., Di Marco, M., Álvarez-Romero, J.G., Januchowski-Hartley, S.R., Pressey, R.L., Weeks, R., Rondinini, C. (2013) Effects of errors and gaps in spatial data sets on assessment of conservation progress. Conservation Biology, 27: 1000–1010.

Watson, J.E.M., Dudley, N., Segan, D.B., Hockings, M. (2014). The performance and potential of protected areas. Nature 515, 67–73.

Westgate, M.J., Likens, G.E., Lindenmayer, D.B. (2013). Adaptive management of biological systems: a review. Biol. Conserv., 158: 128-139.

Wibisono, H.T., Pusparini, W. (2010). Sumatran tiger (*Panthera Tigris Sumatrae*): A review of conservation status. Integrative Zoology 5: 309–318.

Wilson, K.A., Underwood E.C., Morrison S.A. et al. (2007). Conserving biodiversity efficiently: what to do, where, and when. PLoS Biol 5, e223.

Worton, B.J., (1989). Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70: 164–168.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., et al. (2007). How global is the global biodiversity information facility? PLoS ONE 2: e1124.

Yin, D., He, F. (2014). A simple method for estimating species abundance from occurrence maps. Methods in Ecology and Evolution 5: 336–343.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA