



**HAL**  
open science

# Biodiversity occurrences and patterns from the angle of systematics

Julien Troudet

► **To cite this version:**

Julien Troudet. Biodiversity occurrences and patterns from the angle of systematics. Biodiversity. Université Pierre et Marie Curie - Paris VI, 2017. English. NNT : 2017PA066568 . tel-01901059

**HAL Id: tel-01901059**

**<https://theses.hal.science/tel-01901059v1>**

Submitted on 22 Oct 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Université Pierre et Marie Curie

Ecole doctorale 227

*Institut de Systématique, Évolution, Biodiversité (ISYEB)*

*Equipe Evolution fonctionnelle et Systématique (EVOFONCT)*

*Laboratoire Informatique and Systématique (LIS)*

## BIODIVERSITY OCCURRENCES AND PATTERNS FROM THE ANGLE OF SYSTEMATICS

Par Julien TROUDET

Thèse de doctorat de Sciences de la Vie

Dirigée par Régine Vignes-Lebbe et Frédéric Legendre

Présentée et soutenue publiquement le 29/11/2017

Devant un jury composé de :

ANTONELLI Alexandre	Professor	<b>Rapporteur</b>
ARCHAMBEAU Anne-Sophie	Ingénieur de recherche	<b>Examineur</b>
LEGENDRE Frédéric	Maître de conférences	<b>Directeur de thèse</b>
LESSARD Jean-Philippe	Assistant Professor	<b>Rapporteur</b>
PAGE Rod	Professor	<b>Examineur</b>
VIGNES-LEBBE Régine	Professeur	<b>Directeur de thèse</b>





*Je dédicace cette thèse à mes parents pour leur soutien et leur amour.*



# Thanks

I would like to begin by thanking the Institut de Systématique, Évolution, Biodiversité and its director Philippe Grandcolas for welcoming me during this thesis. I am extremely grateful to my thesis directors Régine Vignes-Lebbe and Frédéric Legendre for their help, advice and fantastic involvement during these three exciting years.

Many thanks to the members of my committee and my thesis jury for their time and valuable advice: Alexandre Antonelli, Anne-Sophie Archambeau, Jean-Philippe Lessard, Rod Page, Roseli Pellens, Philippe Grandcolas, Wilfried Thuiller, Samy Gaiji and Jérôme Sueur.

Thanks to all the people I met during my thesis and with whom I exchanged ideas, questions and cheerful moments. Amandine Blin, Michel Baylac, Marie-Elise Lecoq, Sophie Pamerlon, Tim Robertson, Dmitri Schigel, Daniel Currie, Robert Ricklefs, Peter Filzmozer, Mark Judson, Tony Robillard and all members of the UMR ISYEB.

I would also like to thank all my friends and family at the museum with whom I had an exciting discussion, serious or not: Antoine, Vincent, Léo, Camille, Bruno, Maram, Marion, Laura, Gilles, Gaetan, Ninon and many others.

Finally, thank you to my parents and family for their permanent and unconditional love, as well as to all my friends whose support and presence was essential to me.

# Index

Thanks .....	3
Index .....	4
Introduction .....	7
What is Biodiversity? Why it matters? How to study it? .....	8
Biodiversity: a multi-faceted concept.....	9
Species richness, the golden standard of biodiversity measures .....	10
Why studying biodiversity?.....	10
Systematics and ecology: two complementary approaches to study biodiversity .....	11
Species occurrences, biologists' raw material.....	12
Primary biodiversity data .....	13
Datasets: From cabinet of curiosities to databases.....	14
"Big data": a change in scale and practices.....	15
Ecoinformatics, the "big data" of biodiversity .....	16
Global Biodiversity Information Facility .....	17
Data quality and bias .....	19
A global pattern of biodiversity: the latitudinal diversity gradient .....	20
Species richness varies with latitude .....	20
The multiple hypotheses behind the LDG.....	21
Questions addressed in this thesis dissertation.....	23
Can biological diversity be investigated in its entirety?.....	23
How is the practice of biodiversity data gathering evolving? .....	23
Latitudinal Diversity Gradient at large taxonomic scale: which factors shape it?.....	24
Chapter 1: Material and Methods .....	25
Using the GBIF mediated data .....	25
Datasets from the GBIF portal .....	26
The Darwin Core format .....	27
Big data in practice.....	28
Workflow architecture.....	28
Reading and filtering occurrences.....	30
Indexing the table to get a functioning database .....	33
Characterizing a biodiversity dataset: biases and trends.....	34
Species names from the GBIF Backbone Taxonomy and multimedia files.....	34
Public interest and taxonomic research quantity.....	36
Putting the GBIF database into numbers.....	37
Statistical tools .....	39
Working on biodiversity patterns: delving into ecoinformatics .....	40
Cleaning data.....	40
Estimating species richness .....	48
Using our results to understand the LDG .....	55
The species richness covariates .....	56
Statistical analysis of species richness and its covariates.....	57
Chapter 2: The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it? (Troudet et al. Systematic Biology, submitted as a Point of View).....	60

Abstract.....	61
Keywords.....	61
Introduction .....	62
Material and Methods.....	65
Data Set .....	65
Data Quantity .....	66
Data Origin .....	66
Supporting Files .....	66
Evolution of Data Completeness.....	67
Evolution of Taxonomic and Spatial Precision.....	67
Results and Discussion.....	67
A Shift in the Recording of Primary Biodiversity Data .....	67
Primary Biodiversity Data for systematics and evolutionary studies in the 21 <sup>st</sup> Century: Are We There Yet? .....	72
Acknowledgments .....	77
Chapter 3: Taxonomic bias in biodiversity data and societal preferences (Troudet <i>et al.</i> <i>Scientific reports</i> , published 22-08-2017).....	78
Abstract.....	79
Introduction .....	80
Results .....	84
Discussion.....	98
Methods .....	102
Acknowledgements .....	106
Author Contributions.....	107
Chapter 4: Latitudinal Diversity Gradient: Geometric hypotheses revisited using massive biodiversity occurrences in plants and animals of the New World (article in preparation) .....	108
Introduction .....	108
Materiel and Methods.....	110
Species richness estimates.....	110
Explanatory Variables .....	112
Statistical analyses.....	114
Results .....	115
Basic statistics .....	115
Latitudinal diversity gradient .....	115
Environmental hypotheses.....	116
Discussion.....	119
Acknowledgement.....	121
Chapter 5: DwCSP a fast biodiversity occurrence curator (article in preparation)....	122
Introduction .....	122
Software.....	124
Data Enrichment.....	124
Searching for outliers .....	125
Results .....	127
Discussion.....	128
Discussion.....	130
Big-data and biodiversity .....	130
The big-data paradigm.....	131

The genesis of Biodiversity big-data.....	133
A new way of doing science.....	138
Primary Biodiversity data, a proxy to assess the state of the study of biodiversity	141
Biodiversity data are disconnected from specimens .....	142
Taxonomic bias while aiming at investigating the whole biodiversity .....	145
Using biodiversity data to decipher the origin of global biodiversity patterns .....	153
Estimating global species richness from a large and geographically widespread taxonomic sample.....	153
Further into the Latitudinal Diversity Gradient.....	156
The GBIF-mediated data: a fascinating tool for biodiversity analyses .....	159
References .....	161
Appendixes .....	181
Appendix 1: List of used Java Libraries and Dependencies.....	181
Appendix 2: List of indexes of the OCCURRENCES table .....	182
Appendix 3: List of additional database tables.....	183
Appendix 4: VBA script for Web Search Results .....	188
Appendix 5: R script for spatial outlier detection .....	189
Appendix 6: Worldwide species richness maps and plots.....	190
List of figures: .....	199
List of tables: .....	202

# Introduction

In 1807, Alexander von Humboldt wrote:

*"Thus, the nearer we approach the tropics, the greater the increase in the variety of structure, grace of form, and mixture of colours, as also in perpetual youth and vigour of organic life."*

These words were published in « Views of nature, or, Contemplations on the sublime phenomena of creation » after his Latin American expedition, a journey that led him through at least four of the twenty-five hotspots of biodiversity identified far later (Myers *et al.* 2000). Although being a single quote from Humboldt's seminal work, this sentence outlines multiple concepts that have driven the study of life on Earth for the following centuries: the diversity of life on earth (biodiversity), the variability of this diversity across space, and its quantification. Each of these notions has been studied extensively since then and a lot more still awaits investigation.

Unfortunately, most of the habitats Humboldt visited have disappeared or are highly threatened, endangering the biodiversity hosted in these habitats. In 2000, the best-preserved habitat among those Humboldt explored had only 25% of primary vegetation remaining (Myers *et al.* 2000). The study of biodiversity is thus framed in an emergency context, which might explain why ecologists and conservation biologists have mainly grasped it. This sense of urgency led scientists to focus on peculiar taxonomic groups (Lambeck 1997) or habitats (Dixon *et al.* 1994) that were then used to model the rest of biodiversity.

Studying biodiversity is the work of systematists who, since more than 300 years, are the first providers of biodiversity data. This part of their activity resulted in accumulated data that is the groundwork upon which we try to build a better understanding of biodiversity. Less interested in the structures and biotic interactions within an ecosystem than in the biological diversity per se, systematists favour large taxonomic scale wherein each and every species counts.

In the last decades, systematists and ecologists, together with nature enthusiasts, have contributed to accelerate biodiversity data production and sharing. The biodiversity data accumulated is now considerable in volume and international efforts are at play to handle this mass of information. This massive amount of data allows studying countless biodiversity issues at various taxonomic and geographical scales. However, these online databases still contain many gaps (Hortal *et al.* 2015); show data quality issues (Gaiji *et al.* 2013, Troia and McManamay 2016) as well as important biases (Boakes *et al.* 2010, Meyer *et al.* 2016).

It is precisely on those limitations that systematics can shed a new light. The problem of data quality is often approached with ecological and conservation ulterior motives, and is consequently limited to few taxa or areas of interest. In this dissertation I strive to get a broader view on the aforementioned issues, embracing the systematists' point of view. I use the biggest biodiversity dataset available (i.e. GBIF mediated data, see chapter 1) that I consider as a decent representative sample of the global practices of biodiversity data collection. Eventually, with a better understanding of the current limitations of biodiversity data, I come back to Humbolt's bicentenary observation that biodiversity is greater in the tropics and test a few hypotheses that could explain this major biodiversity pattern.

## **What is Biodiversity? Why it matters? How to study it?**

Humboldt described the richness of the living beings as the “variety of structure, grace of form, and mixture of colours”. This concept has persisted until now and was named biological diversity (Dasmann, 1968) before being condensed into “biodiversity”. Biodiversity was first mentioned in 1985 by Walter G. Rosen, and E. O. Wilson popularized this notion in 1988. This term is now widely used in the scientific community, in public media and in many governmental entities as decision-makers grow wary of the services biodiversity provides to humans and of the impact of climate change on them (Nagoya protocol, Accords de Paris...).

The field of conservation biology has grasped the concept of biodiversity and has generated a new interest in its study at large taxonomic and geographical scales, beyond the mere study of model or charismatic organisms. Systematics, the main discipline producing biodiversity data, is obviously another important field for the study of biodiversity. However,

because of practical reasons and taxonomic issues, systematists have rarely engaged in large-scale biodiversity studies. It would yet efficiently complement the dominant ecological approach.

### **Biodiversity: a multi-faceted concept**

Biodiversity takes on many aspects and its definition is still discussed among experts (Harper and Hawksworth 1994, Holt 2006). The word ‘biodiversity’, a contraction of ‘biological diversity’, was formally defined in the United Nations Environment Programme in 1992 (p.27):

*"Biological diversity" means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic systems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems.*

This variability is visible at every level of life, from genomes to entire ecosystems, and is studied at each of these levels:

- Genetic diversity focuses on the genetic differences among individuals, populations or taxa. It is one of the lowest level at which biodiversity is studied (May and Godfrey 1994; Frankham 1995).
- Taxonomic diversity is often measured as the number of species inside a geographic area (species richness). However, it can also take into account species evenness with metrics such as Simpson or Shannon-Wiener indices (Whittaker 1972).
- Phylogenetic diversity takes into account phylogenetic differences between species when estimating biodiversity (Faith 1992).
- Functional diversity is defined as the diversity of functional group (e.g. different locomotion, different trophic level, different feeding mechanisms...) within a population or an ecosystem (Tilman *et al.* 1997).
- Ecosystem diversity studies the variation in ecosystems within a geographical location and its impact on the environment (Lapin and Barnes 1995).

In this thesis dissertation, I use species richness as a mean to estimate biodiversity and both terms are used indifferently.

### **Species richness, the golden standard of biodiversity measures**

The species richness of an area corresponds to the amount of different species in this area. Ordinarily species richness is calculated only for a taxon of interest in a specific region (e.g. avian species richness of South America). No additional computation or manipulation is needed to get this number other than enumerating the number of species. This metric was chosen here considering three main reasons. First, this is one of the simplest biodiversity measurement and “the oldest and most intuitive measure of biological diversity” (Magurran 2004). Second, species richness is widely used (Gotelli and Colwell 2001), which allows checking and comparing my results with other studies. Third, species richness was relatively easy to estimate accurately, using the data at my disposal.

However, species richness as an estimate of biodiversity is not without flaw. There are still debates about species richness and its uses (Spellerberg and Fedor 2003). At a fundamental level some scientists challenge the validity of the “species” concept, in particular for unicellular organisms (Rosselló-Mora and Amann 2001). In the same vein, the use of species richness is based on our capacity to distinguish different species, which can be quite challenging for cryptic species (McNew and Handel 2015), poorly known taxa (a.k.a. Linnean shortfall: Lomolino and Heaney 2004), or taxa with a skill deficit in taxonomy (Cardoso *et al.* 2011). Moreover, as shown in chapter 1, estimating species richness can be done using multiple methods (Bunge and Fitzpatrick 1993) and if not careful with the used tools the results can be misleading or biased (Gotelli and Colwell 2001). Estimating reliably species richness was thus a crucial first step in this work.

### **Why studying biodiversity?**

Most of biodiversity on Earth is confronted to the biggest ecological crisis in human’s history (Brooks *et al.* 2002, Ceballos *et al.* 2015). Human activities destroy entire ecosystems and endanger many species. The leading causes of biodiversity loss are the alteration or destruction of habitats (Margono *et al.* 2014, Haddad *et al.* 2015), importation of invasive



species (Powell *et al.* 2013, Chase and Knight 2013, Katsanevakis *et al.* 2014), pollution (Edinger *et al.* 1998, Vörösmarty *et al.* 2010) and climate change (Thomas *et al.* 2004).

In this worrying context, conservation issues justify the need to study and protect biodiversity. Scientific studies allow a better understanding of the causes and consequences of species extinction and could help mitigating species losses. For example, a better knowledge of ecosystems and their inhabiting species can help to define priority areas for conservation (Myers *et al.* 2000) or to target priority taxa to protect (Lambeck 1997).

On a more academic level, scientists are still working towards a better understanding of our environment. It is estimated that only 14% of terrestrial species and 9% of oceanic species have been discovered by scientists (Mora *et al.* 2011). Those 1.2 million species have taken more than two centuries to be described, meaning that's there is still a colossal gap in our knowledge about life on Earth we need to fill.

Finally, incentives directly impacting human well-being or with economic repercussions are the most concrete to the general public: discovery of new genes and new molecules for the pharmaceutical and GMO industries, control of pathogen vector and crop pests (Cardinale *et al.* 2012), pollination, etc. These outcomes, known as the goods and services provided by biodiversity, are often cited to promote biodiversity study and conservation.

### **Systematics and ecology: two complementary approaches to study biodiversity**

Whereas ecology studies the relationships and processes occurring among organisms and between organisms and their environment, taxonomy describes defines and name taxonomic units, and systematics informs on species relationships and historic diversification (phylogeny). If ecology has gained a tremendous popularity in the last decades (Kormondy 2012), it has not been the same for taxonomy. Many taxonomists deem it necessary to alert about the state of taxonomy (e.g. Rodrigues *et al.* 2010, Costello *et al.* 2013, Vogel Ely *et al.* 2017) and about the considerable workforce still needed in that domain (Lomolino 2004). This situation is worrisome, all the more so that biologists work on biodiversity data mostly produced by systematicists and taxonomists.

This divide between systematists (including taxonomists) and ecologists is old and regularly brought up (e.g. Hagen 1986, Nielsen *et al.* 1998, Bortolus 2008). Differences in objectives and methods drove this fracture, ecology advocating for a more experimental approach and a focus on processes and populations while systematics was more descriptive and interested in individuals and taxa. Both sides had compelling arguments sometimes leading to tensions between the two disciplines (Hagen 1986).

Nowadays the ecology-systematics conflict seems to keep diminishing so that combined approaches are expected to increase (Gotelli 2004, Bortolus 2008). The collaboration between ecology and systematics has proven efficient with the use of taxonomy in community ecology (Gotelli 2004) or the use of systematics measures like phylogenetic diversity (Faith 1992) on conservation and ecological studies. Similarly, biogeography, which studies the distribution of biodiversity over space and time (Brown and Lomolino 1998), is grounded in systematics but uses tools from ecology to infer species interactions with other species or the environment. Biogeography extended at large taxonomic and spatial scales gave birth to macroecology (Brown 1995), aiming notably at studying how the historical distribution of species can impact the actual patterns of distribution of those species.

In fact, the study of the latitudinal diversity gradient and global patterns of biodiversity in general, is a hallmark of macroecology (Beck *et al.* 2012). Global patterns have been mainly studied by ecology scholars (e.g. Chown and Gaston 2000, Allen and Gilgoly 2006, Condamine *et al.* 2012, Pereira 2016). With a formation in systematics and housed in the Institut de Systematique, Evolution, Biodiversité, I tackled this issue from a systematist angle, trying to complement the more ecological current vision. This naturally led me to analyse the data at my disposal once again at a global spatial and taxonomic scale.

## **Species occurrences, biologists' raw material**

In the opening citation of this introduction, Humboldt loosely quantified biodiversity: “the variety of structure, grace of form, and mixture of colours” (emphasis mine). He also mentioned an increase of this variety, suggesting that he compared biodiversity from different locations. Of course, Humboldt had far less tools than scientists have today to quantify

biodiversity, lacking both conceptual knowledge, taxonomy being in its infancy, and concrete information about where occur the different species.

As more and more people studied biodiversity, the data about life on Earth started to accumulate. Scientists recorded and shared part of these data, which mostly came in the form of specimen occurrences: an identified specimen observed or collected at a certain time and place (Johnson 2007).

### **Primary biodiversity data**

Nearly all the results exposed in this manuscript were based on Primary Biodiversity Data. This type of data can be defined as:

*Digital text or multimedia data record detailing facts about the instance of occurrence of an organism, i.e. on the What, Where, When, How, and By Whom of the occurrence and the recording (Borchsenius 2012).*

The digital nature of a primary biodiversity data is debatable, as a lot of data are still only available on paper (Boakes *et al.* 2010). However, I used in this project only digitized data, which fits with this definition. Here, I use primary biodiversity data to refer to multiple records or entire datasets, whereas I use species occurrence to allude to single records. In other words, primary biodiversity data consist of multiple species occurrences.

A species occurrence contains diverse pieces of information but three of them are essential:

- the name of the organism, preferably the complete scientific name (what?),
- the location of the observation, often latitude and longitude (where?)
- the date of the observation (when?).

Other pieces of information have been used in some analyses; they are defined and explained in chapter 1.

## **Datasets: From cabinet of curiosities to databases**

The early history of the study of biodiversity saw the emergence of cabinet of curiosities. The nobles and merchants started to accumulate exotic objects and specimens, often related to natural history, as early as the 16th century (Ferro and Flick 2015). Those heterogeneous collections of items were the precursors of the museums, which gained in popularity following a more rigorous practice of science. The amount of collected items kept growing, supplied by the increasing numbers of explorers and naturalists (Beaman and Cellinese 2012). Today, biologists are still using the data collected and preserved in the museums along with online shared data.

In recent years, the democratization of informatics and, more importantly, the creation of internet allowed scientists to share their own databases and to use the data collected by their colleagues and predecessors. Informatics also gave access to new tools allowing powerful statistical tests: new data visualization, complex processes, relationships modeling and so on. Importantly, aside from the recently collected data that is now routinely added to databases, natural history collection data are also digitalized and added to shared databases.

The omnipresence of informatics for the investigation of biodiversity has obvious advantages, starting with data accessibility. It took 1,500 person-day of gathering for Boakes *et al.* (2010) to find nearly 150 000 usable Galliformes records disseminated in multiple museums, collections, databases and articles. Those numbers are dwarfed by the more than 3 million georeferenced occurrences of Galliformes available in a few seconds on the Global Biodiversity Information Facility portal ([gbif.org](http://gbif.org), consulted 13/02/2017).

However, this evolution is neither complete nor flawless. First, the vast majority of collections are not digitalized. According to Ariño (2010), only 3%, at best, of the 1.2 to 2.1 billion of the specimens deposited in natural history collections are accessible through the biggest biodiversity data portal. Despite working on a popular bird taxon, Boakes *et al.* (2010) still got 24% of their dataset from non-digitalized museum collections, the rest coming from online databases, publications and books. Second, the shift from natural history collection data toward digital databases motivated the production of large quantities of observational

data. This type of data is rarely linked to any material evidence (DNA, material sample) and cannot be verified a posteriori.

Such changes in the way data are created, stored and shared have a heavy impact on the fields of ecology and systematics. In chapter 3, I explore in more details this evolution and its potential consequences for the future of biodiversity studies.

### **“Big data”: a change in scale and practices**

Humboldt wrote in 1807 that “the nearer [he] approach[es] the tropics, the greater the increase in [biodiversity]”. Nearly two centuries later, Platnick (1991) devised about which part of the world had the richest species diversity. His fieldwork in south temperate areas led him thinking that those areas are more diverse than what thought his contemporaries. Both Humboldt and Platnick relied mostly on personal experience and observations to make these hypotheses. Their situation was the norm for most of the history of biology, meaning before the advent on Internet when data retrieval and sharing involved much more considerable effort and time than today (Ferro and Flick 2015). It is only in the last two or three decades that policies and technical advances revolutionized the way biodiversity data is created, maintained, distributed and used (Soberón and Peterson 2004).

In the last decade, accessing a large amount of biodiversity data has become the norm: the era of big data is upon us (Hampton *et al.* 2013). However reducing this evolution to a simple increase in the quantity of available data would be a narrow-minded view. The emergence of big data in biodiversity science influence the entire field, changing how research is conducted (Kelling *et al.* 2009, Kitchin 2014) and what are its end-goals (Devictor and Bensaude-Vincent 2016). These changes are already noticeable as the number of Citizen Science projects increase (Miller-Rushing *et al.* 2012) as well as the number of biodiversity studies performed without new fieldwork (Hampton *et al.* 2013, Rosenheim and Gratton 2017).

My PhD project relied on three aspects of the big-data evolution. Firstly, using a large dataset enabled me to augment the scope of my studies, both at the taxonomic and spatial levels. Even some poorly known taxa feature thousands of occurrences in the publicly

available database and these occurrences cover a large part of Earth surface, although unevenly (Meyer *et al.* 2015). Secondly, working at a broad taxonomical scope allows me getting generalized results or identifying taxonomic exceptions. Thirdly, the large amount of data is also an asset when using statistical tools, increasing their robustness.

### **Ecoinformatics, the “big data” of biodiversity**

The democratization of informatics led to many developments in society and science, among which the large-scale, computer-aided management of biodiversity data: biodiversity informatics (Canhos *et al.* 2004, Peterson *et al.* 2010). Biodiversity informatics is used within the scope of many scientific fields such as macroecology (Miraldo *et al.* 2016), community ecology (Cardillo 2011), biogeography (Devictor *et al.* 2010) and so on. A subfield of biodiversity informatics applied to ecology was called ecoinformatics. Michener and Jones defined ecoinformatics in 2012 as follows: *Ecoinformatics is a framework that enables scientists to generate new knowledge through innovative tools and approaches for discovering, managing, integrating, analyzing, visualizing and preserving relevant biological, environmental, and socioeconomic data and information.*

This definition is subject to debate and Rosenheim and Gratton (2017) provided a simpler version: “*Ecoinformatics [...] refers to ecological studies that use pre-existing data*”. In their article, they also emphasized the “big data” aspect of ecoinformatics. As a central point of my work is the use of publicly available biodiversity data, I use this definition.

Ecoinformatics, as a new field to be explored, shows some very interesting promises. First, there are many advantages to reuse biodiversity data. As underlined earlier, some studies can still be done without the obligation to collect new data if suited data already exist and is available, which saves time and reduces costs. Sharing newly produced data further facilitates this advantage. Using large dataset also benefits the confidence in results as it comes with greater statistical robustness (Benjamin *et al.* 2017; Rosenheim and Gratton 2017). And of course, a larger amount of data allows broadening the scope of the studies as well as repeatability of the analyses by others.

Ecoinformatics also presents many challenges, the most prominent one being the difficulty to manipulate large datasets, which requires adequate computer power (Kumar and

Kumar 2016). As a data-intensive science, and like its name suggests, Ecoinformatics relies heavily on informatics and is therefore limited by hardware and software capabilities. For instance, the raw volume of data used in my project was a limitation as it is too large to handle for common tools, such as R and other statistical software. This difficulty is actually a part of the “big-data” definition. Handling a dataset of >600 million occurrences, I had to come up with a workflow composed of algorithms and programs specifically designed to analyze it. Thus, programming time occupied a large part of my PhD and is discussed in chapter 1.

Beyond technical difficulties, working on pre-existing data means partly losing the control over the collection processes, which can be aggravated by the high heterogeneity typically encountered in large databases (Rosenheim and Gratton 2017). This heterogeneity and the volume of data result from the pooling of diverse sources of data. In addition, replicating studies “big data” is often difficult because no comparable dataset is available (or it would have been included in the former dataset). Without replicates or a control over data collection, it can also be harder to make causal inferences as the sampling could not allow testing for the influence of a specific variable (and re-sampling is not conceivable).

Eliminating those obstacles requires work during the whole data life cycle, from its collect to its use in a publication (Michener and Jones 2012). Scientists must change how they perceive data. They must treat data as a product of research and not only as a step towards publication (Hampton *et al.* 2013), and spend as much effort on data curation as on data gathering (Howe *et al.* 2008). The recent initiatives linking multiple databases and types of data (DNA, occurrences, taxonomy...) and emphasizing the need for well-curated data suggest that this challenge is about to be taken care of (Hampton *et al.* 2013, Bingham *et al.* 2017).

### **Global Biodiversity Information Facility**

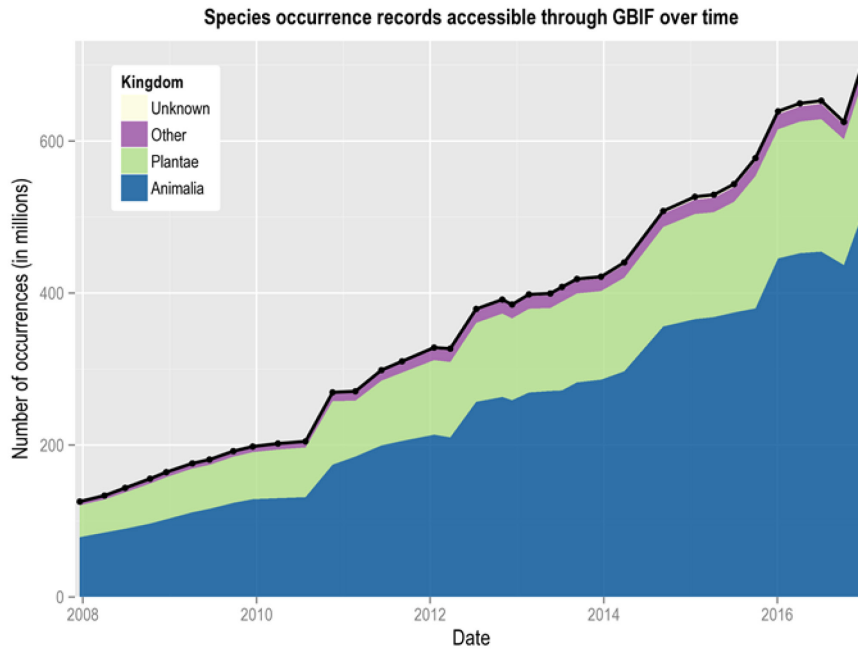
The Global Biodiversity Information Facility (GBIF) is an open data infrastructure funded by governments which aims to make biodiversity information “freely and universally available for science, society and a sustainable future” (gbif.org). This structure provides a central, well-known repository, with standardized data formats, allowing scientists and

institutions to share their data. The specifics of the GBIF infrastructure and data are covered in chapter 1. Briefly, the GBIF gives access to more than 700 million species occurrences and the usefulness of these data has been assessed multiple times (e.g. Beck *et al.* 2013, García-Roselló *et al.* 2015, Ferro and Flick 2015). It is also used as a tool to raise awareness about biodiversity as it provides global information to the public as well as the political actors (Devictor and Bensaude-Vincent 2016).

The GBIF is used here as the primary data provider; it is not, however, the only data source available and there is a vast set of infrastructures sharing biological data. This set comprises small-scale infrastructures such as the repositories of monitoring programs (e.g. Zárbynická *et al.* 2017) or repositories dedicated to a restricted area or taxa (e.g. Cameron *et al.* 2016, Martin and Harvey 2017) but also large-scale data portal associated to Citizen Science programs (e.g. ebird.org) or international structures (e.g. catalogueoflife.org). The diversity of sizes in data portals also come with a diversity of purposes and type of data (e.g. DRYAD, GenBank, DataONE, Map Of Life...).

The GBIF data portal was chosen because it is the biggest open access dataset of primary biodiversity data (Gaiji *et al.* 2013). It is also well known, widely used and provides data in a standardized format. Additional reasons are presented in chapter 1.





**Figure 1: Growth of available species occurrences in the GBIF from 2008 to 2017.**  
(gbif.org accessed on 14/02/2017).

Millions of data are shared each year through the GBIF organization (figure 1). Those data are gathered by professional scientists or people ranging from near professional specialist to naïve and untrained observer in the case of Citizen Sciences. Consequently, the GBIF mediated data are very heterogeneous so that doubts about their quality have been repeatedly reported (e.g. Yesson *et al.* 2007).

### Data quality and bias

As previously underlined, big data does not mean perfect data. Like for every dataset, a critical assessment of the dataset quality and flaw must be done. In this respect, the systematicists' point of view is particularly useful since systematists have formulated criticisms about data in natural history collections (Goodwin *et al.* 2015), gathered through Citizen Science programs (Kosmala *et al.* 2016) or about the tendency to rely more often on digital data (Dubois 2017).

Because the GBIF aggregates diverse datasets, any issue raised about these specific types of data must be tackled. But, there are also additional specific issues, related to the

heterogeneous nature of the GBIF mediated, adding another layer of complexity in the critical data quality procedure. Many studies have tried to assess the quality of the GBIF mediated data and, despite a few discrepancies, they all confirm that GBIF mediated data are not perfect (e.g. Boakes *et al.* 2010, Gaiji *et al.* 2013, Otegui *et al.* 2013, Ferro and Flick 2015, Sikes *et al.* 2016). Incomplete data, duplicate data and other low quality data require checking data quality through filtering and cleaning processes, as explained in chapter 2.

However some issues are not solvable using a simple quality check, in particular the issue involving the composition of the whole dataset. I explore such an issue, the taxonomic bias, in chapter 4. This bias arises when some taxa are consistently treated differently than others: taxa more collected than others, or more studied than others. Some author already pointed out this bias in conservation (Donaldson *et al.* 2016), ethology (Rosenthal *et al.* 2017) and biology in general (Bonnet *et al.* 2002). Here the size and broad taxonomic coverage of the GBIF mediated data is a critical asset. This dataset can be considered as one of the best sample of the global practices of data collection in biodiversity and ecology science.

## **A global pattern of biodiversity: the latitudinal diversity gradient**

Humboldt was among the firsts to reckon what has been later called the latitudinal diversity gradient when he wrote: “the nearer we approach the tropics, the greater the increase in [biodiversity]”. He realized that biodiversity is not distributed uniformly across space and that a higher diversity is observed in tropical regions. Since then, the latitudinal gradient of diversity has been extensively studied event though it remains to be fully understood.

### **Species richness varies with latitude**

Species richness varies greatly across the globe (Pianka 1966, Currie 1991). Some areas are relatively poor in species richness, while others are richer than most other areas on Earth and called hot-spots of biodiversity (Myers *et al.* 2000). Those variations are seen at all spatial scales (Whittaker *et al.* 2001) but I will focus on the largest ones. Those large-scale patterns have aroused interest as early as the 18th century (Ricklefs 2004) and an entire area of ecology called ‘macroecology’ is now dedicated to their study (Brown 1995). Multiple hypotheses have been proposed to explain these patterns (Currie 1991, Willig *et al.* 2003).

One particular pattern, the Latitudinal Diversity Gradient (LDG), has monopolized biogeographers' attention for the last two centuries and is one of the most pervasive patterns of biodiversity (Willig *et al.* 2003).

The LDG has been found at multiple levels of biodiversity, using species richness (Gaston and Blackburn 2008), genetic diversity (Miraldo *et al.* 2016) or functional diversity (Steven *et al.* 2003) to characterize biodiversity. Moreover, the LDG has been tested using a broad range of taxa, from bacteria (Adam *et al.* 2016) and vascular plants (Kreft and Jetz, 2007), to tetrapods (Marin and Hedges 2016), marine bivalves (Jablonski *et al.* 2017), butterflies (Condamine *et al.* 2012) and many others (Willig *et al.* 2003).

### **The multiple hypotheses behind the LDG**

The LDG has been acknowledged a long time ago and biologists rapidly aimed at understanding the mechanisms responsible for this pattern. Numerous works and hypotheses were produced to explain the LDG (Willig *et al.* 2003). Incidentally, Humboldt himself proposed a mechanism explaining the higher richness of the tropics (figure 2): he assumed that winter temperatures (freezing) affect the development of high latitude species and was a too severe constraint for many organisms to thrive.

But although life is everywhere diffused, and although the organic forces are incessantly at work in combining into new forms those elements which have been liberated by death; yet this fulness of life and its renovation differ according to difference of climate. Nature undergoes a periodic stagnation in the frigid zones; for fluidity is essential to life. Animals and plants, excepting indeed mosses and other Cryptogamia, here remain many months buried in a winter sleep. Over a great portion of the earth, therefore, only those organic forms are capable of full development, which have the property of resisting any considerable abstraction of heat, or those which, destitute of leaf-organs, can sustain a protracted interruption of their vital functions. Thus, the nearer we approach the tropics, the greater the increase in variety of structure, grace of form, and mixture of colours, as also in perpetual youth and vigour of organic life.

**Figure 2: Extract from Views of nature, or, Contemplations on the sublime phenomena of creation (P. 215)**

Most of the hypotheses formulated and tested on the LDG have been reviewed in Mittelbach *et al.* (2007) and some of them have received more attention than others. Lately, they have been classified into two broad categories by Jablonski *et al.* (2017): the first category regroups *in situ* hypotheses that focus on the capacity of the local environment to support a certain amount of species (carrying capacity) while the second category contains the historical hypotheses based on spatial and evolutionary processes. In the first category, the prevalent hypotheses are the ambient energy hypothesis (Hawkins *et al.* 2003a), the productivity hypothesis (Hawkins *et al.* 2003b), and the water availability hypothesis (Hawkins *et al.* 2003a). In the second category, speciation, extinction and migrations are of considerable importance and are considered in the most widespread hypotheses such as the tropical niche conservatism hypothesis (Wiens and Donoghue 2004), the climate stability hypothesis (Pianka 1966), and the tropics as a cradle/museum hypotheses (Chown and Gaston 2000).

Historical hypotheses were not tested here, as evolutionary and historical information at a so large taxonomic scale was hardly tractable. However, a third category of hypotheses,

consisting of geometrical hypotheses that were formulated by Colwell and Hurt (1994) and more recently by Gross and Snyder-Beattie (2016), was tested.

Overall, I tested multiple hypotheses, at a global scale and on large and different taxonomic groups. The results of those tests can be found in chapter 4.

## **Questions addressed in this thesis dissertation**

### **Can biological diversity be investigated in its entirety?**

In the current context of global changes and accelerated biodiversity loss, it is of critical importance to study biodiversity in its entirety. Because some taxa might be impacted differently than others by those changes, studies should embrace a large taxonomic scale to achieve generality. That is why the existence and dangers of taxonomic gaps in biology have been emphasized long ago (Stork 1988) and is still a predominant concern across multiple fields (Feeley *et al.* 2016, Oliveira *et al.* 2016, Rosenthal *et al.* 2017). Being advertised over a long period (Clark and May 2002a, Di Marco *et al.* 2017), has taxonomic bias receded lately? Does this bias extend to primary biodiversity data or is it restricted to a few biology fields? If generalized, what are the dangers and how can we reduce this bias?

### **How is the practice of biodiversity data gathering evolving?**

The emergence of informatics and internet brought many changes in the way we do science. Among those changes the shift from “traditional” museum collection toward online databases has a collateral effect. More and more data are produced without being linked to a specimen (observation data) and some collection data are losing the connection to a specimen when digitalized (the digital occurrence has no clear link to the original material). This change deteriorates the link between data and specimens and poses new curation challenges (Howe *et al.* 2008) as well as data quality concern (Santos and Branco 2012). Moreover this change could also be a cause or a consequence of the taxonomic bias. As the data curated and gathered today will be used by the generations to come and we have to look critically at this trend to ensure a maximal usefulness of biodiversity data now and in the future.

## **Latitudinal Diversity Gradient at large taxonomic scale: which factors shape it?**

The Latitudinal Diversity Gradient is a well-known biodiversity pattern. It has been studied multiple times and a few exceptions have been reported (Willig *et al.* 2003). Still, the question at hand is not so much on the LDG existence than on the forces that gave shape to this pattern. More than thirty hypotheses have been formulated to explain its formation and even though some are seen as more likely than others, no consensus has been found (Willig *et al.* 2003). Moreover, when tested, these hypotheses are often restricted on taxonomic or geographical scale (e.g. Albuquerque and Beier 2015, Herk *et al.* 2016, Hanly *et al.* 2017). We then aimed at testing six of these hypotheses using the largest possible taxonomic and geographical scales using GBIF data.

# Chapter 1: Material and Methods

## Using the GBIF mediated data

As previously stated, the GBIF (Global Biodiversity Information Facility) is an international consortium, funded by governments, whose mission is to allow people to share and have access to biodiversity data. In April 2017, the GBIF hosts more than 700 million occurrences, coming from the accumulation of 30,712 datasets shared by 862 institutions and organisms (information available online at: <http://www.gbif.org/resource/81771>). The GBIF constitutes an incredible source of biodiversity data and was the main provider of data during my PhD.

**Green box 1:** In this chapter I will use a single species, the western green lizard *Lacerta bilineata* (image below), to explain most steps of the analysis pipeline. These green boxes will not be referred to in the main text but will serve as additional illustrations.



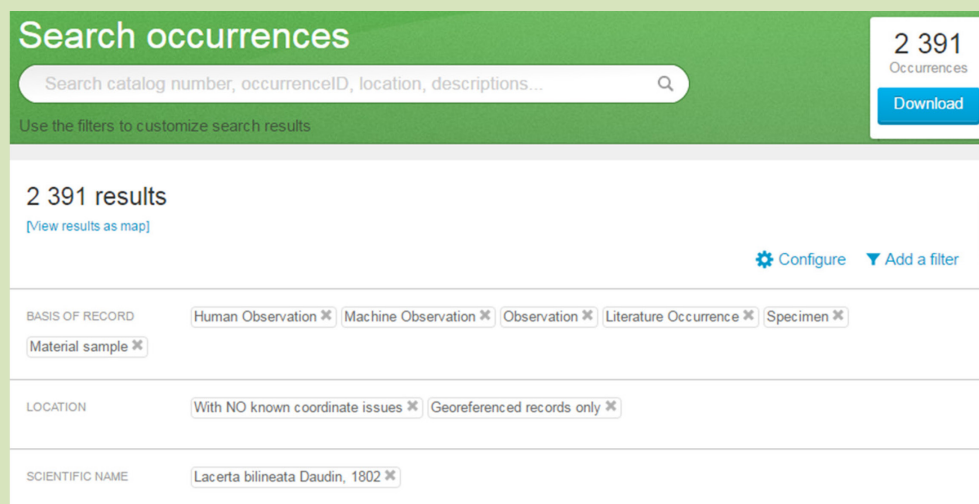
Male specimen of *Lacerta bilineata* (photo from Wikimedia Commons, Se90)

## Datasets from the GBIF portal

I downloaded multiple datasets through the GBIF portal ([www.gbif.org](http://www.gbif.org)), although most of them were used for tests and preliminary analyses during the pipeline set-up. Eventually, almost all analyses were performed on three datasets. The first one includes all GBIF mediated data. It was downloaded on the 7<sup>th</sup> of June 2016 without the use of any filter and comprises 649 767 741 occurrences (<http://doi.org/10.15468/dl.hqesx6>). This dataset was mainly used to characterize the practice of biodiversity data providers, an issue tackled in Chapters 3 and 4. The second one corresponds to the GBIF Taxonomic Backbone. It is a comprehensive dataset of all the taxonomic names used in the GBIF (<http://doi.org/10.15468/39omei>), assembled from 54 taxonomic sources, including the Catalog of Life ([www.catalogueoflife.org](http://www.catalogueoflife.org)). This dataset was mostly used to investigate the taxonomic bias in biodiversity data (chapter 4). The third dataset gathers 547 321 920 occurrences of georeferenced GBIF data (<http://doi.org/10.15468/dl.9goauq>). A filter was used to exclude occurrences flagged in the GBIF as having geospatial issues (`has_geospatial_issue=false`). Occurrences not georeferenced, fossil and living specimens (from zoo, farms, gardens...) were also excluded. This dataset was used to investigate large scale biodiversity patterns (chapter 5). All the datasets downloaded were available in the Darwin Core Archive format.



**Green box 2:** This screenshot shows how 2391 occurrences of *Lacerta bilineata* have been selected using filters. Georeferenced occurrences and occurrences without coordinate issues have been selected (“Location”), meaning that occurrences without coordinates or with coordinate issues have been filtered out. Similarly, fossil, living specimen and unknown origin occurrences have been excluded (“Basis of record”).



The screenshot displays the GBIF search interface for *Lacerta bilineata*. At the top, a green header contains the text "Search occurrences" and a search bar with the placeholder "Search catalog number, occurrenceID, location, descriptions...". To the right of the search bar, a box indicates "2 391 Occurrences" and a "Download" button. Below the header, the text "Use the filters to customize search results" is visible. The main content area shows "2 391 results" with a link to "[View results as map]". On the right side of this area, there are links for "Configure" and "Add a filter". The filter section is divided into three categories: "BASIS OF RECORD" with filters for "Human Observation", "Machine Observation", "Observation", "Literature Occurrence", "Specimen", and "Material sample"; "LOCATION" with filters for "With NO known coordinate issues" and "Georeferenced records only"; and "SCIENTIFIC NAME" with a filter for "Lacerta bilineata Daudin, 1802".

Screenshot of the filtering step on the GBIF portal ([www.gbif.org](http://www.gbif.org)).

## The Darwin Core format

Datasets can be downloaded from the GBIF portal in two file formats: as a simple tabular CSV file that keeps only essential information or as a Darwin Core Archive file (DwC-A). We chose the DwC-A format for its higher information content and because of the following advantages.

First, the DwC-A is a specific archive file format, based on the Darwin Core standards, maintained by the Biodiversity Information Standards group (<http://www.tdwg.org/>). Those standards were created to facilitate the sharing of biodiversity data and metadata. They include a glossary of definitions and standardized terms used to describe the data, along with examples. The Darwin Core supports information about taxa and their occurrence in nature, also called Primary Biodiversity Data (PBData), as well as related information. Second, the DwC-A is an archive file, containing a core dataset file and multiple

supporting files, such as description files, metadata files, etc. In the GBIF, the core file, named `occurrence.txt`, holds information about PBData. Some supporting files were also used such as the file linking Species Occurrences (SpOcc) to multimedia files such as photos or videos.

The main inconvenient of the DwC-A format is the difficulty to use it when a large quantity of data is involved. The complete datasets downloaded from the GBIF had more than 500 million occurrences requiring a significant amount of disc space and computing power.

## **Big data in practice**

This section focuses on the `occurrence.txt` file of the all GBIF mediated data dataset extracted from the DwC-A. It is the core tabular file containing all the occurrences downloaded from the GBIF. The size file exceeds 500 Go of disc space once extracted from the archive. For comparison purposes, it represents half the space of a decent external drive or more than 100 DVD.

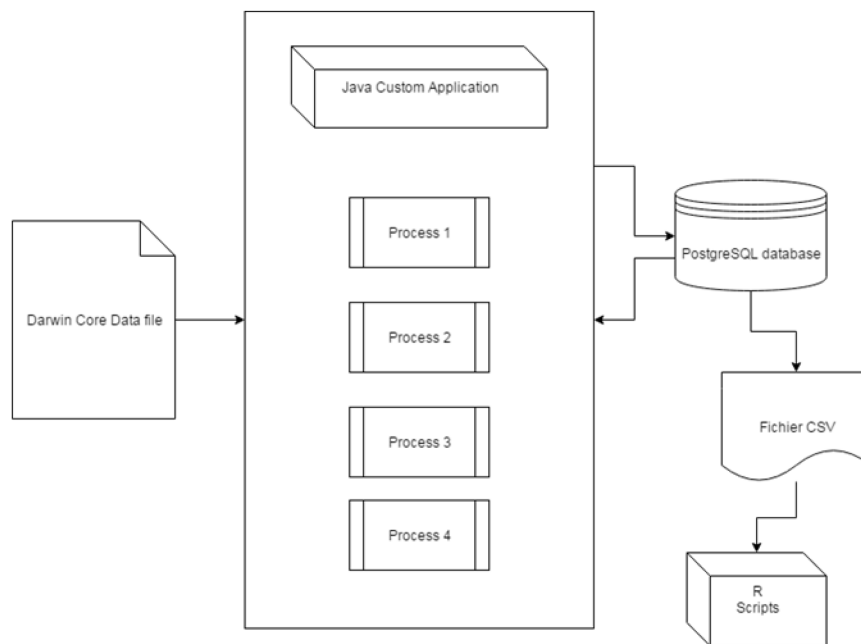
As explained in the introduction the words “big data” qualify datasets that cannot be manipulated with common tools. Many researchers now work on datasets including thousands of species and occurrences (e.g. Andam *et al.* 2016, Boucher-Lalonde and Currie 2016, Nicholson *et al.* 2016), and progresses in informatics facilitate their analyses. One of the most popular software environment used, R (R Development Core Team 2008), allows computing complex operations on thousands to millions of data. Unfortunately, the data used here were still too big to be handled through R alone.

## **Workflow architecture**

In the current era of “big data”, multiple solutions have been created to manipulate very large datasets, including Hadoop (EMC Education Services 2015) and NoSQL databases (Andlinger 2013). Considering the volume of the GBIF mediated data, the analyses planned, as well as my computing skills, I chose a workflow approach. I created a Java application to read data occurrences, do operations on these occurrences and then insert them into a database. This database can be queried and is updated after each operation. From this

database, a csv file can be exported. It contains the data later used with R scripts to compute statistical analyses (Fig 3).

To keep the integrity of the data I downloaded, I chose to “tag” the occurrences after each operation rather than to suppress or update them. The downloaded datasets were thus enriched and new columns with additional information were appended to the database.



**Figure 3: Global workflow organization. The Darwin Core files are read by a custom Java application. The occurrences are enriched with new information and then put into a database. The following processes query subsets of occurrences from the database and update them. Once all the processes have been done, the database can be queried again to export CSV files that are read in R and used to compute statistics.**

The creation of this workflow led me to use many programs and software. I will list here the ones I used during my work to create programs and scripts. The java code was written with the Eclipse Oxygen Release (4.7.0) IDE ([www.eclipse.org](http://www.eclipse.org)), and a complete list of all the Java libraries used is available in the appendix 1. For general code and scripting I used Sublime Text 3 ([www.sublimetext.com](http://www.sublimetext.com)). Many geographical analyses were done using QGIS Desktop 2.8.1 ([www.qgis.org](http://www.qgis.org)). This software was also used to create most of the maps displayed in this manuscript. Most of the geographical computations were done using the

equal-area and pseudo cylindrical map projection Eckert IV. The database engine used was PostgreSQL ([www.postgresql.org](http://www.postgresql.org)) along with pgAdmin III ([www.pgadmin.org](http://www.pgadmin.org)).

I performed all analyses using the R statistical software version 3.3.2 (<https://www.R-project.org>) with associated packages: *ape* (Paradis *et al.* 2004), *biomod2* (Thuiller *et al.* 2009), *FactoMineR* (Husson *et al.* 2016), *geosphere* (Hijmans 2016), *ggplot2* (Wickham 2009), *gmp* (Lucas *et al.* 2017), *gridExtra* (Auguie and Antonov 2017), *GWmodel* (Gollini *et al.* 2013), *MASS* (Venables and Ripley 1994), *plyr* (Wickham 2011), *rgdal* (Bivand *et al.* 2017a), *scales* (Wickham 2015), *spdep* (Bivand and Piras 2015) and *spgwr* (Bivand *et al.* 2017b).

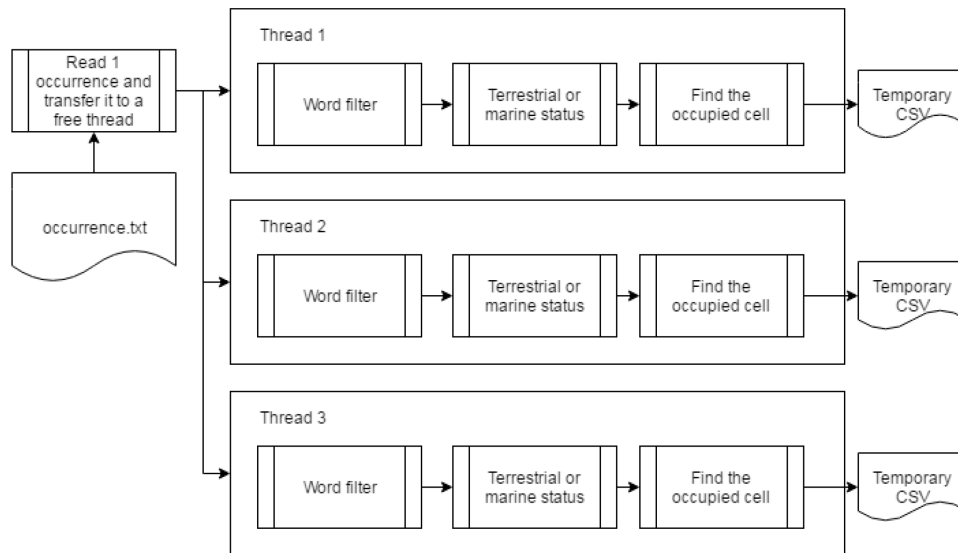
### **Reading and filtering occurrences**

As stated previously the quantity of data to read was a challenge in itself. The `occurrence.txt` file extracted from the DwC-A is tabulated with 234 columns and hundreds of million rows.

Reading all those occurrences could be relatively fast but processing them one by one can require a large amount of time even for a computer program. Moreover, while the direct import of this type of file into a database is usually straightforward, it was impossible here because the CSV file had errors and no indication about the import progression was provided by the SQL server.

I built an application capable of multithreading, meaning that multiple processes can be computed concurrently on the computer. This differs from multiprocessing and it does not require a multi-core processor, even though having multiple processors speed up the computations.

The first function of the application was to read the occurrences one by one from the tabulated file and allocate them to other threads. Each of these threads processes the occurrence and puts the result into a thread-specific CSV file (Fig. 4).



**Figure 4: First computation step of the application.** The occurrence.txt file is read by a single process that reads it line by line. Each line is then processed by a free thread. If no thread is free, the computation waits for a free one. Each thread does three different computations on the occurrence it received. Filter for potential zoo or farm occurrences (word filter), check whether the occurrence is terrestrial or marine, and assign the occurrence to a specific cell of the global 10\*10 km grid (see section Worldwide grid). Once the occurrence is edited, it is added to a temporary CSV file, one for each thread. During computation, as many as 200 threads were used concurrently.

### *Zoo occurrences*

The GBIF provides several filtering options. One of them concerns the column `basisOfRecord` that indicates the origin of the specimen described by the occurrence. It is thus usually possible to determine if a given occurrence comes from an observation, a collected specimen, a photograph, or a living specimen. Typically, farm or zoo data should be labeled as living specimen and I got rid of them using the appropriate filter. Still, I chose to double-check this filter because columns in the occurrence file are not always appropriately filled. Thus, after reading an occurrence, the application checks the columns `Locality` and `occurrenceRemarks` for the words "zoo", "aquarium", "farm" and "captive". When one of these words is found, the occurrence is tagged as "potentially captive specimen".

### ***Tagging terrestrial occurrences***

To investigate the Latitudinal Diversity Gradient, I focused on terrestrial organisms, although GBIF mediated data include both terrestrial and marine occurrences. There is no filter in the GBIF portal to discriminate terrestrial and marine species, in part because these exclusive concepts do not always fit with the reality of organisms living in both habitats and also because the GBIF being a data mediator, it does not seek to assess information about species (D. Shigel pers. comm.). For some occurrences, information about soil, climate or depth could help but, once again, this information is neither standardized nor available for all occurrences. I thus included a computation step in my application to tag land *versus* marine occurrences, when occurrences fall into a terrestrial or marine polygon, respectively. The tag was a word (terrestrial or marine) added to each occurrence. The application used the most precise maps from [www.natureearthdata.com](http://www.natureearthdata.com) (scale 1:10,000,000) and the ArcGis Java library and RTree algorithms to find which polygon contained a given occurrence. I chose to include lakes and other fresh water bodies in the “terrestrial” category.

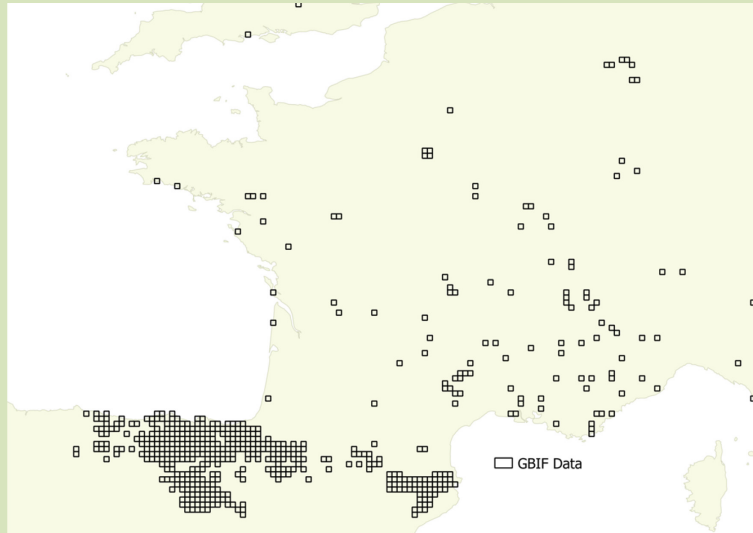
Then, I decided arbitrarily to qualify species as terrestrial when more than 90% of its GBIF mediated occurrences were on land. This probably excluded some terrestrial species (Yesson *et al.* 2007) but insured that very few marine species would be used in the analyses on the LDG.

### ***Worldwide grid***

To investigate the spatial distribution of species richness, I needed a worldwide grid composed of equal area cells. I created a worldwide grid of 10\*10 km cells, using the equal-area pseudocylindrical map projection Eckert IV.

Assigning each occurrence to a geographic cell was done as finding the terrestrial occurrences, but instead of tagging an occurrence with a word, I added a unique identifier – composed of the X and Y coordinates of the cell – to the occurrence. Using these coordinates I could create for each species a list of all the occupied cells. After doing it for all species I ended up with a new table containing all the unique species-cell pairs, or spatially distinct occurrences, for future computations.

**Green box 3:** After allocating the 2391 occurrences of *Lacerta bilineata* to the worldwide grid, we are left with 542 cells.



Repartition of the GBIF mediated data of *Lacerta bilineata* in France and bordering countries. Some data are not represented in this screenshot.

### **Indexing the table to get a functioning database**

At the end of the first step, the application produced up to 200 CSV tabulated files, which are then imported into a database. The application also handled this step using multithreading to import multiple files concurrently.

At this point, all the GBIF mediated occurrences, as well as the additional information previously computed (i.e. terrestrial/marine and geographic cell), have been imported in a PostgreSQL database, as a single table. Each row of this table corresponds to one SpOcc. This table is searchable but its size prevents from reasonably fast searches.

Indexing a table is a common solution to accelerate searches and reduce computation time. I thus created multiple indexes on the main table (see appendix 2 for a list of indexes and the corresponding SQL queries).

## **Characterizing a biodiversity dataset: biases and trends**

Once the GBIF mediated data was filtered, indexed and included in the main database, it was possible to study the dataset itself. To facilitate this investigation, I created many additional tables (See appendix 3), allowing for faster searches and simpler queries. Some additional information was also collected to further study the effect of some variables on the biodiversity data collection. Based on those new tables and additional data I could use scripts and statistical tools to get a better understanding of the biases and the trends affecting the GBIF dataset. Those biases are very important because if they affect the GBIF data which is the biggest primary biodiversity data repository they are likely to affect all biodiversity domains (Powney and Isaac 2015). The same reasoning is applied for the trends in GBIF mediated data. The following section will describe how I quantify those biases and trends in the GBIF data but also try to explain them with external data and statistic tools. The works detailed in chapters 2 and 3 rely on this procedure.

### **Species names from the GBIF Backbone Taxonomy and multimedia files**

Two metrics were not available in the main occurrence.txt file extracted from the DwC-A.

The first metric was the number of multimedia files attached to an occurrence. A SpOcc consists of one observation or collect of a specimen at a specific time and place. However, such specimen could be photographed or its vocalization could be recorded so that a single occurrence can be linked to multiple media files. This 1-N relationship cannot be stored in the occurrence.txt file because an occurrence corresponds to only one row. Instead, a multimedia.txt file containing the list of multimedia files is provided. This file has one row for each multimedia file linked to a SpOcc (using the `gbifid` column) in the GBIF dataset, and a SpOcc can be linked to several rows (i.e. multimedia files). The multimedia.txt file was imported in the database using a simple “copy” query and then queried to get the needed statistics (see chapter 1).

The second metric was the number of described species per taxonomic class, a metric used while investigating the taxonomic bias in biodiversity data (Chapter 4). At first, I



imported these figures from Catalogue of Life ([www.catalogueoflife.org](http://www.catalogueoflife.org)) but many species referenced in the GBIF were not in this catalogue. Indeed, the GBIF created its own classification system, called the GBIF Backbone Taxonomy, using diverse taxonomic databases, including, but not restricted to, Catalogue of Life (Text box 1).

**Text box 1:** The GBIF portal provides the following definition of its Backbone Taxonomy:

The GBIF Backbone Taxonomy, often called the Nub taxonomy, is a single synthetic management classification with the goal of covering all names GBIF is dealing with. It's the taxonomic backbone that allows GBIF to integrate name based information from different resources, no matter if these are occurrence datasets, species pages, and names from nomenclators or external sources like EOL, Genbank or IUCN. This backbone allows taxonomic search, browse and reporting operations across all those resources in a consistent way and to provide means to crosswalk names from one source to another. It is updated regularly through an automated process in which the Catalogue of Life acts as a starting point also providing the complete higher classification above families.

I imported the GBIF Backbone Taxonomy as a new table before using it to get diverse statistics, the most important one being the number of described species in each taxonomic class. Using the `nomenclaturalStatus` column, I excluded synonyms and kept only `accepted` and `doubtful` species. Doubtful species were kept as many species in the GBIF have a `doubtful` name and including only `accepted` species name sometimes led to more than 100 % of known species referenced in the GBIF. This was due to the number of species referenced in the GBIF being higher than the number of `accepted` name in the taxon.

## Public interest and taxonomic research quantity

### *Public interest as a societal influence*

The lay public does not attach the same importance to each and every organism. Among the 1.2 million species described (Mora *et al.* 2011), some are loved, while others generate aversion or are virtually unknown. In this dissertation, the public interest for a taxon is defined as the popularity of the taxa to the public. Because there is no global “likeability” index for every existing species, I chose to use the number of web pages found by a Web Search Engine for a species as a proxy of its public interest. Wilson *et al.* (2007) showed that “many (30–80%) web pages containing the scientific names of species have little or nothing to do with scientific research” so the results obtained are presumably related to societal preferences.

I used a Visual Basic Script (see appendix 4) to get the number of web pages found by the Bing search engine of Microsoft® for a given species. As seen in figure 5, the number needed is displayed just before the results. The script was used to accelerate the process, knowing that more than 48 thousand requests were performed.

The Google search engine was not used because it did not allow for such a large number of searches to be done automatically. However, the two search engines were compared for 1000 species and gave comparable results.



**Figure 5: Screenshot of the results obtained when searching a species in Bing. The number of results I used is circled in red.**

### ***Taxonomic research quantity***

Systematists are one of the main producers of primary biodiversity data. The more systematists for a given taxonomic group there are, the more observations, specimens and species descriptions for this group can be generated. I looked for the number of systematists per taxonomic group, but there was no way to obtain those numbers in a timely manner in available databases, and it was also impossible to get them using Web of Science (WoS). But, from WoS, I used the number of taxonomic papers produced per taxa as an estimate of taxonomic research quantity.

McKenzie and Robertson (2015) similarly measured research quantity on 225 British breeding birds' species using WoS. However, considering the thousands of species in GBIF mediated data and the impossibility to automate the search process, I estimated taxonomic research effort at the order scale. For each of the 454 orders referenced, I searched the WoS portal ([apps.webofknowledge.com](https://apps.webofknowledge.com)) with the following query: “taxonomy” AND (“[order name]” OR “[family names]”), on the 1900-2016 period.

### **Putting the GBIF database into numbers**

Using SQL and R, different statistics were computed to analyze the GBIF mediated data. For each statistics, the computation process was the same: first, I queried the SQL database to obtain a CSV file with the selected data; second, I used R to produce statistics and graphs using the CSV file. Below, I illustrated this process with an example aiming at visualizing the taxonomic bias in biodiversity data.

To investigate the taxonomic bias in biodiversity data, I wanted to visualize the excess and deficit from an ideal repartition of the number of occurrences for each taxonomic class. The ideal occurrence repartition occurs when each class has a number of occurrences proportional to its number of known species, meaning that the more speciose classes would have more data. I worked on the 24 classes having at least one million occurrences in the GBIF mediated dataset.

For each class, I first needed to store, in a CSV tabulated file called `over_under_sampled.csv`, the number of known species and the number of occurrences. I obtain these values with the following queries, respectively:

```
SELECT o_class, COUNT (*) as nb_sp
FROM public.backbone_08_2016
WHERE (o_taxonomicstatus = 'accepted' OR o_taxonomicstatus = 'doubtful')
AND o_genus != ''
AND o_taxonrank = 'species'
AND o_class IN ('Actinopterygii','Agaricomycetes',...)
GROUP BY o_class
ORDER BY o_class;

SELECT o_class, COUNT (DISTINCT(o_specieskey)) as nb_species
FROM public.OCCURRENCES
WHERE o_class IN ('Actinopterygii','Agaricomycetes',...)
AND o_specieskey IS NOT NULL
GROUP BY o_class
ORDER BY o_class;
```

The table 1 shows the first lines of the CSV tabulated file. The global mean of the number of occurrence per species is also the number of occurrence per species if each species was equally sampled in the dataset:

$$\text{Global mean} = \text{total number of occurrences} / \text{total number of species}$$

**Table 1: First lines of the `over_under_sampled.csv` file obtained after querying the database for the number of species and the number of occurrences in each class.**

CLASS	nb sp	nb occurrences	class mean
Actinopterygii	30733	14180324	461.4
Agaricomycetes	23528	3798022	161.4
Amphibia	5887	3941881	669.4
Anthozoa	8637	1027884	118.9

The following R script is a simplified version of the original script that reads the data from the CSV file before producing a graphical representation of the taxonomic bias in the GBIF mediated data (Fig. 6):

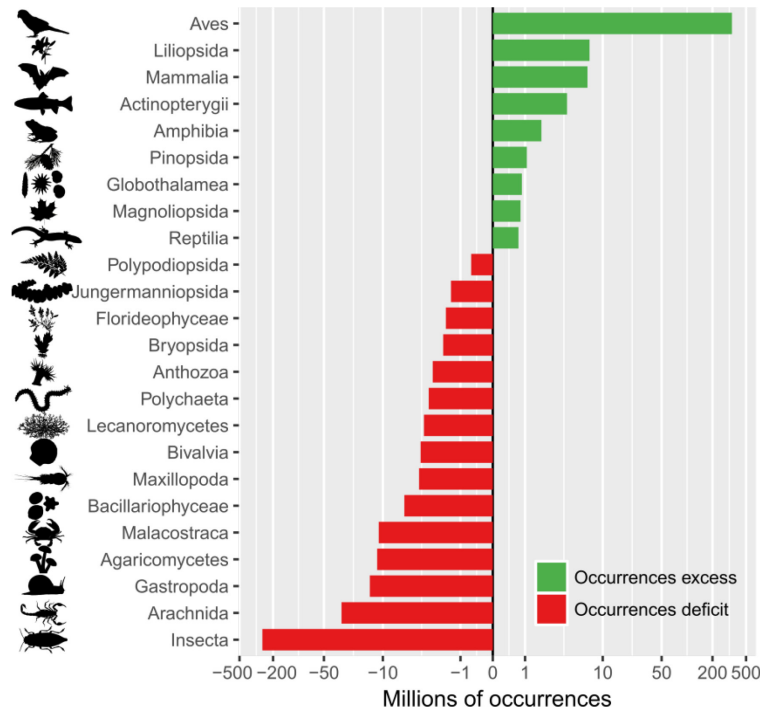
```
# returns the color red if the sign is negative else returns green.
getColor <- function(signVect){
  colorVect = replace(signVect, signVect==-1, "#e41a1c")
  colorVect = replace(colorVect, colorVect==1, "#4daf4a")
  return(colorVect);}
# transformation function similar to a log scale
asinh_trans <- function(){
  trans_new(name = 'asinh', transform = function(x) asinh(x),
  inverse = function(x) sinh(x))}
# read the data from the CSV
datas = read.csv("over_under_sampled.csv", header=TRUE, sep=";")
```

```

# calculate the excess or deficit of data from a perfect data repartition for each class
datas$situation = datas$nb.occurrences - (datas$global.mean*datas$backbone.nb.sp)
# ordering the classes
datas$CLASS <- factor(datas$CLASS, levels = datas$CLASS[order(datas$situation)])

# displaying the plot
ggplot(datas,aes(x=CLASS, y=backbone.nb.sp)) +
geom_hline(yintercept = 0)+
geom_bar(aes(x = CLASS , y = situation/1000000 , fill=getColor(sign(situation)), group=1),
stat="identity", width=0.8 ) +
coord_flip() +
scale_y_continuous(trans = 'asinh', breaks=c(-500,-200,-50,-10,-1,0,1,10,50,200,500))+
scale_fill_manual(...)

```



**Figure 6: Plot of the deficit and excess in occurrences per class in the GBIF dataset, as an illustration of a R script and its output.** More details regarding the significance of this figure are provided in chapter 3.

## Statistical tools

In the process of studying the GBIF dataset I needed to test some hypothesis with more complex statistical tools. Those tools were available as packages in R and allowed me to assess the influence of external variables on the GBIF dataset as well as finding tendencies and trends inside the dataset.

### ***Generalized linear models***

I used generalized linear models (GLM, McCullagh and Nelder 1989) to explore the impact of public interest and taxonomic research quantity on the GBIF dataset. The GLM is a generalization of linear regression that allows the measurement of these variables (and their interaction) effect on other variables in the dataset (number of occurrences per species).

GLM are strongly influenced by extreme value and I had to filter out outliers: species having very high number of occurrences or web search results (public interest). Many GLM were computed and each one was then checked for model validity and its residuals plotted against predicted values.

### ***Multiple correspondence analysis***

The multiple correspondence analysis (MCA) was used to find the relation between categorical variables inside the GBIF dataset. For example I tested for relations between the age of an occurrence (number of years since the observation event), its data origin (categories: specimen, observation and unknown) and the data completeness (categories: no problem, missing temporal information, missing spatial information and missing both). The class of the occurrence can also be projected on the resulting plot. These analyses were done using the FactoMineR package for R (Husson *et al.* 2016). The analysis couldn't be done on all the GBIF occurrences because R couldn't load all the occurrences in memory. I made analyses on multiple 5 million random occurrences samples, and even tried to ventilate categories representing less than 0.5 % of the dataset as they could have altered the results.

## **Working on biodiversity patterns: delving into ecoinformatics**

### **Cleaning data**

Systematics and Ecology are now data-intensive sciences. But “Big Data” does not necessarily mean better data (Boyd and Crawford 2012). Data quality must be ensured before further analyses. Some data are faulty, while others can be insufficient (i.e. under-sampled species) to produce meaningful estimates. These two issues have been tackled on terrestrial

species before computing species richness across the globe to investigate the Latitudinal Diversity Gradient.

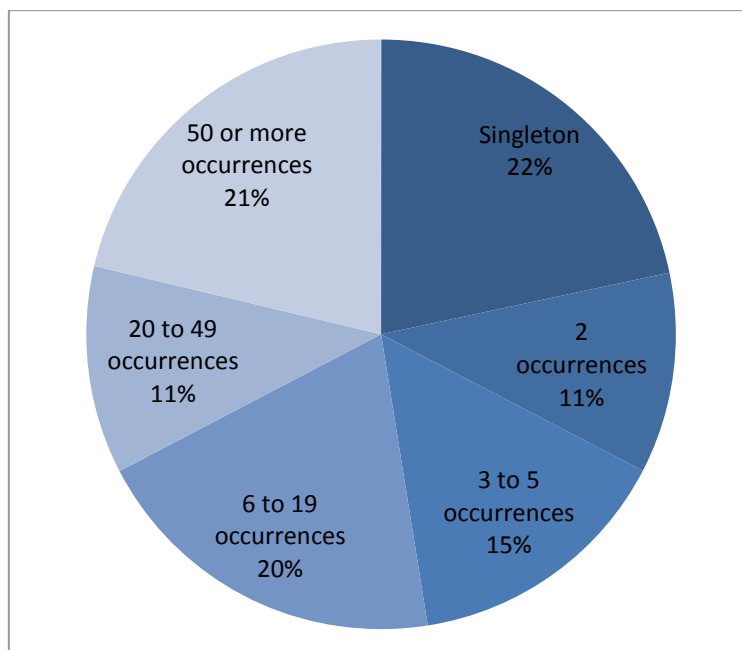
I produced a new table merging occurrences of the same species in the same geographic cell. I kept only the information about the species name, the cell occupied and the number of merged occurrences in the cell. I obtained a new dataset of spatially distinct occurrences to perform computations on. Then, under-sampled species, i.e. species with less than 20 spatially distinct occurrences, were filtered out (see below *Selecting well-sampled species*).

Data can be faulty in several ways: they can be biased, inaccurate or imprecise. The two latter issues were tackled here at the species level. I analyzed each species separately and identified odd occurrences, called *outliers*. The majority of GBIF data being correct, whether collected by scientists or citizens (Yesson *et al.* 2007, Kosmala *et al.* 2016), an algorithm should be able to identify occurrences inconsistent with the others. The selected algorithm used the orthodromic distance between occurrences and the climatic data associated to the occurrences to find potential outliers. Misidentified occurrences mixing two species with different habitats, and input or typing errors would lead to obvious inconsistencies that should be easily detected. On the opposite, erroneous data coherent with the rest of the species dataset would not be found.

The java code allowing spatial and environmental outlier detection was later cleaned and put in new software with a dedicated interface, designed to be easily usable by the scientific community. The resulting software is detailed in chapter 5.

### ***Selecting well-sampled species***

Selecting species with at least N occurrences is straightforward and requires a simple SQL query. On the 1,370,170 species referenced in the complete GBIF dataset 296,487 (22%) are singleton (i.e.  $N = 1$ ), and 447,468 species (32) have at least 20 occurrences (Figure 7).



**Figure 7: Proportions of the 1,370,170 species categorized by their number of occurrences in the GBIF mediated data.** Singletons are species having only 1 occurrence.

In addition, the merging of occurrences occurring in the same 10\*10 km cell (i.e. not spatially distinct) also reduced the number of occurrences per species. This step is the one that eliminated the most legitimate occurrences from our dataset. Chapter 3 provides more elements about the proportion of species having a certain threshold of spatially distinct occurrences. The following outlier detection is done on the species having met this occurrence number threshold.

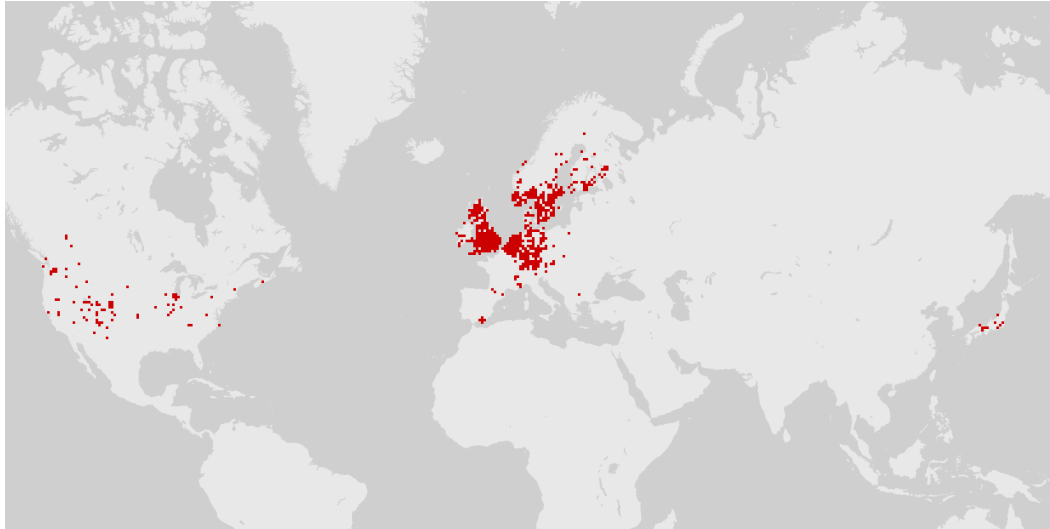
### ***Geographic outliers***

Geographic or spatial outliers are occurrences that are abnormally apart from the rest of the occurrences. Detecting such occurrences can be trivial for the human eye in the case of obvious mistakes when projecting data on a map. However it becomes far more hazardous when dealing with thousands of point, and far more strenuous when repeated for the tens of thousands of species covered by the GBIF mediated data.

The task is further complicated when species occurrences are clusterized, meaning that occurrences from one species are sometimes distributed among multiple patches of data



(Figure 8). Diverse algorithms have been proposed to identify outliers of clusterized data (e.g. Breunig *et al.* 2000, He *et al.* 2003, Hardin and Rojke 2004) but they were not used here because they require additional parameters (such as the number of clusters) that could not be obtained for all the species analyzed.



**Figure 8: Global repartition of the common black ant (*Lasius niger*) occurrences.** A single red square can represent multiple occurrences. The occurrences are distributed among three clusters: North America, Europe and Japan, complicating the detection of outliers.

The worldwide coverage of the data was an additional hindrance. On a spherical world, it is not possible to simply calculate distance between points using latitude and longitude. One point at the coordinates 0,-178 and the other at 0,178 are separated by an angle of 4° and not 356°.

The chosen solution was to use the orthodromic distances, i.e. the shortest distance between two points on a sphere (the earth). For each occurrence in a given cell, I computed the distances to the five nearest cells with conspecific occurrences and summed those distances to get a “spatial eccentricity” value. Then, I flagged as outliers the 1 % cells with the highest spatial eccentricity. This process could potentially flag correct occurrences as outliers but it was the most conservative and fastest method we found. Only the species having met the cell number threshold were tested. The R script used is available in appendix 5 (Nicolas Lebbe, 2015 (*com. pers.*)) and was used on species with at least six spatially distinct

occurrences (= species-cell pairs). An example of spatial outlier detection for the black rat (*Rattus rattus*) occurrences is provided in figure 9.

### ***Climatic outliers***

To maximize data quality and following the same rationale as for geographic outliers, we excluded climatic outliers. Climatic values for each species occurrence were needed to detect these outliers. This type of data was only rarely provided with occurrences and without any standardization between data providers. But, because each occurrence corresponded to a 10\*10 km cell, it was possible to use global climatic data to infer climatic values of each cell of the worldwide grid.

I downloaded the WorldClim data from worldclim.org (Hijmans *et al.* 2005) at the 30 second precision and used Qgis (qgis.org) to assign the climatic data file to each 10\*10 km cell of the worldwide grid. For each of the 20 variables available in WorldClim, I computed the mean value of the WorldClim points inside each cell of the grid.

Then, I compared four of the climatic variables (bio1: annual mean temperature, bio5: max temperature of warmest month, bio6: min temperature of coldest month and bio12: annual precipitation) of all the cells occupied by a given species. At this step, only 4 variables were used because the following computation couldn't be done on species having less occurrences than there are variables and I seek to keep as many species as possible. These specific variables were chosen as they are often cited as important limitations to the species niches (Fine 2015, Ferrer-Castán *et al.* 2016) I used the R package mvoutlier (Filzmoser 2005) to compute the Mahalanobis distance (Mahalanobis 1936) of each cell and find the climatic outliers. The Mahalanobis distance measure the distance between a point and the average value of the distribution. It is often used to detect outliers (De Maesschalck *et al.* 2000). Here is the very short R script used for this:

```
# loading the package
library("mvoutlier")

# bio1, bio5, bio6 and bio12 contain the climatic values of the cells
# containing one occurrence of the species. (excluding spatial outliers)
clim_vals = cbind(bio1, bio5, bio6, bio12)

# creates the plot containing the outliers
# alpha is the maximum thresholding proportion
# quan is the proportion of observations which
```

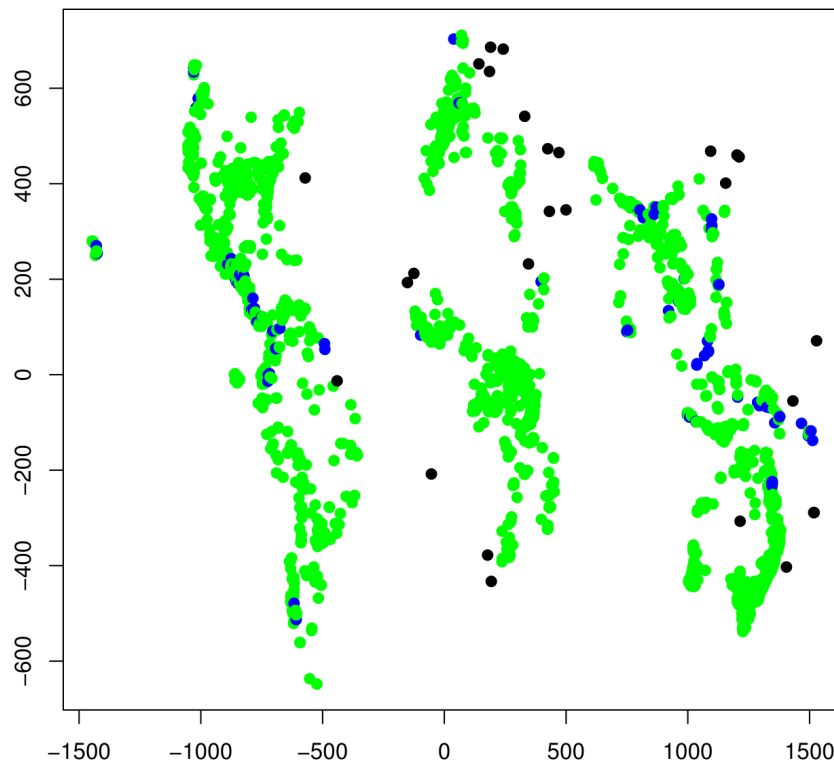
```

# are used for mcd estimations
# those optimal value where kindly provided by the Pr. Filzmoser
# when I asked him for help in using his package
mvout = aq.plot(clim_vals, alpha=0.01, quan=0.75)

# Gets the outlier occurrences array positions
mvout = mvout$outliers
mvout = as.numeric(mvout)

```

An example of climatic outlier detection for the black rat (*Rattus rattus*) occurrences is provided in figure 9.

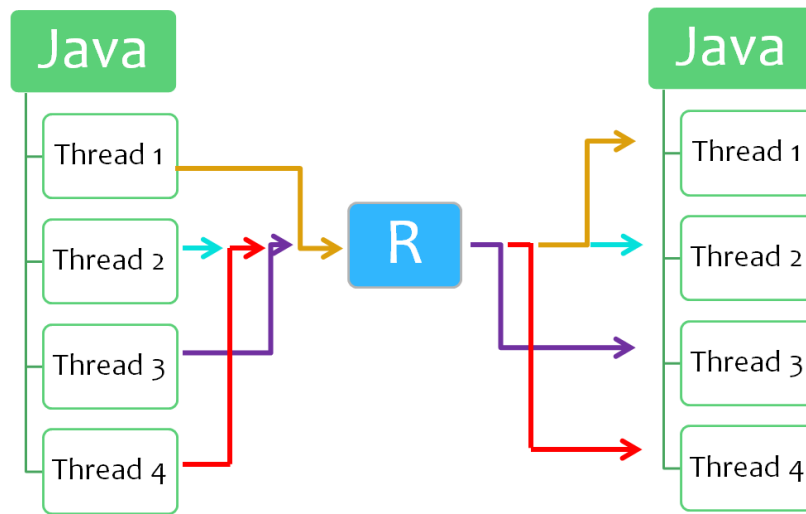


**Figure 9: Visualization of the outlier occurrences for the black rat (*Rattus rattus*).** The occurrences are projected on the plot using their cell coordinate. Each point is an occurrence of the species. In black are the 27 spatial outliers, in blue the 88 climatic outliers and in green the 2473 non-outlier occurrences.

### *Coordinating R and Java*

The methods chosen to detect outliers required coordinating R and Java. Even after filtering the data, it still contained hundreds of thousands of species with potential outliers.

Using only R to fetch the occurrences in the database, find the outliers and then update the database would be feasible but very long.



**Figure 10: Schema of the architecture used to work with R and Java.** Each Java thread work on a single species. Up to 200 threads were working at the same time. When an operation requires the use of an R script, the thread is placed in a queue before transferring data to R. R then runs its scripts sequentially and returns the results to the correct thread.

I chose to use Java and its easy task parallelization to speed up the outlier detection. The basic idea behind this setup was that most of my computations would run one time per species, each instance independent from others. Therefore, I could easily run those computations simultaneously to speed up the process. The first version of the software was not optimized this way and computations would have taken several month using it on the full GBIF dataset, hence the importance of multithreading.

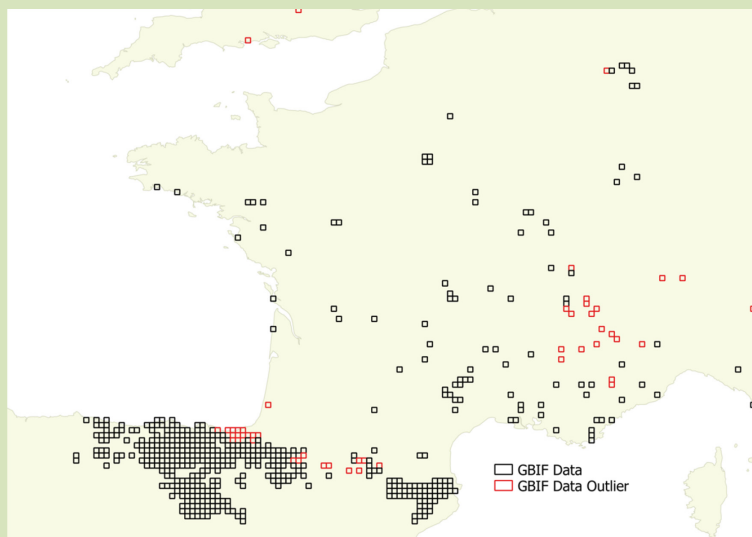
On the other side, R is not designed to run on multiple threads, which mean that I had to restrict the use of R scripts to the bare minimum: the outlier detection. The Java scripts would do all the other computations: querying the database for data, formatting the data, transferring the data to R, getting back the results of the R script and then updating the database. This functioning is detailed in figure 10. Using R was inevitable as many computations could not be re-written in Java in a timely manner.

For outlier detection, the processes were the same for each thread:

- Take the next species in the species list (species A)
- Query the database to get all species A cells
- Format the results returned by the database in a java object compatible with R
- Wait in the Thread Queue for R
- Transfer the data to R and start the execution of the R script (outlier detection)
- Get the results from R and add the outlier status (true or false) to each cell
- Update the corresponding cell in the database
- Take the next species in the species list (species B)

Using my application, I could detect geographic outliers for more than 500,000 species in less than 24 hours.

**Green box 4:** After running the outlier detection scripts on the 542 cells with occurrences of *Lacerta bilineata*, 6 cells were excluded as geographic outliers and 56 as climatic outliers (i.e. 11.4% of the cells).



**Detection of geographic and climatic outliers in the occurrences of *Lacerta bilineata*.** Outliers are represented in empty red squares and the supposedly correct occurrences in empty black squares. Some data are not represented in this screenshot.

## **Estimating species richness**

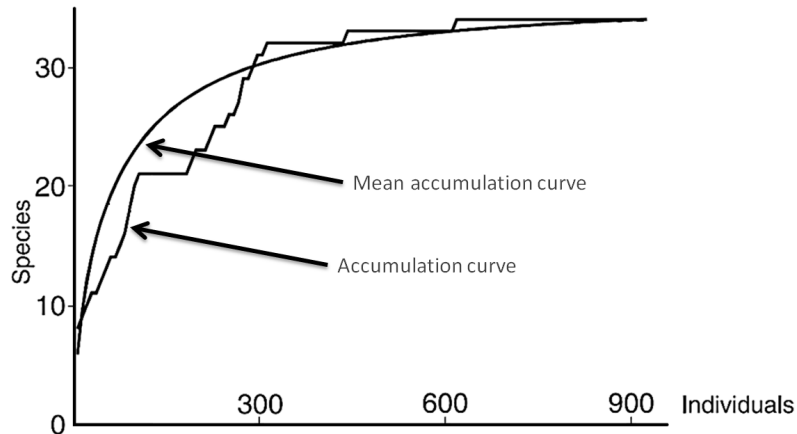
Species richness is the simplest measure to describe community and regional diversity (Magurran 2004) and is an essential statistics for many community ecologist, macroecologists and conservation biologists (Magurran and McGill, 2010).

After obtaining a clean dataset, consisting of all the pairs of species-cell and potential outliers filtered out, I still could not work on species richness patterns or even visualize a worldwide grid of biodiversity richness. Indeed, because of the data fragmentation and poor cover of certain regions most of the classic methods to estimate species richness were inappropriate. To get a better coverage and reduce the limitation of geographically biased samples (Meyer *et al.* 2015), I used niche modeling algorithms.

### ***Classic but unusable methods***

Typically, species richness is calculated for a supra-specific taxon in a given area. For example, one can estimate the species richness of mammals in Madagascar. If we had, for an area, the complete list of a taxon's species, then the species richness of this area would simply be the number of different species of this taxon. However, biologists (almost) never have the complete list of organisms inhabiting a natural area and must use statistics considering the sampling effort to estimate biodiversity. Moreover it is worth underlining that even a supposedly complete sampling would only reflect the species richness of an anthropogenically modified area (Faurby and Svenning 2015).

Of course, sampling a locality does not allow recording for all specimens, unless focusing on a very limited area or taxon. But a sampling effort can be large enough to allow for accurate species richness estimate. As more individuals are sampled, more species will be recorded (Bunge and Fitzpatrick 1993) until an asymptote is reached, meaning all species have been discovered. This accumulation curve can be seen in figure 11 and shows how the sampling effort affects the amount of species found.



**Figure 11: Accumulation curve of species discovered depending on the sampling effort.** The accumulation curve represents a hypothetical sampling (single ordering of individuals) while the mean accumulation curve, also called rarefaction curve, is averaged from repeated re-sampling of all pooled individuals. Figure drawn from Gotelli and Colwell (2001).

The following methods were tried on 100\*100km cells.

A limitation of the GBIF mediated data can be easily understood from this accumulation curve: the sampling is never the same depending on both location and taxa. Considering the intrinsic heterogeneity of the GBIF dataset, there is no way to be at the asymptote (even for a small taxon) or even at a common “minimal” level of sampling in every part of the world. Some areas are even devoid of species occurrences like some regions of central Asia (Meyer *et al.* 2015).

But after all, my aim was not to get the most precise estimation of species richness but having comparable species richness results between different areas. Assuming I could weight species richness with the sampling effort, I would have comparable species richness values. Unfortunately, no attempt to standardize the sampling effort succeeded as the GBIF mediated data were far too heterogeneous.

Another contemplated plan was to do a sub-sampling of the GBIF data to put the over-sampled areas at the same level than the under-sampled ones. Then again, the heterogeneity of the GBIF mediated data remained a problem because sub-sampling would have eliminated the majority of the data.

Finally, I tried to use non-parametric estimators. Those statistical tools are used to estimate species richness and are often based on rarefaction curves. Plenty of those estimators have been tested: Chao1 and 2 (Chao 1984), Jackknife 1 and 2 (Burnham and Overton 1978, 1979; Heltshe and Forrester 1983), Bootstrap (Smith and van Belle 1984), FIDEGAM (Pardo *et al.* 2013), etc. None of them were appropriate for my dataset. I first used them on the brut GBIF data because they require abundance data and are not heavily impacted by outliers. However, due to the high proportion of singleton per cell in the GBIF mediated data species richness was greatly overestimated. As an example, Jackknife1 results suggested there were more than 800 species of mammals in some areas of continental France, which is far more than the 187 species effectively referenced.

### ***Niche modeling***

To compensate for the incompleteness of the data, the final choice was not to estimate the correct species richness inside each cell but rather to compute the supposed repartition of each species. This option was chosen because of the observed incompleteness of each species (which is normal at a 10\*10 km grain). This way, the method is effectively countering, the scattering of the data. By resolving this scattering for each species it should be possible to obtain a better estimation of species richness in each cell.

To resolve data scattering in each species, I used a niche model inside a convex hull as proposed in García-Roselló *et al.* (2015). As previously said, we kept species recorded in 20 or more cells to insure the accuracy of this model (Feeley and Silman, 2010). For each species, I used R to compute a convex hull containing all the non-outlier cells. This step used the multithreading processes described earlier. Then I used the surface range envelope (SRE) model from the package BIOMOD2 (Thuiller *et al.* 2009) on the cells having their center inside the convex hull.

The SRE model was chosen because of both its simplicity and low requirement in computational power (Text Box 2). This model is strongly influenced by extreme values of the variables and the number of variables used. Consequently the more extreme the climatic values for a species the more cells will be compatible with it; and the more environmental variables we use, the more restrictive the model becomes. The 19 climatic values and the

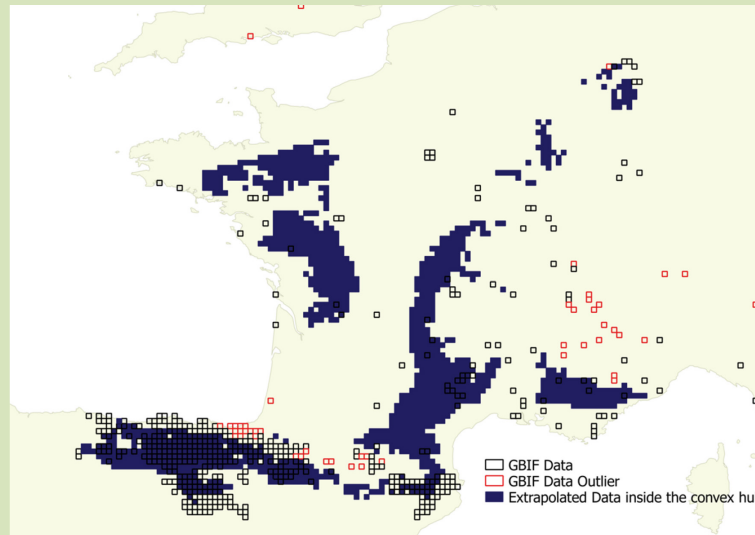


altitude available in WorldClim (Fick and Hijmans 2017) were used. We configured the model to remove the 5% more extreme value of each variable to limit the influence of extreme environmental values.

**Text box 2: Requirements for the niche model.** Many niche modeling algorithms exist, along with many implementations and software to use these models (Wiens *et al.* 2009). However I had many requirements for the model and most of those models didn't meet them all:

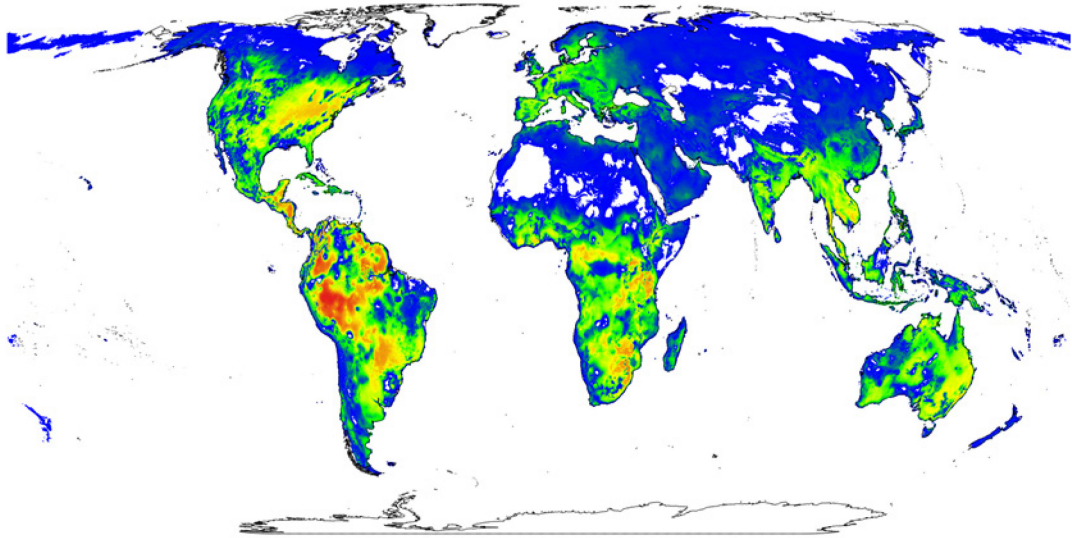
- The model should be simple enough to work on very different species. The GBIF mediated data includes occurrences about primate, conifer, butterfly... Picking a different model for different taxa would have been time consuming, complicated and hazardous.
- The model needed to use only presence and climatic data which was at my disposal.
- I needed a model that could be run by a java application (my software) using command lines for example. A model available as Java library or R package was ideal to run on hundreds of thousands of species.
- The model needed to be able to run in a timely manner. As more than 400,000 species could be processed, the model needed to run in a few second or less for each species.

**Green box 5:** After drawing the convex hull around the non-outlier cells of *Lacerta bilineata* and running the SRE niche model on the cells that are inside this polygon we are left with 1294 niche cells for the species. This is more than double the area (the number of cells) covered by the GBIF mediated data (figure bellow).

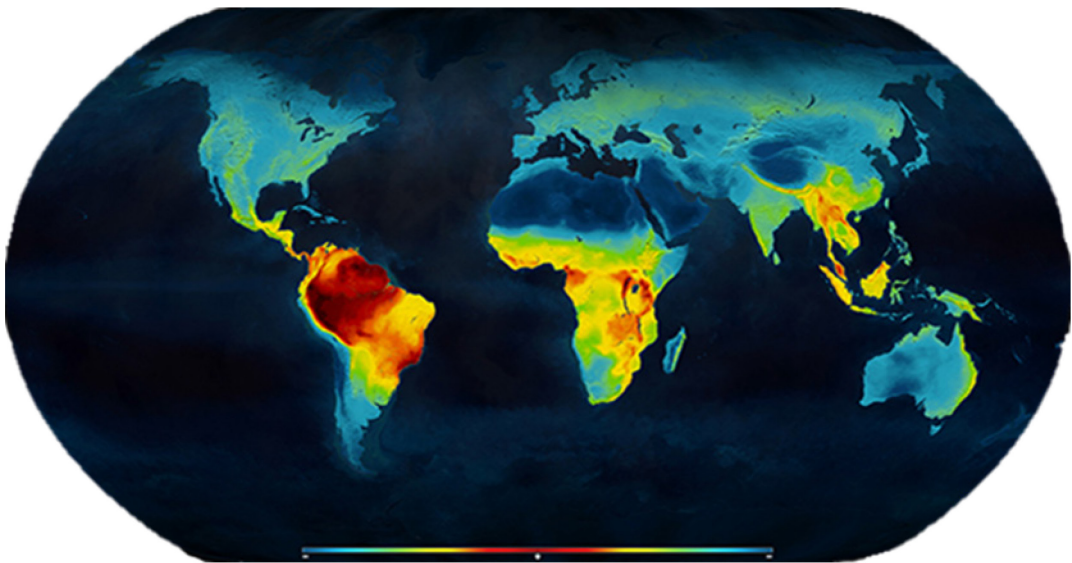


**Result of the niche modelling process on *Lacerta bilineata* non-outlier occurrences.** Outliers are represented in empty red squares, correct GBIF occurrences in empty black squares and potential niche cells in blue full squares.

All the cells determined by the model as potentially occupied by a species were put into a new table of the database. This new dataset includes only the species-cell pairs deduced by the model. As an example of the results of this method, a test on Mammals was performed. The Mammalia dataset contained six million occurrences from the GBIF, which were then simplified into 270 000 species-cell pairs (without outliers). After the niche modeling step, we ended up with a new dataset of 36 million potential species-cell pairs. This new dataset allows computing potential species richness for each cell of the worldwide grid. The data being far less scattered we obtain more readable results (Fig. 12) that are similar to results obtained by other searcher with a different dataset (Fig. 13).



**Figure 12: World map of the Tetrapod species richness (Reptilia, Amphibia, Aves and Mammalia) obtained using the tools I created.** The GBIF data has been filtered, outliers eliminated and each species put into a niche model to obtain all potential species-cell pairs. The cells having the most species are in red and the less species a cell has the more it goes to colder colors. No color means that there were no species referenced.



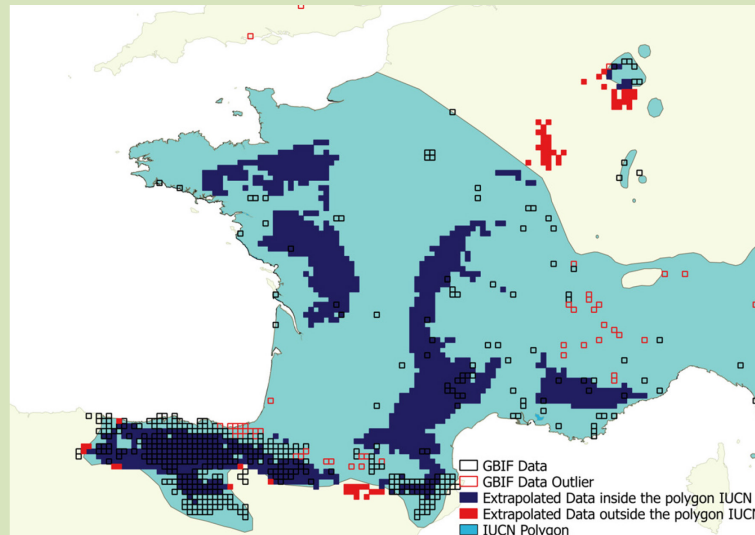
**Figure 13: Map of the Vertebrates diversity proposed by Mannion *et al.* (2014).** The high concentrations of diversity are closer to the red end of the color spectrum.

### ***Validating the methodology***

Using a niche model to infer the presence of a species in a cell is not a new technique but it was important to ensure the validity of the methodology with the GBIF mediated data. Thus, I compared the modeled species repartitions to others often used in biodiversity studies, those provided by the International Union for Conservation of Nature (IUCN). The IUCN Red List species are published with distribution maps and are widely used as references for mammals', amphibians' and reptiles' diversity (e.g. Brooks *et al.* 2002, Brito 2010, Whitton *et al.* 2012) even though they overestimate species distribution (i.e. a species is never present in every cell of a range) and do not perfectly reflect the real species range (IUCN 2009). Still, they were one of the few standard range maps available and easily usable for testing purposes.

I wrote an R script that computed, for each Red List species, the amount of species-cells pairs that were not in the corresponding IUCN polygons. By repeating the process for all the species both in the Red List and the GBIF mediated data, I obtained the proportion of discrepancy between the distributions modeled from GBIF mediated data and the IUCN polygons. The results of this test are explained in the Discussion.

**Green box 6:** Using the *Lacerta bilineata* known range downloaded from the IUCN website we can see which of our potential cells are consistent with it. The results shows that 1229 cells are located in the known range and 65 outside of it which correspond to 5 % of potential errors.



**Comparison of the IUCN (Red List) polygon to the GBIF data and the computed niche.** Outliers are represented in empty red squares, correct GBIF occurrences in empty black squares, potential niche cells in blue full squares and the 65 niche cell outside the IUCN polygon in full red squares. The light blue polygon has been downloaded from the IUCN website ([www.iucnredlist.org](http://www.iucnredlist.org)) and represents the potential repartition of *Lacerta bilineata*.

## Using our results to understand the LDG

Once I obtained a dataset consisting of all the potential cells/species pairs I could easily determine the species richness of each cell, and even choose to compute the species richness of particular taxa. Having this species richness it was once again relatively trivial to compute a latitudinal richness value by averaging the species richness of the cells inside a series of latitudinal ranges (for 10\*10km cells this gives us a species richness value per 100 km<sup>2</sup>).

However, while visualizing the LDG is interesting in its own right; what I really wanted to do was to test hypotheses about the formation of the LDG. Considering the broad taxonomic coverage of my data it would have been complicated to include historical hypothesis such as speciation and extinction rate as well as other phylogenetic hypothesis. However environmental data is easier to obtain and process. Those environmental variables are known to play a role in the latitudinal diversity gradient (Willig et al. 2003) by influencing species richness. I therefore chose to test the influence of those variables on the species richness I computed earlier.

### **The species richness covariates**

The statistical tools I had at my disposal could allow me to test for the correlation between species richness and a set of covariates (explanatory variables). As I had a species richness value for each of my cell, I needed to compute the covariates values for each of those cells. Only after doing this I could use statistical tools on the dataset.

The Ambient Energy (Currie 1991), Productivity (Hutchinson 1959), and Water availability (Hawkins *et al.* 2003a, Hawkins *et al.* 2003b) hypotheses suggest that species richness is influenced by environmental variables. The Ambient Energy hypothesis mainly lay on the assumption that sunshine and temperature are physical requirements of organisms (for thermoregulatory purposes mainly) while the Productivity hypothesis links the productivity (plant biomass) of an area to the number of individual, and therefore the number of species, it can support. The water availability hypothesis is based on the potential limiting factor of water availability on plant biomass. These hypotheses are all related to the energy-richness hypothesis (Currie *et al.* 2004). They suggest that the number of individual in an area is influenced by environmental factors (productivity in particular). As the species richness varies as a function of the number of individuals (Fisher *et al.* 1943), the productivity should consequently influence the species richness.

Those hypotheses were tested using Potential evapotranspiration and Annual Mean temperature, Actual Evapotranspiration and Annual Precipitation values taken from WorldClim ([www.worldclim.org](http://www.worldclim.org)) and Mu *et al.* (2011). All these variables were available as raster files that can be read by the QGIS software. I used this software to transform those files

in a tabulated file format that could be imported into my database. I then computed for each cell the mean values of the environmental variables (the raster files use a finer grid than me). After this operation I had for each cell the species richness and environmental variables values available.

With the data at my disposal I could also test for additional hypotheses formulated on the LDG, the geometrical hypotheses and the Rapoport's effect. The geometrical hypothesis was first formulated by Colwell and Hurtt (1994) who suggested that a latitudinal gradient could arise from the random placement of species ranges across the globe without any influence of environmental variables. This hypothesis, also called mid-domain effect, predicts a species richness peak or plateau in species richness, at the center of a bounded domain, when randomly placing a set of different species ranges within that domain. This hypothesis has, however, been contested by Currie and Kerr (2007, 2008). Later Gross and Snyder-Beattie (2016) added environmental limits concepts to this hypothesis to propose a new model. This new model adds a level of complexity to Colwell and Hurtt's model, and has never been tested on empirical data. Those geometrical hypotheses can also be called null or abiotic models as they imply the LDG could arise as a mathematical artifact, independently of environmental or historical variables.

The Rapoport' rule was formulated by Stevens (1989) who suggested that species latitudinal range sizes tended to increase with latitude. This situation means that latitudinal ranges would be smaller at low latitudes, leaving room for more species. This mechanism was therefore considered a potential factor in the LDG formation. The Rapoport's effect can be calculated using two methods: the original one proposed by Stevens (1989) and the Midpoint method (Rhode *et al.* 1993).

The computation of geometrical hypotheses as well as the Rapoport's effect was done with R and then included in the cell data. This step is covered in more details in chapter 4.

### **Statistical analysis of species richness and its covariates**

Many studies have been done trying to test the effect of environmental variables on species richness (e.g. Ferrer-Castán *et al.* 2016, Rodrigues *et al.* 2017). Many methods are

proposed in the literature to study this kind of spatial relationship and they can be summarized in three steps for most papers:

- The first step is to use a non-spatial analysis. This analysis builds a model assuming all points (in our case, all cells) are independent from one another. It also assumes that the relationship between species richness and its covariates is stationary across space (the model doesn't change depending on the location). In my case I used R and an Ordinary Least Square (OLS) analysis. I followed a manual iterative stepwise method selecting first the best null hypothesis for each class and then kept on adding other explanatory variables. At each step the variable added was evaluated using the coefficient of determination ( $r^2$ ) and the variable was not included when it did not improve the model adjusted  $r^2$  by at least 1 %.
- The second step was to test the model residuals for spatial autocorrelation using Moran's I test. This test is a measure of spatial autocorrelation. If the test find out that the residual are spatially correlated it means that the data is affected by spatial autocorrelation (Tobler, 1970).
- The third step is usually to use a spatial lag model or a spatial error model (Anselin *et al.* 1996) to test the model produced with the OLS analysis. This test will produce a regression model that takes into account spatial autocorrelation and ensure that an explanatory variable is not included only because of it.

Those three steps are often the ones used in paper working on the relation between species richness and environmental variables (e. g. Hawkins *et al.* 2003a, Mora and Robertson 2005). However they assume the model spatial stationarity. Spatial stationarity is rarely tested in such studies (Foody 2004, Mellin *et al.* 2014) mostly because it is a new tool that needs a lot of computing power. However I had at disposal the data and the computing power and decided to test the spatial stationarity of my final model with Geographically Weighted Regression (GWR). GWR is a local regression method that can be used for diagnosing spatial heterogeneity between dependent and explanatory variables over space (Brunsdon *et al.* 1996). GWR is performed within local windows centered on each observation of the dataset. Each observation within the local window is weighted based on its proximity to the center of that window and a regression model is then used on this subset of observations. This analysis



allowed me to test if the relation between species richness and the explanatory variable is constant across space.

## **Chapter 2: The increasing disconnection of primary biodiversity data from specimens: How does it happen and how to handle it? (Troudet *et al.* *Systematic Biology*, submitted as a Point of View)**

High quality data is a pre-requisite for conducting any reliable scientific study but can only be obtained through quality-check and data mining procedures (Cai and Zhu 2015). Quality and quantity being two distinct features, Big Data are not immune to quality issues. Like more sequences are not enough in phylogenomics to avoid systematic errors (Philippe *et al.* 2011), Big Data are not enough to ensure that a global, unbiased pattern would emerge (Boyd and Crawford 2012; Zook *et al.* 2017). Hence, Big Data must be curated even though data quality and mining are even more challenging when the quantity of data increases (Howe *et al.* 2008).

After ensuring a minimal quality for the GBIF mediated data used here, I engaged in data mining analyses, whose results nurtured this chapter and the following. For this chapter, I analysed the data focusing mainly on the column ‘*dwc:basisOfRecord*’ of the DarwinCore format used to manage GBIF mediated data. This column mentions the origin of the biodiversity occurrences and distinguishes occurrences relying on specimens (i.e. with a material evidence of the occurrence) from occurrences relying on observations (i.e. no material evidence of the occurrence). These analyses enabled us to characterize how biodiversity data have been gathered along time and how has evolved this process.

I show below that the practice of biodiversity data gathering has dramatically changed along the last century and that this shift impacts current and future biodiversity studies.

*Julien Troudet<sup>1</sup>, Régine Vignes-Lebbe<sup>1</sup>, Philippe Grandcolas<sup>1</sup>, Frédéric Legendre<sup>1</sup>*

1. Institut de Systématique, Evolution, Biodiversité, ISYEB – UMR 7205 MNHN CNRS UPMC EPHE, Sorbonne Universités, 45 rue Buffon, 75005, Paris, France.

Correspondence and requests for materials should be addressed to J.T. (julien.troudet@mnhn.fr)

## **Abstract**

Primary biodiversity data represent the fundamental elements of any study in systematics and evolution. They are, however, no longer gathered as they used to be, the mass-production of occurrences without any material evidence available (or observation-based occurrences) overthrowing the collection of occurrences based on material evidence such as a specimen or a sample (or specimen-based occurrences). Although this change in practice is a major upheaval with significant consequences in the study of biodiversity, it remains understudied and has not attracted yet the attention it deserves. Analyzing 536 million occurrences from the Global Biodiversity Information Facility (GBIF) mediated data, we show that this spectacular change affects all taxonomic classes (i.e. 24 eukaryote classes studied here). Ethical, practical or legal reasons responsible for this shift are known, and this situation appears unlikely to be reversed. Still, we urge scholars to acknowledge this dramatic change and deal with it, instead of letting it unguided. Specifically, we emphasize why specimen-based occurrences must be gathered, as a warrant to allow both repeating evolutionary studies and conducting rich and diverse investigations. When impossible to secure, voucher specimens must be replaced with observation-based occurrences combined with ancillary data (e.g., pictures, recordings, samples, DNA sequences, etc.). Ancillary data are instrumental for the usefulness of biodiversity occurrences and we show that, despite improving technologies to collate and share them, they remain underused and are rarely gathered. It is yet a small price to pay to ensure that primary biodiversity data collected lately do not quickly become obsolete.

## **Keywords**

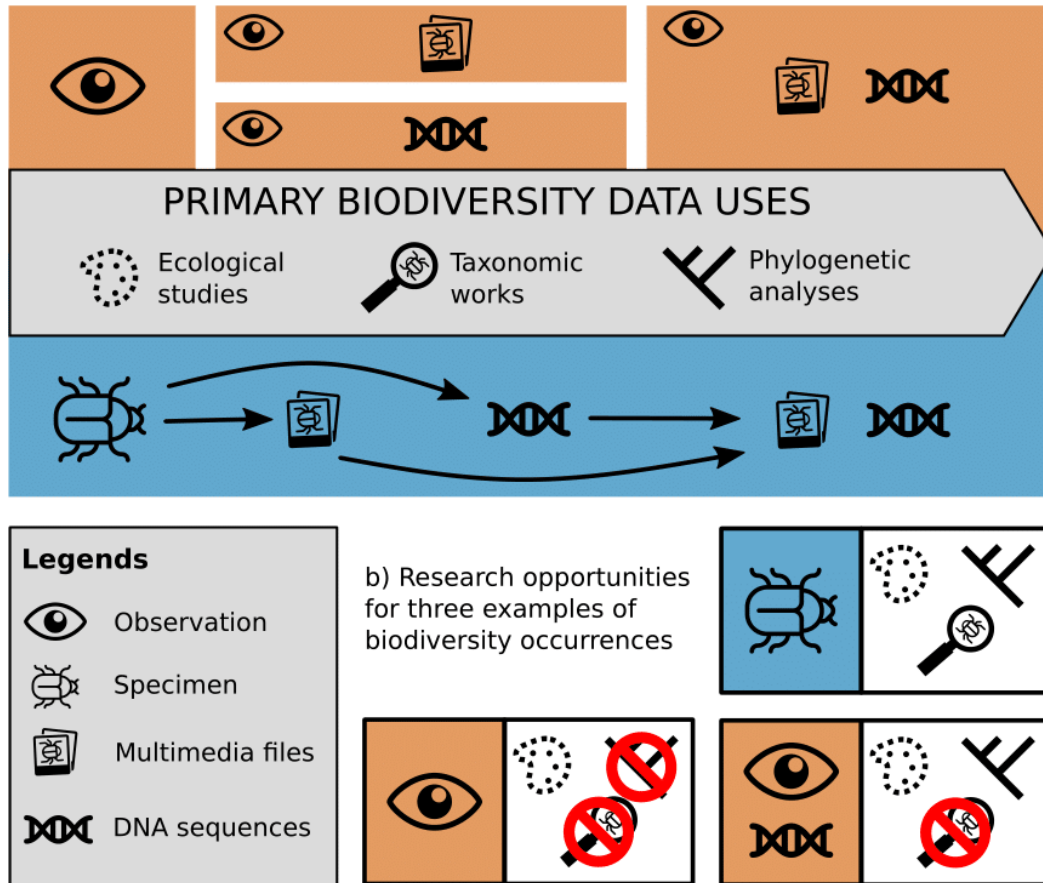
Primary biodiversity data, specimen, observation, database, ancillary data, biodiversity occurrences, big data

## Introduction

Primary biodiversity data, the bricks of systematics and evolutionary studies (May 1990; Funk and Richardson 2002; Hortal *et al.* 2015), are not gathered nowadays as they used to be. In the earliest days of systematics, specimens were collected methodically. Today, because of ethical and practical reasons partly imposed by the current biodiversity crisis, mere observation records, i.e. observations with no link to any tangible material, are mainly collated (Gaiji *et al.* 2013). Mere observations and vouchered specimens are biodiversity occurrences of different fundamental nature, each having assets and liabilities. Mere observations, for instance, are recorded and shared more rapidly than specimens are collected and digitalized. With mere observations, biodiversity data accumulates faster than ever (Bisby 2000; Kitchin 2014), but the link to specimens in natural history collections is being lost. We argue here that the change in biodiversity data gathering [from specimen-based (SB) to observation-based (OB) occurrences] has strong consequences in systematics and evolutionary biology and that it must be acknowledged and dealt with; the sooner, the better.

Biodiversity occurrences are not equivalent to one another and, according to their nature (SB or OB, old or recent, with ancillary data or not, etc.), they offer more or less research opportunities (Fig. 1). Generally, a biodiversity occurrence contains a taxonomic identification, localization and a date (Ariño 2010). These three pieces of information can be provided for SB or OB occurrences, and, in both cases, can be accurate or not, and more or less precise. Accuracy and precision mostly depend on the collector's skills and equipment, but they are also related to the nature of the primary biodiversity occurrence. In addition, a biodiversity occurrence, be it SB or OB, can be complemented with ancillary data such as pictures or samples, increasing the information content of biodiversity occurrences and their usefulness (Gaiji *et al.* 2013; Garrouste 2017; Fig. 1). Most ancillary data, however, cannot be gathered *a posteriori* of an OB occurrence, whereas it can be for a SB occurrence. Thus, the way primary biodiversity data are collected impacts their provided information content for current and future investigations.

a) The different natures and uses of biodiversity occurrences



**Figure 14: Illustrations of observation-based and specimen-based primary biodiversity occurrences and their potential uses.** a) Observations (orange) and voucher specimens (blue) can be complemented with ancillary data such as multimedia files or DNA sequences. For observations, these additional data must be acquired when the observation is performed; it cannot be performed later. On the opposite, for specimens – as long as they are well-curved, ancillary data can be gathered later (this advantage is symbolized through the continuous blue background and the arrows). b) Three hypothetical case studies – Because data can be acquired later, a specimen occurrence offers a wide range of studies and analyses. Conversely, for observation occurrences, the spectrum of analyses depends on the existence or not of ancillary data: a mere observation will not allow as many studies as an observation combined with a DNA sample (the red interdiction signs cover studies that cannot be achieved). Pictograms for specimen, observation, DNA and photos were designed by FreepiK from Flaticon.

This change in practice (from SB to OB occurrences) is a major upheaval with spectacular consequences for systematics and evolutionary studies. Since the very beginning of systematics, specimens have been collected and used to inventory the diversity of life and later to decipher the relationships within the tree of life (Giribet 2015). Natural history collections (NHC), which now support biodiversity, morphology or molecular databases, have been put together and used for species identification and description, comparative anatomy, and phylogenetic studies, to name a few practices embodying their usefulness (Kemp 2015; Buerki and Baker 2016; Fig. 1). Obviously, databases containing mainly mere observations would not be as profitable as data repositories composed of specimens but they have positive sides in return (e.g. the pace at which biodiversity occurrences are shared; datasets with higher statistical value, etc.) and can be complemented with diverse media. Can we then endorse this major change or is it too hazardous? As often, good legacy of previous practices and fruitful innovations must be retained and developed, while bad legacy must be put aside (Godfray 2002).

We argue that specimens belong to the good legacy and are too important to be put aside. Specimens offer a guarantee for repeatability in the study of biodiversity (Huber 1998; Schilthuizen *et al.* 2015; Turney *et al.* 2015), a fact that will resist all future conceptual and technical advances; it is timeless. The recent revolutions in systematics, i.e. the use of DNA and much recently the advent of next generation sequencing (NGS), illustrate this point because they rely on specimens (or samples). Even better, these technical advances are qualified as revolutionary because specimens are available to use them on, enabling us to engage in new research agenda (e.g. Anmarkrud and Lifjeld 2017). Similarly, in the era of phylogenomics, several authors have recently underlined the necessary revival of morphological studies in systematics, which, again, rely on specimens (e.g. Jenner 2004; Wiens 2004; Smith and Turner 2005; Yassin 2013; Pyron 2015; Wanninger 2015; Wipfler *et al.* 2016).

Beyond specimens, good practices about items providing intermediate information content (e.g. samples or pictures) should be advocated to assist the change in biodiversity data gathering (e.g. Garrouste 2017). Every data associated to an occurrence (be either a mere observation or a specimen) is an additional evidence to fight against one or several of the seven currently identified biodiversity shortfalls (Hortal *et al.* 2015). The Linnean shortfall,

the gap between the described species and the actual number of species, undoubtedly requires specimen collection (Dubois 2017; see Pape 2016 for an opposite opinion). But other shortfalls could be filled, in certain cases, as efficiently with samples or pictures rather than with specimens. A picture or a DNA sample of a well-known species would efficiently contribute to reduce the Prestonian shortfall, i.e. the lack of knowledge about the abundance of species and their population dynamics in space and time (Cardoso *et al.* 2011). When doubtful, and unlike with mere observations, the species attribution can be checked consulting the picture or sequencing DNA, so that observational occurrences with ancillary data can constitute appropriate datasets for evolutionary studies.

When a paradigm shift is on the way, measures are required to guide this shift and ensure its maximal usefulness now and in the future. Here, we demonstrate that a shift in the study of biodiversity (i.e. primary data are not SB anymore but mainly OB) is on the rise since several years and that it affects the fields of systematics and evolution. Analysing 536 million occurrences from the GBIF (Global Biodiversity Information Facility) in 24 taxonomic classes, we show empirically that this shift is widely shared across eukaryotes. From then on, because current decisions will shape the future and because one can anticipate negative outcomes for systematics and biodiversity research in general if this observation-trend remains unsupervised, we provide guidelines for primary biodiversity data gathering and sharing, guidelines easily met from individual research to broad citizen science programs.

## **Material and Methods**

### **Data Set**

We downloaded all the data available from the GBIF portal in June 2016 (<http://doi.org/10.15468/dl.hqesx6>). These 649 million occurrences were saved as a Darwin Core archive. Occurrences from this archive were extracted and imported into a SQL database, where data were indexed to reduce computation time of later queries. We focused on 24 taxonomic classes out of the 297 referenced in the GBIF, excluding the classes with less than 1 million occurrences (9.4 million occurrences, distributed into 19 thousands species, had no class affiliation). This filtering reduced the dataset to 626 million of occurrences (NBocc) and 1.01 million species, representing more than 96 % of the total

number of occurrences and 84 % of the total number of species in the GBIF. Finally, because we computed statistics over time, data without a year of collect were excluded. We ended up with 536 million occurrences, which is the dataset used to compute all statistics. A lag exists between an occurrence event recording and its integration in the GBIF database (S. Gaiji comm. pers.) and it might be related to the type of occurrences (i.e. specimen- or observation-based). Consequently, even though we show results until 2016, we avoid interpreting the last five years results to limit the risk of hazardous conclusions.

### **Data Quantity**

To calculate data quantity in the GBIF mediated data, the number of occurrences collected per year was counted. Then, a data accumulation curve was computed.

### **Data Origin**

In the GBIF, the origin of an occurrence can be specified using a controlled vocabulary in the *'basisOfRecord'* field. As in Troudet *et al.* (2017), we distinguished “specimen-based occurrences” linked to tangible material from “observation-based occurrences” (or disconnected observations). The category “specimen” regrouped: fossil specimen, living specimen, material sample, and preserved specimen. The category “observation” regrouped: human observation, machine observation, observation, and literature. A third category, corresponding to the option “unknown”, was also kept.

### **Supporting Files**

Supporting files (or links leading to such files) can be associated to an occurrence in the GBIF. They contribute to improve the traceability between a taxon’s name and a given occurrence. Two kinds of supporting files are mainly used: DNA sequences and multimedia files. For each of those supporting data, we computed 1) the quantity of both DNA sequences and multimedia files per year, and 2) the yearly numbers of DNA sequences and multimedia files divided by the yearly number of occurrences. This last number approximates (because a same occurrence can have several supporting files) the proportion of occurrence with supporting files.



To further understand the structure of the GBIF mediated data we also classified occurrences with supporting files according to their origin (i.e. '*basisOfRecord*'). Thus, we distinguished the number of specimen-based occurrences with multimedia supporting files from the observation-based and unknown occurrences with multimedia supporting files.

### **Evolution of Data Completeness**

Primary biodiversity data are all the more useful than they are associated to a lot of information. The DarwinCore format currently in use in the GBIF (Wieczorek *et al.* 2012) provides 234 columns to record information as diverse as the ethology of a living specimen or the geological strata of a fossil specimen. A complete occurrence would never require these 234 columns to be filled, because there are always inapplicable columns for a given occurrence. Nevertheless, the evolution of data completeness over time can be estimated from the evolution of the proportion of columns containing information. We thus averaged the proportion of non-null (non-empty) columns per occurrence per year.

### **Evolution of Taxonomic and Spatial Precision**

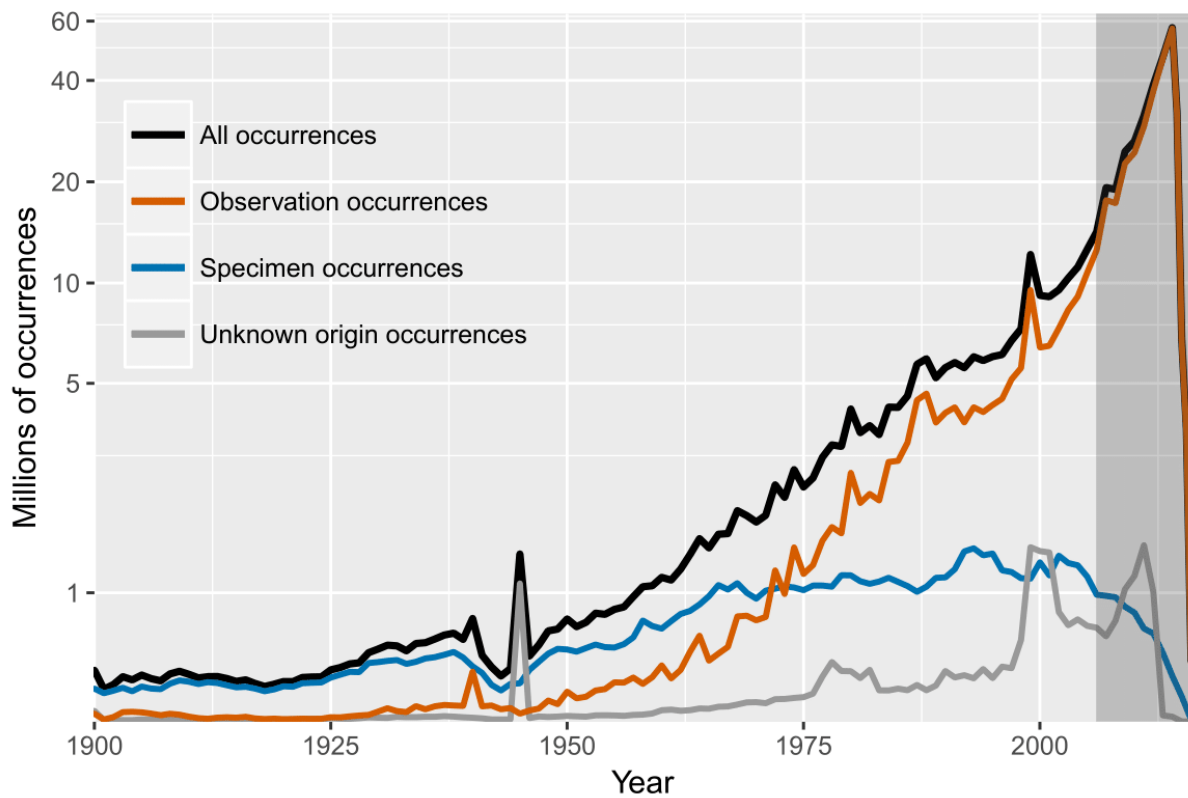
In general, a primary biodiversity data is associated to a scientific name, which can be more or less precise depending on the skills of the identifier but also on the state and availability of taxonomic knowledge. We estimated taxonomic precision (in number and proportion per year) differentiating occurrences identified at least at the species level from supra-specific occurrences. The proportion of occurrences identified at the species or infraspecific level was used to estimate the taxonomic precision of the GBIF mediated occurrences. As for the evolution of spatial imprecision, it was calculated as the number and proportion, per year, of occurrences lacking coordinates or flagged in the GBIF as data with coordinate issues.

## **Results and Discussion**

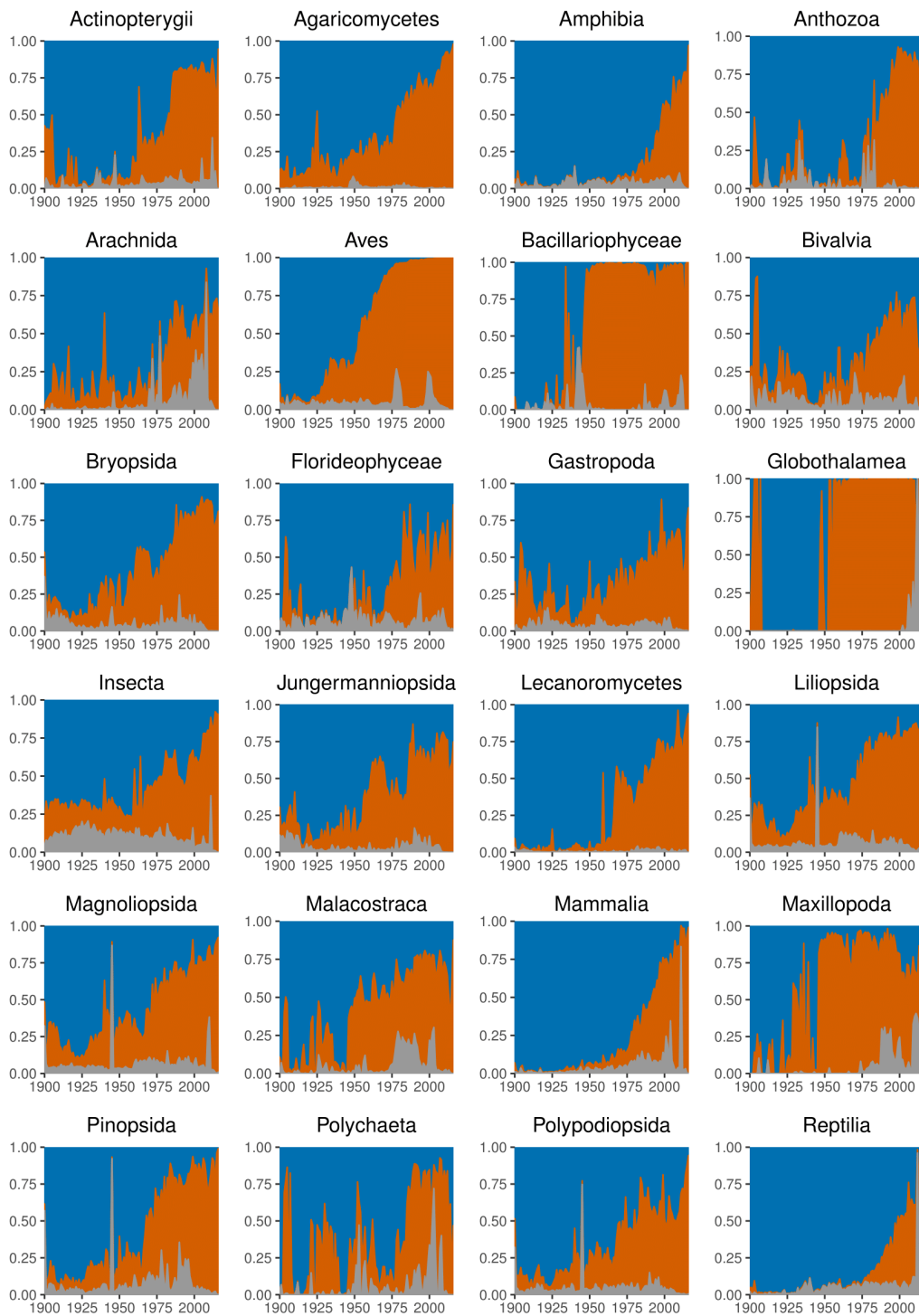
### **A Shift in the Recording of Primary Biodiversity Data**

In the current context of biodiversity crisis, numerous pleas have incited the scientific community to collect as much biodiversity data as possible, out of the fear it might disappear

before we even knew of its existence (May 2004; Butchart *et al.* 2010). These calls have been heard and, indisputably, biodiversity data accumulates faster than ever (Fig. 2 and Supporting Information), a trend most classes of organisms exhibit even though for a few of them the trend is not so strong (Troudet *et al.* 2017). The >57 million occurrences submitted to the GBIF in 2014, more than five times the amount of data submitted ten years earlier (i.e. 11 million occurrences in 2004), embody this report (Supporting Information). With this spectacular acceleration, the amount of data available to scientists is so huge that the study of biodiversity has entered into the “Big Data” era (Hampton *et al.* 2013; Joppa *et al.* 2016; Kelling *et al.* 2009). Multiple benefits followed such as an increased power in statistical analyses because of larger datasets or the possibility to tackle issues at large taxonomical, temporal or spatial scales (Rosenheim and Gratton 2017). However, the large volume of data is also a curation challenge that must be handled to avoid passing on a dubious source of knowledge to future generations because of a fall in data quality (Howe *et al.* 2008), a criticism regularly brought up for GBIF mediated data (e.g. Yesson *et al.* 2007).



**Figure 15: Number of primary biodiversity occurrences per year and origin from 1900 to today.** The plot shows that observation-based occurrences (orange) have outnumbered specimen-based occurrences since 1970 and that this excess is growing. Occurrences from the last ten years are shaded because the pace at which data are added within the GBIF portal, especially for specimen-based occurrences, likely affects them.



**Figure 16: Proportion of occurrences per year of collect and origin for a particular class.** Orange, blue and grey areas represent the proportions of observation-based, specimen-based and unknown origin occurrences, respectively. Contrary to 50 years ago, a majority of observation occurrences is reported whatever the taxonomic class.

This acceleration is triggered, at least partly, by a change in the way biodiversity data are recorded. The origin of biodiversity data has shifted from a majority of specimen-based (SB) to a majority of observation-based (OB) occurrences. This shift has been previously suspected (Gaiji *et al.* 2013) and we show here that, from 1970 to 2016, the proportion of occurrences traceable to tangible material (i.e. specimens) fell from 68 to 18 %. This result applies to the 24 classes studied, except for a few eccentric cases such as Globothalamea and Polychaeta (Figs. 2 and 3). Likely, these exceptions relate to specific practices for observing, collecting or curating these organisms, or to their low volume of primary biodiversity data, which might cast doubt on their atypical trends. Besides, this shift might be slightly inflated because it presumably requires less time to integrate OB than SB occurrences in the GBIF. Still, ignoring the last ten years to limit this potential bias (shaded area in Fig. 2), this shift remains striking. It started, for most of the organisms, in the second half of the 20<sup>th</sup> century and kept intensifying ever since. On the opposite, the number of SB occurrences has stagnated, at best, in the past 40 years. More worrying, most of SB occurrences cannot be readily traced back to a specimen: Only 238 000 occurrences have a filled “*materialsamplid*” column, representing only 0.28 % of the 84 million SB occurrences. This situation hampers the verification process, a founding step in scientific practice (Turney *et al.* 2015). Even though scientists can be delighted with the pace at which biodiversity data accumulates, they cannot be satisfied with a biodiversity research relying mainly on unverifiable observations.

Divergent causes, not necessarily exclusive, may explain this practice shift. In a context of massive biodiversity loss, a sense of urgency fueled the pleas for accelerated data collection (Hampton *et al.* 2013) and encouraged the accumulation of mere observation, less destructive and easier to produce, share and store than specimen-based occurrences. Ethical considerations and conservation issues that hinder specimen collections have commonly been put forward, although they are debatable in some situations (Dubois 2017; Löbl 2017). Concurrently, Grandcolas (2017) suggested that this shift started when biodiversity sciences merged with general biology, more interested in discovering general patterns and laws than in documenting diversity. Others underlined the lack of human and economic resources to ensure both the gathering of specimens and the curation of natural history collections (Kemp 2015). These reasons favored a decrease in specimen collection. On the other hand, the number of observation-based occurrences has dramatically increased with, for instance, the

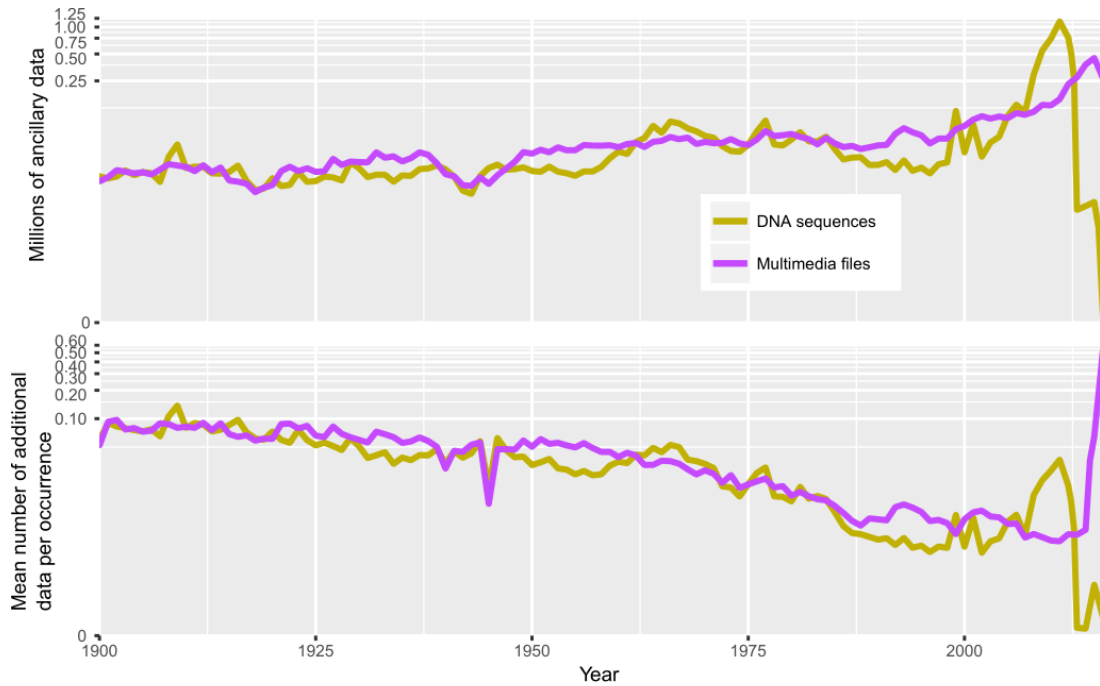
rise of citizen science that enable to rapidly produce a vast amount of observational data (Dickinson *et al.* 2012). Given the multiple origins of this trend, it seems unlikely to be reversed in the near future and must be organized and guided to ensure maximal benefits for the study of biodiversity.

### **Primary Biodiversity Data for systematics and evolutionary studies in the 21<sup>st</sup> Century: Are We There Yet?**

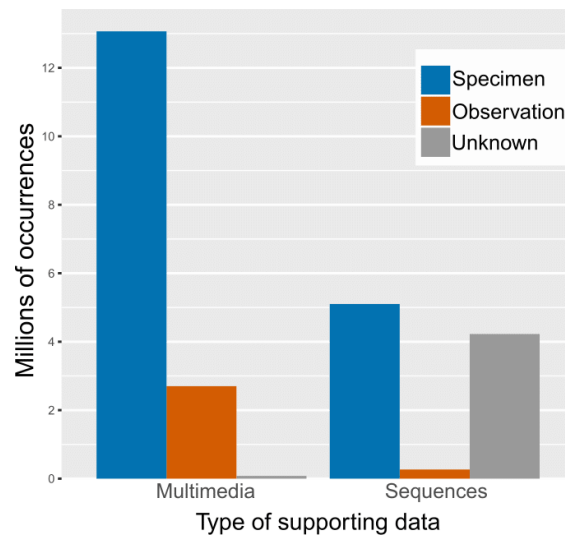
The importance of collecting specimens in taxonomy, evolution and ecology cannot be overemphasized (Huber 1998; Schilthuizen *et al.* 2015) and two main points, previously discussed in the literature, must be reiterated. First, specimens are needed for species description and for the study of biodiversity in general (Dubois 2017 *contra* Pape *et al.* 2017). A crucial argument is the utility of specimens for checking species identification. Goodwin *et al.* (2015) assessed that up to half of tropical plant identifications in museum collections were false. Correcting identification errors can be done after examining specimens, but is impossible for mere observations. If Goodwin *et al.*'s estimation is correct and generalizable to all primary data, the need for specimens, or at least ancillary data to observation occurrences, is critical. Second, the revived focus on morphology advocated lately in systematics requires specimens (Jenner 2004, Wiens 2004, Smith and Turner 2005, Yassin 2013, Pyron 2015, Wanninger 2015, Wipfler *et al.* 2016). Authors recommending this revival underlined that comparative morphology not only brings phylogenetic characters but also allows including fossil taxa in phylogenetic analyses (e.g. Pyron 2011; Wood *et al.* 2013), enabling us to better estimate the structure and branch length of the reconstructed trees (Wiens *et al.* 2010; Pyron 2015). Given that phylogenetic thinking has become of paramount importance in biology, improvements in phylogenetic estimation offer large potentialities in evolutionary studies and in the study of biodiversity in general (Losos *et al.* 2013; Buerki *et al.* 2015).

However, a specimen is not mandatory for a primary biodiversity data to be useful. Instead of specimens, and in complement to mere observations, digital data or molecular data can be collated. New technologies offer a wide range of tools and methods to collect concrete specimen evidence in nature, and it is now relatively easy and affordable to obtain DNA sequences, images and sound recordings. Then, using molecular and digital data should now

be a common practice in the study of biodiversity, as the exponential growth of molecular data and phylogenies, and the development of morphological databases and ontogenies would suggest (Lathe *et al.* 2008; Parr *et al.* 2012; Deans *et al.* 2012, 2015). We show here that digital and DNA data are increasingly used but these data remain patently underemployed (Fig. 4). Only 2.5 % of all the GBIF-mediated occurrences for the 24 focal classes were linked to digital data and 1.5 % to DNA sequences. Worse, proportionally, they become more and more negligible, drowned in the large quantity of observations without supporting data. This situation might be improving lately, but the post-2008 tendency observed demands to be confirmed in future years (Fig. 4). Moreover, and quite inconsistently, digital and DNA data were less used for OB than for SB occurrences (Fig. 5). They would yet be more useful for OB biodiversity data given that they would constitute the only way to independently check or update observation occurrences, whereas one can refer to specimens, as long as those are kept and the traceability chain is not broken, for SB occurrences (Page 2015; Nualart *et al.* 2017). The high proportion of sequences associated to primary biodiversity data of unknown origin could suggest that when a sample is performed, occurrences are often classified in the catch-all class ‘unknown origin’.



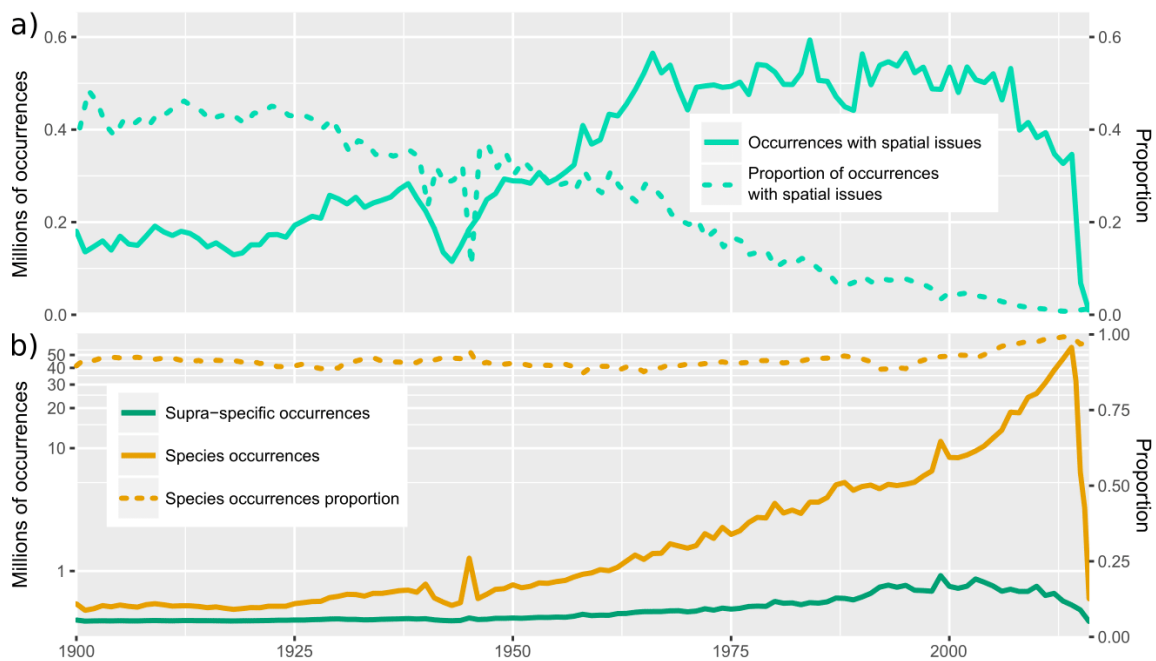
**Figure 17: The increase of ancillary data to biodiversity occurrences does not keep pace with biodiversity data accumulation.** The top plot shows a yearly report of the number of multimedia files (purple curve) and DNA sequences (green curve) linked to occurrences. The bottom plot shows the proportion of occurrences with multimedia files and DNA sequences.



**Figure 18: Occurrences with ancillary data are mainly specimen occurrences.** Occurrences with multimedia files (left) are mainly specimen-based (blue), whereas occurrences with DNA sequences (right) are either specimen-based or of unknown origin (grey). Very few observations-based occurrences (orange) are provided with ancillary data.



In addition to ancillary data, the usefulness of primary biodiversity occurrences can be maximized through a higher level of precision and completeness in recordings. We expect biodiversity data occurrences to be more precise and complete now than before because tools that are more efficient have been developed. Whatever the nature of the occurrence, spatial coordinates for instance can be easily provided with a high precision level given the democratization of GPS. Data completeness should also improve because of the growing awareness that a global and comprehensive picture of biodiversity is needed. Our results showed that, in proportion, data precision does improve but that data completeness stagnates (Fig. 6 and Supporting Information). The proportion of data with geospatial issues in the GBIF (i.e. data with low spatial precision) decreased from 50.2 % in 1900 to 0.6 % in 2014 in spite of a larger number of occurrences with spatial imprecision – this number being quite stable over the past 30 years (Fig. 6A). Over the same period, records identified at the species level augmented from 89.6% to 99.4%, with once again an increase of supra-species records (Fig. 6B). While species identification and spatial precision improves, so does niche modeling results for instance, promising significant advances in historical biogeography (e.g. Meseguer *et al.* 2015; Töpel *et al.* 2016). In this regard, important gains for systematics and evolutionary studies can be anticipated from the increasing level of precision in primary biodiversity data.



**Figure 19: a) Spatial and b) taxonomic precision in the GBIF mediated data improves over time in proportion.** The plot a) shows the number of occurrences collected each year lacking coordinates or tagged as having geospatial issues in the GBIF (plain line). Yet, the proportion of those occurrences is decreasing (dashed line). The plot b) shows the number of occurrences identified at least at the species level (yellow curve) or at a higher taxonomic rank (green curve). The number of occurrences identified at a higher taxonomic rank is increasing with time. Yet, the proportion of occurrences identified at least at the species level (dashed yellow line) is increasing.

Given the progresses of technology and the proportion of people owning smartphones with photo and GPS capabilities, targeting a higher level of completeness in biodiversity data is legitimate but the reasons and the necessity of this objective must be well-advertised, a task that falls to scholars. They have the power to modulate the current trend, demanding a minimal amount of ancillary data when designing their personal or collaborative research projects, including citizen science programs. Taking pictures or samples, not necessarily systematically but more often than now, should be part of the scientific protocol. This will never replace the wealth that specimens in natural history collection offer (Funk and Richardson 2002; Buerki and Baker 2016) but would limit the risk that entire datasets become

useless when data inaccuracy is suspected. Whatever its nature and quantity of ancillary data, primary biodiversity data must be made available, and this evolution would require the adequate infrastructures to support the massive amount of data one can foresee. Several data storage and compression options are currently investigated (e.g. Marx 2013; Numanagic *et al.* 2016), which suggests it will not be an insurmountable hurdle. The costs that should be deployed are substantial but are worth it for evolutionary biologists and for the society. Besides, these efforts would result in large image and DNA databases, whose usefulness, accuracy and search efficiency would augment together with their supply, as a virtuous circle.

The fear of biodiversity disappearance has triggered a vague of biodiversity data accumulation. We are in the middle of a paradigm change where biodiversity data are not anymore gathered like it used to be. This paradigm change has been undergone without any supervision. Even though some aspects of these changes are highly beneficial, others are suboptimal, to say the least, and must not be ignored. We must act now to allow a better monitoring of the biodiversity research agenda and to continue shaping how biodiversity data should be gathered, diversifying the objects of collection (e.g. specimens, samples, DNA, images, etc. – Knapp 2015). We argue that ancillary data (samples, DNA, pictures) must be collected more methodically than today, to avoid disillusionment when we will realize that mere observations were not sufficient to address current and future preoccupying issues about systematics and evolutionary studies (Joppa *et al.* 2016).

## **Acknowledgments**

This study was developed as part of a Ph.D. project and was funded as a grant by the Ministère de la Recherche to JT. We would like to thank Mark Judson for his comments on an early draft of this manuscript. We thank Anne-Sophie Archambeau, Samy Gaiji, Marie-Elise Lecoq, Sophie Pamerlon, Roseli Pellens, Tim Robertson, Dmitri Schigel, Jérôme Sueur and Wilfried Thuiller for fruitful discussions.

### **Chapter 3: Taxonomic bias in biodiversity data and societal preferences (Troudet *et al. Scientific reports*, published 22-08-2017)**

The trend of primary biodiversity data to rely increasingly on observations is worrying. However, this situation has the merit of allowing the rapid production of these data. Logically, this should benefit the study of biodiversity as a whole. However, analyses of GBIF mediated data have again revealed issues that may affect such studies, particularly for certain geographical areas and taxa.

The first of these trends that I have been able to characterize in the GBIF is an important spatial bias in favor of certain areas of the globe. This sampling bias, which was very prevalent in GBIF-managed data, made Europe and North America particularly rich in occurrences, while other areas such as Russia and Africa were much less explored. However, I did not continue to explore this bias as it was characterized in detail by Meyer *et al.* (2015) shortly after my first work on it. However, this was not the only bias present in this dataset.

Some taxa are better known than others, especially for model species that have extensive literature on them (Grandcolas 2017). However, even at wider taxonomic scales a bias in favor of certain taxa has been observed for a long time (Gaston and May 1992). This taxonomic bias was never characterized on a very large scale and is often treated at the discipline level (e. g. Feeley *et al.* 2016, Di Marco *et al.* 2017) or in the case of a particular taxon (Cardoso *et al.* 2011). The study of GBIF data was therefore an ideal situation to work on this bias, especially since primary biodiversity data are the basis of ecological and systematic research, which makes this bias potentially disabling for the knowledge of biodiversity as a whole.

Given the potential consequences of this bias, I chose to characterize it before working on biodiversity patterns. I and the co-authors of the following study did not stop at characterizing this bias. We have attempted to find an explanation for it by exploring the influence of research and the general public on the amount of data available for a taxon. Once again, this work was carried out with my two thesis directors Régine Vignes-Lebbe and

Frédéric Legendre as well as Philippe Grandcolas and Amandine Blin. This work was submitted to Scientific Reports in April 2017 and published in the same journal in August 2017. The full article is therefore the essence of this chapter. I integrated the figures originally proposed in the supplementary materials of the article.

*Julien Troudet<sup>1</sup>, Philippe Grandcolas<sup>1</sup>, Amandine Blin<sup>2</sup>, Régine Vignes-Lebbe<sup>1</sup> and Frédéric Legendre<sup>1</sup>*

1. Institut de Systématique, Evolution, Biodiversité, ISYEB – UMR 7205 MNHN CNRS UPMC EPHE, Sorbonne Universités, 45 rue Buffon, 75005, Paris, France.
2. Outils et Méthodes de la Systématique Intégrative, OMSI – UMS 2700 CNRS MNHN, Muséum national d’Histoire naturelle, CP26, 57 rue Cuvier, 75231, Paris Cedex 05, France.

Régine Vignes-Lebbe and Frédéric Legendre jointly supervised this work. Correspondence and requests for materials should be addressed to J.T. (email: [julien.troudet@mnhn.fr](mailto:julien.troudet@mnhn.fr))

## **Abstract**

Studying and protecting each and every living species on Earth is a major challenge of the 21st century. Yet, most species remain unknown or unstudied, while others attract most of the public, scientific and government attention. Although known to be detrimental, this taxonomic bias continues to be pervasive in the scientific literature, but is still poorly studied and understood. Here, we used 626 million occurrences from the Global Biodiversity Information Facility (GBIF), the biggest biodiversity data portal, to characterize the taxonomic bias in biodiversity data. We also investigated how societal preferences and taxonomic research relate to biodiversity data gathering. For each species belonging to 24 taxonomic classes, we used the number of publications from Web of Science and the number of web pages from Bing searches to approximate research activity and societal preferences. Our results show that societal preferences, rather than research activity, strongly correlate with taxonomic bias, which lead us to assert that scientists should advertise less charismatic species and develop societal initiatives (e.g. citizen science) that specifically target neglected

organisms. Ensuring that biodiversity is representatively sampled while this is still possible is an urgent prerequisite for achieving efficient conservation plans and a global understanding of our surrounding environment.

## **Introduction**

Since the first Convention on Biological Diversity in 1992, biodiversity and the consequences of its destruction have become a central concern for biologists (Díaz *et al.* 2006, Cardinale *et al.* 2012, Gascon *et al.* 2015). From scientists to the lay public or policy makers and practitioners, the need to study and protect biodiversity is growing, and scientists have shown that it must be tackled at the gene, species and ecosystem level (Mace *et al.* 2012). Within a context of global change and accelerated biodiversity loss, this necessity has become a major concern and challenge for the 21st century (Dirzo and Raven 2003, Ceballos *et al.* 2015). However, discussions on biodiversity often only focus on a small subset of species, while most of the eukaryotic biodiversity remains unknown or ignored (Feeley *et al.* 2016, Di *et al.* 2017).

Taxonomic bias, also referred to as taxonomic chauvinism (Bonnet *et al.* 2002), is pervasive in biodiversity research. This bias stems from disparities in our knowledge of different organisms, and in the extent to which they are the focus of scientific research, across a wide range of biological disciplines. Some organisms – mostly plants and vertebrates – are over-represented in various scientific fields (Feeley *et al.* 2016, Bonnet *et al.* 2002, Clark and May 2002a), are more likely to raise funds (Leather 2009), or are considered ecologically more important than others (Ford *et al.* 2017). It has been shown, however, that focusing on a few, often charismatic, species, prevents reaching global conclusions and developing efficient conservation plans (Feeley *et al.* 2016, McKinney *et al.* 1999, Seddon *et al.* 2005). Rare, small or uncharismatic creatures do play pivotal functions in ecosystems (Lawler *et al.* 2003, Mouillot *et al.* 2013). In addition, biomimicry, i.e. the application of the properties of living organisms to technology, and bioprospecting activities, i.e. the search for new natural products in wild species, cannot be performed efficiently when less than 1% of known species have been carefully studied (Wilson 2000). Thus, given its scientific and societal significance, describing taxonomic bias in the study of biodiversity and understanding its underlying causes are undeniable priorities.

Taxonomic bias in science has long been recognized (Clark and May 2002b, May 1988, Gaston and May. 1992) but its origin is less clear. Obviously, some organisms are more difficult to study than others because they live in remote habitats, are local endemics or are microscopic and difficult to identify (Pawar 2003). But these intrinsic features alone cannot fully explain the pervasive taxonomic bias observed in science. Two hypotheses on the role of two extrinsic factors can then be put forward: the ‘taxonomic research’ hypothesis and the ‘societal preferences’ hypothesis. The ‘societal preferences’ hypothesis suggests that societal interests influence and bias the choice of study organisms (Stahlschmidt 2011, Wilson *et al.* 2007). The ‘taxonomic research’ hypothesis implies that scientific reasons and limitations will lead and orientate biodiversity data gathering (Pawar 2003). Because of the intricate interactions between scientists, citizens and funding agencies, and their mixed influence (Martín-López *et al.* 2009), the underlying mechanisms are unclear. Nevertheless, these hypotheses deserve to be explored and confronted at a global taxonomic scale. Moreover, the recent development of citizen science (Chandler *et al.* 2017) may have increased the impact of societal preferences. Here, to investigate the relative impact of ‘societal preferences’ and ‘taxonomic research’ on biodiversity data, we used the number of webpages from Bing searches and the number of publications retrieved from Web of Science, as proxies, respectively (see Methods).

The study of biodiversity is a daunting task – ca. 10 million species are estimated to inhabit the planet – that requires deploying a considerable workforce to gather and analyse biodiversity data (Costello *et al.* 2013). Fortunately, for ethical and scientific reasons (Michener and Jones 2012, Duke and Porter 2013, Peterson *et al.* 2015), data sharing practices and tools like the Global Biodiversity Information Facility (GBIF) were developed, facilitating access to species occurrence records. The GBIF portal was chosen because it hosts the biggest open access primary biodiversity database and, even though the big data paradigm does not mean that big datasets are devoid of flaws, they offer a significant potential for new and broad insights (Boyd and Crawford 2012). Moreover, although open access primary biodiversity data are heterogeneous, resulting from the good will of contributors and not from a well-planned sampling protocol<sup>30</sup>, they reflect our knowledge and practices in the study of biodiversity. Thus, they can be used to investigate taxonomic bias on a large geographical and taxonomic scale.

Here, we aim to quantify taxonomic bias in biodiversity data using 626 million of GBIF-mediated occurrences covering 24 classes of organisms. After careful data validation procedures, we characterized biodiversity gaps, a necessary first step before trying to bridge these gaps (Faith *et al.* 2013). We did not assess the validity of GBIF mediated data, which is an issue that has already been raised and discussed repeatedly (Yesson *et al.* 2007, Gaiji *et al.* 2013, García-Roselló *et al.* 2014, Maldonado *et al.* 2015, Sikes *et al.* 2016). Instead, we quantified taxonomic bias and imprecision (i.e. when an occurrence has been identified not at the species level but only at a higher taxonomic level) and related them to information provided in the occurrence records information (data origin, record date and locality). We tested the relative impact of societal preferences and taxonomic research on taxonomic bias, using public interest (i.e. the number of webpages) and taxonomic research quantity (i.e. the number of publications) as explaining variables, respectively. Opposing these hypotheses enabled us to suggest future directions for developing strategies for representative sampling of biodiversity while this is still possible.



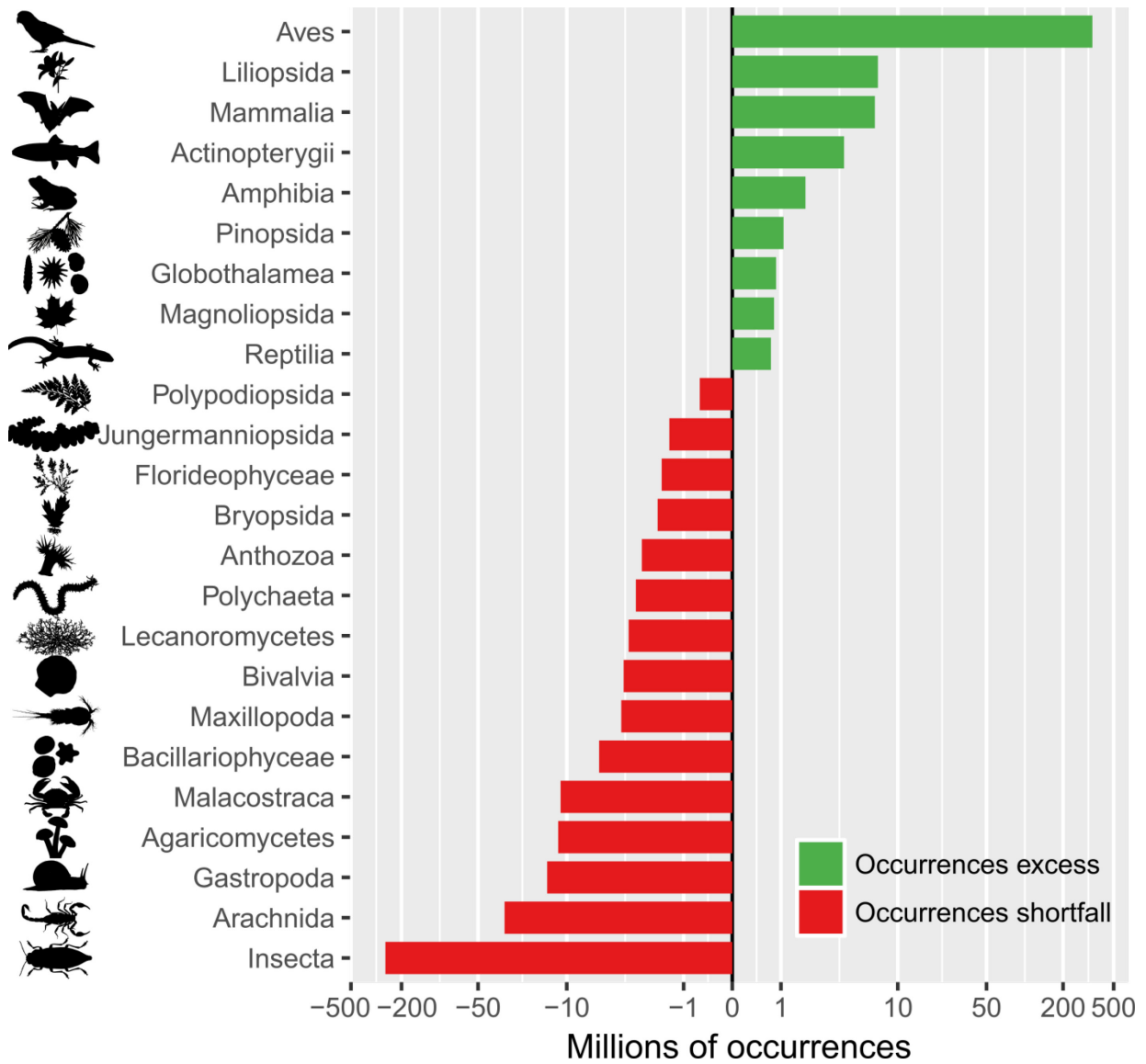
**Table 2. Biodiversity occurrence data statistics for 24 taxonomic classes.** The number of occurrences ( $nb_{occ}$ ) was obtained before the use of any filter. The number of species ( $p_{>1}$ ) corresponds to the number of unique scientific names having at least one occurrence. In bold are the eight classes selected to study the taxonomic bias at the ordinal level.  $med_{sp}$  is the median number of occurrences per species and  $mad$  is the associated median deviation. Taxonomic precision is the proportion of taxa identified at least at the species level.

	$nb_{occ}$ (millions)	$p_{>1}$ (thousands)	$med_{sp}$ ( $mad$ )	Taxonomic precision
<b>Aves</b>	345.11	12.82	371 (541)	0.99
<b>Magnoliopsida</b>	118.21	261.01	19 (25)	0.92
<b>Insecta</b>	46.78	352.78	3 (3)	0.77
Liliopsida	36.75	68.99	15 (19)	0.95
Actinopterygii	14.18	30.73	27 (37)	0.92
<b>Mammalia</b>	10.78	11.53	15 (21)	0.88
Bryopsida	6.06	18.85	7 (9)	0.95
Gastropoda	5.85	46.99	7 (9)	0.69
<b>Reptilia</b>	4.98	11.30	24 (34)	0.88
<b>Lecanoromycetes</b>	4.97	17.79	8 (10)	0.93
Polypodiopsida	4.91	12.65	23 (31)	0.95
<b>Amphibia</b>	3.94	5.89	54 (76)	0.91
<b>Agaricomycetes</b>	3.80	23.53	4 (4)	0.93
Malacostraca	2.73	30.16	6 (7)	0.73
Globothalamea	2.68	4.07	10 (13)	0.74
Arachnida	2.17	38.11	3 (3)	0.77
Bivalvia	2.02	14.02	9 (12)	0.70
Bacillariophyceae	1.96	11.19	2 (1)	0.70
Maxillopoda	1.87	9.98	4 (4)	0.58
Pinopsida	1.57	0.91	110 (160)	0.95
Jungermannniopsida	1.41	6.93	7 (9)	0.91
Polychaeta	1.29	8.77	6 (7)	0.73
Florideophyceae	1.07	5.78	17 (24)	0.88
Anthozoa	1.03	8.64	7 (9)	0.59
<i>Total for 24 classes</i>	<i>626.13</i>	<i>1013.39</i>	<i>7 (9)</i>	<i>0.94</i>
<i>Total in the GBIF</i>	<i>649.79</i>	<i>1200.38</i>	<i>6 (7)</i>	<i>0.93</i>

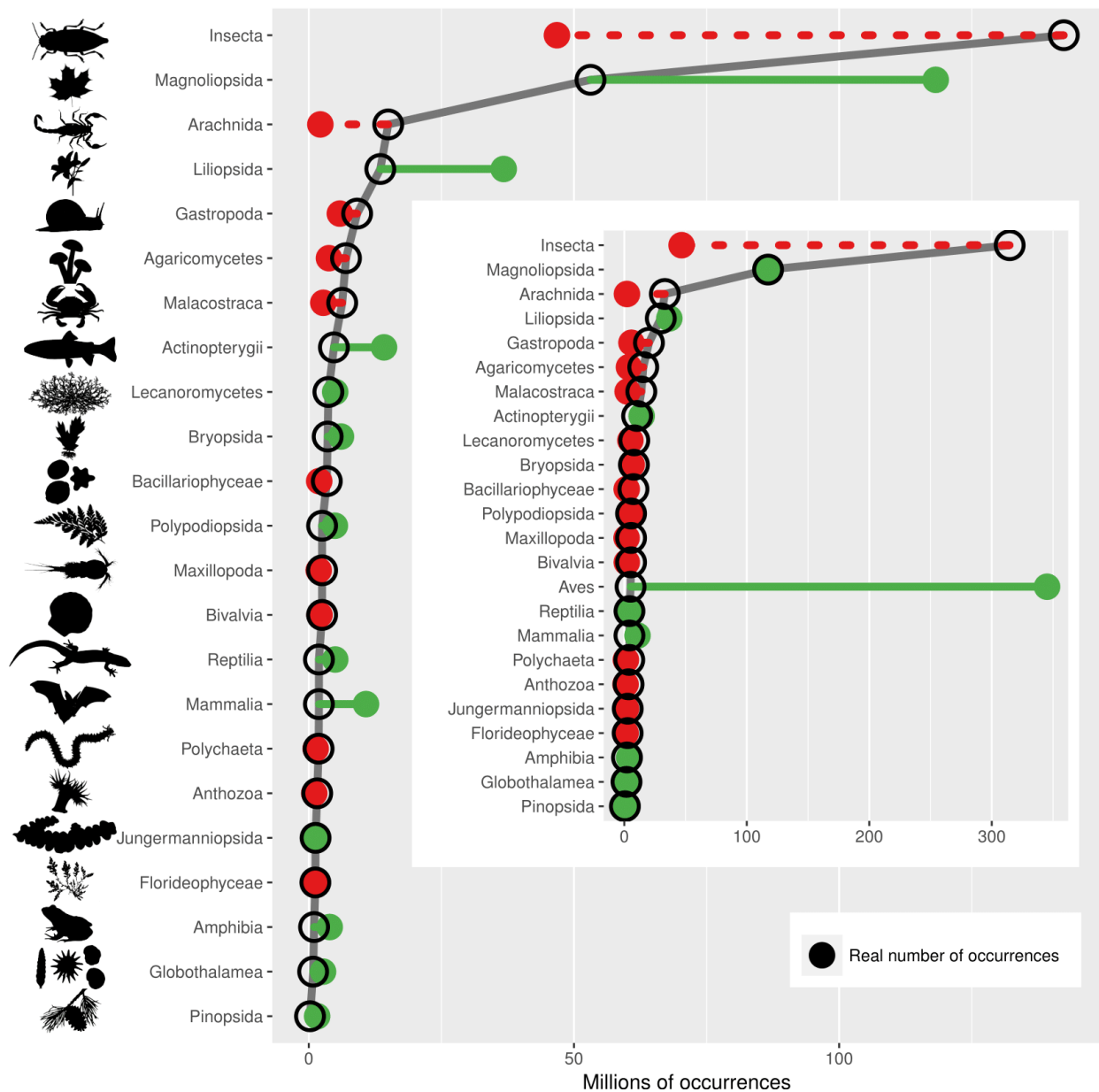
## Results

Global taxonomic coverage and taxonomic precision. 24 classes of organisms recorded in the GBIF database had more than 1 million occurrences, with widely variable numbers of occurrence recordings (Table 2). More than half of the records were bird (Aves) occurrences (345 million occurrences; 53% of the GBIF mediated data), even though birds represent only 1% of the total number of species catalogued in GBIF. Aves was also the class with the highest median number of occurrences per species (med/sp = 371). By contrast, and despite being three times more speciose, Arachnida had only 2.17 million occurrences and one of the lowest median numbers of occurrences per species (med/sp = 3). The lowest values of the median number of occurrences per species (i.e. below 7) were found for several classes of Arthropods (Insecta, Maxillopoda, Arachnida, Malacostraca), some fungi (Agaricomycetes) and diatoms (Bacillariophyceae). Magnoliopsida and Insecta, two highly speciose classes, were the ones with the highest number of species recorded. Only six of the 24 classes had a median number of occurrences per species higher than 20.

With regard to taxonomic precision, 94% of GBIF occurrences were identified (at least) at the species level (88% not counting Aves). The lowest levels of taxonomic precision were found in Maxillopoda and Anthozoa (58% and 59% of occurrences, respectively), whereas the highest levels were found in the different classes of Plantae (91 to 95% of occurrences in Magnoliopsida, Liliopsida and Pinopsida), Fungi (93% in Agaricomycetes and Lecanoromycetes) and Aves (99%).



**Figure 20. Taxonomic bias in biodiversity occurrence data.** The vertical line at  $x = 0$  depicts the ‘ideal’ number of occurrences per class, where each class is sampled proportionally to its number of known species. Green and red bars show the classes that are over- and under-represented in the GBIF mediated database compared to this ‘ideal’ sampling, respectively. Insects lack >200 millions occurrences and birds have an excess of >200 millions occurrences compared to an unbiased taxonomic sampling. Because birds and insects are greatly over- and under-represented, respectively, an inverse hyperbolic sine transformation was used for the x-axis.



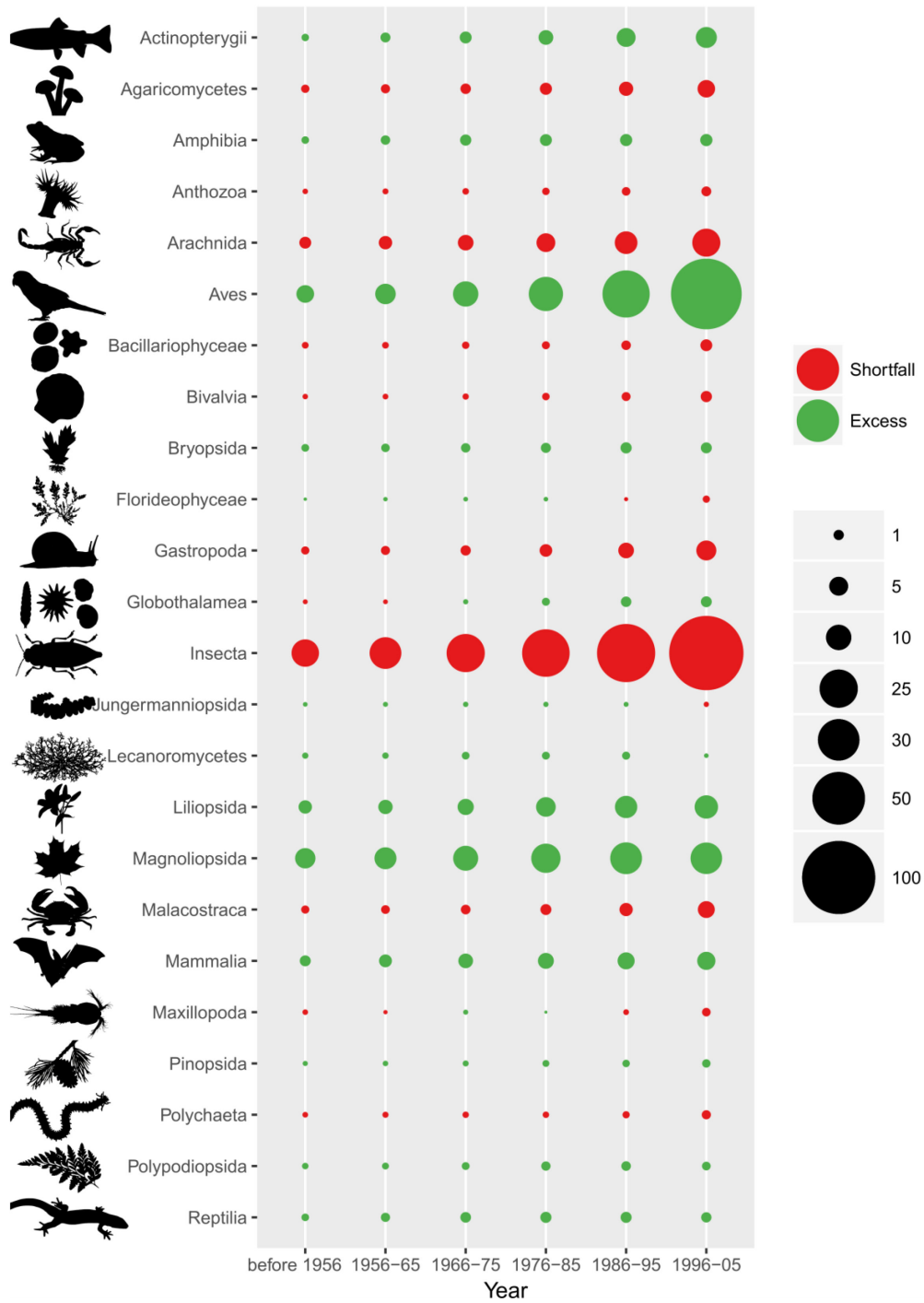
**Figure 21: Taxonomic bias in biodiversity data occurrences.** The grey line and black circles represent the ‘ideal’ number of occurrences per class, wherein each class is sampled proportionally to its number of known species. Green and red symbols show classes that are over- and under-represented in GBIF mediated data with regard to this ‘ideal’ sampling, respectively. The green and red dots represent the real number of data in the GBIF. The insert also depicts the taxonomic bias but includes Aves, the most over-represented class, in the calculation.

**Taxonomic bias.** Of the 2.2 million of species referenced in the GBIF taxonomic backbone, 1.2 million species can be found in the GBIF published datasets and 1.01 million belong to the 24 classes selected here. The number of recorded species per class was not proportional to their known species richness, highlighting a strong taxonomic bias. Aves and Insecta were, by far, the most over- and under-represented classes, respectively. Mammalia, Liliopsida, Actinopterygii, Amphibia and Magnoliopsida were also over-represented, whereas Arachnida, Gastropoda, Agaricomycetes, Malacostraca and Bacillariophyceae were under-represented (Fig. 20 and 21). This taxonomic bias was already apparent more than 50 years ago, meaning that classes that were over- or under-represented in the 1950's are still over- or under-represented today (Fig. 22). Nonetheless, we found an increase in taxonomic bias over time, mostly due to the faster accumulation of data for birds compared to other classes (Fig. 23 top; 283 million bird occurrences recorded between 2000 and 2016). Recently, data has accumulated faster than ever before for most classes (Fig. 23 top, middle and Fig. 24) however, for Amphibia, Reptilia and Florideophyceae, the number of occurrences recorded per year has stagnated or even declined over the past 40 years (Fig. 23 bottom).

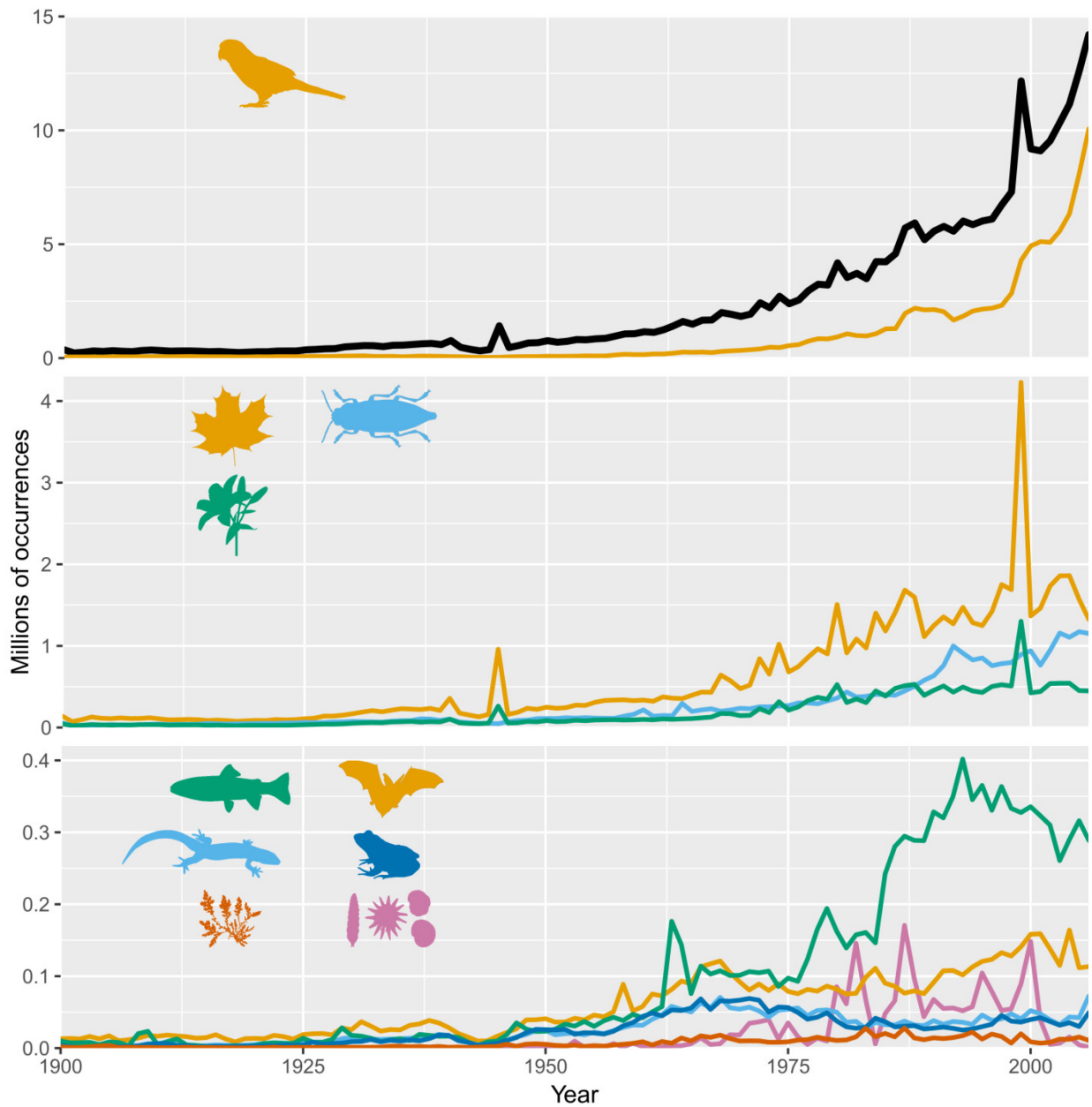
Twenty out of 24 classes had more than 50% of their described species referenced at least once in GBIF, and, for 14 of these classes, these statistics rose to 70% or more. By contrast, only 35% of Insecta and 36% of Arachnida species were referenced at least once in GBIF (Fig. 25 top). Furthermore, species were more or less intensely recorded in GBIF: 21% had only one occurrence (i.e. 212,911 species), 44% had between 2 and 19 occurrences (i.e. 446,643 species), and 35% had 20 or more occurrences (i.e. 353,843 species). This density of recordings per species was unevenly distributed between classes (Fig. 25 top). Only three classes (Aves, Amphibia and Actinopterygii) had more than half of their species with at least 20 occurrences, and only Aves had more than half of its species “decently” sampled (i.e. with 20 spatially distinct occurrences). This contrasted strikingly with the Arthropod classes, where, at best, 9% of species were “decently” sampled, even though Malacostraca had 68% of its species recorded in the GBIF.

This taxonomic bias recurs at a lower taxonomic scales. We selected eight classes and showed that, for all of them, some orders were better represented in the GBIF-mediated database than others (Table 3). For instance, the median number of occurrences varied largely within each class, some orders having medians that were more than 50 times higher than those

of other orders of the same class (e.g.  $m_{\text{Phaethontiformes}} = 5504$  vs  $m_{\text{Sphenisciformes}} = 2$ ;  $m_{\text{Chiroptera}} = 107$  vs  $m_{\text{Cetacea}} = 2$ ). The smallest difference in medians was found within poorly represented classes, in which all orders have medians less than 20. Taxonomic precision was also estimated and found to be highly heterogeneous between orders of the same class. The largest differences were observed within Insecta. More than 90% of occurrences were identified at the species level for four orders (Siphonaptera, Odonata, Orthoptera and Psocodea), whereas taxonomic precision ranged from 35 to 0.5% for Grylloblattodea, Mantophasmatodea and Strepsiptera. Taxonomic precision within Mammalia was also very heterogeneous ranging from 22% (Perissodactyla) to 99% (Monotrema and Notoryctemorphia). Conversely, taxonomic precision was less variable between orders of Lecanoromycetes (over 89% taxonomic precision for all orders), Magnolopsida (82% and above) and Aves (77% and above).

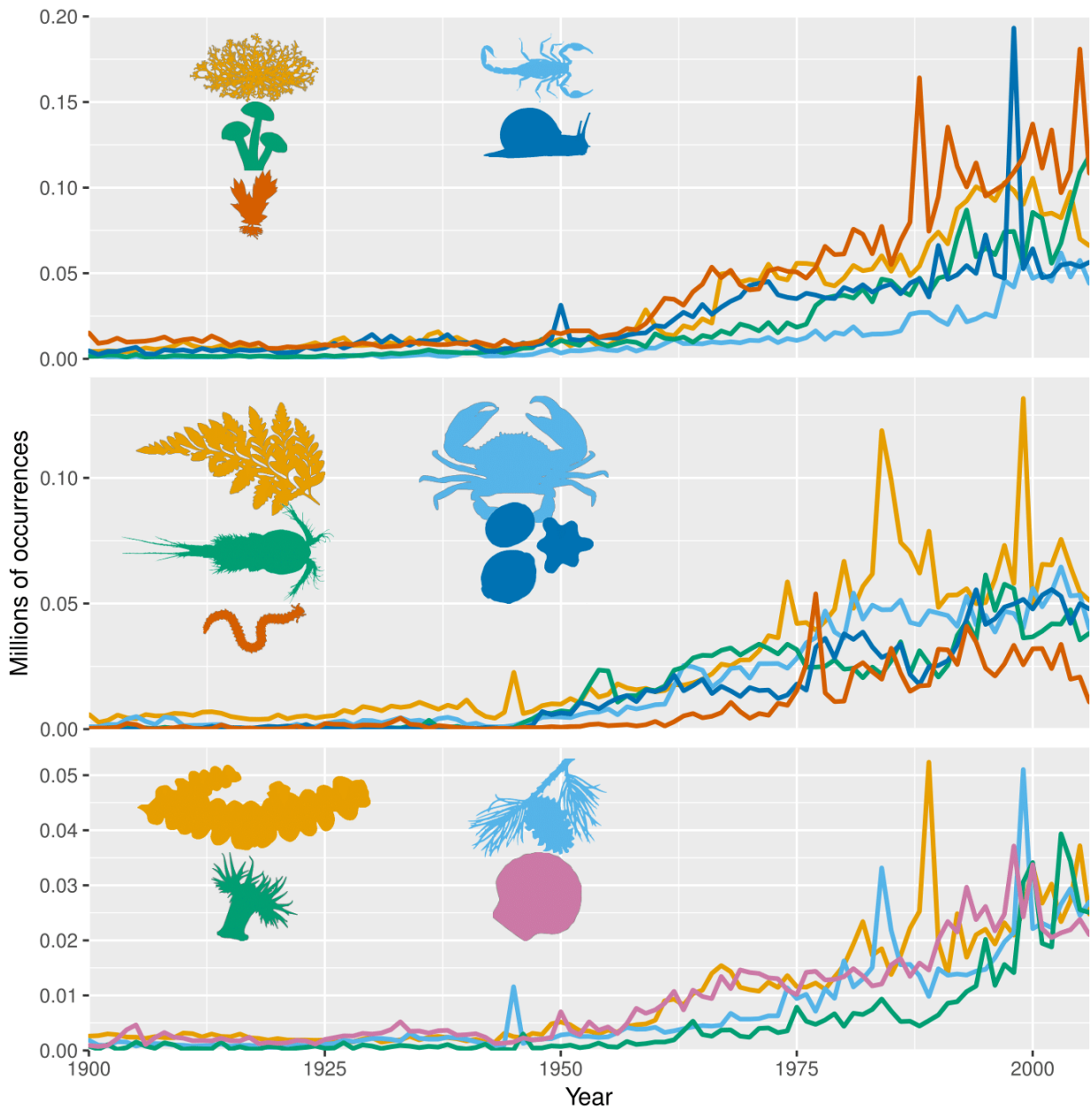


**Figure 22. Evolution over time of the taxonomic bias for each class.** The larger the circle, the higher the deviation from I, the ‘ideal’ number of occurrences per class if no taxonomic bias is observed. Red dots indicate negative deviations (i.e. shortfall in occurrences = under-represented classes); green dots indicate positive deviations (i.e. excess of occurrences = over-represented classes).



**Figure 23. Biodiversity occurrences recorded in GBIF between 1900 and 2006.** For each curve, the number of occurrences was plotted yearly. Top: black = all 24 classes considered together, yellow = Aves; Middle: yellow = Magnoliopsida, blue = Insecta, green = Liliopsida; Bottom: green = Actinopterygii, yellow = Mammalia, light blue = Reptilia, dark blue = Amphibia, orange = Florideophyceae, purple = Globothalamea.





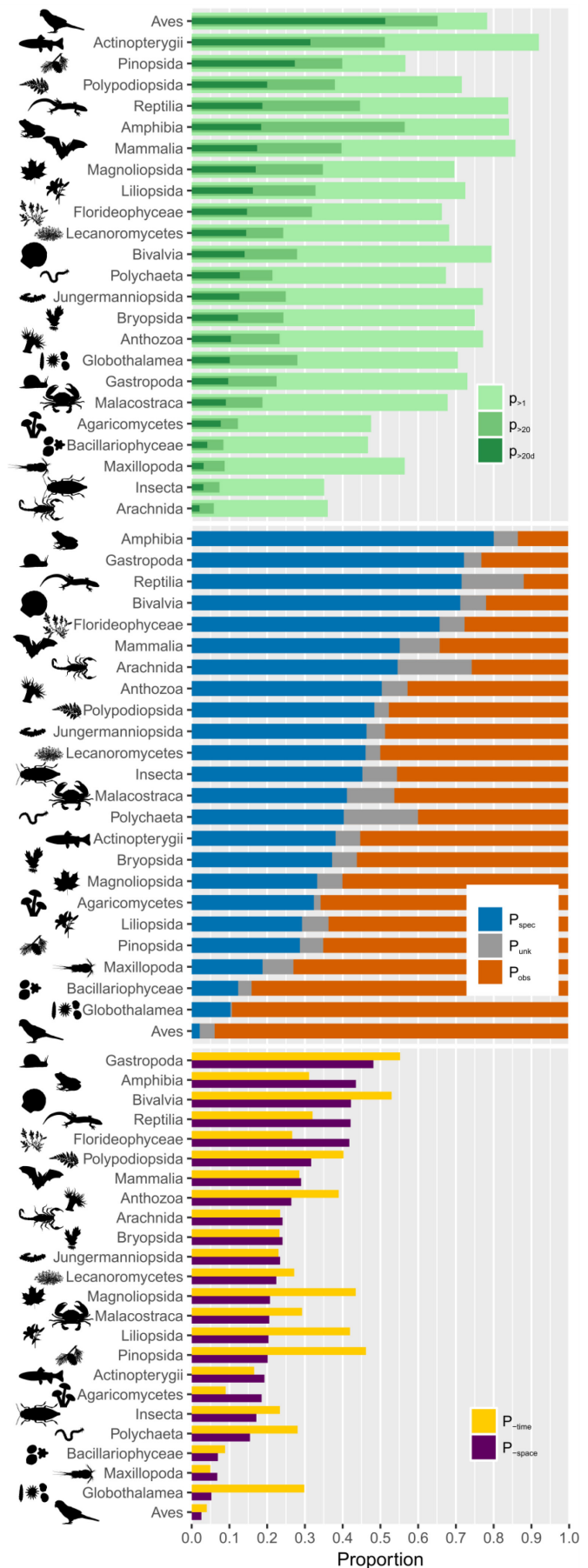
**Figure 24. Biodiversity occurrences recorded in the GBIF between 1900 and 2006.** For each curve, the number of occurrences is displayed year by year. Top: yellow = Lecanoromycetes; light blue = Arachnida; green = Agaricomycetes; dark blue = Gastropoda; orange = Bryopsida; Middle: yellow = Florideophyceae; light blue = Malacostraca; green = Maxillopoda; dark blue = Bacillariophyceae; orange = Polychaeta; Bottom: yellow = Jungermanniosida; light blue = Pinopsida; green = Anthozoa; purple = Bivalvia.

**Figure 25. Taxonomic heterogeneity in sampling, occurrence data origin and quality for 24 taxonomic classes.**

*Top:* Proportion of species per class recorded in GBIF with at least one occurrence (light green:  $p_{>1}$ ), with more than 20 occurrences (green:  $p_{>20}$ ), and with more than 20 spatially distinct occurrences (i.e. “decently” sampled – dark green:  $p_{>20d}$ ). For all classes, except Aves, less than 1/3 of all species are “decently” sampled. Classes are ranked according to their proportion of “decently” sampled species.

*Middle:* Occurrence origin (*basisOfRecord*) for each class. Some classes like Amphibia have a high proportion of occurrences based on specimens (blue: living or preserved specimen, material samples or fossils), whereas others like Aves have a majority of occurrences based on observation (orange: machine or human observation, literature). Grey bars show occurrences where the record basis is unknown. Classes are ranked according to their proportion of specimenbased occurrences.

*Bottom:* Data incompleteness. Proportion of occurrences with spatial (purple) or temporal (yellow) inaccuracies for each class. Spatial inaccuracy corresponds to an occurrence lacking coordinates or tagged has having geospatial issues by GBIF. Temporal inaccuracy corresponds to a sampling event with no specified month or year. Classes are ranked according to their proportion of occurrences with spatial issues.



**Explanatory variables.** In GBIF, recorded occurrences can refer to a collected specimen (or object) or an observation. The proportion of specimen- vs observation-based occurrences differed greatly between classes (Fig. 25 middle). Some classes had 90% or more of their occurrences based on observation (e.g. Globothalamea, Aves), whereas others had between 70 and 80% of occurrences based on specimens (e.g. Amphibia, Gastropoda, Reptilia and Bivalvia). Between these extremes, the relative proportion of specimen- vs observation-based occurrences in the 24 classes formed a continuum, with a few classes having an almost equivalent number of occurrences of both origins (e.g. Insecta). Three of the four groups of Tetrapods (Amphibia, Reptilia and Mammalia) had occurrences based mainly on specimens, whereas birds had the highest proportion of observation-based occurrences (94%).

Although temporal and geographical information can also be added to a record, these fields are informed with more or less precision. The proportion of spatial and temporal inaccuracies (p-time and p-space) differed greatly between classes (Fig. 25 bottom). Only 4% of Aves occurrences had temporal and/or spatial inaccuracies, whereas 48% and 55% of Gastropoda occurrences had spatial and temporal inaccuracies, respectively. Along with Gastropoda, the classes with the highest inaccuracy rates were Amphibia, Bivalvia and Reptilia, and these four classes were the ones with the highest proportion of specimen-based occurrences.

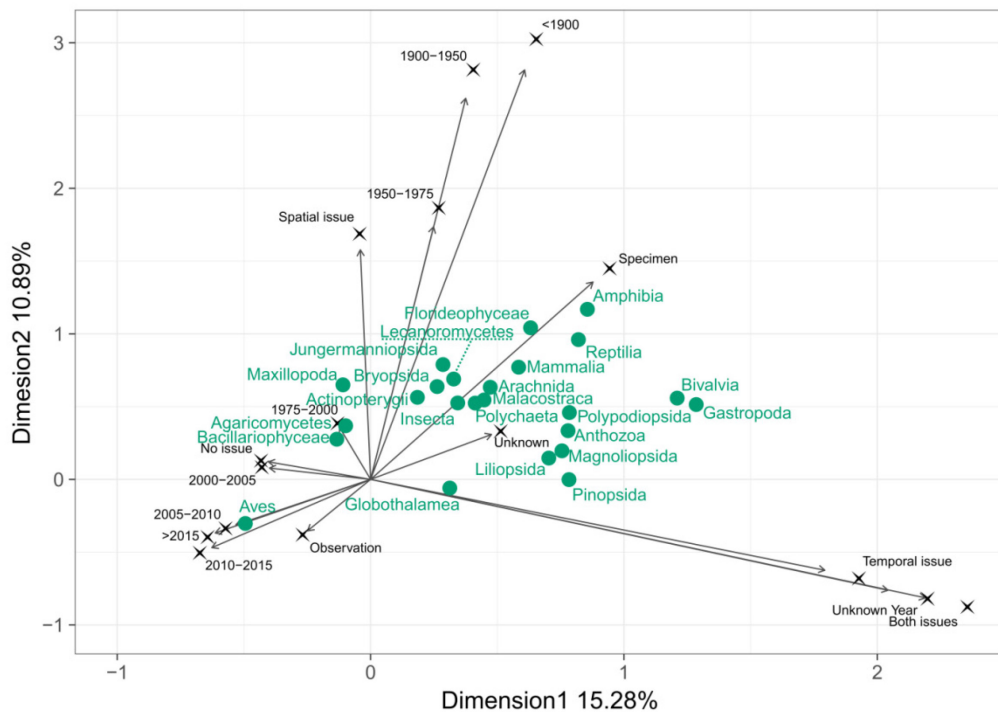
All Multiple Correspondence Analyses (MCA) showed that occurrences recorded before 1975 were grouped with specimen-based occurrences and with occurrences with spatial issues (Fig. 26). Conversely, more recent occurrences were grouped with complete and observation-based occurrences. Most of the classes, and in particular Amphibia, Reptilia and Florideophyceae, were in the upper right section of the graph (old, incomplete specimen-based occurrences), whereas Aves was in the lower left section, characterized by recent and complete observations.

Public interest (inferred from the number of web pages referenced by a search engine) and taxonomic research effort (inferred from the number of publications in Web of Science) were assessed and used in Generalized Linear Models (GLM). The number of web pages (with the keyword “species” added to the species’ scientific name) ranged from 0 to 1.8 million with a median number of 1,480 pages for the 24,000 best-represented species (1,000

species for each class) and 22 pages for the randomly chosen species. The number of publications, tallied for 453 orders, ranged from 0 (for eight orders) to 72 426 for Coleoptera, with a median number of 229 publications. For most classes, GLMs suggested a positive and significant correlation between public interest and the number of occurrences in GBIF (Table 4). A few negative correlations were found but were never significant. The quantity of research was not significantly correlated with the number of occurrences for most classes, and, when the correlation was significant, it was either positive (e.g. Mammalia) or negative (e.g. Agaricomycetes). A significant correlation between public interest and research quantity was found in 10 out of 47 cases.

**Table 3. Biodiversity occurrence data statistics for the orders (maximum 10) with the most occurrences within eight selected classes. Statistics and abbreviations as in Table 2.**

Order		nb <sub>occ</sub> (millions)	p>1 (thousands)	med/sp (mad)	Taxonomic precision	Order		nb <sub>occ</sub> (millions)	p>1 (thousands)	med/sp (mad)	Taxonomic precision
Agaricomycetes	Agaricales	1.80	14.14	4 (4)	0.93	Lecanoromycetes	Lecanorales	2.86	9.24	8 (10)	0.93
	Russulales	0.53	1.96	7 (9)	0.94		Teloschistales	0.75	2.38	11 (15)	0.94
	Polyporales	0.53	2.72	4 (9)	0.92		Peltigerales	0.58	1.45	15 (21)	0.92
	Hymenochaetales	0.27	0.74	7 (9)	0.91		Pertusariales	0.26	1.10	7 (9)	0.91
	Boletales	0.25	1.39	4 (9)	0.90		Ostropales	0.20	2.75	5 (6)	0.90
	Cantharellales	0.10	0.59	4 (4)	0.98		Umbilicariales	0.09	0.14	21 (30)	0.98
	Thelephorales	0.08	0.41	5 (6)	0.97		Baeomycetales	0.08	0.16	15 (21)	0.97
	Auriculariales	0.05	0.26	4 (4)	0.98		Candelariales	0.07	0.08	19 (27)	0.98
	Gomphales	0.04	0.30	7 (9)	0.92		Acarosporales	0.05	0.35	6 (7)	0.92
	Corticiales	0.02	0.25	4 (4)	0.95		Agryiales	0.01	0.06	12 (16)	0.95
Amphibia	Anura	2.85	5.03	54 (76)	0.82	Magnoliopsida	Asterales	17.02	33.12	16 (21)	0.82
	Caudata	1.06	0.60	172 (246)	0.93		Lamiales	13.03	28.01	17 (22)	0.93
	Gymnophiona	0.02	0.16	14 (19)	0.89		Fabales	11.27	24.13	25 (34)	0.89
Aves	Passeriformes	185.57	7.34	368 (525)	0.95		Rosales	9.45	13.76	12 (16)	0.95
	Charadriiformes	37.63	0.48	2538 (3710)	0.94		Malpighiales	7.06	19.68	22 (30)	0.94
	Anseriformes	34.12	0.21	5135 (7609)	0.93		Gentianales	5.06	21.70	17 (22)	0.93
	Accipitriformes	14.83	0.34	579 (855)	0.95		Ericales	4.74	14.07	20 (27)	0.95
	Piciformes	13.81	0.48	467.5 (650)	0.93		Myrtales	4.63	14.71	26 (34)	0.93
	Columbiformes	11.41	0.38	261 (366)	0.93		Apiales	4.55	5.93	20 (27)	0.93
	Pelecaniformes	11.01	0.16	1517 (2248)	0.90		Mammalia	Rodentia	3.62	3.59	25 (36)
	Gruiformes	4.98	0.28	148.5 (219)	0.94	Chiroptera		2.23	1.31	107 (154)	0.94
	Suliformes	4.57	0.09	777 (1150)	0.93	Carnivora		1.62	0.89	10 (13)	0.93
	Apodiformes	4.17	0.53	565 (802)	0.97	Diprotodontia		0.63	0.22	48.5 (70)	0.97
Insecta	Lepidoptera	17.41	64.11	3 (3)	0.76	Artiodactyla		0.63	0.98	8 (10)	0.76
	Coleoptera	9.77	96.27	3 (3)	0.93	Soricomorpha		0.48	0.67	16 (22)	0.93
	Hymenoptera	8.23	58.02	3 (3)	0.88	Cetacea		0.40	0.54	2 (1)	0.88
	Diptera	4.70	63.99	2 (1)	0.90	Lagomorpha		0.28	0.19	20.5 (29)	0.90
	Hemiptera	1.97	33.81	2 (1)	0.22	Perissodactyla		0.20	0.52	9 (12)	0.22
	Trichoptera	1.27	6.69	3 (3)	0.78	Primates		0.12	0.80	12 (16)	0.78
	Odonata	1.20	3.48	11 (15)	0.90	Reptilia	Squamata	4.49	9.16	37 (52)	0.90
	Orthoptera	0.96	9.57	3 (3)	0.78		Testudines	0.37	0.63	16 (22)	0.78
	Ephemeroptera	0.40	1.27	4 (4)	0.61		Crocodylia	0.05	0.16	3 (3)	0.61
	Plecoptera	0.23	1.92	4 (4)	0.88		Rhynchocephalia	0.00	0.02	2 (1)	0.88



**Figure 26. Relation between age, origin and quality of the occurrence data for 24 taxonomic classes.** Graph showing the first two axes of a Multiple Correspondence Analysis (MCA) performed on 5 million random occurrences. Labels in black represent the categories considered for all occurrences. Classes' names (in green) are placed at the average position of the class occurrences. Occurrence age contains eight time intervals and an Unknown Year category; data origin contains three categories: *Specimen* for specimen-based occurrences, *Observation* for observation-based occurrences, and *Unknown* for unknown origins; data quality contains four categories: *Temporal issue* for the lack of year or month, *Spatial issues* for the lack of coordinates, *Both issues* and *No issue*.

**Table 4. GLM results assessing the link between research quantity, public interest and their combined interaction on the amount of biodiversity data per class.** A positive correlation between public interest and the number of occurrences was found in most classes.

Green cells have a significant p-value at a 5 % threshold. Blue text indicates a positive influence while a text in red indicates a negative influence of the variable on the number of occurrences. Nb species = number of species used in the GLM after removing outliers; pval = p-values; NA = not available (because no order information and therefore no research quantity was available for Pinopsida).

Class	Selected species	Nb species	Public interest influence pval	Research influence pval	Interaction influence pval
Actinopterygii	Best	930	0.000	0.780	0.023
	Random	883	0.000	0.004	0.014
Agaricomycetes	Best	951	0.000	0.002	0.055
	Random	738	0.000	0.032	0.659
Amphibia	Best	916	0.573	0.000	0.058
	Random	875	0.076	0.024	0.714
Anthozoa	Best	910	0.304	0.273	0.000
	Random	744	0.002	0.101	0.198
Arachnida	Best	930	0.376	0.021	0.000
	Random	799	0.029	0.624	0.632
Aves	Best	930	0.376	0.021	0.000
	Random	850	0.000	0.182	0.277
Bacillariophyceae	Best	885	0.000	0.616	0.174
	Random	780	0.000	0.011	0.230
Bivalvia	Best	928	0.000	0.082	0.160
	Random	755	0.000	0.313	0.087
Bryopsida	Best	905	0.000	0.000	0.079
	Random	846	0.000	0.366	0.672
Florideophyceae	Best	904	0.000	0.070	0.000
	Random	818	0.000	0.002	0.665
Gastropoda	Best	718	0.683	0.183	0.045
	Random	521	0.033	0.110	0.738
Globothalamea	Best	886	0.005	0.000	0.599
	Random	793	0.015	0.310	0.106
Insecta	Best	967	0.000	0.246	0.216
	Random	769	0.013	0.369	0.601
Jungermanniopsida	Best	905	0.000	0.405	0.013
	Random	850	0.001	0.999	0.558
Lecanoromycetes	Best	961	0.000	0.667	0.851
	Random	804	0.000	0.584	0.560
Liliopsida	Best	931	0.000	0.060	0.168
	Random	856	0.000	0.000	0.615
Magnoliopsida	Best	959	0.000	0.003	0.205
	Random	768	0.001	0.170	0.863
Malacostraca	Best	906	0.000	0.002	0.001
	Random	757	0.156	0.392	0.154
Mammalia	Best	913	0.000	0.024	0.000
	Random	800	0.000	0.049	0.100
Maxillopoda	Best	889	0.000	0.017	0.540
	Random	835	0.012	0.898	0.510
Pinopsida		796	0.000	NA	NA
Polychaeta	Best	790	0.000	0.053	0.389
	Random	712	0.010	0.212	0.519
Polypodiopsida	Best	938	0.000	0.174	0.335
	Random	785	0.000	0.048	0.473
Reptilia	Best	940	0.180	0.627	0.190
	Random	794	0.040	0.104	0.448

## Discussion

Taxonomic bias, i.e. the fact that some taxa are more investigated than others, is a well-known problem for the study of biodiversity. How can we infer general principles and put in place effective strategies for biodiversity conservation when some taxa are over-studied while others are ignored? Although known for a long time, taxonomic bias is currently receiving an increasing attention. However most studies on taxonomic bias have been restricted to a few taxa or areas (Bonnet *et al.* 2002, Gaston and May. 1992, Troia and McManamay 2016, McKenzie and Robertson 2015, Donaldson *et al.* 2016, Pérez-Ponce de León and Poulin 2016). By analysing data from the biggest biodiversity data repository available, we emphasize here the prevalence of taxonomic bias in biodiversity data.

Unsurprisingly, and as previously reported regarding GBIF mediated data (Gaiji *et al.* 2013), we show that birds are over-represented in biodiversity data. Some studies highlighted the over-representation of birds in diverse disciplines ranging from behavioural ecology to evolution and conservation (Bonnet *et al.* 2002, Driscoll *et al.* 2014). The ever-growing number of observations that bird enthusiasts report undoubtedly amplify bias. Other vertebrate classes (Actinopterygii and Mammalia, and to a lesser extent Reptilia and Amphibia) are relatively well represented in the GBIF-mediated database, as are most Plantae classes, especially Liliopsida and Magnoliopsida. On the other hand, Arthropods (Insecta, Arachnida, Malacostraca and Maxillopoda) and Mollusca (Gastropoda and Bivalvia) are under-represented, with insects being particularly mis-represented. Birds and insects are obvious outliers but, beyond these two classes, the taxonomic bias in biodiversity data remains blatant.

Taxonomic bias is even more apparent when considering “decently” sampled species, namely species sampled in at least 20 different points on the globe. For any study requiring a number of different sampling points, like those relying on niche modelling, the field of investigation is restricted to vertebrates and plants on land and Actinopterygii in aquatic habitats. Invertebrates and fungi, on the other hand, have to be virtually ignored because of insufficient data at the scale of the planet. Given that these neglected organisms have a high diversity and play crucial roles in diverse ecosystems (Cardinale *et al.* 2012, Gascon *et al.* 2015, Lawler *et al.* 2003), this situation will inevitably result in an unbalanced fundamental



knowledge of biodiversity, risky guesses and uninformed conservation decisions (Feeley *et al.* 2016, Seddon *et al.* 2005, Gaston and May. 1992, Hortal *et al.* 2007, Yang *et al.* 2013). A similar taxonomic bias, with equivalent outcomes, is found between orders within each class.

More disturbingly, we show that the taxonomic bias in biodiversity data, although known for a few decades (May 1988), has remained broadly the same since the 1950's. The evolution of taxonomic bias over time has rarely been investigated, and never at a large taxonomic scale. Bonnet *et al.* (2002), focusing on vertebrates, showed there had been no changes in taxonomic chauvinism in ecology and behavioural research. Similarly, Stahlschmidt (2011) reported a static taxonomic bias from 2001 to 2010 in parental care research. He noted, however, that the absolute number of publications on parental care in birds increased significantly over this period. Along the same lines, Di Marco *et al.* (2017) emphasized that, in conservation science, some historically under-studied taxa were receiving more attention today, but underlined that a taxonomic bias toward taxa that are threatened or less rich in biodiversity still exists. Our results confirm this status quo situation at a larger taxonomic scale: most classes that were under- or over-represented in the GBIF mediated database in 1950 are still under- or over-represented today. Even though most classes are better recorded today than before, the gap between birds and the rest of biodiversity (i.e. ~99% of known biodiversity) increases with time because bird occurrences accumulate much faster than other class occurrences. Thus, while most of biodiversity remains to be described (Costello *et al.* 2013), the same taxa are preferentially studied and recorded over and over again.

The large taxonomic scale approach we used here comes with a few limitations. First, it must be emphasized that big datasets, like all sampling, are biased so that conclusions must be drawn accordingly (Boyd and Crawford 2012). Second, this large-scale approach implies that each species is equivalent and directly comparable, which is obviously arguable. Third, it neglects scale effects: species richness in insects is so large that whatever the means used, this class is always at risk of being understudied. Still, this approach enabled us to highlight the pervasiveness of taxonomic bias and bring new insights into the nature of this bias.

The underlying causes of taxonomic bias must be identified if one wants to reverse it. We suggest here that societal preferences, and not taxonomic research, orientate which

biodiversity data are gathered. The most popular species on the web are also the species with the most records in GBIF. Moreover, the best-supported model, where the interaction between taxonomic research effort and the number of web pages was taken into account, indicated a significant effect of public interest on biodiversity data gathering. The role played by the general public in the study and conservation of biodiversity has already been established: positive links exist between public opinion, scientific productions and conservation policies, however the directionality of these interactions remains unclear (Martín-López *et al.* 2009, Ressurreição *et al.* 2012). Our analyses confirm these interactions but do not allow us to clarify the causality issue. Although inevitable biases occur when using internet searches, such as the inability to distinguish scientific web pages from other web pages, particularly at such a broad taxonomic scale, “many (30–80%) web pages containing the scientific names of species have little or nothing to do with scientific research“ (Wilson *et al.* 2007) indicating that our results are presumably related to societal preferences. Surveys to determine public preferences could help counteract this issue but should be carried out at large taxonomic scales.

Studying invasive alien species, Wilson *et al.* (2007) concluded that “the choice of research subject in biology reflects the interests of society”. Because of public interest, and not specifically for their scientific interest, studies of ‘public-aware’ taxa are more likely to be funded and receive more funding (Leather 2009, Martín-López *et al.* 2009, Stein *et al.* 2002). Our results provide further evidence of this trend, highlighting the active role of the general public in biodiversity data collecting, given that, for instance, the biggest dataset was provided by eBird (211 million occurrences), a collective enterprise devoted to birds and partly relying on citizen science<sup>45</sup>. For multiple reasons (e.g. the difficulty of obtaining permits, more and more endangered species, citizen science programmes, population decline, etc.), less specimen-based occurrences are now reported. Amphibia, Gastropoda and Reptilia, the three classes with the highest proportion of specimen-based occurrences, are also the classes with a decreasing or stabilizing trend in data accumulation. We thus anticipate an increasing bias between taxa mostly known from observation-based occurrences and taxa mostly known from specimen-based occurrences. In addition, a lot of records are old and incomplete, and could soon, or already, be obsolete (Escribano *et al.* 2016), which risks reinforcing the taxonomic bias against classes with relatively few recent occurrences.

The good news is that the observed taxonomic bias can be corrected. Shine and Bonnet (2000) showed how snakes, which were under-represented in ecology among terrestrial vertebrates until 1990, have grown in popularity in this scientific field, illustrating that acting on taxonomic bias is possible. Similarly, for most classes, occurrences accumulate at a much faster rate now than 50 or 30 years ago, which is an encouraging trend. Obviously, this trend can also result from changes in data-sharing practices, and not simply from overall data collection. Still, as we are accumulating more and more biodiversity data, the question of how to efficiently sample the whole of biodiversity remains open. The biodiversity knowledge chain is complex and its links influence one another. Scientists play a key role in this chain. However, our results show that they alone cannot ensure that biodiversity is sampled adequately and that societal preferences are too important to be ignored. Scientists must reach out to the lay audience (Wilson *et al.* 2007, Martín-López *et al.* 2009, Turpie 2003) and advertise under-represented organisms to the general public. For instance, the crucial role of protists in ecosystem functioning probably seems too obscure to generate any interest from the general public (Cotterill *et al.* 2007). New practices or methods, from citizen science to metagenomics, should also help increase public awareness and would have even more impact if programmes were developed jointly between science and society (Pawar 2003, Hochachka *et al.* 2012). The expected gain would be colossal and would achieve more than a well-balanced sampling of biodiversity: new vocations in science, more efficient citizen sciences programmes, influence on funding and political decisions, etc.

Citizen science and data gathering by non-professionals might be decisive in the near future. The contribution of citizen science to the most over-represented class of GBIF-mediated data, birds, dates back more than a hundred years (Miller *et al.* 2012). Different fields of research from molecular engineering (Eiben *et al.* 2012) to quantum science (Lieberoth *et al.* 2014) and neurosciences (Marx 2013) have greatly benefited from the involvement of non-professionals, and it has been shown that a well-made citizen science programme can produce in two years the same amount of data that scientists can produce in a decade (Zapponi *et al.* 2016). Yet, the use of citizen science for studying taxa that are not as charismatic as birds or mammals is still in its infancy (Zapponi *et al.* 2016, Gardiner *et al.* 2012). Efforts must be made to develop such initiatives, probably by relying on new technologies such as smartphones and dedicated applications (Zapponi *et al.* 2016,

Newman *et al.* 2012). Citizen science cannot, and must not, replace standard scientific practices (Kamp *et al.* 2016); they are complementary approaches with different strengths and limitations. However, citizen science could substantially contribute to our knowledge of biodiversity, especially if adapted programmes devoted to neglected taxa are highlighted (Chandler *et al.* 2017).

Considering the whole of biodiversity, and not only charismatic organisms, is a prerequisite for the development of efficient conservation plans, of prolific bioprospecting activities, and for enhancing our understanding of biodiversity on a global scale (Di *et al.* 2017, Wilson 2000, Cardoso *et al.* 2011). Many international projects have been developed since the Convention on Biological Diversity, illustrating an increased awareness of the astonishing diversity of functions and services that biodiversity supports (Cardinale *et al.* 2012, Gascon *et al.* 2015). Nevertheless, while biodiversity declines at an unprecedented rate (Barnosky *et al.* 2011), taxonomic bias is still a burden on biodiversity studies. It is urgent that we get rid of this burden and that we start embracing the whole of biodiversity.

## Methods

**Dataset.** We downloaded all available occurrence records from the GBIF data portal in June 2016 (<http://doi.org/10.15468/dl.hqesx6>). 649 million occurrences were saved as a Darwin Core archive. Occurrences from this archive were extracted and imported into a SQL database, where data were indexed to reduce the computation time of subsequent queries. We focused on 24 taxonomic classes out of the 297 referenced in GBIF, excluding classes with less than 1 million occurrences (9.4 million occurrences from 19,000 species, had no class affiliation). We ended up with 626 million occurrences (NBocc) and 1.01 million species, representing more than 96% of the total number of occurrences and 84% of the total number of species in GBIF. All statistics were computed from this dataset.

**Taxonomic errors: imprecision and bias.** For each class, we quantified the level of taxonomic precision as the proportion of occurrences with information at the species level or lower. We assessed taxonomic bias by computing and comparing the following statistics for each class: the total number of occurrences (nbocc), the median number of occurrences per species ( $med_{sp}$ ) and the median absolute deviation, the proportion of species with at least one

occurrence ( $p_{>1} = n_{>1}/N$ ), and the proportion of species with at least 20 occurrences ( $p_{>20} = n_{>20}/N$ ), where  $n_{>i}$  is the number of species with at least  $i$  occurrences and  $N$  is the number of known species for a given class.  $N$  was obtained using the GBIF taxonomic backbone (accessible at: <http://doi.org/10.15468/39omei>), by counting the number of distinct species with either the ‘accepted’ or ‘doubtful’ taxonomic status. This method excluded synonyms. Furthermore, we computed  $p_{>20d}$ , the proportion of species with at least 20 spatially distinct occurrences. Two occurrences were considered spatially distinct when, using a global grid of 10\*10 km cells based on the pseudocylindrical equal-area map projection Eckert IV, they fell in two different cells. We chose a threshold of 20 spatially distinct occurrences because it is a common threshold in niche modelling analyses (Feeley and Silman 2010). Occurrences without spatial coordinates were excluded when computing the number of spatially distinct occurrences. We calculated how each class deviates from an ‘ideal’ sampling  $I$ , where each class is sampled proportionally to its number of known species ( $N$ ).  $I = NB_{occ}*(N/N_{tot})$  where  $N_{tot}$  is the total number of known species. To investigate the evolution of taxonomic bias over time, we excluded i) occurrences without a collection year and ii) occurrences recorded during the last 10 years because of the lag between recording and integration in the GBIF database (S. Gaiji, pers. comm.). The ‘ideal’ sampling  $I$  was calculated every ten years between 1956–2006 and deviations from these ‘ideal’ samplings were plotted for each class.

Statistics were computed at the ordinal level for Agaricomycetes, Amphibia, Aves, Insecta, Lecanoromycetes, Magnoliopsida, Mammalia and Reptilia using the same methods. These classes were chosen due to their relatively high number of occurrences and/or species, and because of the diversity of patterns they exhibited in our preliminary results. We also tried to cover a large taxonomic range (Tetrapods, Arthropoda, Plantae, Fungi) to include as much biodiversity as possible.

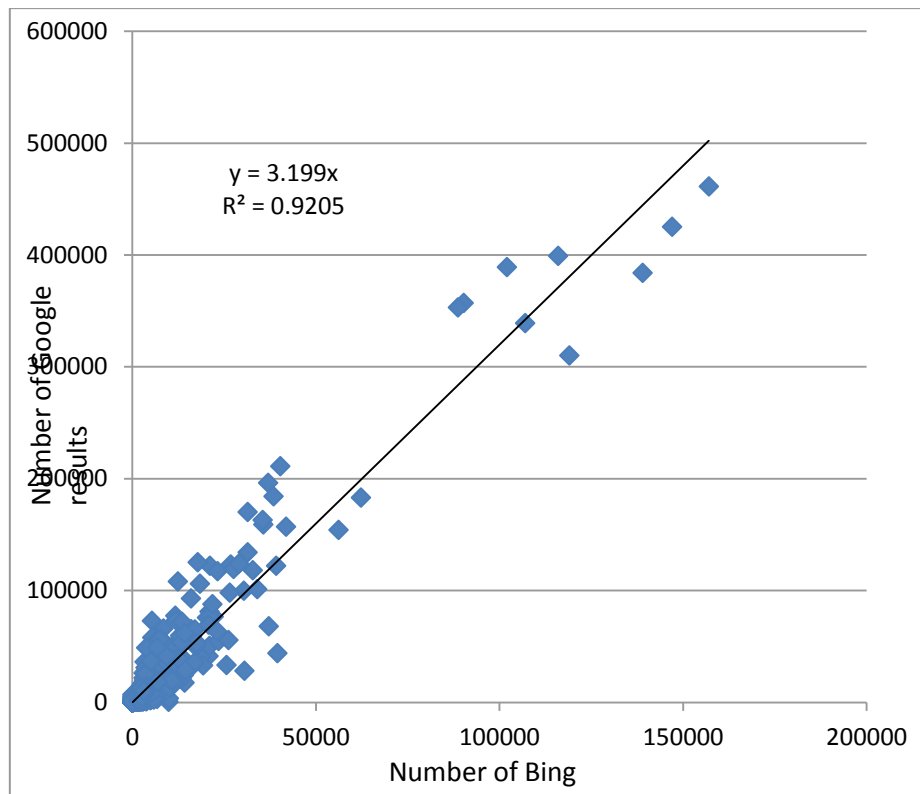
**Explanatory variables computed from the GBIF dataset.** Data origin. In GBIF, the origin of an occurrence can be specified using a controlled vocabulary in the ‘basisOfRecord’ field. We delimited three categories, depending on whether recorded occurrences refer to a specimen (or object), an observation, or was of unknown origin. The “specimen” category ( $O_{spec}$ ) contained: fossil specimens, living specimens, material samples and preserved specimens. The “observation” category ( $O_{obs}$ ) consisted of: human observations, machine

observations, unclassified observation and literature. The third category corresponded to the “unknown” option ( $O_{\text{unk}}$ ).

**Date and Locality precision (Data completeness).** For each class, the proportion of temporal ( $p_{\text{time}}$ ) and spatial inaccuracies ( $p_{\text{space}}$ ) was computed as follows:  $p_{\text{time}} = O_{\text{time}}/nb_{\text{occ}}$  and  $p_{\text{space}} = O_{\text{space}}/nb_{\text{occ}}$ , where  $O_{\text{time}}$  is the number of occurrences lacking information regarding either the month, year or both, and  $O_{\text{space}}$  is the number of occurrences missing coordinates or flagged as having geospatial issues in GBIF.

**External explanatory variables. Taxonomic research and societal preferences.** Taxonomic research was quantified through the number of publications. We searched the Web of Science portal (apps.webofknowledge.com) with the following query for each order: “taxonomy” AND (“[order name]” OR “[family names]”), over the 1900–2016 period. The number of systematists, who are the producers of primary biodiversity data, would have been a better indicator but this could not be obtained due to the current architecture of Web of Science. We therefore used the publication metrics for taxonomic research from Web of Science as done previously (McKenzie and Robertson 2015).

Public interest for a given species was estimated through the number of web page results, a proxy that has been proven to be reliable (Wilson *et al.* 2007). These numbers were obtained from Bing searches using the exact Latin name (e.g. “Corvus corax”) or a combination of the Latin name and the keyword “species” (e.g. “Corvus corax” + species). Bing and Google searches yielded similar results for the 4,000 species tested with both search engines (Fig. 27), but only Bing allowed us to carry out a high number of searches. For each class, these searches were performed on the 1,000 species with the most occurrences (except for Pinopsida, which only had 902 species recorded in the GBIF) and then on a further 1,000 randomly chosen species. Each search was run twice to check for consistency.



**Figure 27. Relation between the number of Google search results and Bing search results for 4000 random species.** We compared the number of web search results for two popular search engines. Using 4000 species from four classes (Aves, Magnoliopsida, Insecta and Liliopsida – 1000 species each), we found that the two search engines gave comparable results, with Bing returning fewer results than Google in general. Using Google for more requests was impossible because the script used was detected as potential spam.

**Statistical analyses.** We favored medians (m) over means because of their robustness to outliers. For the same reason, we used the median absolute deviation (mad), which represents the median of the absolute deviation from the median, as a measure of statistical dispersion. In all analyses needing spatial or temporal information,  $O_{\text{-space}}$  and  $O_{\text{-time}}$  occurrences were ignored, respectively.

The relationship between data origin, completeness and year of record was investigated using multiple correspondence analyses (MCA). Analyses were done on three samples of five million random occurrences from our dataset. The variables were: class (24 categories), year of the record (categories: '<1900', '1900–1949', '1950–1974', '1975–1999', '2000–2004', '2005–2009', '2010–2014', '> = 2015'), data origin (categories: specimen,

observation, unknown), data completeness (categories: no problem, missing temporal information, missing spatial information, missing both). Because results can be hard to interpret when categories with very few observations are used (Cardoso *et al.* 2011), each analysis was performed a second time ventilating the categories represented in less than 0.5% of the dataset.

To explore the relative impact of public interest and taxonomic research quantity on taxonomic bias, we used generalized linear models (McCullagh and Nelder 1989, Zuur *et al.* 2009) (GLM). For each of the 24 classes, we looked at the effect of these two variables and their interaction on the number of occurrences per species in GBIF. We used an identical model for all classes, which was fitted using a negative binomial distribution to take into account overdispersion. Half of the GLMs were computed using the 1,000 best-represented species in GBIF (Best), while the other half used 1,000 random species referenced in GBIF (Random). Only one GLM was computed for Pinopsida because they had less than 1,000 species. Initial models were strongly influenced by extreme values and had poor resolution. Therefore we excluded outliers, which were identified when the number of occurrences or web search results was  $>Q_3 + 4 * IQR$ , where  $Q_3$  is the third quartile value and IQR is the interquartile range. For each GLM, we checked the validity of the model by plotting the values of residuals against predicted values to test the homogeneity of residuals.

We performed all analyses using the R statistical software version 3.3.2 (<https://www.R-project.org>) with associated packages: FactoMineR (Husson *et al.* 2016), ggplot2 (Wickham 2016), gridExtra (Auguie 2017), MASS (Venables and Ripley 2013) and plyr (Wickham 2011).

## **Acknowledgements**

This study was developed as part of a Ph.D. project and was funded as a grant by the Ministère de la Recherche to JT. It was in part presented at the IBS Special Meeting in Beijing, China. We thank the organizers for giving us the opportunity to present these results and the attendees for their useful questions and suggestions. We thank Anne-Sophie Archambeau, Michel Baylac, Samy Gaiji, Marie-Elise Lecoq, Sophie Pamerlon, Roseli Pellens, Tim Robertson, Dmitri Schigel, Jérôme Sueur and Wilfried Thuiller for fruitful



discussions. We thank H el ene Citerne for revising the English. We thank two anonymous reviewers who helped improving the manuscript.

### **Author Contributions**

J.T., F.L., R.V.L. and P.G. designed this study, J.T. performed the analyses, A.B. contributed to the statistical analyses, J.T. and F.L. wrote the first draft, and all authors discussed the results and provided input on the manuscript.

# **Chapter 4: Latitudinal Diversity Gradient: Geometric hypotheses revisited using massive biodiversity occurrences in plants and animals of the New World (article in preparation)**

The two previous chapters present an exploration of the data themselves. It is therefore now possible for me to make optimal use of the GBIF data while being aware of their strengths and weaknesses. The final objective of the study of these data is to enable me to work on the major patterns of biodiversity and in particular the latitudinal diversity gradient (LDG). This gradient is still one of the favorite puzzles of biogeographers and macroecologists (Hawkins 2001).

The exploration of this gradient is continuing, and more than 30 hypotheses have been formulated to explain the existence of this gradient (Willig *et al.* 2003). Among these, the geometric hypotheses have generated quite a lot of debate. Formulated by Colwell and Hurtt (1994) for the first time, the null-hypothesis was later discarded by Currie and Kerr (2008) as an explanation of the LDG. However Gross and Snyder-Beattie (2016) recently proposed an improved version of these hypotheses which was not tested against the gradient. In the following chapter I tested this hypothesis by studying LDG in 8 taxonomic classes as well as five additional hypotheses about its formation.

## **Introduction**

Species are not randomly distributed on earth (Pianka 1966, Hawkins 2001), which results in biodiversity patterns, the most pervasive one being the Latitudinal Diversity Gradient (LDG; Willig *et al.* 2003). The LDG refers to the higher species richness found at lower than higher latitudes. Although reported a long time ago (Hawkins 2001), this biodiversity pattern still fascinates and challenges ecologists and evolutionary biologists (Gaston 2000; Willig *et al.* 2003). Multiple hypotheses have been suggested to explain this pattern, some relying on environmental factors while others underline evolutionary factors or geometric constraints (Willig *et al.* 2003; Jablonski *et al.* 2006; Mittelbach *et al.* 2007). None

of them, however, proved satisfactory and, today, multifactor hypotheses hold the best position to apprehend the LDG.

Among the different hypotheses behind the LDG, the geometric constraints hypothesis reflects an original approach to this pattern. It was first proposed at the end of the 20<sup>th</sup> century (Colwell and Hurtt 1994) and suggests that biological parameters (environmental or evolutionary) are not needed to generate a latitudinal gradient. A “nonbiological” gradient could be caused by the random repartition of species ranges in a bounded domain like earth. This sometimes called null hypothesis has been largely debated (Colwell and Lees 2000, Lees and Colwell 2007, Currie and Kerr 2007) and is still investigated (Meza-Joya and Torres 2016) even though, for some authors, opposing arguments almost wiped it out (Currie and Kerr 2008, Fine 2015).

The geometric constraints hypothesis has yet known a recent revival. An improved version of this null hypothesis has been proposed and its authors suggested that its role in latitudinal formation should be revisited (Gross and Snyder-Beattie 2016). Gross and Snyder-Beattie underlined that our ability to evaluate the contribution of geometric hypotheses to biodiversity gradients was limited so far and they proposed an extended mathematical framework to improve this situation. They demonstrated how their model fits with empirical biodiversity gradients but did not formally test it with biodiversity data. In this study, we test for the first time this modern version of the geometric constraints hypothesis, which seems to propose a response to criticisms of the earliest version of the geometric hypothesis but has never been tested on empirical data.

To implement this test we used eight taxonomic classes of plants and animals from the New World. This biodiversity sample was selected from GBIF-mediated data according to two criteria designed to circumvent classical limitations in the study of LDG. First, we covered a large taxonomic scale to increase the generality of our analyses (Amphibia, Aves, Liliopsida, Magnoliopsida, Mammalia, Polypodiopsida, Pinopsida, Reptilia). These classes were the one with the best sampling and data quality as reported in Troudet *et al.* (2017). Second, we covered a large latitudinal range as studying biodiversity gradients at small geographic scale may alter the species richness signal (Rahbek 2005) and results in diminished or inversed gradients (Willig 2003). We therefore investigated biodiversity pattern

in the New World, an area covering a very large latitudinal range and more evenly sampled than others in the GBIF-mediated data (Meyer *et al.* 2015). Because no single mechanism is, in all probability, responsible for biodiversity gradients, we compared the revised version of the geometric constraints hypothesis (Gross and Snyder-Beattie 2016) with the original one (Colwell and Hurtt 1994) and with in situ hypotheses (Jablonski *et al.* 2017) that mostly focus on the capacity of the local environment to support a certain amount of species (carrying capacity).

The statistical approach we used integrates the spatial dimension of biodiversity gradient, a dimension sometimes neglected but paramount. In Geography, Tobler's first law specifies “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). This means that species richness data, as well as environmental data, are spatially autocorrelated and that it must be accounted for (Legendre 1993, Dormann 2007, Boucher-Lalonde and Currie 2016). In addition, the spatial stationarity of the statistical model must be considered. A classic regression model assumes constancy across space (here between an explanatory variable and species richness), also called model spatial stationarity (Dormann *et al.* 2007). As the geographical scope of a study increases, the spatial stationarity assumption becomes largely dubious, a point to be considered through Geographical Weighted Regressions (Foody 2004).

## **Material and Methods**

### **Species richness estimates**

Species richness was computed using the data from the Global Biodiversity Information Facility (GBIF) portal at “gbif.org” (<http://doi.org/10.15468/dl.9goauq>). We excluded occurrences flagged in the GBIF as having geospatial issues as well as occurrences not georeferenced, fossil and living specimens (from zoo, farms, gardens...). We also discarded occurrences having one of four key words ("zoo", "aquarium", "farm", "captive") inside the “Locality” or “OccurrenceRemarks” columns. We targeted eight classes given their amount of data available (Troudet *et al.* 2017; Amphibia, Aves, Liliopsida, Magnoliopsida, Mammalia, Pinopsida, Polypodiopsida and Reptilia) and restricted the study to the New World, a relatively well-sampled region (Meyer *et al.*, 2015). All the subsequent geographical

computations were done using the equal-area and pseudo cylindrical map projection Eckert IV.

Using maps from [www.naturalearthdata.com](http://www.naturalearthdata.com) we removed species with  $\geq 10\%$  occurrences in the oceans flagging them as marine species. Occurrences for the remaining species were allocated to cells on a worldwide 10\*10km grid, which included climatic values from Bioclim (Fick and Hijmans 2017) that were averaged inside each cell.

We searched for geographic and climatic outliers for each species. First, for each pair of cells occupied by a species, the orthodromic distance, i.e. the shortest distance between two points on a sphere (the earth), was computed. We flagged as outliers the most distant cells, those for which the distance mean to the five nearest cells was in the last centile. This process is bound to flag “good” cells as outliers, but it is a fast and conservative solution. We then used climatic variables and the R package *mvoutlier* (Filzmoser 2005) to compute the Mahalanobis distance (Mahalanobis 1936) of each cell and identify climatic outliers.

Even though the New World is better sampled than other continents in the GBIF-mediated data, this sampling is uneven and incomplete. To compensate for these limitations, we used niche modeling algorithms, as proposed in García-Roselló *et al.* (2015), using only species recorded in  $\geq 20$  cells (Feeley and Silman, 2010). For each species, we computed a convex hull and used the Surface Range Envelop (SRE) model from the package *BIOMOD2* (Thuiller *et al.* 2009) on the cells having their center inside the convex hull. This model marks a cell as compatible with a species when the environmental values of the cell are consistent with the values in the cells actually occupied by the species. The SRE model was chosen because of both its simplicity and low requirement in computational power. The 19 climatic values and the altitude available in Bioclim (Fick and Hijmans 2017) were included in the model. We then configured it to remove the 5% more extreme values of each variable to limit the influence of extreme environmental values. This process resulted in a potential niche for each species, a list of compatible cells per species. Species richness of a geographic cell was computed from these lists by counting the number of putative species per cell (for each class or in total). Although imperfect, this method allowed us having a better spatial coverage for each species. Moreover, because we restricted niche modeling to the convex hull of the species (after filtering outliers), the latitudinal range of a species was not overestimated.

Finally, because the SRE model is more restrictive as more environmental variables are included, the use of 20 variables greatly restricted the number of potential cells per species, limiting the risk of artificially spreading species distribution.

### **Explanatory Variables**

We used eight variables (null-CH, null-GSB, PET, BIO1, AET, BIO12, RAPO-Stevens and RAPO-mid) covering six of the hypotheses formulated to explain the LDG. Those hypotheses correspond to the “environmental” explanations of the LDG.

### ***Null hypotheses***

**null-CH:** Colwell and Hurtt (1994) suggested that a latitudinal gradient could arise from the random placement of species ranges across the globe, without any influence of environmental variables. This phenomenon, later called mid-domain effect, was used as a first null hypothesis. For each class, all the actual species ranges were randomly placed within the latitudinal boundaries of the class (Colwell and Lees 2000). The randomization was done using the default pseudorandom number generator in R. We then counted the number of species ranges inside latitudinal intervals of  $0.5^\circ$  to get species richness in each of these intervals. For each class, we repeated this randomization 1000 times and averaged the results to robustly estimate the mid-domain effect.

**null-GSB:** As a second null hypothesis, we relied on the model of Gross and Snyder-Beattie (2016), which uses additional concepts such as species environmental niches and range size limits. The addition of these parameters adds more complexity to the patterns produced by the model. Those patterns are more nuanced than those produced by the null-CH hypothesis and exhibit tropical plateaus or mid latitude inflexion points in species richness, which is more in agreement with the commonly observed gradients. The model required the environmental tolerance (ET), the distance limitation (DL) and the ice cap extent (ICE) as inputs. For each class, i) the DL was calculated as the average species latitudinal range, ii) the ICE was set, separately for both hemispheres, to the maximum absolute latitude value (in radians) occupied by the class (latitudinal boundaries), and iii) the ET was selected among eighteen ET values (from 0.01 to 0.35 with 0.02 increments) by taking the one better correlated with the species richness of the class (higher  $R^2$ ) using ordinary least square (OLS)

regression. The “grain” of the model was fixed at 180 insuring a precision of at least 0.5°. The whole process resulted in 16 selected models (8 classes with two ICE values, one for north and one for south boundaries).

### ***Ambient Energy***

**PET & BIO1:** High latitude environments are supposed to have lower inputs of solar energy and lower climatic stability, which would result in smaller species richness according to the Ambient Energy hypothesis. Ambient energy was estimated using two variables: potential evapotranspiration (PET), calculated using the data from Mu *et al.* (2011); annual Mean Temperature from WorldClim (BIO1). Other temperature variables were not used to prevent colinearity between explicative variables.

### ***Productivity***

**AET:** The link between productivity and species richness can be attributed to Hutchinson (1959) and has been regularly mentioned afterwards (Hawkins *et al.* 2003a, Currie *et al.* 2004, Gillman *et al.* 2015). This hypothesis suggests that a higher productivity in a given area generates more individuals, which would result in a higher species richness. We used Actual Evapotranspiration (AET), from Mu *et al.* (2011), to estimate productivity.

### ***Water availability***

**BIO12:** Water availability was suggested as an important factor of species distribution, in particular in lower latitudes (Hawkins *et al.* 2003a, Hawkins *et al.* 2003b). We used Annual Precipitation from Bioclim (BIO12) as a mean to study the impact of water availability on species richness.

### ***Rapoport's effect***

**RAPO-Stevens & RAPO-mid:** The Rapoport's rule (Stevens 1989) suggests that species latitudinal ranges sizes tend to increase with latitude, partly explaining the LDG. Smaller range sizes at low latitudes would allow more species to share an area, resulting in higher species richness compared to higher latitudes. Although contested (e.g. Rohde *et al.* 1993, Gaston *et al.* 1998, Gaston and Chown 1999), we tested this effect once again. The

Rapoport effect was calculated according to both Stevens' Rule (Stevens 1989) and the Midpoint method (Rhode *et al.* 1993) using 0.5° intervals and raw GBIF-mediated data. Both results were then included in the models (RAPO-Stevens and RAPO-mid, respectively).

### **Statistical analyses**

For each class, relationships between species richness and the explanatory variables were modeled using ordinary least squares analyses (OLS) on 10,000 random cells. This random selection provided nearly identical results to those based on the whole dataset and was necessary because spatial regression models required too much computing power to be performed on all cells. The dependent variable is the number of species inside each 100km<sup>2</sup> cell while the covariates include all the explanatory variables listed above. Following a manual iterative stepwise method, we first selected the best null hypothesis for each class and then added other explanatory variables. At each step, the benefit obtained from the added variable was evaluated using the coefficient of determination ( $r^2$ ) and we stopped the process when the added variable did not improve the  $r^2$  by at least 1 %.

OLS analyses did not take into account the spatial dimension of the data, which is a limitation. We thus used the residuals from the best OLS models to test for spatial autocorrelation. We used Moran's I test to assess whether the variables were spatially dependent. When spatial dependency was detected, we used spatial lag and spatial error regressions, two models integrating a spatial dimension. Residuals of these spatial regressions were tested with Moran's I to ensure spatial autocorrelation has been properly dealt with.

Finally, we computed geographically weighted regressions (GWR). GWR is a local regression method that can be used for diagnosing spatial heterogeneity between dependent and explanatory variables over space (Brunsdon *et al.* 1996). GWR is performed within local windows centered on each observation of the dataset. In a local window, each observation is weighted according to its proximity to the center of that window and a regression model is then used within the window, hence on a subset of observations. The main advantage of GWR is that it allows for spatial variation within a given area. For instance, local coefficient of determination ( $R^2$ ) can be assessed for most GWR, enabling us to identify where the tested



variables have greater or lesser explanatory powers (Charlton and Fotheringham 2009; Lloyd, 2010).

We used the R package *spgwr* to compute a GWR on 40,000 random cells (from a total of 435,086 cells on the New World) for the classes showing a latitudinal diversity gradient. The GWR could not be computed with the whole dataset due to computing limitations and 40,000 cells were estimated as a good trade-off in terms of data quantity and computation time. For each class, the model used in the GWR was the model selected using OLS. GWR required two other parameters: the weighting formula and the size of the local window. The weighting formula choice has a low impact (Charlton and Fotheringham 2009) and the classic bisquare formula was chosen. The window size, called bandwidth in the *spgwr* package, is more important and depends on the clusterization of the data. Selected randomly, the points were most often regularly spaced but not always (due to America's geography), which could impact heavily on the analyses. After multiple tests on each class (i.e. adaptive bandwidth of 5, 10 and 15% and fixed bandwidth of 2000 and 5000 kilometers), we empirically selected an adaptive bandwidth of 15 % to counteract this clustering, and we chose a Poisson implementation of the GWR as we dealt with species counts.

## **Results**

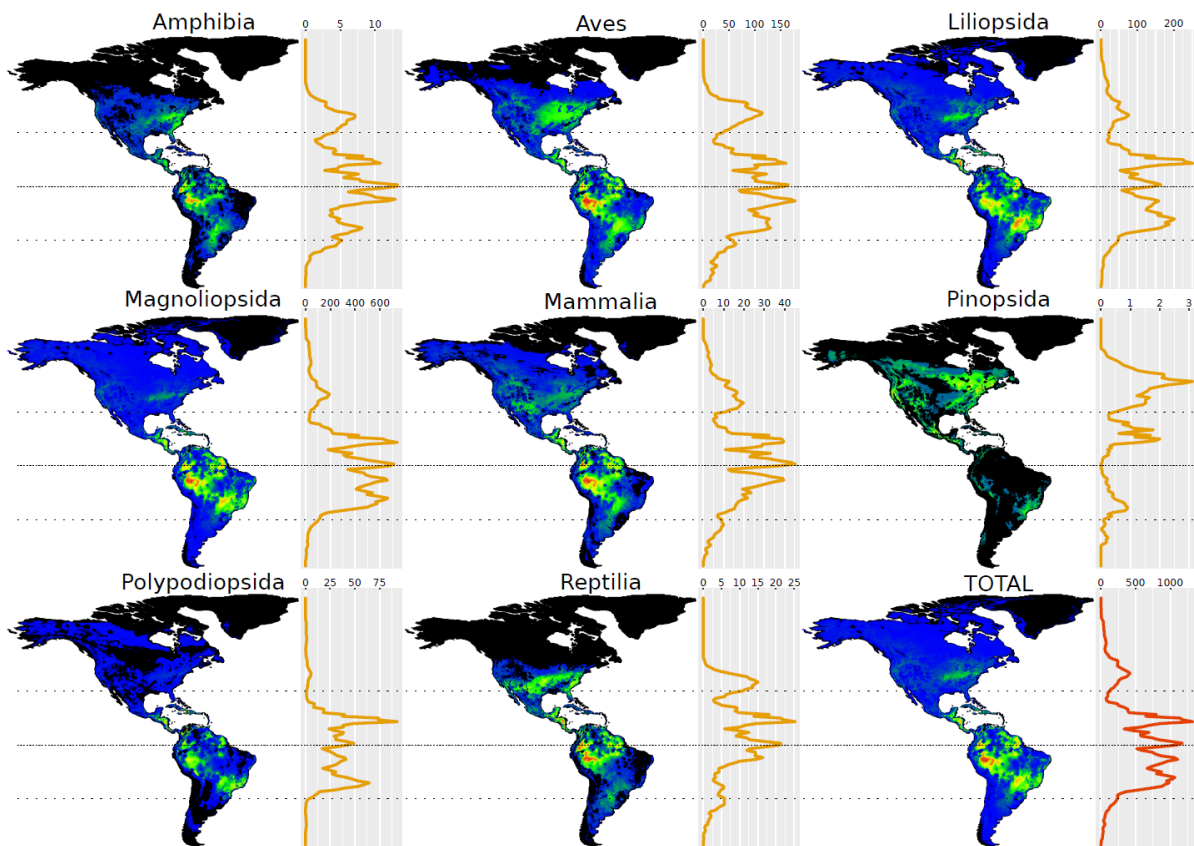
### **Basic statistics**

The dataset downloaded from the GBIF portal contained 547 million occurrences, 462 million of which belonging to 323,044 species of Amphibia, Aves, Liliopsida, Magnoliopsida, Mammalia, Polypodiopsida or Reptilia. 421 million of those occurrences were terrestrial, and 208 million occurrences (for 149,243 species) were in the New World. 62,099 species had  $\geq 20$  spatially distinct occurrences (cells of  $10 \times 10$  km). In total, the dataset contained 11 million unique species-cell pairs in the New World for the eight selected classes, but, after computing SRE models, the dataset contained 208 million potential unique species-cell pairs.

### **Latitudinal diversity gradient**

For each class, specific richness was mapped according to the species distribution obtained through SRE models (Fig. 28). A LDG is visible for all classes, except Pinopsida, as

well as for the map summing the species richness for the eight classes combined. The plots of average species richness per latitude ( $1^\circ$  precision) confirm these patterns.



**Figure 28:** For 7 out of 8 tested classes the Latitudinal Diversity Gradient is clearly visible. Pinopsida was the only taxa that did not comply with the classic pattern. The TOTAL plot uses the sums of species richness from the 8 taxa. On the maps hotter colors correspond to higher species richness. Each side plot shows the mean species richness per 100km<sup>2</sup> per latitude. Every plot has a different species richness scale. The dotted lines correspond to the equator and the  $+30^\circ$  and  $-30^\circ$  latitudes.

### Environmental hypotheses

OLS models were performed on each class but no variable explain the species richness distribution of Pinopsida, the only targeted class with no latitudinal diversity gradient. For the seven other classes, the best null model was the one proposed by Gross and Snyder-Beattie (null-GSB; Table 5). The forward stepwise regression revealed that Actual Evapotranspiration

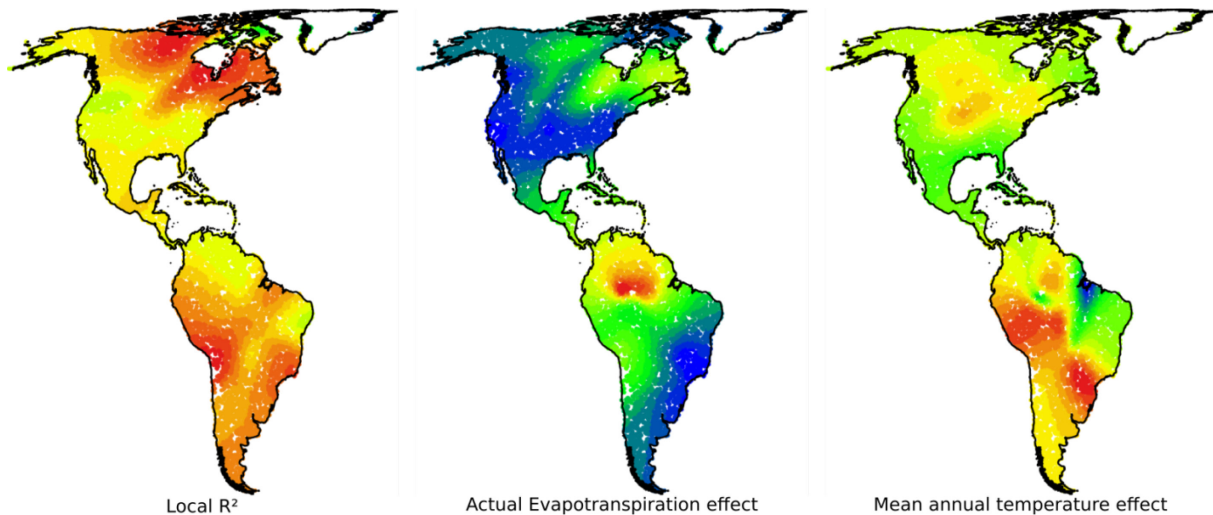
(AET) was the variable adding the most explaining power to the OLS models for all classes. Amphibia, Mammalia and Reptilia also exhibited a positive relationship between species richness and annual mean temperature (BIO1), while Liliopsida and Magnoliopsida species richness had a positive relationship with the Rapoport effect values computed with Steven's Method (RAPO-Stevens). Other explanatory variables did not significantly improve the models.

Moran's I tests were significantly positive for the seven classes, showing that spatial autocorrelation was detected. Spatial lag and spatial error models showed that the Gross and Snyder-Beattie hypothesis was not significant anymore, suggesting that its importance was overestimated when spatial autocorrelation was not considered (Table 5).

Once mapped, the GWR showed a strong spatial structuration effect of the variables on the species richness (Fig. 29). For most classes and variables, one or more 'hot' or 'cold' spots were visible, indicating an important deviation from global models that do not take into account the fact that a given variable has not the same impact over a large latitudinal gradient. No clear pattern could be drawn from all the maps but they all showed, for the seven classes, an absence of spatial stationarity of the explanatory variables.

**Table 5: Results of the regression models taking into account spatial autocorrelation or not.** For each class there are 2 model rows, the top one for the classic OLS model and the bottom one for the spatial regression model. The Gross and Snyder-Beattie model (GSB-null) was found significant only in the models that do not consider spatial autocorrelation (i.e. OLS models). Environmental tolerance (ET), distance limitation (DL) and ice cap extent (ICE) values, required for the GSB-null model, are provided for each class. AET = actual evapotranspiration (productivity hypothesis); BIO1 = annual mean temperature (ambient energy hypothesis); RAPO-Stevens = Steven’s original rule (Rapoport’s rule). No variable correlated with Pinopsida species richness.

Species	Model values for GSB-null			Best OLS model	R <sup>2</sup> adjusted (P-val)
	DL	ICE	ET	Best model after spatial regression	(P-val lag) (P-val error)
Amphibia	0.11	(69.00 ; -2.00)	0.33	Species richness ~ GSB-null + AET + BIO1	0.19 (<2.2e-16)
				Species richness ~ AET + BIO1	(<2.2e-16) (<2.2e-16)
Aves	0.46	(83.50 ; -6.00)	0.11	Species richness ~ GSB-null + AET	0.19 (<2.2e-16)
				Species richness ~ AET	(<2.2e-16) (<2.2e-16)
Liliopsida	0.21	(83.50 ; -6.00)	0.09	Species richness ~ GSB-null + AET + RAPO-Stevens	0.22 (<2.2e-16)
				Species richness ~ AET	(<2.2e-16) (<2.2e-16)
Magnoliopsida	0.17	(83.50 ; -6.00)	0.05	Species richness ~ GSB-null + AET + RAPO-Stevens	0.30 (<2.2e-16)
				Species richness ~ AET	(<2.2e-16) (<2.2e-16)
Mammalia	0.30	(83.00 ; -5.50)	0.03	Species richness ~ GSB-null + AET + BIO1	0.21 (<2.2e-16)
				Species richness ~ AET + BIO1	(<2.2e-16) (<2.2e-16)
Polypodiopsida	0.23	(83.00 ; -6.00)	0.07	Species richness ~ GSB-null + AET	0.22 (<2.2e-16)
				Species richness ~ AET	(<2.2e-16) (<2.2e-16)
Reptilia	0.15	(62.50 ; -4.00)	0.35	Species richness ~ GSB-null + AET + BIO1	0.22 (<2.2e-16)
				Species richness ~ AET	(<2.2e-16) (<2.2e-16)



**Figure 29: GWR results for Mammalia and 10,000 cells, projected as a map.** Hotter colors are for higher values. As explained the Local  $R^2$  (Right map) varies across space indicating that the regression model has varying explanatory power across space. The AET (productivity) effect also varies and seems to be of particular importance in the Amazonian basin, while the annual temperature (ambient energy) could have more importance in South-America.

## Discussion

Using GBIF-mediated data for eight taxonomic classes comprising plants and animals, we have investigated the Latitudinal Diversity Gradient over the New World. All classes but Pinopsida, a known exception (Stevens and Enquist 2000), show a clear LDG pattern, with a higher specific richness at lower latitudes, as previously reported (Willig *et al.* 2003). With regard to the LDG, the newly formulated geometric constraints hypothesis (GSB-null) has a greater explanatory power than the original geometric constraints hypothesis (*sensu* Colwell and Hurtt 1994). However, two important spatial issues (spatial auto-correlation and spatial non-stationarity) minimize this power and we underline here that these issues must be considered when geographic patterns are investigated.

For all classes with a LDG pattern, the GSB-null hypothesis surpasses the original geometric constraints hypothesis. Our results suggest that the GSB-null hypothesis should be considered as a null hypothesis in the study of biodiversity gradients, even though it is clearly

insufficient to explain species richness gradients and that our results unambiguously support a non-random distribution of species ranges. Species richness, however, is not the only parameter to investigate latitudinal gradients. Such gradients have indeed been highlighted using genetic diversity (Miraldo *et al.* 2016) or functional diversity (Stevens *et al.* 2003) and their study would benefit from the GSB-null hypothesis. The same reasoning holds for elevational gradients. Those remain studied with regard to the original geometric constraints hypothesis (e.g. Hu *et al.* 2016), whereas the GSB-null hypothesis has also been developed with a mathematical formalization for elevational gradients (Gross and Snyder-Beattie 2016).

While mechanisms for biodiversity patterns should be revised with regard to the GSB-null hypothesis, spatial autocorrelation and spatial non-stationarity must be included in statistical analyses more systematically than they used to be. When spatial autocorrelation is considered, the explanatory power of the GSB-null hypothesis decreases dramatically. Similarly, for all the variables and classes we examined, spatial non-stationarity has been found. Equivalent results were previously reported for the sub-Saharan endemic avifauna (Foody 2004) but GWR are still underused in biodiversity analyses (Mellin *et al.* 2014). This is an important issue to be considered and explored further because it suggests that prediction of species richness from environmental variables is not so straightforward, which might have consequences on conservation plans.

Besides these important methodological considerations, our results strongly support the productivity hypothesis, as implemented from the actual evapotranspiration (AET). A meta-analysis of 297 publications (Field *et al.* 2009) suggested that AET is a strong predictor of species richness. A recent contribution of vascular plants drew the same conclusion (Gillman *et al.* 2015). Here, using a large taxonomic coverage, we confirm the relative importance of the productivity hypothesis. At a lesser extent, the ambient energy hypothesis is supported for a few vertebrate classes, a result previously underlined (Hawkins *et al.* 2003a; Belmaker and Jetz 2015). Its more marginal role than the productivity analysis could result from the fact that ambient energy is supposed to be mostly influential at high latitudes (Hawkins *et al.* 2003a).

Other hypotheses, notably evolutionary hypotheses, should be considered in the future (Jablonski *et al.* 2017). Those hypotheses are among the most favored lately. They have

become easier to consider with the latest phylogenies and diversification analyses (Marin and Hedges 2016) but their inclusion remains difficult at large taxonomic scale analyses. They could help, however, discriminating whether the tropics act mainly as cradles or museums in species diversification and distribution (Rivadeneira *et al.* 2015, Pulido-Santacruz and Weir 2016, Siqueira *et al.* 2016, Hanly *et al.* 2017, Schluter and Pennell 2017), or as both as in the “out of the tropics” hypothesis (Jansson *et al.* 2013, Rolland *et al.* 2014).

Building a better model to estimate species richness should take into account the aforementioned elements. Our capacity to understand what influence species richness impacts our ability to design better conservation policies (Caesar *et al.* 2017) and predict future changes in species diversity and distribution. The support for the productivity hypothesis, for instance, is of peculiar importance because studies predicted that productivity is also dependent from species richness (Duffy *et al.* 2017), both variables acting as in a positive feedback loop. A better understanding of these complex mechanisms is pivotal in the current context of biodiversity crisis.

## **Acknowledgement**

This study was developed as part of a Ph.D. project and was funded as a grant by the Ministère de la Recherche to JT. We thank Peter Filzmozer, Anne-Sophie Archambeau, Samy Gaiji, Marie-Elise Lecoq, Sophie Pamerlon, Roseli Pellens, Tim Robertson, Dmitri Schigel, Jérôme Sueur and Wilfried Thuiller for fruitful discussions.

## **Chapter 5: DwCSP a fast biodiversity occurrence curator (article in preparation)**

Bioinformatics would benefit to various research programs but most biologists are not formally trained to bioinformatics (List *et al.* 2017). This well-known deficiency has become even more striking with the advent of Big Data because the mere quantity of data requires automatized procedures and so, at least basic bioinformatics skills. A lack of expertise can be offset through collaborating or using well-designed tools developed elsewhere. However, most software developed by biologists remains unpublished, restraining advances in research (Prins *et al.* 2015).

Arguably, unpublished software stays so because it has not been developed robustly or the developers think that their programs are not ‘good enough’ for publication and they apply self-censorship (Taschuk and Wilson 2017). This attitude, although humble, is counterproductive and is a waste of time and energy. Perfectionism must sometimes be put aside as an ‘imperfect’ tool would still improve reproducibility and accelerates research (Prlic and Procter 2012; Taschuk and Wilson 2017).

In this context, I developed a bioinformatics tool to handle massive biodiversity data occurrences along my PhD project. Even if this tool is imperfect, I aimed at sharing it widely. I invested time to increase its usability, looking for feedback from prospective users, providing logging information and using standard formats, and, above all, designing it to handle large amounts of data (List *et al.* 2017).

### **Introduction**

Biodiversity databases are becoming increasingly numerous and comprehensive (Hampton *et al.* 2013). Primary biodiversity data, or occurrence data, which indicates the presence of a taxon at a given location and time, is not exempt from this rule (La Salle *et al.* 2016). More and more of this data is being produced by scientists, but also by citizen science and amateur naturalists, who are grouped into vast networks (Bingham *et al.* 2017). Consequently scientists have access to increasingly large datasets that allow for new types of



analysis and more comprehensive studies. In addition, these analyses are becoming feasible for more and more taxa, locations and time periods. However, these changes are not without consequences and many biologists have called for caution regarding the quality of the data and therefore the quality of the studies using it (Yesson *et al.* 2007, Leonelli 2014).

As in many other sectors, scientific or not, the big-data era has arrived in the sciences of biodiversity and with it many new problems in particular data management and curation issues (Howe *et al.*). Data quality problems are reinforced by those curation issues. Indeed tools to improve data quality exist, but few are intended to process tens of millions of data. The statistical software R, for example, needs to load the data in memory to manipulate it and is therefore limited by the RAM of the computer on which it is running.

Yesson *et al.* (2007) showed that more than 15% of the data they downloaded from GBIF could be easily classified as invalid and disposed of using simple control systems. The authors further eliminated non-terrestrial occurrences, as well as occurrences in countries or regions where they were impossible according to a control database. However, they did this with the help of database management software as well as a spatial data management plugin. These tools are not trivial to use and, although very powerful and versatile, they require substantial technical expertise. To make better use of primary biodiversity data tools exist to enhance it by adding information or detecting outliers. The detection of outliers is used in multiple domains (He *et al.* 2003) and allows the efficient filtering of most of the trivial errors present in a dataset (error of data entry, coordinate inversion, identification error...).

We present here a program, called DwCSP (Darwin Core Spatial Processor), designed to process large volumes of primary biodiversity data. The software is capable of enriching a tabulated occurrence file (csv, DarwinCore, text file) with spatial data from polygon files (ESRI Shapefile) or Rasters file (geotiff). The software is also capable of detecting and tagging outliers based on their geographic coordinates (latitude/longitude) or numerical variables (environmental variables for example). The software has been designed to be fast, operational on very large datasets, and easy to use. It is available as a stand-alone java archive and executable on all machines with a java environment.

## **Software**

The program is available as an executable java archive. No installation is required if the java environment is already present. At runtime, the software proposes a presentation window to the user and a tab for each operation available in the application. Each tab also includes a short user guide.

The only usable tab when launching the application is the data selection tab. This tab allows you to select an occurrences file (csv, text, DarwinCore, Excel) and to fill in the information about it, in particular the names of columns containing the latitude and the longitudes of the occurrences. The other tabs cannot be used until the file has been checked. A file is deemed valid if it is readable by the application and contains the latitude and longitude columns indicated by the user.

## **Data Enrichment**

The DwCSP program makes it possible to enrich biodiversity data by matching the coordinates of each occurrence with the information contained in files for geographic information systems (GIS) projected in WGS84. One of the formats used by the program is the shapefile format initially developed by the Environmental Systems Research Institute (ESRI). This type of file contains geometry entities and diverse information about them like names or statistics. These entities can be represented as points, lines or polygons. The program accepts shapefiles containing polygons that do not intersect. It then computes the position of each occurrence in relation to the shapefile and, if the occurrence is located in a polygon, it attaches the intersecting polygon information to the occurrence. The user must indicate to the program what types of information he wants to add by naming each variable of the shapefile he wants to include. If a shapefile entity does not have one of the selected variables, the program produces an error when checking the file.

In order to speed up the mapping between occurrences and the polygon file, the program uses R-Tree structures (Guttman 1984) and multithreading. R-tree structures are used to index the polygons inside the shapefile. This data structure makes it possible to greatly accelerate the mapping of a point (an occurrence) and a geographical entity (polygon).

Matching is further accelerated by the use of multithreading, which parallelize the tasks performed by the program. In this case the program creates the number of threads desired by the user and then runs the occurrence file line by line. It assigns each occurrence to a free thread and waits in case no thread is available. Each thread has its own R-Tree spatial index because R-Tree objects do not support simultaneous access. Each thread writes the results line by line in a temporary text file. These temporary files are then collected in the output file once all occurrences have been processed. Note that the program does not store any occurrences in memory after processing, which allows managing very large datasets.

A similar process is used when matching a raster file with occurrences. A raster file is an image containing geo-referencing information. The program is compatible with files in GeoTIFF format (Sazid and Ramakrishnan 2003). Each pixel of the image represents a geographical area and contains numerical information about that area (temperature, humidity, altitude...). The program searches for the pixel corresponding to the occurrence's coordinates and adds the value of this pixel to the variables of the occurrence. It is possible to provide the program with several raster files. The occurrence is compared to each of the files and one variable per file is added to the occurrences. In the final tabulated file the name of the new columns will correspond to the name of the raster files. The use of an R-Tree index is not possible in the case of raster files, but parallelization of calculations between multiple threads is used. The temporary file system described above for shapefile is also used here.

### **Searching for outliers**

The second type of operation proposed by the program is the search for outliers. Outliers are often the result of an error in entry or identification in primary biodiversity data. The DwCSP program proposes two common and relatively simple methods for detecting outliers.

Before presenting these methods, it is necessary to explain how the application sorts the data to form the groups in which the outliers will be searched. Data is only aberrant within an otherwise coherent set of data, so these errors must be looked for within a species or taxon. Here the application suggests that the user designates one of the columns in the occurrences file to be used as an identifier for the application, for example a column containing the

scientific name of the observed species or a numerical identifier specific to each taxon. The application then copies the data into temporary files, one file for each unique identifier. The rest of the computations are done on these files.

The first method of outlier detection is a spatial method based on the occurrences' coordinates, meaning only the occurrences file is needed. The method consists in calculating a measure of spatial eccentricity for each occurrence. The user indicates to the application the number  $N$  of neighboring occurrences to be used as well as a percentage  $P$  of occurrences to be classified as outliers. The application then calculates the cumulative orthodromic distance (distance to the surface of the globe) in km between the studied occurrence and the nearest  $N$  occurrences. Once all eccentricity measurements have been performed for a group, the application designates as outliers the  $P$  % of data with the highest eccentricity. This value can be adjusted by the user but we recommend using low values (1 to 5%) to keep most of the data. As the eccentricity value will be put into the output file, the user can also "handpick" the outliers by removing only the points with the highest eccentricity. Once the operation has been performed on all taxonomic groups, the temporary files are merged into a single results file. The result file includes two additional columns, the first contains the value of the spatial eccentricity and the second the status: outlier or non-outlier. In order to speed up this operation the multithreading is used there, each thread being taken care of a particular group to treat several groups in parallel. A closer neighbor search algorithm called KD-Tree (Bentley 1975) is used to quickly find the nearest  $N$ 's closest neighbors to an occurrence.

The second method available in the program uses another distance measurement called Mahalanobis distance (Mahalanobis 1936). This distance is computed using a set of numerical variables (typically climatic variables) which can be the ones added during the data enrichment steps. Each occurrence is compared with the entire occurrence distribution of the group. The program therefore computes the distance of Mahalanobis from each occurrence to the rest of the group and designates as outliers the proportion of occurrences with the highest distance from Mahalanobis. Again multi-threading and temporary file methods are used to speed up calculations. For now the software can't exclude a data when doing the computations, meaning that during the environmental outlier steps, spatial outliers will not be filtered out. Consequently the two computations can flag the same occurrences as outliers or produce different results.

## Results

For all computation we used a computer with a 12 cores processor with a speed of 1.90 GHz and 64 GB of RAM. The performance of such a computer is clearly superior to a conventional desktop computer. However the dataset used is also far bigger than a "normal" dataset (e.g. García-Roselló *et al.* 2015 Charbonnier *et al.* 2016, Chaudhary *et al.* 2016). The software was tested on a dataset consisting of 45,948,943 occurrences downloaded from the GBIF portal (<https://doi.org/10.15468/dl.hickgt>). This dataset contains all geo-referenced data collected in 2012. It represents 20.6 GB of data in the form of a text file in DarwinCore format.

For the first step the shapefile was downloaded from [www.natureearthdata.com](http://www.natureearthdata.com). Data at 1:10m scale containing the 247 countries covering the Earth's surface were downloaded (Admin file 0 - Countries). We have set up the application to add the 'SOVEREIGNT' and 'TYPE' columns containing the polygon name (name of the country) and the polygon type (Country, Sovereign Country, Dependence...) to the occurrences, respectively. The other parameters used are a number of 100 threads and no line number limit. The application was started with the `-Xmx42g` command to allow the Java virtual machine to use more RAM than the default values. Adding this information to the occurrences required a total of 66 minutes, including 29 minutes of computation and 36 minutes of CSV file manipulation. This gives an average of 1.45 minutes per million occurrences.

For the second step, we downloaded all the raster data available on Worldclim (Fick and Hijmans 2017) as 19 Raster files. We have chosen raster files defined at 2.5 degrees. 19 columns were therefore added to the occurrences by this operation. Once again the command `-Xmx42g` was used and this time 20 threads were performing parallel calculations. It took 39.1 hours of computation to add the information from the 19 raster files to all occurrences, including 38.5 hours of computation. This gives an average of 2.6 minutes of computation per million occurrences per raster file.

To search for spatial outliers we used the following parameters: the 'species' column was used as the identifier column, 20 occurrences required to start the search, 5% of occurrences to be classified as outliers, 5 neighbors are used to compute spatial eccentricity

and a 100 thread limit was chosen. The computation time of this step was 104 hours for 123,526 species. This is an average of 2.3 hours per million occurrences. Some optimization work is still in work for this step.

Finally the search for environmental outliers was done using the following parameters: the 'species' column was used as the identifier column, 20 occurrences required to start the search, 5% of occurrences to be classified as outliers, the bio1, bio2 and bio12 Bioclim values (added previously as column) are used to compute Mahalanobis distance and a well as a 20 thread limit. For now the process can't be completed as there is a problem of singular matrix when computing the Mahalanobis distance.

## **Discussion**

With the accelerating production of biodiversity data (Devictor and Bensaude-Vincent 2016) it becomes urgent to develop, in the same time, tools capable of using this data. There are several barriers to the use of large datasets, but two of these issues are particularly relevant. First of all, available software solutions are rare for very large datasets and those solutions often require significant computer skills (Gaiji *et al.* 2013). Secondly, large datasets are often heterogeneous because they consist of an accumulation of data from different producers (Gaiji *et al.* 2013). This diversity is the source of errors and often cited as a major concern when using these data (Troudet *et al.* 2017).

The application presented here addresses both of these issues. It does not require the installation or use of third-party software and has a simple graphical user interface that requires no computer knowledge. The application makes it possible to enrich biodiversity data and also to ensure its quality by eliminating the most abnormal data. Moreover, the processing speed of this data allows the software to be used on several tens/ hundreds of millions of occurrences at once with a relatively short processing time.

Therefore the DwCSP program has great potential to enhance the value of large datasets. It not only enables scientists who use the data to ensure their quality, but also enables people who maintain biodiversity databases to ensure the quality of their data and even improve their management systems by integrating this software.



## Discussion

Life on Earth can be studied at various scales (e.g. ecosystems, species, behaviors, structures, molecules, etc.). Many scientists spend their careers studying the finest elements of life with a magnifying glass, literally and metaphorically. Others choose to take a step back and investigate life at species or regional levels. From the largest to the narrowest scale, there is no wrong way to try deciphering how biodiversity originated and thrived. These approaches complement each other and contribute bringing a better understanding of biodiversity.

Even though the elements of life investigated differ according to the scale chosen, the way these elements are sampled is always paramount. All researchers studying life on earth, being ecologists, biogeographers or systematists, pay special care to how they delineate the samples they will analyze at a later stage. They do not, however, design their samples following the same rationale because their aims differ. Yet again, practices in one field should benefit to other fields, and reciprocally (Bortolus 2008; Schilthuizen *et al.* 2015; Ward *et al.* 2015), as the success of trans-disciplinarity as shown (Tress *et al.* 2005).

Museum collections and primary biodiversity databases are incredible tools to investigate life on earth (Knapp 2017) and should be grasped by the whole scientific community interested in biodiversity. Although built by scholars from different fields, including systematics, they seem mainly used by ecologists (Hampton *et al.* 2013), biogeographers (Brown and Lomolino 1998), or conservationists (Joppa *et al.* 2016). The study of biodiversity is yet a task shared by systematists and non-systematists alike. Arguably, systematists could bring a complementary perspective on sampling, an issue they have addressed since a long time (Greene 2017).

With this context in mind, I tried to analyze biodiversity data and patterns from a systematics perspective, using the GBIF mediated data. Working at large taxonomic and geographical scales has drawbacks that a more focused approach would alleviate, but we are not yet done taking advantage of the strength it also brings.

### Big-data and biodiversity



The emergence of big-data in biodiversity sciences is both a boon and a challenge that requires technical adjustments. Biodiversity data are accumulated and shared faster than ever allowing global and powerful studies. Big datasets, however, are more difficult to handle and analyze and the deficit in biocuration identified ten years ago has not been filled yet (Howe *et al.* 2008). The study of life is not the first field to have entered into the big-data paradigm so that it could learn from previous experience, even if biodiversity data has its specificities in terms of data acquisition, storage, distribution, and analysis.

### **The big-data paradigm**

The shift towards what has been called the big-data paradigm, and now encountered in the study of biodiversity, is not new in science (Kitchin 2014). The impact of big-data on science has been shown to be massive in various fields of research (Boyd and Crawford 2012). In Biology, DNA is a typical example of a kind of data that has blown out in terms of quantity and availability, to the point that genomics could soon exceed other big-data domains in terms of data quantity (Stephens *et al.* 2015). For biodiversity studies, this change takes more time because producing primary biodiversity data of quality is a difficult and lengthy process (May 2004). Consequently, biodiversity is still at the very beginning of its paradigm shift, an opportune time to shape how this new paradigm should be embraced.

Big-data refers to datasets so large that traditional data processing tools cannot handle them. However, big-data is not merely a matter of data quantity and other features characterize big-data. Kitchin (2013) defines big-data mainly according to its i) volume of data, ii) great speed of data production, iii) high data heterogeneity, iv) data exhaustiveness and wide coverage. These multiple features justify why big-data are seen as a new scientific paradigm, sometimes called exploratory science, that change the scientific approach (Kelling *et al.* 2009, Boyd and Crawford 2012, Kitchin 2014).

A striking change is how the scientific logic seems to have evolved with big-data. In a "classical" scientific approach, a hypothesis is formulated to explain an observation and then data are collected and analyzed to confirm or refute the hypothesis. With big-data the approach is "data-driven": we already have the data and are looking for remarkable patterns within the data. These patterns are then compared with assumptions about data production processes. Authors have talked of "born from the data" patterns (Kelling *et al.* 2009). This

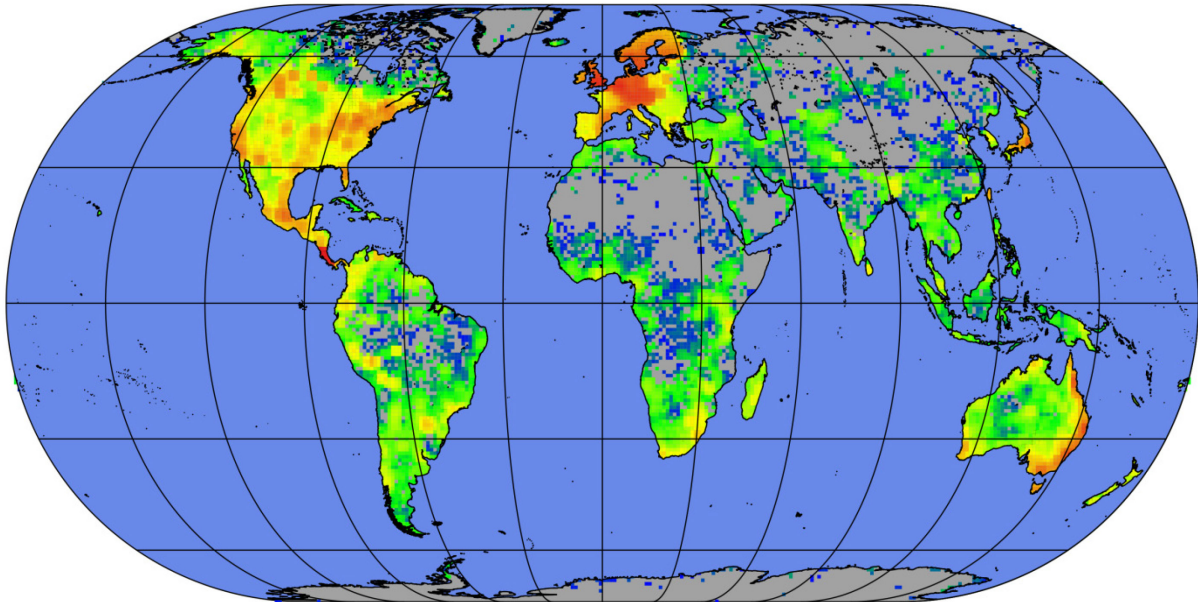
approach is more exploratory than the “classical” one yet not bad or better, and it has been argued that it will form a new field within ecology (Michener and Jones 2012, Canhos *et al.* 2004).

The shift of focus from fine and local mechanisms to a more global vision of the studied mechanisms has triggered a ‘datafication’ process (Devictor and Bensaude-Vincent 2016). The exploratory approach is an incentive to and justifies producing more data. Data accumulation becomes an objective rather than items to study a phenomenon (Boyd and Crawford 2012), at the risk of accumulating data for the sake of data accumulation. Gathering data without a precise objective, analyze these data and draw conclusions is not necessarily harmful for science but it should not replace standard scientific practice either. It can be a serious issue for domains that produce large amounts of data as a by-product of their core business (e. g., Twitter and other social networks). For the study of biodiversity, the consequence could be highly detrimental if the data collection trend is not uniform, as we have shown, as it would deepen the taxonomic bias.

Also, as promising as the big-data paradigm might sound (La Salle *et al.* 2016), it must be seized with cautions because any dataset, big or not, remain a sample of the reality. The main sources of big-data today, generating billions of data every day about millions of users, are social platforms on the web such as Facebook, Twitter and YouTube (Stephens *et al.* 2015). However, a societal study based on those data would be a study of the subset of humans using those platforms, and not of all humans. The same rationale holds for biodiversity data: the outcomes of a study relying on the largest biodiversity dataset would be inapplicable to most of biodiversity because it has not been sampled yet (Larsen *et al.* 2017). Unfortunately, the high quantity of data might sometimes obscure this reality, giving a false impression of objectivity and accuracy (Boyd and Crawford 2012). A quick look the distribution of insects specific richness derived from raw GBIF mediated-data (fig. 30) would yet clearly show that large datasets also contain biases.

All those changes prompted Kitchin (2014) to argue that: “*Big Data and new data analytics are disruptive innovations which are reconfiguring in many instances how research is conducted; and (2) there is an urgent need for wider critical reflection within the academy on the epistemological implications of the unfolding data revolution, a task that has barely*

*begun to be tackled despite the rapid changes in research practices presently taking place”.* These big-data related problems have been identified in various fields and they undoubtedly apply as well to biodiversity. Lessons have been drawn and should be applied to the study of biodiversity to avoid repeating the same mistakes.



**Figure 30: Insecta species density across the globe according to raw GBIF-mediated data shows a sampling bias.** Warmer colors indicate higher species richness and grey area a lack of data. This map was obtained when simply computing the number of species in 100\*100km cells using the GBIF data without any filtering or correction of the sampling bias.

## **The genesis of Biodiversity big-data**

### ***Producing biodiversity data***

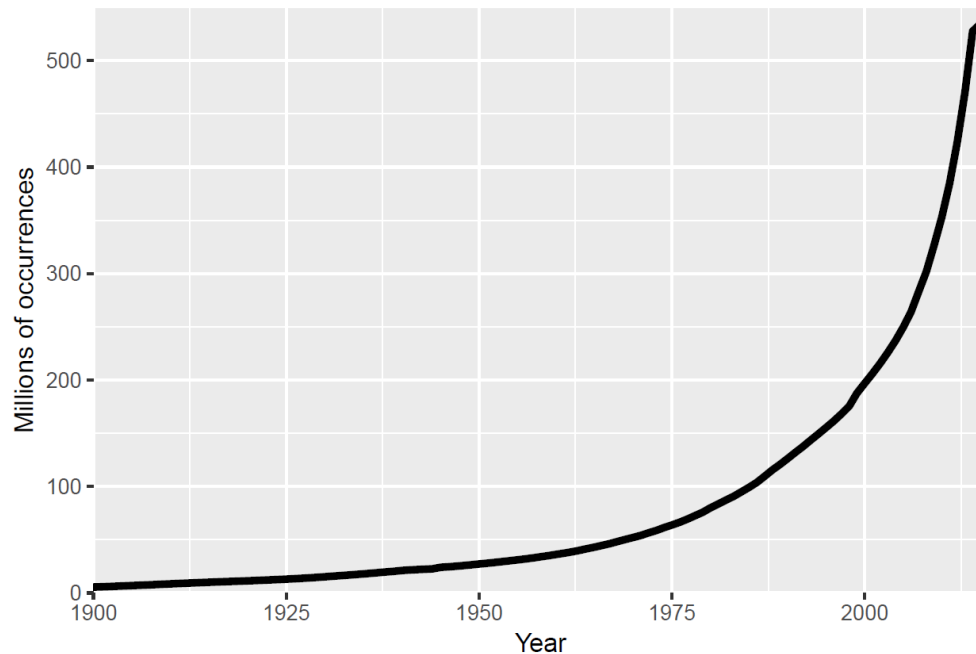
Big-data sciences like astrophysics and societal sciences based on web data rely on datasets with a fine resolution and a fast creation process (Kitchin 2014), because they are either very recent, created from an informatics infrastructure (Twitter, YouTube...) or were acquired according to a process tuned from the very beginning to produce high quality and high quantity data (Leonelli 2014). For biodiversity, the data production process strikingly differs from these data intensive sciences. Biodiversity data have been produced from the early days of systematics so that some date back to a few centuries (e.g. 60,000 GBIF-

mediated data were collected in the 17<sup>th</sup> century) while others are produced today. Thus, the GBIF portal contains data from the digitization of museum collections (Sikes *et al.* 2016), data produced as part of ecological studies (Hampton *et al.* 2013), observations from networks of amateur naturalists (Sullivan *et al.* 2009) and citizen sciences (Chandler *et al.* 2017). At the time of writing (20-09-2017), GBIF has 1,360 data publishers who provided data in almost every country. These large timespans and diverse origins result in biodiversity data far more heterogeneous than in other big-data sciences, which poses one of the biggest obstacles to their use.

Even though available biodiversity data is expanding at an exponential rate (Fig. 31, Isaac and Pocock, 2015; Gaiji *et al.* 2013), the production of new biodiversity data is still an important limitation (May 2004). It is perhaps the most distinctive feature of primary biodiversity data: it is hard to produce. As opposed to many other big-science data sources, primary biodiversity data producers are still mostly humans (May 2004) and creating a species occurrence has not been yet easily automated. Only 9.4 million of the 837.3 million occurrences of the GBIF-mediated data are classified as “machine observation” (20-09-2017).

In an effort to produce and share as biodiversity data as possible, numerous projects have been developed: citizen science, naturalists’ networks, and mass digitalization of museum collections (Le Bras *et al.* 2017). Citizen science and naturalists’ networks are the biggest data producers at the moment. The ebird network (ebird.org), composed of a multitude of amateur ornithologists, represents the biggest data provider in the GBIF (i.e. 275.7 million occurrences). As one of the main limitations in the study of biodiversity is the lack of people (Godfray 2002, Rodrigues *et al.* 2010), the use of citizen science is especially effective and can serve different purposes such as taxa identification (Silvertown *et al.* 2015, Martin and Harvey 2017) or data production (Chandler *et al.* 2017, Tiago *et al.* 2017). As for museum collections, it is estimated that they contain more than a billion specimens (Ariño, 2010), whose digitization would result in a biodiversity dataset much more varied and faster to produce than an equivalent collection of new data (Beaman and Cellinese 2012, Blagoderov *et al.* 2012). Collection objects are troves of “new” data (Knapp 2017) with great values as they inform us about past ecosystems (Escribano *et al.* 2016) while being the foundation to be complemented with modern data (Sosef *et al.* 2017). Museum collections and citizen science are not exclusive sources of biodiversity data. In the Muséum national

d’Histoire naturelle, for instance, volunteers were asked to read tags of botanical plates (on photos) from the herbarium collection before rewriting them using a web interface, allowing for the digitization of 5.4 million herbarium specimens (Le Bras *et al.* 2017).



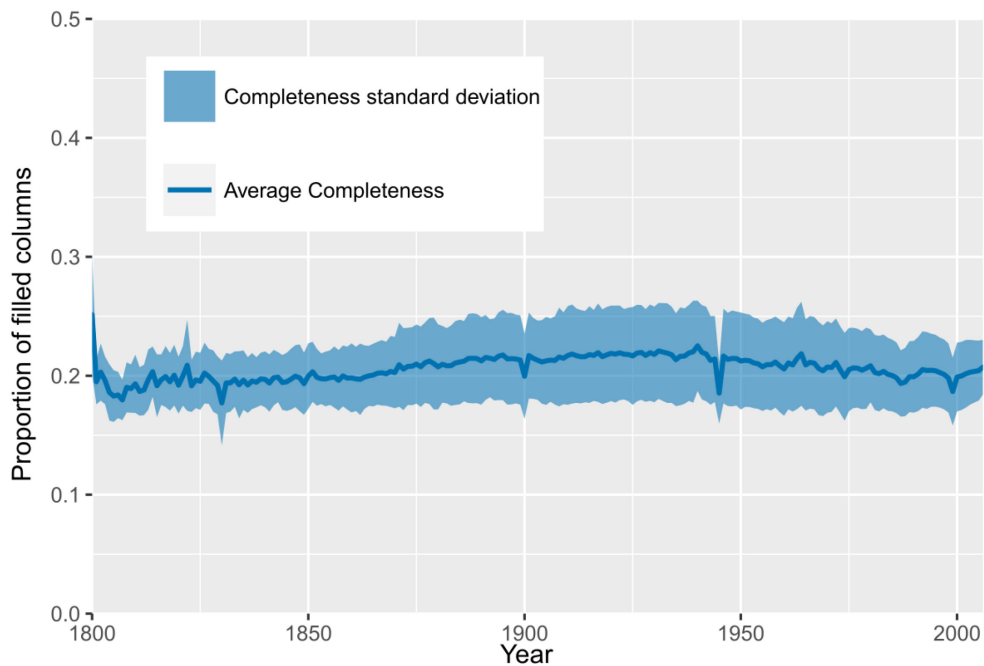
**Figure 31: Accumulation curve of the number of occurrences available in the GBIF.**

Two other “troves” could produce massive amount of data: ecology and grey data. Ecology is more and more a data intensive science (Michener and Jones 2012) and sometimes relies on available data such as those found through the GBIF portal. However, many ecology studies still depend on data specifically collected. Given the high number of ecology papers produced, an enormous amount of data should be produced. Unfortunately, while biodiversity occurrences are effectively collected, very few of them are shared after publication (Ward *et al.* 2015), resulting in what has been called dark data (Hampton *et al.* 2013). This unavailable data adds up to grey data or grey literature (Boakes *et al.* 2010), defined as data existing but not shared because of a lack of time or interest (Heidorn 2009). This situation is unfortunately too well-known for most researchers: everyone has at least one file on a computer or a shelf that could be put online, but would need some time to be formatted before being shared.

### ***Maintaining biodiversity data***

Once collected, a biodiversity occurrence should be stored and shared, tasks one could think to be easy to perform with the advent of online databases facilitating them (gbif.org, genbank, map of life, iNaturalist...). However, it is not possible to fully entrust the data management task to these entities. The GBIF, for example, is able to store and make accessible all the occurrence data of a project or collection, but it will not correct or change the data, only flag them in some circumstances (Dmitry Shiggel *pers. com.*). It is up to the data provider to ensure "after-sales service", a task far from being trivial.

The GBIF provides biodiversity occurrences of varying quality and quantity across space, time, and taxa. Biodiversity occurrences can lack basic information such as time and place of the observation as well as additional information that are not mandatory in the GBIF (elevation, sex, sampling protocol, *etc.*). With a DarwinCore format composing of 230 columns, GBIF-mediated data cannot be complete, and most of the columns stay empty (Fig. 32). Some of these columns are inappropriate for a given occurrence but others are important like the one providing the link to the specimen (less than 5 % of the specimen-based occurrences had an identification number recorded). If not provided when recorded, this information will probably not be added posteriorly. Very few of the Data Providers edit the data after it has been deposited in the GBIF (Robertson T. *pers. com.*). The data creation effort has to be thorough from the beginning, as data will unlikely be improved later. In addition to a lack of information, biodiversity occurrences may suffer from errors introduced when creating or digitalizing the data (Yesson *et al.* 2007; Goodwin *et al.* 2015). When errors accumulate in biodiversity databases, their usefulness is jeopardized. The most common errors are misidentification of specimens and typing errors during computerization of the data (inversion of coordinates, typing error...) (Yesson *et al.* 2007). Most of the time, they can easily be identified and corrected, but it still time-consuming.



**Figure 32: Average completeness of the GBIF mediated data per year does not evolve along time.** The blue line represents the average proportion of columns filled in the DarwinCore format. The blue area represents the standard deviation of this value. The average completeness of the data doesn't change much over the years and is never above 25 %.

Adding missing information and correcting datasets are boring and tedious tasks for which researchers are offered very little compensation. Encouraging the publication of data papers, even for data correction, and increasing their value and appeal by citing them when using the data would be a first step to improve the final quality of produced data sets. A further improvement would be to offer researchers optimized data entry interfaces (more adapted to the current uses and standardized than the commonly used spreadsheet software), including automatic verification tools for the most common errors. A more time-consuming solution would be to set up a system for reviewing the data before putting it in the database, similarly to a publication. However, this solution is undesirable if we consider the workload that is already weighing on researchers.

In addition to these suggestions occurring while creating a biodiversity occurrence, a system of data control *a posteriori* could also be set up. Again, there are many possible solutions, but given the large volume of data to be verified, I think two options are really

worth considering. The first solution would be to implement a computer system capable of automatically verifying the data, such as, for instance, detecting outliers in a dataset (see Chapter 5). The second solution would be to engage the public and create a crowdsourcing platform for verifying biodiversity data according to previously established rules.

With cleaner and more complete occurrences, biodiversity datasets will become more and more useful. To facilitate the transition towards an efficient use of large datasets, Howe *et al.* (2008) proposed three main lines of action : *First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, researchers, academic institutions and funding agencies should, in the next ten years, increase the visibility and support of scientific curation as a professional career.*

Drawing on my PhD experience, this 3<sup>rd</sup> point strikes me as the most crucial. It took me a very long time to take hold of the GBIF mediated dataset and be able to use it. Most ecologists or systematists, already overloaded with work, cannot easily invest so much time and efforts to understand all the subtleties of the DarwinCore format and its different versions. Most of them will not take either the time to clean, share and maintain their data beyond what they already do in their current practice. Even with good incentives they would have to make a choice between doing original research work and data management. The support of scientific curation as a professional career has already been pleaded (Howe *et al.* 2008) and should be promoted in every structure producing or storing biodiversity data. This new demand for data-scientists could however enter in conflict with other big-science domains and the private sector. The big-data paradigm being new and full of opportunities in various fields, finding skilled people to work in the biodiversity domain could be tricky (La Salle *et al.* 2016).

### **A new way of doing science**

The way biodiversity data is produced and maintained has inevitable consequences on its uses. Kitchin (2014) summarized the challenge of analyzing big-data as “*coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty,*



*high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity”.*

The heterogeneous nature of biodiversity data has been one of the main reasons why data from the GBIF, for instance, has been criticized or negatively reported (e.g. Yesson *et al.* 2007). This heterogeneity is indeed an issue to be mentioned but it does mean that the data, and the studies relying on them, must be ignored. GBIF mediated-data reflect the practices of all the protagonists involved in the study of biodiversity, practices notoriously imperfect. Consequently, the very first step when using large datasets is to understand its gaps and biases before ascertaining whether the data can be used to conduct a study. This task requires data-mining and data-filtering. Conscious of these needs, the GBIF portal provides filtering options using many criteria (Fig. 33), a process that must be complemented with others because the large volume of data makes it insufficient otherwise.

Search all fields 🔍

**Simple** Advanced

Record License ▼

Scientific Name ▼

Basis Of Record ▲

<input type="checkbox"/> Observation	30,393,351
<input type="checkbox"/> Literature	574,203
<input type="checkbox"/> Preserved Specimen	126,119,103
<input type="checkbox"/> Fossil Specimen	6,685,567
<input type="checkbox"/> Living Specimen	1,465,322
<input type="checkbox"/> Human Observation	640,774,541
<input type="checkbox"/> Machine Observation	9,843,450
<input type="checkbox"/> Material Sample	508,439
<input type="checkbox"/> Unknown	31,759,089

Location ▼

Year ▼

Month ▼

Dataset ▼

Country ▼

Issue ▼

Media Type ▼

Publisher ▼

Institution Code ▼

Collection Code ▼

Catalogue Number ▼

Type Status ▼

**Figure 33: Filtering occurrences in the GBIF.** The GBIF portal allows for filtering the occurrences using many parameters. Here are just a few of them displayed by the simple interface.

A second important reason that has slowed down the use of big-data in biodiversity is the computing power required to handle the data (Rosenheim and Gratton 2017), from verifying its quality to conducting analyses. In the last two decades, however, technological progress and software innovation allowed for the analysis of very big datasets (Kumar and Kumar 2016), even though the amount of data available increases faster than the computing

power available to most researchers (Walker 2014). The situation is significant in biodiversity analyses: most computer programs commonly used in ecology and systematics (Spreadsheet, R, Maxent...) are not intended to handle tens of millions of occurrences.

Fortunately, like for data curation, other fields have faced the problem of big-data analyses before the study of biodiversity turned in the big-data paradigm. As previously underlined, biodiversity data takes a long time to produce compared to other sciences, meaning that large biodiversity datasets have not yet reached the massive quantities of other sciences (Leonelli 2014). Once downloaded and uncompressed, the GBIF data set represents more than 500 GB of data, which is not so large when compared to the 15GB of data the Hubble telescope produces per day or the 50GB of data the Gaia project produced per mission (Jordan 2008). This relatively limited amount of data is an opportunity for biodiversity science because big-data is at a more advanced stage in other fields, which are at the forefront for creating tools to manipulate this data (Rosenheim and Gratton 2017).

While relying on advances from other fields, specific tools for the study of biodiversity need to be developed and are currently flourishing (e.g. Cameron *et al.* 2016; Deck *et al.* 2017, Martin and Harvey 2017; Töpel *et al.* 2017). In addition to the specific program presented in chapter 5, I have participated to the creation of an online biodiversity data curator application, Biodatascreen, which displays a workflow checking coordinates, localities, and taxonomy of primary biodiversity occurrences. This creation echoes the development of similar tools such as BioVel (Vicario *et al.* 2011) or speciesgeocodeR (Töpel *et al.* 2017) but emphasized simplicity of use to target as many and diverse users as possible.

## **Primary Biodiversity data, a proxy to assess the state of the study of biodiversity**

The core of a primary biodiversity occurrence is found in three pieces of information: a name, a place and a date. Since a very long time, biodiversity occurrences have been gathered with these elements and they have supported a large spectrum of biodiversity analyses. By its ubiquity and importance, this data is an effective tool for studying the practice of biodiversity sciences. Primary biodiversity data is truly the foundation on which biodiversity knowledge is built.

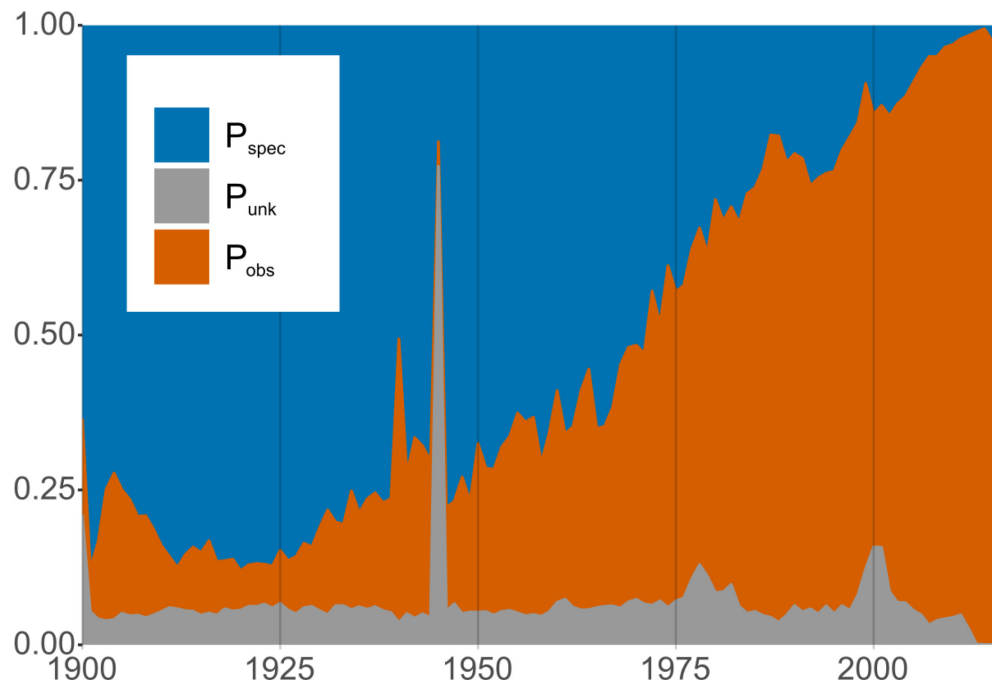
Data occurrences have been collected for several centuries (Le Bras *et al.* 2017) and are still produced today, but I showed that biodiversity data collating has evolved from a specimen-based to an observation-based practice. Thus, I will first delve on the particular relationship between biodiversity occurrence data and specimens (i.e. items in collections). Then, even though this practice shift applies to all the eukaryotic groups investigated, biodiversity data occurrences do not accumulate similarly across taxonomic groups. In a second phase, I will concentrate on the differences of treatment between different taxonomic groups as well as the reason and consequences of such differences.

### **Biodiversity data are disconnected from specimens**

Concomitantly to the big-data revolution, the study of biodiversity faces a new trend: specimens and biodiversity occurrences are disconnected. While more observation occurrences are produced (Gaiji *et al.* 2013), less specimen vouchers are collected (Turney *et al.* 2015). Collecting specimens is, however, one of the oldest traditions of naturalist sciences (Nualart *et al.* 2017), a habit that has made possible to start cataloguing and classifying life on earth. Although the first collections were mere objects of curiosity, they became very powerful study tools (Buerki and Baker 2016), recently compared to the gigantic technological tools used in other sciences (Knapp 2017). And yet, increasingly overwhelming amounts of biodiversity data are produced without supplying specimen collections (Chapter 2, Gaiji *et al.* 2013).

This is a relatively recent change, consistent with four main modifications that occurred in the naturalist and scientific environment in the last few decades (Fig. 34). The first change is the increase in public and scientific awareness of biodiversity erosion (Cardinale *et al.* 2012). This realization has been accompanied, for ethical and legal reasons, by profound changes in the way specimens are harvested (Dubois 2017). The second change relates to the development of computers and Internet. With computers and web technologies, scientists have new tools to manipulate and use biodiversity data, which obviously must be digitalized (Grandcolas 2017). The third factor could result from the success of ecology, a discipline that, unlike taxonomy and phylogeny, is more independent to specimens (Grandcolas 2017). The fourth factor contributes mainly to the most recent part of the increase in observational-based occurrences: the boom of citizen sciences and amateur

naturalist networks (ebird.org, iNaturalist.org...). These four changes did not happen concomitantly but have likely contributed to reinforce the observational trend in biodiversity data gathering.



**Figure 34: Proportion of occurrences per year of collect and origin cumulated for 24 classes.** Orange, blue and grey areas represent the proportions of observation-based ( $P_{\text{obs}}$ ), specimen-based ( $P_{\text{spec}}$ ) and unknown origin ( $P_{\text{unk}}$ ) occurrences, respectively. Contrary to 50 years ago, a majority of observation occurrences is reported.

Even though specimen-based occurrences are not devoid of issues (Goodwin *et al.* 2015) and that observation-based occurrences are not necessarily harmful, the latter are more prone to a lack of reliability (Bortolus 2008, Turney *et al.* 2015), reusability (Ferro and Flick 2015) and versatility (Buerki and Baker 2016) than the former. The absence of voucher specimens in sometimes enormous datasets is thus worrying, especially for group of organisms known for their difficulty of identification such as insects. Yet, some insect species are referenced with thousands of observation-based occurrences in the GBIF and not a single specimen-based occurrence (Table 6). Some of these insects are linked to health, economic or agricultural issues and previous examples of misidentifications in similar circumstances have

proved disastrous (Bortolus 2008). This concern is less pronounced, but not absent, for most vertebrate taxa because they are less prone to identification errors.

**Table 6: Top 10 list of species with the most data and only observation occurrences in the GBIF.** Those ten species have at least two thousand occurrences in the GBIF but not a single specimen-based one. Therefore it could be very hard to test the accuracy of those observations.

Species	Class	# Observation-based occurrences
<i>Etropus crossotus</i>	<b>Actinopterygii</b>	<b>5740</b>
<i>Blastobasis adustella</i>	<b>Insecta</b>	<b>5150</b>
<i>Ablabesmyia longistyla</i>	<b>Insecta</b>	<b>4584</b>
<i>Acanthis cabaret</i>	<b>Aves</b>	<b>2895</b>
<i>Ameletus alpinus</i>	<b>Insecta</b>	<b>2821</b>
<i>Tockus rufirostris</i>	<b>Aves</b>	<b>2292</b>
<i>Zentrygon frenata</i>	<b>Aves</b>	<b>2183</b>
<i>Apsectrotanypus trifascipennis</i>	<b>Insecta</b>	<b>2132</b>
<i>Ablabesmyia phatta</i>	<b>Insecta</b>	<b>2067</b>
<i>Cubaris plasticus</i>	<b>Malacostraca</b>	<b>2047</b>

Moving from specimen to observation-based occurrences has also consequences on data curation and database management methods. In both cases, the longevity of the database is correlated to the way it is managed and updated (Howe *et al.* 2008), which entails a cost. One could reasonably assume that observation-based occurrences and databases would be less costly because they do not include specimens to manage and the associated strenuous curation. Natural history collections, however, have already proven their longevity, tested along centuries despite inevitable damaged or lost specimens. Collections are also a way to bypass the problem of forgotten data (dark data) and unreadable data (floppy disks, old files formats, forgotten literature...). Never a specimen will go unreadable (unless destroyed) and specimens are usually treated with more care than digital data and are less likely to be lost (Heidorn 2009).

The use limit of an occurrence (its "expiration date") is also an important parameter but is more difficult to estimate. Data of any age is always useful; however some of its uses can dramatically wear off after a while. It can be argued that specimen data has a longer

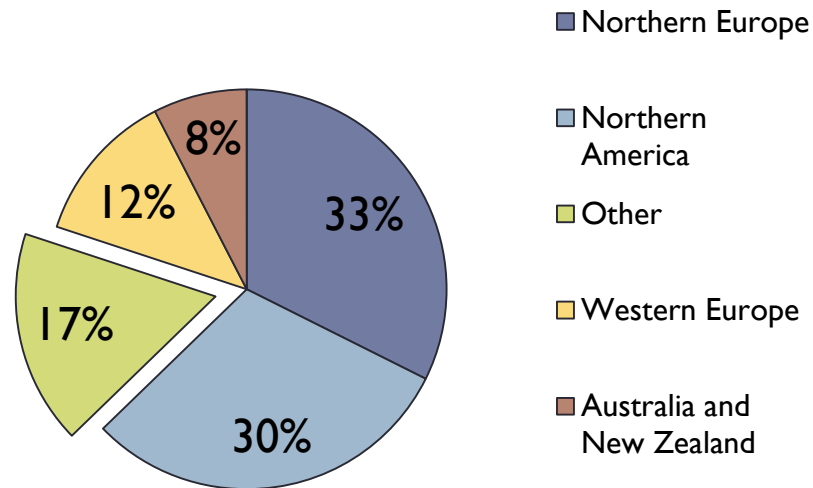
expiration date, a scientific potential that will last longer than observation based data (Turney *et al.* 2015). All biodiversity occurrences, whatever their natures, have possible uses when recorded and in the future, what could be called a heritage value. This heritage value, however, may lose part of its relevance after a while and the data becomes obsolete. Escribano *et al.* (2016) have shown that biodiversity data can reach obsolescence, for instance, when the environment in which the data was collected changed too much. In this situation, observation-based occurrences become merely indications of the species present in the past. But for specimen-based occurrences, a physical object remains to be studied, unless it was lost or destroyed, extending its usefulness long after the species has disappeared from a given place or has faced extinction (Turney *et al.* 2015).

Consequences of the paradigm shift in biodiversity data gathering can be envisioned but they should be quantified in a near future to be better apprehended. How many studies and researchers use specimen-based or observation-based occurrences? How many studies using specimens would not be possible to conduct with mere observations, and reciprocally? Similarly, if the average time before description for a specimen is 21 years (Fontaine *et al.* 2012), what is the average time for an observation before its first use in a scientific study? Observations recorded in the scope of a particular study are likely used relatively rapidly, but it is possible that some observations, like those recorded by naturalists enthusiasts (e.g. from the ebird network or through iNaturalist), have never been used yet. The “expiration date” mentioned earlier should also be quantified. Escribano *et al.* (2016) have shown that >75% of the biodiversity occurrences of the Navarra region have become obsolete between 1956 and 2012. Are these impressive figures transferable to other geographic regions? GBIF mediated-data, with their large geographic and temporal coverage, enables us initiating studies aiming at quantifying some of the consequences envisioned because of the shift from specimen to observation-based occurrences.

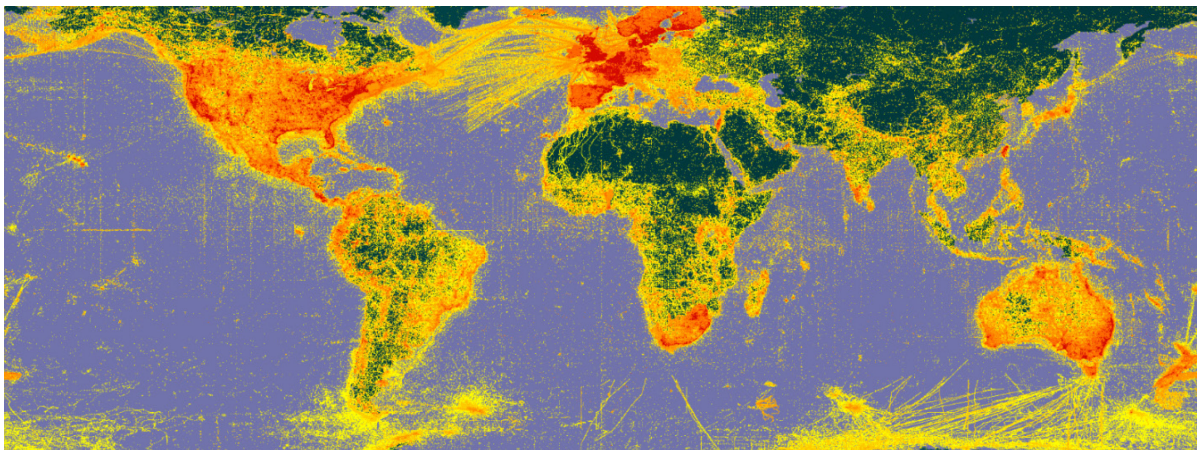
### **Taxonomic bias while aiming at investigating the whole biodiversity**

Biases in science are inevitable and must be limited and acknowledged to reduce distorted conclusions. In the study of biodiversity, spatial and taxonomic biases are the most renown. The spatial bias refers to the fact that some geographical regions are more studied than others. This bias is easily visualized from the GBIF website (Figures 35 and 36), has

been studied elsewhere (Meyer *et al.* 2015, Caesar 2017), and will not be discussed further here. The taxonomic bias refers to the fact that some organisms are more studied than others. It is also known for a long time but remains misunderstood (Gaston and May 1992).



**Figure 35: Proportion of the number of occurrences per region of the world in the GBIF mediated data.** Figure drawn using Gaiji *et al.* (2013) statistics done in December 2010 on 267 million occurrences.

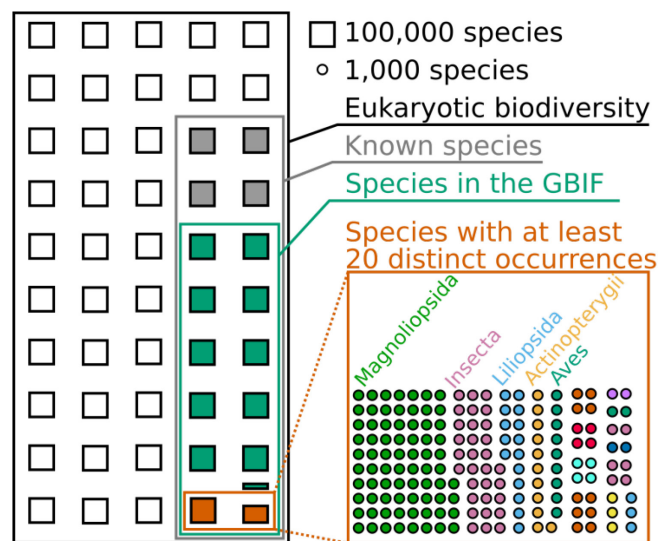


**Figure 36: The global repartition of GBIF-mediated occurrences is uneven.** Warmer colors indicate a higher density of occurrences (www.gbif.org on 18-09-2017).

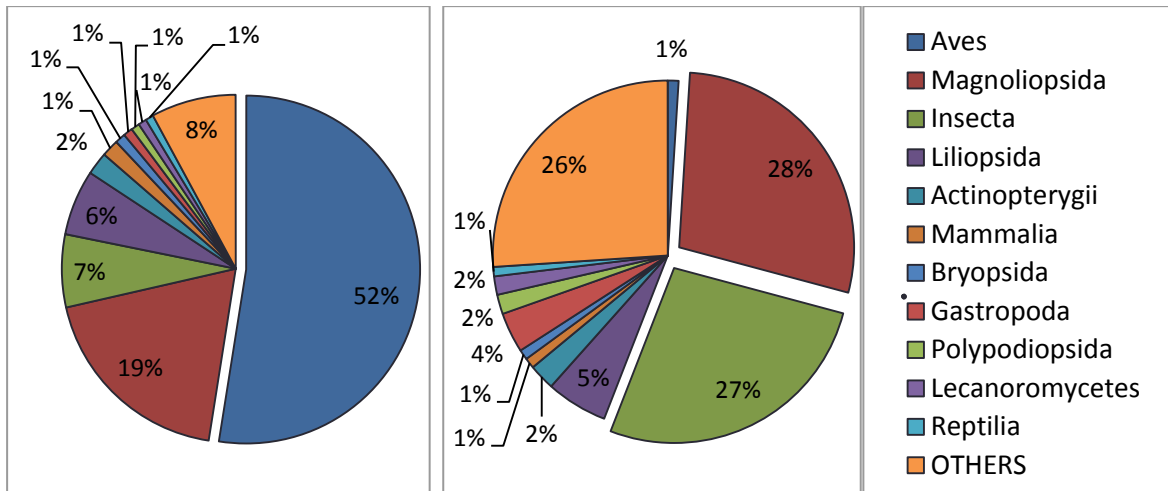


## The taxonomic bias and its societal origin

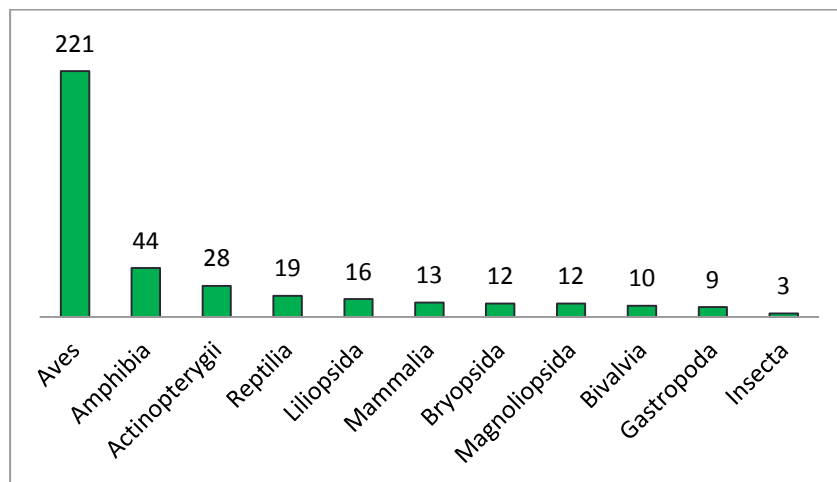
The taxonomic bias is not the preserve of biodiversity studies; it has been shown in ethology (Stahlschmidt 2011), in conservation biology (Donaldson *et al.* 2016), in the distribution of cryptic diversity (Pérez-Ponce de León and Poulin 2016) or in its own study as it mainly focused on vertebrates (e.g. Bonnet *et al.* 2002, Stahlschmidt 2011, McKenzie and Robertson 2015). This bias is often highlighted between higher taxonomic groups (e.g. birds, mammals, invertebrates; Gaston and May 1992, Bonnet *et al.* 2002, Donaldson *et al.* 2016) but is also present at smaller taxonomic scales (chapter 3, McKenzie and Robertson 2015). In the GBIF-mediated data, the enormous difference between the number of occurrences in birds and insects embodies this taxonomic bias (Fig. 38 and 39), which extends much further in terms of taxonomic groups, quality (completeness and accuracy) and origins (specimen or observation) of the occurrences. While most of biodiversity remains unknown (Fig. 37, Larsen *et al.* 2017) we have to keep in mind that the described diversity is not evenly distributed in the tree of life (Fig. 37)



**Figure 37: The current knowledge on eukaryotic species diversity is incomplete and biased.** Most species remain to be described (white square). Colored squares represent described species (grey), species referenced in the GBIF (green) and species with at least 20 spatially distinct occurrences in the GBIF (orange; “decently” sampled species). Details of the “decently” sampled species are provided in the orange rectangle, showing that most of these species belong to a few classes.



**Figure 38: Discrepancy between the proportion of occurrences per class in the GBIF mediated data (left) and the proportion of species per class (right).** The class Aves represents 52 % of the GBIF occurrences while accounting for only 1 % of the species, whereas the class Insecta accounts for only 7 % of the GBIF occurrences and 27 % of the species.



**Figure 39: The median number of occurrences per species in the GBIF-mediated data differs according to taxonomic classes.** Given that, for diverse studies, several occurrences per species are necessary; this graph suggests that some studies cannot be performed for most species in some classes.

This situation is worrying because, in a context of climate change and biodiversity erosion (Díaz *et al.* 2006), i) part of what is unknown will remain so and ii) a global knowledge on ecosystem functioning, mandatory for effective conservation plans, will remain unreachable. Bridging the Linnean shortfall (Brito 2010), the fact that some species have not been described yet, is a first way to produce primary biodiversity data, including on uncharismatic taxa such as invertebrates (Cardoso *et al.* 2011), but it requires a higher manpower in systematics (May 2004). As for effective conservation plans, some conservationists are already advocating for concentrating and increasing efforts on charismatic taxa (e.g. Pérez-Ponce de León and Poulin 2016; Ripple *et al.* 2016). This practice neglects, however, least studied taxa that yet play crucial roles in ecosystems (Ford *et al.* 2017).

Multiple hypotheses have been proposed to explain the formation and maintenance of a taxonomic bias and they can be categorized into “internal” and “external” causes, whether they are properties of the taxa or not, respectively. I focused on two external hypotheses, the amount of scientific research and societal influence. The influence of scientists and the public are not easily untangled and their role was shown in funding conservation programs and decision-making about conservation policies (Martín-López *et al.* 2007, 2009). My results emphasized particularly the role of societal preferences but should be explored further.

The main difficulty to identify causes of taxonomic bias is that potential factors are intertwined and a given factor acts at many levels. Public influence, for instance, impacts political choices, and politicians orientate research funding policies (Martín-López *et al.* 2009). Politicians as well as scientists may have preferences for some taxa for personal, and not scientific, reasons. Finally, the preferences of the public are also those of future scientists, and often the desire to exercise a scientific profession linked to natural history is fuelled by a previous interest in a particular taxon (Leather 2009). In this regard, my conclusions about the impact of societal preferences on taxonomic bias should be taken with cautions.

To better analyze societal influence, a public poll questioning why some taxa are more attractive than others might be useful. Martín-López *et al.* (2007) demonstrated that the public was more likely to pay for the conservation of organisms judged closer to humans (anthropomorphism) or considered useful (anthropocentrism). But this hardly explains why

birds have much more data than mammals. There are therefore potentially other hypotheses to explore, such as the presence in the media (films, novels, art...), economic interest, geographical proximity, and so on, that could be identified through large public polls, and hopefully deciphered.

The causes behind taxonomic bias must be understood before planning research policies to counteract or reduce taxonomic bias. Making insects more anthropomorphic is impossible, but their anthropocentric value can be enhanced through scientific programs explaining the role of insects in ecosystem (Cardoso *et al.* 2011). Also, reinstating or developing courses about sidelined taxa could get the future generation of scientists more interested into those taxa (Balmford *et al.* 2002, Leather 2009). Correcting the taxonomic bias will undoubtedly require tremendous efforts and must scientists must tackle this task head on. Along the way, scientists should not ignore the manpower represents the public. Thus, citizen science might be a remarkable option in this respect, as it can raise the public interest for and increase our knowledge on a given taxa.

### ***Further exploring the taxonomic bias***

Besides “external” causes, “internal” factors contribute to taxonomic bias. Internal factors are based on taxon characteristics that may favor or limit its study, and their role is uncontroversial (Cardoso *et al.* 2011). Unfortunately, studying internal factors at a large taxonomic scale is very time-consuming, a task I could not carry out during my PhD project. I distinguished two main categories of internal factors, those altering the detectability of a specimen and those complicating its identification. Specimen detection depends on several factors such as population density or habitats, whereas identification is mainly correlated to the size of organisms and the existence of cryptic species. Some of these factors are listed in table 7 and briefly developed below.

**Table 7: Summary of the effect of different “internal” taxon characteristics.** Each line represents a characteristic. A green cell indicates a potential positive effect, a red one a potential negative effect. ‘NA’ stands for mixed or indistinct effect.

	Detectability		Ease of Identification
	Encounter probability	Ease of Observation	
Body size increase	-	+	+
Reachable habitat	+	NA	NA
High abundance	+	NA	NA
Discrete behavior	NA	-	NA
Large range	+	NA	NA
Diurnal taxon	+	+	NA
Species similarity	NA	-	-
High Speciosity	NA	NA	-

Body size is a feature that impacts the detectability of a taxon. Large taxa are easier to detect and observed. However, large animals often have low abundance (Robinson and Redford 1986) and may be quite sensitive to disturbance (Blumstein 2006), which reduces the probability to encounter them. In addition, larger specimens are more difficult to harvest due to logistical constraints.

The localization of a specimen, both its geographic origin and its habitat, influences its detectability. Species living in marine or hostile environment, as well as species with narrow distribution (e.g. endemic species) are more complicated to collect than other species. Besides, taxa highly sensitive to human disturbance live in remote places, located farther from urban centers, which also reduces the probability of encounter (Meyer *et al.* 2015).

The abundance and discretion of a taxon are two variables that also affect its detectability. Evidently, taxa with high abundances are encountered more often than taxa with low abundance and discrete taxa (e.g. camouflage, mimicry, vibrational communication...) are more difficult to observe than conspicuous taxa. Other behaviors such as anti-predator behaviors (e.g. freezing *vs.* escape behavior) or circadian activity (diurnal *vs.* nocturnal; Burton 2012) impact the detectability of specimens.

The difficulty of identification is affected by fewer variables and mainly depends on the size of the organisms, the degree of similarity with other species (mimicry, cryptic species) and the number of species described in a taxonomic group.

All these internal factors may strengthen or counteract the impact of external factors on taxonomic bias. They could help understanding why taxonomic bias is not fully explained by the societal hypothesis for instance. Mammals are less sampled than birds (Table 8, chapter 3) despite their higher anthropomorphism value, a result easily explained considering that birds are more easily observed than mammals (Law and Lynch 1988).

**Table 8: Top 10 list of species with the most occurrences in the GBIF.** All of the 100 most sampled species in the GBIF are birds belonging to diverse families (data not shown); the 100th species has 741 thousands occurrences.

Species	Family	Class	Millions of occurrences
<i>Sturnus vulgaris</i>	Sturnidae	Aves	4.57
<i>Anas platyrhynchos</i>	Anatidae	Aves	4.54
<i>Zenaida macroura</i>	Columbidae	Aves	4.41
<i>Corvus brachyrhynchos</i>	Corvidae	Aves	4.13
<i>Turdus migratorius</i>	Turdidae	Aves	4.07
<i>Cardinalis cardinalis</i>	Cardinalidae	Aves	3.96
<i>Passer domesticus</i>	Passeridae	Aves	3.57
<i>Branta canadensis</i>	Anatidae	Aves	3.49
<i>Cyanocitta cristata</i>	Corvidae	Aves	3.38
<i>Spinus tristis</i>	Fringillidae	Aves	3.25

Finally, even if I worked on 24 classes, I neglected numerous taxa, especially unicellular organisms. Only taxa with sufficient occurrences were kept, which excludes a huge part of the living world and, consequently, the case of unicellular organisms was only brushed on. Among the 24 studied classes, only a few consisted of unicellular organisms. Microbes are among the least known groups of organisms (Larsen *et al.* 2017) and the taxonomic bias against this organisms is colossal, whatever the estimates on microscopic species richness (Larsen *et al.* 2017). Most of those organisms are symbionts of macroscopic species (gut biotas) whose ignorance or extinction leads to a lack of knowledge or extinction

of multiple microscopic species (Larsen *et al.* 2017). This statement adds to the urgency of understanding and correcting taxonomic bias to really embrace the whole biodiversity.

## **Using biodiversity data to decipher the origin of global biodiversity patterns**

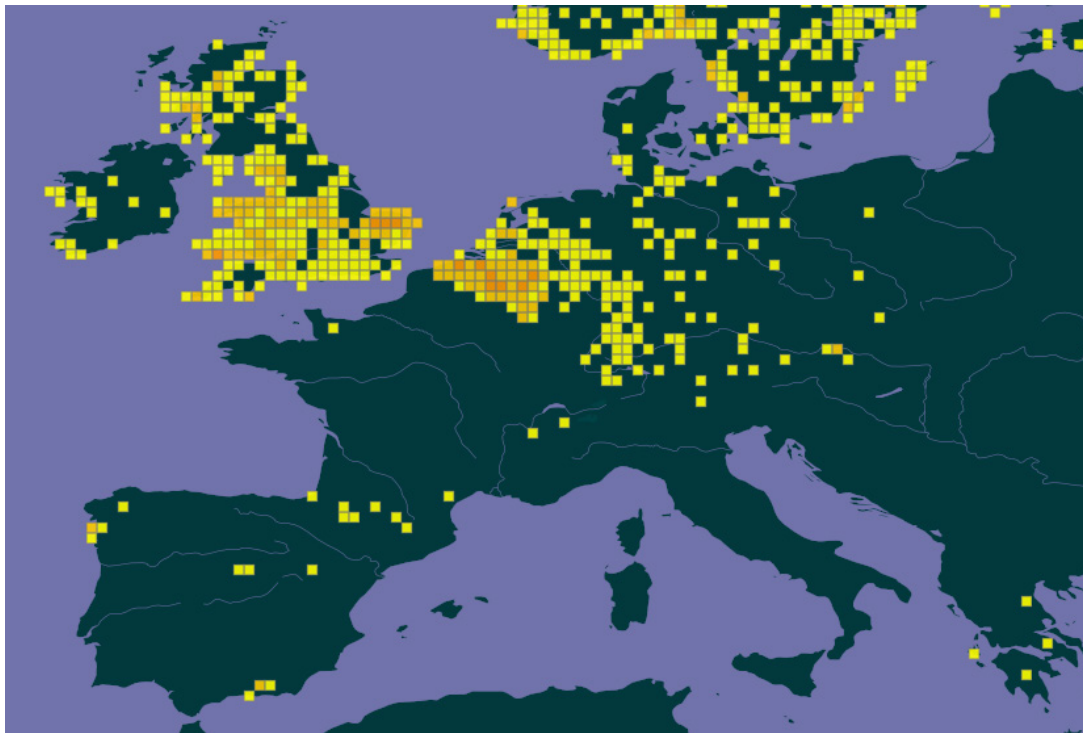
The outcomes of the data mining conducted in a first step provided insights about how GBIF-mediated biodiversity occurrences have been collated and which occurrences can be used, in a second step, to investigate what is arguably the most fascinating pattern of biodiversity: the Latitudinal Diversity Gradient (LDG). The LDG refers to the fact that species richness increases when latitude decreases. Despite decades of interest, the LDG still puzzles ecologists and a few tens of hypotheses have been formulated to explain its origin.

These hypotheses can be sorted in three broad categories: *in situ* hypotheses that focus on the capacity of the local environment to support a certain amount of species (carrying capacity), historical hypotheses based on spatial and evolutionary processes, and geometric hypotheses that rely on mathematical effects. All authors agree that these different hypotheses are non-exclusive and that all of them may have contributed to the LDG. The real challenge is to disentangle their relative contribution. Results obtained so far are mitigated and appear to depend on the taxonomic group and their geographic origin.

Following a recent contribution on the geometric hypothesis (Gross and Snyder-Beattie, 2016), I re-investigated the LDG in the New World using eight classes of plants and animals. Together with the revisited formulation of the geometric hypothesis, this large taxonomic and geographic coverage is an original strength of this study because the LDG has been mainly investigated at narrower taxonomic and geographic scales. The results underline the importance of considering spatial autocorrelation of and non-stationarity in the environmental variables. They also confirm the role of the productivity hypothesis (*in situ* hypothesis) in the LDG formation (Hawkins *et al.* 2003b, Gillman *et al.* 2015, Duffy *et al.* 2017), while the roles of the geometric hypothesis (Colwell and Hurtt 1994), of water availability (Hawkins *et al.* 2003b) and of the Rapoport effect (Stevens 1989) are down-weighted.

## **Estimating global species richness from a large and geographically widespread taxonomic sample**

The biggest assets of using GBIF-mediated data are the wealth of data and their large-scale coverage. But they correlate with its main liability, the difficulty of getting evenly spread data geographically and of performing detailed analyses for multiple groups of organisms. Occurrences distribution is irregular and some regions where a given species is present are not recorded as (Fig. 40). Consequently, estimating species richness of vast areas, the basic information required to dwell on the LDG, is particularly challenging and solutions to counterbalance these limitations must be looked for.



**Figure 40: Repartition of the Yellow Ant (*Lasius flavus*) occurrences across Europe in the GBIF dataset.** Data scattering is striking with some areas less sampled than others. The distribution of *Lasius flavus* encompasses France, where it is a common species, but this species have rarely been sampled and shared through the GBIF (gbif.org 06-09-2017).

I first planned to use non-parametric estimators, especially those developed by Chao and collaborators because they seem the most used and robust (e.g. Burnham and Overton 1979, Chao 1984, Chao and Bunge 2002). Those estimators mostly rely on species accumulation curves (Gotelli and Colwell 2001) and use the number of occurrences per species in a given region to estimate the total species richness of the region. Those estimators



are, however, very sensitive to singletons, i.e. species that were sampled only once, a common situation for GBIF-mediated data. Thus, this solution was not kept because it would have largely overestimated species richness.

Resorting to subsampling procedures was also discarded. In subsampling analyses, a subset of the occurrences is used and adjusted to get the same occurrences density across the targeted region. Nevertheless, many species and areas were too poorly sampled to conduct subsampling procedures. As for keeping only well-sampled species, it would have discarded the vast majority of the GBIF data – nullifying a strength of GBIF-mediated data – and species richness estimates would have been misleading.

Niche modeling, which aims at estimating the potential distribution of a species, is a common method to compensate for biased geographic sampling. This method uses environmental variables and species occurrences to compute which other area are compatible with the species preferences. Many algorithms are available, some specifically tuned for a taxon or a peculiar type of data (e.g. Peterson *et al.* 2002, McNyset 2005, Lisòn and Calvo 2013). Here, the model used was the SRE model as implemented in biomod2 (Thuiller *et al.* 2009), which is a very simple model and was only used within the convex hull delimited from species occurrences. We used a commonly accepted threshold of at least 20 occurrences for modeling a species (Feeley and Silman 2010), even though it remains a rule of thumb. Yet, almost 600,000 species listed in the GBIF have fewer than 20 occurrences. So this method is not perfect and alternatives must be looked for in the future.

Conscious of the limitations of the SRE modeling combined with the conservative convex hull approach, I tested whether this methodology was robust enough to be used. Unfortunately, testing how the modeled distribution fits with our knowledge for all species was impossible because of time constraints and of our partial knowledge of most species distribution. Still, I compared modeled some niches with the distribution shown from the IUCN Red List data (IUCN Species Survival Commission 2001) but it reveals inconclusive and, although overlapping, niches for several species differ. IUCN and GBIF-mediated data are imperfect and some flaws cannot be detected through automatic data quality check procedures but only manually (area stopping at borders in the IUCN; fossil occurrences referenced as observations in the GBIF...).

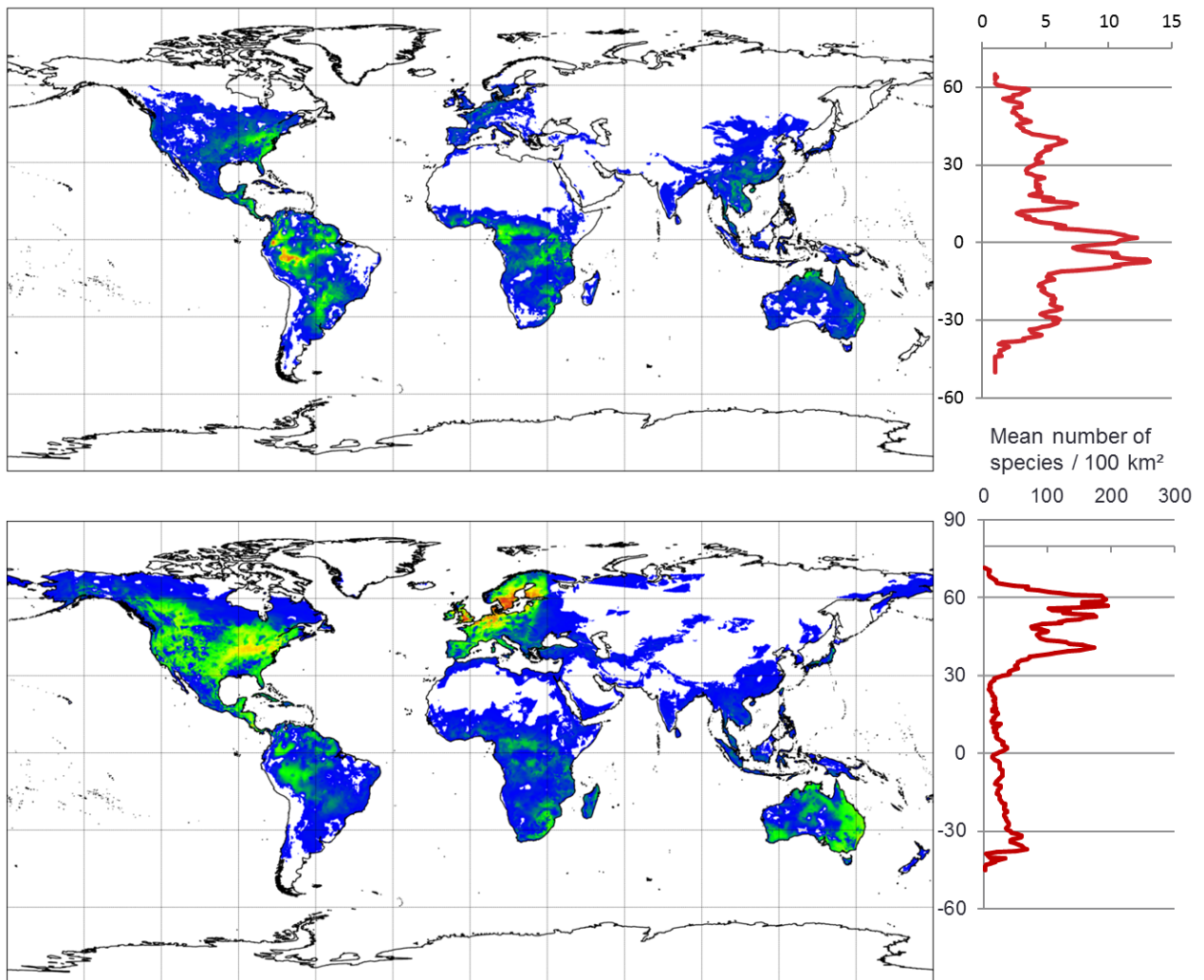
## Further into the Latitudinal Diversity Gradient

To better understand the LDG, geometric and in situ hypotheses were considered, but no historical hypotheses. Integrating a historical dimension in future analyses will be important as several recent studies emphasized its potential role in species richness distribution of different organisms (e.g. Weir and Schluter 2007, Brown 2014, Machac and Graham 2016). A large taxonomic coverage to test a few hypotheses was favored over a smaller taxonomic coverage to test more hypotheses.

The LDG was studied here in the New World but future studies will also have to focus on other geographic regions. Some areas are better candidates than others because they have been more densely and evenly sampled. South Africa and Australia, for instance, have a large number of data (Fig. 36, Meyer *et al.* 2015) and could provide new insights, if not on the LDG because they do not encompass so many latitudes, at least on the origin of species distribution. Similarly, other taxa could be investigated as long as they are estimated as sufficiently sampled in a given region. Condamine *et al.* (2012) have, for example, studied LDG at a global geographic scale for a family of butterflies. In Chapter 3, we found that within classes, some orders are better sampled than others. Even in insects, the most-underrepresented class, odonates have more data than other Arthropods and could be used to investigate the LDG (Pearson and Boreyo 2009).

The main difficulty, as already underlined, is to identify geographic and taxonomic bias too pronounced to be counterbalanced with statistical analyses. Establishing a threshold between taxa to keep or remove in macroecological analyses is difficult, especially because geographic and taxonomic biases work together and, obviously, the raw number of occurrences for a taxonomic group does not help establishing this threshold. Although Chapter 4 focuses on the New World, I ran some analyses at a worldwide scale and produced several species richness maps and plots as in Figures 41 and 42. These maps – several of which are available in Appendix 6 – could help identifying knowledge gaps due to sampling

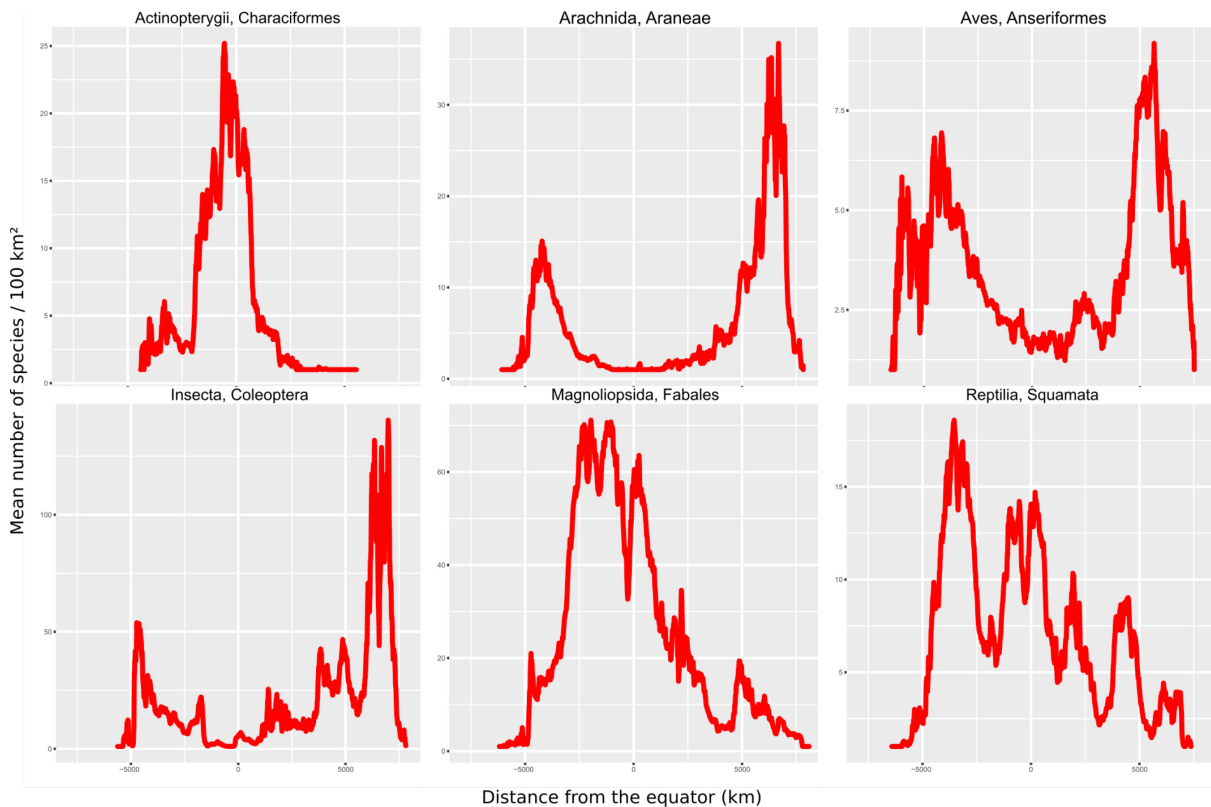
biases.



**Figure 41: Different effects of the spatial bias depending on the taxa.** Top) Amphibia species richness map and plots computed using 2 million occurrences from 1 183 species with 20 or more distinct occurrences; bottom) Insecta species richness map and plots computed using 34 million occurrences from 26 273 species with 20 or more distinct occurrences. On the map the hotter colors represents higher species richness. The plots show the mean number of species per 100 km<sup>2</sup> per latitude and have different scales. As far as we know, both groups should exhibit a latitudinal diversity gradient, but it is not the case for insects due to obvious geographic sampling bias.

Finally, for taxa with no detrimental bias, the existence of a LDG could be systematically checked. Even though the LDG is a global pattern, some taxa showing another

geographic pattern are known (Willig et al. 2003). How many taxa are in this case? Do the taxa belong to a few taxonomic groups or are they widely spread across living beings? How can we explain these alternative patterns? I initiated this work, using GBIF data and niche modeling as described above, for more than 400 orders. I produced graphs showing the specific richness according to the distance to the equator (Fig. 42). This approach, illustrating the exploratory approach of big-data (Kelling *et al.* 2009), could allow discovering taxa with particular evolutionary histories or very unusual responses to ecological constraints.



**Figure 42: Species richness patterns derived from the GBIF-mediated for 6 orders.**

These patterns might reflect biases or represent the actual diversity but sorting them out is not always easy. The Characiformes and Fabales show a typical LDG pattern while others are very different. Arachnida, Anseriformes and Coleoptera show a higher species richness in temperate latitude, which undoubtedly results from a sampling bias for the Arthropods. But Anseriformes, one of the most sampled Aves orders, might legitimately display an unusual latitudinal diversity gradient, while Squamata, with decreasing species richness as we go north, appears at odds with common biodiversity patterns.

## **The GBIF-mediated data: a fascinating tool for biodiversity analyses**

The GBIF-mediated data were used to investigate the LDG but they can be used for countless studies. Because of analogies with LDG, the first type of study that comes to mind deals with altitudinal gradients. The most common altitudinal gradient is very similar to the LDG as it implies a higher specific richness at mid-elevation than high or low elevations. It would be very simple to adapt the workflow developed during my PhD in order to study this pattern. Better yet, the model proposed by Gross and Snyder-Beattie (2016) includes a version compatible with the altitudinal diversity gradient. It is therefore possible to test this model, with alternatives hypotheses, in an altitudinal context. It should be noted, however, that only a portion of the GBIF data was collected in the mountains and that even fewer of these data have accurate information on harvest altitude. Hence stringent data filtering and cleaning should be performed before.

Geographical gradients could also be investigated using other variables than species richness. For example, Miraldo *et al.* 2016 focused on genetic diversity and they reported a latitudinal gradient at a global scale. Using primary biodiversity occurrences, phylogenetic diversity (Faith 1992) or functional diversity (Lamanna *et al.* 2014), for instance, could be investigated with regard to latitude.

Species concomitances could also be tested from the GBIF-mediated data. Because of the large taxonomic coverage, it is possible to identify species that tend to have the same ecological niches, and thus are concomitant (e.g. as in host/parasite relationships), or that tend to avoid one another (e.g. as in competition relationships). It would potentially enable us to identify biotic relationships between species (MacKenzie *et al.* 2005) at a large geographic scale. Conversely, known interactions between species could be used to find sampling gaps. Plant specific richness is, for instance, a good indicator of insect specific richness (Siemann *et al.* 1998, Lewinsohn and Roslin 2008, Dinnage *et al.* 2012). Since plants taxa are much better sampled than insects in general, it may be possible to compare the specific richness of insects with the one "predicted" from the specific richness of plants. This would contribute to identify areas to promptly sample.

The GBIF-mediated occurrences are much more than an accumulation of biodiversity occurrences. They constitute a colossal scientific heritage that brings together the work of

millions of people, researchers, amateurs and citizens. The GBIF-mediated data is a great tool to take the pulse of the study of biodiversity. The dramatic increase in the amount of data available in GBIF is a proof of its success and appeal: between the beginning of my thesis and the writing of these concluding lines, almost 300 million data have been shared through the GBIF portal. Nevertheless, this impressive wealth of data and the speed at which it grows should not make us lose sight of the ultimate goal of a better understanding of biodiversity as a whole. This accumulation of data must occur in the best possible conditions to ensure the sustainability and high quality of each occurrence. I hope that this work will help better understanding biodiversity occurrences and ensure their high usefulness now and in the future.

## References

- Albuquerque F., Beier P. 2015. **Global patterns and environmental correlates of high-priority conservation areas for vertebrates.** *J. Biogeogr.*:n/a-n/a.
- Allen A.P., Gillooly J.F. 2006. **Assessing latitudinal gradients in speciation rates and biodiversity at the global scale.** *Ecology Letters*. 9:947–954.
- Andam C.P., Doroghazi J.R., Campbell A.N., Kelly P.J., Choudoir M.J., Buckley D.H. 2016. **A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus *Streptomyces*.** *mBio*. 7.
- Andlinger P. 2013. **RDBMS dominate the database market, but NoSQL systems are catching up.** Available from [https://db-engines.com/en/blog\\_post/23](https://db-engines.com/en/blog_post/23).
- Anmarkrud J.A., Lifjeld J.T. 2017. **Complete mitochondrial genomes of eleven extinct or possibly extinct bird species.** *Molecular Ecology Resources*. 17:334–341.
- Anselin L., Bera A.K., Florax R., Yoon M.J. 1996. **Simple diagnostic tests for spatial dependence.** *Regional Science and Urban Economics*. 26:77–104.
- Ariño A.H. 2010. **Approaches to estimating the universe of natural history collections data.** *Biodiversity Informatics*. 7.
- Auguie B., 2017. **gridExtra: Miscellaneous functions for “grid” graphics.** *Cran.r-project.org*. at <http://CRAN.R-project.org/package=gridExtra>, package version 2.2.1
- Balmford A., Clegg L., Coulson T., Taylor J. 2002. **Why Conservationists Should Heed Pokémon.** *Science*. 295:2367–2367.
- Barnosky A.D., Matzke N., Tomiya S., Wogan G.O.U., Swartz B., Quental T.B., Marshall C., McGuire J.L., Lindsey E.L., Maguire K.C., Mersey B., Ferrer E.A. 2011. **Has the Earth’s sixth mass extinction already arrived?** *Nature*. 471:51–57.
- Beaman R.S., Cellinese N. 2012. **Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science.** *Zookeys*. 7–17.
- Beck J., Ballesteros-Mejia L., Buchmann C.M., Dengler J., Fritz S.A., Gruber B., Hof C., Jansen F., Knapp S., Kreft H., Schneider A.-K., Winter M., Dormann C.F. 2012. **What’s on the horizon for macroecology?** *Ecography*. 35:673–683.
- Beck J., 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 Distrib.* 19:1043–1050.
- Belmaker J., Jetz W. 2015. **Relative roles of ecological and energetic constraints, diversification rates and region history on global species richness gradients.** *Ecol Lett.*:n/a-n/a.
- Benjamin D.J., Berger J.O., Johannesson M., Nosek B.A., Wagenmakers E.-J., Berk R., Bollen K.A., Brems B., Brown L., Camerer C., Cesarini D., Chambers C.D., Clyde M., Cook T.D., De Boeck P., Dienes Z., Dreber A., Easwaran K., Efferson C., Fehr E., Fidler F., Field A.P., Forster M., George E.I., Gonzalez R., Goodman S., Green E., Green D.P., Greenwald A.G., Hadfield J.D., Hedges L.V., Held L., Hua Ho T., Hoijtink H., Hruschka D.J., Imai K., Imbens G., Ioannidis J.P.A., Jeon M., Jones J.H., Kirchler M., Laibson D., List J., Little R., Lupia A., Machery E., Maxwell S.E., McCarthy M., Moore D.A., Morgan S.L., Munafó M., Nakagawa S., Nyhan B., Parker T.H., Pericchi L., Perugini M., Rouder J., Rousseau J., Savalei V., Schönbrodt F.D., Sellke T., Sinclair B., Tingley D., Van Zandt T.,

- Vazire S., Watts D.J., Winship C., Wolpert R.L., Xie Y., Young C., Zinman J., Johnson V.E. 2017. **Redefine statistical significance.**
- Bentley J.L. 1975. **Multidimensional Binary Search Trees Used for Associative Searching.** Commun. ACM. 18:509–517.
  - Bingham H., Weatherdon L., Despot-Belmonte K., Wetzel F., Martin C. 2017. **The Biodiversity Informatics Landscape: Elements, Connections and Opportunities.** Research Ideas and Outcomes. 3:e14059.
  - Bisby F.A. 2000. **The Quiet Revolution: Biodiversity Informatics and the Internet.** Science. 289:2309–2312.
  - Bivand R., Keitt T., Rowlingson B., Pebesma E., Sumner M., Hijmans R., Rouault E. 2017a. **Bindings for the “Geospatial” Data Abstraction Library.**
  - Bivand R., Piras G. 2015. **Comparing Implementations of Estimation Methods for Spatial Econometrics.** 63.
  - Bivand R., Yu D., Nakaya T., Garcia-Lopez M.-A. 2017b. Package ‘spgwr’ Geographically Weighted Regression. Available from <https://cran.r-project.org/web/packages/spgwr/spgwr.pdf>.
  - Blagoderov V., Kitching I.J., Livermore L., Simonsen T.J., Smith V.S. 2012. **No specimen left behind: industrial scale digitization of natural history collections.** Zookeys.:133–146.
  - Blumstein D.T. 2006. **Developing an evolutionary ecology of fear: how life history and natural history traits affect disturbance tolerance in birds.** Animal Behaviour. 71:389–399.
  - Boakes E.H., McGowan P.J.K., 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 Biol. 8:e1000385.
  - Bonnet X., Shine R., Lourdaïs O. 2002. **Taxonomic chauvinism.** Trends in Ecology and Evolution. 17:1–3.
  - Borchsenius, F. 2012. **GBIF and IAVD.** Presented at International Arctic Vegetation Database Workshop. Roskilde, Denmark May 29-31. (Presentation)
  - Bortolus A. 2008. **Error Cascades in the Biological Sciences: The Unwanted Consequences of Using Bad Taxonomy in Ecology.** AMBIO: A Journal of the Human Environment. 37:114–118.
  - Boucher-Lalonde V., Currie D.J. 2016. **Spatial Autocorrelation Can Generate Stronger Correlations between Range Size and Climatic Niches Than the Biological Signal — A Demonstration Using Bird and Mammal Range Maps.** PLOS ONE. 11:e0166243.
  - Boyd D., Crawford K. 2012. **Critical Questions for Big Data. Information, Communication and Society.** 15:662–679.
  - Breunig M., Kriegel H.-P., Ng R.T., Sander J. 2000. LOF: Identifying Density-Based Local Outliers. **PROCEEDINGS OF THE 2000 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA.**:93–104.
  - Brito D. 2010. **Overcoming the Linnean shortfall: Data deficiency and biological survey priorities.** Basic and Applied Ecology. 11:709–713.
  - Brooks T.M., Mittermeier R.A., Mittermeier C.G., Da Fonseca G.A.B., Rylands A.B., Konstant W.R., Flick P., Pilgrim J., Oldfield S., Magin G., Hilton-Taylor C. 2002. **Habitat Loss and Extinction in the Hotspots of Biodiversity.** Conservation Biology. 16:909–923.
  - Brown J.H. 1995. **Macroecology.** University of Chicago Press.



- Brown J.H. 2014. **Why are there so many species in the tropics?** *J. Biogeogr.* 41:8–22.
- Brown J.H., Lomolino M.V. 1998. **Biogeography.** Sinauer Associates.
- Brunsdon C., Fotheringham A.S., Charlton M.E. 1996. **Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity.** *Geographical Analysis.* 28:281–298.
- Buerki S., Baker W.J. 2016. **Collections-based research in the genomic era.** *Biol. J. Linn. Soc.* 117:5–10.
- Buerki S., Callmander M.W., Bachman S., Moat J., Labat J.-N., Forest F. 2015. **Incorporating evolutionary history into conservation planning in biodiversity hotspots.** *Phil. Trans. R. Soc. B.* 370:20140014.
- Bunge J., Fitzpatrick M. 1993. **Estimating the Number of Species: A Review.** *Journal of the American Statistical Association.* 88:364–373.
- Burnham K.P., Overton W.S. 1978. **Estimation of the Size of a Closed Population when Capture Probabilities vary Among Animals.** *Biometrika.* 65:625–633.
- Burnham K.P., Overton W.S. 1979. **Robust Estimation of Population Size When Capture Probabilities Vary Among Animals.** *Ecology.* 60:927–936.
- Burton A.C. 2012. **Critical evaluation of a long-term, locally-based wildlife monitoring program in West Africa.** *Biodivers Conserv.* 21:3079–3094.
- Caesar M., Grandcolas P., Pellens R. 2017. **Outstanding micro-endemism in New Caledonia: More than one out of ten animal species have a very restricted distribution range.** *PLOS ONE.* 12:e0181437.
- Cai L., Zhu Y. 2015. **The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.** *Data Science Journal.* 14.
- Cameron E.K., Decaëns T., Lapiéd E., Porco D., Eisenhauer N. 2016. **Earthworm databases and ecological theory: Synthesis of current initiatives and main research directions.** *Applied Soil Ecology.* 104:85–90.
- Canhos V.P., Souza S. de, Giovanni R.D., Canhos D.A.L. 2004. **Global Biodiversity Informatics: setting the scene for a “new world” of ecological forecasting.** *Biodiversity Informatics.* 1.
- Cardillo M. 2011. **Phylogenetic structure of mammal assemblages at large geographical scales: linking phylogenetic community ecology with macroecology.** *Philosophical Transactions of the Royal Society of London B: Biological Sciences.* 366:2545–2553.
- Cardinale B.J., Duffy J.E., Gonzalez A., Hooper D.U., Perrings C., Venail P., Narwani A., Mace G.M., Tilman D., Wardle D.A., Kinzig A.P., Daily G.C., Loreau M., Grace J.B., Larigauderie A., Srivastava D.S., Naeem S. 2012. Biodiversity loss and its impact on humanity. *Nature.* 486:59–67.
- Cardoso P., Erwin T.L., Borges P.A.V., New T.R. 2011. **The seven impediments in invertebrate conservation and how to overcome them.** *Biological Conservation.* 144:2647–2655.
- Ceballos G., Ehrlich P.R., Barnosky A.D., García A., Pringle R.M., Palmer T.M. 2015. **Accelerated modern human-induced species losses: Entering the sixth mass extinction.** *Science Advances.* 1:e1400253.
- Chandler M., See L., Copas K., Bonde A.M.Z., López B.C., Danielsen F., Legind J.K., Masinde S., Miller-Rushing A.J., Newman G., Rosemartin A., Turak E. 2016. **Contribution of citizen science towards international biodiversity monitoring.** *Biological Conservation.*

- Chandler M., See L., Copas K., Bonde A.M.Z., López B.C., Danielsen F., Legind J.K., Masinde S., Miller-Rushing A.J., Newman G., Rosemartin A., Turak E. 2017. **Contribution of citizen science towards international biodiversity monitoring.** *Biological Conservation*. 213:280–294.
- Chao A. 1984. **Nonparametric Estimation of the Number of Classes in a Population.** *Scandinavian Journal of Statistics*. 11:265–270.
- Chao A., Bunge J. 2002. **Estimating the Number of Species in a Stochastic Abundance Model.** *Biometrics*. 58:531–539.
- Charbonnier Y.M., Barbaro L., Barnagaud J.-Y., Ampoorter E., Nezan J., Verheyen K., Jactel H. 2016. **Bat and bird diversity along independent gradients of latitude and tree composition in European forests.** *Oecologia*:1–9.
- Charlton M., Fotheringham S. 2009. **GEOGRAPHICALLY WEIGHTED REGRESSION WHITE PAPER.** .
- Chase J.M., Knight T.M. 2013. **Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough.** *Ecol Lett*. 16:17–26.
- Chaudhary C., Saeedi H., Costello M.J. 2016. **Bimodality of Latitudinal Gradients in Marine Species Richness.** *Trends in Ecology and Evolution*.
- Chown S.L., Gaston K.J. 2000. **Areas, cradles and museums: the latitudinal gradient in species richness.** *Trends in Ecology and Evolution*. 15:311–315.
- Clark J.A., May R.M. 2002. **Taxonomic Bias in Conservation Research.** *Science*. 297:191–192.
- Clark J.A., May R.M. 2002. **How biased are we?: Even now, conservation research is still lopsided.** *Conserv. Practice* 3(3), 28–29
- Colwell R.K., Hurtt G.C. 1994. **Nonbiological Gradients in Species Richness and a Spurious Rapoport Effect.** *The American Naturalist*. 144:570–595.
- Colwell R.K., Lees D.C. 2000. **The mid-domain effect: geometric constraints on the geography of species richness.** *Trends in Ecology and Evolution*. 15:70–76.
- Condamine F.L., Sperling F.A.H., Wahlberg N., Rasplus J.-Y., Kergoat G.J. 2012. **What causes latitudinal gradients in species diversity? Evolutionary processes and ecological constraints on swallowtail biodiversity.** *Ecology Letters*. 15:267–277.
- Costello M.J., May R.M., Stork N.E. 2013. **Can We Name Earth's Species Before They Go Extinct?** *Science*. 339:413–416.
- Cotterill F.P.D., Al-Rasheid K., Foissner W. 2008. **Conservation of protists: is it needed at all?** *Biodivers Conserv*. 17:427–443.
- Currie D.J. 1991. **Energy and Large-Scale Patterns of Animal- and Plant-Species Richness.** *The American Naturalist*. 137:27–49.
- Currie D.J., David J., Gary G. Mittelbach, Howard V. Cornell, Richard Field, Jean-Francois Guégan, Bradford A. Hawkins, Dawn M. Kaufman, et al. (2004) **Predictions and Tests of Climate-Based Hypotheses of Broad-Scale Variation in Taxonomic Richness.** *Ecology Letters* 7, 12:1121–34.
- Currie D.J., Kerr J.T. 2007. **Testing, as opposed to supporting, the Mid-domain Hypothesis: a response to Lees and Colwell (2007).** *Ecology Letters*. 10:E9–E10.
- Currie D.J., Kerr J.T. 2008. **Tests of the Mid-Domain Hypothesis: A Review of the Evidence.** *Ecological Monographs*. 78:3–18.
- Dasmann R.F. 1968. **A Different Kind of Country.** Collier Books.

- De Maesschalck R., Jouan-Rimbaud D., Massart D.L. 2000. **The Mahalanobis distance.** *Chemometrics and Intelligent Laboratory Systems.* 50:1–18.
- Deans A.R., Lewis S.E., Huala E., Anzaldo S.S., Ashburner M., Balhoff J.P., Blackburn D.C., Blake J.A., Burleigh J.G., Chanet B., Cooper L.D., Courtot M., Csösz S., Cui H., Dahdul W., Das S., Dececchi T.A., Dettai A., Diogo R., Druzinsky R.E., Dumontier M., Franz N.M., Friedrich F., Gkoutos G.V., Haendel M., Harmon L.J., Hayamizu T.F., He Y., Hines H.M., Ibrahim N., Jackson L.M., Jaiswal P., James-Zorn C., Köhler S., Lecointre G., Lapp H., Lawrence C.J., Novère N.L., Lundberg J.G., Macklin J., Mast A.R., Midford P.E., Mikó I., Mungall C.J., Oellrich A., Osumi-Sutherland D., Parkinson H., Ramírez M.J., Richter S., Robinson P.N., Ruttenberg A., Schulz K.S., Segerdell E., Seltmann K.C., Sharkey M.J., Smith A.D., Smith B., Specht C.D., Squires R.B., Thacker R.W., Thessen A., Fernandez-Triana J., Vihinen M., Vize P.D., Vogt L., Wall C.E., Walls R.L., Westerfeld M., Wharton R.A., Wirkner C.S., Woolley J.B., Yoder M.J., Zorn A.M., Mabee P. 2015. **Finding Our Way through Phenotypes.** *PLOS Biology.* 13:e1002033.
- Deans A.R., Yoder M.J., Balhoff J.P. 2012. **Time to change how we describe biodiversity.** *Trends in Ecology and Evolution.* 27:78–84.
- Deck J., Gaither M.R., Ewing R., Bird C.E., Davies N., Meyer C., Riginos C., Toonen R.J., Crandall E.D. 2017. **The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples.** *PLOS Biology.* 15:e2002925.
- Devictor V., Bensaude-Vincent B. 2016. **From ecological records to big data: the invention of global biodiversity.** *HPLS.* 38:13.
- Devictor V., Whittaker R.J., Beltrame C. 2010. **Beyond scarcity: citizen science programmes as useful tools for conservation biogeography.** *Diversity and Distributions.* 16:354–362.
- Di Marco M., Chapman S., Althor G., Kearney S., Besancon C., Butt N., Maina J.M., Possingham H.P., Rogalla von Bieberstein K., Venter O., Watson J.E.M. 2017. **Changing trends and persisting biases in three decades of conservation science.** *Global Ecology and Conservation.* 10:32–42.
- Díaz S., Fargione J., Ili F.S.C., Tilman D. 2006. **Biodiversity Loss Threatens Human Well-Being.** *PLOS Biology.* 4:e277.
- Dickinson J.L., Shirk J., Bonter D., Bonney R., Crain R.L., Martin J., Phillips T., Purcell K. 2012. **The current state of citizen science as a tool for ecological research and public engagement.** *Frontiers in Ecology and the Environment.* 10:291–297.
- Dinnage R., Cadotte M.W., Haddad N.M., Crutsinger G.M., Tilman D. 2012. **Diversity of plant evolutionary lineages promotes arthropod diversity.** *Ecol. Lett.* 15:1308–1317.
- Dirzo R., Raven P.H. 2003. **Global State of Biodiversity and Loss.** *Annual Review of Environment and Resources.* 28:137–167.
- Dixon R.K., Brown S., Houghton R.A., Solomon A.M., Trexler M.C., Wisniewski J. 1994. **Carbon Pools and Flux of Global Forest Ecosystems.** *Science.* 263:185–190.
- Donaldson M.R., Burnett N.J., Braun D.C., Suski C.D., Hinch S.G., Cooke S.J., Kerr J.T. 2016. **Taxonomic bias and international biodiversity conservation research.** *FACETS.*
- Dormann C.F. 2007. **Effects of incorporating spatial autocorrelation into the analysis of species distribution data.** *Global Ecology and Biogeography.* 16:129–138.
- Driscoll D.A., Banks S.C., Barton P.S., Ikin K., Lentini P., Lindenmayer D.B., Smith A.L., Berry L.E., Burns E.L., Edworthy A., Evans M.J., Gibson R., Heinsohn R., Howland B., Kay G., Munro N.,

- Scheele B.C., Stirnemann I., Stojanovic D., Sweaney N., Villaseñor N.R., Westgate M.J. 2014. **The Trajectory of Dispersal Research in Conservation Biology**. Systematic Review. PLOS ONE. 9:e95053.
- Dubois A. 2017. **The need for reference specimens in zoological taxonomy and nomenclature**. Bionomina. 12:4–38.
  - Duffy J.E., Godwin C.M., Cardinale B.J. 2017. **Biodiversity effects in the wild are common and as strong as key drivers of productivity**. Nature. 549:261–264.
  - Duke C.S., Porter J.H. 2013. **The Ethics of Data Sharing and Reuse in Biology**. BioScience. 63:483–489.
  - Edinger E.N., Jompa J., Limmon G.V., Widjatmoko W., Risk M.J. 1998. **Reef degradation and coral biodiversity in indonesia: Effects of land-based pollution, destructive fishing practices and changes over time**. Marine Pollution Bulletin. 36:617–630.
  - Eiben C.B., Siegel J.B., Bale J.B., Cooper S., Khatib F., Shen B.W., Players F., Stoddard B.L., Popovic Z., Baker D. 2012. **Increased Diels-Alderase activity through backbone remodeling guided by Foldit players**. Nat Biotech. 30:190–192.
  - EMC Education Services. 2015. **Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data**. John Wiley and Sons.
  - Escribano N., Ariño A.H., Galicia D. 2016. **Biodiversity data obsolescence and land uses changes**. PeerJ. 4:e2743.
  - F. Dormann C., M. McPherson J., B. Araújo M., Bivand R., Bolliger J., Carl G., G. Davies R., Hirzel A., Jetz W., Daniel Kissling W., Kühn I., Ohlemüller R., R. Peres-Neto P., Reineking B., Schröder B., M. Schurr F., Wilson R. 2007. **Methods to account for spatial autocorrelation in the analysis of species distributional data: a review**. Ecography. 30:609–628.
  - Faith D., Collen B., Ariño A., Koleff P.K.P., Guinotte J., Kerr J., Chavan V. 2013. **Bridging the biodiversity data gaps: Recommendations to meet users' data needs**. Biodiversity Informatics. 8.
  - Faith D.P. 1992. **Conservation evaluation and phylogenetic diversity**. Biological Conservation. 61:1–10.
  - Faurby, S., and J.-C. Svenning. (2015) **Historic and Prehistoric Human-driven Extinctions Have Reshaped Global Mammal Diversity Patterns**. Diversity and Distributions 21, 10: 1155–66.
  - Feeley K.J., Silman M.R. 2010. **The data void in modeling current and future distributions of tropical species**. Global Change Biology. 17:626–630.
  - Ferrer-Castán D., Morales-Barbero J., Vetaas O.R. 2016. **Water-energy dynamics, habitat heterogeneity, history, and broad-scale patterns of mammal diversity**. Acta Oecologica. 77:176–186.
  - Ferro M.L., Flick A.J. 2015. **“Collection Bias” and the Importance of Natural History Collections in Species Habitat Modeling: A Case Study Using Thoracophorus costalis Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.org**. The Coleopterists Bulletin. 69:415–425.
  - Fick S.E., Hijmans R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37:4302–4315.

- Field R., Hawkins B.A., Cornell H.V., Currie D.J., Diniz-Filho J.A.F., Guégan J.-F., Kaufman D.M., Kerr J.T., Mittelbach G.G., Oberdorff T., O'Brien E.M., Turner J.R.G. 2009. **Spatial species-richness gradients across scales: a meta-analysis.** *Journal of Biogeography*. 36:132–147.
- Filzmoser P. 2005. **Identification of multivariate outliers: A performance study.** *AUSTRIAN JOURNAL OF STATISTICS* Volume. 34:127–138.
- Fine P.V.A. 2015. **Ecological and Evolutionary Drivers of Geographic Variation in Species Diversity.** *Annual Review of Ecology, Evolution, and Systematics*. 46:369–392.
- Fisher R. A., Corbet A. S., Williams C. B. 1943. **The relation between the number of species and the number of individuals in a random sample of an animal population.** *Journal of Animal Ecology* 12: 42–58.
- Fontaine B., Perrard A., Bouchet P. 2012. **21 years of shelf life between discovery and description of new species.** *Current Biology*. 22:R943–R944.
- Foody G.M. 2004. **Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna.** *Global Ecology and Biogeography*. 13:315–320.
- Ford A.T., Cooke S.J., Goheen J.R., Young T.P. 2017. **Conserving Megafauna or Sacrificing Biodiversity?** *BioScience*. 67:193–196.
- Frankham R. 1995. **Inbreeding and Extinction: A Threshold Effect.** *Conservation Biology*. 9:792–799.
- Funk V.A., Richardson K.S. 2002. **Systematic Data in Biodiversity Studies: Use It or Lose It.** *Systematic Biology*. 51:303–316.
- Gaiji S., Chavan V., Ariño A.H., Otegui J., Hobern D., Sood R., Robles E. 2013. **Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials.** *Biodiversity Informatics*. 8.
- García-Roselló E., Guisande C., Manjarrés-Hernández A., González-Dacosta J., Heine J., Pelayo-Villamil P., González-Vilas L., Vari R.P., Vaamonde A., Granado-Lorencio C., Lobo J.M. 2015. **Can we derive macroecological patterns from primary Global Biodiversity Information Facility data?** *Global Ecology and Biogeography*. 24:335–347.
- Gardiner M.M., Allee L.L., Brown P.M., Losey J.E., Roy H.E., Smyth R.R. 2012. **Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs.** *Frontiers in Ecology and the Environment*. 10:471–476.
- Garrouste R. 2017. **The “wild shot”: photography for more biology in natural history collections, not for replacing vouchers.** *Zootaxa*. 4269:453.
- Gascon C., Brooks T.M., Contreras-MacBeath T., Heard N., Konstant W., Lamoreux J., Launay F., Maunder M., Mittermeier R.A., Molur S., Al Mubarak R.K., Parr M.J., Rhodin A.G.J., Rylands A.B., Soorae P., Sanderson J.G., Vié J.-C. 2015. **The Importance and Benefits of Species.** *Current Biology*. 25:R431–R438.
- Gaston K., Blackburn T. 2008. **Pattern and Process in Macroecology.** John Wiley and Sons.
- Gaston K.J. 2000. **Global patterns in biodiversity.** *Nature*. 405:220–227.
- Gaston K.J., Blackburn T.M., Spicer J.I. 1998. **Rapoport's rule: time for an epitaph?** *Trends in Ecology and Evolution*. 13:70–74.
- Gaston K.J., Chown S.L. 1999. **Why Rapoport's Rule Does Not Generalise.** *Oikos*. 84:309–312.

- Gaston K.J., May R.M. 1992. **Taxonomy of taxonomists**. *Nature*. 356:281–282.
- Gillman L.N., Wright S.D., Cusens J., McBride P.D., Malhi Y., Whittaker R.J. 2015. **Latitude, productivity and species richness**. *Global Ecology and Biogeography*. 24:107–117.
- Giribet G. 2016. **New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics**. *Organisms Diversity and Evolution*. 16:419–426.
- Godfray H.C.J. 2002. **Challenges for taxonomy**. *Nature*. 417:17–19.
- Gollini I., Lu B., Charlton M., Brunsdon C., Harris P. 2013. **GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models**. arXiv:1306.0413 [stat].
- Goodwin Z.A., Harris D.J., Filer D., Wood J.R.I., Scotland R.W. 2015. **Widespread mistaken identity in tropical plant collections**. *Current Biology*. 25:R1066–R1067.
- Gotelli N.J. 2004. **A taxonomic wish–list for community ecology**. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 359:585–597.
- Gotelli N.J., Colwell R.K. 2001. **Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness**. *Ecology Letters*. 4:379–391.
- Grandcolas P. 2017. **Loosing the connection between the observation and the specimen: a by-product of the digital era or a trend inherited from general biology?** *Bionomina*. 12:57–62.
- Greene H.W. 2017. **Evolutionary Scenarios and Primate Natural History**. *The American Naturalist*. 190:S69–S86.
- Gross K., Snyder-Beattie A. 2016. **A General, Synthetic Model for Predicting Biodiversity Gradients from Environmental Geometry**. *The American Naturalist*. 188:E85–E97.
- Guttman A. 1984. **R-trees: A Dynamic Index Structure for Spatial Searching**. *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*.:47–57.
- Haddad N.M., Brudvig L.A., Clobert J., Davies K.F., Gonzalez A., Holt R.D., Lovejoy T.E., Sexton J.O., Austin M.P., Collins C.D., Cook W.M., Damschen E.I., Ewers R.M., Foster B.L., Jenkins C.N., King A.J., Laurance W.F., Levey D.J., Margules C.R., Melbourne B.A., Nicholls A.O., Orrock J.L., Song D.-X., Townshend J.R. 2015. **Habitat fragmentation and its lasting impact on Earth’s ecosystems**. *Science Advances*. 1:e1500052.
- Hagen J.B. 1986. **Ecologists and taxonomists: Divergent traditions in twentieth-century plant geography**. *J Hist Biol*. 19:197–214.
- Hampton S.E., Strasser C.A., Tewksbury J.J., Gram W.K., Budden A.E., Batcheller A.L., Duke C.S., Porter J.H. 2013. **Big data and the future of ecology**. *Frontiers in Ecology and the Environment*. 11:156–162.
- Hanly P.J., Mittelbach G.G., Schemske D.W., Harrison S., Winn A.A. 2017. **Speciation and the Latitudinal Diversity Gradient: Insights from the Global Distribution of Endemic Fish**. *The American Naturalist*.:000–000.
- Hardin J., Rocke D.M. 2004. **Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator**. *Computational Statistics and Data Analysis*. 44:625–638.
- Harper J.L., Hawksworth D.L. 1994. **Preface**. *Philosophical Transactions: Biological Sciences*. 345:5–12.
- Hawkins B.A. 2001. **Ecology’s oldest pattern?** *Endeavour*. 25:133–134.

- Hawkins B.A., Field R., Cornell H.V., Currie D.J., Guégan J.-F., Kaufman D.M., Kerr J.T., Mittelbach G.G., Oberdorff T., O'Brien E.M., Porter E.E., Turner J.R.G. 2003a. **Energy, water, and broad-scale geographic patterns of species richness.** *Ecology*. 84:3105–3117.
- Hawkins B.A., Porter E.E., Felizola Diniz-Filho J.A. 2003b. **Productivity and History as Predictors of the Latitudinal Diversity Gradient of Terrestrial Birds.** *Ecology*. 84:1608–1623.
- He Z., Xu X., Deng S. 2003. **Discovering cluster-based local outliers.** *Pattern Recognition Letters*. 24:1641–1650.
- Heidorn P.B. 2009. **Shedding Light on the Dark Data in the Long Tail of Science.** *Library Trends*. 57:280–299.
- Heltshe J.F., Forrester N.E. 1983. **Estimating species richness using the jackknife procedure.** *Biometrics*. 39:1–11.
- Herkt K.M.B., Barnikel G., Skidmore A.K., Fahr J. 2016. **A high-resolution model of bat diversity and endemism for continental Africa.** *Ecological Modelling*. 320:9–28.
- Hijmans R.J. 2016. **geosphere.pdf.** Available from <https://cran.r-project.org/web/packages/geosphere/vignettes/geosphere.pdf>.
- Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., Jarvis A. 2005. **Very high resolution interpolated climate surfaces for global land areas.** *Int. J. Climatol*. 25:1965–1978.
- Hochachka W.M., Fink D., Hutchinson R.A., Sheldon D., Wong W.-K., Kelling S. 2012. **Data-intensive science applied to broad-scale citizen science.** *Trends in Ecology and Evolution*. 27:130–137.
- Holt A. 2006. **Biodiversity definitions vary within the discipline.** *Nature*. 444:146–146.
- Hortal J., Bello F. de, Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. **Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity.** *Annual Review of Ecology, Evolution, and Systematics*. 46:null.
- Hortal J., Lobo J.M., Jiménez-Valverde A. 2007. **Limitations of Biodiversity Databases: Case Study on Seed-Plant Diversity in Tenerife, Canary Islands.** *Conservation Biology*. 21:853–863.
- Howe D., Costanzo M., Fey P., Gojobori T., Hannick L., Hide W., Hill D.P., Kania R., Schaeffer M., St Pierre S., Twigger S., White O., Yon Rhee S. 2008. **Big data: The future of biocuration.** *Nature*. 455:47–50.
- Hu W., Wu F., Gao J., Yan D., Liu L., Yang X. 2017. **Influences of interpolation of species ranges on elevational species richness gradients.** *Ecography*. 40:1231–1241.
- Huber J.T. 1998. **The importance of voucher specimens, with practical guidelines for preserving specimens of the major invertebrate phyla for identification.** *Journal of Natural History*. 32:367–385.
- Husson F., Lê S., Pagès J. 2016. **Analyse de données avec R.** Presses Universitaires de Rennes.
- Hutchinson G.E. 1959. **Homage to Santa Rosalia or Why Are There So Many Kinds of Animals?** *The American Naturalist*. 93:145–159.
- Isaac N.J.B., Pocock M.J.O. 2015. **Bias and information in biological records.** *Biol J Linn Soc Lond*. 115:522–531.
- IUCN Species Survival Commission, 2001. **Categories, I. I. R. L. Criteria: Version 3.1.** Gland, Switzerland.
- IUCN (2009). **2009 IUCN Red List of threatened species.** <http://www.iucnredlist.org>.

- Jablonski D., Huang S., Roy K., Valentine J.W., Bronstein J.L. 2017. **Shaping the Latitudinal Diversity Gradient: New Perspectives from a Synthesis of Paleobiology and Biogeography.** *The American Naturalist*.:000–000.
- Jablonski D., Roy K., Valentine J.W. 2006. **Out of the Tropics: Evolutionary Dynamics of the Latitudinal Diversity Gradient.** *Science*. 314:102–106.
- Jansson R., Rodríguez-Castañeda G., Harding L.E. 2013. **What Can Multiple Phylogenies Say About the Latitudinal Diversity Gradient? A New Look at the Tropical Conservatism, Out of the Tropics, and Diversification Rate Hypotheses.** *Evolution*. 67:1741–1755.
- Jenner R.A., Steel M. 2004. **Accepting Partnership by Submission? Morphological Phylogenetics in a Molecular Millennium.** *Systematic Biology*. 53:333–359.
- Johnson N.F. 2007. **Biodiversity Informatics.** *Annu. Rev. Entomol.* 52:421–438.
- Joppa L.N., O'Connor B., Visconti P., Smith C., Geldmann J., Hoffmann M., Watson J.E.M., Butchart S.H.M., Virah-Sawmy M., Halpern B.S., Ahmed S.E., Balmford A., Sutherland W.J., Harfoot M., Hilton-Taylor C., Foden W., Minin E.D., Pagad S., Genovesi P., Hutton J., Burgess N.D. 2016. **Filling in biodiversity threat gaps.** *Science*. 352:416–418.
- Jordan S. 2008. **The Gaia project: Technique, performance and status.** *Astron. Nachr.* 329:875–880.
- Kamp J., Oppel S., Heldbjerg H., Nyegaard T., Donald P.F. 2016. **Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark.** *Diversity Distrib.* 22:1024–1035.
- Katsanevakis S., Wallentinus I., Zenetos A., Leppäkoski E., Ertan Çinar M., Oztürk B., Grabowski M., Golani D., Cardoso A.C. 2014. **Impacts of invasive alien marine species on ecosystem services and biodiversity: a pan-European review.** *Aquatic Invasions*. 9:pp 391-423.
- Kelling S., Hochachka W.M., Fink D., Riedewald M., Caruana R., Ballard G., Hooker G. 2009. **Data-Intensive Science: A New Paradigm for Biodiversity Studies.** *BioScience*. 59:613–620.
- Kemp C. 2015. **Museums: The endangered dead.** *Nature News*. 518:292.
- Kitchin R. 2013. **Big data and human geography: Opportunities, challenges and risks.** *Dialogues in Human Geography*. 3:262–267.
- Kitchin R. 2014. **Big Data, new epistemologies and paradigm shifts.** *Big Data and Society*. 1:2053951714528481.
- Knapp S. 2015. **The changing role of collections and field research.** *Descriptive Taxonomy: The Foundation of Biodiversity Research.* Cambridge University Press.
- Knapp S. 2017. **Using the “Natural History Large Hadron Collider” to tell us about plant diversity.** Available from <http://blogs.biomedcentral.com/on-biology/2017/03/07/using-the-natural-history-large-hadron-collider-to-tell-us-about-plant-diversity/>.
- Kormondy. 2012. **A Brief Introduction to the History of Ecology.** *The American Biology Teacher*. 74:441–443.
- Kosmala M., Wiggins A., Swanson A., Simmons B. 2016. **Assessing data quality in citizen science.** *Front Ecol Environ*. 14:551–560.
- Kreft H., Jetz W. 2007. **Global patterns and determinants of vascular plant diversity.** *PNAS*. 104:5925–5930.



- Kumar S. Manoj, Kumar G. Dileep. 2016. **Effective Big Data Management and Opportunities for Implementation**. IGI Global.
- La Salle J., Williams K.J., Moritz C. 2016. **Biodiversity analysis in the digital era**. *Phil. Trans. R. Soc. B.* 371:20150337.
- Lamanna C., Blonder B., Violle C., Kraft N.J.B., Sandel B., Šímová I., Donoghue J.C., Svenning J.-C., McGill B.J., Boyle B., Buzzard V., Dolins S., Jørgensen P.M., Marcuse-Kubitza A., Morueta-Holme N., Peet R.K., Piel W.H., Regetz J., Schildhauer M., Spencer N., Thiers B., Wisser S.K., Enquist B.J. 2014. **Functional trait space and the latitudinal diversity gradient**. *PNAS.* 111:13745–13750.
- Lambeck R.J. 1997. **Focal Species: A Multi-Species Umbrella for Nature Conservation**. *Biology.* 11:849–856.
- Lapin M., Barnes B.V. 1995. **Using the Landscape Ecosystem Approach to Assess Species and Ecosystem Diversity**. *Conservation Biology.* 9:1148–1158.
- Larsen B.B., Miller E.C., Rhodes M.K., Wiens J.J. 2017. **Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life**. *The Quarterly Review of Biology.* 92:229–265.
- Lathe W., Williams J., Mangan M., Karolchik D. 2008. **Genomic data resources: challenges and promises**. *Nat. Educ.* 1:2
- Law J., Lynch M. 1988. **Lists, field guides, and the descriptive organization of seeing: Birdwatching as an exemplary observational activity**. *Hum Stud.* 11:271–303.
- Lawler J.J., White D., Sifneos J.C., Master L.L. 2003. **Rare Species and the Use of Indicator Groups for Conservation Planning**. *Conservation Biology.* 17:875–882.
- Le Bras G., Pignal M., Jeanson M.L., Muller S., Aupic C., Carré B., Flament G., Gaudeul M., Gonçalves C., Invernón V.R., Jabbour F., Lerat E., Lowry P.P., Offroy B., Pimparé E.P., Poncy O., Rouhan G., Haeevermans T. 2017. **The French Muséum national d’histoire naturelle vascular plant herbarium collection dataset**. *Scientific Data.* 4:sdata201716.
- Leather S. R., 2009. **Taxonomic chauvinism threatens the future of entomology**. *Biologist* 56, 10–13.
- Lees D.C., Colwell R.K. 2007. **A strong Madagascan rainforest MDE and no equatorward increase in species richness: re-analysis of ‘The missing Madagascan mid-domain effect’, by Kerr J.T., Perring M. and Currie D.J. (Ecology Letters 9:149–159, 2006)**. *Ecology Letters.* 10:E4–E8.
- Legendre P. 1993. **Spatial Autocorrelation: Trouble or New Paradigm?** *Ecology.* 74:1659–1673.
- Leonelli S. 2014. **What difference does quantity make? On the epistemology of Big Data in biology**. *Big Data and Society.* 1:2053951714534395.
- Lewinsohn T.M., Roslin T. 2008. **Four ways towards tropical herbivore megadiversity**. *Ecology Letters.* 11:398–416.
- Lieberoth A., Pedersen M.K., Marin A.C., Planke T., Sherson J.F. 2014. **Getting Humans to do Quantum Optimization - User Acquisition, Engagement and Early Results from the Citizen Cyberscience Game Quantum Moves**. *Human Computation.* 1.
- Lisón F., Calvo J.F. 2013. **Ecological niche modelling of three pipistrelle bat species in semiarid Mediterranean landscapes**. *Acta Oecologica.* 47:68–73.
- List M., Ebert P., Albrecht F. 2017. **Ten Simple Rules for Developing Usable Software in Computational Biology**. *PLOS Computational Biology.* 13:e1005265.

- Lloyd C.D. 2010. **Nonstationary models for exploring and mapping monthly precipitation in the United Kingdom.** *Int. J. Climatol.* 30:390–405.
- Löbl I. 2017. **Assessing biodiversity: a pain in the neck.** *Bionomina.* 12:39.
- Lomolino M.V., Heaney L.R. 2004. **Frontiers of Biogeography: New Directions in the Geography of Nature.** Sinauer Associates.
- Losos J.B., Arnold S.J., Bejerano G., Iii E.D.B., Hibbett D., Hoekstra H.E., Mindell D.P., Monteiro A., Moritz C., Orr H.A., Petrov D.A., Renner S.S., Ricklefs R.E., Soltis P.S., Turner T.L. 2013. **Evolutionary Biology for the 21st Century.** *PLOS Biology.* 11:e1001466.
- Lucas A., Scholz I., Boehme R., Jasson S., Maechler M. 2017. **Package ‘gmp.’** Available from <https://cran.r-project.org/web/packages/gmp/gmp.pdf>.
- Mace G.M., Norris K., Fitter A.H. 2012. **Biodiversity and ecosystem services: a multilayered relationship.** *Trends in Ecology and Evolution.* 27:19–26.
- Machac A., Graham C.H. 2016. **Regional Diversity and Diversification in Mammals.** *The American Naturalist.* 189:E1–E13.
- MacKenzie D.I., Nichols J.D., Sutton N., Kawanishi K., Bailey L.L. 2005. **Improving Inferences in Population Studies of Rare Species That Are Detected Imperfectly.** *Ecology.* 86:1101–1113.
- Magurran A.E. (1955-) A. 2004. **Measuring biological diversity.** Blackwell. Malden (Mass.), Oxford, UK, Carlton (Victoria).
- Magurran A.E., McGill B.J., editors. 2010. **Biological Diversity: Frontiers in Measurement and Assessment.**
- Mahalanobis P., 1936. **On the generalized distance in statistics.** *Proc. Nat. Inst. Sci. India (Calcutta)* 49–55
- Maldonado C., Molina C.I., Zizka A., Persson C., Taylor C.M., Albán J., Chilquillo E., Rønsted N., Antonelli A. 2015. **Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?** *Global Ecology and Biogeography.* 24:973–984.
- Mannion P.D., Upchurch P., Benson R.B.J., Goswami A. 2014. **The latitudinal biodiversity gradient through deep time.** *Trends in Ecology and Evolution.* 29:42–50.
- Margono B.A., Potapov P.V., Turubanova S., Stolle F., Hansen M.C. 2014. **Primary forest cover loss in Indonesia over 2000–2012.** *Nature Clim. Change.* 4:730–735.
- Marin J., Hedges S.B. 2016. **Time best explains global variation in species richness of amphibians, birds and mammals.** *J. Biogeogr.*
- Marshall S.A., Evenhuis N.L. 2015. **New species without dead bodies: a case for photo-based descriptions, illustrated by a striking new species of *Marleyimyia* Hesse (Diptera, Bombyliidae) from South Africa.** *Zookeys.* 117–127.
- Martin A.C., Harvey W.J. 2017. **The Global Pollen Project: A New Tool for Pollen Identification and the Dissemination of Physical Reference Collections.** *Methods Ecol Evol.*
- Martín-López B., Montes C., Benayas J. 2007. **The non-economic motives behind the willingness to pay for biodiversity conservation.** *Biological Conservation.* 139:67–82.
- Martín-López B., Montes C., Ramírez L., Benayas J. 2009. **What drives policy decision-making related to species conservation?** *Biological Conservation.* 142:1370–1380.
- Marx V. 2013a. **Biology: The big challenges of big data.** *Nature.* 498:255–260.
- Marx V. 2013b. **Neuroscience waves to the crowd.** *Nat Meth.* 10:1069–1074.

- May R.M. 1988. **How Many Species Are There on Earth?** *Science*. 241:1441–1449.
- May R.M. 1990. **Taxonomy as destiny.** *Nature*. 347:129–130.
- May R.M. 2004. **Tomorrow's taxonomy: collecting new species in the field will remain the rate-limiting step.** *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 359:733–734.
- May R.M., Godfrey J. 1994. **Biological Diversity: Differences between Land and Sea [and Discussion].** *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 343:105–111.
- McCullagh P., Nelder J.A. 1989. **Generalized Linear Models, Second Edition.** CRC Press.
- McKenzie A.J., Robertson P.A. 2015. **Which Species Are We Researching and Why? A Case Study of the Ecology of British Breeding Birds.** *PLOS ONE*. 10:e0131004.
- McKinney M.L. 1999. **High Rates of Extinction and Threat in Poorly Studied Taxa.** *Conservation Biology*. 13:1273–1281.
- McNew L.B., Handel C.M. 2015. **Evaluating species richness: biased ecological inference results from spatial heterogeneity in detection probabilities.** *Ecological Applications*.
- McNyset K.M. 2005. **Use of ecological niche modelling to predict distributions of freshwater fish species in Kansas.** *Ecology of Freshwater Fish*. 14:243–255.
- Mellin C., Mengersen K., Bradshaw C.J.A., Caley M.J. 2014. **Generalizing the use of geographical weights in biodiversity modelling.** *Global Ecology and Biogeography*. 23:1314–1323.
- Meseguer A.S., Lobo J.M., Ree R., Beerling D.J., Sanmartín I. 2015. **Integrating Fossils, Phylogenies, and Niche Models into Biogeography to Reveal Ancient Evolutionary History: The Case of *Hypericum* (Hypericaceae).** *Syst Biol*. 64:215–232.
- Meyer C., Kreft H., Guralnick R., Jetz W. 2015. **Global priorities for an effective information basis of biodiversity distributions.** *Nat Commun*. 6:8221.
- Meyer C., Weigelt P., Kreft H. 2016. **Multidimensional biases, gaps and uncertainties in global plant occurrence information.** *Ecology Letters*. 19:992–1006.
- Meza-Joya F.L., Torres M. 2016. **Spatial diversity patterns of *Pristimantis* frogs in the Tropical Andes.** *Ecol Evol.*:n/a-n/a.
- Michener W.K., Jones M.B. 2012. **Ecoinformatics: supporting ecology as a data-intensive science.** *Trends in Ecology and Evolution*. 27:85–93.
- Miller-Rushing A., Primack R., Bonney R. 2012. **The history of public participation in ecological research.** *Frontiers in Ecology and the Environment*. 10:285–290.
- Miraldo A., Li S., Borregaard M.K., Flórez-Rodríguez A., Gopalakrishnan S., Rizvanovic M., Wang Z., Rahbek C., Marske K.A., Nogués-Bravo D. 2016. **An Anthropocene map of genetic diversity.** *Science*. 353:1532–1535.
- Mittelbach G.G., Schemske D.W., Cornell H.V., Allen A.P., Brown J.M., Bush M.B., Harrison S.P., Hurlbert A.H., Knowlton N., Lessios H.A., McCain C.M., McCune A.R., McDade L.A., McPeck M.A., Near T.J., Price T.D., Ricklefs R.E., Roy K., Sax D.F., Schluter D., Sobel J.M., Turelli M. 2007. **Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography.** *Ecology Letters*. 10:315–331.
- Mora C., Robertson D.R. 2005. **Causes of Latitudinal Gradients in Species Richness: A Test with Fishes of the Tropical Eastern Pacific.** *Ecology*. 86:1771–1782.

- Mora C., Tittensor D.P., Adl S., Simpson A.G.B., Worm B. 2011. **How Many Species Are There on Earth and in the Ocean?** PLOS Biology. 9:e1001127.
- Mouillot D., Bellwood D.R., Baraloto C., Chave J., Galzin R., Harmelin-Vivien M., Kulbicki M., Lavergne S., Lavorel S., Mouquet N., Paine C.E.T., Renaud J., Thuiller W. 2013. **Rare Species Support Vulnerable Functions in High-Diversity Ecosystems.** PLOS Biology. 11:e1001569.
- Mu Q., Zhao M., Running S.W. 2011. **Improvements to a MODIS global terrestrial evapotranspiration algorithm.** Remote Sensing of Environment. 115:1781–1800.
- Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A.B., Kent J. 2000. **Biodiversity hotspots for conservation priorities.** Nature. 403:853–858.
- Newman G., Wiggins A., Crall A., Graham E., Newman S., Crowston K. 2012. **The future of citizen science: emerging technologies and shifting paradigms.** Frontiers in Ecology and the Environment. 10:298–304.
- Nicholson D.B., Holroyd P.A., Valdes P., Barrett P.M. 2016. **Latitudinal diversity gradients in Mesozoic non-marine turtles.** Royal Society Open Science. 3:160581.
- Nielsen D.L., Shiel R.J., Smith F.J. 1998. **Ecology versus taxonomy: is there a middle ground?** Hydrobiologia. 387–388:451–457.
- Nualart N., Ibáñez N., Soriano I., López-Pujol J. 2017. **Assessing the Relevance of Herbarium Collections as Tools for Conservation Biology.** Bot. Rev.:1–23.
- Numanagić I., Bonfield J.K., Hach F., Voges J., Ostermann J., Alberti C., Mattavelli M., Sahinalp S.C. 2016. **Comparison of high-throughput sequencing data compression tools.** Nature Methods. 13:1005–1008.
- Oliveira U., Paglia A.P., Brescovit A.D., de Carvalho C.J.B., Silva D.P., Rezende D.T., Leite F.S.F., Batista J.A.N., Barbosa J.P.P.P., Stehmann J.R., Ascher J.S., de Vasconcelos M.F., De Marco P., Löwenberg-Neto P., Dias P.G., Ferro V.G., Santos A.J. 2016. **The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity.** Diversity Distrib. Otegui J., Ariño A.H., Encinas M.A., Pando F. 2013. **Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF).** PLoS ONE. 8:e55144.
- Page R.D.M. 2016. **DNA barcoding and taxonomy: dark taxa and dark texts.** Phil. Trans. R. Soc. B. 371:20150334.
- Pape T. 2016. **Taxonomy: Species can be named from photos.** Nature. 537:307–307.
- Paradis E., Claude J., Strimmer K. 2004. **APE: Analyses of Phylogenetics and Evolution in R language.** Bioinformatics. 20:289–290.
- Pardo I., Pata M.P., Gómez D., García M.B. 2013. **A Novel Method to Handle the Effect of Uneven Sampling Effort in Biodiversity Databases.** PLoS One. 8.
- Parr C.S., Guralnick R., Cellinese N., Page R.D.M. 2012. **Evolutionary informatics: unifying knowledge about the diversity of life.** Trends in Ecology and Evolution. 27:94–103.
- Pawar S. 2003. **Taxonomic Chauvinism and the Methodologically Challenged.** BioScience. 53:861–864.
- Pearson R.G., Boyero L. 2009. **Gradients in regional diversity of freshwater taxa.** Journal of the North American Benthological Society. 28:504–514.
- Pereira H.M. 2016. **A latitudinal gradient for genetic diversity.** Science. 353:1494–1495.

- Pérez-Ponce de León G., Poulin R. 2016. **Taxonomic distribution of cryptic diversity among metazoans: not so homogeneous after all.** *Biology Letters*. 12:20160371.
- Peterson A.T., Ball L.G., Cohoon K.P. 2002. **Predicting distributions of Mexican birds using ecological niche modelling methods.** *Ibis*. 144:E27–E32.
- Peterson A.T., Knapp S., Guralnick R., Soberón J., Holder M.T. 2010. **The big questions for biodiversity informatics.** *Systematics and Biodiversity*. 8:159–168.
- Peterson A.T., Soberón J., Krishtalka L. 2015. **A global perspective on decadal challenges and priorities in biodiversity informatics.** *BMC Ecology*. 15:15.
- Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. **Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough.** *PLOS Biology*. 9:e1000602.
- Pianka E.R. 1966. **Latitudinal Gradients in Species Diversity: A Review of Concepts.** *The American Naturalist*. 100:33–46.
- Platnick N.I. 1991. **Patterns of biodiversity: tropical vs temperate.** *Journal of Natural History*. 25:1083–1088.
- Powell K.I., Chase J.M., Knight T.M. 2013. **Invasive Plants Have Scale-Dependent Effects on Diversity by Altering Species-Area Relationships.** *Science*. 339:316–318.
- Powney G.D., Isaac N.J.B. 2015. **Beyond maps: a review of the applications of biological records.** *Biol J Linn Soc Lond*. 115:532–542.
- Prins P., de Ligt J., Tarasov A., Jansen R.C., Cuppen E., Bourne P.E. 2015. **Toward effective software solutions for big biology.** *Nat Biotech*. 33:686–687.
- Prlić A., Procter J.B. 2012. **Ten Simple Rules for the Open Development of Scientific Software.** *PLOS Computational Biology*. 8:e1002802.
- Pulido-Santacruz P., Weir J.T. 2016. **Extinction as a driver of avian latitudinal diversity gradients.** *Evolution*. 70:860–872.
- Pyron R.A. 2011. **Divergence Time Estimation Using Fossils as Terminal Taxa and the Origins of Lissamphibia.** *Systematic Biology*. 60:466–481.
- Pyron R.A. 2015. **Post-molecular systematics and the future of phylogenetics.** *Trends in Ecology and Evolution*. 30:384–389.
- R Development Core Team 2008. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Rahbek C. 2005. **The role of spatial scale and the perception of large-scale species-richness patterns.** *Ecology Letters*. 8:224–239.
- Ressurreição A., Gibbons J., Kaiser M., Dentinho T.P., Zarzycki T., Bentley C., Austen M., Burdon D., Atkins J., Santos R.S., Edwards-Jones G. 2012. **Different cultures, different values: The role of cultural variation in public's WTP for marine species conservation.** *Biological Conservation*. 145:148–159.
- Ricklefs R.E. 2004. **A comprehensive framework for global patterns in biodiversity.** *Ecology Letters*. 7:1–15.
- Ripple W.J., Chapron G., López-Bao J.V., Durant S.M., Macdonald D.W., Lindsey P.A., Bennett E.L., Beschta R.L., Bruskotter J.T., Campos-Arceiz A., Corlett R.T., Darimont C.T., Dickman A.J., Dirzo R.,

- Dublin H.T., Estes J.A., Everatt K.T., Galetti M., Goswami V.R., Hayward M.W., Hedges S., Hoffmann M., Hunter L.T.B., Kerley G.I.H., Letnic M., Levi T., Maisels F., Morrison J.C., Nelson M.P., Newsome T.M., Painter L., Pringle R.M., Sandom C.J., Terborgh J., Treves A., Van Valkenburgh B., Vucetich J.A., Wirsing A.J., Wallach A.D., Wolf C., Woodroffe R., Young H., Zhang L. 2016. **Saving the World's Terrestrial Megafauna**. *BioScience*. 66:807–812.
- Rivadeneira M.M., Alballay A.H., Villafaña J.A., Raimondi P.T., Blanchette C.A., Fenberg P.B. 2015. **Geographic patterns of diversification and the latitudinal gradient of richness of rocky intertidal gastropods: the 'into the tropical museum' hypothesis**. *Global Ecology and Biogeography*.:n/a-n/a.
  - Robinson J.G., Redford K.H. 1986. **Body Size, Diet, and Population Density of Neotropical Forest Mammals**. *The American Naturalist*. 128:665–680.
  - Rodrigues A.S.L., Gray C.L., Crowter B.J., Ewers R.M., Stuart S.N., Whitten T., Manica A. 2010. **A Global Assessment of Amphibian Taxonomic Effort and Expertise**. *BioScience*. 60:798–806.
  - 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*.
  - Rohde K., Heap M., Heap D. 1993. **Rapoport's Rule Does Not Apply to Marine Teleosts and Cannot Explain Latitudinal Gradients in Species Richness**. *The American Naturalist*. 142:1–16.
  - Rolland J., Condamine F.L., Jiguet F., Morlon H. 2014. **Faster Speciation and Reduced Extinction in the Tropics Contribute to the Mammalian Latitudinal Diversity Gradient**. *PLOS Biology*. 12:e1001775.
  - Rosenheim J.A., Gratton C. 2017. **Ecoinformatics (Big Data) for Agricultural Entomology: Pitfalls, Progress, and Promise**. *Annual Review of Entomology*. 62:399–417.
  - Rosenthal M.F., Gertler M., Hamilton A.D., Prasad S., Andrade M.C.B. 2017. **Taxonomic bias in animal behaviour publications**. *Animal Behaviour*. 127:83–89.
  - Rosselló-Mora R., Amann R. 2001. **The species concept for prokaryotes**. *FEMS Microbiol Rev*. 25:39–67.
  - Santos A.M., Branco M. 2012. **The quality of name-based species records in databases**. *Trends in Ecology and Evolution*. 27:6–7.
  - Sazid M., Ramakrishnan R. 2003. **GeoTIFF - A standard image file format for GIS applications**. *Geospatial World*.
  - Schilthuizen M., Vairappan C.S., Slade E.M., Mann D.J., Miller J.A. 2015. **Specimens as primary data: museums and 'open science.'** *Trends in Ecology and Evolution*. 30:237–238.
  - Schluter D., Pennell M.W. 2017. **Speciation gradients and the distribution of biodiversity**. *Nature*. 546:48–55.
  - Seddon P.J., Soorae P.S., Launay F. 2005. **Taxonomic bias in reintroduction projects**. *Animal Conservation*. 8:51–58.
  - Shine null, Bonnet null. 2000. **Snakes: a new "model organism" in ecological research?** *Trends Ecol. Evol. (Amst.)*. 15:221–222.
  - Siemann E., Tilman D., Haarstad J., Ritchie M. 1998. **Experimental Tests of the Dependence of Arthropod Diversity on Plant Diversity**. *The American Naturalist*. 152:738–750.

- Sikes, Copas, Hirsch, Longino, Schigel. 2016. **On natural history collections, digitized and not: a response to Ferro and Flick.**
- Silvertown J., Harvey M., Greenwood R., Dodd M., Rosewell J., Rebelo T., Ansine J., McConway K. 2015. **Crowdsourcing the identification of organisms: A case-study of iSpot.** Zookeys.:125–146.
- Siqueira A.C., Oliveira-Santos L.G.R., Cowman P.F., Floeter S.R. 2016. **Evolutionary processes underlying latitudinal differences in reef fish biodiversity.** Global Ecol. Biogeogr.
- Smith E.P., Belle G. van. 1984. **Nonparametric Estimation of Species Richness.** Biometrics. 40:119–129.
- Smith N.D., Turner A.H., Macleod N. 2005. **Morphology's Role in Phylogeny Reconstruction: Perspectives from Paleontology.** Systematic Biology. 54:166–173.
- Soberón J., Peterson T. 2004. **Biodiversity informatics: managing and applying primary biodiversity data.** Philosophical Transactions of the Royal Society of London B: Biological Sciences. 359:689–698.
- Sosef M.S.M., Dauby G., Blach-Overgaard A., van der Burgt X., Catarino L., Damen T., Deblauwe V., Dessein S., Dransfield J., Droissart V., Duarte M.C., Engledow H., Fadeur G., Figueira R., Gereau R.E., Hardy O.J., Harris D.J., de Heij J., Janssens S., Klomberg Y., Ley A.C., Mackinder B.A., Meerts P., van de Poel J.L., Sonké B., Stévant T., Stoffelen P., Svenning J.-C., Sepulchre P., Zaiss R., Wieringa J.J., Couvreur T.L.P. 2017. **Exploring the floristic diversity of tropical Africa.** BMC Biology. 15:15.
- Spellerberg I.F., Fedor P.J. 2003. **A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index.** Global Ecology and Biogeography. 12:177–179.
- Stahlschmidt Z.R. 2011. **Taxonomic Chauvinism Revisited: Insight from Parental Care Research.** PLOS ONE. 6:e24192.
- Stein B.A., Master L.L., Morse L.E. 2002. **Taxonomic Bias and Vulnerable Species.** Science. 297:1807–1807.
- Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E. 2015. **Big Data: Astronomical or Genomical?** PLOS Biology. 13:e1002195.
- Stevens G.C. 1989. **The Latitudinal Gradient in Geographical Range: How so Many Species Coexist in the Tropics.** The American Naturalist. 133:240–256.
- Stevens R.D., Cox S.B., Strauss R.E., Willig M.R. 2003. **Patterns of functional diversity across an extensive environmental gradient: vertebrate consumers, hidden treatments and latitudinal trends.** Ecology Letters. 6:1099–1108.
- Stork N.E. 1988. **Insect diversity: facts, fiction and speculation\*.** Biological Journal of the Linnean Society. 35:321–337.
- Sullivan B.L., Wood C.L., Iliff M.J., Bonney R.E., Fink D., Kelling S. 2009. **eBird: A citizen-based bird observation network in the biological sciences.** Biological Conservation. 142:2282–2292.
- Taschuk M., Wilson G. 2017. **Ten simple rules for making research software more robust.** PLOS Computational Biology. 13:e1005412.
- Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., de Siqueira M.F., Grainger A., Hannah L., Hughes L., Huntley B., van Jaarsveld A.S., Midgley

- G.F., Miles L., Ortega-Huerta M.A., Townsend Peterson A., Phillips O.L., Williams S.E. 2004. **Extinction risk from climate change**. *Nature*. 427:145–148.
- Thuiller W., Lafourcade B., Engler R., Araújo M.B. 2009. **BIOMOD – a platform for ensemble forecasting of species distributions**. *Ecography*. 32:369–373.
  - Tiago P., Pereira H.M., Capinha C. 2017. **Using citizen science data to estimate climatic niches and species distributions**. *Basic and Applied Ecology*. 20:75–85.
  - Tilman D., Knops J., Wedin D., Reich P., Ritchie M., Siemann E. 1997. **The Influence of Functional Diversity and Composition on Ecosystem Processes**. *Science*. 277:1300–1302.
  - Tobler W.R. 1970. **A Computer Movie Simulating Urban Growth in the Detroit Region**. *Economic Geography*. 46:234–240.
  - Töpel M., Zizka A., Calió M.F., Scharn R., Silvestro D., Antonelli A. 2017. **SpeciesGeoCoder: Fast Categorization of Species Occurrences for Analyses of Biodiversity, Biogeography, Ecology, and Evolution**. *Syst Biol*. 66:145–151.
  - Tress G., Tress B., Fry G. 2005. **Clarifying Integrative Research Concepts in Landscape Ecology**. *Landscape Ecol*. 20:479–493.
  - Troia M.J., McManamay R.A. 2016. **Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States**. *Ecol Evol*.
  - Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R., Legendre F. 2017. **Taxonomic bias in biodiversity data and societal preferences**. *Scientific Reports*. 7:9132.
  - Turney S., Cameron E.R., Cloutier C.A., Buddle C.M. 2015. **Non-repeatable science: assessing the frequency of voucher specimen deposition reveals that most arthropod research cannot be verified**. *PeerJ*. 3.
  - Turpie J.K. 2003. **The existence value of biodiversity in South Africa: how interest, experience, knowledge, income and perceived level of threat influence local willingness to pay**. *Ecological Economics*. 46:199–216.
  - Venables W.N., Ripley B.D. 1994. **Modern Applied Statistics with S-PLUS**. New York: Springer.
  - Venables W.N., Ripley B.D. 2013. **Modern Applied Statistics with S-PLUS**. Springer Science and Business Media.
  - Vicario S., Hardisty A., Haitas N. 2011. **BioVeL: Biodiversity Virtual e-Laboratory**. *EMBnet.journal*. 17:5–6.
  - Vogel Ely C., de Loreto Bordignon S.A., Trevisan R., Boldrini I.I. **Implications of poor taxonomy in conservation**. *Journal for Nature Conservation*.
  - Vörösmarty C.J., McIntyre P.B., Gessner M.O., Dudgeon D., Prusevich A., Green P., Glidden S., Bunn S.E., Sullivan C.A., Liermann C.R., Davies P.M. 2010. **Global threats to human water security and river biodiversity**. *Nature*. 467:555–561.
  - Walker S.J. 2014. **Big Data: A Revolution That Will Transform How We Live, Work, and Think**. *International Journal of Advertising*. 33:181–183.
  - Wanninger A. 2015. **Morphology is dead, long live morphology! Integrating MorphoEvoDevo into molecular EvoDevo and phylogenomics**. *Frontiers in Ecology and Evolution*. 3.
  - Ward E.J., Marshall K.N., Ross T., Sedgley A., Hass T., Pearson S.F., Joyce G., Hamel N.J., Hodum P.J., Faucett R. 2015. **Using citizen-science data to identify local hotspots of seabird occurrence**. *PeerJ*. 3:e704.



- Weir J.T., Schluter D. 2007. **The Latitudinal Gradient in Recent Speciation and Extinction Rates of Birds and Mammals.** *Science*. 315:1574–1576.
- Whittaker R.H. 1972. **Evolution and Measurement of Species Diversity.** *Taxon*. 21:213–251.
- Whittaker R.J., Willis K.J., Field R. 2001. **Scale and species richness: towards a general, hierarchical theory of species diversity.** *Journal of Biogeography*. 28:453–470.
- Whitton F.J.S., Purvis A., Orme C.D.L., Olalla-Tárraga M.Á. 2012. **Understanding global patterns in amphibian geographic range size: does Rapoport rule?** *Global Ecology and Biogeography*. 21:179–190.
- Wickham. H., 2011. **The Split-Apply-Combine Strategy for Data Analysis.** *Journal of Statistical Software*. 40:1–29.
- Wickham H., 2015. **S: Scale functions for visualization.** Cran.r-project.org. at <http://CRAN.R-project.org/package=scales>, package version 0.4.1.
- Wickham H., 2016. **ggplot2: Elegant Graphics for Data Analysis.** Springer.
- Wieczorek J., Bloom D., Guralnick R., Blum S., Döring M., Giovanni R., Robertson T., Vieglais D. 2012. **Darwin Core: An Evolving Community-Developed Biodiversity Data Standard.** *PLoS ONE*. 7:e29715.
- Wiens, John J., Michael J. Donoghue. (2004) **Historical Biogeography, Ecology and Species Richness.** *Trends in Ecology & Evolution*. 12: 639–44.
- Wiens J.A., Stralberg D., Jongsomjit D., Howell C.A., Snyder M.A. 2009. **Niches, models, and climate change: Assessing the assumptions and uncertainties.** *PNAS*. 106:19729–19736.
- Wiens J.J., Collins T. 2004. **The Role of Morphological Data in Phylogeny Reconstruction.** *Systematic Biology*. 53:653–661.
- Wiens J.J., Kuczynski C.A., Townsend T., Reeder T.W., Mulcahy D.G., Sites J.W. 2010. **Combining Phylogenomics and Fossils in Higher-Level Squamate Reptile Phylogeny: Molecular Data Change the Placement of Fossil Taxa.** *Systematic Biology*. 59:674–688.
- Willig M.R., Kaufman D.M., Stevens R.D. 2003. **LATITUDINAL GRADIENTS OF BIODIVERSITY: Pattern, Process, Scale, and Synthesis.** *Annual Review of Ecology, Evolution, and Systematics*. 34:273–309.
- Wilson E.O. 1988. **Biodiversity.** National Academies Press.
- Wilson E.O. 2000. **A Global Biodiversity Map.** *Science*. 289:2279–2279.
- Wilson J.R., Procheş Ş., Braschler B., Dixon E.S., Richardson D.M. 2007. **The (bio)diversity of science reflects the interests of society.** *Frontiers in Ecology and the Environment*. 5:409–414.
- Wipfler B., Pohl H., Yavorskaya M.I., Beutel R.G. 2016. **A review of methods for analysing insect structures — the role of morphology in the age of phylogenomics.** *Current Opinion in Insect Science*. 18:60–68.
- Wood H.M., Matzke N.J., Gillespie R.G., Griswold C.E. 2013. **Treating Fossils as Terminal Taxa in Divergence Time Estimation Reveals Ancient Vicariance Patterns in the Palpimanoid Spiders.** *Syst Biol*. 62:264–284.
- Yang W., Ma K., Kreft H. 2013. **Geographical sampling bias in a large distributional database and its effects on species richness–environment models.** *J. Biogeogr*. 40:1415–1426.
- Yassin A. 2013. **Phylogenetic classification of the Drosophilidae Rondani (Diptera): the role of morphology in the postgenomic era.** *Systematic Entomology*. 38:349–364.

- Yesson C., Brewer P.W., Sutton T., Caithness N., Pahwa J.S., Burgess M., Gray W.A., White R.J., Jones A.C., Bisby F.A., Culham A. 2007. **How Global Is the Global Biodiversity Information Facility?** PLoS ONE. 2:e1124.
- Zapponi L., Cini A., Bardiani M., Hardersen S., Maura M., Maurizi E., Redolfi De Zan L., Audisio P., Bologna M.A., Carpaneto G.M., Roversi P.F., Sabbatini Peverieri G., Mason F., Campanaro A. **Citizen science data as an efficient tool for mapping protected saproxylic beetles.** Biological Conservation.
- Zarybnicka M., Sklenicka P., Tryjanowski P. 2017. **A Webcast of Bird Nesting as a State-of-the-Art Citizen Science.** PLOS Biology. 15:e2001132.
- Zook M., Barocas S., Boyd D., Crawford K., Keller E., Gangadharan S.P., Goodman A., Hollander R., Koenig B.A., Metcalf J., Narayanan A., Nelson A., Pasquale F. 2017. **Ten simple rules for responsible big data research.** PLOS Computational Biology. 13:e1005399.
- Zuur A.F., Ieno E.N., Walker N., Saveliev A.A., Smith G.M. 2009. **Mixed Effects Models and Extensions in Ecology with R.** Springer

# Appendixes

## Appendix 1: List of used Java Libraries and Dependencies

This appendix regroups the list of all the Java libraries used in my custom programs.

- ArcGIS Runtime SDK for Java
- Java Runtime Environment 1.8.0
- org\apache\commons\commons-csv\1.1\commons-csv-1.1.jar
- commons-lang\commons-lang\2.6\commons-lang-2.6.jar
- org\postgresql\postgresql\9.4-1200-jdbc41\postgresql-9.4-1200-jdbc41.jar
- com\github\dblock\waffle\waffle-jna\1.7\waffle-jna-1.7.jar
- net\java\dev\jna\jna\4.1.0\jna-4.1.0.jar
- net\java\dev\jna\jna-platform\4.1.0\jna-platform-4.1.0.jar
- org\slf4j\slf4j-api\1.7.7\slf4j-api-1.7.7.jar
- com\google\guava\guava\18.0\guava-18.0.jar
- org\slf4j\slf4j-simple\1.7.7\slf4j-simple-1.7.7.jar
- mysql\mysql-connector-java\5.1.34\mysql-connector-java-5.1.34.jar
- uk\com\robust-it\cloning\1.9.1\cloning-1.9.1.jar
- org\objenesis\objenesis\2.1\objenesis-2.1.jar
- org\apache\commons\commons-math3\3.5\commons-math3-3.5.jar
- org\apache\commons\commons-dbc2\2.1.1\commons-dbc2-2.1.1.jar
- org\apache\commons\commons-pool2\2.4.2\commons-pool2-2.4.2.jar
- commons-logging\commons-logging\1.2\commons-logging-1.2.jar
- jsi-1.0.0-javadoc.jar
- jsi-1.0.0-sources.jar
- jsi-1.0.0.jar
- slf4j-api-1.6.3.jar
- trove4j-2.0.2.jar

## Appendix 2: List of indexes of the OCCURRENCES table

Here is the list of all the indexes created for the main OCCURRENCES table. Those indexes are used by the database engine to shorten computation time when querying the table.

```
CREATE INDEX basisofrecord_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_basisofrecord COLLATE pg_catalog."default");

CREATE INDEX class_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_class COLLATE pg_catalog."default");

CREATE INDEX computedhabitat_OCCURRENCES_idx ON public.OCCURRENCES USING btree (computedhabitat COLLATE pg_catalog."default");

CREATE INDEX maille10_OCCURRENCES_idx ON public.OCCURRENCES USING btree (x, y);

CREATE INDEX o_classkey_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_classkey COLLATE pg_catalog."default");

CREATE INDEX o_date_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_date COLLATE pg_catalog."default");

CREATE INDEX o_decimallatitude_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_decimallatitude);

CREATE INDEX o_genus_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_genus COLLATE pg_catalog."default");

CREATE INDEX o_genuskey_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_genuskey COLLATE pg_catalog."default");

CREATE INDEX o_grid_species_query_OCCURRENCES_idx ON public.OCCURRENCES USING btree (x, y, o_specieskey COLLATE pg_catalog."default", o_taxonrank COLLATE pg_catalog."default");

CREATE INDEX o_month_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_month COLLATE pg_catalog."default");

CREATE INDEX o_specieskey_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_specieskey COLLATE pg_catalog."default");

CREATE INDEX o_taxonrank_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_taxonrank COLLATE pg_catalog."default");

CREATE INDEX o_year_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_year COLLATE pg_catalog."default");

CREATE INDEX suspectedcaptive_OCCURRENCES_idx ON public.OCCURRENCES USING btree (suspectedcaptive);

CREATE INDEX taxonrank_OCCURRENCES_idx ON public.OCCURRENCES USING btree (o_taxonrank COLLATE pg_catalog."default");
```

## Appendix 3: List of additional database tables

This appendix regroups the list of all the additional tables created to study the structure of the GBIF mediated data. Many of those tables have redundancy because they were created as and when needed. For Each table a brief description, as well as the query used to create it is included. Many of those tables also had indexes to speed up searches but I didn't include the indexing queries as they are trivial and non-informative.

All the columns names start with "o\_" because some Darwin Core terms were identical to reserved SQL terms and couldn't be used as columns names. Adding a prefix to all column names was the simplest and fastest solution.

### **OCCURRENCES\_list\_species**

This table lists all the species in the OCCURRENCES table and counts their number of terrestrial and marine occurrences

```
CREATE TABLE OCCURRENCES_list_species
AS SELECT
    o_specieskey, o_species, o_class, o_classkey,
    o_order, o_orderkey, o_family, o_familykey, o_genus, o_genuskey,
    sum(case computedhabitat when 'LAND' then 1 else 0 end) as nb_land,
    sum(case computedhabitat when 'OCEAN' then 1 else 0 end) as nb_ocean
FROM public.OCCURRENCES
GROUP BY
    o_specieskey, o_species, o_class, o_classkey,
    o_order, o_orderkey, o_family, o_familykey, o_genus, o_genuskey
```

### **OCCURRENCES\_stat\_species**

This table lists all the species in the OCCURRENCES table and counts their total number of occurrences.

```
CREATE TABLE OCCURRENCES_stat_species AS
SELECT
    o_species, o_specieskey,
    o_genus, o_genuskey,
    o_family, o_familykey,
    o_order, o_orderkey,
    o_class, o_classkey,
COUNT(*) as nb_occ
FROM public.OCCURRENCES
GROUP BY
    o_species, o_specieskey,
    o_genus, o_genuskey,
    o_family, o_familykey,
    o_order, o_orderkey,
    o_class, o_classkey
```

## **OCCURRENCES\_dates\_orders**

This table lists all the orders in the OCCURRENCES table and counts their total number of occurrences per Year of collect. There is one row per order.

```
CREATE TABLE OCCURRENCES_dates_orders
AS SELECT o_year,
         o_kingdomkey, o_kingdom,
         o_phylumkey, o_phylum,
         o_classkey, o_class,
         o_orderkey, o_order,
         o_basisofrecord,
count(*) as nb_occ
FROM public.OCCURRENCES
GROUP BY o_year,
         o_kingdomkey, o_kingdom,
         o_phylumkey, o_phylum,
         o_classkey, o_class,
         o_orderkey, o_order,
         o_basisofrecord;
```

## **OCCURRENCES\_species\_list\_temporal**

This table lists all the orders in the OCCURRENCES table and finds the description date of the species using its scientific name.

```
CREATE TABLE OCCURRENCES_species_list_temporal
AS SELECT
         o_specieskey, o_year,
         o_scientificname,
         o_species,
         o_class,
         o_classkey,
         o_order,
         o_orderkey,
substring(o_scientificname from '%#[0-9]{4}#%' for '#') as desc_date
FROM public.OCCURRENCES
WHERE o_taxonrank = 'SPECIES'
GROUP BY
         o_specieskey,
         o_year,
         o_scientificname,
         o_species,
         o_class,
         o_classkey,
         o_order,
         o_orderkey;
```

## **OCCURRENCES\_species\_list\_temporal\_merged**

This table lists all the species in the OCCURRENCES table as well as the description date and the year of collect. There is one row per species per year of collect.

Merges the species with different scientific names who are the same (synonyms)

```

CREATE TABLE OCCURRENCES_species_list_temporal_merged
AS SELECT
o_specieskey,
o_year,
o_species,
o_class,
o_classkey,
o_order,
o_orderkey,
MIN(desc_date::int) as desc_date
FROM public.OCCURRENCES_species_list_temporal
GROUP BY
o_specieskey,
o_year,
o_species,
o_class,
o_classkey,
o_order,
o_orderkey;

```

### **OCCURRENCES\_species\_list\_temporal\_stats**

This table lists all the species in the OCCURRENCES table and uses OCCURRENCES\_species\_list\_temporal\_merged to compute 6 statistics for each species: the description date, the number of year for which there has been an occurrence event, the year of the first occurrence, the year of the last occurrence, the number of year between first observation and description (can be positive or negative) the number of years since the description (in 2017).

```

CREATE TABLE OCCURRENCES_species_list_temporal_stats
AS SELECT
MIN(desc_date::int) as desc_date,
count(o_year) as nb_year,
MIN(o_year::int) as first_data_year,
MAX(o_year::int) as last_data_year,
(MIN(o_year::int)) - (MIN(desc_date::int)) as time_desc_to_data,
2017-(MIN(desc_date::int)) AS time_since_desc,
o_species,
o_class,
o_order
FROM public.OCCURRENCES_species_list_temporal_merged
WHERE o_year IS NOT NULL
AND desc_date IS NOT NULL
GROUP BY
o_species,
o_class,
o_order;

```

### **OCCURRENCES\_reccords\_stats**

This table lists all the classes in the OCCURRENCES table and computes 3 statistics for each class per year: the number of people and organization producing occurrences (creator and recordedby), the number of country where there was occurrences and the number of

datasets (id, key and names). All those statistics are computed for each year meaning the number of rows is the number of unique class-year pairs.

```
CREATE TABLE OCCURRENCES_reccords_stats
AS SELECT
o_year, o_class,
COUNT (DISTINCT(o_creator)) AS o_creator_nb,
COUNT (DISTINCT(o_countrycode)) AS o_countrycode_nb,
COUNT (DISTINCT(o_datasetid)) AS o_datasetid_nb,
COUNT (DISTINCT(o_datasetkey)) AS o_datasetkey_nb,
COUNT (DISTINCT(o_datasetname)) AS o_datasetname_nb,
COUNT (DISTINCT(o_recordedby)) AS o_recordedby_nb
FROM public.OCCURRENCES
GROUP BY o_year, o_class;
```

### **OCCURRENCES\_stat\_species\_no\_spatial\_duplicate**

This table lists all the species in the OCCURRENCES table and counts all the distinct cells where the species is referenced. This number was used as the number of spatially distinct occurrences for each species.

```
CREATE TABLE OCCURRENCES_stat_species_no_spatial_duplicate AS
SELECT
o_species, o_specieskey,
o_genus, o_genuskey,
o_family, o_familykey,
o_order, o_orderkey,
o_class, o_classkey,
COUNT(DISTINCT(x,y)) as nb_occ
FROM OCCURRENCES
WHERE x IS NOT NULL
AND y IS NOT NULL
GROUP BY
o_species, o_specieskey,
o_genus, o_genuskey,
o_family, o_familykey,
o_order, o_orderkey,
o_class, o_classkey;
```

### **OCCURRENCES\_species\_list\_basisofrecord**

This table lists all the species in the OCCURRENCES table and count the number of occurrence per type of origin. The origin or basis of an occurrence is stored in the basisofrecord column. There are 9 categories for a species occurrence to fall in: MACHINE\_OBSERVATION, FOSSIL\_SPECIMEN, PRESERVED\_SPECIMEN, MATERIAL\_SAMPLE, LIVING\_SPECIMEN, HUMAN\_OBSERVATION, LITERATURE, OBSERVATION and UNKNOWN.

```
CREATE TABLE OCCURRENCES_species_list_basisofrecord
AS SELECT
o_specieskey, o_species,
o_class, o_classkey,
```



```
o_order, o_orderkey,  
sum(case o_basisofrecord when 'MACHINE_OBSERVATION' then 1 else 0 end) as machine_observation,  
sum(case o_basisofrecord when 'FOSSIL_SPECIMEN' then 1 else 0 end) as fossil_specimen,  
sum(case o_basisofrecord when 'PRESERVED_SPECIMEN' then 1 else 0 end) as preserved_specimen,  
sum(case o_basisofrecord when 'MATERIAL_SAMPLE' then 1 else 0 end) as material_sample,  
sum(case o_basisofrecord when 'LIVING_SPECIMEN' then 1 else 0 end) as living_specimen,  
sum(case o_basisofrecord when 'HUMAN_OBSERVATION' then 1 else 0 end) as human_observation,  
sum(case o_basisofrecord when 'LITERATURE' then 1 else 0 end) as literature,  
sum(case o_basisofrecord when 'OBSERVATION' then 1 else 0 end) as observation,  
sum(case o_basisofrecord when 'UNKNOWN' then 1 else 0 end) as unknown  
FROM OCCURRENCES  
GROUP BY o_specieskey, o_species, o_class, o_classkey, o_order, o_orderkey;
```

## Appendix 4: VBA script for Web Search Results

This appendix show the Visual basic code used to count the number of web pages found for a species using the Bing search engine. This code uses a list of species name in the first column as the input and will output the number of results in the second column. This code also worked with the Google search engine for a short time but was then detected as spamming and couldn't run with this search engine.

```
Sub XMLHTTP_bing()  
  
    Dim url As String, lastRow As Long  
    Dim XMLHTTP As Object, html As Object, objResultDiv As Object, objH3 As Object, link As  
Object  
    Dim start_time As Date  
    Dim end_time As Date  
    'counts the number of row in column A  
    lastRow = Range("A" and Rows.Count).End(xlUp).Row  
    Dim cookie As String  
    Dim result_cookie As String  
    'record the starting time  
    start_time = Time  
    Debug.Print "start_time:" and start_time  
    'Loop from the first row to the last  
    For i = 1 To lastRow  
  
        'Using the web engine URL we can build a query for this engine to search the term  
located in the cell (A,i)  
        'url = "https://www.bing.com/search?q=%2B" and Cells(i, 1) and "andgo=Valider"  
        'The previous query has been modified to also include the word "species" and elimitate  
mani false positive  
        url = "https://www.bing.com/search?q=%2B" and Cells(i, 1) and  
"+%2Bspeciesandgo=Valider"  
        'Construction and sending of the HTTP request  
        Set XMLHTTP = CreateObject("MSXML2.serverXMLHTTP")  
        XMLHTTP.Open "GET", url, False  
        XMLHTTP.setRequestHeader "Content-Type", "text/xml"  
        XMLHTTP.setRequestHeader "User-Agent", "Mozilla/5.0 (Windows NT 6.1; WOW64  
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.102 Safari/537.36"  
        XMLHTTP.setRequestHeader "X-Client-Data", "CIa2yQEIorbJAQjEtskBCP2VygEI7ZzKAQ=="  
        XMLHTTP.send  
  
        'The result returned by the http request are put in a html object  
        Set html = CreateObject("htmlfile")  
        html.body.innerHTML = XMLHTTP.ResponseText  
        'The code search for the element of the file containing the number of results  
        Set objResultDiv = html.getelementbyid("b_tween")  
        'If this element is empty then we put 0 as a result in the cell (B,i)  
        If objResultDiv Is Nothing Then  
            Cells(i, 2) = 0  
        'If the Element is not emty we put its contents into the cell (B,i)  
        Else  
            Cells(i, 2) = objResultDiv.FirstChild.FirstChild.NodeValue  
        End If  
        DoEvents  
    Next  
    'After all rows have been processed the code output the duration of execution  
    end_time = Time  
    Debug.Print "end_time:" and end_time  
    Debug.Print "done" and "Time taken : " and DateDiff("n", start_time, end_time)  
    MsgBox "done" and "Time taken : " and DateDiff("n", start_time, end_time)  
End Sub
```

## Appendix 5: R script for spatial outlier detection

This appendix show the R code used to detect spatial outliers in a species' occurrences. This code needs at least 6 occurrences to work and calculate the orthodromic distance between one occurrence and the 5 nearest to it. If the mean of this distance is in the last centile of all the distance the point is considered an outlier. The script was written by Nicolas Lebbe.

```
# \file script.R
library("proj4")

# On récupère les coordonnées de nos points
# en eckert IV et on les convertis en latitude/longitude
xy = cbind(xEckert, yEckert)
coords = project(xy, "+proj=eck4", inverse = TRUE)

# on récupère les latitudes et longitudes
# des points que l'on converti en radian
x = coords[,1]/180*pi
y = coords[,2]/180*pi
n = length(x)
# rayon de la terre pour obtenir distance orthodromique en mètre
R = 6374892.5

dist = rep(0, n)
iter = seq(1, n)
k = 5
for(i in iter) {
  others = (iter != i)
  #list = (x[others] - x[i])^2 + (y[others] - y[i])^2
  # distance orthodromique
  list = R*acos(sin(x[others])*sin(x[i]) +
               cos(x[others])*cos(x[i])*cos(y[others] - y[i]))
  # trouve la valeur du maximum des k plus petits
  max_min = quantile(list, k/(n-1))
  # extrait les k plus petites valeurs
  kmins = list[list <= max_min]
  # mesure "d'excentricité"
  dist[i] = mean(kmins)
}

q = 99##%
# on extrait ceux qui sont dans le quantile à q%
#ok = dist < quantile(dist, q/100)
outlier = dist >= quantile(dist, q/100)

outlier = as.numeric(outlier)

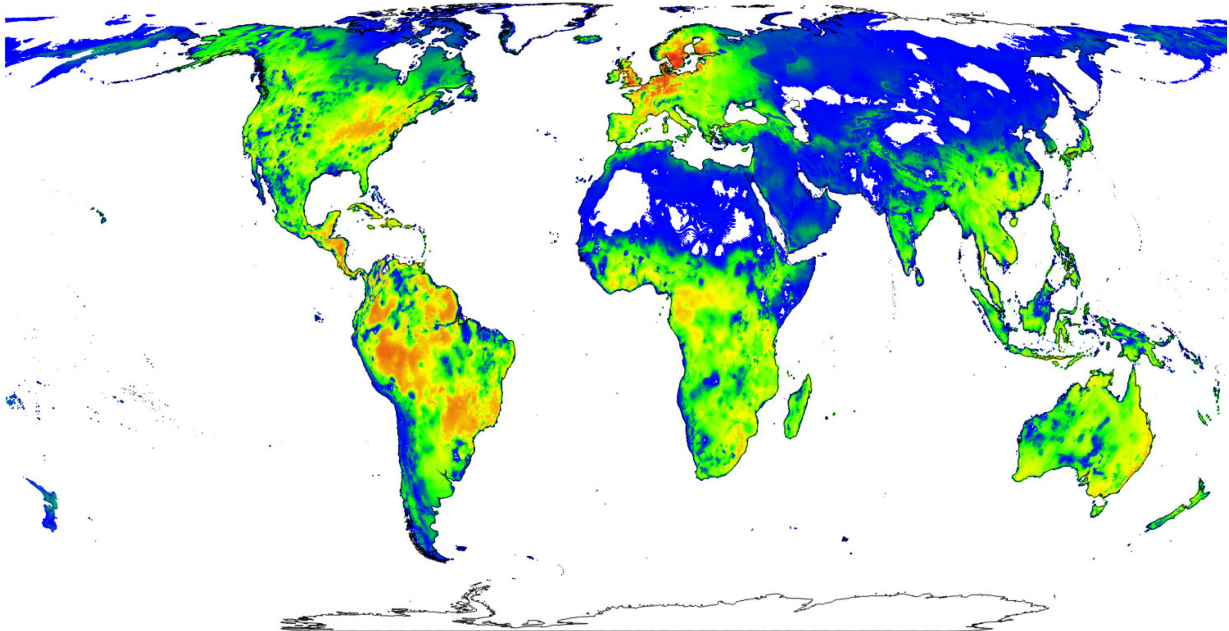
# affichage avec projection de Robinson conservant les distances
# library("proj4")
# xy = cbind(data["coords_x"], data["coords_y"])
# coords = project(xy, "+proj=robin +lon_0=90w")
# plot(coords)
# points(coords$x[!ok], coords$y[!ok], pch=22, col="red", bg="red")

# pour abscisse et longitude pour ordonnée
# plot(x, y)
# points(x[!ok], y[!ok], pch=22, col="red", bg="red")

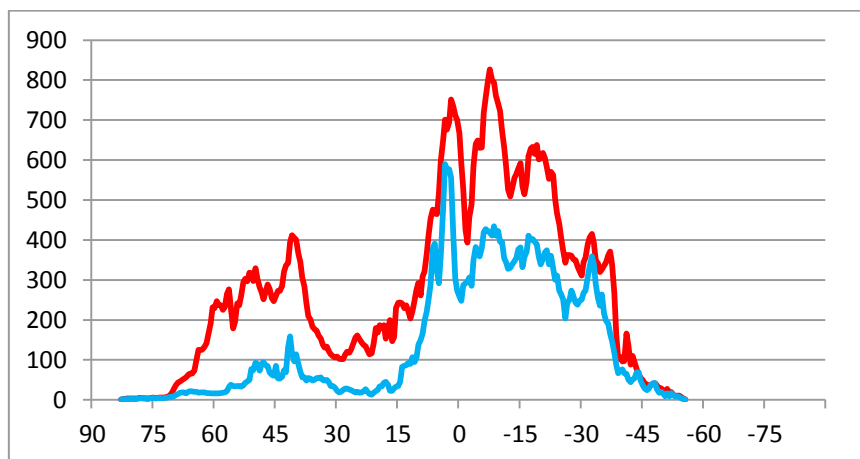
# création nouvelle table sans ceux qui sont "loin des autres"
# write.table(subset(data, ok), file="Sorex_cinereus_ok.csv", sep=";")
```

## Appendix 6: Worldwide species richness maps and plots

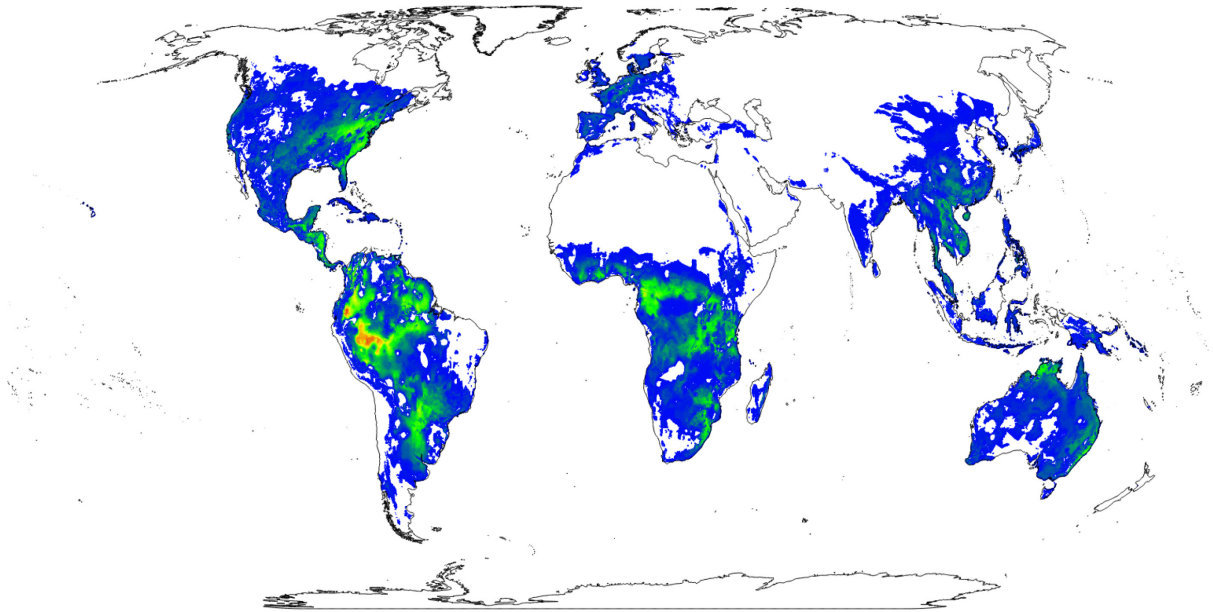
This appendix contains all the global species richness maps created with the methods exposed in the chapter 1.



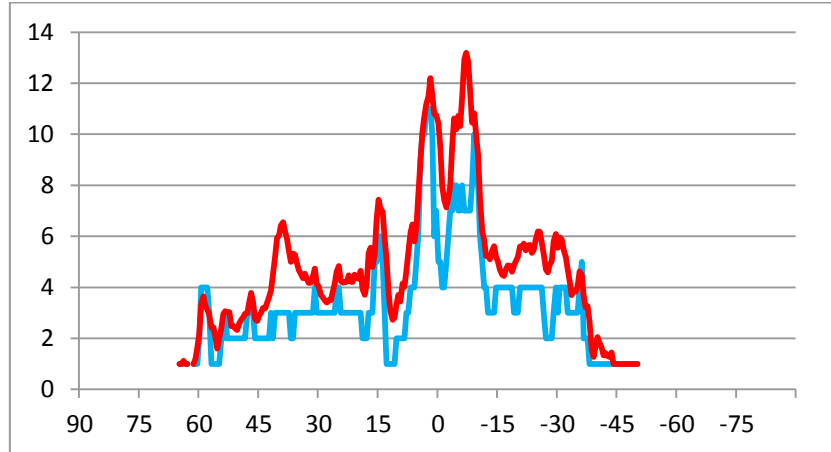
**Map GBIF:** Species richness including all the GBIF species with at least 20 cells (no taxonomic filtering). The hotter colours indicate higher species richness.



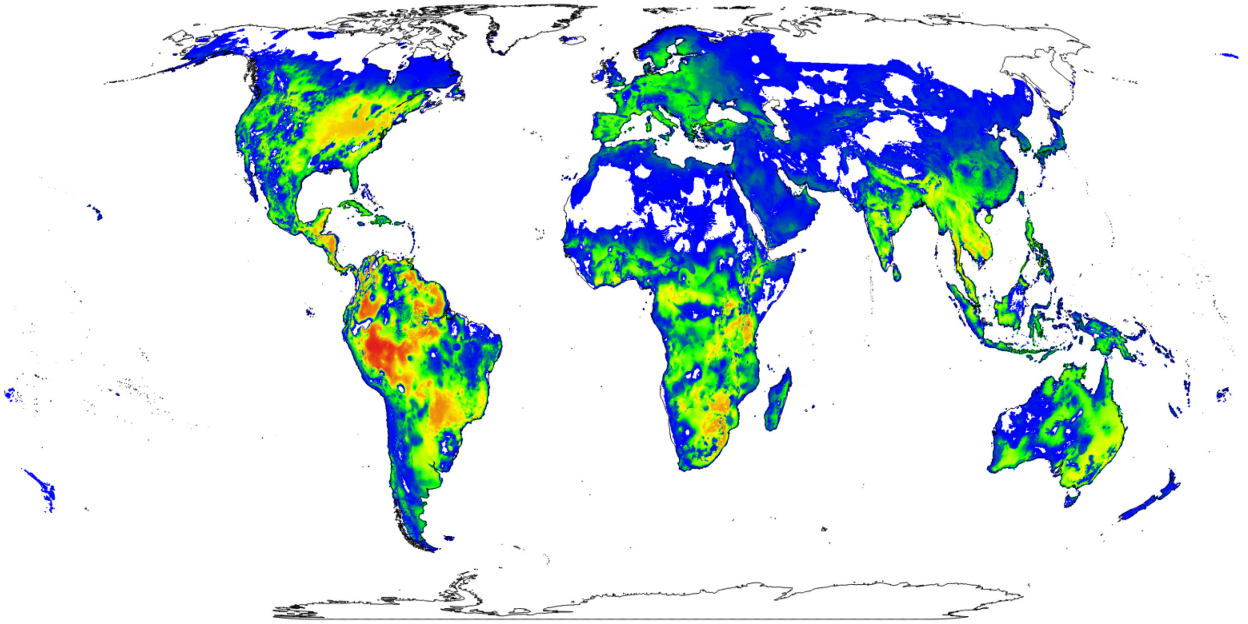
**Plot GBIF:** Species richness per latitude including all the GBIF species with at least 20 cells (no taxonomic filtering). The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



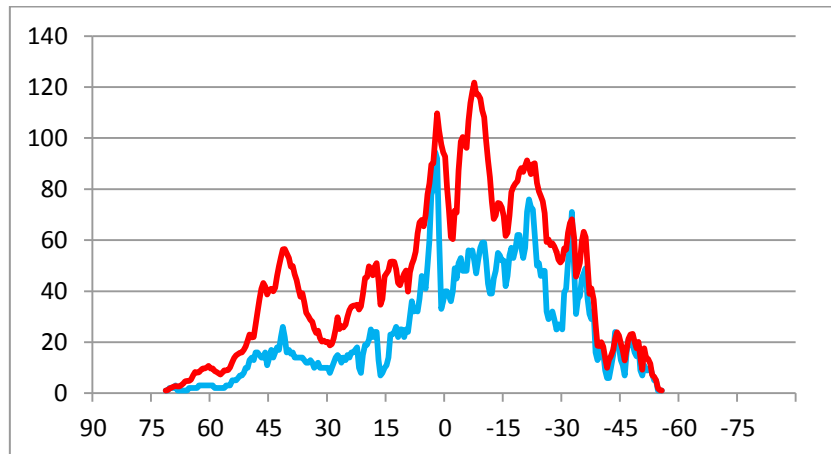
**Map Amphibia:** Species richness including all the GBIF **Amphibia** species with at least 20 cells. The hotter colours indicate higher species richness.



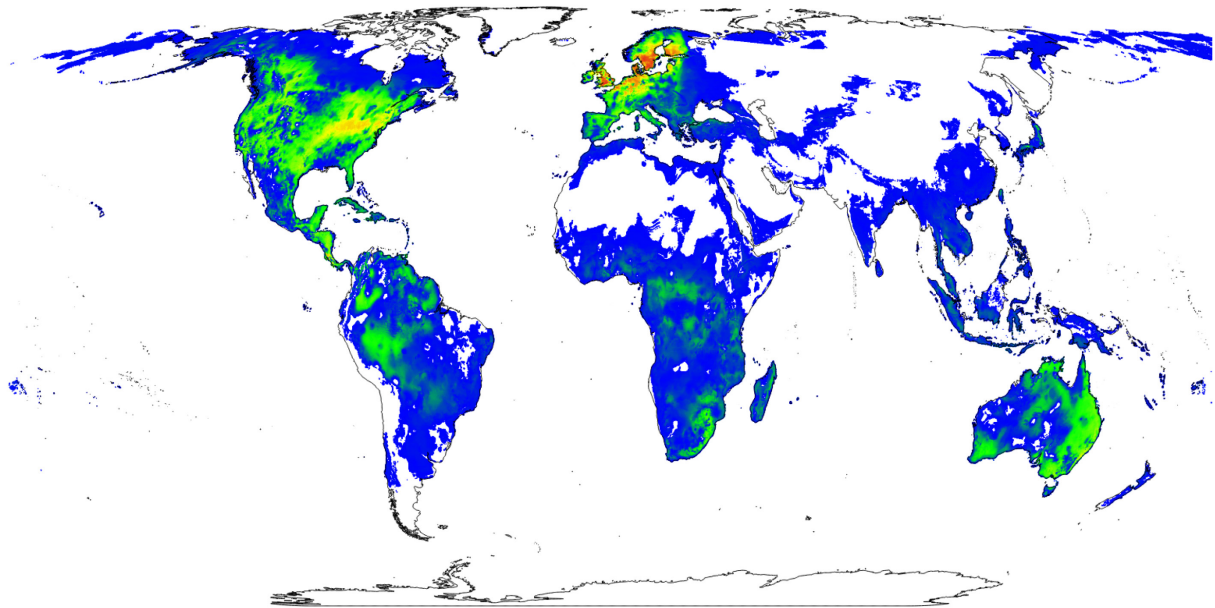
**Plot Amphibia:** Species richness per latitude including all the GBIF **Amphibia** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



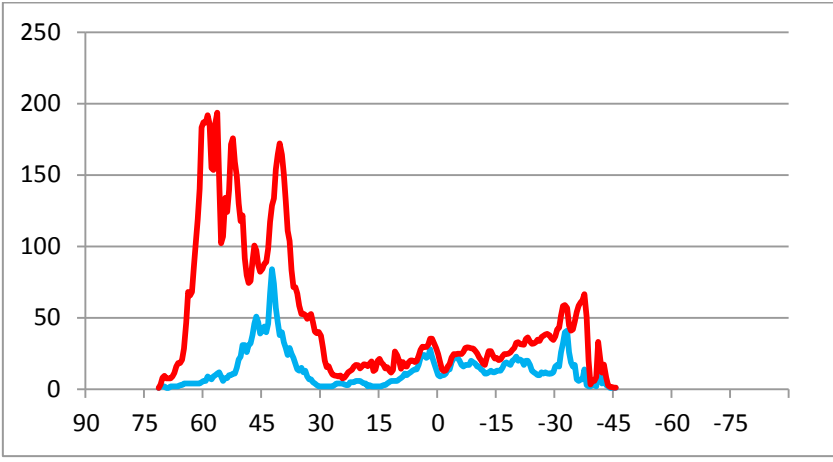
**Map Aves:** Species richness including all the GBIF **Aves** species with at least 20 cells. The hotter colours indicate higher species richness.



**Plot Aves:** Species richness per latitude including all the GBIF **Aves** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.

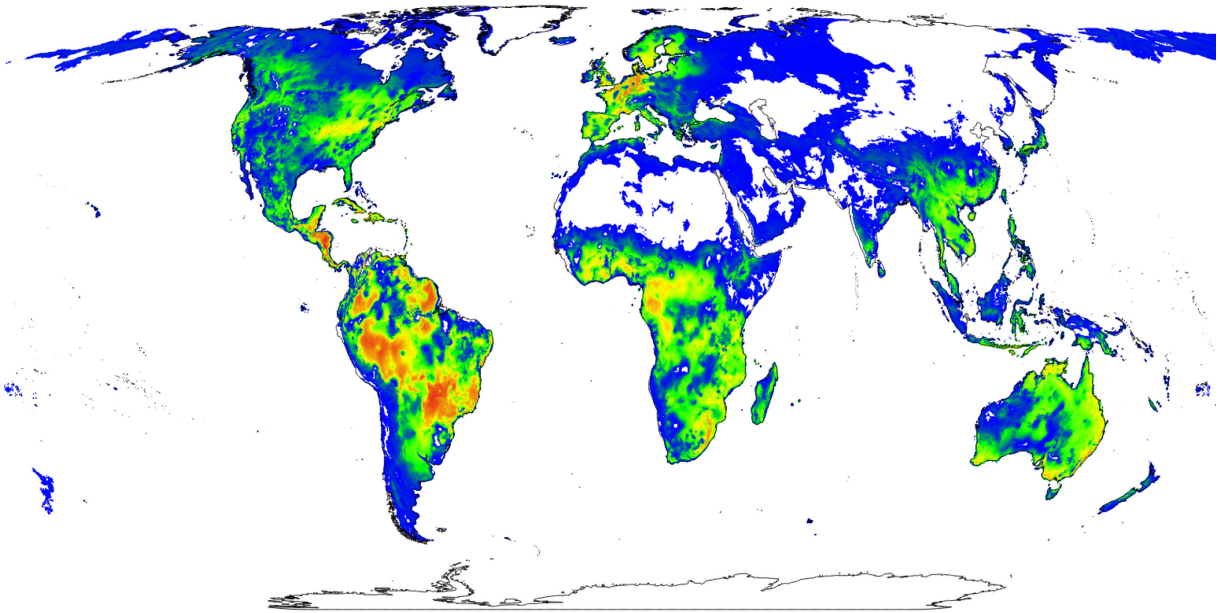


**Map Insecta:** Species richness including all the GBIF **Insecta** species with at least 20 cells. The hotter colours indicate higher species richness.

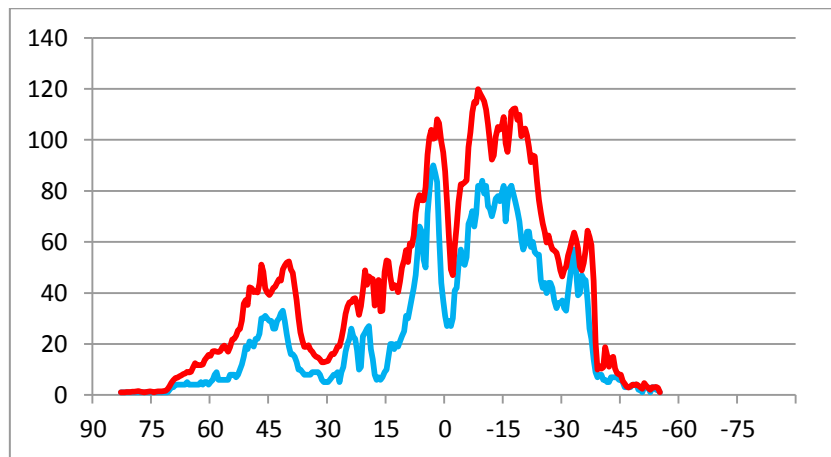


**Plot Insecta:** Species richness per latitude including all the GBIF **Insecta** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



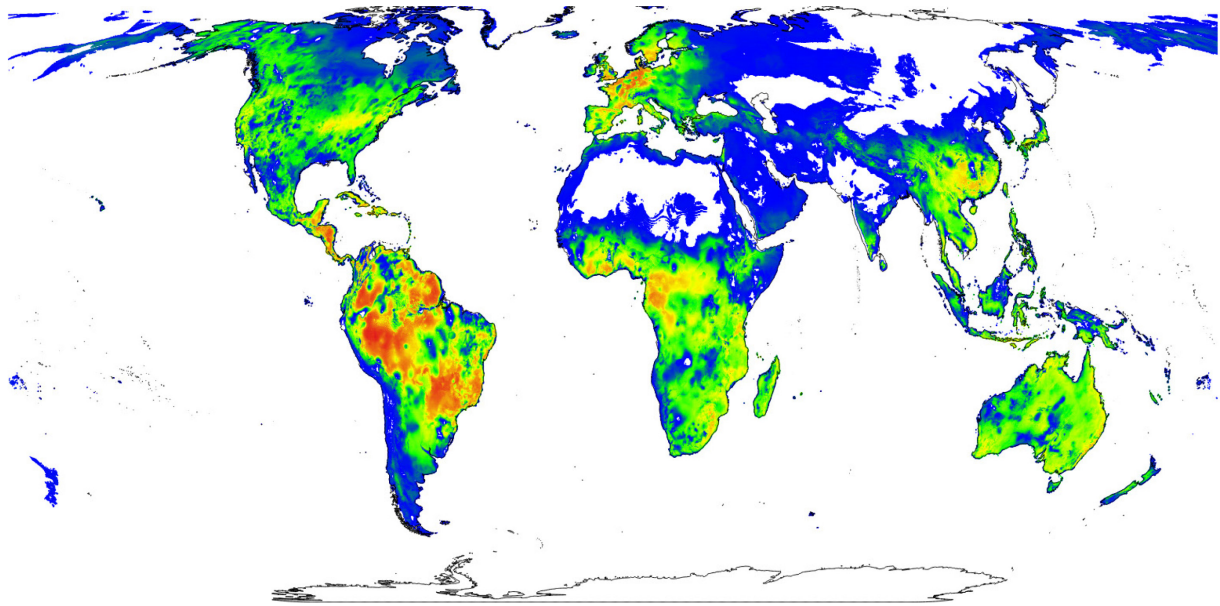


**Map Liliopsida:** Species richness including all the GBIF **Liliopsida** species with at least 20 cells. The hotter colours indicate higher species richness.

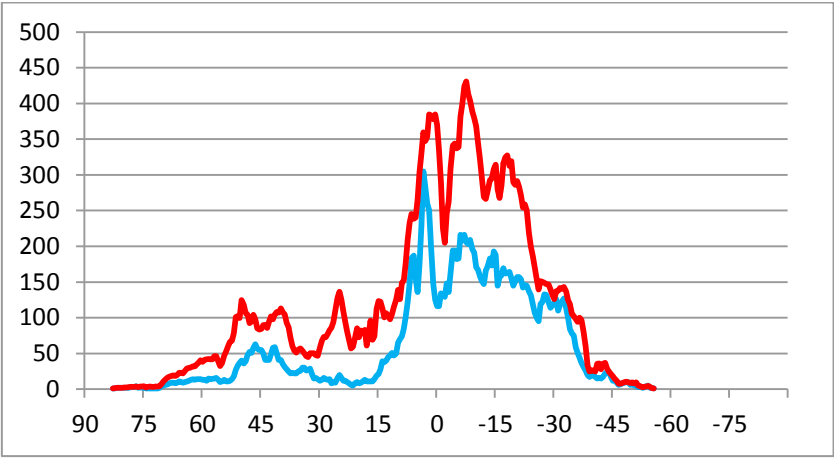


**Plot Liliopsida:** Species richness per latitude including all the GBIF **Liliopsida** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.

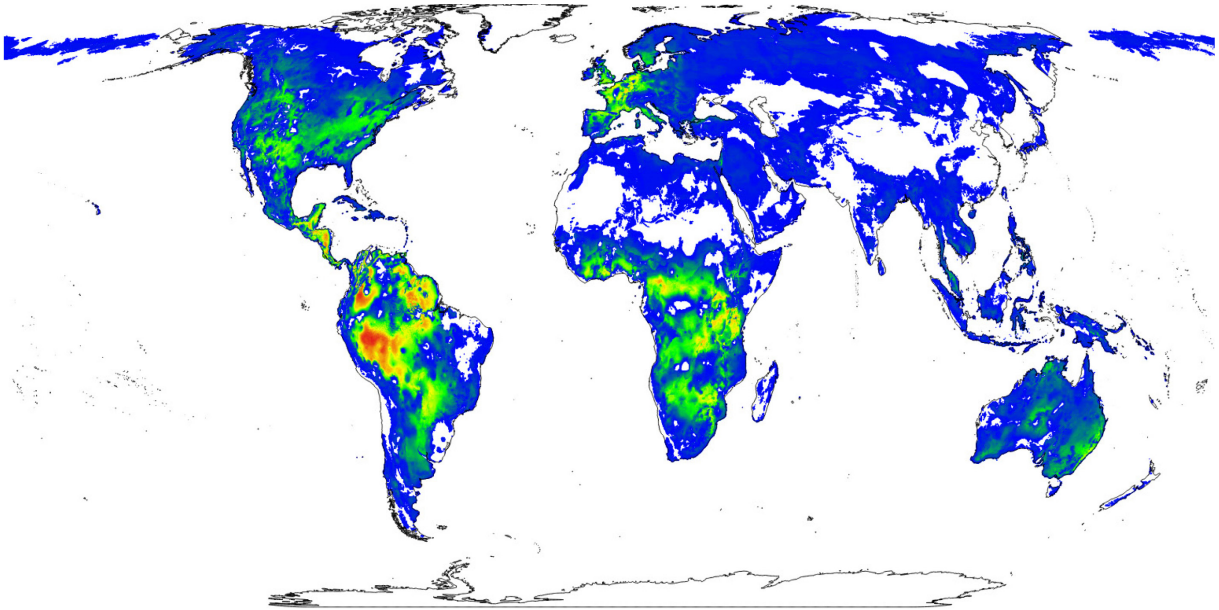




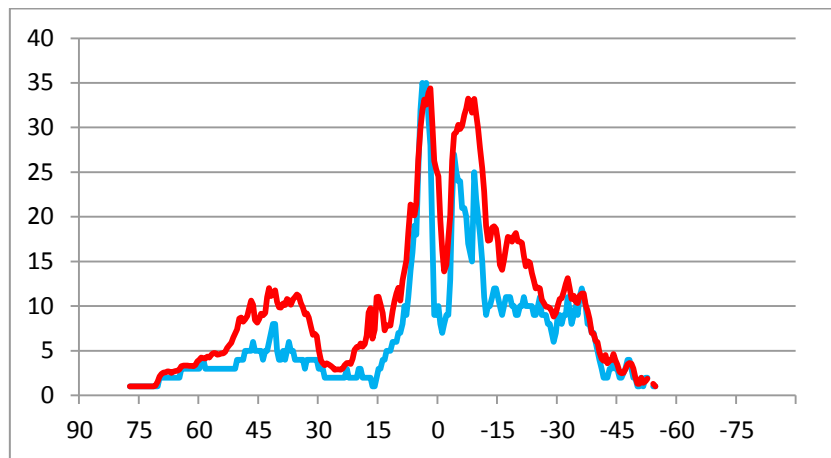
**Map Magnoliopsida:** Species richness including all the GBIF **Magnoliopsida** species with at least 20 cells. The hotter colours indicate higher species richness.



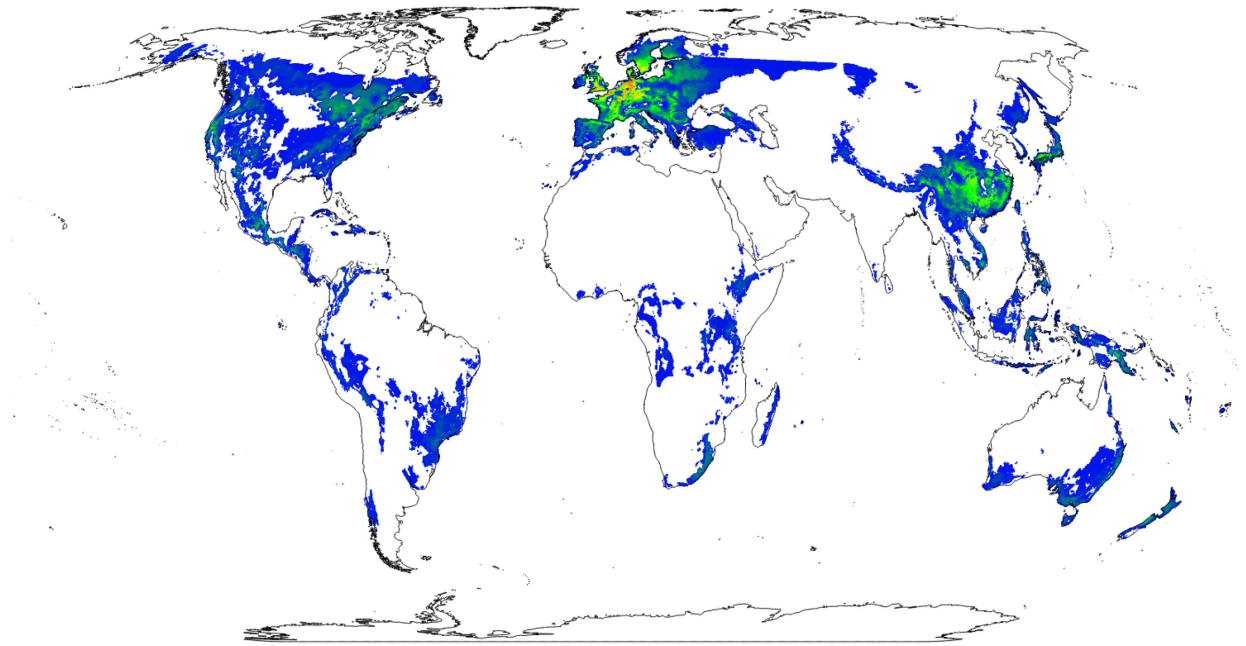
**Plot Magnoliopsida:** Species richness per latitude including all the GBIF **Magnoliopsida** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



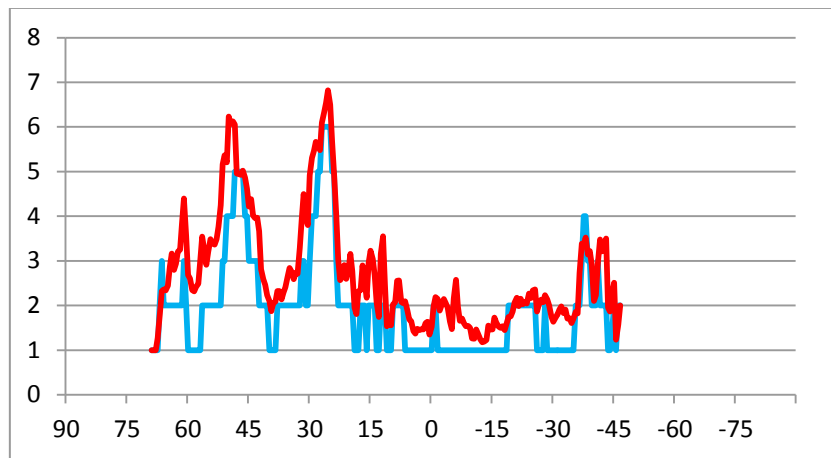
**Map Mammalia:** Species richness including all the GBIF **Mammalia** species with at least 20 cells. The hotter colours indicate higher species richness.



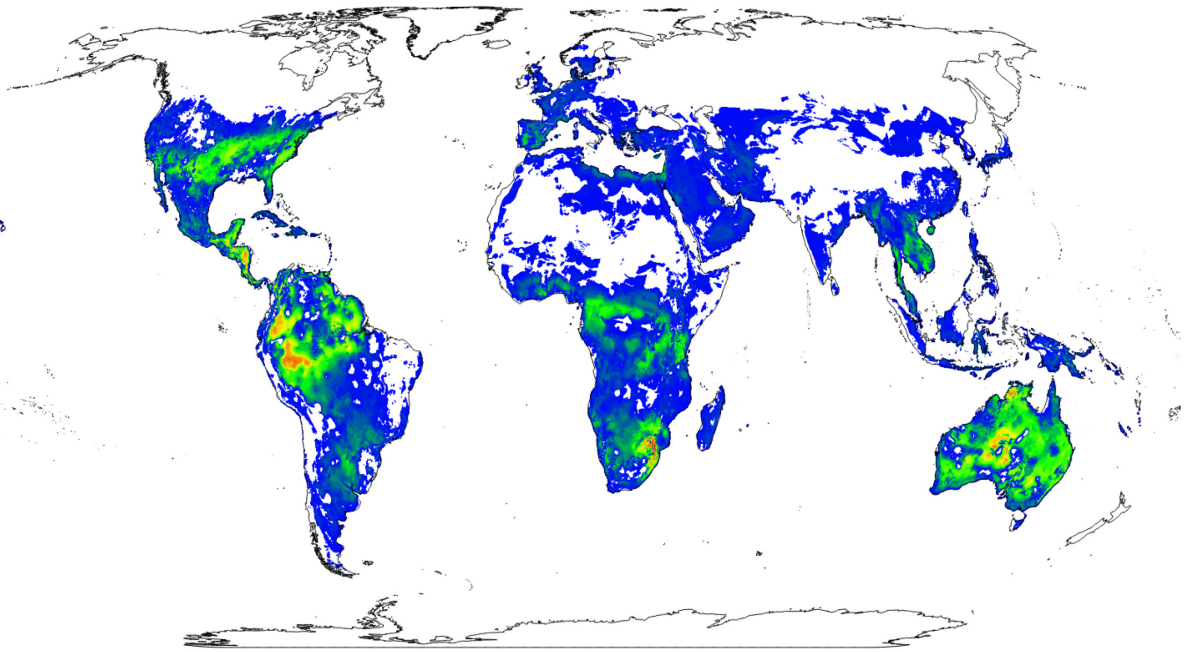
**Plot Mammalia:** Species richness per latitude including all the GBIF **Mammalia** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



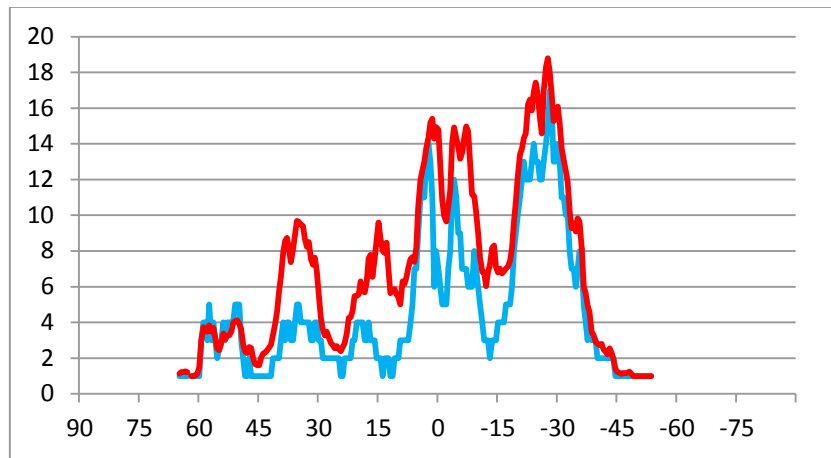
**Map Pinopsida:** Species richness including all the GBIF **Pinopsida** species with at least 20 cells. The hotter colours indicate higher species richness.



**Plot Pinopsida:** Species richness per latitude including all the GBIF **Pinopsida** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.



**Map Reptilia:** Species richness including all the GBIF **Reptilia** species with at least 20 cells. The hotter colours indicate higher species richness.



**Plot Reptilia:** Species richness per latitude including all the GBIF **Reptilia** species with at least 20 cells. The red line indicates the mean species richness per 100 km<sup>2</sup> and the blue curve indicates the median species richness per 100 km<sup>2</sup>.

## List of figures:

Figure 1: Growth of available species occurrences in the GBIF from 2008 to 2017.....	19
Figure 2: Extract from Views of nature, or, Contemplations on the sublime phenomena of creation.....	22
Figure 3: Global workflow organization.....	29
Figure 4: First computation step of the application.....	31
Figure 5: Screenshot of the results obtained when searching a species in Bing.....	36
Figure 6: Plot of the deficit and excess in occurrences per class in the GBIF dataset.....	39
Figure 7: Proportions of the 1,370,170 species categorized by their number of occurrences in the GBIF mediated data.....	42
Figure 8: Global repartition of the common black ant ( <i>Lasius niger</i> ) occurrences.....	43
Figure 9: Visualization of the outlier occurrences for the black rat ( <i>Rattus rattus</i> ).....	45
Figure 10: Schema of the architecture used to work with R and Java.....	46
Figure 11: Accumulation curve of species discovered depending on the sampling effort.....	49
Figure 12: World map of the Tetrapoda species richness (Reptilia, Amphibia, Aves and Mammalia) obtained using the tools I created.....	53
Figure 13: Map of the Vertebrates diversity proposed by Mannion <i>et al.</i> (2014).....	53
Figure 14: Illustrations of observation-based and specimen-based primary biodiversity occurrences and their potential uses.....	63
Figure 15: Number of primary biodiversity occurrences per year and origin from 1900 to today.....	69
Figure 16: Proportion of occurrences per year of collect and origin for a particular class.....	70
Figure 17: The increase of ancillary data to biodiversity occurrences does not keep pace with biodiversity data accumulation.....	74
Figure 18: Occurrences with ancillary data are mainly specimen occurrences.....	74
Figure 19: a) Spatial and b) taxonomic precision in the GBIF mediated data improves over time in proportion.....	76

Figure 20. Taxonomic bias in biodiversity occurrence data. ....	85
Figure 21: Taxonomic bias in biodiversity data occurrences.....	86
Figure 22. Evolution over time of the taxonomic bias for each class. ....	89
Figure 23. Biodiversity occurrences recorded in GBIF between 1900 and 2006. ....	90
Figure 24. Biodiversity occurrences recorded in the GBIF between 1900 and 2006. ....	91
Figure 25. Taxonomic heterogeneity in sampling, occurrence data origin and quality for 24 taxonomic classes.....	92
Figure 26. Relation between age, origin and quality of the occurrence data for 24 taxonomic classes.....	96
Figure 27. Relation between the number of Google search results and Bing search results for 4000 random species. ....	105
Figure 28: For 7 out of 8 tested classes the Latitudinal Diversity Gradient is clearly visible. ....	116
Figure 29: GWR results for Mammalia and 10,000 occurrences, projected as a map.....	119
Figure 30: Insecta species density across the globe according to raw GBIF-mediated data shows a sampling bias .....	133
Figure 31: Accumulation curve of the number of occurrences available in the GBIF. ....	135
Figure 32: Average completeness of the GBIF mediated data per year does not evolve along time.....	137
Figure 33: Filtering occurrences in the GBIF .....	140
Figure 34: Proportion of occurrences per year of collect and origin cumulated for 24 classes .....	143
Figure 35: Proportion of the number of occurrences per region of the world in the GBIF mediated data.....	146
Figure 36: The global repartition of GBIF-mediated occurrences is uneven.....	146
Figure 37. The current knowledge on eukaryotic species diversity is incomplete and biased. ....	147
Figure 38: Discrepancy between the proportion of occurrences per class in the GBIF mediated data (left) and the proportion of species per class (right).....	148

Figure 39: The median number of occurrences per species in the GBIF-mediated data differs according to taxonomic classes. ....	148
Figure 40: Repartition of the Yellow Ant ( <i>Lasius flavus</i> ) occurrences across Europe in the GBIF dataset.....	154
Figure 41: Different effects of the spatial bias depending on the taxa.....	157
Figure 42: Species richness patterns derived from the GBIF-mediated for 6 orders.....	158

## List of tables:

Table 1: First lines of the over_under_sampled.csv file obtained after querying the database for the number of species and the number of occurrences in each class.....	38
Table 2. Biodiversity occurrence data statistics for 24 taxonomic classes. ....	83
Table 3. Biodiversity occurrence data statistics for the orders (maximum 10) with the most occurrences within eight selected classes.....	95
Table 4. GLM results assessing the link between research quantity, public interest and their combined interaction on the amount of biodiversity data per class. ....	97
Table 5: Results of the regression models taking into account spatial autocorrelation or not. ....	118
Table 6: Top 10 list of species with the most data and only observation occurrences in the GBIF.....	144
Table 7: Summary of the effect of different “internal” taxon characteristics. ....	151
Table 8: Top 10 list of species with the most occurrences in the GBIF. ....	152



## Occurrences et patrons de biodiversités sous l'œil de la systématique

### Résumé :

Dans le contexte actuel de crise de biodiversité, il est primordial de comprendre où et comment se distribuent les êtres vivants. En utilisant les données de biodiversité gérées par le GBIF (> 640 millions d'occurrences) et couvrant 24 classes taxonomiques, j'ai étudié un patron de biodiversité remarquable et qui se caractérise par une augmentation de la richesse spécifique lorsque l'on se rapproche de l'équateur : le gradient latitudinal de diversité (LDG). Cet objectif m'a d'abord amené à produire des outils informatiques afin de manipuler ces données massives de biodiversité (paradigme du Big Data), puis à évaluer la qualité des données primaires de biodiversité. J'ai alors mis en évidence deux phénomènes importants. Premièrement, un fort biais taxonomique existe dans les données d'occurrences de biodiversité. Certains taxons sont plus étudiés que d'autres, créant un déficit de connaissance pour certains groupes et se révélant problématique pour notre compréhension de la biodiversité dans son ensemble. Ce biais semble s'expliquer par l'impact des préférences sociétales plutôt que par l'activité de recherche scientifique. Deuxièmement, un changement radical dans les pratiques de collecte de ces données se produit : de plus en plus de données primaires de biodiversité sont de simples observations et non plus des spécimens récoltés et mis en collection. Les dangers et avantages liés à ce changement de pratique sont discutés, le rôle de spécimens vouchers est rappelé et, en l'absence de spécimens, la nécessité d'acquérir des données supplémentaires est soulignée. Enfin, fort de cette analyse critique des données primaires de biodiversité, six hypothèses pouvant expliquer le LDG sont testées sur un jeu de données nettoyées couvrant huit classes taxonomiques. Ce test permet de réfuter une hypothèse de contrainte géométrique récente mais jamais testée pour finalement révéler que l'hypothèse de productivité est la mieux soutenue.

Mots clés : [bioinformatique, données primaires de biodiversité, biais taxonomique, spécimens, observations, bases de données, big-data, gradient latitudinal de diversité]

### Abstract:

In the current context of biodiversity crisis, it is essential to understand where and how life is distributed. Using biodiversity data managed by the GBIF (>640 million occurrences) covering 24 taxonomic classes, I investigated one of the best-known biodiversity patterns: the Latitudinal Diversity Gradient (LDG), which is characterized by an increase in specific richness as we approach the equator. This objective first led me to produce informatics tools for handling large amount of data (Big data paradigm), before evaluating the quality of primary biodiversity data. Two important outcomes resulted from this evaluation. First, I highlight that a strong taxonomic bias exists in biodiversity occurrences. This bias implies that some taxa are more studied than others, creating a knowledge gap detrimental to our understanding of biodiversity as a whole. This bias is strongly impacted by societal preferences rather than research activity. Second, a radical change in biodiversity data gathering practices is happening: primary biodiversity data are now mostly observation-based and not specimen-based. Assets and liabilities of this shift are discussed, while the role of voucher specimens is reiterated and, for observations, the need for ancillary data is underlined. Finally, six hypotheses proposed to explain the LDG are tested on a cleaned dataset encompassing eight taxonomic classes. A recent, but never tested, version of the geometric constraint hypothesis is refuted, while the productivity hypothesis is strongly supported.

Keywords: [bioinformatics, primary biodiversity data, taxonomic bias, specimens, observations, database, big-data, latitudinal diversity gradient]