INAUGURAL – DISSERTATION

zur Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen
Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

Vorgelegt von

Diplom-Geograph Clemens Jacobs

aus: Karlsruhe

Tag der mündlichen Prüfung: 21. Mai 2019

Thema

Data Quality of Citizen Science Observations of Organisms: Plausibility Estimation Based on Volunteered Geographic Information Context

Gutachter: Prof. Dr. Alexander Zipf, Universität Heidelberg

Prof. Dr. Michael Wink, Universität Heidelberg

Acknowledgements

I would first like to thank my supervisor, Professor Alexander Zipf, who welcomed me to the GIScience research group at Heidelberg University and provided many years of invaluable scientific and professional support. He gave me the unique opportunity to combine my professional and scientific careers, and to fulfill my long-standing desire to undertake a PhD. Thank you very much!

I owe my sincere thanks to Professor Michael Wink for his willingness to review this work and to act as an examiner. Professor Ulrike Gerhard and Professor Olaf Bubenzer also agreed to examine the defense of this thesis, for which I would like to express my gratitude.

Throughout my PhD, Professor Bernhard Höfle never failed to provide me with most welcome advice. He also entrusted me with tasks and positions that have helped to broaden my perspective and to expand my capabilities, for which I offer my sincere thanks.

This work would not have been possible without the support of many people. A big thank you for proofreading, enlightening and motivating discussions, and technical support in different stages of my scientific journey goes to Dr. Sven Lautenbach, Martin Raifer, Dr. Franz-Benjamin Mocnik, Dr. René Westerholt, Dr. Johannes Lauer, Dr. Yair Grinberger, Dr. Martin Hämmerle, Dr. Michael Schultz, Michael Auer and Dr. Jack Williams. I also received help, useful hints, and valuable support in many forms from Professor Bernd Resch, Professor Joao Porto de Albuquerque, Professor Fan Hongchao, Dr. Yang Mon-Shieh (thanks for all the tea!), Fabian Kowatsch, Rafael Troilo, Lin Diao, and Dr. Wang Zhiyong. These and all the other great people in the GIScience research group at Heidelberg University created a unique atmosphere of companionship which carried me through my time as a doctoral student. Special thanks are owed to Sebastian Döring, who patiently carried the burden of sharing offices with me over several years. I know that I cost him a lot of time and nerves, bugging him with my problems and merciless conversation - I promise to make up for it somehow!

Bettina Knorr is the administrative and organizing genius at the heart of the GIScience research group. Thank you for your help in so many ways!

A Chinese proverb says that a journey of a thousand miles begins with a single step. Sometimes you need someone to help you make this first step. I therefore give very special thanks to Dr. Richard Leiner. Dear Linus, ever since my time as a student of geography at Heidelberg University, when you put a lot of effort into helping me with my Diploma thesis, you never lost sight of me. On several occasions, you provided the impulse for very important turning points in my career. Unfailingly, you also played a crucial role in creating the opportunity for me to undertake my PhD. I am deeply grateful.

I thank Rüdiger Wolff of Leiner & Wolff for his open-mindedness, his support and his good counsel on many occasions.

Annalena Schotthöfer and Dominik Frank of KoNat gave me the opportunity to discuss my ideas and results with them, and provided very valuable feedback based on their expertise and knowledge.

Finally, this work builds on the time and effort of volunteers who spend many hours of their lives on observing organisms, or on mapping our world, and who create data in projects that are available to everyone. These projects are being organized and developed by farsighted scientists, software engineers, and other great people. Thank you!

Kurzfassung

Eine wachsende Zahl von Citizen-Science-Projekten sammelt große Mengen an Naturbeobachtungsdaten von Freiwilligen. Solche Daten sind ein wichtiger Beitrag zur Biodiversitätsforschung, weil sie Informationen über die Verbreitung von Arten für große Gebiete und über lange Zeiträume hinweg zur Verfügung stellen. Sie werden in der derzeitigen weltweiten Biodiversitätskrise dringend für die Erforschung und den Schutz der Biodiversität benötigt. Eines der wichtigsten Probleme, die gelöst werden müssen, damit diese Daten tatsächlich verwendet werden können, ist ihre Datenqualität. Dieses Problem besteht v.a. bei Daten aus Zufallsbeobachtungen, die ohne strenge Vorgaben zur Sicherung bestimmter Datenqualitätsstandards vor oder während der Datensammlung entstehen. Es besteht ein großer Bedarf an methodischen Ansätzen zur automatischen Qualitätseinschätzung von Zufallsbeobachtungen von Arten, die es erlauben, die großen Datenmengen zu verarbeiten, die in den betreffenden Citizen-Science-Projekten entstehen.

Zufallsbeobachtungen von Arten aus Citizen-Science-Projekten sind biologische, aber auch geographische Daten, da sie immer einen Ortsbezug besitzen. Da sie zumeist im Internet und von nicht geschulten Freiwilligen gesammelt werden, gehören sie zu jenen geographischen Daten, die als Volunteered Geographic Information (VGI, Goodchild 2007) bezeichnet werden. Sie eignen sich daher besonders zur Anwendung geographischer Kriterien für die Entwicklung entsprechender Methoden zur Einschätzung der Qualität. Zugleich sind Naturbeobachtungen eine besondere Art von VGI, da sie zumeist keine beständigen Objekte widergeben, sondern den Charakter von einmaligen Ereignissen haben, die nicht als korrekt oder fehlerhaft erwiesen werden können. Die Qualitätsprüfung muss daher auf Annäherungen zurückgreifen, wie beispielsweise auf die Plausibilitätsprüfung mithilfe bestimmter Vergleichsinformationen.

In dieser Arbeit wurden neuartige Verfahren zur Qualitätsprüfung von Zufallsbeobachtungen von Arten entwickelt, die deren Plausibilität mithilfe ihres VGI-Kontexts einschätzen. Dabei wurden Daten aus zwei Citizen-Science-Projekten genutzt, die Zufallsbeobachtungen von Arten sammeln, und die aus zwei Untersuchungsgebieten stammen: Daten von ArtenFinder Rheinlan-Pfalz (Deutschland), und Daten des weltweiten Projekts iNaturalist aus Kalifornien (USA). In einem intrinsischen Verfahren wird dabei der geographische Kontext aus benachbarten Beobachtungen desselben Datensatzes erzeugt, indem artspezifische Beobachtungsgemeinschaften gebildet werden, die die typischerweise in der Umgebung einer bestimmten Art beobachteten Arten widergeben. Ein ebenfalls entwickeltes extrinsisches Verfahren nutzt als Quelle des geographischen Kontexts die OpenStreetMap, ein weit fortgeschrittenes VGI-Projekt, das detaillierte geographische Informationen über die physische Umwelt zur Verfügung stellt. Der geographische Kontext einer Art wird hier in Form einer OSM-Umgebung beschrieben, in der jene OSM-Elemente zusmamengefasst sind, die typischerweise in der Umgebung einer Beobachtung der betr. Art zu finden sind. Die Plausibilität einer neu hinzukommenden Beobachtung wird nun abgeschätzt, indem der Kontext dieser Beobachtung in Form benachbarter Beobachtungen, oder in Form der sie umgebenden OSM-Elemente, mit der Beobachtungsgemeinschaft oder der OSM-Umgebung der betreffenden Art abgeglichen wird. Dieser Abgleich wird mithilfe von Ähnlichkeitsindices durchgeführt.

Sowohl der intrinsische Ansatz mit Beobachtungsgemeinschaften als auch der extrinsische Ansatz mit OSM-Umgebungen wurden evaluiert, indem beide auf plausible und unplausible Beobachtungen angewendet wurden. Hierbei wurden sowohl reale, bestätigte oder verworfene Beobachtungen verwendet, als auch plausible und unplausible Beobachtungen, die für diesen Zweck eigens künstlich erzeugt wurden. Die Evaluierung konnte zeigen, dass beide Ansätze in der Lage sind, plausible von unplausiblen Beobachtungen anhand ihres VGI-Kontexts und mithilfe von Ähnlichkeitsindices zu unterscheiden. Sie schätzen dabei die Plausibilität des Meldungsortes hinsichtlich der ihn umgebenden Meldun-

gen bzw. der umgebenden OSM-Elemente sowie der vom Beobachter angegebenen Art ein. Die nähere Betrachtung der Evaluierungsergebnisse zeigte Unterschiede im Verhalten der Methoden bei Verwendung unterschiedlicher Ähnlichkeitsindices. Auch zwischen den beiden Anwendungsfällen von Beobachtungsdaten unterscheiden sich die Ergebnisse teilweise. Die ungleiche räumliche Verteilung der Beobachtungsdaten und der OSM-Daten nimmt Einfluss auf die Werte der berechneten Ähnlichkeitsindices. Beobachtungsgemeinschaften geben zumeist die biologischen und ökologischen Eigenschaften der Arten wider, zu denen sie gehören, während dies bei OSM-Umgebungen nur selten der Fall ist. Die methodischen Ansätze wurden mit einer Reihe von Paramater-Einstellungen getestet und erwiesen sich als weitgehend stabil, was ihre Funktionsweise angeht. Einige ebenfalls untersuchte methodische Erweiterungen, wie die Anwendung von Landnutzungs- oder Landbedeckungsdaten zur Schärfung des geographischen Kontexts, oder die Berücksichtigung von Beobachtungshäufigkeiten in der Ähnlichkeitsberechnung, eignen sich möglicherweise zur Verbesserung der Ergebnisse.

In der zukünftigen Forschungsarbeit müssen Lösungen für einige Schwächen und Nachteile der hier untersuchten Ansätze zur Einschätzung der Plausibilität von Zufallsbeobachtungen von Arten gefunden werden. So können die hier untersuchten Methoden nur für Arten angewendet werden, die eine ausreichende Informationsbasis in Form vorhandener Beobachtungen aufweisen, und nur für zu prüfende Meldungen, die in Gebieten mit ausreichend Umgebungsbeobachtungen oder OSM-Daten liegen. Der Einfluss wechselnder räumlicher Dichte der geographischen Kontextinformationen ist vor allem im extrinsischen Ansatz mit OSM-Umgebungen ein Problem. Beide Ansätze sollten darüber hinaus mit anderen Methoden kombiniert werden, die weitere Informatinen über eine Naturbeobachtung nutzen, wie beispielsweise das Beobachtungsdatum, oder den Grad an Erfahrung, die der jeweilige Beobachter besitzt.

Abstract

In a growing number of Citizen Science projects, volunteers from the general public collect large amounts of observation data of organisms. Such data are an important contribution to biodiversity research, providing information on the distribution of species over large areas and long periods of time. In the current global biodiversity crisis, such information is urgently needed to support research and conservation efforts. One of the most important issues which must be addressed before these data can effectively be used, is data quality. This is a concern especially with data which are being collected in a casual way, without strict, formal protocols ensuring certain standards of data quality before or during the data collection process. There is great need for approaches which allow for assessing data quality of casual citizen science observations of organisms automatically, to cope with the large amounts of observation data which are produced by casual biodiversity citizen science projects.

Casual citizen science observations of organisms are biological, but also geographical data, because they always possess location information. Collected mostly online and by untrained volunteers, they belong to the emerging domain of geographic information called Volunteered Geographic Information (VGI, Goodchild 2007). Approaches which are based on geographical criteria are therefore a promising avenue towards providing suitable methods for quality assessment. At the same time, casual citizen science observations of organisms are a special kind of VGI, because they mostly do not represent permanent objects, but rather have the nature of events which cannot be proven to be correct or incorrect. Quality assessment must therefore resort to proxy approaches such as estimating the plausibility of an observation in light of certain reference information.

This thesis developed and evaluated novel approaches to quality assessment of casual citizen science observations of organisms based on estimating the plausibility of observations in light of VGI context. It employed two use cases of casual citizen science projects with two different areas of interest: ArtenFinder Rheinland-Pfalz (Germany), and the global project iNaturalist, of which data from California (USA) were used. In an intrinsic approach, geographic context is provided by neighboring observations from the same dataset which are transformed into species-specific observed communities, describing a species' typical context of other species usually observed close-by. An extrinsic approach uses OpenStreetMap (OSM), a well-established global VGI project providing detailed geographic information on physical objects, for describing a species' geographic context in the form of an OSM environment, consisting of the OSM features typically found in close proximity to a species' observations. Plausibility of a new observation added to the dataset is estimated by comparing its context of neighboring observations or of OSM features to the species' observed community or OSM environment. This comparison is achieved by using similarity indices.

The intrinsic observed communities approach as well as the extrinsic OSM environments approach were evaluated by estimating the plausibility of plausible or implausible observations. This was done with real approved or rejected observations from the respective projects, but also with synthetic plausible and implausible observations created for this purpose. Evaluation proved that both approaches are able to distinguish between plausible and implausible observations based on VGI context, using similarity index values. The approaches estimate the plausibility of the location of an observation in light of surrounding observations or OSM context, and in light of the species identification given for the observation by the volunteer. Careful examination of evaluation results revealed differences in behavior of both approaches depending on the similarity index used. Results also partly differed between the data use cases. Variable spatial density of observations and OSM data has an influence on similarity index values. Observed communities were found to reflect biological and ecological properties of species, while OSM environments rarely do so. Both methods were also tested with a number of parameter changes, and results found basically stable with different parameter settings. Some modi-

fications to the basic methodology of the approaches, such as applying auxiliary land cover data for focusing relevant geographic context or using observation frequency in similarity calculation, showed potential of improving results.

Future work must seek to overcome the most important drawbacks and weaknesses of the approaches to plausibility estimation of casual citizen science observations of organisms developed in this work. They can be used only for species with an adequate base of previous observations, and for candidate observations in locations providing an adequate geographic context of observations or OSM data. Influence of variable spatial density of context information on plausibility estimation is a problem especially in the extrinsic OSM environments approach. Both methods should be combined with other approaches using other information about an observation, such as the observation date, or the observers' experience.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Biodiversity on earth is currently experiencing severe loss through extinction of species. This issue has been receiving rising amounts of attention from science as well as from the public since the 1980s (Wilson 1988). Some authors compare its magnitude, in terms of numbers of species lost, and in terms of the speed at which this loss occurs, with mass extinction events in earlier geologic ages (Novacek & Cleland 2001, Ceballos et al. 2015, Johnson et al. 2017). However, other than these earlier events, which were triggered by natural processes in earth's history, the current extinction crisis (Ehrlich 1988, Novacek & Cleland 2001) is caused by human activities. The most important factors are habitat loss or degradation (e.g., transformation of forests to agricultural land), overexploitation of natural resources (e.g., of water or soils), and invasive species harming local species communities (Ehrlich 1988, Pereira et al. 2010, Pettorelli et al. 2014). These factors combine and interact with human-induced climate change, which leads to shifts in species distributions (e.g., Steinbauer et al. 2018). Besides actual loss of species through extinction, another important facet of this crisis is decline in populations. For instance, decline in insect populations received much attention in recent years (Van Swaay et al. 2015, Hallmann et al. 2017), and negative impacts on other species groups such as birds, where many species rely on insects as a food source, have also been found (Schrauth & Wink 2018).

International initiatives soon addressed these concerns, such as the Convention on Biological Diversity (CBD), which was initiated by the United Nations Environment Progamme (UNEP) and opened for signature on the 1992 United Nations Conference on Environment and Development at Rio de Janeiro[1]. One of its principle objectives is the conservation of biological diversity, supported, among other things, by suitable monitoring efforts. It was followed, in 2002, by a Strategic Plan which defined the 10 so-called Aichi Biodiversity Targets, which were aimed at significantly reducing global biodiversity loss until 2010. This goal was, however, not reached (Butchart et al. 2010). A follow-up Strategic Plan 2011-2020 was installed with an extended set of targets[2] (Buckland & Johnston 2017), but with the same overall goal of reducing biodiversity loss. There is some indication that this goal will again be missed (Tittensor et al. 2014).

At the same time, gaps in the knowledge about biodiversity were identified as one of the most important difficulties which must be overcome to effectively face the biodiversity crisis (e.g., Balmford et al. 2005, Butchart et al. 2010, Schmeller et al. 2017). Gaps are found in geographic coverage, in temporal coverage, and in coverage of different species groups (Amano et al. 2016). Knowledge on biodiversity depends mostly on monitoring programs which collect data on species occurrence in space and time, as well as other variables (Proença et al. 2017). However, existing biodiversity monitoring programs, which often rely not only on professional scientists, but also on local volunteers with high expertise who support the programs for long periods of time, suffer from effects such as population shifts from rural to urban areas, aging of volunteers, and societal changes which make it more and more difficult to recruit new long-term volunteers for such programs (Franke & Eissing 2010, Walz et al. 2013, Kettunen et al. 2016).

Collaboration with volunteers in biodiversity research is part of what, today, is called citizen science. Despite the difficulties which traditional biodiversity monitoring is facing, citizen science is a very promising avenue in tackling the challenges presented by biodiversity data gaps. This is mostly due to

[1] https://www.cbd.int/history/, last accessed on 2018-11-11

[2] https://www.cbd.int/sp/default.shtml, last accessed on 2018-11-11

the profound changes which have been taking place in the last 10 to 15 years in citizen science through technological progress (Newman et al. 2012, Dickinson et al. 2012, Kosmala et al. 2016), and which still continue (Brenton et al. 2018). Advanced internet technologies were the most important factor (Dickinson et al. 2010, Wiersma 2010, Newell et al. 2012, Liu et al. 2017, Mazumdar et al. 2018). In particular, technologies which allow for collaborative use of the web transform internet users from mere consumers to active producers of content (Budhathoki et al. 2008). They facilitated the emergence of numerous online projects in recent years, where people can report observations of organisms. More technologies, such as Global Positioning System (GPS) receivers, smart phones and other mobile devices supporting the use of mobile software applications (apps), and digital cameras are also important in the field of biodiversity citizen science (August et al. 2015, Pocock et al. 2018). Such projects are very varied (Pocock et al. 2017). Projects employing the general public to collect observations of species occurrences have the potential to produce very large amounts of occurrence data over broad geographic scales and long periods of time, and to do so at relatively low cost (Connors et al. 2012, Dickinson & Bonney 2012, Hochachka et al. 2012, Tulloch et al. 2013, Theobald et al. 2015, Schmeller et al. 2017).

Most studies which advocate citizen science as a promising way of improving the data basis for biodiversity research do not fail to point at serious issues which need to be addressed if such data are to be used in meaningful research (Hochachka et al. 2012). Most prominent among these is data quality (Kosmala et al. 2016, Pocock et al. 2018, Williams et al 2018). Data from citizen science sources suffer from a lack of trust by potential data users in science and authorities (Conrad & Hilchey 2011, Kosmala et al. 2016), because data are often collected by persons who do not have formal expertise in the respective field. Projects, however, strive to enhance data quality and to build trust in their data outcome in many different ways (Wiggins et al. 2011, Freitag et al 2016, Kosmala et al. 2016). Efforts have been undertaken to establish data quality as a primary goal in citizen science and to make project organizers aware of its importance, e.g., in the European Citizen Science Association's Ten Principles of Citizen Science (ECSA 2015, Robinson et al. 2018) or the Green Paper of Germany's "Bürger schaffen Wissen" initiative (Bonn et al. 2016, Richter et al 2018). Moreover, many studies found that adequate analysis methods can be successfully used to overcome data quality issues, such as biases of different kinds (Encarnaçao et al. 2012, van Strien et al. 2013, Bird et al. 2014, Isaac et al. 2014, Kosmala et al. 2016, Botella et al. 2018, Robinson et al. 2018). In this way, such data can be put to many different uses in research (Powney & Isaac 2015). Data collected by volunteers in citizen science projects may therefore indeed present a way of mitigating data gaps in biodiversity research, but there are more issues to be taken into account. For instance, citizen science projects exhibit global heterogeneity in distribution of recording. Global regions with high recording (mostly in the developed countries), and regions with globally high species richness (found mostly in the tropics) do not necessarily match (Pocock et al. 2018). Also, citizen science observation data of organisms are sometimes found to fail in providing a suitable data basis for answering certain questions (Kamp et al. 2016).

Many biodiversity citizen science projects have a relatively low level of structure or protocol in their data acquisition processes, which allows them to keep entry barriers for volunteers low (Wiggins et al. 2011, Freitag et al 2016, Minghini et al. 2017) and to attract large numbers of volunteers. Often, large amounts of observation data are produced in relatively short periods of time. There is an urgent need of adequate approaches and methods for quality assessment of these observations (Idris et al. 2014). This work is aimed at contributing to the development of methods which allow for automatic plausibility estimation of observations of organisms from such citizen science sources. They are an important component in making citizen science observations of organisms accessible for research in biodiversity and ecology, results of which contribute to stemming the depletion of biodiversity on earth.

1.2 Citizen Science and Volunteered Geographic Information: A Conceptual Landscape

Citizen science was recently taken to a new level by the internet, mobile devices and other technological innovations. It is situated within an extensive landscape of related phenomena and concepts, for which a number of different terms are used in scientific literature. This section presents the concept and term of citizen science and related concepts and terms which are basic for this thesis. It explains their origins and the relations among them. Citizen science observations of organisms, which are the primary object of this research, are geographical and biological data at the same time. This thesis uses a geographical perspective which views observations of organisms by citizen scientists as a form of Volunteered Geographic Information, or VGI (Goodchild 2007). It is therefore very important to clarify the relations between citizen science, VGI and other related concepts which complement this conceptual landscape. Concluding, this section presents a graphical representation illustrating relationships, such as overlaps and inclusions, of the terms which are discussed here (Figure 1.2.1).

The *phenomenon* of citizen science has been existing for a long time, at least for some centuries (Miller-Rushing et al. 2012, Liu et al. 2017). However, the *term* 'citizen science' is relatively new (e.g., Silvertown 2009, Dickinson & Bonney 2012). Finke (2014) cites Irwin's 1995 "Citizen Science: A Study of People, Expertise and Sustainable Development" as the first book having that term in its title. There is a lively discussion going on about what should, or, maybe even more important, what should not be called citizen science (Eitzel et al. 2017). Attempts to define its boundaries use a number of different aspects of the phenomenon. Some ask for the presence of a clear scientific agenda in the form of a scientific hypothesis formulated before the actual start of the citizen science project. Such projects are then specifically designed to prove or disprove that hypothesis, following a scientific standard model (Silvertown 2009). Further aspects used to define citizen science include the scientific nature of the object examined, or the use of scientific methods, protocols or tools by participants (Haklay 2013). In the context of environmental monitoring by citizens, Resch (2013) highlights the importance of project participants' local expertise in a certain scientific domain to be considered citizen scientists. In a typology of citizen science which uses the level of participation by citizens in a scientific endeavor as the main criterion, Haklay (2013) coined the term 'extreme citizen science' for projects with the highest possible level of citizen participation. In a related effort, Bonney et al. (2009) used the term 'public participation in scientific research' (PPSR, see also See et al. 2016). They categorize citizen science projects by degree of participation into contributory, collaborative and co-created projects, a concept further extended later by Shirk et al. (2012).

These aspects can be illustrated by looking at biodiversity citizen science. Many citizen science projects in the biodiversity domain examine the occurrence and distribution of organisms. Thus, their object is certainly scientific in nature (Haklay 2013). Most of these projects are of contributory type (Pocock et al. 2017). They are designed by scientists and receive data contributions by volunteers (Bonney et al. 2009). However, their approaches at how to collect data and the goals what to do with them are very diverse (Pocock et al. 2018). Some projects provide the online infrastructure to collect and manage reports of species occurrences by volunteers without pursuing a single, narrowly defined research agenda or hypothesis (Dickinson 2010). They do not impose strict protocols or rules on their participants, while of course striving for certain properties and quality standards for the data produced. ArtenFinder Rheinland-Pfalz[3] and iNaturalist[4], the data use cases of this work, are just two examples for this type of project. Their mode of data collection is often called casual or opportunistic. Other projects implement very specific procedures which they require their volunteers to adhere to, such as

[3] https://artenfinder.rlp.de

[4] https://www.inaturalist.org/

fixed places to be monitored for occurrences of a limited set of species, with a fixed amount of effort, and often at closely defined times or intervals. Examples include butterfly monitoring programs in many countries, for instance, in Germany[5] or Great Britain[6]. These projects often have equally specific research goals and closely defined uses which the data are put to, one of which is reliable long-term monitoring of population development and of spatial distribution of certain species or species groups. Another difference can be found in the way projects acquire participants. Some projects invite anyone to take part, regardless of prior experience, personal skills or local expertise. Other projects make an effort to select and train participants to make them fit for the tasks they are asked to perform. Projects such as eBird[7] (Sullivan et al. 2009) try to find a middle course of using some degree of protocol to produce data with increased value, e.g., data allowing for absence information to be derived. At the same time, they try to keep hurdles for participation low enough to attract large numbers of participants. We can see that the answer to the question whether a project qualifies as citizen science clearly depends on the aspects which are deemed crucial. This question is more often in debate for the projects on the "low protocol and open participation" end of the scale.

However, for the purpose of this work, an insight is helpful which was phrased by Haklay (2013, p. 107) as follows: "While it is possible to try to formulate a definition that delineates the boundaries of what should or should not be considered citizen science, a much more fruitful approach is to understand the general properties of citizen science and its overlap with VGI". Haklay thereby steers the discussion away from the necessity for, and difficulties associated with, defining the boundaries of citizen science. He rather puts it in relation with VGI. Contrary to citizen science, the origins and definition of this term are quite clear. Goodchild introduced it in his 2007 GeoJournal paper. He describes VGI as the result of "the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information" (Goodchild 2007, p. 212). He also discusses several examples of projects producing different kinds of VGI, among them OpenStreetMap (OSM), now one of the most successful projects in the domain of VGI (Neis & Zielstra 2014). Haklay (2013) introduces the term 'geographical citizen science' for the part of citizen science overlapping with VGI. It includes all kinds of citizen science projects "… where the collection of location information is an integral part of the activity" (Haklay 2013, p. 112). Among biodiversity citizen science projects are many examples for projects of this type. Observations of species from volunteers always carry location information. They can therefore be considered to be a special kind of VGI. See et al. (2016) also list the term 'collaboratively contributed geospatial information' (CCGI) and trace it back to Bishr & Kuhn (2007) and to Keßler et al. (2009). They identify a higher degree of collaboration between individuals as the most important characteristic distinguishing it from VGI.

There are also a large number of citizen science projects without a geographical element. Prominent examples include Galaxy Zoo[8] and Foldit[9], from the astronomy and medicine domains, respectively. However, they also use web technologies and the internet. Grey (2009) created the term 'citizen cyberscience' for this type of citizen science projects and the two examples cited here. Clearly, the term also matches most geographical citizen science projects, although Grey (2009) does not use geographical citizen science examples to illustrate it. Haklay (2013) extends Grey's (2009) discussion of citizen cyberscience by adding the field of 'participatory sensing'. Technical sensors and GPS receivers in volunteers' mobile phones are used to collect georeferenced environmental measurement data. Grey mentions this concept (but not the term) already briefly in his article as one of the future developments anticipated for citizen science. It is related to the concept of 'people as sensors'. Here, "people act as

[5] http://www.tagfalter-monitoring.de/
[6] http://www.ukbms.org/
[7] https://ebird.org/home
[8] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/
[9] https://fold.it/portal/

non-technical sensors with contextual intelligence and comprehensive knowledge" (Resch 2013, p. 393), providing, according to Resch (2013), observations rather than measurements. This description in turn is closely related to what Goodchild (2007) states for the term 'humans as sensors', speaking of a sensor network consisting "of humans themselves, each equipped with some working subset of the five senses and with the intelligence to compile and interpret what they sense" (Goodchild 2007, p. 218). Therefore, the terms 'people as sensors' and 'humans as sensors' can be and actually are used interchangeably (Resch 2013), along with a third variant, 'citizens as sensors', found in the title of Goodchild's 2007 GeoJournal article. Finally, all of these concepts denote approaches which lead to the production of VGI or geographical citizen science data.

The concepts discussed so far build on using the general public to perform important tasks in a project. In most cases, they use an internet application as a common platform for the contributors and the project managers. Such activities are often subsumed under the term 'crowdsourcing', created by J. Howe in 2006 (Howe 2006). It has been eagerly accepted for use in academic scientific discussion about VGI and citizen science, although it was not coined in the (academic) scientific community. Most of the examples used by Howe in his original article to describe the phenomenon have not an academic, but rather a commercial, business, industry, or media background. They present crowdsourcing as a new form of the business practice of 'outsourcing'. Such activities produce 'user-generated content' along the way. This places the concept of crowdsourcing within that broader concept also used by Goodchild in his description of VGI as "a special case of the more general Web phenomenon of user generated content" (Goodchild 2007, p. 212).

The new processes and tools which lead to the generation of VGI have also been subsumed under the term 'neogeography' (Turner 2006, Haklay et al. 2008, Connors et al. 2012). While overlapping heavily with VGI, geographical citizen science and user-generated content, Haklay et al. (2008, p. 2022) state that "Neogeography websites do not necessarily rely on user-generated content to supply innovative services and instead some supply data which they collect from disparate or difficult to access sources."

For citizen science, it should be noted that there is a "non-internet" sector in the form of projects which use traditional, analog methods, at least for parts of their activities. Examples include the collection of botanical or zoological specimens from citizen scientists, which simply can't be collected, transmitted, stored or shared in digital form, or the use of paper forms for information transmission. For instance, the German Mückenatlas project[10] collects specimens of mosquitos along with a paper questionnaire (Kampen et al. 2015). What's more, some projects or programs do not use the crowdsourcing approach to draw on the general public for acquiring participants. They use a more or less fixed network of volunteers who reliably contribute on a regular basis, often over many decades, something that could be termed "non-crowd" citizen science. Many meteorological monitoring programs work that way (Elwood et al. 2012; Haklay 2013, World Meteorological Organisation 2001). Eitzel et al. (2017, p. 10) also conclude that "not all citizen science is crowdsourcing, and not all crowdsourcing is citizen science". Finally, there is a "dark figure" of citizen scientists who do not take part in collective efforts or projects which cross-link them with other citizen or professional scientists via the internet or in other ways. They engage in scientific research on their own and on their own account (Wink 2017).

To conclude this discussion, Figure 1.2.1 wraps up graphically the relations between most of the terms discussed here, especially overlaps and inclusions in one another. As the general discussion is going

[10] https://mueckenatlas.com/

on, new concepts and terms will not fail to emerge in the future, modifying the landscape which is described here.

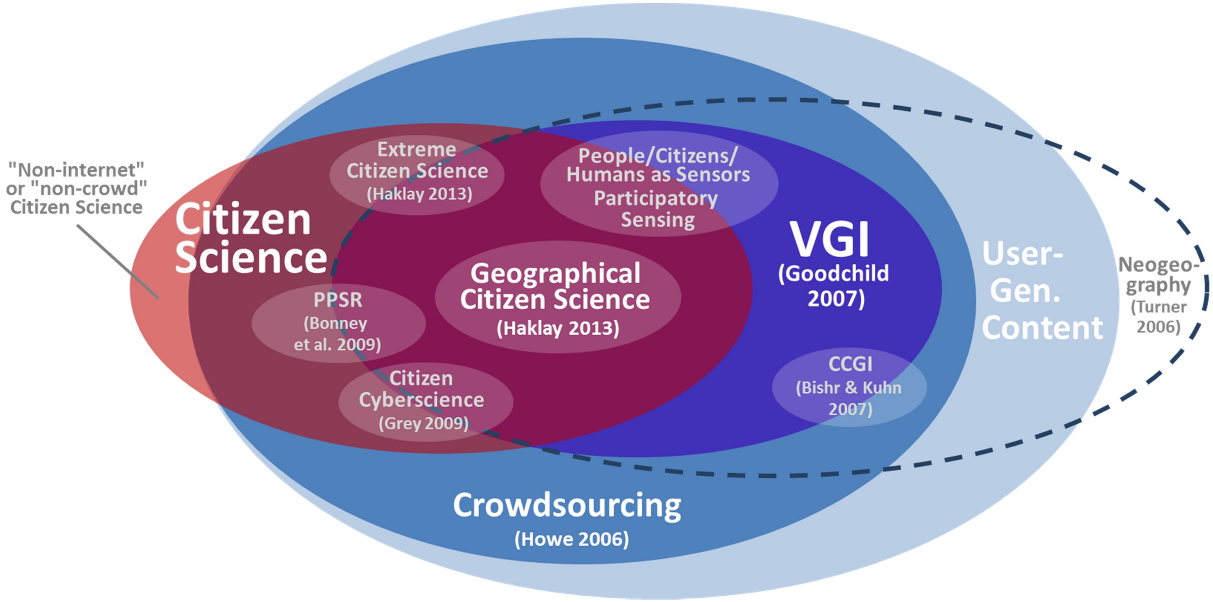

"Non-internet" or "non-crowd" Citizen Science
Citizen Science
Extreme Citizen Science (Haklay 2013)
People/Citizens/ Humans as Sensors Participatory Sensing
VGI (Goodchild 2007)
User-Gen. Content
Neogeo-graphy (Turner 2006)
PPSR (Bonney et al. 2009)
Geographical Citizen Science (Haklay 2013)
Citizen Cyberscience (Grey 2009)
CCGI (Bishr & Kuhn 2007)
Crowdsourcing (Howe 2006)

Figure 1.2.1: Graphical representation of the conceptual landscape of citizen science and VGI.

Citizen science is a universal phenomenon which interacts and overlaps with a large number of scientific domains and social phenomena. Therefore, there are a large number of other perspectives possible, besides the geographical perspective used here. Also, this discussion and the graphical rendering of its results may be questionable, because it includes terms and concepts which are not all of the same nature. For instance, VGI is rather the product of certain approaches to collecting geographical information, whereas citizen science or crowdsourcing are terms which denote processes rather than their results or outcomes (See et al. 2016). However, the overview provided in this section is still useful, because most authors do not explicitly consider those differences when using the terms in a common context. For this work, its main purpose is to put the terms ‘citizen science’ and ‘VGI’ in relation with one another and with some more specific and some more general concepts, because this work deals with citizen science data which are a special kind of VGI, and with projects which are part of geographical citizen science.

1.3 Assessing Quality of Citizen Science Observation Data of Organisms: Principles, Approaches, and Current Practice

Quality of citizen science observations of organisms has been the object of research ever since new web technologies led to a growing availability of such data. In a review of contributory citizen science projects, Wiggins et al. (2011) categorize the different ways of handling data quality issues by the stage of the data collection process in which they are implemented. The authors differentiate between methods introduced before data collection, methods which take effect during the data collection process, and finally methods which are used on the resulting data. In another survey of strategies employed by citizen science projects to support the credibility of their data outcomes, Freitag et al. (2016) adopt this useful principle in a similar approach of categorization and group their survey results into "early actions", "in the field", and "in the office" approaches. Moreover, Bordogna et al. (2014 and 2016a) review strategies of data quality improvement according to when they are adopted (before or after VGI data creation) and divide them into "ex ante" (before data creation) and "ex post" (after data creation) strategies respectively.

Approaches to ensure data quality or to build credibility before the start of the data collection process typically include training of volunteers or implementing standards which volunteers are required to meet. This aims to avoid sources of error connected to the volunteers (e.g., Kosmala et al. 2016). Freitag et al. (2016) also list advice by a scientific institution or by scientists during project development, while Wiggins et al. (2011, p. 18) list the implementation of "quality assurance project plans", Both strategies are aimed at strengthening project design and at avoiding sources of error connected to project structure. Some of these approaches are also used during the process of data collection, especially volunteer training and selection of participants by pre-defined levels of skill (Wiggins et al. 2011). In this stage, more mechanisms are taking effect, such as supervision of participants by experts or experienced volunteers, or technological tools which support the process of data recording (Freitag et al. 2016). Also, repeated sampling, the use of standardized equipment, requirements to submit evidence (e.g., photos) or additional paper forms along with online reports all support data quality in the stage of data acquisition (Wiggins et al. 2011). Often, experiences made in the early stages of data acquisition are used to improve project design even after the proper project development has ended, with the aim of enhancing the quality of resulting data (Kosmala et al. 2016). Again, some strategies employed in this second stage of the data collection process are also used in the third stage, i.e. after data collection. One of these strategies is the identification of observations which appear unusual for some reason (Wiggins et al. 2011). Some projects check observations against predefined threshold values of known parameters on submission, such as seasonal occurrence or plausible numbers of individuals, and provide feedback to the volunteer right away. Jacobs (2016) discusses some examples of this type. Another technique which can be applied during or after data collection is evaluation of a volunteer's performance, which may require background information on the volunteer (Wiggins et al. 2011), but which can also be done on the basis of the volunteer's contributions. The latter approach is used by ArtenFinder Rheinland-Pfalz, which records the numbers of observations per observer that were accepted or rejected in the project's validation process. This allows for the calculation of an observer-specific performance parameter (Jacobs & Schotthöfer 2015). Many projects practice more methods of quality control in the post-data-collection stage, including manual data validation by experts, image recognition, checking against reference data, or computational methods which use statistics or data mining (Wiggins et al. 2011). Freitag et al. (2016) add to this list the publication of results or data in outlets with peer review as another means of enhancing credibility. Finally, Wiggins et al. (2011) also list documentation of quality assurance mechanisms as an important element.

Building on their former work, Wiggins et al. (2013) provide a data management guide for project managers. They propose use of the term ‘quality assurance’ for all mechanisms employed to assure data quality in the stages before and during data collection, and the term ‘quality control’ for the post-data-collection stage. This is a very useful step towards disambiguating these terms, which are so far not used in a systematic manner in the relevant scientific literature. A third term which does not appear in the sources cited above, but which is widely used in the literature on data quality of citizen science and VGI data, is ‘quality assessment’ (e.g., Crall et al. 2011, Foody et al. 2013, Samy et al. 2013, Bordogna et al. 2014, Fan et al. 2014, Arsanjani et al. 2015). It can be recognized as a necessary element of quality control (in the sense introduced by Wiggins et al. 2013), because assessment of quality can only be performed on already existing data. Compared to ‘quality assurance’ and ‘quality control’, the term is more neutral because it does not contain any element of quality enhancement or any suggestion of that a certain level of quality should be insured. It rather denotes the examination of the quality of a piece of information or of a dataset, the results of which can be used for quality control, in the dual sense of error identification and of certification of (high or adequate) quality. Results of quality assessment are, of course, a necessary basis for quality improvement. Chapman (2005) used the term ‘data cleaning’ to denote a combination of quality assessment and subsequent quality improvement based on quality assessment results.

Goodchild & Li (2012) present three main mechanisms of quality control which are suitable for VGI projects, and which are also used in practice of data quality control. Citizen science observations of organisms is a form of VGI (see section 1.2), such that these mechanisms are also applicable to such data. The mechanisms identified by Goodchild & Li (2012) are the following:

- The Crowd-sourcing approach, referring to the principle that data errors are quickly found when many people work on the same dataset (also sometimes called “Linus’ Law”),
- the social approach, where certain experienced users take the role of controllers watching over data quality, deciding editing disputes, and so forth, and
- the geographic approach, where the power of geographic context is used to check data using formalized rules which build on geographic knowledge.

All three mechanisms are well established in quality control practice for citizen science observations of organisms. Goodchild (2013) points to serious drawbacks of the first two approaches. For instance, the crowd-sourcing approach may fail if a piece of information is relatively obscure, that is, if it is not well known by other users who might correct it if it is represented erroneously in the data. In the case of observations of organisms, this concern translates to the problem that they do not, in many cases, represent geographic facts but transient events which cannot be verified, an important issue which will be discussed in detail in chapter 2. The social approach has the main problem of involving manual checking of data by persons, which makes it slow (Goodchild 2013). In projects where large amounts of information are acquired in relatively short periods of time, this factor compromises the feasibility of this approach. Goodchild (2013) concludes that the geographic approach, with its ability to provide tools which can be used for automatic quality assessment of data, holds great potential to solve these issues, and he identifies a great research need in this area, especially for geographic information science. Yan et al. (2017) also cite two of Goodchild & Li’s (2012) approaches (the crowd-sourced approach and the geographical approach), but relate them with two other strategies which are also common in VGI data quality assessment. The first of these is based on the provenance or lineage of a feature in a VGI dataset and therefore builds on information about its history (e.g., Mooney & Corcoran 2012a). The second approach they mention is the use of trust, centering on a VGI contributor’s experience, reputation and trustworthiness which can be used as a proxy to assess the quality of a volunteer’s contributions to a dataset (see also section 2.3).

Research, as well as quality control practice in citizen science projects, is already tapping into the field of geographical methods for VGI data quality assessment. The geographic context which is necessary for using this approach can come from two principle types of sources. On the one hand, geographic data can be used which describe environmental aspects of the geographic context of the observation data whose quality is to be assessed. On the other hand, other observation data within a project provide context. A potential third source of geographic context data for observations of organisms would be provided by authoritative reference data of species occurrences collected by scientific standards (Snäll et al. 2014). For other types of VGI, such as OSM, using authoritative reference data has been a much-used strategy for assessing the quality of datasets (Neis & Zielstra 2014). However, suitable reference data for applying this quality assessment strategy to citizen science observation data of organisms do not exist for most species or regions (Snäll et al. 2014, Yan et al. 2017, Vahidi et al. 2018). In a recent study, Vahidi et al. (2018) use a combination of context sources to assess the quality of citizen science observations of a plant species from iNaturalist and E-Flora BC[11]. They use a niche-modelling approach to assess an observation's consistency with the target species' ecological requirements. This approach builds on geographic data about environmental variables, such as climate parameters and elevation, to calculate the probability of the occurrence of the target species at a location. In a second indicator an observation's distance to the center of a cluster of previous observations of the target species is used as a measure of trust in the observation. The data basis for finding these clusters consists not only of citizen science observations from the same dataset, but also of records from museum collections and herbaria, which are considered authoritative. Thus, their first approach (based on niche modelling) makes use of geographic information from outside the citizen science project whose observations are evaluated, while the second approach uses, at least in part, information about geographic context which comes from the project whose observations are assessed, as well as external observations which are considered authoritative. This second indicator builds on work by Yan et al (2017), who introduce a quality indicator for crop pest reports from farmers in China which uses the distance of a report to the center of a cluster of similar reports. Both studies complement their indicator (or indicators) based on geographical context by a trust indicator that evaluates the volunteers' abilities. Mülligann et al. (2011) propose a method for evaluating the plausibility of new points of interest (POI) in OSM. They establish a semantic similarity between OSM tags based on the similarity of tagging histories of OSM objects and on their proximity to existing POI of the same type. Interestingly, the different nature of OSM data when compared to citizen science observations (see also section 2.2) leads to an inversion of the plausibility evaluation in certain cases: while an observation of a species or phenomenon appears more plausible if it is closer to a cluster of already existing observations of the same species or phenomenon (Yan et al 2017, Vahidi et al. 2018), a new POI which is added to OSM in a location which is close to an already existing POI of the same type may appear implausible, especially with certain POI types which tend to be spatially dispersed, such as post offices (Mülligann et al. 2011).

Quality control practice in citizen science projects collecting casual observations of organisms makes use of a number of filtering methods based on heuristics which use general knowledge about a species, including geographic criteria. Some examples are also discussed in Jacobs (2016). For instance, the well-known American bird reporting portal eBird, the German bird reporting portal ornitho.de[12] and the German multi-species portal naturgucker.de[13] check observations for accordance of region and season, based on known species properties in this respect. The "cleaner rules" employed in the National Biodiversity Network[14] in the UK are a similar example which uses known spatial and temporal

[11] http://ibis.geog.ubc.ca/biodiversity/eflora/

[12] https://www.ornitho.de/

[13] https://naturgucker.de/natur.dll/$/

[14] https://nbn.org.uk/news/new-record-cleaner-rules-now-available/, last accessed on 2018-10-31

distribution of species (Spyratos et al. 2014), and so is the North American project Feeder Watch[15] (Bonter & Cooper 2012). First steps towards using previous observation data can be found in naturgucker, which checks for previous observations of a species in a region when a new observation of the species is submitted for that region. eBird tunes filter parameters of their heuristics based on existing observation data (Sullivan et al. 2014). In an attempt to make use of geographical criteria in order to support the expert validation of observations, ArtenFinder Rheinland-Pfalz, one of the data use cases of this work, implemented a plausibility tool which, among other things, offers information about species observed around a candidate observation (Jacobs & Schotthöfer 2015). This tool is designed to support validation decisions on new ArtenFinder observations. However, the information on species found around a candidate is so far not used in quality assessment practice (personal information D. Frank (ArtenFinder/KoNat), 07-2017) because it is too complex and therefore too time-consuming to use (see also section 2.1.1). The same plausibility tool also uses another geographic approach by checking for previous observations of the same species in spatial proximity to a new observation, which is related to similar checks in, e.g., naturgucker (see above). It is complemented by evaluation of observer trustworthiness and a temporal criterion (Jacobs & Schotthöfer 2015). ArtenFinder also allows for observers to add information on geographical context to an observation, by adding a photo of the habitat where the observation was made, which can be very helpful to experts who check the observation and evaluate its plausibility. However, this option is rarely used by ArtenFinder observers: in the most recent 500 observations added to the system before 2018-10-12, just 6% provided a habitat photo. The method of asking observers to provide context with an observation was also studied by Bordogna et al. (2016b), albeit not with a focus on the geographical context. Rather, they asked observers of crops to provide a self-assessment of how certain they were about the crop classification of their contribution, and to describe the conditions under which the observation was made (e.g., distance to the reported object). Many citizen science projects in the biodiversity domain allow for contextual information to be added,whether geographic or of some other form. iNaturalist, another data use case of this work, provides a wide range of additional input fields to be added to an observation and even allows for users to create new input fields (see also section 2.2). Ornitho.de (already mentioned above) allows observers to add context such as time, age and sex of the observed individual, land use, behavior and more from a fixed set of additional input fields, but all of these are optional.

The difficulty of finding and accessing suitable external reference data for VGI data quality assessment has triggered a number of efforts to provide intrinsic approaches to the problem. For instance, the OSMatrix (Roick et al. 2011) allows for visual assessment of relative data completeness of OSM within a spatial grid structure. Barron et al. (2014) present a comprehensive framework for assessing the quality of extracts of OSM data. This framework mainly examines completeness aspects of objects as well as of attribute information, and currentness. It builds on the history of the OSM extracts which are examined. Their saturation-based approach to estimating the completeness of the road network is adopted and developed further by Barrington-Leigh & Millard-Ball (2017). An intrinsic approach to assess OSM class completeness was the subject of a study by Ballatore & Zipf (2015). These works mostly examine the quality of datasets, rather than of individual objects, but there are also examples for intrinsic data quality assessment of individual objects, such as Mooney & Corcoran's (2012a) approach which uses the editing history of OSM objects. Some studies use intrinsic geographic context, such as Mülligann et al.'s (2011) spatial semantic approach for OSM POIs or Yan et al.'s (2017) cluster-based approach for reports of crop diseases already cited above, but they are rare.

Several aspects of research need can be identified from this review of the research and practice in quality assurance and quality control of VGI, which guided the direction of research of this work. Many citizen science projects build trust in their data by employing quality assurance strategies which

[15] https://feederwatch.org/

take effect before or during data acquisition in an often successful effort to ensure certain standards and procedures bearing on the quality of the data outcome. Another type of projects, whose data are often called casual or opportunistic data, use no or only few such strategies. Such projects are often faced with the criticism of not being adequately scientific (see section 1.2), but can produce very valuable observation data if proper strategies for quality control after data collection are in place. A widespread approach is checks of the plausibility of observations which are perfomed by experts (Wiggins et al. 2011) based on their expert knowledge about species and their distribution. Low entry barriers lead to large numbers of observers taking part in projects collecting casual observations. They produce large amounts of observations which need to be checked. However, expert checking can be slow (Goodchild 2013) and does not scale well. There is, therefore, a great need for methods for filtering unusual observations which need special attention. Such methods can be based on different aspects of an observation, such as trust in the observer, temporal aspects, and geographical criteria. Among these, geographical criteria have been identified to hold especially great potential in the VGI domain (Goodchild 2013), to which casual citizen science observations of organisms belong. These methods can support experts who are in charge of validating observations, but they can also be applied by any data user who needs to decide which observations to select for the purpose he or she has in mind, and which to disregard. What is more, such methods cannot only be used to identify unusual observations which are potentially of low quality, but they can also be used to prove high quality of observations, and they can help to objectify data quality assessment.

1.4 Objectives and Research Questions

In response to the research need identified in section 1.3, the work presented here focuses on casual citizen science observations of organisms. There is particular need for data quality assessment approaches for this type of data, due to a lack of data quality assurance procedures employed before and during data collection, which is a common characteristic of projects collecting such data. In these projects, observers are free to decide about all core aspects of the observation process:

- "what": which organisms to report and which to ignore,
- "when": time and frequency of observation activity,
- "where": which locations to choose for observation activity, and
- "how": amount of effort invested (duration of time invested, distance or area covered), resources used (e.g., equipment such as cameras or binoculars, auxiliary information material such as guide books or online resources for species identification), modes of observation (e.g., moving around, staying at one location, or a mix of both).

Quality assessment approaches for such data necessarily target the ex-post stage of data collection, assessing the quality of observations after they were submitted to the dataset by the observer. These approaches have the potential to reduce the workload in quality control regimes based on expert validation of observations, by providing filtering mechanisms which identify unusual or doubtful observations, and may even be used to prompt volunteers not to submit doubtful observations.

This work uses the geographical context of observations as its starting point, because there is great potential for novel approaches to quality assessment of casual citizen science observations of organisms in this field. Geographical context can be used to develop plausibility indicators which estimate the plausibility of an observation in light of that context, while verification of the truth of an observation is mostly not possible due to the event nature of such observations, an important aspect which will be further discussed in chapter 2.

What is more, approaches should be examined which make use of the full information content of multi-species datasets and thus go beyond just using geographic relations within observations of a single species, such as the cluster-based method presented by Yan et al. (2017) and used by Vahidi et al. (2018), cited above (section 1.3). Multi-species approaches hold the potential of working on observations of a species even in the absence of previous observations of that species in the vicinity of the observation which is examined. Of course, they are also restricted to data environments providing the necessary multi-species context.

Further, this work proposes approaches to solve the problem of assessing the quality of individual observations, not of entire datasets or subsets (Keßler & de Groot 2013), because assessing the quality of individual observations is at the basis of all analyses and uses the data may be put to later, including quality assessment of the dataset as a whole (Pocock et al. 2018, Spyratos et al. 2014). All subsequent uses must in some way rely on the quality of individual observations, which is also why most projects collecting casual citizen science observations of organisms have some procedure in place for assessing the quality of individual observations (see section 1.3).

In a domain where there is a general lack of authoritative reference data for extrinsic quality assessment (see section 1.3), it is of special interest to look for intrinsic approaches to quality assessment. In this work, an approach is considered to be intrinsic if it requires no data other than those provided by the project whose observations are assessed. However, while intrinsic approaches have important advantages, especially skirting the difficulty of finding appropriate external sources of geographic context information, they also have disadvantages based on the limits of context information which is intrinsically available. Overcoming these limits requires combination of multiple sources of geographic context in plausibility assessment, including extrinsic ones. From the perspective of casual citizen science observation data in the biodiversity domain, OSM is a novel source of extrinsic geographic context information within the VGI domain, whose value for use in quality control of casual citizen science observations of organisms has so far not been thoroughly explored.

All of the above considerations condense into the following principle research questions, which are tackled in this work:

1. Principle research questions:

> **a) How can geographic context be used for intrinsic assessment of the plausibility of casual citizen science observations of organisms?**
>
> a1) How can casual observations be turned into an intrinsic source of geographic context?
>
> a2) How can this intrinsic context information be used to estimate plausibility of a candidate observation?
>
> **b) How can extrinsic VGI data be used for assessing the plausibility of casual citizen science observations of organisms?**
>
> b1) How can OSM data be used as an extrinsic source of geographic context?
>
> b2) How can this extrinsic context information be used to estimate plausibility of a candidate observation?

Answering the principle research questions requires the following steps:

- identify approaches which are able to make use of geographic relations between casual citizen science observations, or between these observations and OSM data, to produce meaningful geographic context,
- develop these approaches into methodologies which allow for using them as plausibility indicators, and
- conduct thorough studies to examine the properties and behavior of these plausibility indicators.

Studies on the properties and behavior of approaches for plausibility assessment, whether based on intrinsic geographic context from observations, or on extrinsic VGI context from OSM, raise a number of in-depth research questions which need to be answered to grasp the value and limits of these approaches:

2. In-depth research questions:

a) What are the effects of the spatial properties of geographic context data on plausibility estimation?

Both casual citizen science observations of organisms and OSM data are VGI. Such data have special properties concerning spatial structure, especially uneven distribution in space. How does this structure affect results obtained with approaches to plausibility estimation based on geographic context from such data?

b) How do species properties affect results?

Observations of biological species are a very special kind of VGI because species have various properties reflecting their biology and ecology. How do such factors bear on the results and their interpretation?

c) How do changes to parameters and methodological modifications affect results?

How exactly do parameter changes influence the approaches' behaviors? Which methodological modifications and extensions are promising, and what are the tradeoffs?

d) What are the extent and limits of indicative power of the obtained approaches to plausibility estimation?

In order to gauge the usefulness of any plausibility indicator, it is important to determine as far as possible the exact extent and limitations of the information it is able to convey. What exactly does such an indicator tell us about the plausibility of an observation? What is it *not* able to tell us?

Chapter 2 of this thesis introduces the projects whose data were selected as use cases for this work. In a detailed comparison, it presents relevant properties of the projects and their data and discusses consequences for quality assessment of these data. Chapter 3 presents two new approaches to plausibility estimation of casual citizen science observations of organisms: the intrinsic observed communities approach and the extrinsic OSM environments approach. It explains the methodological considerations which form their basis, and provides step-by-step descriptions of their functionality. Both approaches are evaluated by using data from the use cases described before, and evaluation results are presented in chapter 4. Detailed discussions in chapter 5 examine the evaluation results, especially effects of the spatial properties of context data and of species properties. Effects of parameter changes and methodological modifications are also discussed in detail. Extent and limits of the approaches' indicative power concerning a candidate case's plausibility are also treated in chapter 5, using a number of practical examples. Chapter 6 provides general conclusions from this work and proposes numerous paths to follow in future research work building on what was achieved here.

2 Properties of Casual CitizenScience Observation Data and OSM: The VGI Perspective

A large number of web-based citizen science projects which are collecting observations of organisms from volunteers went online in recent years. There are virtually hundreds of such projects (See et al. 2017). This multitude is accurately reflected in a number of reviews which used different criteria to analyze their properties and outcomes, with different study goals. For instance, Theobald et al. (2015) reviewed 388 projects to quantify the actual impact of biodiversity citizen science on biodiversity research. Chandler et al. (2017) analyzed 420 citizen science programs in the biodiversity domain concerning their coverage of essential biodiversity variables and these programs' geographical and taxonomic coverage. This work uses casual citizen science observation data of organisms from two projects: ArtenFinder Rheinland-Pfalz and iNaturalist. ArtenFinder Rheinland-Pfalz and iNaturalist are both geographical citizen science projects producing casual observations of organisms which they make publicly available. From the multitude of possible use cases, these two biodiversity citizen science projects were selected because they collect observations of a wide range of species in a casual way. Data of this type are the focus of this work for reasons explained in sections 1.3 and 1.4. Both projects' data collection follows similar procedures, giving data properties a number of important common characteristics which allow for using basically the same methodologies on both, and to compare results. Also, both use cases provide adequate amounts of data within the areas of interest selected (reasons of which are explained below). A the same time, the use cases possess interesting differences, e.g., in species composition of data, geographic characteristics of areas of interest, organizational background of projects etc., which were expected to lead to insightful differences in results obtained with the one or the other use case. However, it can certainly be said that this work would have been equally possible with other project and data use cases, producing comparable results. In the case of ArtenFinder, this research builds on previous work which aimed to support validation of observations by plausibility tools (Jacobs & Schotthöfer 2015).

Data from OSM are used in this work as an additional source of geographic context. OSM is one of the most successful VGI projects and also one of the most researched in recent years (Elwood et al. 2012, Neis & Zielstra 2014). It provides VGI context in an accessible way on a global scale. While sharing some characteristics with the citizen science projects and data used here, it also presents characteristics which distinguish the OSM project and its data. It is important to understand the properties of the projects used here, as well as the properties of their data, in order to be able to understand the results obtained with the approaches to plausibility estimation for casual citizen science observation data of organisms developed in this work. After portraying the projects and their data one by one in section 2.1, this chapter engages in a comparative discussion of their properties (section 2.2). The chapter concludes by identifying important consequences for quality assessment of casual citizen science observation data (section 2.3).

2.1 Project and Data Properties

2.1.1 ArtenFinder Rheinland-Pfalz

The project and its area of interest

ArtenFinder Rheinland-Pfalz provides a web portal and an app for collecting observations of organisms from the general public. It went officially online early in 2011. Compared to most initiatives of this kind, it has a rather narrow spatial scope, focusing its efforts on the German federal state of Rheinland-Pfalz (see Figure 2.1.1 and Figure 2.1.2). Another property setting the project apart is the fact that observations, after expert validation, are incorporated in the official species distribution data of the federal state of Rheinland-Pfalz and serve a number of purposes in conservation administration and planning. This use of citizen science data in government, administration and planning is still rather unusual, but there is an increase in the use of citizen science approaches by environmental authorities worldwide (Owen & Parker 2018). The European Union also promoted collaboration of citizen science and environmental policy makers by supporting so-called Citizens' Observatories (Liu et al 2017).

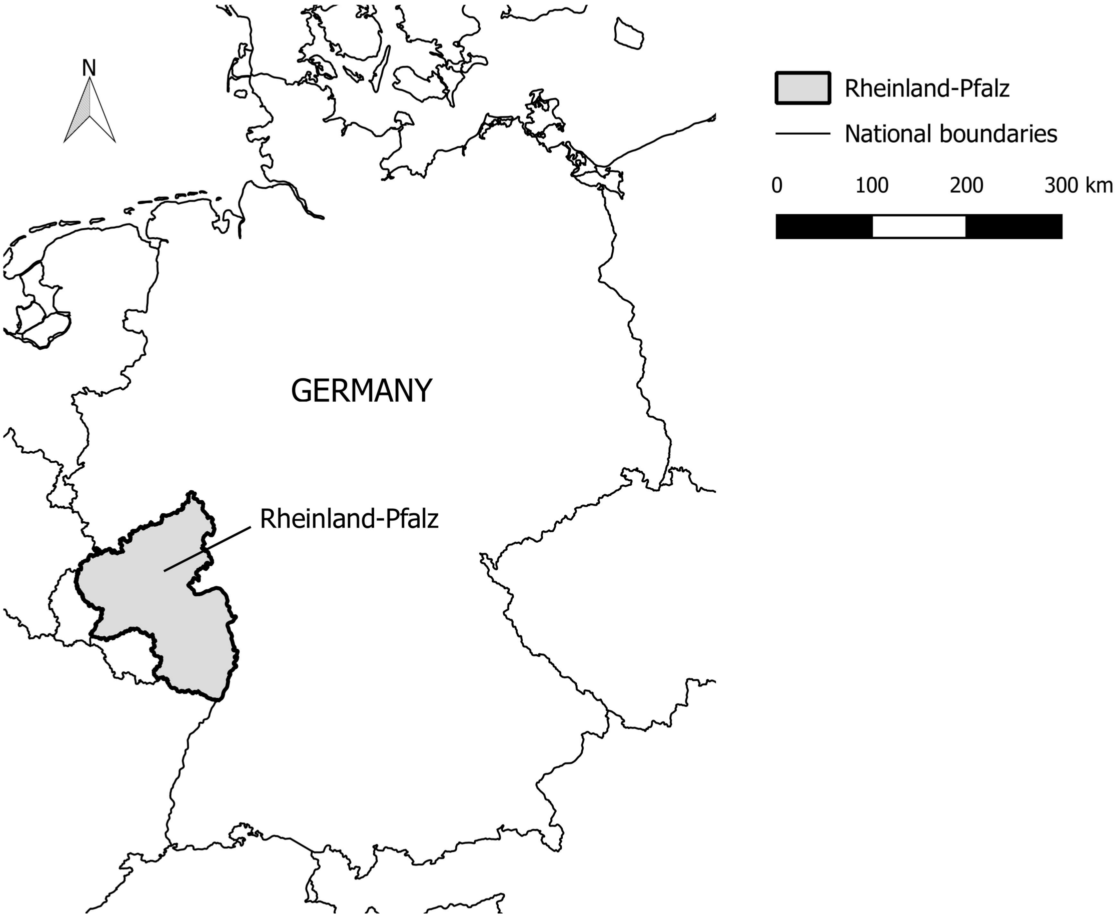


Figure 2.1.1: Geographic situation of the federal state of Rheinland-Pfalz. Area: 19,854 km^2. (Source of national boundaries: ESRI[16]. Source of Rheinland-Pfalz state line: Lanschaftsinformationssystem der Naturschutzverwaltung (LANIS) Rheinland-Pfalz.)

[16] Portions of this document include intellectual property of ESRI and are used herein by permission. Copyright © 2018 Environmental Systems Research Institute, Inc. All Rights Reserved;

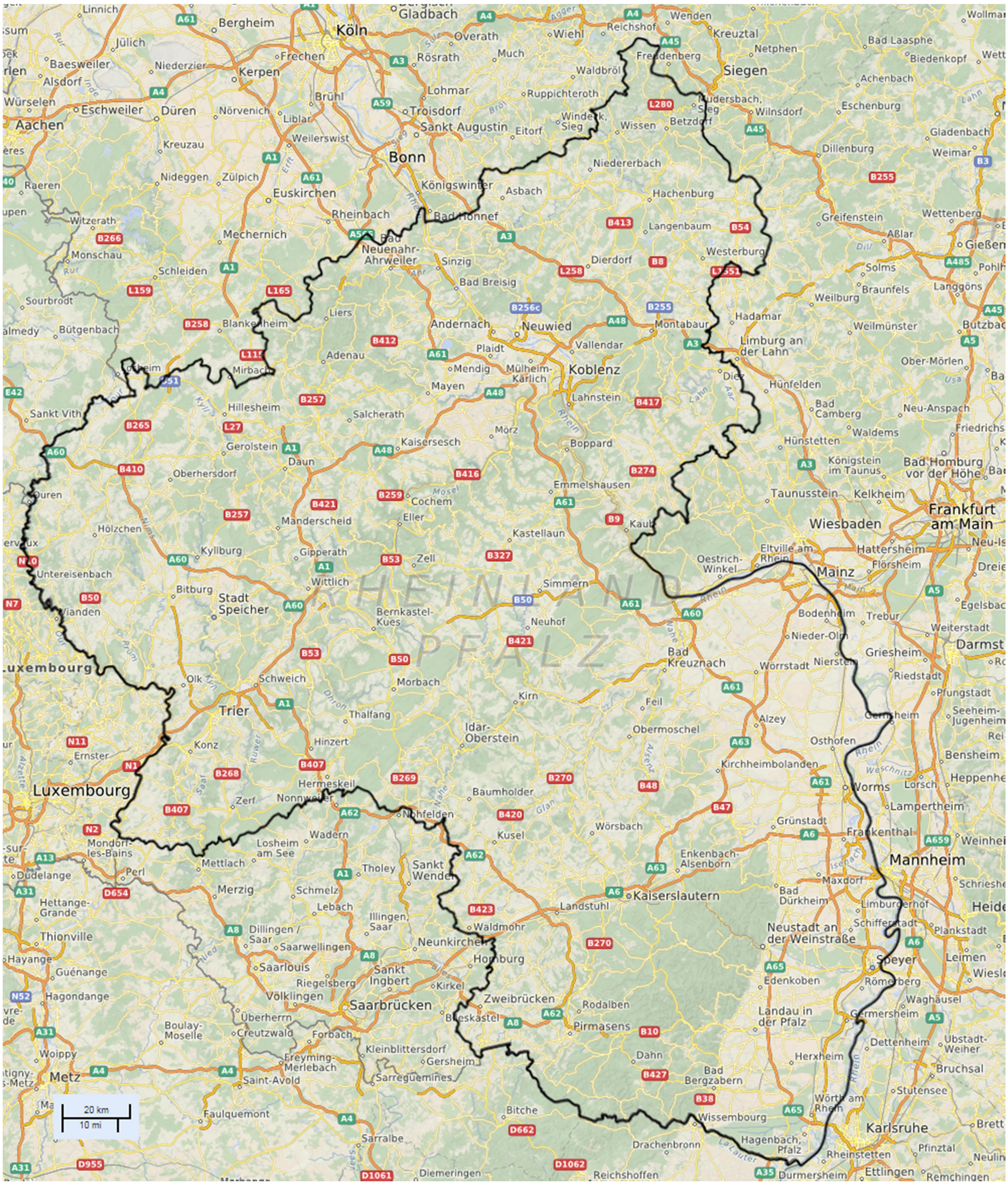


Figure 2.1.2: Rheinland-Pfalz, topographic overview. (Source of base map: OpenStreetMap[17]. Source of Rheinland-Pfalz state line (black): Lanschaftsinformationssystem der Naturschutzverwaltung (LANIS) Rheinland-Pfalz.)

For a report of an observation to ArtenFinder's online portal, an observer has to provide the following mandatory information:

- observer (a user has to be logged in to be able to add an observation, so that this information is automatically added to the dataset),

[17] © OpenStreetMap contributors; OpenMapSurfer; further data sources used there: CIAT-CSI SRTM, ASTER GDEM2, ETOPO1, GeoNames, GlobCover, NaturalEarthData

- location (coordinates, provided either by clicking in a map, or typing the coordinates into the proper fields, in ETRS89 UTM32 format)
- species (supported by an automatic suggestion functionality)
- observation date, and
- status of the observation (only observations marked as "public" will be validated by experts; the alternative is to add the observation to the observer's private list).

Optional information includes a remark, the number of individuals observed, and photos of the observed individual and of the setting it was observed in, or an audio recording. There is also an ArtenFinder app for submitting observations by using a smart phone or other mobile device. It adds an estimation of coordinate precision (based on the GPS precision) to the observation. Beginners or observers unknown to the validating expert are consistently asked to provide a photo proof with their public observations.

The use of the data in government and administration makes a quality assurance strategy necessary which not only ensures a minimum error rate in the data, but which also makes the data credible to potential and actual users in government and administration. In some cases, the data even need to be justiciable if involved in a trial, e.g. connected to infrastructure planning etc. At ArtenFinder Rheinland-Pfalz, the need for this level of quality assurance led to the implementation of a validation regime in which each and every public observation is checked by a regional and species group expert, who either accepts or rejects it. These experts are organized in a corporation, the KoNat UG (a limited liability company). Many of them are volunteers with prior experience and occupations in the environmental sector, but also some ArtenFinder volunteers were promoted to validating experts because of their experience acquired as ArtenFinder observers.

All observations which observers release as public observations have initially the status "Under Scrutiny", German "In Prüfung". Experts have basically three alternatives. The first is to accept the observation, if they judge it to be probably correct. The second option is to reject it, if there are reasons for that. These reasons can be a judgement that the observation is probably wrong, that is, that either the location, or the date of observation, or the species identification is judged to be probably incorrect. Other reasons may be a missing photo proof (which is only but consistently asked from beginners), or a photo proof which does not show important characteristics of the species. There is also the possibility to put the observation "on hold" (German "pausiert"). In all cases where an observation is put on hold or rejected, the observer is informed about the reasons, provided with information on the species in question, and is asked for more information which may lead to acceptance. This communication also generates a learning effect on the side of the observer. Field excursions with experts, which also have a training effect for participants, are frequently offered but are not mandatory for persons who want to take part in the project as observers. Online eLearning tools offered on the project's homepage, such as species portraits and dichotomous keys for many species, also support observers in species identification.

The experts' judgement is based on their knowledge and experience, both regional and for the species group they are working on. In many cases, the expert's knowledge of an observer's experience and reputation also plays a major role in validation decisions. Since early 2016, ArtenFinder's quality assurance process is supported by a plausibility tool which extracts certain information about the candidate observation, using currently existing, accepted observations as context, and which provides this information back to the validating experts. The tool has four elements. The first evaluates the observer's experience, based on a record of quality assurance decisions which allows for calculating the ratios of accepted and rejected observations for the observer. The second element visualizes the candidate observation's date in light of species observation frequency over the year exposed by previous observations of the same species. The third element visualizes spatial proximity of previous observations of the same species. Finally, the fourth element provides a list of species observed so far around the candidate observation's location (sorted by observation frequency). Experts can trigger an evaluation by the plausibility tool for each observation they check, but it is not mandatory. In practice, the tool is considered as a helpful support to the validation process, but has also shown some weaknesses (Jacobs & Schotthöfer 2015). One of these concerns the list of species surrounding a candidate: this list is often very long (depending on the search radius applied and on observation density in proximity to the candidate observation) and therefore hard to handle. The list contains species from all species groups. A validating expert is usually specialized in one or two species groups. Evaluating all species in the list as to their relevance for the candidate species at hand is therefore a task which no single expert is able to perform. Also, such a species-by-species evaluation would take up way too much time. Therefore, this element of the plausibility tool plays almost no role in the verification process (personal information by D. Frank (KoNat), 07-2017). Depending on the target species and its properties, a validation decision of an expert is thus mostly based on the observation date and its match with the species' life cycle or temporal migration patterns, the expert's judgement of the observer's experience in identifying the target species, and on the expert's general knowledge of the spatial distribution of the target species within Rheinland-Pfalz. Photo proofs very often play a major role in the expert's validation decision.

Thematic data properties

ArtenFinder categorizes species into species groups which represent taxonomic units on different taxonomic levels, such as birds (or Aves, a class) and butterflies and moths (or Lepidoptera, an order). This categorization was mostly kept in this work. The species groups of spermatophytes and pteridophytes were consolidated into the species group 'plants', dominated by spermatophyte observations. Figure 2.1.3 illustrates that, in accepted observations up to 2016, 43.50% were of birds, followed by butterfly and moth observations (27.78%), dragonflies and damselflies (9.48%), and plants (6.48%). Overall, plants therefore play a minor role in the ArtenFinder dataset. This is important because the ArtenFinder observation data are thus dominated by species which are at least to some degree mobile, some (especially many bird species) even to a high degree.

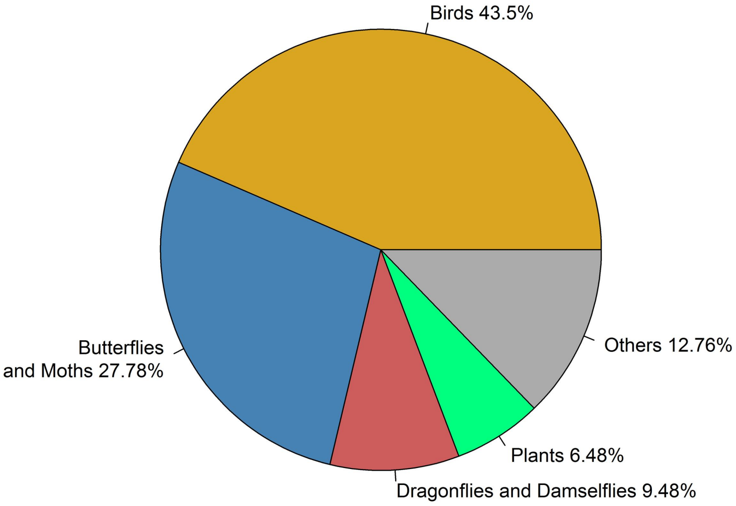
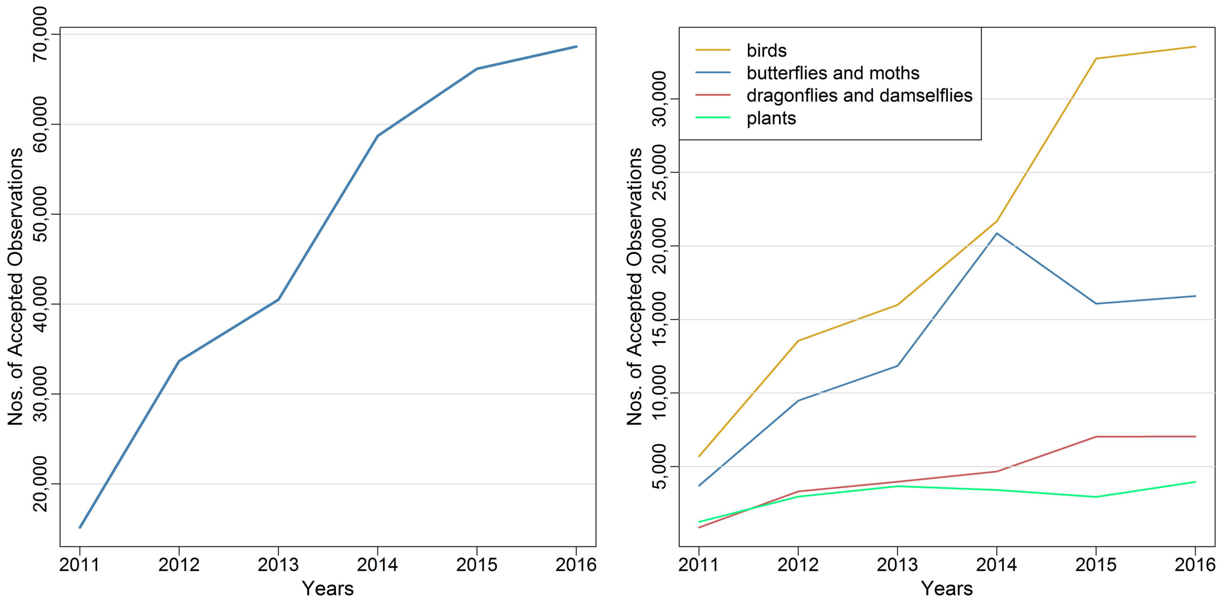


Figure 2.1.3: ArtenFinder, portions of species groups. (Based on Accepted observations in Rheinland-Pfalz up to 2016).

Yearly numbers of observations contributed to ArtenFinder have been going up since the project went online in 2011 (see Figure 2.1.4). Growth rates have mostly been declining, especially in recent years. KoNat is publishing a series of identification guides for species groups. The 2013 appearance of a birds guide (Rößner et al. 2013) may have contributed to the accelerated rise of bird observations in 2014 and 2015. In 2014, the appearance of the guide for butterflies (Schotthöfer et al. 2014) likely triggered the observed rise in yearly butterfly observations, but observation numbers dropped back down afterwards. The 2017 appearance of a guide for dragonflies and damselflies (Ott et al. 2017) caused a rise in observations from this species group in 2017 (personal information by D. Frank, KoNat, 07-2017), not represented in Figure 2.1.4, because 2017 data are not represented in the data use case. ArtenFinder uses a project-specific list of species which volunteers can report. Species not on the list cannot be reported. This list was extended several times since the project went online. At the time the ArtenFinder observation data for this work were last downloaded (February 2017), it held 12,492 species. The list was extended since to hold 16,163 species (as of August 2018). ArtenFinder focuses on protected or threatened species, which are especially relevant for uses in government, conservation and planning. The project's species list also holds common and widespread species, but descriptions of some of these species in the project's eLearning materials contain appropriate notes discouraging observers from reporting such species in ArtenFinder. There also have been several campaigns calling for observations of certain species or species groups which certainly had an impact on the species composition of the observation data. In 2011, a call for observations of Stag Beetle (*Lucanus cervus*) produced ca. 600 observations of this large and striking beetle. Another campaign for Red Kite (*Milvus milvus*) observations in the same year (and continued the following years) helped to collect more than 4,000 accepted observations of this bird of prey until 2014 (Röller 2015). This campaign made Red Kite the most frequently observed species in the years 2011-2015.


a) Sum of all species groups

b) Most frequent species groups

Figure 2.1.4: ArtenFinder, yearly numbers of accepted observations 2011-2016.

The accepted ArtenFinder observations used in this work were reported by 1,342 different observers. Typically for crowdsourcing projects, the level of observation activity per user, in this case measurable as the number of accepted observations contributed to the project, shows a high concentration on few very active observers, while most contributed very little. Table 2.1.1 shows the distribution of ArtenFinder observations per user for all users who contributed at least one public, accepted observation in the years 2011 to 2016. Numbers are based on the date of observation, which represents the date of the actual observation activity, not of the submission to the ArtenFinder portal.

Table 2.1.1: ArtenFinder, number of accepted observations per user. (Based on active users in 2011-2016 with at least one public, accepted observation.)

No. of observations per observer	No. of observers	% of observers	% of observations
10,000 and more	6	0.45	39.37
1,000 to 9,999	41	3.05	46.45
100 to 999	95	7.07	10.44
10 to 99	281	20.92	2.96
1 to 9	920	68.50	0.78

The development of the number of active observers per year (users who contributed at least one public, accepted observation in a given year) is visualized in Figure 2.1.5. It is interesting to note that 2012, the second year of the ArtenFinder project, saw a maximum of 646 observers submitting observations which were accepted, while numbers of active observers have been dropping since then. The 2012 maximum was probably boosted by campaigns asking observers to report sightings of Red Kite and of Stag Beetle (see also above). The campaigns were quite successful and motivated hundreds of citizens to take part[18]. However, many observers contributed just one or a few observations to the campaign, but did not become long-term observers in the ArtenFinder project. For instance, 199 users reported one sighting of Red Kite or Stag Beetle in 2012, but did not contribute any more after that.

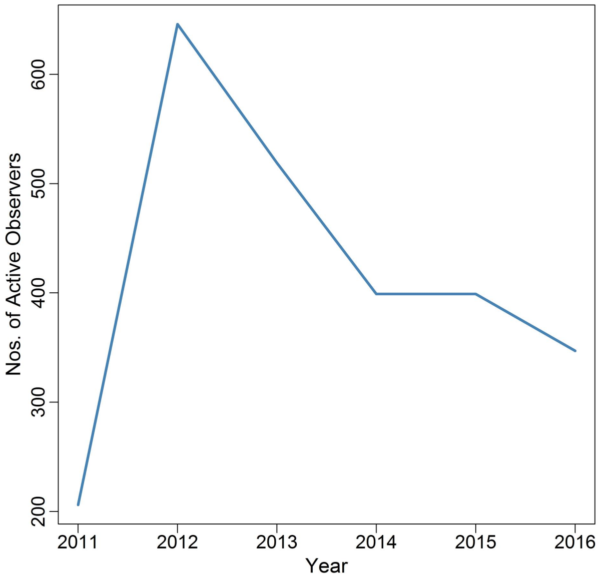


Figure 2.1.5: ArtenFinder, number of active observers. (Based on users with at least one public, accepted observation per year).

Spatial data properties

ArtenFinder data used in this work have an overall density of 14.4 accepted observations per square kilometer. However, one of the most important characteristics of the dataset is a pronounced inequality in spatial distribution of observations which follows roughly a northwest to southeast trend of increasing observation density. This trend can easily be visualized by a quadrat count map, as shown in Figure 2.1.6.

[18] https://artenfinder.rlp.de/node/3, last accessed on 2017-07-13

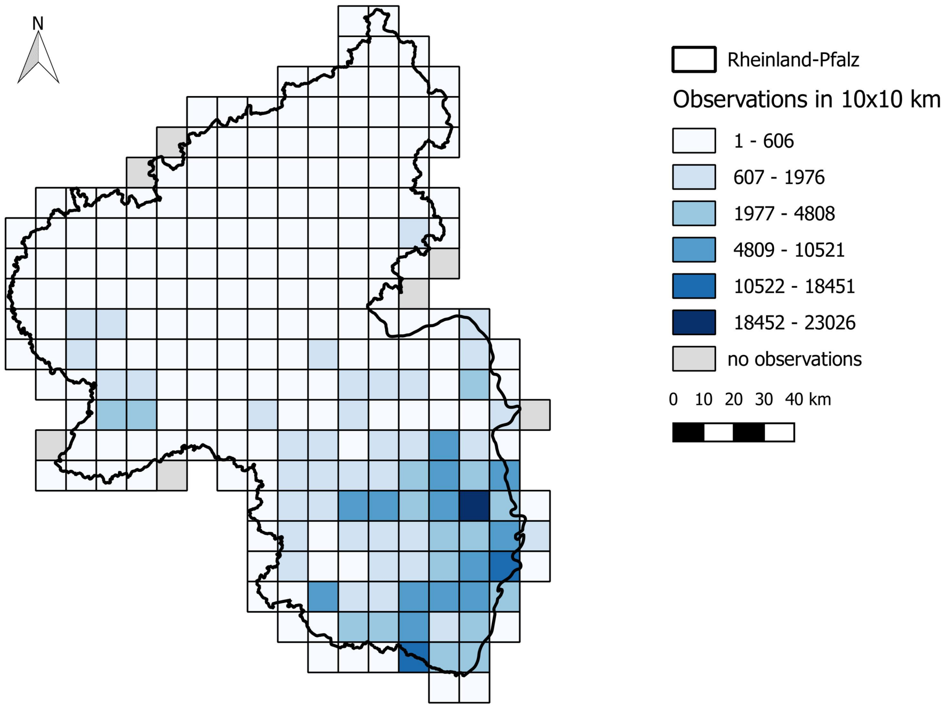


Figure 2.1.6: Regional differences in the density of accepted ArtenFinder observations. (Based on observations up to and including 2016. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.) n = 284,962.

Box 1: A Note on the Modifiable Areal Unit Problem

The quadrat count map in Figure 2.1.6 is subject to the modifiable area unit problem (MAUP, a well-known problem discussed in detail, for example, in O’Sullivan & Unwin 2010), and so are all following maps of this kind in this work. Choosing different quadrat sizes and positions would modify the outcome. Also, edge effects were not corrected, so that quadrats at the margins of the area of interest underestimate density because they contain areas outside of the area of interest. However, these maps are used to visualize only general spatial trends in observation distribution. The drawbacks of these map types are therefore considered to be of minor importance, but should be kept in mind when reading these maps. Regular grids are also a common means of visualizing spatial distributions of species in so-called distribution atlases (e.g., Wink 1987, Ebert & Rennwald 1991, Netzwerk Phytodiversität Deutschland e.V. & Bundesamt für Naturschutz 2013). In Germany, a standard for quadrat size and position of such grids is the ca. 10x10 km size and sheet line system of the official topographic maps of scale 1:25,000. The grid used here for ArtenFinder observation density maps has the same width. However, it does not follow the sheet line system of this map series, but was specifically created for this work to fit Rheinland-Pfalz state borders. The same considerations apply to California observation density maps in this work (e.g., section 2.1.2). MAUP is also present in choropleth maps (e.g., Figure 2.1.7).

Comparison of spatial density of ArtenFinder observations with population density in Rheinland-Pfalz (Figure 2.1.7) shows that there is an impact of the latter on the former in ArtenFinder observation data, but that other factors are also important. Higher observation density in the southeastern part of the state coincides with somewhat higher population density when compared to the northwest of the state, where both observation and population density are low. However, relatively high population density in the north of the state is not reflected in the spatial distribution of observation data at all. A probable cause is that the ArtenFinder project was initiated and is closely cooperating with Pollichia e.V., an association for natural history research which is active mostly in southern Rheinland-Pfalz. Many ArtenFinder volunteers are members of this association.

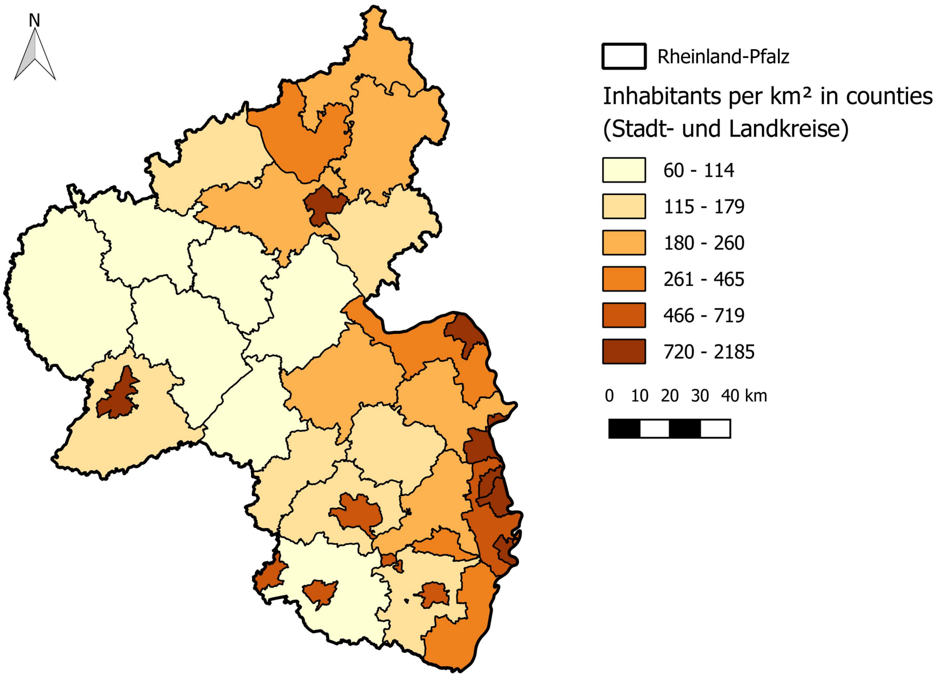


Figure 2.1.7: Regional differences in population density in Rheinland-Pfalz. State of population data: 2016. (Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz. Source of county borders: Landesamt für Vermessung und Geobasisinformation Rheinland-Pfalz[19]*. Source of population data: Statistisches Landesamt Rheinland-Pfalz).*

Embedded in the spatial trend in observation density described above, we find a pronounced clustering of observations. It is caused by some of the principle biases which are inherent in almost all casual citizen science datasets from the biodiversity domain. Observations cluster at places where observers are, or where they go. These include recreational areas, but also traffic routes (including footpaths, hiking trails etc.), settled areas, and local biodiversity hotspots such as conservation areas, water bodies, and the like. Clustering of observations is clearly visible in a dot map of spatial observation distribution (Figure 2.1.8). There are several approaches to statistically assess the clustering in spatial distributions of points (O'Sullivan & Unwin 2010). One of the more elaborate methods to test whether a point pattern shows a tendency towards clustering uses the so-called L function (for details about this method see O'Sullivan & Unwin 2010 and Baddeley et al. 2015). Stationarity of the examined point process is required for the L function to perform properly as a method to find out whether a point pattern is "consistent with" clustering (a wording proposed by Baddeley et al. 2015, p. 207). This is not the case over the whole ArtenFinder project region. Therefore, two data windows were selected. Figure 2.1.9 shows the empirical L function (based on the actual observation points) and the theoretical L function for a Poisson process (i.e., with points distributed randomly in space), for both data windows indicated in Figure 2.1.8 (calculated with R's "spatstat" package, with edge effect correction, duplicate points removed from the pattern). For the northern data window, the process can be considered to be close to stationary and the data are clearly consistent with clustering (see notes on this procedure in Box 2). It is important to stress that results of the L function analyses presented here are dependent on the position and size of the data windows employed. Changing size and/or position would lead to different results. This is illustrated by Figure 2.1.9b showing the empirical L function and the theoretical

[19] ©GeoBasis-DE / LVermGeoRP 2018, dl-de/by-2-0, www.lvermgeo.rlp.de [Data modified]

L function for a Poisson process, for the southern data window indicated in Figure 2.1.8. Following the discussion in Baddeley et al. 2015, this graph may provide evidence that the point process is not stationary within the southeastern data window. With all due caution however, we can say that accepted observations in ArtenFinder concentrate at certain localities, that these concentrations have different sizes with a spatial trend towards larger concentrations in the southeast of the project region. For the behavior of the potential plausibility indicators examined here, it is not important to formally prove that the data are clustered. It is, however, of importance that observations concentrate locally. This is undoubtedly the case.

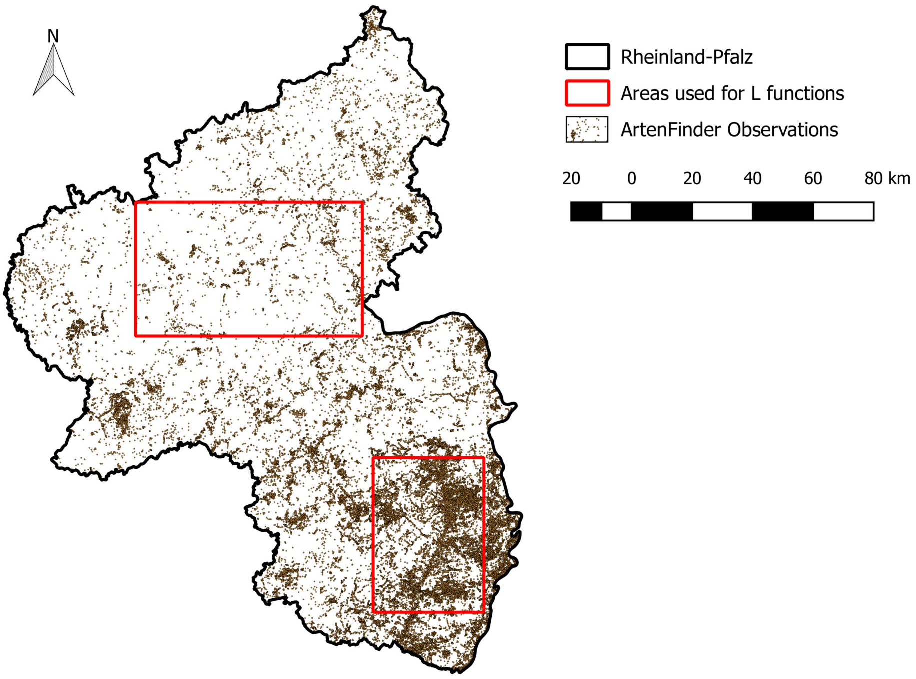
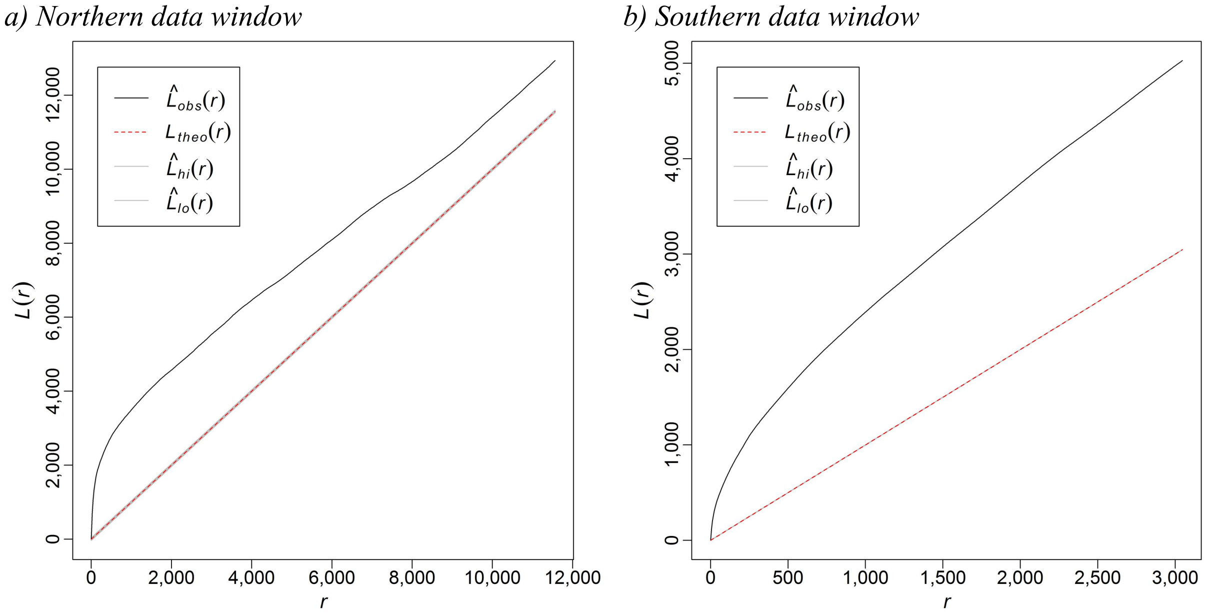


Figure 2.1.8: Dot map of ArtenFinder observations. Red rectangles: windows used for calculation of L functions (see text and Figure 2.1.9). (Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)

Box 2: A Note on L Functions

L functions are a modification of the K function, a method introduced by Ripley (1976), to determine whether a point pattern tends towards clustering, whether its points are evenly distributed, or whether they are randomly distributed. The underlying principle is counting points within a radius around a point, to do so for all points in the pattern and for successively growing radii, and to see whether the mean number of points in the radius is smaller or larger than expected for a random point pattern. If it is larger for relatively small radii, the point pattern is probably clustered. In this case, the graph of the empirical L function (based on the actual point pattern) initially takes a course above the graph of the theoretical L function, converging back to the theoretical L function for larger radii. One important assumption for this method is that the point process' intensity is homogeneous throughout the whole pattern examined (Baddeley et al. 2015).


a) Northern data window

b) Southern data window

Figure 2.1.9: ArtenFinder, empirical and theoretical L functions for the data windows indicated in Figure 2.1.8. Black line ($L_{obs}(r)$): empirical L function. Red dashed line ($L_{theo}(r)$): theoretical L function for a Poisson process. The gray envelope shows the global max. ($L_{hi}(r)$) and min. ($L_{lo}(r)$) deviations from the theoretical L function, using 100 simulated realizations of complete spatial randomness. r = radius in m. Duplicate points (points with identical coordinates) were removed for this calculation.

2.1.2 iNaturalist

The project and its area of interest

Like ArtenFinder Rheinland-Pfalz, iNaturalist provides a web portal and an app for collecting observations of organisms from the general public. The project began as part of a master thesis at UC Berkeley, was formally organized in 2011, adopted by the California Academy of Sciences in 2014, and is also a joint initiative with National Geographic Society since 2017[20]. In contrast to the ArtenFinder project, iNaturalist has a worldwide scope, but, as an American initiative, has a pronounced focus on the U.S. (which produce ca. 80% of observations) and especially California (with ca. 30% of U.S. observations). California (see Figure 2.1.10 and Figure 2.1.11) was therefore selected as the area of interest for the iNaturalist data use case.

In iNaturalist, verification of data, especially correctness of species identification, is exclusively governed by mutual confirmation or disagreement within the community of volunteers. In fact, species identification by the observer is not a mandatory part of an observation submission. Species identification can be entirely left to the community. Due to the nature of casual citizen science observations of organisms (which will be discussed in detail in section 2.2) this is, of course, only possible if sufficient evidence is attached to an observation, usually in the form of photographs. iNaturalist observations can therefore exist in three different quality grades: “casual”, “Needs ID”, and “Research Grade”. A “casual” observation misses one or more of the information components required for an observation to be considered verifiable (and to eventually become research grade). In most cases, such records do not provide evidence in the form of photographs. An observation which is submitted with a location, a date, and photo or sound evidence, has the category “Needs ID”. This happens regardless of whether the observer provided a species guess or not. Other members of the iNaturalist community can then

[20] https://www.inaturalist.org/pages/about, last visited 2018-09-13

add their own species identifications, in agreement or disagreement with the record's observer or other persons. "Research grade" status will be reached if and when two thirds of the people involved agree on the same species-level identification. Agreement can also be reached on higher taxonomic levels, which will give the record a community identification, but will usually not lead to research grade status. For research grade, there need to be at least two concordant species identifications (e.g., the observer's and one other person's). With the "identotron", iNaturalist provides a tool for looking up checklist information and existing observations for an observation location to help observers with species identification and to support co-observers in confirming species identifications. The portal also provides general information on species from wikipedia, range maps, taxonomy, conservation status from the IUCN Red List and other organizations, and references to physically similar species.

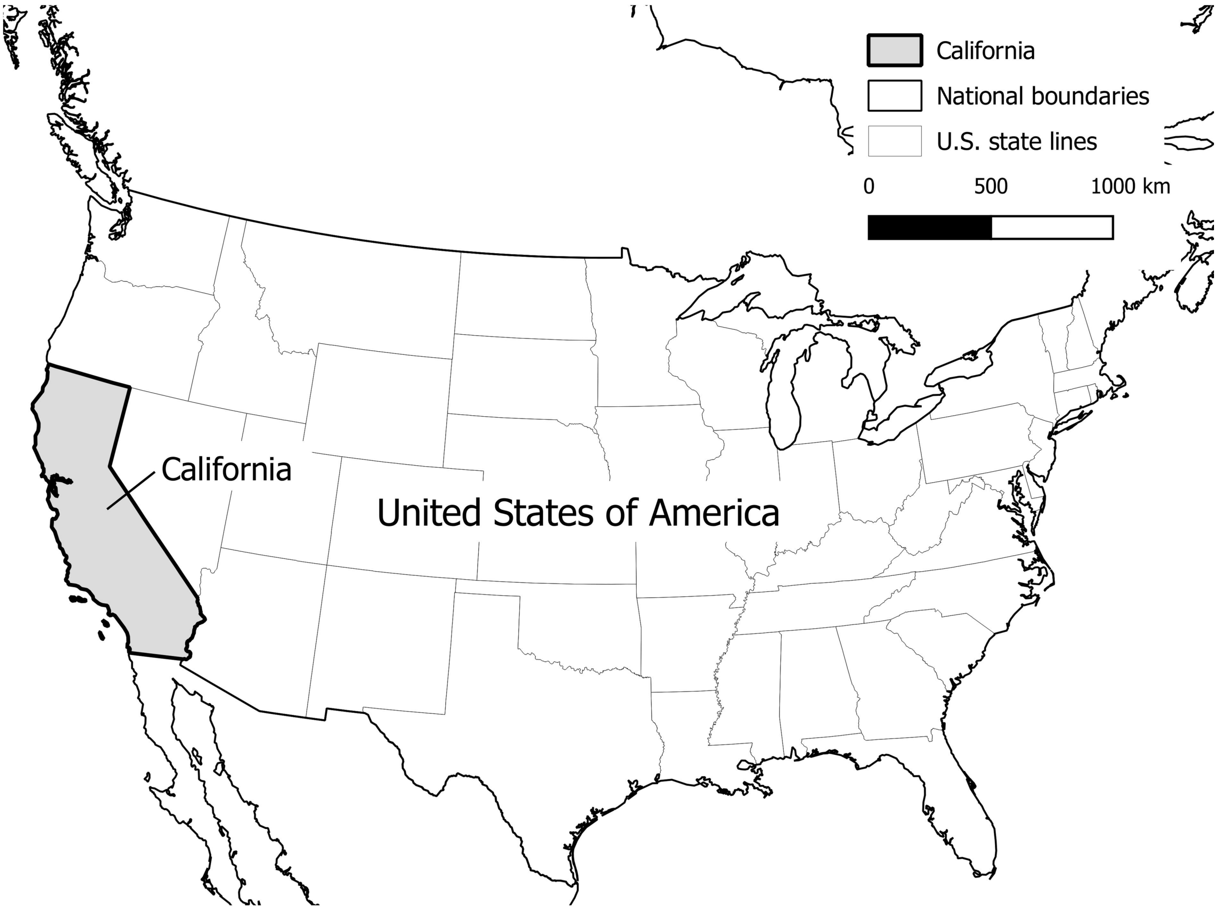


Figure 2.1.10: Geographic situation of the U.S. state of California. Area: 423,970 km^2. (Source of national and state boundaries: ESRI[21]*.*

Technically, there is no mandatory information that an observation submitted to iNaturalist must carry. Volunteers trying to submit observations without a date or location are warned about this fact, but can choose to go on. Species identification by the community is an important part of the project concept of iNaturalist. It is therefore o.k. for a volunteer to submit observations without specifying a species name. Identification by the community, of course, needs evidence. Submission of one or more photos with a report is therefore strongly encouraged. For instance, to start the submission process of observations via the online portal in its current state, a volunteer first has to select the photographs belonging to the observations he or she wants to submit. Photos can be removed before final submission, so that records without photos can exist, but these records then cannot become research grade, even if many participants agree on the species identification (which is, of course, improbable anyway without prop-

[21] Portions of this document include intellectual property of ESRI and are used herein by permission. Copyright © 2018 Environmental Systems Research Institute, Inc. All Rights Reserved

er evidence). Photos are therefore a very important element in iNaturalist, to the point that some authors call it a photo sharing platform (e.g., Jackson et al. 2015).

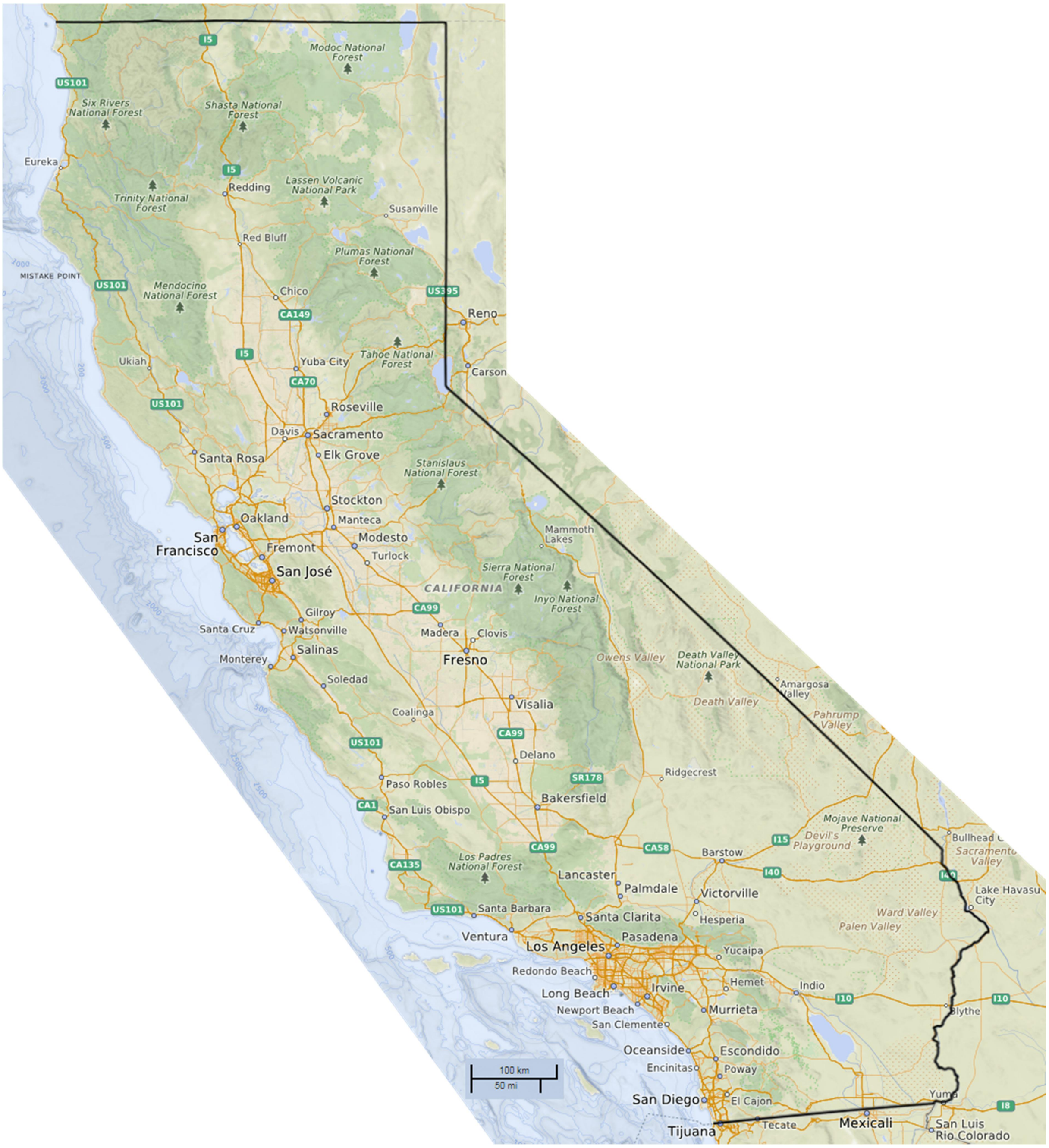


Figure 2.1.11: California, topographic overview. (Source of base map: OpenStreetMap[22]. Source of state line (black): U.S. Geological Survey 2016.)

This has, of course, an important impact on the selection of species which are reported to iNaturalist and on the type of observations which can reach research grade. Species reported to iNaturalist will predominantly be species which are suitable for photography, and species reaching research grade will be those whose most important and characteristic features can be recognized from a photograph. An observation even of the most skilled of volunteers and with a very plausible location and observation date cannot reach research grade if it does not provide photographic evidence. Alternatively, a sound

[22] © OpenStreetMap contributors; OpenMapSurfer; further data sources used there: CIAT-CSI SRTM, ASTER GDEM2, ETOPO1, GeoNames, GlobCover, NaturalEarthData, Scripps Institution of Oceanography, UC San Diego

recording can be submitted, but this is as yet very rare in iNaturalist observations; as of August 8th 2017, there were only 7,298 observations carrying a sound recording in iNaturalist (1,594 also having one or more photos), 5,504 of which had reached research grade, 679 of them in California, 481 observed 2016 or earlier and thus part of the data used here. For many species, it is very hard or not at all possible to provide conclusive evidence for species identification in a photograph. These species are virtually excluded from the research grade data pool produced by the project.

Thematic data properties

To be comparable to ArtenFinder data, iNaturalist species were grouped into the species groups used in the ArtenFinder project (see section 2.1.1). Project properties described above may be responsible for a distribution of observations in species groups which is markedly different from ArtenFinder Rheinland-Pfalz (see Figure 2.1.12). Most research grade observations are of plants at 33.63%. Birds (which are leading the ArtenFinder record) rank a close second (29.72%). Other species groups make up much smaller portions of the data: Butterflies and moths, mollusks and reptiles follow at 5.12-6.47%, mammals at 4.06%. Overall, the iNaturalist dataset is thus strongly dominated by two species groups (with plants and birds making up over 60% of the data), while in the ArtenFinder dataset where nearly 50% of observations are of birds. Plants leading the iNaturalist record may be a consequence of the importance of providing photographs with observations, which are, of course, much more easily taken of sessile organisms, but there may be other reasons. Birds are generally popular with observers, in most projects

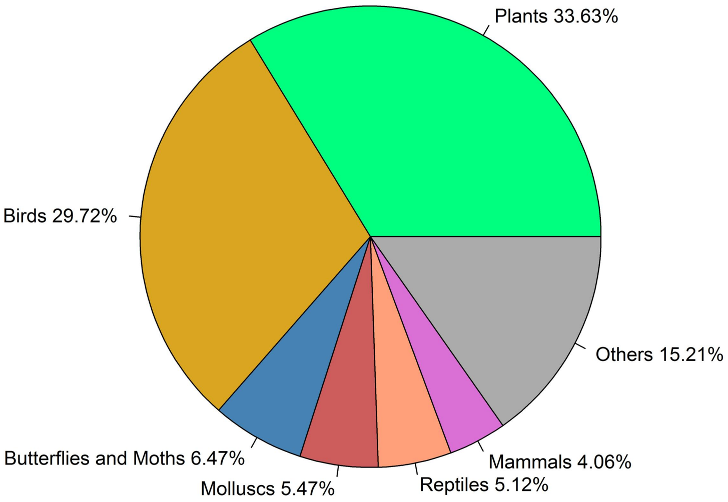
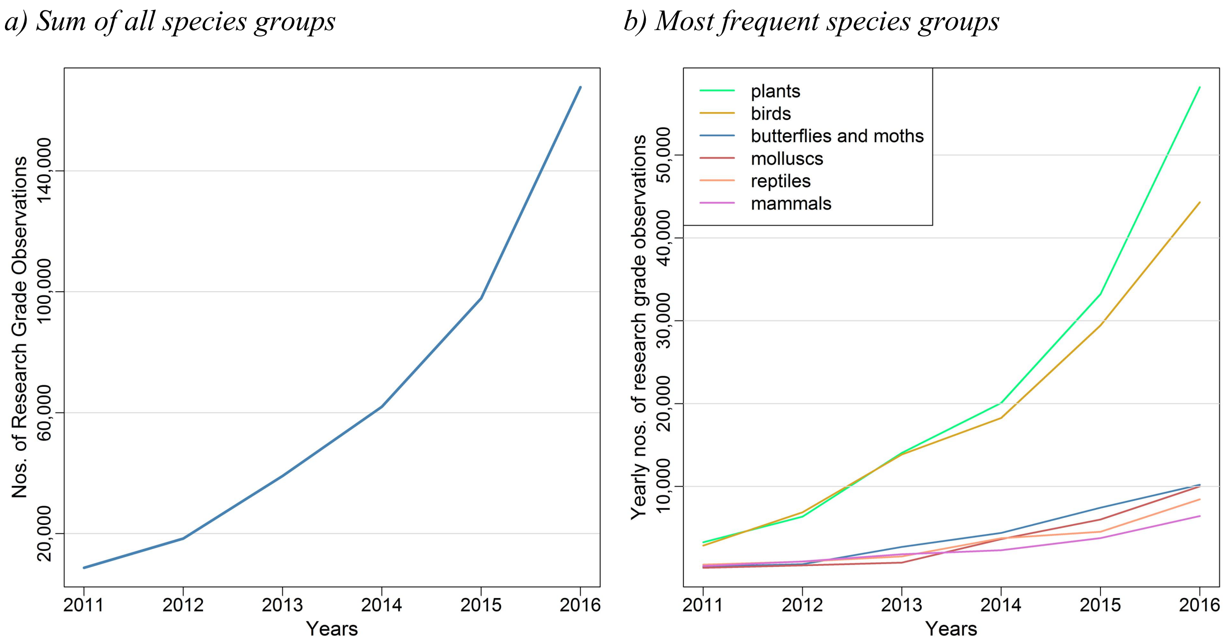


Figure 2.1.12: iNaturalist, portions of species groups. (Based on research grade observations from California up to 2016).

Figure 2.1.13 shows the development of yearly numbers of research grade observations. Numbers have been going up, as has the growth rate in yearly research grade observations. The development shows that growth rates of yearly research grade observations are also highest for plants and birds.



a) Sum of all species groups

b) Most frequent species groups

Figure 2.1.13: iNaturalist, yearly numbers of accepted observations 2011-2016.

iNaturalist currently has over one million users signed up (as of 2018-12-19; current numbers are published continuously on the project site start page). Table 2.1.2 shows the usual concentration of observation and reporting activity on a small portion of all users, in this case for all users having contributed at least one research grade observation in California in 2011-2016. Concentration of observations on observers is even stronger than in the ArtenFinder dataset: portions of observers with high observation numbers are smaller and those of observers with low observations numbers larger, respectively. Reporting observations is not the only possible form of contribution in iNaturalist. Adding species identifications to observations contributed by other observers is another important form of activity, which is not reflected in observation numbers per user. Numbers of iNaturalist observers contributing observations in California are growing fast, see Figure 2.1.14. In contrast to ArtenFinder, growth rate was still accelerating in 2016.

Table 2.1.2: iNaturalist, number of research grade observations per user. (Based on active users in California in 2011-2016 with at least one research grade observation).

No. of observations per observer	No. of observers	% of observers	% of observations
10,000 and more	2	0.02	8.59
1,000 to 9.999	64	0.50	44.42
100 to 999	340	2.67	23.82
10 to 99	2448	19.22	16.36
1 to 9	9881	77.59	6.82

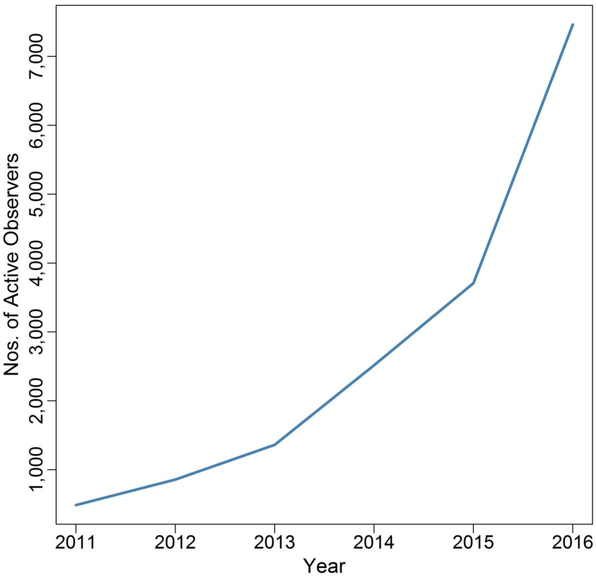


Figure 2.1.14: iNaturalist, number of active observers in California. (Based on users with at least one research graded observation per year.)

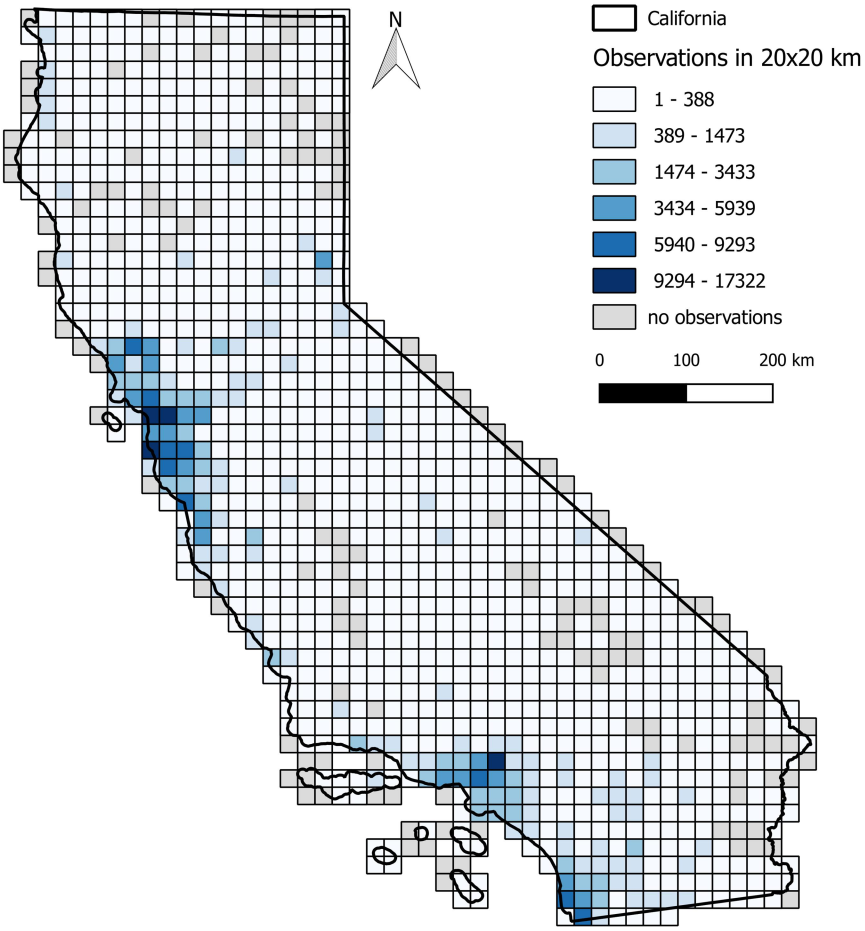


Figure 2.1.15: Regional differences in the density of research grade iNaturalist observations in California. (Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016.)

Spatial data properties

California has 0.97 research grade iNaturalist observations per square kilometer. This number was calculated using the state area value of 423,970 km^2 given in the attributes of the state line in the USGS National Boundary Dataset (U.S. Geological Survey 2016). These data include the US State Submerged Lands (a maritime zone of 5.57 km width) along the pacific coast. Only seven observations used in the observations dataset are located in the sea outside of this zone. The map in Figure 2.1.15 visualizes observation density in a quadrat count map. For this map (and all following maps of this kind in this work) the same considerations concerning MAUP apply which were already discussed for the according Rheinland-Pfalz maps in section 2.1.1 (see Box 1). A larger grid size of 20x20 km was chosen for California maps, to maintain readability at smaller scale. iNaturalist research grade observations concentrate in roughly two regions, around San Francisco and around Los Angeles. Large parts of California have rather low observation densities, with lowest concentrations in southeastern and northern California, as well as in the southwestern Long Valley, where large areas can be found which do not have research grade observations at all. In contrast to ArtenFinder data from Rheinland-Pfalz, spatial distribution of iNaturalist observations fits the pattern of population density of California (Figure 2.1.16) quite well.

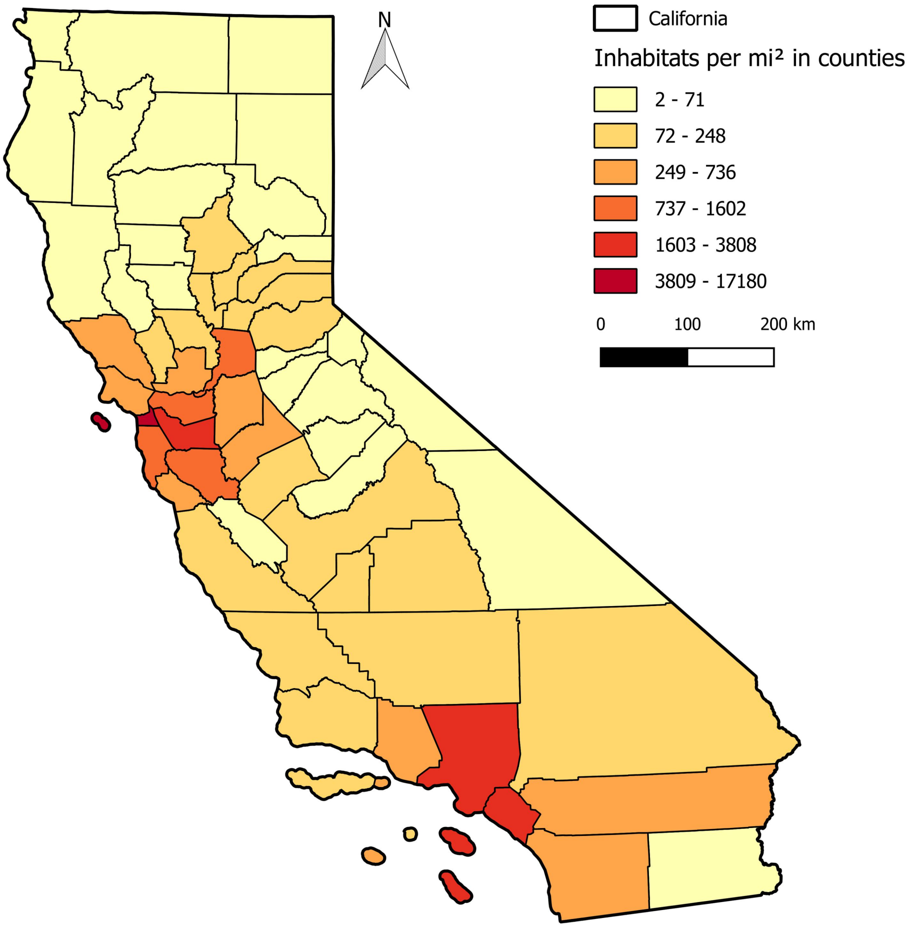


Figure 2.1.16: Regional differences in population density in California. (Classified by Natural Breaks. Source of state line and county borders: U.S. Geological Survey 2016. Source of population data: U.S. Census Bureau 2012.)

Like ArtenFinder data, California iNaturalist observation data show clustering, for the same reasons (see section 2.1.1). Again, clustering of observations is clearly visible in a dot map of observations (Figure 2.1.17). Figure 2.1.18 shows empirical L functions and theoretical L functions for a Poisson process, for the areas drawn in red in figure 3.1.21 (again, calculated and plotted using the R's "spatstat" package, with edge effect correction, duplicate points removed from the pattern). They indicate that the point pattern is consistent with clustering (Baddeley et al. 2015) within the test areas (test areas were again chosen for approximate stationarity of the observation process within these areas, which is not the case for the point pattern as a whole). Again, it is important to stress that results of the L function analyses presented here are dependent on the position and size of the data windows (test areas) employed. Changing size and/or position would lead to different results. However, the test results presented here are able to indicate a tendency towards clustering, which is an important property of the iNaturalist data used in this data use case. See also section 2.1.1, Box 2 for some notes on the principle underlying the L function.

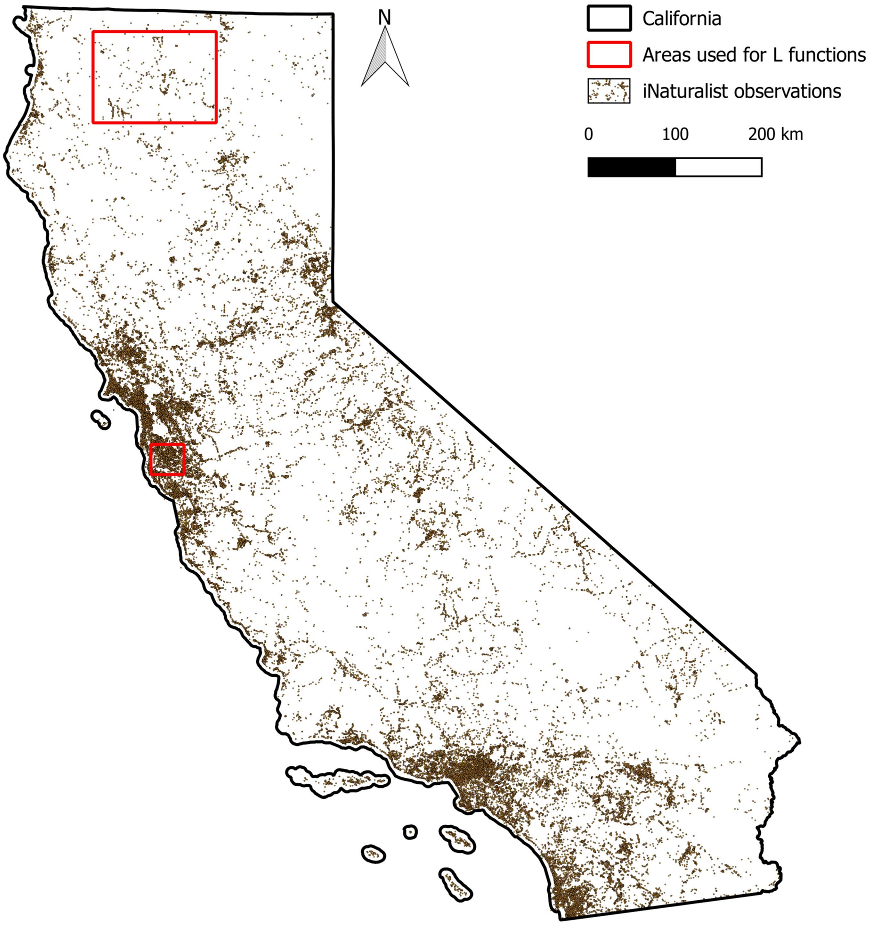
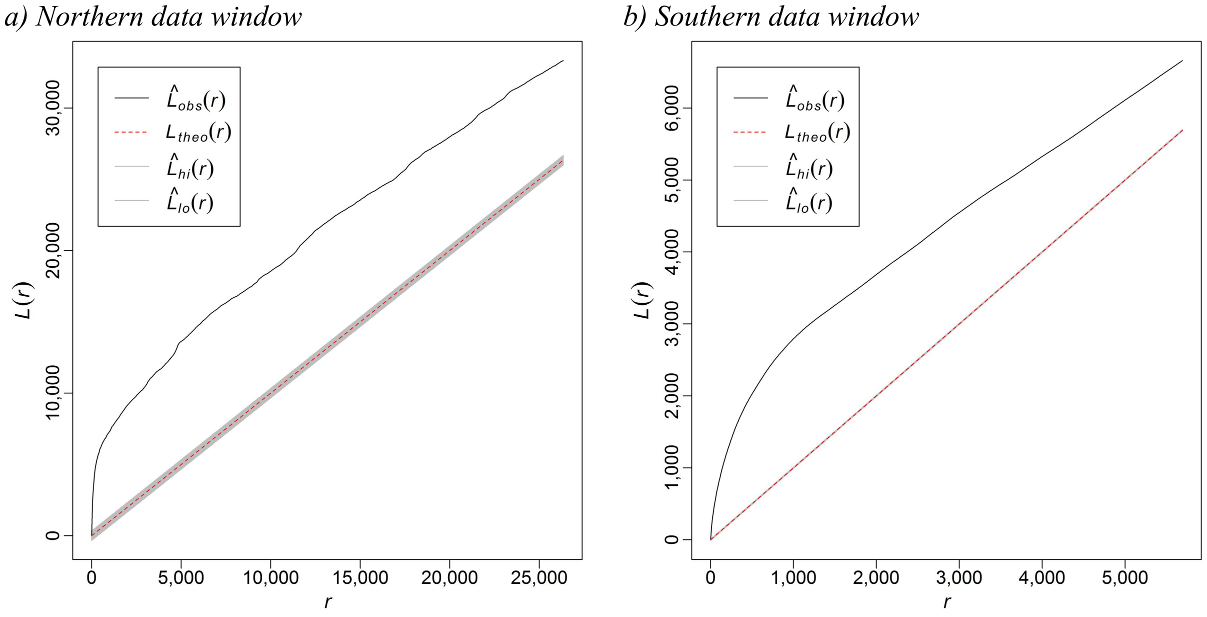


Figure 2.1.17: Dot map of iNaturalist observations in California. Red rectangles: windows used for calculation of L functions (see text and Figure 2.1.18). (Source of state line: U.S. Geological Survey 2016.)



a) Northern data window *b) Southern data window*

Figure 2.1.18: iNaturalist, empirical and theoretical L functions for the data windows indicated in Figure 2.1.17. Black line ($L_{obs}(r)$): empirical L function. Red dashed line ($L_{theo}(r)$): theoretical L function for a Poisson process. The gray envelope shows the global max. ($L_{hi}(r)$) and min. ($L_{lo}(r)$) deviations from the theoretical L function, using 100 simulated realizations of complete spatial randomness. r = radius in m. Duplicate points (points with identical coordinates) were removed for this calculation.

2.1.3 OpenStreetMap

The project

The OSM project, its history, goals, and general properties have been described in a large number of publications (e.g., Neis & Zipf 2012, Budhathoki & Haythornthwaite 2013, Neis & Zielstra 2014). Only the most important facts will therefore be reiterated here, but more will emerge in the discussion following this overview (section 2.2).

Founded by Steve Coast, the OSM project started out in 2004 at University College London. The project was soon organized in a foundation and gained worldwide scope. Creating a new object in OSM basically requires the same steps as in iNaturalist or ArtenFinder, namely creating a geometry object and adding attribute information to it in the form of tags. Tags are pairs of a key and a value, e.g. “highway=residential”. At first, contributions were mainly based on GPS tracks collected by volunteers in the field, but modes of data production changed when several Internet companies and public authorities provided satellite and aerial imagery for tracing objects for OSM (Neis & Zielstra 2014). Besides contributions by individual volunteers, OSM data are sometimes also created by bulk imports of geographic data from open data sources. For the data used here, the most important of these imports occurred in 2008, adding to OSM parts of the U.S. census bureau’s TIGER (Topologically Integrated Geographic Encoding and Referencing) data[23], such as roads, railroads, and administrative boundaries, but also statistical geographic areas (Zielstra et al. 2013).

OSM’s quality assurance strategy is a classic example of what Goodchild & Li (2012) call the “crowdsourcing approach”, which means that quality control is placed in the hands of the community of volunteers, and based on the principle that errors cannot persist if many volunteers work on the data and

[23] https://www.census.gov/geo/maps-data/data/tiger.html, last accessed on 2018-09-20

eventually correct errors other volunteers might have made. OSM's quality assurance strategy is therefore closely related to iNaturalist's, where the community of observers reviews, comments, confirms, or contradicts the information given in an observation record (especially species identification), albeit with important differences concerning the nature of information, which will be discussed in section 2.2. It is, however, in strong contrast to ArtenFinder's quality assurance strategy, which employs privileged experts who validate observations, while normal users are not able to comment directly on other observers' reports.

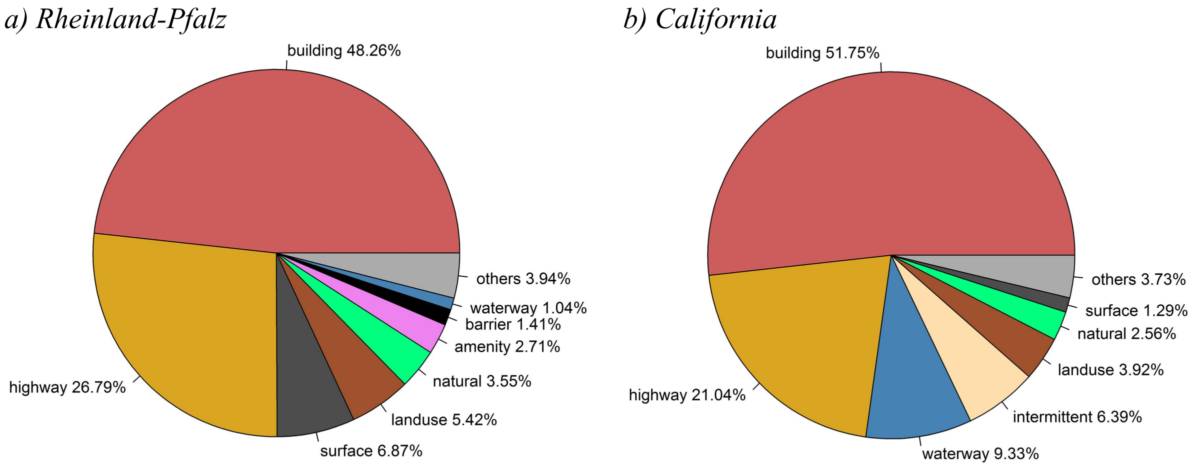
Thematic data properties

Thematic information in OSM is represented by tags in the form of key-value pairs, see above. Thematic information in OSM is different and also more varied in nature than observations of organisms. This makes describing relevant thematic properties of OSM data more complex, compared to describing properties of biodiversity observation data. For instance, an observation of a species has just one species identification, belonging to just one species group (e.g., a bird or a butterfly). It is therefore straightforward to aggregate data by calculating portions of observations belonging to a certain species group. An OSM element, on the other hand, may carry several tags at the same time. Analyzing the thematic properties of the OSM data used in this work requires some form of aggregation (analog to aggregation by species group), which was done by key. The data structure explained here must be kept in mind when interpreting statistics describing these data. Also, not all available OSM tags were used in this work, but only tags holding information which was judged relevant for the intended use of OSM information as a geographic context source. Details are explained in the Methods chapter, section 3.3.4, and tags selected as relevant listed in the appendix, section 7.2. The following analyses refer only to the data carrying these selected tags.

The Heidelberg Center for Geoinformation Technology (HeiGIT[24]) provides a service called the Ohsome API[25]. One of its functionalities is counting, in a set of OSM elements, the elements which carry a certain tag. This tool was used to examine the thematic properties of OSM data in the two areas of interest concerning the relevant tags used in this work. Due to the fact that an OSM element may carry several relevant tags at the same time, the basic population of such a count is not the number of elements, but the number of occurrences of tags (where some elements are counted several times). This count is able to convey some information on which tags can be encountered more often in the data than others. This is what Figure 2.1.19 tries to do. It shows a dominance of "building=*" and "highway=*" tags in all relevant tag occurrences in both Rheinland-Pfalz and California OSM data. Rheinland-Pfalz data also have notable portions of "surface=*" (mostly of some kind of highway, street or path), "landuse=*", "natural=*", "amenity=*", "barrier=*", and "waterway=*" tags. In California OSM data, "waterway=*" tags are more prominent than in Rheinland-Pfalz, and a relatively high portion of "intermittent=*" tags (all of them "intermittent=yes") reveal seasonally dry conditions in large parts of California. "landuse=*", "natural=*", and "surface=*" tags make up relatively small portions of all tags here.

[24] https://heigit.org/

[25] https://api.ohsome.org/v0.9

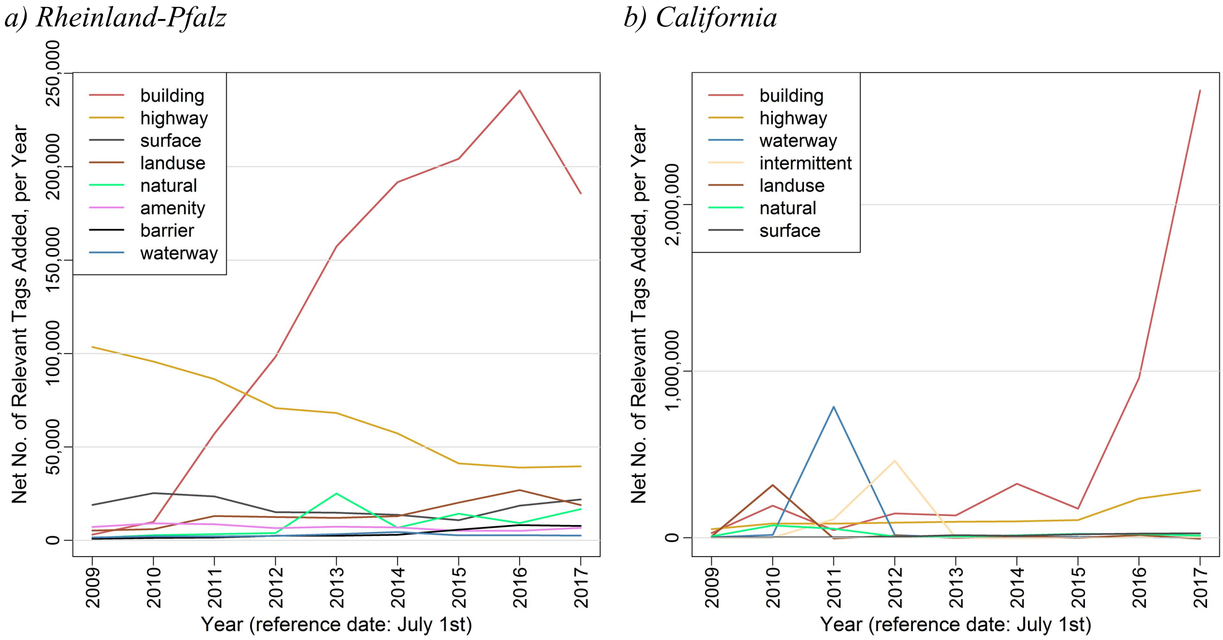


a) Rheinland-Pfalz

b) California

Figure 2.1.19: Portions of the most frequent keys in OSM extracts from Rheinland-Pfalz and California. State of OSM: 2017-07-01. n(a, occurrences of tags) = 2,380,703. n(b, occurrences of tags) = 9,071,835. (Analysis includes only tags used in this work.)

Figure 2.1.20 shows the development of yearly numbers of tags added to OSM (aggregated by keys as well). While most tags were added at steady and rather low rates to the Rheinland-Pfalz OSM, “highway=*” tags declined steadily from relatively high rates. “building=*” tags experienced a strong incline peaking in 2015 with a much higher rate than the other keys, and have declined since. The California extract of OSM data shows a more recent increase in “building=*” tags. “waterway=*” tags were mostly added before 2012-07-01, followed by “intermittent=*” tags. “highway=*” tags experienced a raise in recent years. Another difference to Rheinland-Pfalz data is that negative numbers occur, that is, a net removal of “intermittent=*” and “landuse=*” tags from the data in some years.



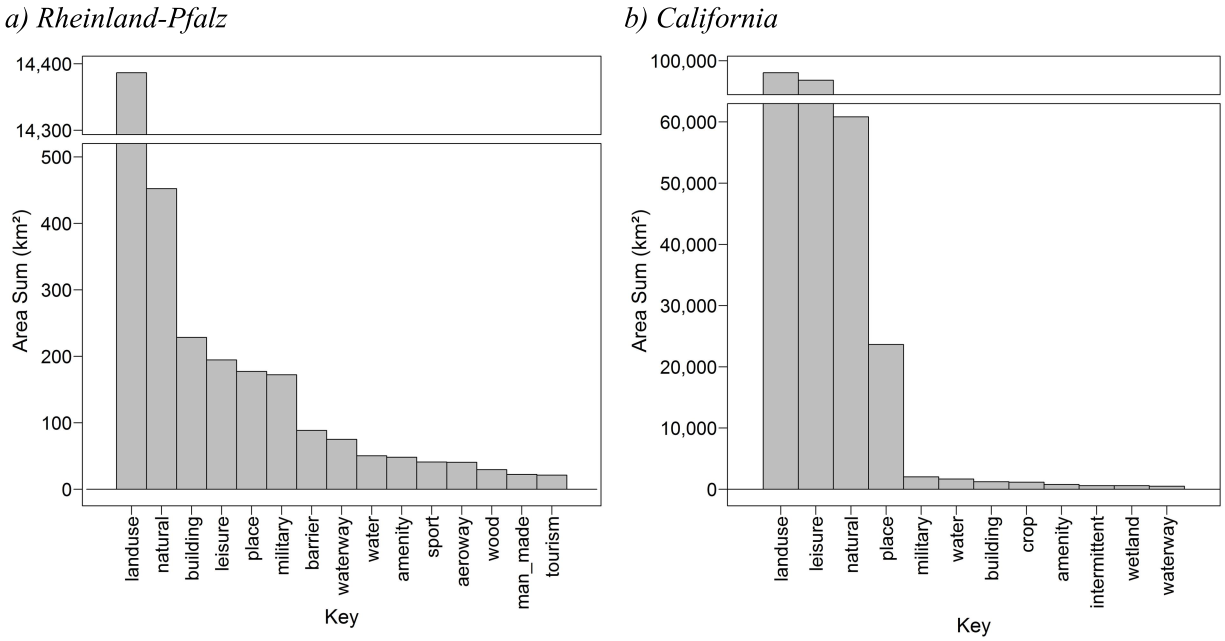
a) Rheinland-Pfalz

b) California

Figure 2.1.20: Yearly net numbers of tags from the most frequent keys added to OSM elements in Rheinland-Pfalz and California. (Analysis includes only selected tags used in this work).

However, caution has to be used with these analyses. Difficulties are introduced by the complex geometrical nature of OSM data, which consist of point elements (“nodes”), line elements (“ways”) and polygon elements (represented by some closed “way” objects or some “relations”). While “node” elements have properties which are comparable to point objects, ways and relations may or may not be segmented into a larger or smaller number of geometric objects. For instance, a highway may be represented in relatively long segments of several kilometers of length, or be split into many relatively short segments of just a few tens of meters. The latter case boosts tag counts in the above statistic. Also, buildings are relatively small and many, and consequently represented by many small, individual geometric elements, while land use units are usually relatively large and often also represented by large geometric elements in the map, boosting tag counts for the “buildings” key relatively to that of the “landuse” key. It makes sense, therefore, to also look at other ways of examining thematic data properties, especially by using area or length instead of counts, but always keeping in mind that a single element may carry several relevant tags and sums of areas or lengths incorporate overlaps, thematically as well as geometrically. This view on the data was implemented by looking only at elements of type way and relation and by looking at the most important keys representing mostly polygon elements. Figure 2.1.21 presents results. They may contain thematic overlaps between elements: a polygon my carry tags from several of the relevant keys and may therefore be contained in several of the area sums per key. Also, a polygon may carry several tags of the same key and may therefore be contained several times in the area sum for a key. The same considerations apply to linear objects which were examined by using lengths of way type elements of the most important keys representing mostly linear elements. Results ae presented in Figure 2.1.22.

Results of area calculations for Rheinland-Pfalz show that “landuse=*” tags, not very prominent in element counts, are by far the most important in terms of coverage of area, while tags which were prominent in element counts, especially “building=*”, are less important. Results with California OSM data are similar: “landuse=*”, “leisure=*” or “natural=*” tags cover large areas, while “building=*” tags (and others) are insignificant. In element length results for Rheinland-Pfalz and for California, keys are prominent which also reached high tag count results, such as “highway=*” or “waterway=*”.


a) Rheinland-Pfalz

b) California

Figure 2.1.21: Area of OSM elements, aggregated by keys, in Rheinland-Pfalz and California. (Sums of areas of way and relation elements, no correction for thematic and geometric overlaps). State of OSM: 2017-07-01. (Analysis includes only selected tags used in this work.)

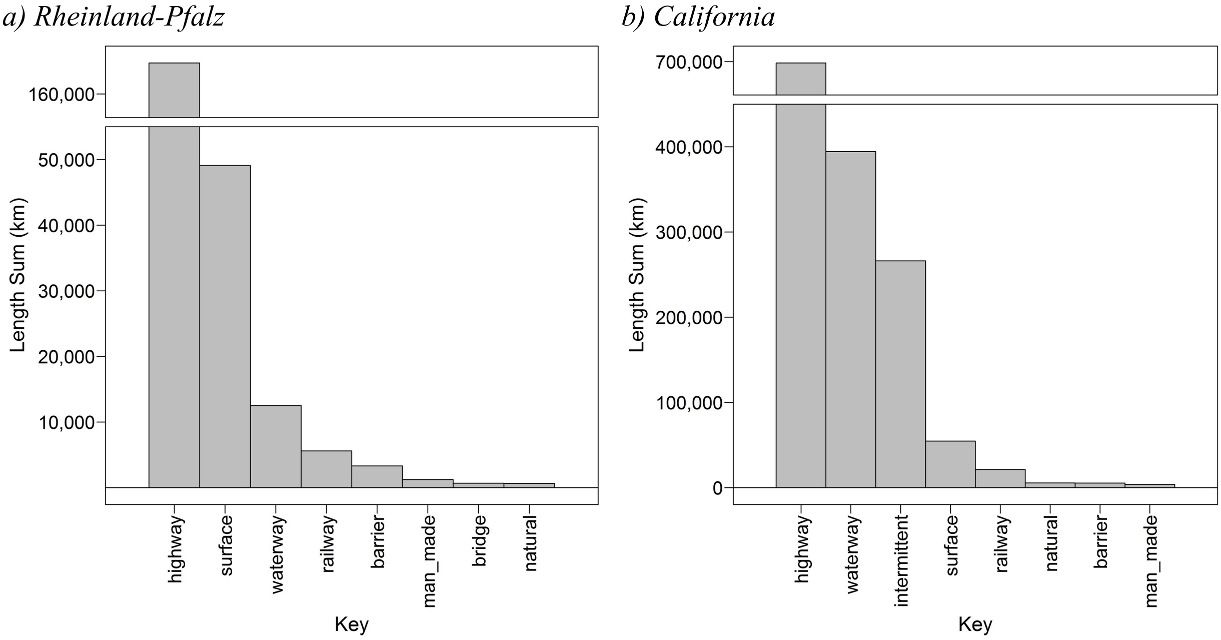


Figure 2.1.22: Length of OSM elements, aggregated by keys, in Rheinland-Pfalz and California. (Sums of length of way elements, no correction for thematic and geometric overlaps). State of OSM: 2017-07-01. (Analysis includes only selected tags used in this work.)

Numbers of active mappers were more or less constant in Rheinland-Pfalz in the years the OSM data used here were produced, with a rise in mapper numbers in earlier years and a slight drop towards later years (see Figure 2.1.23). In California, numbers of mappers have been rising constantly and have recently reached numbers comparable to Rheinland-Pfalz, but in a far larger area with a much higher population. In absolute numbers, participation in OSM is far superior to numbers of volunteers in ArtenFinder and iNaturalist. The fact that in OSM, just as in all VGI projects, a small portion of volunteers is responsible for most of the contributions, has been found and documented by many research works on OSM data (e.g., Neis & Zipf 2012). It can certainly be expected for the OSM data used in this work as well and was not examined here in detail.

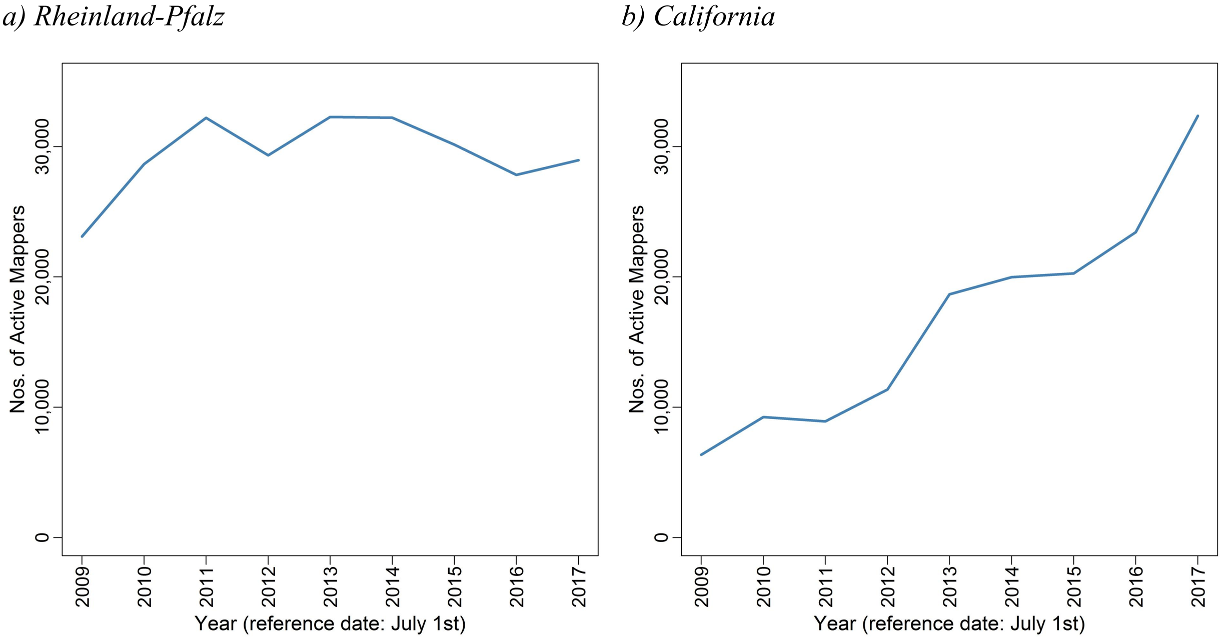


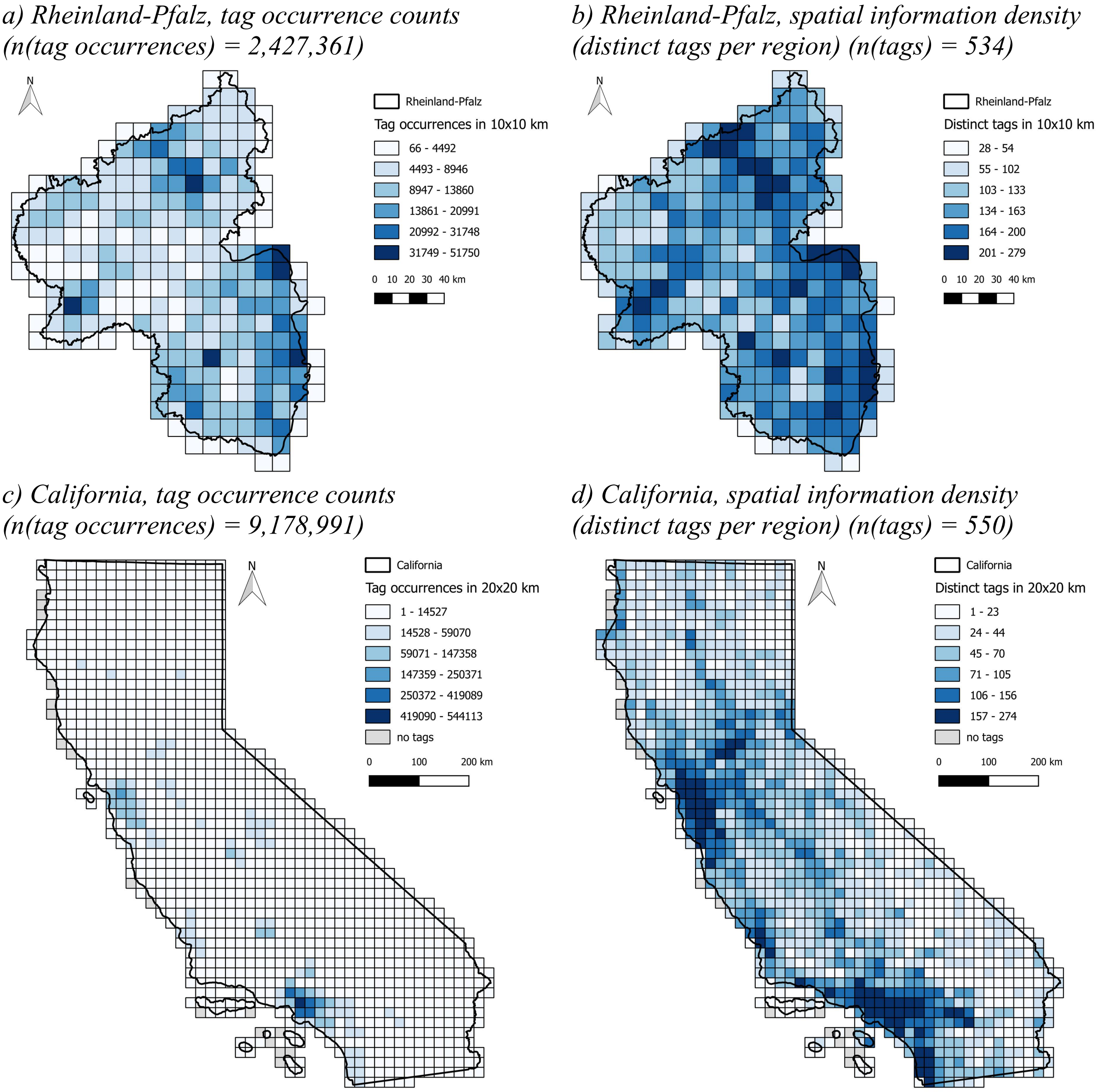
Figure 2.1.23: Numbers of active OSM mappers per year in Rheinland-Pfalz and California. (For technical reasons, only node and way objects were evaluated here.)

Spatial data properties

Spatial properties of OSM data cannot be described in the same way as spatial properties of casual citizen science observation data, due to reasons already explained, which are rooted in the OSM data's complex data structure. It is, of course, possible to count numbers of occurrences of a tag in an area, but the sum of these counts over all tags will be larger than the number of objects involved, because any object may carry more than one tag (and many do). A density map of this kind may still convey some relevant information about spatial properties of OSM data and is represented in Figure 2.1.24 (parts a and c). Keep in mind, however, that counts of tags depend on the structure of the objects which carry them, as was already explained above. If a tag is attached to strongly segmented objects, this will raise its count. This effect can be avoided by looking at the number of distinct tags in an area, a parameter which represents spatial information density, rather than actual density of geometric objects or tag occurrences. This is represented in Figure 2.1.24 as well (parts b and d).

Parts a and c of Figure 2.1.24 illustrate the well-known fact that OSM data tend to concentrate in urban areas (e.g., Haklay 2010). In both areas of interest, counts of tag occurrences very distinctly highlight the most important urban centers, such as Mainz, Ludwigshafen, Kaiserslautern, Trier and Koblenz in Rheinland-Pfalz, and San Francisco and Los Angeles in California. Maps of spatial information density (number of distinct tags in a region) do the same, but high-value regions include areas adjoining urban centers, and smaller cities also stand out, such as Landau or Neustadt an der Weinstraße in Rheinland-Pfalz, and San Diego, the Indio/Palm Springs area, Oxnard/Ventura, Santa Barbara, San Louis Obispo/Grover Beach, and Sacramento in the Long Valley in California. Contrast is weaker here, because spatial information density is still higher in urban centers, but contrast to other areas is not so strong in this parameter. However, spatial properties are basically the same with both parameters.

In Rheinland-Pfalz OSM data, there is no pronounced spatial trend like it was found in observation data. OSM data therefore are possibly able to provide sufficient geographic context for plausibility estimation also in regions of Rheinland-Pfalz where observation data are scarce. California OSM data, however, show a pattern which is quite similar to iNaturalist observation data spatial distribution, with concentrations in the San Francisco Bay and Los Angeles areas in both tag count and spatial information density. The latter parameter does also show (smaller) areas with higher values outside these areas.



a) Rheinland-Pfalz, tag occurrence counts (n(tag occurrences) = 2,427,361)

b) Rheinland-Pfalz, spatial information density (distinct tags per region) (n(tags) = 534)

c) California, tag occurrence counts (n(tag occurrences) = 9,178,991)

d) California, spatial information density (distinct tags per region) (n(tags) = 550)

Figure 2.1.24: Spatial distribution of OSM tags in Rheinland-Pfalz and California. (Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz. Source of CA state line: U.S. Geological Survey 2016.)

2.2 Comparative Discussion of Projects and Data Use Cases: Common Ground and Important Distinctions

The previous sections presented project properties, as well as the thematic and spatial properties of the data produced by ArtenFinder, iNauralist, and OSM. What are the most important aspects of commonality of these different kinds of VGI projects and their data, and where do they differ in important ways? What are the consequences which have to be taken into account when using these projects and their data? Comparison of the projects used in this work, and comparison of their data, is an effective way of revealing more of their properties of critical importance to this work and to deepen the understanding of geographical citizen science and VGI in general.

Both ArtenFinder and iNaturalist are dedicated to opportunistic data collection, producing so-called casual observation data (see section 1.3). OSM is very similar in this respect. Besides a general agreement among the members of the mapping community as to the kind of information which should be included in OSM, volunteers are not bound to any temporal, spatial, or other restrictions concerning their contribution activities. This is a very important common characteristic of the projects discussed here. The contributors and their behavior and choices are major factors which determine the properties of the data produced in this way. Probably the most important consequences of this are issues of spatial, temporal, and thematic completeness in the data which are the effect of several biases. They were briefly reviewed and some aspects further examined in Jacobs & Zipf (2017). Useful overviews and discussions of the major biases occurring in opportunistic citizen science observation processes are presented, among others, by van Strien et al. (2013) and by Isaac & Pocock (2015).

Volunteers taking part in casual citizen science observation projects usually do not, either individually or collectively, generate a regular pattern of sites from which observations are contributed. Also, sites are not distributed randomly. Several factors determine the volunteers' choice of locations for observation activity. Among these are the places of residence of observers and also, more generally, population density (Dennis et al. 1999), both clearly visible in the spatial distribution of ArtenFinder and iNaturalist data (discussed in sections 2.1.1 and 2.1.2). Observers are also attracted by locations which feature rare or uncommon species, or a high biodiversity providing the opportunity to observe many different species (Dennis & Thomas 2000). Such effects lead to what van Strien et al. (2013) call geographic bias, resulting in uneven spatial coverage (Isaac & Pocock 2015). In the absence of standardized field protocols, there will usually also be variance in the amount of time and in the size of an area a volunteer will cover in an observation activity (also called effort), resulting in so-called observation bias (van Strien et al. 2013). Another bias listed by van Strien et al. (2013), so-called reporting bias, is caused by the fact that many observers tend to be selective in which observations they actually report, and which they don't. Observers tend to concentrate on certain species groups, such as the very popular birds, or on rare or uncommon species, neglecting unpopular species groups or common species. This produces a taxonomic bias in the resulting data, an effect which Troudet et al. (2017) show not to be specific to citizen science data. However, it can be traced very well in the data use cases of this work. For instance, in ArtenFinder observation data up to 2015, 75% of observers with at least 100 observations contributed 50% or more of their observations from one species group, predominantly from birds or butterflies. iNaturalist observers are somewhat less specialized, with 68% of observers with 100 or more observations up to 2015 reporting predominantly one species group, most of them plants (followed by birds). Reporting bias may also arise from project properties, e.g., requirement of a photo proof, which excludes species for which it is difficult or impossible to provide it. Finally, not all species are equally easy to detect. This factor may even change in a species with seasons and life cycle. For instance, many insects, such as butterflies or dragonflies, are much easier to be found as imagines (fully developed adults) than as larvae. Many songbirds sing predominantly during their breeding season and are then much easier to detect than at other times of the year. Then of course, whole groups of species, although present at a location, are hard to be found at all times, due to behavior, habitat preferences, or simply size. Detection bias (van Strien et al. 2013) is therefore also an important factor for species composition of casual citizen science observation datasets. Closely related is the problem of difficulty of species identification. Many species require special techniques and methods for identification, which cannot be easily learned or used by non-professional observers (Isaac & Pocock 2015). Uneven sampling over time is another important bias in biodiversity observation data pointed out by Isaac & Pocock (2015), caused by rising yearly numbers of observations which have been collected over the last decades. This effect was also found within the relatively recent ArtenFinder and iNaturalist datasets used in this work. It should be noted that the biases described here are not restricted to casual citizen science. Williams et al. (2002) cite examples for spatial distributions of museum collection data which represent road or river networks, because these structures provide accessibility, and therefore these data show geographic bias as well. They also discuss cases of observation bias in these data. Boakes et al. (2010) provide more examples from museum collection data, but also from other sources, such as species distribution atlases and scientific literature on biodiversity.

Most factors described for casual citizen science data also take effect on OSM data. The most important of these is probably geographic bias, which results mainly in spatial inhomogeneity of completeness of OSM data and which was extensively examined in a large number of studies. Neis and Zielstra (2014) review most of these, which found OSM data to be much more complete in urban areas than in rural areas in several European countries. As already discussed in section 1.3, most studies used the road network as a completeness indicator and compared OSM data to some official or commercial dataset (e.g. Haklay 2010; Zielstra & Zipf 2010; Neis et al. 2011). Another study from the U.S. found a contrary trend for Florida (Hochmair 2011), which was interpreted as an effect caused by the TIGER bulk data import (see also section 2.1.3). OSM data of California, used in this work, seem rather to concentrate in population centers (see section 2.1.3). The bottom line for both casual citizen science projects and OSM is that the manner of data acquisition in the absence of strict protocols leads to biases in the resulting data, especially in their spatial distribution, which, as we will see later, is of some importance for this work.

More common ground of OSM, iNaturalist and ArtenFinder emerges when we look at the statements that these projects provide about their visions concerning the data they produce. OSM states that "We started it [the OSM project] because most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways"[26]. This statement reveals that there are no predefined data uses implied or intended. iNaturalist describes its vision as following: "Get connected with a community of over 750,000 scientists and naturalists who can help you learn more about nature! What's more, by recording and sharing your observations, you'll create research quality data for scientists working to better understand and protect nature."[27] The statement reflects the more specific nature of the data collected here when compared to OSM, but still opens a very broad field of goals and possible uses which the project's data can be put to. These statements show that both projects share a common approach in collecting data from the public and providing them back to the public, leaving the choices of what to do with these data entirely to the users. They stress that the data are free to use, not only in the sense of free of charge, but free to anyone interested in using the data, and free to any intended use, within the boundaries of the licenses that apply. ArtenFinder, in the same spirit, calls to the general public "to report observations of animals and plants, thereby creating a valuable contribution to research on nature and to nature con-

[26] https://wiki.openstreetmap.org/wiki/Main_Pagem, last accessed on 2018-09-20
[27] http://www.inaturalist.org/pages/about, last accessed on 2018-09-20

servation"[28]. An important reason for founding this project was tapping the potential of crowdsourcing for generating species distribution data for administrative use by the regional government of the federal state of Rheinland-Pfalz. However, the project also provides online tools to the public for analyzing ArtenFinder data in many different ways, as well as an API for downloading all public data, thereby confirming the spirit of free use of the data.

Another feature of both OSM and iNaturalist (but not ArtenFinder) is that the data produced are open to be worked on by the community and therefore, in many cases, are actually worked on by the community, not only by a single user. There is a slight difference between OSM and iNaturalist in the actual way this can be done, as, in OSM, users can directly edit all aspects of an object created by another user, whereas in iNaturalist, users can only comment on a contribution, confirm or disagree about a species identification (or identify a species in the first place, if the contributor didn't do so), but cannot edit contributions directly which are not their own. However, the principle and the consequences are the same: most of the data go back to the efforts of not only one but several users, and this constitutes, for both datasets, an important factor for quality assurance, based on the often-cited principle of Linus' law, stating that "given enough eyes, all bugs are shallow" (Goodchild & Li 2012). Also, the possibility for contributors to use the platforms to communicate with each other over a common subject gives projects such as OSM and iNaturalist a social network dimension (Mooney & Corcoran 2012b). This is augmented by an open approach to data contribution (Neis & Zielstra 2014) shared by both projects, requiring from the contributor nothing more than creating a user account, with the immediate possibility to start contributing and becoming a part of the community. While the latter property is also present in ArtenFinder, this project does not provide the possibility for users to interact on concrete observations, or to interact socially on the online platform itself. Contacts and interaction do take place to a limited extent between observers and experts charged with checking observations, mainly in the form of feedback from experts to observers on certain observations.

OSM as well as biodiversity citizen science projects have developed ways to take the social network aspect from the virtual level to the "real world". For OSM, this comes in the form of so-called mapping parties or mapathons, which bring mappers together at a certain location, often to work on a common task, for instance, mapping a certain region or area to increase data availability and fill up gaps in the map. This can include outdoor activity, but, especially when based on mapping from aerial or satellite images, can also be an event held exclusively indoors (then also called "armchair mapping"[29]). A good example for events belonging to the latter category are the crisis mapping events which were held at the Geography Department of Heidelberg University and in many other places around the world after 2013 Typhoon Hayan over the Philippines and the 2015 Nepal earthquake. Organized in collaboration with the Humanitarian OpenStreetMap Team (HOT), both mapping events significantly contributed to augment the content of OSM for regions affected by natural disasters. They supported rescue efforts of relief organizations by providing up-to-date and complete road networks as well as crisis-related information such as damage on buildings or locations of provisionary accommodation such as tents. All of this information was mapped onscreen from up-to-date satellite or aerial imagery. On both occasions, large groups of volunteers met at Heidelberg University's Institute of Geography. The events were organized by volunteers who also gave some instructions and auxiliary information on the tasks to be performed, pointing mapping efforts at tasks and areas most in need. They also provided information on specific cultural, economic and infrastructural characteristics of the affected countries, because most participants didn't have prior knowledge or personal experience concerning these countries. Help was also given concerning the mapping process, including spe-

[28] translated from https://artenfinder.rlp.de, last accessed on 2018-09-20
[29] https://wiki.openstreetmap.org/wiki/Mapathon, last accessed on 2018-11-18

cific information about the technique of mapping from aerial or satellite images, as well as on the use of OSM editor software (on both occasions, a number of participants were new to the OSM project).

In biodiversity citizen science, something similar to OSM mapping events (albeit with a less serious background than the examples given above) exists in the so-called bioblitzes. These are events where a number of volunteers meet to map the fauna and flora in an area. The motivation of such events, similar to OSM mapping events, is mostly to fill gaps in the data of a project, but also educational purposes (e.g., as part of school biology education), or to raise public awareness and engage people (Novacek 2008). Unlike OSM mapping events, bioblitzes are always outdoor events (at least to a large degree), as mapping from aerial or satellite images is not possible in this domain (as will be discussed in more detail below). For instance, iNaturalist staged a bioblitz in downtown San Jose CA during the Citizen Science Association's Citizen Science 2015 conference, resulting in over 800 observations from parks and other areas[30]. In a more virtual form, many biodiversity citizen science projects also steer contributions towards a certain goal (again, mostly for filling spatial or taxonomic gaps in the data) by launching campaigns, specifically calling for contributions from certain regions or of certain species. Campaigns are mostly launched by newsletters to participants and by appropriate calls on a project's web portal. For instance, ArtenFinder regularly launches campaigns of this kind (see also section 2.1.1). All of these approaches can have a considerable influence on the properties of the data, as mapping events as well as campaigns can produce spatial and temporal clusters of data which cannot be explained without taking into account the events which produced them.

Concerning the social network aspect as such and its role in overall project motivation and goals, many citizen science projects even put the social network dimension in first place. iNaturalist is such a case, declaring that "iNaturalist is an online social network of people sharing biodiversity information to help each other learn about nature", and their "primary goal in operating iNaturalist is to connect people to nature"[31]. Generation of data is defined as a secondary goal (same source). In this, we find a characteristic distinguishing iNaturalist from the OSM project, which declares that "OpenStreetMap is a free, editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license"[32], thereby putting the outcome, which is the map and the data it is built from, in first place. ArtenFinder is rather similar to the OSM project in this respect: producing useful observation data is the primary goal. However, a social network component is largely missing in this project.

Thus, after discussing mostly common ground of OSM and casual biodiversity citizen science projects and data, which is found predominantly in the open and unrestricted way in which data collection is organized and in which these data are made available, it is important to look at fundamental differences between them. Most important among these is the difference in nature of the information which is gathered by OSM on the one hand, and by casual citizen science projects in the biodiversity domain on the other hand. This aspect was discussed very briefly in Jacobs & Zipf (2017), where it forms the backdrop of different conceptualizations of completeness in the two data domains. Here, this aspect is shown to be important from the perspective of data quality assessment. OSM deals with "physical features on the ground"[33], which includes topographic features such as rivers or highways, but also less objective features, for example, shops, crafts, or emergency assembly points. Such features are not (or not explicitly) present in classic cartographic products such as topographic maps or road maps. This may be due the fact that they are less permanent in nature and therefore not suitable for maps

[30] https://www.inaturalist.org/projects/citizen-science-association-2015-san-jose-bioblitz, last accessed on 2018-11-18

[31] http://www.inaturalist.org/pages/what+is+it, last accessed on 2018-09-28

[32] https://wiki.openstreetmap.org/wiki/About_OpenStreetMap, last accessed on 2018-11-18

[33] http://wiki.openstreetmap.org/wiki/Map_Features, last accessed on 2018-11-18

which typically have update cycles of several years, as opposed to OSM, which could, theoretically, be almost up-to-date at all times, given a sufficiently active contributor community. Still, all of these features are physical features, representing real-world objects (buildings, signs, etc.). This means that another volunteer can observe and report these features in exactly the same way, at least within a certain period of time. This is the fact which allows the "crowdsourcing approach" (Goodchild & Li 2012) quality assurance regime to work in OSM. Another volunteer can find (and thereby verify the correctness of) a mapped object, or may prove a feature in the map to be erroneous or obsolete by not finding the object in question at the place indicated.

An observation of an organism, especially if it is of a mobile animal species, is of very different nature. In most cases, there is no physical, spatially or temporally permanent object in the field which can be observed in exactly the same way by another volunteer. The report of an observation usually has rather the character of a witness report (Munzinger et al. 2017). It is the report of an event, a volunteer accidentally observing the presence of an organism, which cannot be repeated in exactly the same way by another volunteer. There are, of course, exceptions, such as sessile and perennial organisms (Jacobs & Zipf 2017). Also many mobile species are bound to certain habitats or territories which they do not leave (at least temporarily), so that several volunteers may observe even the same individual at different points in time at approximately the same location. Going back to OSM, the same is possible for typical OSM objects: e.g., a highway may be repeatedly observed by many OSM volunteers. However, while this repeated observation usually does not lead to repeated reports (in this case, repeated mapping) of the object, which would be considered an error, casual citizen science projects do not refuse repeated reports of the same species from the same place at different points in time or from different observers. Each individual observation is considered an original piece of information. This underlines the event character of a casual citizen science observation. One of the most important consequences for this work is that the truth of a casual citizen science observation of an organism cannot, in most cases, be proved or disproved (Munzinger et al. 2017). Implications of this fact for this work will be discussed in section 2.3.

Another difference between casual citizen science in the biodiversity domain and OSM lies in the fact that most features represented in OSM do not require special knowledge or expertise to be correctly classified by the contributors, as they belong to their every-day environments and are part of the set of experiences most people share (at least, within the same cultural context). Not so most objects in biodiversity citizen science: most ordinary people do not habitually observe organisms and identify species. This requires a special interest and a certain amount of effort beyond every-day activities and experience. As a consequence, the threshold to be overcome for taking part in a biodiversity citizen science project may be higher than that for taking part in the OSM project. An ordinary person might have more confidence in his or her ability to map a highway, than in getting a bird species right. However, where the OSM project may be less "discouraging" concerning the objects that it deals with, it may be more so when it comes to the technical tools that a contributor has to deal with. Contributing to OSM requires the use of special editor software, which, although designed for use by non-experts in GIS, is still somewhat complex to various degrees, depending on the editor software. This is also due to the complex nature of the process of creating digital geometric representations of geographical objects in the form of lines, polygons, and points. In contrast to this, interfaces of casual biodiversity citizen science projects, which come mostly as smart phone apps or internet portals used with ordinary web browsers, are generally very easy to use and functionally much closer to other web applications which the ordinary internet user knows from every-day experience. But then, the task to be performed is not complex (technically), consisting basically of entering text into form fields, selecting items from drop down lists, uploading an image file, and finding and placing a point location on a map similar to well-known web maps such as google map. Both projects whose data were used in this work fall into this category of easy-to-use applications. However, the mapping element has been shown to be still somewhat outside the usual experience of most potential users, and to pose particular difficulties (Newman et al. 2010).

The OSM project practices a way of producing contributions that is totally alien to citizen science projects in the biodiversity domain: contributors to OSM can use satellite and aerial imagery to extract features by tracing them onscreen with drawing tools provided by the editing software already mentioned. In this way, people can contribute real-world features without actually having been on-site. This practice is called remote mapping and has become a very important source of contribution ever since the first sets of suitable imagery were made available. Yahoo! imagery was available between 2007 and 2011, and Bing Aerial images can be used to extract geographic features for OSM since 2010 (Neis & Zielstra 2014) for that very purpose. It is out of scope of this work to discuss the many consequences for data properties and quality that this practice brings along (see, for instance, Eckle & Porto de Albuquerque 2015). However, the reasons why it is possible for OSM reside, again, in the nature of the features which are the object of the OSM project. The very same reasons prevent this practice from being applicable in biodiversity citizen science projects. While certain features, such as buildings, roads, units of land use, and so forth, can be extracted from satellite or aerial imagery (given a sufficient quality of the images), this is impossible for almost all occurrences of organisms, for the good reason that the aforementioned features can in many cases be identified from satellite or aerial imagery, while the latter are mostly not even visible or discernable. In biodiversity citizen science projects, satellite or aerial imagery is often used for orientation purposes in map viewers imbedded in reporting tools (such as apps or web portals), but serves here only as information which supports the localization of an observation.

Another source of data in OSM, much debated in the community, are bulk imports of data from external datasets, often data previously published as open data. Such data often provide a basic stock of features (e.g., a baseline road network), which the contributors can build on by adding features conventionally or by remote mapping. Again, it is out of the scope of this work to discuss the many aspects and effects this has on the project itself, on the contributing process and on its data, or to elaborate on the debate which it sparked within the OSM community (Zielstra et al. 2013). The reason why it is mentioned here is that such practices are rather uncommon in citizen science projects of the biodiversity domain, constituting another notable difference between the OSM project and most citizen science initiatives in the biodiversity domain. Such projects usually do not import observation data from other sources, be it data produced by professional scientists, or data from other citizen science initiatives. One of the exceptions doing so is ArtenFinder Rheinland-Pfalz, where observations from naturgucker.de, another German casual biodiversity citizen science project (see also above), are regularly imported to complement ArtenFinder's dataset. There are also efforts to pool species occurrence data from diverse sources. One of the most prominent of these is the Global Biodiversity Information Facility (GBIF). This may also serve as a good example to illustrate difficulties which are connected to this endeavor. The most important for this work is the high variability of data quality and properties (especially, spatial accuracy and precision of observation coordinates) in data from different providers (Samy et al. 2013), making it difficult or impossible to use data which were aggregated from several sources. In the case of naturgucker imports to ArtenFinder, a practical difficulty is that many naturgucker observations' positions refer to center points of observation areas arbitrarily defined by users, or to map quadrat center points. Thus, they often do not represent actual observation locations. They are therefore different in this respect from original ArtenFinder data, and were not used in this work (see also section 3.3.3). Problems of this kind are also an important property of data pools such as GBIF's. To be very clear, GBIF's goal is not to transfer data from one project's dataset to another one (e.g., from a digitized museum collection to a citizen science project collecting observations of plants and animals), but to make data from very different sources available (downloadable) in one single format. For instance, a user can download a single list of observations from a certain region which consists of observation data from sources as different as a digitized collection of specimens from a natural history museum or a bioblitz which was staged to give a group of high school students a better understanding of ecology. GBIF leaves the responsibility for data quality issues concerning a certain dataset firmly with the data providers, making it a responsibility of the data user to take the highly heterogeneous nature of the data into account, and enables the user to do so by keeping data provenance transparent at all times.

A distinction between OSM and casual biodiversity citizen science data can also be found in the way the history of the information contained in the dataset is managed. The OSM project strives to keep its product, the map, up-to-date. If information becomes completely or partly obsolete because its real-world state changed (e.g., a building is torn down, or a land use changed), this change will (ideally) be implemented also in the OSM data representing the object in question. OSM preserves the full history of every object it contains, so that all changes can be retraced. However, the map itself will always represent as closely as possible the current state of reality. Again, the different nature of casual citizen science observations of organisms, representing events instead of permanent objects, leads to a difference in the resulting data in this respect. All observations remain part of the dataset at all times, no matter how far in the past the underlying observation event took place. In a quality assurance regime such as ArtenFinder's, where experts check observations, rejected observations are often removed from the dataset and are completely lost. In iNaturalist, where other volunteers can comment on observations, add species identification etc., an observation can acquire a history somewhat similar to an OSM feature, but still cannot become obsolete in the same sense as an OSM feature, because, as a report of an event, it cannot be disproved (see above). It can, however, be deleted from the dataset by the observer and then disappears without leaving a trace.

Finally, there are also notable differences between projects in casual biodiversity citizen science and OSM, in areas related to the way information is organized. One of the most prominent characteristics of the OSM project is the possibility, for every contributor, to add new tags to the project, thereby expanding its scope concerning the types of objects which are represented in OSM, as well as the properties which can be assigned to these objects. Consequently, the list of object types that the project deals with and also the list of properties which the representations of these objects can have in OSM are not fixed. To prevent the project from developing an uncontrolled growth of the number of tags, "The community agrees on certain key and value combinations for the most commonly used tags, which act as informal standards"[34]. In most citizen science projects in the biodiversity domain, users can generally contribute observations out of a fixed list of species, with various scopes. Some projects restrict reporting to certain species groups (e.g., birds, see the eBird project or ornitho.de), while others strive to provide a species list which is as comprehensive as possible. Also, most projects provide a fixed set of fields to enter certain aspects about an observation, centered on the basic aspects of location, date, species, and observer. There are exceptions: the iNaturalist project allows its users to create their own fields for entering additional information beyond the core set of aspects which make up the report of an observation. However, iNaturalist is also a good example to illustrate the dangers inherent in such functionality: at the time of writing (researched 29.09.2018), there were 30 user-generated fields for information related to "Behavior/behavior" alone, some created to receive very specific behavioral information, e.g., to be used for a single species, but many simply duplicates with similar names and/or similar descriptions concerning their purpose. Although these fields contain potentially valuable additional information for many observations, to actually use them is difficult under these circumstances. As the more or less consistent set of tags in OSM proves, the community consensus

[34] https://wiki.openstreetmap.org/wiki/Map_Features, last accessed on 2018-11-18

approach practiced by the OSM project is a not always perfect, but overall effective way to prevent such aberration. ArtenFinder does not provide the possibility to add user-defined input fields.

In summary, this section used a comparative discussion to show important properties of casual biodiversity citizen science as a special kind of VGI, using the project and data use cases of ArtenFinder Rheinland-Pfalz and iNaturalist. Comparison among one another and to OSM allowed for highlighting important characteristics of these projects and their data. Among the most important for this work are the opportunistic character of the data collection process and its effect on the data produced, in the form of various biases, and the nature of information collected by geographic citizen science projects in the biodiversity domain, with their reports of observation having the character of non-permanent events. Besides these main findings, more common ground and differences came to light which support our understanding of the projects and data on which this work is based. They are conflated in Table 2.2.1.

Table 2.2.1: Summary of comparison of project and data properties.

Aspect	ArtenFinder	iNaturalist	OSM
Character of data collection	casual/opportunistic		
Data use	free and open, no specific intended use		
Social activities	bioblitzes, campaigns		mapathons
Principle nature of information	non-permanent events		permanent objects
Knowledge involved	specific scientific knowledge		common knowledge
Technical tools	online portals or apps with relatively common functionality		specific editor software
Data provenance	contributions of field observations by volunteers		field and remote mapping, bulk data imports
Essential data output	all observations equally important, including past		current map as main product (history available)
Collaboration, social interaction among volunteers	not on the project platform	collaboration and various ways of social interaction on the project platform	
Thematic scope	static	volunteers may add more parameters/attributes	

2.3 Consequences for Quality Assessment of Casual Citizen Science Observation Data

The properties of casual citizen science biodiversity observation data laid out in sections 2.1 and 2.2 imply consequences for, and entail restrictions on, quality assessment of such observations. This is also true for the approaches developed, evaluated and discussed in this work. A fundamental consequence resulting from the event nature of casual observations of organisms is that most of these observations cannot be proven to be correct, or proven to be incorrect. In a detailed discussion of this problem, Munzinger et al. (2017) draw parallels to Popper's theories on the nature of certain types of scientific findings and also to the nature of witness reports. They conclude that "it is only possible to subjectively assess the degree of their credibility" (translated from Munzinger et al. 2017, p. 230). 'Credibility' is a term often used in connection to VGI data quality (e.g., Gouveia et al. 2004, Flanagin & Metzger 2008, Ostermann & Spinsanti 2011, Freitag et al. 2016). Other terms often appearing in this context are 'plausibility' (e.g., Mülligann et al. 2011, Ali et al. 2014) and 'trust' or sometimes also 'trustworthiness' (e.g., Bishr & Kuhn 2007 and 2013, Bishr & Mantelas 2008, Keßler & de Groot 2013, D'Antonio et al. 2014, Vahidi et al. 2018). So far, these terms are not used in a consistent way. The term 'trust' seems to be used mostly in connection to social aspects and to aspects connected to the volunteers and their properties and relations, especially volunteers' specific experience, knowledge and abilities, as well as social interaction between them. E.g., Bishr & Kuhn (2007) found their work on finding quality measures for VGI on the notion of trust between users within social networks. The term 'plausibility' seems to be used predominantly in connection to more technical aspects, e.g., with plausibility tests of data which are based on factual knowledge and data properties as their grounding (Mocnik et al. 2018). For instance, Ali et al. (2014), who employ geometric properties, topological relations and statistical properties of context in an effort to disambiguate OSM objects, use the term 'plausibility'. The work presented in this thesis falls into this category, and that is why the term 'plausibility' is used throughout. However, there are also examples of contradicting use. For instance, Vahidi et al. (2018) use consistency with habitat and consistency with surroundings (along with reputation of contributor) in a fuzzy trust model and call these indicators 'trust indicators', not 'plausibility indicators'. The ArtenFinder project conducts plausibility checks on its observation data and uses the term 'plausibility' (German "Plausibilität"[35]) although the observer's experience and reputation is often an important factor in validation decisions (see section 2.1.1). There is certainly a need for more work on the clarification of relations between terms such as 'credibility', 'trust', and 'plausibility' (and maybe more terms, such as 'reputation' and 'reliability' also used in literature, e.g., Bishr & Mantelas 2008, Rossiter et al. 2015), which would allow for a more confident use. All of these concepts have been shown to be useful for assessing VGI data quality, but they necessarily remain indirect, so-called proxy approaches (Bishr & Kuhn 2007, Keßler & de Groot 2013, Vahidi et al. 2018), which is also the case for the approaches developed and explored in this work.

Biases inherent in the observation datasets limit their usefulness as a source of geographic context. Several relevant types of bias where described in section 2.2. Geographic and observation bias, leading to inhomogeneous distribution of context information in space, implies that geographic context will not be adequate for plausibility estimation in all locations. Also, spatial inhomogeneity of geographic context may have an influence on plausibility estimation results in locations where this estimation is basically possible. Reporting bias and detection bias have the effect of completely excluding certain species or even whole species groups from the observation dataset, or at least of reducing observation numbers of certain species or species groups. Casual citizen science observation data of organisms, such as the data used in this work, indicate the presence of a species at a place and time, but

[35] https://artenfinder.rlp.de/node/1, last accessed on 2018-10-03

do not allow for inferring that other species which were not reported were absent. This must necessarily affect the thematic properties of the geographic context used for plausibility estimation, which will be biased towards species and species groups which are more popular, or more easily detected, than others. Also, some species might not have enough observations to provide an adequate information basis for plausibility estimation. All of these issues will be duly considered and their effects discussed in detail in following chapters, which explore plausibility indicators for casual citizen science observations of organisms.

3 Methods

3.1 Intrinsic Approach: Observed Communities

The focus of this work lies on the development and evaluation of plausibility indicators for new observations which are added to the data in a biodiversity citizen science portal collecting casual observations. The methodology described in this section is aimed at estimating the plausibility of a new observation in light of existing approved observations and is therefore an intrinsic approach. To achieve this, the set of species observed around the candidate observation, that is, a new observation which is to be tested for plausibility, is compared to the set of species *usually* observed close to observations of the same species. It is well known that species form communities (biocenoses) whose composition is determined by species distribution, which is in turn governed by environmental conditions (e.g., climate, elevation, soil properties, habitat structures of all kinds, and so on) or by the relationships between species (Wittig & Niekisch 2014). Section 2.2 explained that the spatial distribution of casual citizen science observations of species deviates from the natural distribution in many ways due to a number of biases introduced by the VGI data acquisition process. The set of species that are usually observed close to a certain target species is therefore determined by the biocenosis (or several biocenoses) the species belongs to, as well as by factors arising from the observation process.

I call the set of species frequently observed in proximity to observations of a certain target species the *observed community* of that target species. Observed communities do not represent the target species' biocenoses, but rather the target species' typical context of observation within the observation dataset. In my approach to plausibility estimation, this is used as a basis for comparison with the set of species observed in proximity to a candidate observation of the same species. This comparison appraises how well the context of the candidate observation fits the usual context of the target species within the same dataset of observations. I use similarity measures, which were developed for the comparison of biocenoses, to compare the candidate observation's context with the target species' observed community. I argue that a high level of similarity between the proper observed community and a candidate's observed context indicates a high plausibility of the candidate observation. A low level of similarity between candidate context and observed community, however, makes either the observation's location or the species identification given by the observer appear implausible.

The principle elements of this method are the following (see Figure 3.1.1):

- Extraction of species-specific lists of other species frequently observed around them from existing observation data (observed communities),
- extraction of the species observed around the candidate observation which is tested for plausibility (candidate context),
- comparison of the candidate context to the proper observed community by means of a similarity measure, and
- interpretation of the resulting similarity value as an indicator of plausibility of the candidate observation.

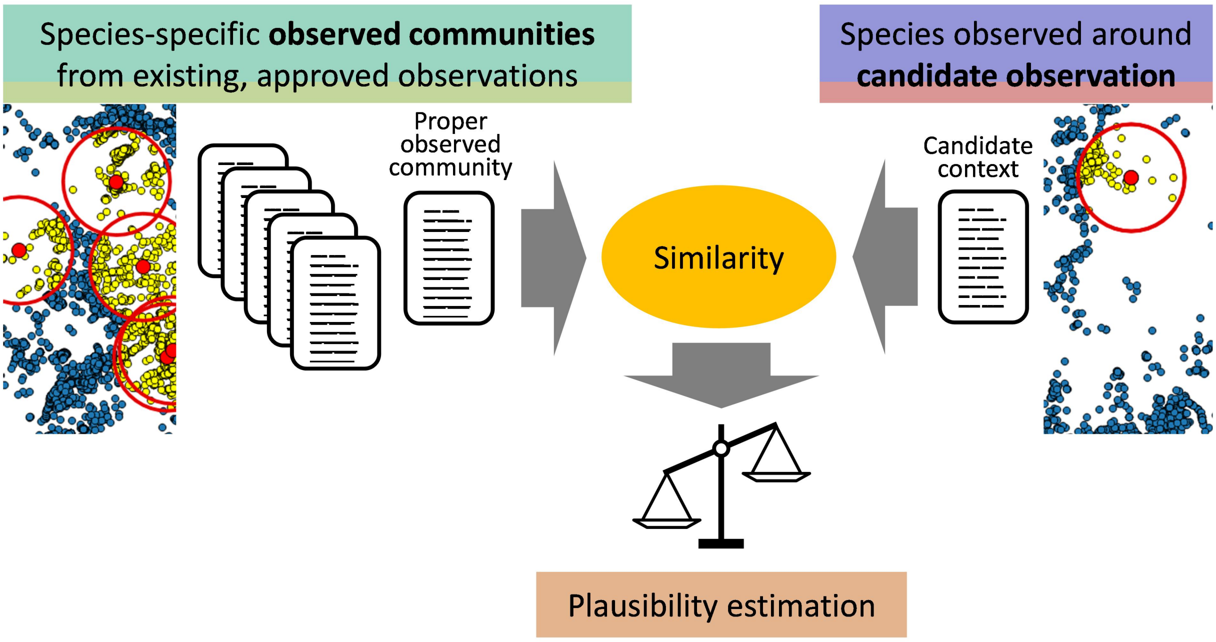

Species-specific **observed communities** from existing, approved observations
Proper observed community
Species observed around **candidate observation**
Candidate context
Similarity
Plausibility estimation

Figure 3.1.1: Observed communities approach, schematic overview of principle elements. A new candidate observation's context of species observed around it (the candidate context) is compared to the typical context of the observed species (the observed community).

In the following, this section describes the basic principles and steps of analysis employed in the observed communities approach to plausibility estimation of casual citizen science observations of organisms. In its basic form, the approach is entirely intrinsic, because candidate observations are tested for plausibility based on earlier, approved observations from the same dataset, and no external data whatsoever are used.

Step 1: Extraction and preprocessing of observed communities

Extraction of the observed community for a certain target species is based on existing, approved observations of this target species and on approved observations of all other species in the same observation dataset. In an iterative process, species observed in a defined neighborhood around each target species' observation are extracted (see Figure 3.1.2). The relevant neighborhood is defined by a search radius and is therefore a circular area around the target species observation. For the purposes of evaluation, the size of this radius was set to 1,000 m as a compromise value between smaller radii which render a more precise view on the geographic context, but which are strongly limiting of context observations numbers, and larger radii which may lead to unspecific contexts. In sensitivity analysis, experiments were also conducted which use land cover and other information to focus neighborhoods in a meaningful way (see section 3.4.4).

Neighborhoods of target species observation cases frequently overlap, especially in regions with high observation density of the target species. A single observation of a context species may therefore produce co-observation cases with several observations of the target species. These cases are considered valid co-observation cases and are counted accordingly. Each coincidence of the target species and a context species within the search radius around a target species observations results in a co-observation of the target species with the context species. A context species becomes part of the list of species associated with the target species if just one such co-observation occurs. The frequency of co-observation is also recorded for each context species, that is, the portion of target species observations having at least one observation of the context species in its neighborhood. This allows for finding context species which are frequently found close to a target species (see below). The example target species in Figure 3.1.2, with six target species observations, has three co-observations with context species A, and six co-observations with context species B, resulting in a co-observation frequency of 0.5 for species A, and of 1.0 for species B.

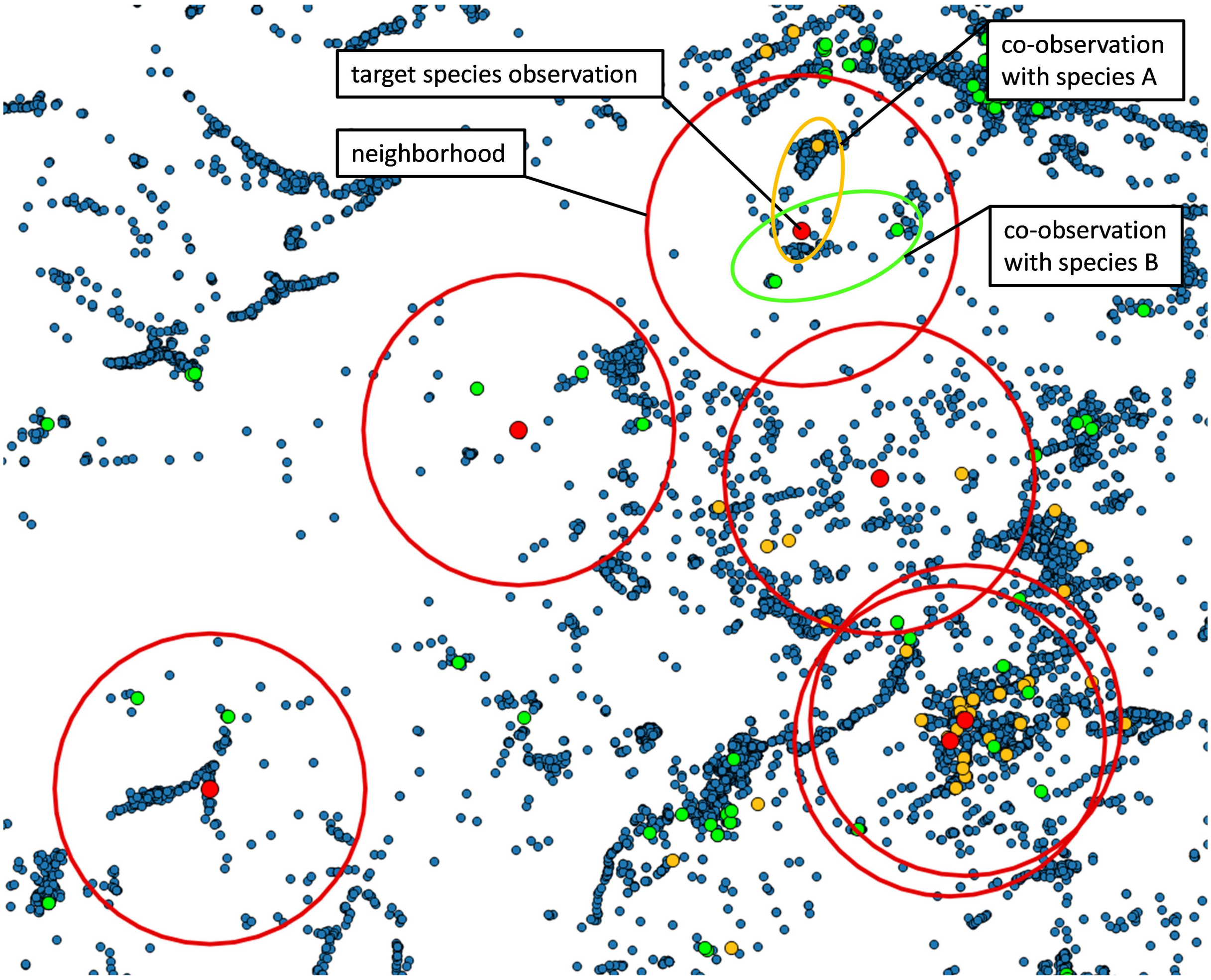

target species observation
neighborhood
co-observation with species A
co-observation with species B

Figure 3.1.2: Schematic view of the basic principles of extraction of observed communities. (Source of observations: ArtenFinder Rheinland-Pfalz. Red points: target species observations; red circles: relevant neighborhood of target species observations; orange points: observations of context species A; green points: observations context species B; blue points: observations of more potential context species not highlighted in this example.)

For each target species, this procedure results in a list of context species which were observed within a certain spatial proximity of the observations of the target species. Principally, all species in a dataset can be used as target species one after the other, while all other species serve as their context species, and roles are reversed when moving to the next target species. However, for evaluation only species with at least 100 approved observations were used as target species for observed communities extraction. The reason for this is that, in cases with very few available approved observations of a target species, these represent only a few situations of that target species in which it was actually observed, which may lead to an observed community which is biased towards these few observation situations.

The resulting list of species co-observed with the target species is further processed in two ways (see Figure 3.1.3). First, the lists are restricted to context species with a frequency value of co-observation above a certain threshold value. In this way the lists are reduced to context species which are frequently observed in association with the target species, while disregarding species which are only occasionally observed in proximity to a target species observation. This step is based on the assumption that frequent co-observations are more meaningful in describing a species' observation environment than infrequent ones. It is important to stress that this procedure will retain only associations which are frequently observed, and not necessarily associations which frequently occur in the natural environment. Ecologically important associations, such as a butterfly species and its food plant, or a predator and its preferred prey, may be lost if one of the species involved is not frequently observed in conjunction with the other. This is one of the important factors which make observed communities distinct from true natural species communities. This step also serves to reduce chance co-observations of species which are probably not important for describing a target species' observed environment, e.g. chance flyovers of birds. There is no obvious way to determine the frequency threshold for this step. It was set to a medium value of 0.5 for evaluation, but effects of higher or lower values were also tested (see section 3.4.3).

Up to this point, this procedure is in fact a spatial association analysis with spatial transactions defined by reference features as proposed by Koperski & Han (1995). This results in lists of spatial association rules of the form "target species x → context species y". Selecting only target species with a minimum number of approved observations means that only species with a minimum support are used. The rate of co-observation of a species with the target species in all considered observations of the target species is equal in these terms to the confidence of the association rule "target species x → species y". Filtering for species with a high rate of co-observation reduces the rules to those with a high confidence. In this way, strong association rules are identified (Koperski & Han 1995).

The resulting species lists are further processed by eliminating context species which occur in many lists at the same time. To this end, each context species is checked for the portion of lists it is part of, and context species with a rate above a certain threshold value are eliminated from the lists. This step is based on the concept of so-called companion species, which is used in plant sociology to describe species that are part of many different species communities and that therefore do not add to the dissimilarity of these communities (Wilmanns 1998). The concept is applied to observed communities for the same reason: to eliminate species which do not add to the dissimilarity of observed communities. Such species are called nonspecific species in this work, because the use of the concept does not correspond to the proper meaning of the term 'companion species', which is only used in plant sociology. A frequency threshold value of 0.5 was employed in evaluation, but higher or lower values were also tested for effects on results (see section 3.4.3).

The product of this procedure is what I call the observed community of a target species. It lists all context species frequently observed in spatial proximity to the target species, excluding context species which are frequently associated with many target species at the same time. The observed community describes the geographic context of a species within the observation dataset. For further analysis, only observed communities listing 10 or more context species were used in order to avoid erratic results in similarity calculation. Observed communities meeting this requirement are subsequently called valid observed communities. It is important to note that this condition, as well as the other parameters and thresholds employed in creating observed communities (see above), restrict the number of species for which a valid observed community can be extracted. For example, a target species with 120 approved observations may produce an initial list of 20 associated species. Let us say that only 11 of these are associated with 50% or more of the target species observations, and identification and removal of nonspecific species removes two more species from the list of associated species. The resulting observed community of this example target species has now nine species left and is therefore not used in subsequent analysis.

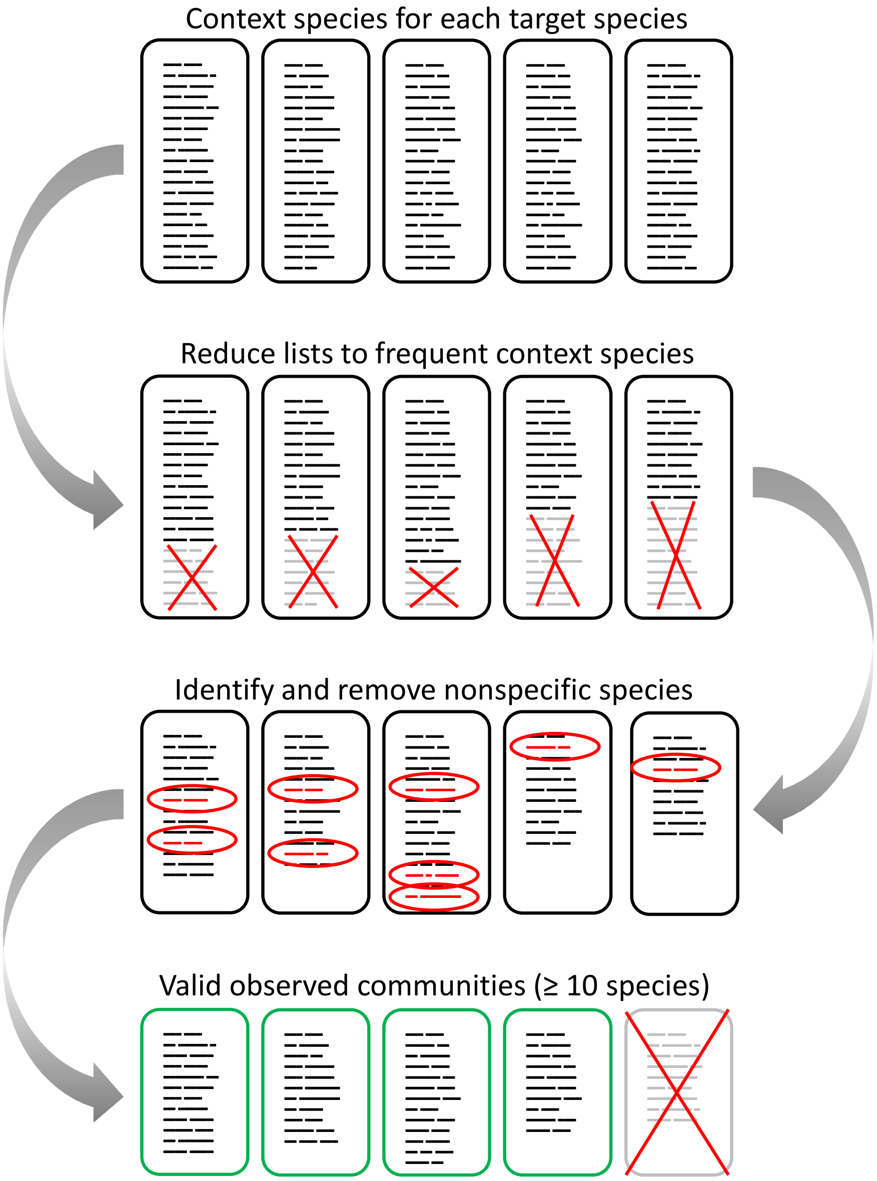

Context species for each target species
Reduce lists to frequent context species
Identify and remove nonspecific species
Valid observed communities (≥ 10 species)

Figure 3.1.3: Processing of observed communities.

Step 2: Extraction of species observed close to the candidate observation

The procedure for extracting the context species observed around the candidate observation is necessarily very similar to the extraction process of observed communities. The same radius is used to define the neighborhood, and the context species are extracted from the approved observations found there. As we deal with just one observation in each candidate case, a rate of co-occurrence of context species can of course not be calculated in the same way, and all context species with at least one observation in the candidate's context become part of the list of context species. Nonspecific species known from the observed community extraction step (see above) are removed from the resulting species list. Similar to observed communities, context species lists with less than 10 species are discarded, so that the subsequent comparison by means of similarity coefficients (see below) is only conducted on two species lists with at least 10 species each. Candidate observations meeting this requirement are subsequently called valid candidate observations.

Step 3: Comparison by means of similarity indices

Similarity indices or coefficients present a widespread method used to measure the similarity between species assemblages in different sites, at different scales, or between different points in time (so-called beta diversity, Wittig & Niekisch 2014, Whittaker 1960), but also for measuring similarity between objects in other domains (Zuur et al. 2007, Legendre & Legendre 1998, Lennon et al. 2001, Koleff et al. 2003). Here, similarity coefficients are used to compare the observed community of the target species to a candidate observation's observed species context. In other words, the comparison is drawn between a synthetic typical observation site of the target species and a single real observation site that was not part of the creation process of the synthetic site. Good overviews and useful discussions of similarity coefficients and their properties can be found, for instance, in Legendre & Legendre 1998, Koleff et al. 2003, Magurran 2004, and Zuur et al. 2007.

When selecting a suitable similarity coefficient, a first and very important decision has to be made concerning the treatment of so-called double zeros, that is, species missing at both sites (in both lists) (Zuur et al. 2007). Some similarity coefficients, called symmetric, include double zeros as a factor which raises similarity, while other coefficients, called asymmetric disregard double zeros. The important question here is to ask which approach makes sense in the actual case. In the present case, including double zeros would mean counting the absence of all species which were observed in the area of interest of the data use case, but which are absent from the observed community as well as from the candidate context, as double zeros contributing to the similarity of the two. This obviously does not make sense. It would lead to very high similarity values, because the observed community and the candidate's species context are both much smaller than the entirety of species observed in the area of interest. In a VGI environment of casual observations, the absence of a species at a site may be due to the species not occurring at the site, or due to not having been detected or reported by the observers, representing detection or reporting bias in the data (see section 2.2). Species missing both from an observed community and from the candidate context compared to it are therefore irrelevant for that comparison and should not contribute to similarity. Legendre & Legendre (1998, p. 253) also state that "it is thus preferable to abstain from drawing any ecological conclusion from the absence of a species at two sites". To this, I would add that this is so except in cases where this explicitly makes sense, which is not the case here. Therefore, this work uses asymmetrical similarity coefficients.

Another important distinction is between similarity coefficients which use only the presence-absence information at two sites, which are called binary coefficients, and coefficients which use quantitative aspects of the data, and which are called quantitative coefficients. For quantitative coefficients, abundance information is usually used as the quantitative aspect, that is, the number of individuals of a species at a site. A common problem with opportunistic citizen science observations of species is that abundance information is often either missing or else quite unreliable. Mostly, this is because abundance information is often not mandatory information explicitly demanded from the observer. Also, giving accurate numbers of individuals can be challenging in some cases, while in other cases it is less difficult (Kosmala et al. 2016). Observations therefore often do not contain any information about the numbers of individuals observed, or these numbers, if given, are a rough estimate. Both data characteristics occur in the data use cases used here, making them typical for opportunistic citizen science data cases in this respect. For this reason, this information was not used for the calculation of quantitative coefficients, and the method was rather evaluated with binary similarity coefficients. It has to be noted, as Koleff et al. (2003, p. 368) also state, that "the vast majority of explicit studies of beta diversity have focused on presence/absence data". For data with missing or uncertain abundance information, this is a natural choice. Considering the VGI nature of the data, however, a quantitative similarity index was tested in sensitivity analysis, substituting the abundance information with observation frequency or a geographic distance criterion (see section 3.4.5). The goal here was not to miss advantages for plausibility estimation which might be obtained by choosing a quantified approach to similarity calculation.

In order to demonstrate the concept of the observed communities plausibility indicator method presented here, a well-established asymmetrical, binary similarity coefficient was used, namely the Jaccard index (Zuur et al. 2007). Legendre and Legendre (1998) present more similarity coefficients which are mostly variants of the Jaccard index, and which do not introduce significantly different approaches to similarity calculation. They were therefore not evaluated. The Jaccard index is represented in the following formula, where *a* is the number of species which both sites have in common, *b* the number of species unique to site one and *c* the number of species unique to site two.

F1: Jaccard index $$\mathrm{J} = \frac{a}{a+b+c}$$

The coefficient renders values between zero (no species identical at both sites) and one (perfect similarity with identical species at both sites).

A further consideration, especially for binary coefficients, is their sensitivity to large species richness differences between the two sites compared, that is, large differences in length between compared species lists. Most standard binary similarity coefficients are quite sensitive to such richness differences: if one of the lists is much longer than the other, there will always be a great number of species not present in one of the lists, and resulting similarity values will be generally low. This is also the case for the Jaccard index presented above. Lennon et al. (2001) present a similarity coefficient which they derived from a coefficient used by G. G. Simpson (1943), and which decreases the influence of richness differences between the two sites compared, because it compares the number of species which both sites have in common, with the smaller of the two sites. Baselga et al. (2007) call this coefficient the Simpson index[36]. In the approach presented here, there are usually marked differences in numbers of species between observed communities and candidate contexts. In most cases, observed communities, which are reduced to frequent associations, are smaller than candidate contexts, except in cases where a candidate observation is situated in a region with low observation density. A coefficient which is able to reduce the influence of species richness differences is therefore potentially useful here.

Using the same notation as for Jaccard index (see F1), the Simpson index is represented below, closely following a notation by Lennon et al. 2001 and Koleff et al. 2003, where, as above, *a* is the number of species which both sites have in common, *b* is the number of species unique to site one, and *c* is the number of species unique to site two:

F2: Simpson index $$\mathrm{S} = \frac{a}{a+\min(b,c)}$$

Comparisons between an observed community of a target species and the species occurring around a candidate observation of that species are conducted only with observed communities and candidate contexts with 10 or more species each. This was done to avoid insufficient data basis for similarity calculation, which might lead to erratic results.

[36] This index not to be confused with a measure of species diversity also called the Simpson index, and which was introduced by E. H. Simpson in 1949 (Simpson 1949, Allaby 2004)

Step 4: Interpretation of the similarity as an indicator of plausibility of the candidate observation

The goal of the approach and method presented here is to use a similarity value as an indicator of the plausibility of a candidate observation in light of existing, approved observation data. The similarity value obtained by comparing a candidate observation's species context to the proper species' observed community is a measure for how well the candidate observation fits the existing, approved observations of the same species. However, for similarity to be able to perform in this way, it is necessary that observations which fit the existing data well have similarity values which are different from similarity values of observations which do not fit well, and therefore allow for distinguishing between these observations. In other words, the approach should be able to identify unusual observations with the help of similarity values. The natural expectation is that similarities for well-fitting observations are higher than for unusual observations.

It is important to stress here that observations identified as unusual and thus implausible are not necessarily erroneous: both species identification and reported location may represent a real observation event at the place specified and of the species given by the observer. There are several possible causes for an observation not fitting the existing approved data well, for instance, an observation coming from a place where (for whatever reason) the species in question was not observed before, or shifts in the range of a species (e.g., induced by climate change), just to name two reasons. The first reason is rooted in the VGI nature of the data collection process, while the second is related to the species' biological and ecological properties. Unusual observations, whether identified by the indicator presented here, or by any other means, should therefore never be subject to automatic removal from a dataset, but rather be marked as unusual, and further scrutinized.

The stock of observation data available for extracting observed communities or candidate contexts usually grows over time, as more observations are reported and validated. Therefore, observed communities and candidate contexts change if extracted at a later point in time. Plausibility estimations for the same candidate observation may therefore change also if repeated, because their data context changed. However, species inventories of areas will reach, at some point, a level at which adding more observations will not add many more species so far unobserved in the area (Colwell et al. 2004, Jacobs & Zipf 2017). At this point, observed communities and candidate contexts will also almost cease to change. Spatial heterogeneity of VGI data implicates that this stage is reached in different areas at different points in time. Another factor for changes in the species composition of contexts is the fact that some species change their range, e.g., due to climate change (Munzinger et al. 2017). An observation of a species which is spreading into new territory will appear unusual at first, considering its context, but may well become plausible at a later point in time, when more observations transform the target species' observed community. It can therefore be argued that such observations should not be removed from the dataset at all (Munzinger et al 2017), but rather tested again later on.

3.2 Extrinsic Approach: OSM Environments

Section 3.1 presents a methodology for estimating the plausibility of casual citizen science observations of organisms which uses existing, approved observations within the same data source as the source of geographic context. Such purely intrinsic approaches to plausibility estimation are useful and have several advantages, but should be complemented with approaches which introduce extrinsic context information into the process of plausibility estimation. While the former rely solely on data from the same dataset as geographic context source and therefore can work only for candidate observations which have a minimum number of context observations, the latter use some kind of external geographic reference data for this purpose, potentially providing geographic context also in regions with insufficient numbers of context observations.

In this work, the suitability of OSM data as a novel source of geographic context is examined. There are many commonalities between OSM data and casual citizen science observations of organisms, because they are both VGI data, but they also show important differences (see chapter 2 for a detailed discussion). For the most part, the methodology laid out in section 3.1 is retained, but adapted where necessary to the use of OSM data as the source of geographic context, instead of existing, approved observations. This method takes a step, therefore, from observed communities to OSM environments. The approach is not intrinsic in the sense discussed in section 1.4, because it uses an extrinsic source of geographic context. It retains a partially intrinsic nature in using previous target species observations to extract a typical environment of a target species in terms of OSM tags frequently mapped in spatial proximity to observations of that target species.

OSM objects and their tags contain a wealth of information about the environment, in the narrower sense, describing elements of the physical environment, both natural and anthropogenic, and in a broader sense, including also information on cultural, social, economic, and other human aspects of the environment. For the goals of this work, OSM data on the physical environment are certainly more important, although careful analysis might reveal interesting connections between elements of the cultural, social and economic environment especially with certain properties of citizen science observations which are determined by their VGI nature. However, this work concentrates on using OSM data on the physical environment as a source of context for describing typical environments of species. To this end, the species found around an observation are replaced with the OSM tags attributing the geospatial objects found around observations. The rationale behind this approach exhibits a certain methodological proximity to niche modelling (e.g., Vahidi et al. 2018), where a number of different parameters characterizing the natural environment of a species and multivariate regression models are used to extrapolate probabilities of occurrence from the locations of existing, authoritative observations to areas where no observation data are available. However, the method presented here constructs the typical environment of a species in a different way, uses different information for describing this environment, and estimates the plausibility of candidate observations (that is, observations actually made), rather than extrapolating to unobserved areas.

Thematic information is attached to OSM objects (also called features) in the form of so-called tags[37], which consist of a key and a value, e.g., “landuse=forest”. Each object can have one or more such tags. The OSM environments approach exploits these tags to describe a target species’ or candidate observation’s environment. OSM tags tell us something about the presence of certain basic habitat elements, natural or man-made, such as a waterbody, buildings, or forests. Most of these elements can have a multitude of different shapes and are often not even clearly defined. Bennet discusses the latter problem for forests (Bennet 2001). OSM is therefore certainly unsuitable as a source of proper habitat in-

[37] https://wiki.openstreetmap.org/wiki/Tags, last accessed on 2018-11-18

formation, which is also true for land cover data sources such as CORINE Land Cover (used in a methodological modification to the observed communities approach, see below, section 3.4.4). In contrast to these data sources, however, OSM contains information and elements belonging to many different domains, not just land use or land cover. Information in OSM is far more detailed, both thematically and geometrically, than in these land cover datasets. This is also the reason why OSM was used in this work as an alternative source of geographic context in a separate, extrinsic approach to plausibility estimation of observations which exploits OSM's thematic information content. A majority of OSM's features relate to man-made objects, but it also includes data about natural objects as small as a single tree, a spring or a small pond, along with larger elements of land use or land cover, such as a forested area or a lake. It is, however, also very important to remember that OSM data are spatially heterogeneous, a result of their VGI nature. For instance, trees in streets may have been mapped in one part of a city and are therefore available as geographic context information when OSM is used, but may be missing in OSM in another part of the same city, because they were not mapped so far.

While the processing steps already used in the observed communities approach are kept, replacing surrounding observations of species with surrounding OSM objects and their tags requires some modifications within the processing steps. The following description of the methodology therefore follows closely the description in section 3.1, but introduces necessary modifications.

Step 1: Extraction and preprocessing of OSM environments

Extraction of the OSM environment of a target species is, again, based on existing, approved observations of the target species. Geographic context information is provided by OSM. The analysis uses all geometric types of objects: nodes objects as well as ways and relations (forming lines and polygons). To identify tags frequently associated with observations of a target species, the rate of co-observation of a tag with the target species in all considered observations of the target species is calculated. For each target species, this procedure results in a list of tags which were mapped within a certain spatial proximity of observations of the target species. Frequency of association of a tag with the target species is determined by the number of target species observations having that tag within their search area, regardless of how often that tag occurs within the search area. A circular area around the target species observation is used to define the relevant neighborhood. Figure 3.2.1 illustrates the procedure. The description of the properties of OSM data in section 2.1.3 discussed difficulties introduced by geometric element segmentation, leading to an artificial multiplication of tag occurrences, e.g., because a highway may be split into many small segments all having the same tag. This problem does not affect the OSM environments approach because it employs a binary view on the geographic context extracted from OSM: a context will contain a tag regardless of how often it actually occurs nearby. However, a tag will be part of many OSM environments or candidate contexts when it is frequent, or is attached to elements which cover large areas or long distances.

The list of tags is then restricted to those that are frequently associated with the target species, that means, tags which are found in many places where the target species was observed. Next, tags frequently occurring around many different species are also removed. Hereafter, these tags are called *nonspecific tags*. It is obvious that this last step does not lead, in the OSM environments approach, to a list of nonspecific species, that is, species which are frequently associated with many other species. However, based on the assumption that nonspecific species identified in the observed communities approach are widespread and therefore have rather unspecific OSM environments, they were excluded from evaluation of the OSM environments approach. If it makes sense to exclude such species in the observed communities approach (which it does), this is also the case for the OSM environments approach. In evaluation of the OSM environments approach, nonspecific species lists were used which were obtained in observed communities evaluation with similar parameters.

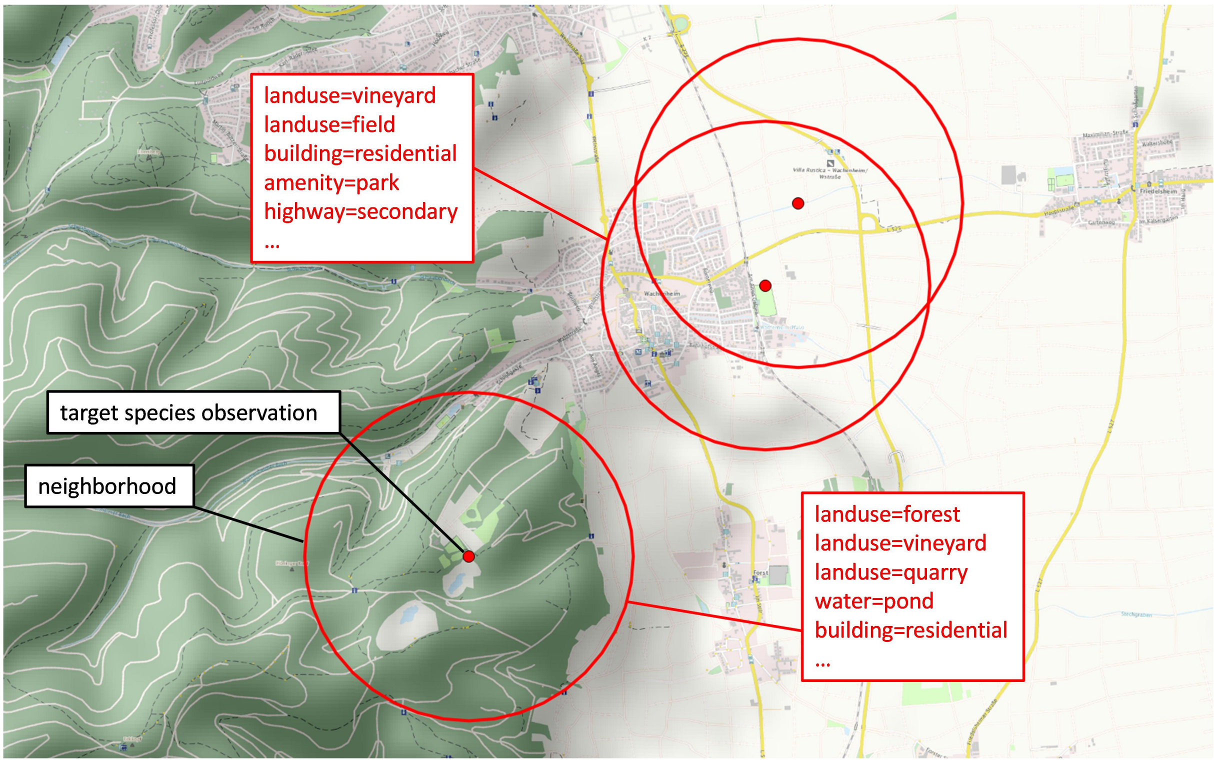


Figure 3.2.1: Schematic view of the basic principles of extraction of OSM environments. (Source of base map: OpenStreetMap[38]; observations: ArtenFinder Rheinland-Pfalz. Original data for hillshade: CIAT-CSI SRTM.)

A list of tags extracted and filtered in this way is what I call the OSM environment of a target species. It lists all tags frequently mapped in close spatial proximity to the target species, excluding tags which are frequently associated with many target species at once. The OSM environment describes the geographic context of a species as this context is mapped in OSM. For further analysis, only OSM environments with 10 or more tags were used, to avoid erratic results in similarity calculation.

Step 2: Extraction of OSM objects and their attributes close to the candidate observation

The procedure for extracting OSM tags mapped around a candidate observation is similar to the process described for extraction of context tags around target observations for OSM environments extraction (see above). A circular area is used to define the neighborhood, and the tags are extracted from the OSM objects found there. Tags which were identified in the preceding step to be associated with many different species (so-called nonspecific tags), are removed from the resulting list. Candidate context tag lists with less than 10 tags are discarded, so that the subsequent comparison by means of similarity coefficients (see below) is only conducted on two lists with at least 10 tags each.

[38] © OpenStreetMap contributors; osm-wms.de

Step 3: Comparison by means of similarity coefficients

Comparison between OSM environments and candidate context tag lists are conducted with the same similarity coefficients used for the observed communities approach, namely the Simpson and the Jaccard index (see section 3.1). Again, comparison is between a synthetic typical observation site of the target species represented by the OSM environment and a real observation site which was not part of the creation process of the synthetic site, namely the candidate context. Although the geographic context retrieved from OSM is of different nature than observed communities of species, careful examination of all considerations for the selection of the similarity indices as laid out in section 3.1 hold also for the OSM environments approach.

Step 4: Interpretation of similarity as an indicator of plausibility of the candidate observation

The overall goal of the approach and method presented here is to use a similarity value as an indicator of the plausibility of a candidate observation in light of the species' OSM environment. The similarity value obtained by comparing the OSM tags mapped around a candidate observation to the proper species' OSM environment is still (as with the observed communities approach) a measure for how well the candidate observation fits the existing, approved observations of the same species. Again, the expectation is that similarities for well-fitting observations are higher than for unusual observations, which is a fundamental prerequisite for the OSM environments approach to work as a plausibility indicator. As before, only approved observations (by the standards of the quality assurance strategies employed in the data use cases) are used in the process of extracting OSM environments, to make sure that they are based, as far as possible, on correct observations.

3.3 Evaluation Methods

3.3.1 Principles and Workflow of Evaluation

The observed communities approach and the OSM environments approach to plausibility estimation for casual citizen science observations of organisms can provide indicators of the plausibility of observations under the following condition: plausible candidate observations must show a higher similarity of their context to the target species' observed community or OSM environment, than do implausible candidate observations. Both approaches were therefore evaluated by comparing similarity values of plausible and implausible observations. For this purpose, suitable sets of candidate observations were selected from real observations. Also, sets of synthetic plausible or implausible candidate observations were created. Generating these candidate sets of candidates was the first step in evaluation, and it is described in detail in section 3.3.2.

Evaluation then proceeded by calculating similarity values for all candidate observations in the sets, using both the observed communities approach and the OSM environments approach. From these similarity values, distributions of values were derived for sets consisting of either plausible or implausible candidate observations. These distributions of similarity values were visualized in the form of boxplots. Additionally, kernel density estimations of the distributions were calculated to obtain smoothed renderings of the probability density of similarity values of the different sets of candidate observations. These two different ways of visualizing the evaluation results were chosen because they are able to convey different and complementary information on the distributions of similarity values. It is important to note here that both kernel density estimations (being normalized to area under the curve = 1) and boxplots obscure differences in sizes between the sets. In results (chapter 4), sizes of sets (numbers of candidate observations) are therefore always given in the captions. In boxplots of results used in the results chapter (chapter 4), boxes represent the interquartile range, dots represent means, horizontal lines in the boxes represent medians, and whiskers extend to max. 1.5 times the range of the box.

Finally, statistical tests were conducted to examine whether differences between the distributions of similarity values of plausible observations and those of implausible observations are significant. Similarity values found in this work are not normally distributed. Therefore, the Mann-Whitney-U-Test (Mann & Whitney 1947) was used to test differences between distributions. Basically, the test allows for examining whether differences between distributions are statistically significant or not. However, strictly correct formulation of the hypotheses of this test depends on whether variances of the distributions involved are homogeneous or not. This was tested with the Fligner-Killeen test (Fligner & Killeen 1976), a test suitable for non-normal distributions. Mostly, variances were found not to be homogeneous. In this case, the null hypothesis of the Mann-Whitney-U-Test is as follows: The probability that a similarity value in the first distribution is greater than a similarity value in the second distribution is not different from the probability that a similarity in the second distribution is greater than a similarity value in the first distribution. The alternative hypothesis is as follows: The probability that a similarity value in the first distribution is greater than a similarity value in the second distribution is different from the probability that a similarity value in the second distribution is greater than a similarity value in the first distribution. See MacFarland & Yates (2016) for details on this test procedure. In rare cases where variances are homogeneous, the test can be interpreted to show the significance of differences between the medians of the distributions involved. Note that p-values are often given as "< $2.2*10^{-16}$", this number representing the minimum value of the R functions used for calculation (R package "stats", functions "wilcox.test" and fligner.test"). Figure 3.3.1 summarizes the principle steps of the evaluation methodology explained above.

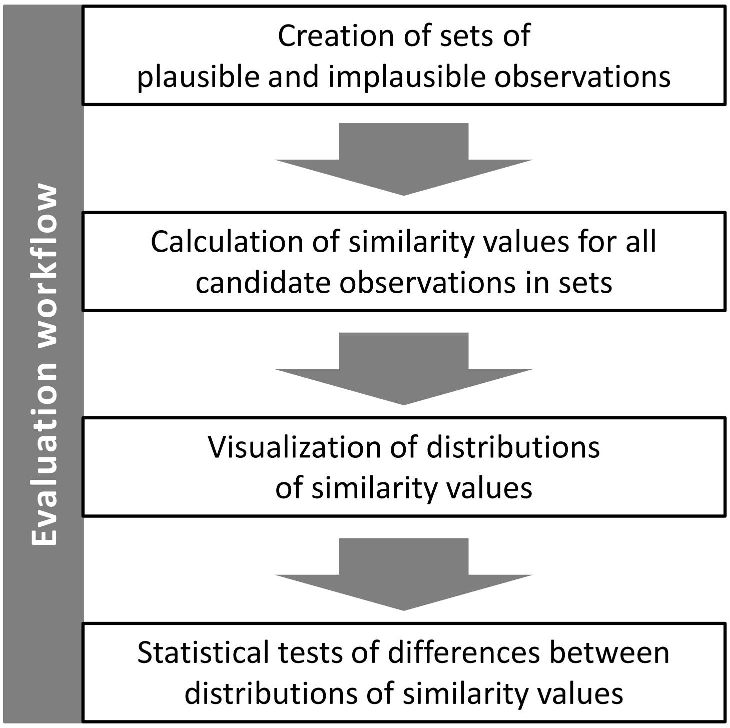

Evaluation workflow
Creation of sets of plausible and implausible observations
Calculation of similarity values for all candidate observations in sets
Visualization of distributions of similarity values
Statistical tests of differences between distributions of similarity values

Figure 3.3.1: Workflow used in evaluation of the observed communities and the OSM environments approach.

3.3.2 Generating Sets of Candidate Observations for Evaluation

A key factor in the evaluation methodology is the generation of suitable sets of candidate observations. These sets were produced in several different ways which are explained in this section. They comprise sets of real observations from the two citizen science projects' data pools used in this work, as well as sets of synthetic observations which were specifically generated for this purpose.

Real observations which were either approved or disapproved in their project's validation process could possibly be used as plausible or implausible candidate observations. Sets of such observations can be considered to contain largely plausible or implausible observations, respectively, because plausibility plays an important role in the validation process. However, there are several difficulties to be considered here. First, while approval or disapproval of observations is largely based on plausibility estimations on the side of the persons making that decision, it is not necessarily the geographic observation context playing a role here, but also other considerations, such as the observer's experience, the observation date, or a photo proof. For the ArtenFinder use case, we have seen that decisions to reject an observation may even be based on reasons other than plausibility, such as a missing photo proof in an observation from an unexperienced observer. Such reasons cause sets of approved or disapproved observations to contain cases which will appear implausible although approved, or plausible although rejected, if only the observations' geographic context is considered. Second, while observations which pass the validation mechanism become part of the dataset of the respective project and are therefore available for use, disapproved observations are in most cases quickly corrected, deleted, recommitted to the observer's private data pool, or in some other way removed from the dataset. They are therefore, in most cases, not available for analysis. Despite these difficulties, sets of real approved and rejected (where available) observations were used in evaluation, to test the approaches' performance on such data. Details on the generation of these sets are explained in section 3.3.3. Evaluation results with these sets tell us something about how far the approaches' plausibility estimations for certain candidate cases are in accordance or discordance with actual validation decisions in these cases.

To overcome the shortcomings of real approved and disapproved observations for the purposes of evaluation, strategies and methods for synthesizing plausible and implausible candidate observations were developed and applied. A first method of synthesizing implausible observations was developed in close cooperation with domain experts from the ArtenFinder project. With real error cases in mind which frequently occur in the observation process, they proposed to use species closely resembling each other physically, but living in different habitats or regions. Such species should have a differing spatial distribution, which can be expected to lead to different observed communities for the two species. Swapping species identification for the correct observations of such species was expected to produce implausible observations which are realistic in that they resemble cases often occurring in the real observation process: lacking expert knowledge, participants often mix up species which closely resemble each other, but which could be distinguished when taking their typical environments into account. For both data use cases, lists of species pairs were developed and potentially implausible observations synthesized by swapping species identifications between the accepted candidate observations of these species pairs. ArtenFinder species for this set of candidates (see appendix, section 7.1.1) were selected in cooperation with experts from the ArtenFinder project. Their positions were extracted from accepted observations of 2016. For the iNaturalist use case, a set of artificial implausible observations was extracted in the same way from research grade iNaturalist observations of 2016. The selection of suitable pairs of species was grounded in two online resources: the Audubon Guide to North American Birds[39] and the Jepson Flora Project[40]. These sources allow for identifying species which are physically similar, but live in different habitats or regions. The resulting set of candidate observations also has a species group composition differing from the candidate set of research grade observations.

This method of synthesizing implausible observations has two major advantages: it is grounded in biological and/or ecological knowledge about the species involved and in knowledge about real errors occurring in the observation process. Its main disadvantage is that it is quite involved and time-consuming. Also, it considers only a small and rather arbitrary (although well-founded) selection of species, so that the thematic properties of this set of observation data are potentially not well comparable to those of other sets of candidate observations. Finally, there is still some probability that observations of two species, whose observation points are exchanged among one another, are observed in spatial proximity to one another, resulting in cases with high similarities.

Therefore more methods to generate sets of candidates for evaluation were employed which do not involve biological or ecological expert knowledge, or knowledge about the observation process, and which are, potentially, better able to produce sets with a high content in plausible or implausible observations. To this end, the basic properties of the two approaches to plausibility estimation were reconsidered. A candidate observation should appear plausible when it is situated in any location where its observation or OSM context is similar to its species' observed community or OSM environment. A candidate observation should appear implausible if it is situated in any location where its observation or OSM context is dissimilar to its species' observed community or OSM environment. These considerations can be used to create synthetic sets of candidate observations whose members should mostly be identified as plausible or implausible, if the approaches are working properly, and are therefore suitable to evaluate the approaches.

Producing a set of synthetic implausible candidate observations for a certain target species is pretty straightforward: it may consist of observation points which are located away from known observations of the same species. This will exclude most locations with a context similar to the target species' observed community or OSM environment. A certain unknown probability remains, though, that such a

[39] https://www.audubon.org/bird-guide
[40] Jepson Flora Project (eds.) 2018. Jepson eFlora, http://ucjeps.berkeley.edu/eflora

point will still be located in a matching context, in which the target species was so far not observed. This technique was used to generate two sets of implausible candidate observations. The first was designed to make sure that the spatial properties of the synthetic implausible observations match the spatial properties of the real observation data. This is relevant because, as we will see, similarity values correlate positively with the observation or tag density around the candidate under certain circumstances. A synthetic set of implausible candidates complying with this condition was created by using valid approved candidate observations and newly created random points. Synthetic implausible candidate observations were created by finding, for each valid, approved candidate observation, a random point situated away from known observations of the same species, but in a location with similar observation or tag density. This point was then assigned the same species. In this way, the spatial properties of the resulting synthetic candidate observations are similar to those of real observation data. Also, the thematic properties of this synthetic set in terms of species group composition and number of observations per species are similar to the set of approved candidate observations. Results with this set of synthetic candidate observations tell us something about the approach's ability to identify implausible observations which were made in areas with observation or tag density similar to neighborhoods of approved cases, but with probably differing observation or OSM context.

Another set of synthetic implausible observations was generated in a similar way, but without any considerations concerning observation or tag density. Existing observations are mostly placed within clusters, and therefore points of this set are often placed in situations where observation density is lower, because random points within clusters are mostly ruled out due to their proximity to existing observations. Results with this set of synthetic candidate observations tell us something about the approach's ability to identify implausible observations also in areas less visited by observes than the usual observation in a certain dataset.

In contrast to the above considerations for synthetic implausible candidate observations, producing a set of synthetic plausible observations requires knowledge about locations where such a synthetic candidate would indeed be evaluated as plausible. It is not enough to place synthetic candidates for a species simply close to known, approved observations of that target species, because approved observations are not necessarily plausible in light of their observation or tag context. A possible approach to find suitable locations is simply to apply the observed communities or OSM environments approach to plausibility estimation to a set of candidate observations (e.g., real, approved observations), then to select candidates which have a high plausibility, and finally placing synthetic candidates close to these plausible candidates. It is quite obvious that, with such an approach, it would be possible to produce a set of synthetic candidates with any degree of plausibility. It is, however, still useful to test the approach's ability to identify plausible candidate observations. To this end, synthetic candidate observations were produced by finding, for each real, approved candidate observation which was evaluated as plausible, a random point close-by and assigning to this random point the same species as the respective plausible approved candidate. Thresholds for similarity values representing approved observations with a high plausibility were chosen so that they represent ca. 20% of valid approved observations in evaluation. Results with this set of synthetic candidate observations tell us something about the approach's ability to identify plausible candidate observations in locations which should have an observation or OSM context reasonably similar to the candidate species' observed community or OSM environment.

Table 3.3.1 gives an overview of sets of plausible and implausible observations for both data use cases, whose origin was explained above. In subsequent text, diagrams, charts etc., the set codes indicated in the table will be used.

Table 3.3.1: Sets of Candidate observation used for evaluation.

		ArtenFinder Rheinland-Pfalz		**iNaturalist (California)**	
	Set of candidates	**Data**	*Set code*	**Data**	*Set code*
Sets of plausible observations	**Approved observations**	Observations accepted by experts	*AF_A*	Research grade observations	*iNat_A*
	Synthetic plausible observations	Random points in the vicinity of plausible accepted observations	*AF_SP*	Random points in the vicinity of plausible research grade observations	*iNat_SP*
Sets of implausible observations	**Observations rejected by experts**	Observations rejected by experts	*AF_R*	Not available	-
	Synthetic implausible observations based on physically similar species	Swapped species identifications between accepted observations of physically similar species living in different habitats	*AF_SI1*	Swapped species identifications between research grade observations of physically similar species living in different habitats	*iNat_SI1*
	Synthetic implausible observations in similar density situations	Random points away from accepted observations of a species, but in locations with comparable observation or tag density	*AF_SI2*	Random points away from research grade observations of a species, but in locations with comparable observation or tag density	*iNat_SI2*
	Synthetic implausible observations in any location	Random points away from accepted observations of a species	*AF_SI3*	Random points away from research grade observations of a species	*iNat_SI3*

3.3.3 Acquisition, Preprocessing, Selection, and Partitioning of Observation Data for Evaluation

ArtenFinder data used for evaluation

ArtenFinder provides several possibilities of retrieving observation data and other information. Users can download their own observations as a csv file. There is also a REST API. Depending on the user's role, it allows for downloading public data of all users, but also for changing or deleting data in the database. The API was used to download all public ArtenFinder observations up to 2016, with their validation status on February 24th, 2017 (latest download of the data). The data were placed in a local spatial database for further processing and analysis. ArtenFinder receives a few observations from regions adjacent to the federal state of Rheinland-Pfalz, which were discarded. In cooperation with naturgucker.de, another German citizen science initiative collecting observations of organisms, ArtenFinder regularly imports naturgucker observations from Rheinland-Pfalz. These were also removed from the dataset used for analysis, because they have, at least in part, different properties, such as rasterized observation locations (coordinates representing map quadrat center points rather than the original observation locations to protect occurrences of sensitive species), or locations referring to the center point of an arbitrary area rather than to an exact observation location.

Accepted observations up to 2015 (216,316 observations) were used to generate observed communities and OSM environments. Accepted observations of the year 2016 (68,646 observations) were used for producing a set of candidate observations expected to contain predominantly plausible observations (set AF_A, see Table 3.3.1). This approach of partitioning the data into older observations for extraction of observed communities or OSM environments, and new observations used as candidates, reflects the fact that the assessment of the plausibility of a recent candidate observation with the observed communities or OSM environments approach is always and necessarily based on older approved observations. This way of partitioning the data therefore is more appropriate here than, for instance, selecting a random sample of accepted observations to be used as candidate observations from the whole time period. Evaluation should show whether the approaches to plausibility estimation are suitable for estimating the plausibility of new observations based on older pre-existing observations, which can be achieved in this way. Moreover, using a whole year of observations as candidates in evaluation minimizes any seasonal biases in evaluation results: if only data of a certain part of a year would be used, this would introduce a bias towards species observed in that season.

The data properties of the two data portions of approved observations (up to 2015, and from 2016) do not exhibit important differences. This shows that the observation process did not change in a critical way between these two time periods. In both sets of observations, birds, butterflies, dragonflies and plants make up most of the observations, with the same ranking of species groups. In 2016, birds make up a somewhat higher part of observations when compared to the data up to 2015, while the rate of butterflies (and some groups with smaller portions in the data) was slightly lower, see Table 3.3.2. More information about the composition of the dataset concerning species groups and development over time can be found earlier in section 2.1.1.

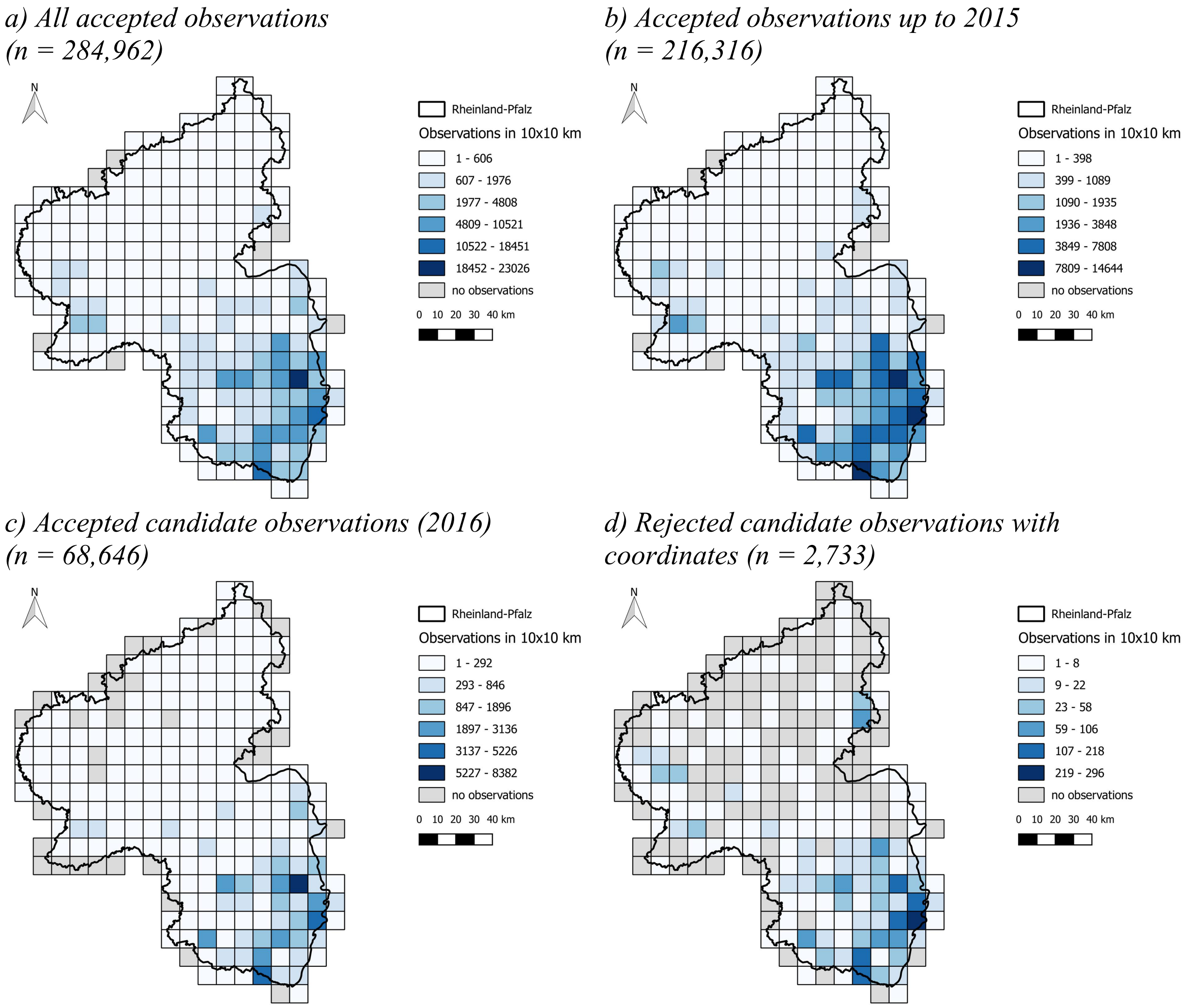
Table 3.3.2: Portions of species groups in sets of accepted ArtenFinder observations. Comparison of values for accepted observations up to 2015 and accepted observations from 2016.

Species group	Portion (%) up to 2015	Portion (%) 2016
plants	6.7	5.5
fungi	1.9	3.4
mammals	2.4	1.9
birds	41.8	48.9
reptiles	1.7	1.1
amphibians	1.6	1.0
modern bony fishes	0.1	0.0
butterflies and moths	28.9	24.2
hymenopterans	0.7	0.8
beetles	0.8	0.6
dragonflies and damselflies	9.2	10.3
mantids	0.1	0.1
locusts	3.4	1.5
mollusks	0.3	0.2
true bugs	0.1	0.1
spiders	0.0	0.1

The ArtenFinder API was also used to retrieve quality assurance protocol data. These data store status changes for all observations processed in the project's quality assurance process, providing the opportunity to find IDs (identification numbers) of rejected observations. This is rarely possible in projects of this kind, because observations which are rejected are usually quickly corrected or deleted, or referred back into the private data spaces of the observers, where they cannot be easily accessed. In the case of ArtenFinder, it was possible to access the IDs of rejected observations, by means of quality assurance protocol data (kindly made available to the author by the project lead). They do not, however, provide the positions of these observations, or their species identification. It was therefore necessary to collect, at regular intervals, observations not yet validated by experts (available via the project's public API), which contain coordinates and species identifications, and later harvest from this list the observations which were eventually rejected using the observation ID numbers from the quality assurance data. It was also possible to filter out observations which were rejected in the first place, but later accepted, e.g., because the observer provided more information. In this way, observations coukld be retrieved which were permanently rejected by the experts in the validation process to form set AF_R. They provide a valuable basis of analysis, especially for the evaluation of the plausibility estimation approaches laid out here, because they allow for comparing similarity values of real observations which were accepted as correct, with real observations which were rejected as incorrect (or for other reasons, see section 2.1.1). These two sets of candidate observations allow for analyzing whether approved or rejected observations differ in their plausibility estimations.

Extraction of rejected observations used all available rejected observations, including 2016 as well as earlier observations. Therefore, there is no clear partition between recent and older observations here. However, this was necessary to arrive at useful numbers of valid observation cases: 6,845 rejected observations were available, and for 2,733 of them coordinates could be retrieved. As rejected observations are not at all used for extracting observed communities, there was no conflict here concerning data partitioning. The composition of species groups in this set is markedly different from, e.g., approved observations: butterflies are leading at 25%, followed by observations of "other species" (24%), dragonflies (18%), birds (6%), plants (6%), mushrooms (5%), and locusts (5%). This may be caused by a higher rate of species which are hard to identify in the insect species groups, causing a higher rate of rejections in observations of these groups when compared, for instance, to birds.

All sets of candidate obsevations presented here show a similar overall spatial distribution of observations, reflecting the same northwest to southeast trend of increasing observation density. This is demonstrated by comparing quadrat count maps for these sets of observations, see Figure 3.3.2. Sectors of maximum concentration of observations are, however, slightly different.


a) All accepted observations (n = 284,962)

b) Accepted observations up to 2015 (n = 216,316)

c) Accepted candidate observations (2016) (n = 68,646)

d) Rejected candidate observations with coordinates (n = 2,733)

Figure 3.3.2: Spatial distribution of observations in sets of ArtenFinder data. (No. of points in 10x10 km raster. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)

iNaturalist data used for Evaluation

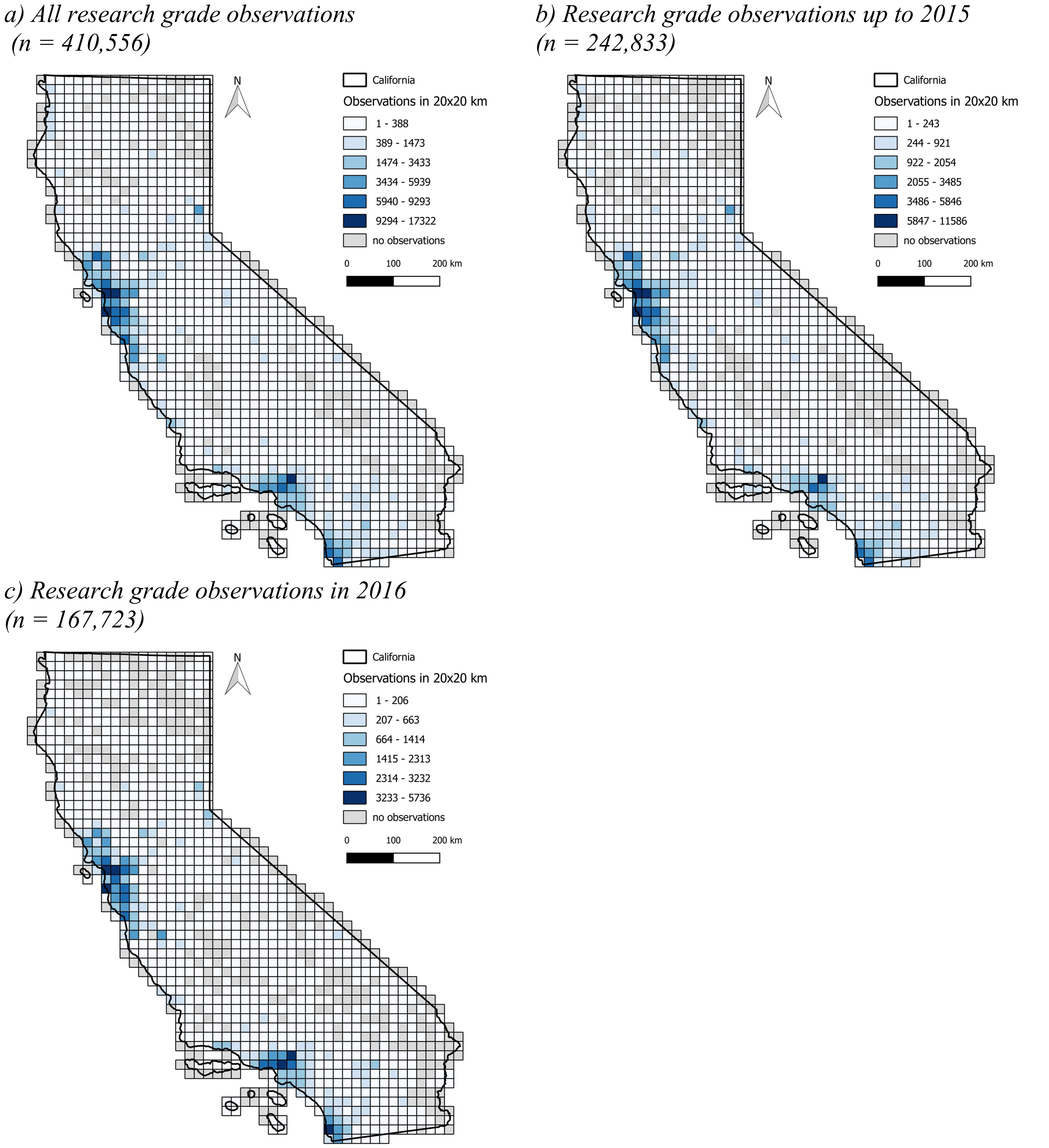
iNaturalist provides the possibility to download their data as a csv file. This was used to retrieve all iNaturalist observations for the state of California on March 3rd 2017. California was chosen as the area of interest of this data use case, because the iNaturalist data record is strongest there (see also section 2.1.2). Observation numbers are comparable to those of ArtenFinder, albeit spread over a much larger area. Again, the data were placed in a local spatial database for further processing and analysis. iNaturalist obscures coordinates of certain observations, mostly to protect rare or sensitive species. Also, observers can choose to obscure coordinates of observations for privacy reasons. Observations with obscured coordinates were removed from the dataset used here, because their coordinates do not represent the true location of observation.

Data partitioning followed the same principles for iNaturalist data, as already used for ArtenFinder data. Research grade observations from California up to 2015 (242,833 observations) were used to generate observed communities or OSM environments. Research grade observations of the year 2016 (167,723 observations) were used as a set of plausible candidate observations based on approved observations. There is no way of identifying observations rejected in the iNaturalist dataset, due to the differing quality assurance strategy employed here.

Table 3.3.3: Portions of species groups in sets of research grade iNaturalist observations. Comparison of values for research grade observations up to 2015 and research grade observations from 2016.

Species group	Portion (%) up to 2015	Portion (%) 2016
plants	33.4	35.1
fungi	2.3	4.7
mammals	4.2	3.9
birds	32.0	26.7
reptiles	5.2	5.1
amphibians	1.6	1.8
modern bony fishes	0.4	0.3
butterflies and moths	6.7	6.1
hymenopterans	1.1	1.5
beetles	1.2	1.8
dragonflies and damselflies	1.7	1.4
earwigs	0.1	0.1
mantids	0.1	0.1
cockroaches	0.0	0.1
locusts	0.3	0.5
crustaceans	0.8	1.3
mollusks	5.1	6.0
other species	2.1	2.3
true bugs	0.6	0.9
flies	0.2	0.3
spiders	0.8	0.0

Although yearly observation numbers are strongly increasing in iNaturalist (see section 2.1.2), data properties of the two sets of research grade observations (data up to 2015, and from 2016) are not critically different, which shows, again, that the observation process did not change over time in a critical way (see Table 3.3.3). In 2016, plants make up a slightly larger part of observations when compared to the data up to 2015, while the rate of birds was somewhat lower. Figure 3.3.3 demonstrates that these two sets of data also show a similar spatial structure, with similar regions of higher and/or lower point density.


a) All research grade observations (n = 410,556)

b) Research grade observations up to 2015 (n = 242,833)

c) Research grade observations in 2016 (n = 167,723)

Figure 3.3.3: Spatial distribution of observations in sets of iNaturalist data. (No. of points in 20x20 km raster. Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016).

3.3.4 Source of OSM Data and Tag Selection

For the purpose of this work, OSM data from the two study areas (Rheinland-Pfalz and California) were used, paralleling the areas of interest from which the data use cases of casual citizen science observations were taken. The general properties of the OSM project and its data are described, discussed, and compared to those of ArtenFinder and iNaturalist in chapter 2. Evaluation of the OSM environments approach was conducted with OSHDB[41], a tool for analyzing OSM full-history data. This tool allows for running queries that produce results grouped by thematic or temporal parameters. The OSHDB software is developed and maintained by HeiGIT (see also section 2.1.3). Regional OSM extracts provided by HeiGIT and used in this work were produced in January 2018 (for California) and

[41] https://github.com/giscience/oshdb

March 2018 (for Rheinland-Pfalz). For evaluation of the OSM environments approach, a consistent state of OSM at 2017-07-01 was used.

From the entirety of tags which are currently supported or used in OSM, such tags were selected which are potentially relevant for describing an observation's environment in the form of OSM features and their attributes. This excludes, for the most part, tags subsumed under “Additional properties”[42], such as names, address information, opening hours etc., which do not characterize the physical environment. The bulk of tags selected for this work come from the “Primary features” group of keys. These describe, for the most part, elements of the physical environment and their properties. For some of these keys, such as “Natural” or “Landuse”, all available values were deemed relevant and thus all tags included in the selection of relevant tags. For other keys, a selection was made from available values, again excluding mostly values which do not characterize the (relevant) physical environment. Examples for such cases are “amenity=parking” (designating a parking lot, which is a relevant physical element of settled areas, and therefore selected) vs. “amenity=atm” (designating an automatic teller machine, and therefore not selected). The “properties” group of keys also has several keys which are obviously relevant, such as “leaf_cycle=*” (evergreen, deciduous, etc.) or “leaf_type=*” (broadleaved, needle leaved, etc.) and some others which are not so obvious but may still be relevant, such as “bridge=*” (with different types) or “cutting=*” (for roads incised in the land surface). Selection resulted in 757 tags with 44 distinct keys, see appendix, section 7.2.

[42] https://wiki.openstreetmap.org/wiki/Map_Features, last accessed on 2018-11-18

3.4 Sensitivity Analysis: Effects of Modifications of Input Parameters and the Methodology

The observed communities approach and the OSM environments approach to plausibility estimation of casual citizen science observations of organisms were evaluated with a basic methodology and a specific set of parameter settings (presented in sections 3.1 and 3.2), to understand their principle functionality and to prove the general concepts of the approaches. The evaluation methods and workflow were explained in detail in section 3.3. Parameter settings are, however, not necessarily fixed to the values used in evaluation. Application of the approaches with other data use cases might require other settings, or domain experts might ask for other parameter settings for reasons arising from certain domain-specific considerations. For instance, a higher threshold for identifying nonspecific species might be required to avoid that certain species are considered as nonspecific, or larger search radii might make sense in certain data contexts with a lower spatial data density, to allow for more candidate observations to be assessed. Also, there are certain methodological modifications which might make sense, such as using quantitative information in similarity calculation, instead of the binary indices used in evaluation in this work. But what effects will these changes and modifications have on the approaches' behavior and consequently on evaluation results? And will the approaches basically continue to work under these conditions?

Effects of changes in input parameters and effects of methodological modifications can be tested with so called sensitivity analysis. In a review of this field with a focus on sensitivity analysis of environmental models, Pianosi et al. (2016) found that there are many possible approaches to this task. Here, sensitivity analysis consisted in conducting evaluation experiments with changes to certain input factors and with some methodological modifications, and in comparing results of these experiments mostly visually to the original evaluation results. Pianosi et al. (2016) call this approach to sensitivity analysis a "One-At-a-Time (OAT) method" and classify the approach chosen here as a "perturbation method", with one-by-one variation (perturbation) of input factors and visual inspection of results (Pianosi et al 2016, p. 219). This approach to sensitivity analysis was chosen because it is a suitable procedure for cases with a relatively low number of input factors or methodological modifications which are to be examined. The main goals were to gain an insight into the mechanics taking effect when parameter settings are changed or the methodology is modified in certain ways, and to see how robust the methods are against these parameter changes and methodological modifications. Some of these may also hold the potential to improve the approaches' performance as plausibility indicators in general or at least for the data use cases at hand.

Sensitivity analysis was mostly conducted on the observed communities approach only (except one specific methodological modification which was tested in the OSM environments approach, see below), because both approaches have analog methodological structures, and at least some of the results can be expected to be analogous for both approaches. However, future work should certainly examine also the sensitivity of the OSM environments approach more closely, to prove or disprove this assumption.

Changes to the following input factors were introduced to the observed communities approach one by one, and effects on results examined:

- minimum number of approved observations of a target species necessary for extracting an observed community,
- size of the search radii applied for defining the relevant context during observed community extraction, as well as for candidate context extraction,

- frequency threshold for identifying frequently associated species in observed communities, and
- frequency threshold for identifying nonspecific species in observed communities.

The following changes to the methodology were tested to assess their effects on the behavior of the respective approach:

- use of auxiliary land cover related information when defining the relevant search area for observed community extraction, as well as for candidate contexts,
- use of quantitative information about the species in observed communities and those found around a candidate observation, and use of a suitable quantitative similarity index,
- introduction of a guard zone for edge effect correction in the observed communities approach,
- use of a modified Simpson index with a more straightforward interpretation of index values in the observed communities approach, and
- using date-specific OSM context in the OSM environments approach.

The following subsections explain the details of parameter changes and methodological modifications whose effects were examined.

3.4.1 Using a Lower Minimum Requirement for Target Species Observations

In evaluation, a conservative minimum number of 100 observations were required from a target species for observed communities extraction, to avoid bias towards few observation situations represented in these observations. While this is a valid concern, it can be argued that, from a VGI perspective, the observed community cannot be biased, because it always represents the observation situation of the species in the dataset. An observed community based on just a few observations is realistic in that sense, if only a few such observations are available. There are also ecological arguments for a not too high threshold here (personal information D. Frank, ArtenFinder project, 07-2017): species with only a few observations may be rare species strongly specialized in a certain habitat, so that only a few observations are available. In such a case a few observations may be enough to describe the species' context adequately also from an ecological perspective, because the species occurs only in the habitat the observations come from. Still, with a smaller number of target observations, an observation with a rather untypical context can exert a stronger influence on the properties of the resulting observed community than in cases with more target observations. Therefore, a minimum threshold of just 10 approved observations for a species to qualify as a target species for observed community extraction was tested, to investigate the effect of such a change to the results achieved with the approach.

3.4.2 Using Variable Search Radii

The observed communities approach was evaluated with a search radius (uniform for all species groups) of 1,000 m, for defining the relevant neighborhood of an observation both when extracting observed communities and the species observed around a candidate observation. However, species groups (and, strictly speaking, even single species within them) have different properties as to the size of habitats, mobility, and behavior, factors which might influence the size of the area in which an individual interacts with its environment. This might in turn have an influence on which species are associated with it, and thus which species are observed in association with it. It might therefore be beneficial to use different search radii for different species groups.

The plausibility tool developed for the quality assurance process of the ArtenFinder project (Jacobs & Schotthöfer 2015) provides the possibility to set different search radii for different species groups, and even for single species, when extracting the list of species surrounding a candidate observation. The rationale behind this feature is that a specific search radius might lead to a more specific list of associated species, which describes the candidate observation's environment more accurately. This reasoning is backed by the professional beliefs of the ArtenFinder experts who are charged with validating the project's observations and who were consulted in the development phase of the plausibility tool. However, the feature has so far not been actually used (no specific search radii set for species groups or species) because the results provided by the tool component in question have proved to be difficult to use due to several reasons (already explained in section 2.1.1). The component, although operational, is therefore not actually used in ArtenFinder quality assurance practice so far, and therefore did not affect ArtenFinder data in any significant way. The validity of the arguments for variable search radii quoted above is therefore untested so far.

To test the effects of assigning group-specific search radii in this work, a larger radius of 3,000 m was chosen for birds and mammals. Insect groups such as butterflies and dragonflies were assigned a search radius of 2,000 m, and the search radius for other species groups, especially plants, was left at 1,000 m. These values reflect results of discussions with experts from the ArtenFinder project conducted during development of the plausibility tool for that project. Larger search radii can be expected to produce larger observed communities and also larger candidate contexts, because more context observations are considered.

3.4.3 Shifting Frequency Thresholds for Frequent Co-Observations and Nonspecific Species

Two important steps in the process of creating observed communities are their restriction to context species which are frequently associated with the target species, and filtering of nonspecific species which are part of many observed communities at the same time. It was already said that there is no obvious way to determine the frequency thresholds used for these two processing steps. In evaluation, an intermediate value of 0.5 was used for both. However, other threshold values might be required for various reasons, when the observed communities approach is applied to different use cases. Basically, a higher frequency threshold to find species frequently associated with a target species will exclude more context species from an observed community, while a lower threshold value will allow for more context species to be included. A more strict frequency criterion for nonspecific species will lead to a smaller number of species being classified as nonspecific species, while a less strict criterion will take effect to the contrary. How does a shift of these thresholds, either upwards or downwards, affect the properties of results? Effects on evaluation results were examined with new threshold values of 0.25 and of 0.75.

3.4.4 Using Auxiliary Land Cover and Ecological Land Unit Information

Besides examining the effects of parameter changes on evaluation results, experiments introducing alterations to the methodology itself were also conducted. While the general methodological framework of the observed communities approach was kept intact, the first of these experiments examined the effects of introducing auxiliary land cover information into the way relevant search areas are defined when looking for context observations around a target observation in observed community extraction, or when extracting the context of a candidate observation.

In evaluation, the method used a circular search area to find relevant context observations both for observed community extraction and for finding a candidate context. This procedure will always include context observations which were made in the same habitat in which the target observation is situated (e.g., a stretch of forest, or a meadow), while some context observations included in the search area may be situated in neighboring habitats, so that they potentially do not belong to the target observation's relevant context. A classic example would be a species which is typically living in open habitats such as fields or meadows and is therefore part of a natural community of species which share this preference. Most search areas will include only context observations which are situated in open habitats as well, but some contexts of target observations situated at the edge of the habitat may include observations in a neighboring forest or settled area. If these are numerous enough, they will introduce associations into the target species' observed community which do not make sense from a biological or ecological point of view.

Such considerations depend, of course, on a number of factors. Target observations situated in small habitats, such as small ponds or streams, will always have many context observations from neighboring habitats (given that neighboring areas do actually have observations) because small habitats themselves cover only a small portion of the search area. As already pointed out, target observations on the edge of a habitat will have context observations in neighboring habitats, while target observations well within (large) habitats do not. Many species use diverse habitats during different times (daytime or season), or for different behaviors (resting, feeding, breeding, etc.) and their relevant context therefore comprises different situations. Also, observations may accidentally be placed by the observer outside of the habitat where the observation was actually made. Such a misplacement may be also be caused by imprecision of the location device (e.g., a GPS device in a smart phone), or by inaccuracies of the base maps which are provided by the application used for reporting (Spyratos et al. 2014). Often, there is also a spatial offset between the observer's location and the location of the observed individual, which may or may not be corrected by a (more or less accurate) estimation of the individual's true location by the observer. With (very) mobile animal species, all of these considerations are overlain by the fact that individuals may be observed in places which do not provide habitat functions to them at all. For instance, a high-flying flock of migratory birds may be observed above almost any kind of habitat. On the other hand, observations of species with low or no mobility, especially sessile animals and plants, suffer only to some degree from some of the factors discussed above. Still, these considerations add up to a quite impressive list of uncertainties affecting locations of casual citizen science observations of organisms relative to habitats, and thus also affecting results of spatial analyses of these data.

The methodological change which is introduced here consists in the additional use of a suitable data source representing habitat-related boundaries in defining search areas around an observation. In this approach, relevant search areas include only that part of the circular search area defined by the search radius which is also part of the habitat-related area in which the target observation is situated. Figure 3.4.1 gives an example: an ArtenFinder observation of the butterfly species Common Blue (*Polyommatus icarus*) is situated in a grassland area (light brown), with neighboring forest (greens), fields (tan) and an industrial area (purple) within 1,000 m search radius. Now, only observations in the grassland area are considered to be associated. In this procedure, the actual land cover type does not play a role. Only the boundary, or, technically speaking, the polygon geometry within which a target observation is situated is used to restrict the relevant search area for context observations. This principle was applied in observed community extraction and in candidate context extraction in the same way.

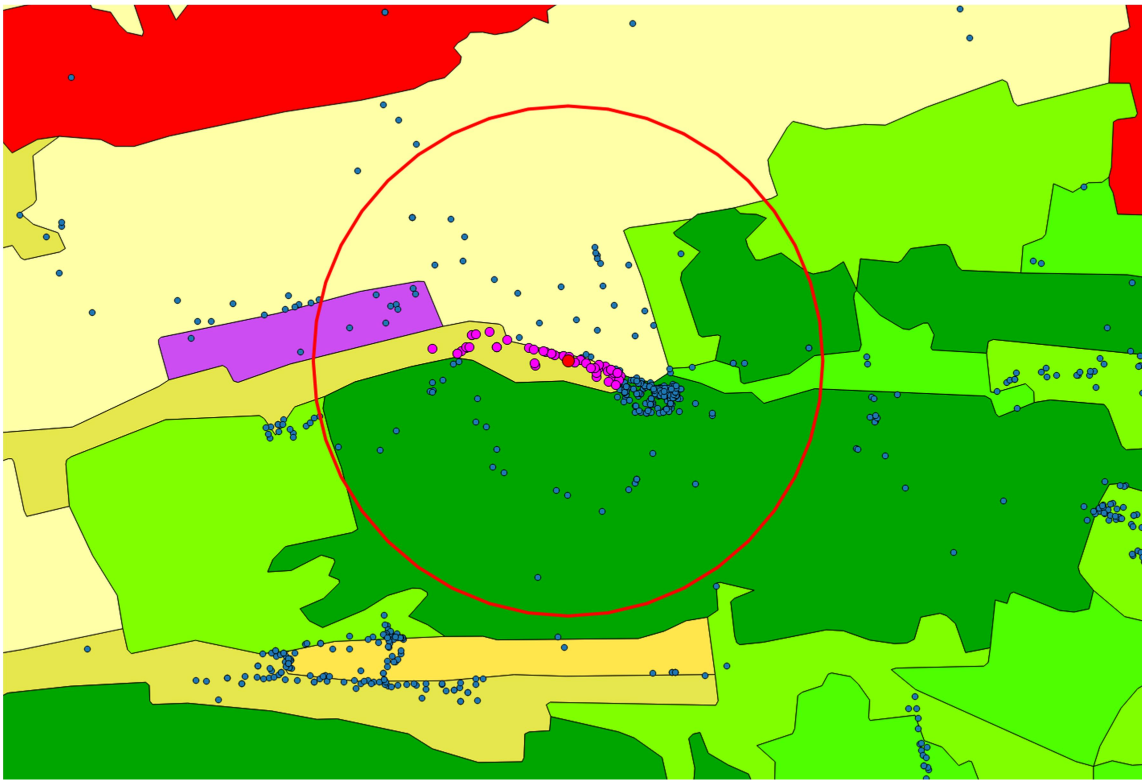


Figure 3.4.1: Schematic view of the basic principles of extraction of co-observations, using auxiliary polygons for focusing the search area. (Source of observations: ArtenFinder Rheinland-Pfalz. Red point: target observation; red circle: 1,000 m search radius; pink points: resulting associated observations; blue points: other observations. Source of polygons: CORINE land cover 2012 CLC10[43])

True habitat information covering large areas without gaps is not available. To produce it is time-consuming and costly, and such information is therefore mostly captured for small areas such as, at least in Germany, one or several neighboring districts ("Gemarkungen") at most, which is often done for planning purposes. While some municipalities (again, at least in Germany) possess such information, often even in digital form, bringing it together and harmonizing it would be a worthwhile yet very costly endeavor. Public authorities also provide data on certain protected habitat types (e.g., habitats protected by German nature protection law, or by the European Natura 2000 system), but they are not comprehensive, covering only certain habitat types. It was therefore necessary to choose auxiliary data which provide an approximation to this kind of information and which is comprehensive. A possible substitute which can be considered for this approach is land cover data. For the ArtenFinder data use case, which is situated in Europe, CORINE ("Coordination of Information on the Environment") land cover data provide a relatively up-to-date source of such information. Moreover, the freely available 2012 CLC10 product, with its 10 ha minimum size of area units, can be considered to have a suitable geometric resolution for the task at hand ("CLC10" stands for CORINE Land Cover, 10 ha as smallest area units). Of course, any land cover unit with small dimensions below this threshold is not represented in the data. This leads to an underrepresentation of land cover units which have small extents, such as small to medium-sized waterbodies, hedges, groves, isolated buildings, etc., in favor of large, coherent units such as forests, larger bodies of settlements, agricultural land, and others. Therefore it is far from representing true habitat information. However, it can still serve for providing polygon geometries which might be useful for the purpose at which this experiment is aimed: to render the search area for a target observation's context more precise than just using a circle, and to do so in a meaningful way. Data with too fine a geometric resolution might restrict resulting search areas too much, leading to too small context observation numbers for the approach as a whole to work, except,

[43] © GeoBasis-DE / BKG 2017

maybe, in regions with very high spatial observation densities. CORINE land cover data for Germany are available from the download site of the Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodäsie, BKG), from where a shapefile was downloaded, then transferred to a spatial database for further processing. Documentation for the data is available from the same source. The dataset provides a continuous set of polygon geometries based on land cover information derived from Landsat multispectral remote sensing data. CORINE land cover data use 44 distinct land cover classes, of which 27 classes are present in Rheinland-Pfalz. Table 7.3.11 in the appendix lists these classes.

CORINE land cover data are only available for Europe. For the iNaturalist data use case, the most suitable source of polygon geometries based on land cover is the NLCD (National Land Cover Database) 2011 Land Cover (2011 Edition, amended 2014). Similar to CORINE, this dataset also derived from Landsat satellite imagery. It has a somewhat smaller thematic resolution, with 20 classes, of which 16 occur in California. However, the original NLCD dataset has a much higher spatial resolution, with smallest units of ca. 0.1 ha (corresponding to the 30x30m spatial resolution of Landsat imagery). It was therefore necessary, for reasons explained above, to raise minimum size of polygon geometries. This was achieved by converting the spatial resolution to 316.23x316.23 m, corresponding to ca. 10 ha, the minimum spatial unit size of the CORINE data used. In this way, the sensitivity analysis could be conducted with land cover data with smallest polygons having the same size for both data use cases, but at the cost of a lower-than original geometric resolution in NLCD data. Thematic resolution in NLCD data was not reduced by the applied conversion. Table 7.3.12 in the appendix lists the relevant NLCD land cover classes.

Another interesting dataset, which is globally available, are Ecological Land Units (ELUs), developed in a joint effort by USGS and ESRI. In this dataset, geographic information on bioclimates, landforms, lithology, and land cover are consolidated into a single information layer. Sayre et al. (2014) describe in detail the dataset's source data, methods of generation, and properties. They also provide a detailed discussion of the terms ecosystem and habitat. The authors elucidate that the ELUs are regarded as ecosystems, and that ecosystems are mostly encompassing multiple habitats, while habitats are a concept mostly referring to a particular species. They also observe that land cover information usually has an emphasis on vegetation structure, at least outside settled areas. Thus, both ELUs and land cover units are only approximations to the habitat structure of an area. The ELU data used for analysis here have a geometric resolution of 250 m, which is comparable to the CORINE and (processed) NLCD data described above. What distinguishes the ELU data is their far greater thematic resolution of 3,923 classes worldwide, of which 140 occur within Rheinland-Pfalz, and 789 within California. However, as explained above, the thematic resolution is of little consequence as the thematic information is not directly used in this methodological modification. In the Rheinland-Pfalz area of interest, ELUs have a mean size of 780,989 m^2, which is very close to the CORINE data's value for that region (854,762 m^2). In California, both the processed NLCD data (1,956,905 m^2) and ELUs (2,821,663 m^2) have notably larger areal units, on average.

The experiments conducted with the data described above allowed both for examining the effects of the use of auxiliary geometries for search areas in general, as well as for investigating whether ELUs, with their extended thematic basis, hold a higher potential for positively influencing results, than does "pure" land cover information. However, it should be noted that this modification of the methodology introduces an extrinsic element into the observed communities approach, which therefore cannot be considered to be an altogether intrinsic approach anymore (following the definition used in this work and laid out in section 1.4) if this modification is applied. Also, restricting search areas by using auxiliary polygon geometries will generally lead to smaller search areas and thus smaller amounts of context observations which can be evaluated, from which a drop in numbers of valid observed communities and candidate observations can be expected.

3.4.5 Using a Quantitative Similarity Index

The observed communities approach was evaluated with two so-called binary similarity indices, which means that the information they are using is just the presence or absence of a species in the species lists which are compared. There also exist similarity coefficients which use some quantitative information about the observations involved. The goal in the sensitivity analysis described here was to analyze whether quantitative instead of binary similarity calculation is able to produce better results in terms of discrimination of plausible or implausible observations.

Zuur et al. (2007) present a quantitative similarity coefficient, called the Similarity Ratio, which can be viewed as the quantitative counterpart of the Jaccard index, because its results are identical to those of the Jaccard index if used with binary data. Using the notation from Zuur et al. (2007), this coefficient can be represented as follows:

F3: Similarity Ratio
$$\mathrm{SR} = \frac{\sum_k Y_\mathrm{k} X_\mathrm{k}}{\sum_k Y_k^2 + \sum_k X_k^2 - \sum_k Y_\mathrm{k} X_\mathrm{k}}$$

Y_k and X_k represent quantitative information about a species k at the sites Y and X. Usually abundance is used for quantification, that is, the observed number of individuals of a species. Reliable abundance information is usually not available from casual citizen science observations of organisms, and this is also the case for the two datasets used in this work. What is available, however, is the frequency of observation of a context species close to a target or candidate observation. This parameter is, of course, quite different from real abundance, in that it measures the frequency at which a species was observed by volunteers. It does not represent the number of observed individuals of a context species. Each observation represents at least one individual, but it is unknown how many individuals were actually present. In many cases, it is even quite probable that the same individual was observed many times at the same place, by the same observer, or by several observers, during a shorter or longer period of time. This creates a seeming abundance which has its roots in the perpetual nature of the observation process and ultimately in the VGI nature of the data collection process. Observation frequency measures the abundance of observation events, not the abundance of the species. It can certainly be argued that observation frequency is a function of species abundance: more abundant species are more likely to be observed than less abundant species, especially by relatively untrained volunteers. Many other factors might also play a role, especially detectability, observer-related factors (e.g., their interests and abilities), and more. Using this parameter therefore adds information originating from the biological and ecological properties of the species, as well as from the VGI nature of the data collection process. In similarity calculation, this parameter gives a larger weight to species which are observed more often in the context of an observation, for various reasons. Context species which were more frequently observed are thus more important in describing the context as a whole, than are less frequently observed species.

In observed community extraction, the mean observation frequency of a context species in all target species contexts (including cases with no observation of a context species) was used to set this parameter. In candidate contexts, where a single candidate observation is considered in each candidate case, observation frequency simply represents the frequency of observations of a context species around the candidate case. This parameter was used as a substitute for abundance in calculating the Similarity

Ratio between candidate contexts and observed communities in a sensitivity analysis for examining the effects of this modification to evaluation results.

A quite different approach of adding a quantitative dimension to the observation data used for similarity calculation is to use distances of context observations to the target or candidate observation. For geographic data, this is a natural notion of weighting the relationship of objects. Context observations which are closer to a target or candidate observation could be considered to be more important to describe the context, than are observations which are further away. To test the effect of this information on evaluation results in a proper sensitivity analysis, an inverse distance weighting was introduced, by calculating the distance of a context observation to the target or candidate observation, converting it to percent of search radius (thus making the value independent of the size of the search radius used), and inverting it. There is thus a linear growth of the weight with decreasing distance to the observation. For context species with more than one context observation, the mean distance of all context observations to the target or candidate observation was calculated. These values were then used to weigh species in Similarity Ratio calculation. A unified search radius of 1,000 m was kept as the maximum distance of relevant context observations used in this procedure.

3.4.6 Examining Edge Effects

An issue which should always be considered in analyses such as the one presented here, are edge effects. These arise in this methodology because any observation which is closer than the search radius to the edge of the area of interest has a neighborhood including a part which lies outside of the area of interest and therefore does not have context observations. Evaluation was conducted without any edge effect correction, because no critical effects on evaluation results were expected. O’Sullivan and Unwin (2010) discuss edge effects on distance based point pattern measures and point out that edge effects are usually relevant only in cases where the number of events (here: the number of observations) is small. This is certainly not the case in this study, which is why edge effects were not corrected in evaluation. The ArtenFinder project focuses on the area of the federal state of Rheinland-Pfalz. There are a limited number of observations in areas adjacent to Rheinland-Pfalz, because volunteers’ interest in observing species does not necessarily stop at the state line. The project portal allows for submitting observations also from areas outside of Rheinland-Pfalz. However, observation numbers decrease rapidly in these areas, and observations outside of Rheinland-Pfalz were not used in this work. In the iNaturalist use case, California was chosen as the area of interest. However, iNaturalist observations rather concentrate in the coastal areas of California (see section 2.1.2), so that edge effects caused by choosing the state line as the spatial limit of the data use case can be expected to be low, because there are only relatively few observations close to the state line.

Edge effects caused by state lines should be examined for both data use cases, to gain an insight in how far they might influence evaluation results presented here. This was done in a sensitivity analysis using a guard zone approach (O’Sullivan and Unwin 2010). Guard zones were produced by creating an inner buffer in the area of interest of 1,000 m along state lines. Only observations not situated in this zone were then used as target observations in observed community and candidate context extraction, while context observations could also be situated within the guard zone. As the new target and candidate observation data have a reduced spatial extent, results using edge effect correction cannot be compared to the original evaluation results. A proper set of evaluation results with reduced spatial extent and without edge effect correction was calculated to provide a suitable basis of comparison to evaluate effects of the edge effect correction. iNaturalist data with their global scope, which are theoretically able to provide a true guard zone outside the California state line, were treated likewise to maintain methodological consistency.

3.4.7 Overcoming Ambivalent Information of Simpson Index Values

The two binary similarity indices used in evaluation of the observed communities approach have different structures. The Jaccard index, on the one hand, calculates the ratio of the size of the intersection of two species lists and the size of their union. Therefore, all Jaccard index values are based on calculations using the same elements, and the resulting index value can always be interpreted as the rate of species which the two lists have in common in all species found in the two lists. The Simpson index, on the other hand, calculates the ratio of the size of the intersection of two species lists and the size of the smaller of the two lists involved. For the Simpson index, it is therefore not clearly determined which element is used as denominator. For the use cases and data used here, there are cases where the size of the observed community is used as denominator and other cases where the size of the candidate context is used. In the former cases, the index value is therefore the rate of observed community species covered by the species found around the candidate observation. In the latter case, the index value represents the rate of observed community species found in the species surrounding the candidate observation. This problem arises from the fact that the two species lists involved are of different origin and (usually) of different size: observed communities represent a typical situation for a species' observation which is aggregated from many observations of that species and reduced to frequent co-observations, while the list of all species found around a candidate observation represents a single observation's environment. The former is smaller, in most cases, than the latter.

For the sake of a more straightforward interpretation of index values, it might make sense to derive, from the Simpson index, a variant by fixing the denominator to the size of the observed community. It can be defined as follows

F4: Variant of Simpson index

$$\text{Svar} = \frac{a}{a+b}$$

where *a* is the number of species that both lists have in common (their intersection), and *b* the number of species unique to the observed community. Behavior of this variant of the Simpson index was tested in a sensitivity analysis using the index variant, and results were compared to those with the original Simpson index.

3.4.8 Using Date-Specific OSM Context

Sensitivity analysis was not conducted with the OSM environments approach to the same extent as conducted with the observed communities approach. However, sensitivity if the OSM environments approach was examined with respect to a methodological modification which probably has special relevance for this approach: the use of time-specific geographic OSM context. OSM is a highly dynamic data source. The dataset changes for several reasons:

- OSM is still a young project whose data are still developing and becoming more complete.
- The OSM project is itself a dynamic process accompanied by vivid discussions of many aspects within the OSM community, which lead to changes in project policy as well as in technology.
- The real world is changing constantly, and many of the changes relevant to OSM are constantly updated in the dataset by volunteers.

Using OSM at a recent state allows for taking advantage of the most complete OSM data available. However, the above factors, especially real world changes reflected in changes to OSM data, may offer an advantage which may improve results. OSM allows for accessing the full history of all elements ever recorded in its data. This provides the opportunity to examine the state of OSM at a past point in time. OSM might therefore reflect an observation's geographic context as it was at the observation date. This might present an advantage over using a recent OSM state, because OSM context at the observation date might fit the actual observation situation better.

Going back in time in this way has several consequences. The desired effect is, of course, that real world changes which happened since an observation was made in a certain place (e.g., construction of a residential area where fields used to be) are corrected for. This is especially relevant for observations used for OSM environment extraction, because these are always "historic" observations which may be several years old. Candidate observations are usually younger, at least in schemes such as ArtenFinder where new observations are soon evaluated. However, to use the full potential of the methodological change proposed here, it makes sense to also extract candidate contexts from OSM data specific to the observation date. Using an older state of OSM may also have undesirable effects, especially a lower completeness of data. Objects may not yet have been mapped, or improperly tagged. Also, errors may still have existed in earlier OSM data stages which were later corrected. Therefore, there are also good reasons for using OSM in its current state. A sensitivity analysis with date-specific OSM environments and candidate OSM contexts was conducted to test the behavior of the OSM environments approach with time-adapted OSM information and to compare results to those with recent OSM data used in evaluation. Making use of HeiGIT's OSHDB (see section 3.3.4) allows for "turning back time" when extracting an OSM environment, or a candidate OSM context. It contains the full history of each OSM object, that is, all edits made to the data, back to 2007-11-01. Any query detecting the OSM content found around an observation can be set to the point in time at which the observation was actually made. In this way, all spatial queries on OSM concerning the neighborhood of an observation were set to the observation date, looking at OSM in the state it was in at the observation date. For technical reasons (mostly for reducing computing time), the actual query date was always set to the first of the month the observation was made in, leading to a maximum offset between the state of OSM and the date of observation of one month.

3.5 Taxonomy

An important issue when dealing with biodiversity data is taxonomy. Different citizen science projects use different taxonomical systems. In analyses where observation data of two (or more) data sources are joined or mixed, this fact requires harmonization of taxonomy, e.g., by using appropriate online services (Jacobs & Zipf 2017). This problem is of no consequence in analyses which do not mix observation data of organisms from several different sources. The observed communities approach and the OSM environments approach are of this kind, and so is the evaluation methodology explained above. ArtenFinder Rheinland-Pfalz uses a taxonomy which corresponds to the species occurrence databases of the federal state of Rheinland-Pfalz, because ArtenFinder observations are collected with the goal of integrating them into these databases (see section 2.1.1). iNaturalist uses an extensive list of global and regional taxonomic authorities to build its taxonomic system and carefully explains the reasons for doing so in its curator guide[44]. One of the main reasons is founded in the citizen science nature of the project. Most taxonomic authorities' information is freely and publicly available, so that all potential users, including amateurs, have access to it, other than, e.g., to many scientific publications on taxonomic issues. In this way, taxonomy has a maximum degree of transparency. iNaturalist curators, many of whom are recruited from experienced iNaturalist users, are provided with the authority to change iNaturalist's taxonomy and are encouraged to collaborate with other curators in doing so. They are asked to adhere to certain rules explained in the curator guide mentioned above.

In this work, observation data of the different projects were not used together in the same analyses, but always treated separately. It was therefore not necessary to harmonize taxonomies of the observation data used in this work, allowing for respecting the projects' taxonomies. This implies that in subsequent chapters, some species names may appear which are considered to be outdated by most taxonomic authorities, especially in ArtenFinder data. For instance, ArtenFinder still uses *Parus caeruleus* for Blue Tit or *Parus palustris* for Marsh Tit. For these species, *Cyanistes caeruleus* and *Poecile palustris* are mostly used nowadays, which is also the case in iNaturalist. This remains, however, without consequence for analysis here, because both datasets are not joined or mixed.

Scientific names of species and of higher taxonomic levels may change over time. If a scientific name in a project's taxonomy is changed, this change must be applied to the whole dataset, to avoid inconsistencies. However, for some changes, this is not a trivial task. iNaturalist's curator guide elaborates that changes can be of several different types, which implicate different levels of difficulty: while a 1:1 change in a (sub)species, genus, or higher level taxon is basically quite straightforward, a taxon split presents the most difficulties, because often the system cannot automatically decide to which of the new taxa an observation belongs. Where possible, iNaturalist uses information about the spatial distribution of species to resolve such cases, because often species are split by geographic criteria. However, problems introduced by overlaps of species ranges and other issues remain. Also, iNaturalist users can opt out of automatic taxonomy changes for their data, so that even in cases which can be automatically resolved, there is no guarantee that this will happen for all data.

Taxon changes introduce some measure of uncertainty into the internal taxonomic consistency of the datasets, especially in relatively open systems such as the one applied in iNaturalist. The extent of this uncertainty is hard to gauge, but is probably small. For instance, certain users deciding to abide with an old taxon will probably remain the exception. Also, curators can be expected (and are asked by the curator guide cited above) to manually resolve most uncertain cases. To make an example from iNaturalist, the species *Cervus elaphus*, which was formerly used for Red Deer in Europe as well as for Wapiti in North America, was split into *C. elaphus* (for Europe) and *C. canadensis* (for North Ameri-

[44] https://www.inaturalist.org/pages/curator+guide, last accessed on 2018-11-09

ca) in October 2017. As of 2018-11-09, just one observation of *C. elaphus* in North America (situated in Mexico) remained in the dataset, as opposed to 3,885 that were switched to *C. canadensis*. The observer reasoned in this case that the species was introduced in the area the observation comes from and therefore properly labeled with *C. elaphus*. Thus, at least in this example unresolved cases did not persist.

4 Results

Distributions of similarity values of sets of candidate observations, and the differences between them, are the principle results of the evaluation. They demonstrate whether the observed communities approach and the OSM environments approach are basically able to estimate the plausibility of observations using similarity values. Evaluation also rendered more aspects about the observed communities, about the OSM environments and about the sets of candidate observations. They hold important information for the interpretation of the evaluation results and their discussion (see chapter 5) and are therefore also presented here.

4.1 Evaluation of the Observed Communities Approach

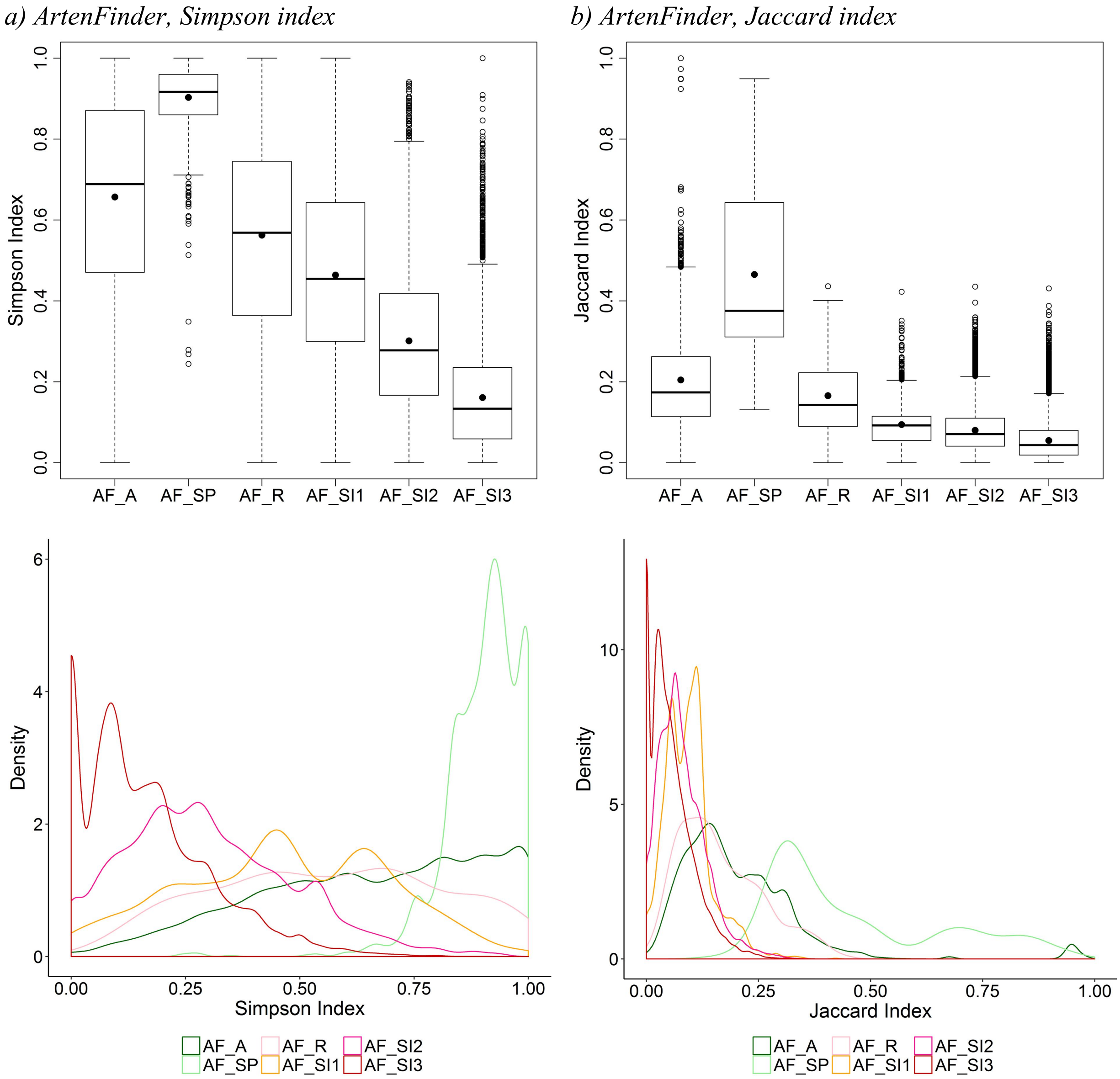
Evaluation of the observed communities approach was conducted using the methods described in sections 3.1 and 3.3, for both data use cases, with two different binary similarity indices (the Simpson index and the Jaccard index). Parameter settings used in evaluation were also described in section 3.1, and are summarized as follows:

- Extraction of OSM environments:
    - Uniform search radius of 1,000 m for relevant neighborhood for all species groups,
    - target species: only species with 100 or more approved observations up to 2015,
    - frequency threshold for context species frequently observed in proximity to the target species: 0.5.
    - frequency threshold for nonspecific species (context species which are part of many observed communities at the same time, and therefore removed from the final observed communities): 0.5.
- Extraction of candidate context:
    - Uniform search radius of 1,000 m for relevant neighborhood for all species groups,
    - Nonspecific species removed from candidate contexts.
- Comparison of similarity values:
    - Candidate cases restricted to cases where the size of both the observed community and of the candidate context is 10 or larger,
    - candidate observations of nonspecific species not used in analysis.

4.1.1 Evaluation Results, Observed Communities Approach, ArtenFinder Data

Similarity values

Distributions of Simpson similarity index values for ArtenFinder data are shown in Figure 4.1.1, part a. Differences in distributions of similarity values are largest between approved observation data (set AF_A) or synthetic plausible candidate observations (set AF_SP) on the one hand, and sets of random synthetic implausible candidates (sets AF_SI2 and AF_SI3) on the other hand. Rejected observations (set AF_R) and synthetic implausible candidates based on physically similar species (set AF_SI1) are closer to AF_A and AF_SP. Table 4.1.1 summarizes the results of the statistical tests employed to examine these differences. Test results allow for accepting the alternative hypothesis ($p \leq 0.05$) of the Mann-Whitney-U-Test for all comparisons of sets of accepted or synthetic plausible candidate observations and rejected or synthetic implausible candidate observations. Distributions of Simpson index values of accepted and synthetic plausible candidate observations are statistically different from distributions of Simpson index values of rejected or synthetic implausible observations according to this test.



a) ArtenFinder, Simpson index

b) ArtenFinder, Jaccard index

Figure 4.1.1: ArtenFinder, distributions of Simpson and Jaccard similarity index values, observed communities approach. n(AF_A) = 22,426; n(AF_SP) = 1,486; n(AF_R) = 362; n(AF_SI1) = 1,718; n(AF_SI2) = 22,197; n(AF_SI3) = 22,896.

Evaluation results with the Jaccard index, presented in Figure 4.1.1, part b, show that index values are overall much lower than with the Simpson index, which is to be expected considering the index’ structure (see section 3.1). AF_SP distribution, with high index values, is well distinguished from AF_SI1, AF_SI2 and AF_SI3, the latter forming a group of distributions with very low index values. AF_A Jaccard index distribution is closer to this group than to AF_SP, and AF_R is very close to AF_A. Analog to Simpson index results, results with the Jaccard index exhibit significant differences in distributions of similarity values between candidate observations accepted by validating experts (AF_A) or synthetic plausible candidates (AF_SP) on the one hand, and the sets of rejected candidate observations (AF_R) or synthetic implausible candidates (AF_SI1, AF_SI2, AF_SI3) on the other hand. Table 4.1.2 summarizes these test results. Distributions of Jaccard index values of accepted and synthetic plausible candidate observations are statistically different from those of rejected or synthetic implausible observations according to this test. Figure 4.1.1, however, also shows that there are overlaps between the distributions with both indices, with no perfect separation between sets of plausible or implausible observations.

Table 4.1.1 and Table 4.1.2 also give the results of the Fligner-Killeen-Tests conducted to examine whether variances in the distributions which are compared are homogeneous or not. This bears on the interpretation of the test results of the Mann-Whitney-U-Tests which were used to examine the differences between these distributions, see section 3.3.1. Results of the Fligner-Killeen-Test are significant in almost all cases, except AF_A vs. AF_R. This means that variances are not homogeneous in most cases.

Table 4.1.1: ArtenFinder, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with AF_A, AF_SP and the four different sets of rejected or synthetic implausible candidate observations, Simpson index, observed communities approach.

Simpson Index, AF_A vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	0.3922	$1.57*10^{-10}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$1.155*10^{-12}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Simpson Index, AF_SP vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Table 4.1.2: ArtenFinder, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with AF_A, AF_SP and the four different sets of rejected or synthetic implausible candidate observations, Jaccard index, observed communities approach.

Jaccard Index, AF_A vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	0.0005092	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$1.657*10^{-7}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Jaccard Index, AF_SP vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Properties of Observed communities

There are 291 (of 2,957) species with 100 or more accepted observations in the ArtenFinder data up to 2015. After their lists of surrounding species are filtered for frequently associated species and for nonspecific species (both with a threshold of 0.5), 216 observed communities with 10 or more species remain. The list of species which are identified as nonspecific species has 46 species (see Table 5.1.7 in the discussion chapter). These species and their observed communities are themselves not used for evaluation purposes, because, as nonspecific species, they can be considered to be species with rather unspecific list of associated species. Species composition of observed communities will be discussed in more detail later. 13 of them have observed communities with less than 10 species anyway, so that this consideration reduces the number of valid observed communities further to 183 species left for use in evaluation (with observed community sizes ranging from 10 to 119 species, mean: 31.7). Table 4.1.3 gives an overview of these numbers for ArtenFinder data.

Table 4.1.3: ArtenFinder, key numbers describing valid observed communities.

No. of valid observed communities	**Mean no. of species in observed communities**	**No. of nonspecific species**
183	31.7	46

Valid observed communities belong to 13 species groups. Most of them are birds (33.9%), followed by butterflies and moths (26.8%) and dragonflies and damselflies (16.4%). Together, the target species of all valid observed communities represent 23,475 of the accepted candidate observations (AF_A), or roughly one third of the candidates in this set. Similar ratios apply to the other sets of candidate observations. The number of valid candidate observations is further reduced to a minor degree by the fact that some candidate observations have less than 10 context species.

Properties of valid candidate observations

Table 4.1.4 lists some key numbers describing the sets of valid candidate observations resulting from evaluation with ArtenFinder data. Sets of valid candidates are different in size because their source sets have different sizes (see section 3.3.2). Mean numbers of species in candidate contexts and mean numbers of species in observed communities associated to these cases also differ between sets: relative to the AF_A set, they are higher for AF_R (real rejected observations) and lower for AF_SP and AF_SI1. Mean numbers of context observations are lower for AF_R and higher for AF_SP and AF_SI1. Valid AF_R observations are thus situated in places with lower observation density, but higher observed species diversity, on average. AF_SP and AF_SI1 observations are found in contrarious situations, if compared to the average AF_A case. Set AF_SI2 shows properties close to AF_A, because valid accepted observations form the base for its generation, and it was designed to mirror AF_A's spatial properties (see section 3.3.2). AF_SI3 deviates strongly in the mean size of its candidate contexts and in observation density around candidates, which are both much lower due to the method of creation of this set (see section 3.3.2).

Table 4.1.4: ArtenFinder, key numbers describing sets of valid candidate observations, observed communities approach.

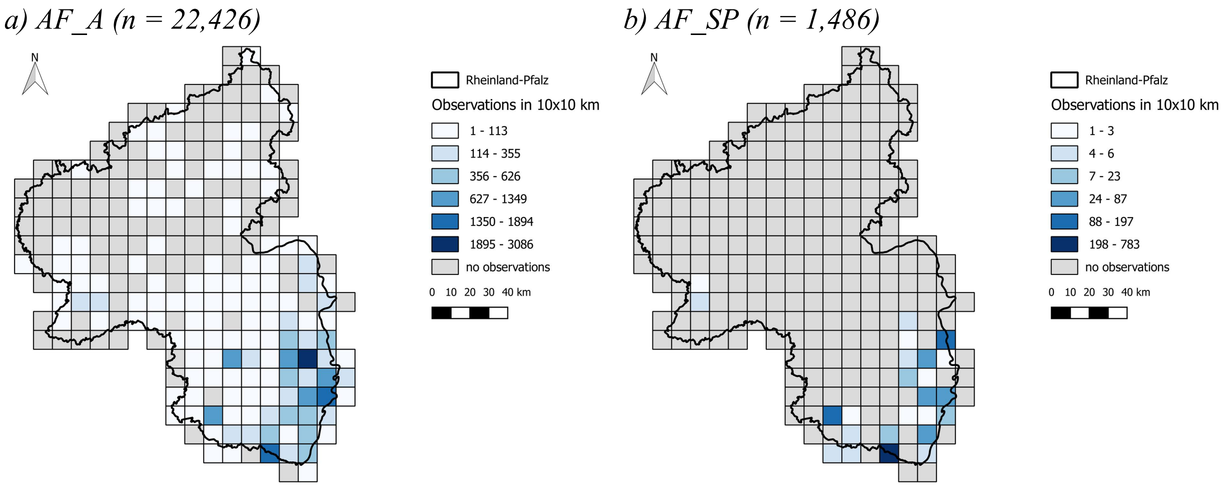
Set of candidates	**No. of valid candidate cases**	**Mean no. of species in candidate contexts**	**Mean no. of species in observed communities, per set of candidates**	**Mean no. of context obs.**
AF_A	22,426	108.2	34.7	1,225.4
AF_SP	1,486	96.5	54.0	3,090.5
AF_R	362	115.1	42.0	898.3
AF_SI1	1,718	99.5	21.6	1,511.8
AF_SI2	22,179	102.5	34.7	1,086.2
AF_SI3	22,896	31.2	35.2	112.7

Table 4.1.5 summarizes the species group compositions of the sets of valid candidate observations (that is, candidate observations actually used in evaluation). Due to the origin of AF_SI2 and AF_SI3, they are almost identical to AF_A. AF_SI1 and AF_SP have more birds and less butterflies than these two, and reptiles rank second in AF_SI1, instead of the group of dragonflies and damselflies. AF_R is dominated by dragonflies and damselflies, and all other species groups also strongly differ here from the other sets. Compared to the original composition of the accepted observations used as candidates (see Table 3.3.2), AF_A remains dominated by birds, while experiencing a boost in the group of dragonflies and damselflies. Butterflies and plants decline.

Table 4.1.5: ArtenFinder, portions of species groups in sets of valid candidate observations, observed communities approach.

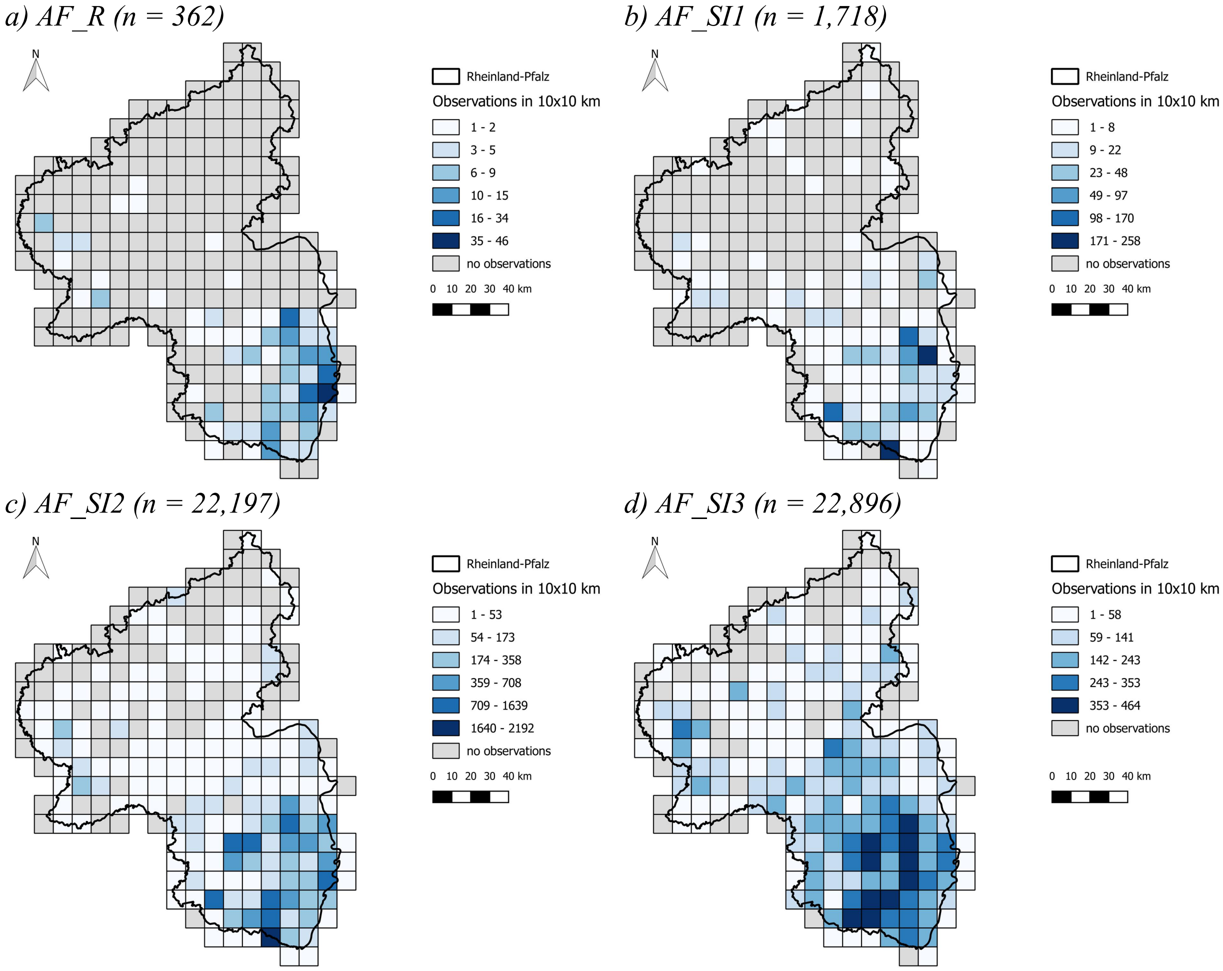
Species group	AF_A (%)	AF_SP (%)	AF_R (%)	AF_SI1 (%)	AF_SI2 (%)	AF_SI3 (%)
plants	0.8	0.1	0.0	0.0	0.7	1.0
fungi	0.3	0.0	0.0	0.0	0.3	0.4
mammals	1.9	1.3	0.3	0.0	1.8	2.0
birds	47.2	73.0	7.5	60.5	47.1	48.0
reptiles	0.8	0.0	0.6	22.9	0.8	0.8
amphibians	2.2	0.6	8.0	0.0	2.2	2.4
butterflies and moths	15.4	2.6	19.3	8.8	15.3	15.4
hymenopterans	1.0	0.0	0.3	0.0	1.0	1.0
dragonflies and damselflies	27.6	22.3	54.4	3.0	27.7	26.0
mantids	0.3	0.0	0.0	0.0	0.3	0.3
locusts	2.1	0.1	9.7	0.0	2.1	2.2
mollusks	0.2	0.0	0.0	4.8	0.2	0.2
true bugs	0.3	0.0	0.0	0.0	0.3	0.3

Figure 4.1.2 and Figure 4.1.3 examine the spatial distribution of valid candidate observation (that is, using only observations which passed the filter effects occurring in evaluation). All sets show the same general northwest to southeast trend of growing observation density. AF_SP candidate observations concenter in the high-observation-density areas in the southeast, while AF_SI3 observations are more dispersed than in the other sets, but are also predominantly found in the southeastern part of the area of interest, because valid candidate cases with a sufficient number of context species can only occur in areas with sufficient numbers of context observations.


a) AF_A (n = 22,426)

b) AF_SP (n = 1,486)

Figure 4.1.2: Spatial distribution of valid candidate observations in sets of accepted and synthetic plausible ArtenFinder candidates, observed communities approach. (No. of points in 10x10 km raster. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)



a) AF_R (n = 362)

b) AF_SI1 (n = 1,718)

c) AF_SI2 (n = 22,197)

d) AF_SI3 (n = 22,896)

Figure 4.1.3: Spatial distribution of valid candidate observations in sets of rejected and synthetic implausible ArtenFinder candidates, observed communities approach. (No. of points in 10x10 km raster. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)

4.1.2 Evaluation Results, Observed Communities Approach, iNaturalist Data

Similarity values

Evaluation results for the observed communities approach with iNaturalist data are comparable, in all respects, to results obtained with ArtenFinder data. Figure 4.1.4 shows the distributions of Simpson similarity index values with iNaturalist data, and Table 4.1.6 presents numeric results of the statistical test used for assessing differences between distributions of similarity values of plausible or implausible observations. The data exhibit significant differences in distributions of similarity values between approved or synthetic plausible observations (iNat_A and iNat_SP) and the sets of synthetic implausible observations (iNat_SI1, iNat_SI2 and iNat_SI3). They allow for accepting the alternative hypothesis ($p \leq 0.05$) of the Mann-Whitney-U-Test for all comparisons of sets. According to this evaluation, distributions of Simpson index values of accepted and synthetic plausible candidate observations are statistically different from distributions of Simpson index values of synthetic implausible observations. The same is true for Jaccard index values, see Figure 4.1.4, and Table 4.1.7. Variances of the distributions which were compared were found to be not homogeneous by the Fligner-Killeen-Test (see section 3.3.1 for details).

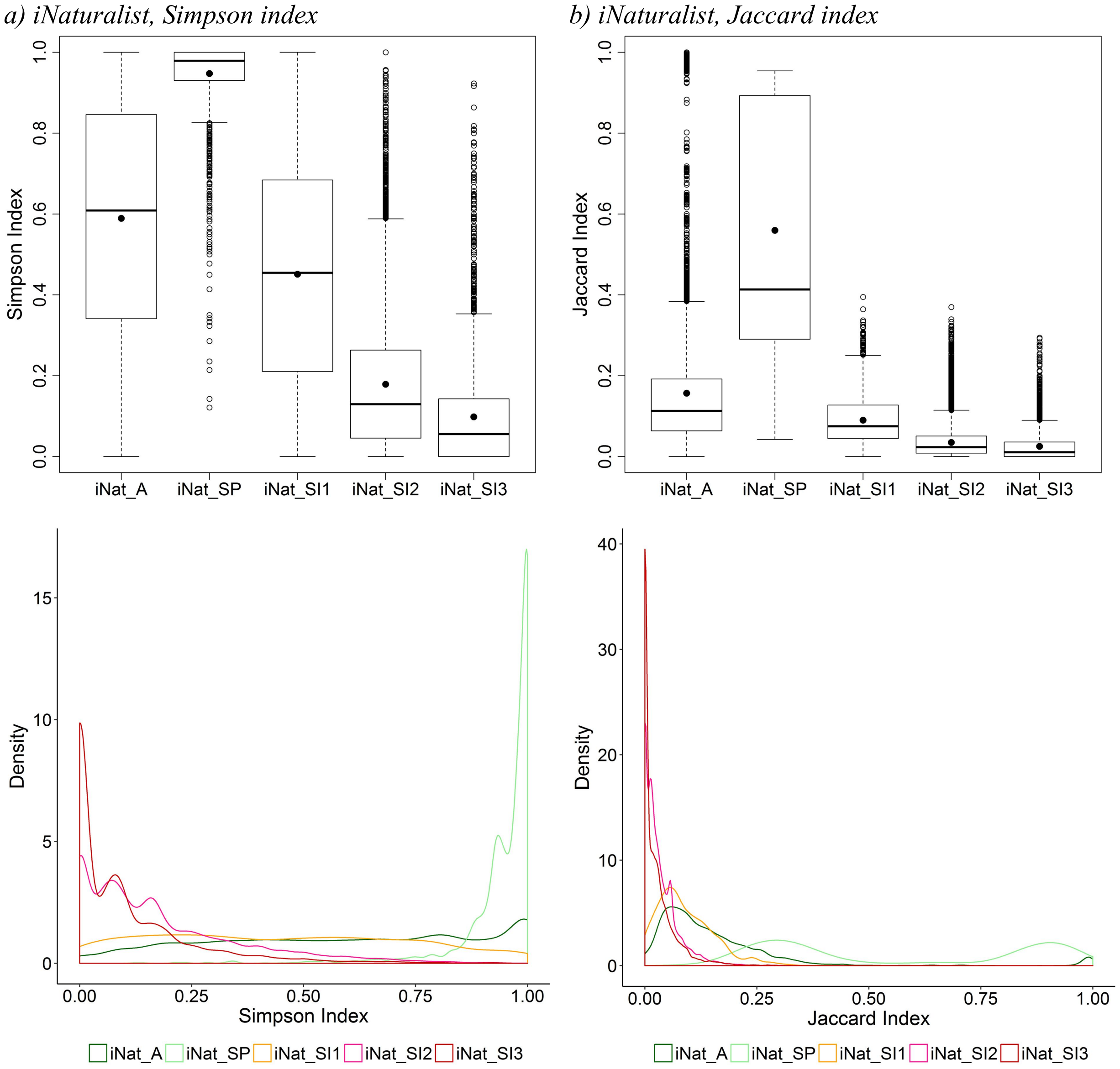


Figure 4.1.4: iNaturalist, distributions of Simpson and Jaccard similarity index values, observed communities approach. n(iNat_A) = 34,821; n(iNat_SP) = 2,415; n(iNat_SI1) = 2,216; n(iNat_SI2) = 34,485; n(iNat_SI3) = 4,768.

Table 4.1.6: iNaturalist, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with iNat_A, iNat_SP and the three different sets of synthetic implausible candidate observations, Simpson index, observed communities approach.

Simpson Index, iNat_A vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	0.0002067	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Simpson Index, iNat_SP vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Table 4.1.7: iNaturalist, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with iNat_A, iNat_SP and the three different sets of synthetic implausible candidate observations, Jaccard index, observed communities approach.

Jaccard Index, iNat_A vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Jaccard Index, iNat_SP vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Properties of Observed communities

iNaturalist data from California up to 2015 have 549 species with 100 or more research grade observations. Filtering for frequently associated species (with a threshold of 0.5) leaves 484 non-zero observed communities and 234 observed communities with 10 or more species (ranging from 10 to 534 species, mean: 86.9 species per observed community). A threshold of 0.5 for finding and eliminating nonspecific species results in no nonspecific species in the observed communities for this dataset. This means that no species is present in 50% or more of observed communities. Compared to the ArtenFinder data, this points to higher species diversity in the observed communities, already visible in the much higher number of species represented in the data, and in the higher mean number of species in observed communities. This is certainly to be expected in a much larger and much more ecologically diverse area of interest. To keep the methodology comparable to the evaluation with ArtenFinder data, the threshold for nonspecific species was not changed to a lower value. Table 4.1.8 provides an overview of the numbers cited above.

Table 4.1.8: iNaturalist, key numbers describing valid observed communities.

No. of valid observed communities	**Mean no. of species in observed communities**	**No. of nonspecific species**
234	86.9	0

41.0% of valid observed communities are of birds, followed by plants (29.5%) and mollusks (mostly marine mollusks, 14.5% of observed communities). The target species of the valid observed communities represent 41,576 of the accepted candidate observations (iNat_A), or roughly one fourth of these candidates.

Properties of sets of candidate observations

Some key parameters characterizing the sets of valid candidate observations resulting from the evaluation with iNaturalist data are presented in Table 4.1.9. Source sets having different sizes (see section 3.3.2), sets of valid candidates are also different in size. iNat_SP has, on average, the largest candidate contexts, as well as the largest observed communities associated to these cases. Also, its candidates are situated in locations with very high observation densities. iNat_SI1, iNat_SI2 and iNat_SI3 all have lower values for mean numbers of species in candidate contexts, and of context observations, if compared to the average iNat_A case (iNat_SI3 extremely so). iNat_SI1 also has a lower number of species in observed communities associated to its cases.

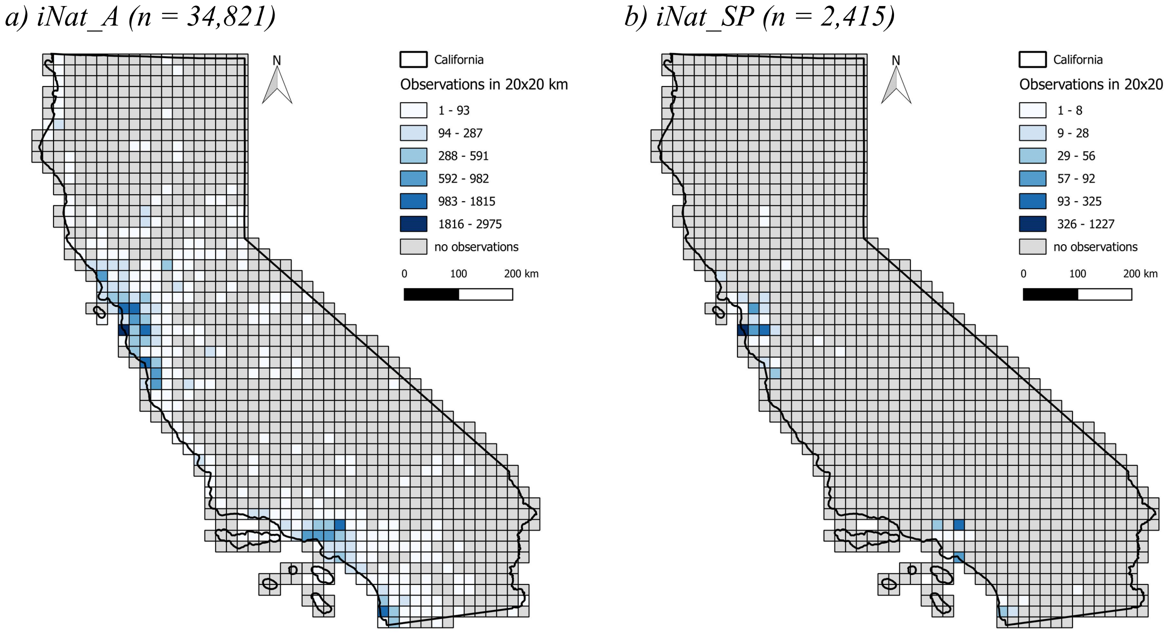
Table 4.1.9: iNaturalist, key numbers describing sets of valid candidate observations, observed communities approach.

Set of candidates	No. of valid candidate cases	Mean no. of species in candidate contexts	Mean no. of species in observed communities, per set of candidates	Mean no. of context obs.
iNat_A	34,821	132.3	65.2	753.1
iNat_SP	2,415	316.9	260.1	3,957.5
iNat_SI1	2,216	106.4	23.7	377.3
iNat_SI2	34,485	127.6	65.5	487.9
iNat_SI3	4,768	34.4	71.8	56.0

Looking at species group compositions of the sets of valid candidate observations (that is, observations actually used in the evaluation, Table 4.1.10), iNat_A, iNat_SI2 and iNat_SI3 have again very similar species group compositions, which is due to the method used for creating iNat_SI2 and iNat_SI3. Selection of valid research-grade candidate observations in the process of evaluation brought a pronounced change to the thematic properties of iNat_A: they are now clearly dominated by bird observations, which make up about half of the valid candidate observations. Plants now rank only second, followed by mollusks, butterflies, and mammals. The class of "other species" (species not assignable to any of the groups used in these data) holds ca. 6-7% of valid candidates in this set and in iNat_SI2 and iNat_SI3. Reptiles, beetles and others make up the remaining candidates. Set iNat_SI1, containing only species selected for their special properties (physically similar species living in different habitats) contains mostly birds as well as some plant species. iNat_SP is dominated by mollusks, followed by birds, plants and "other species".

Table 4.1.10: iNaturalist, portions of species groups in sets of valid candidate observations, observed communities approach.

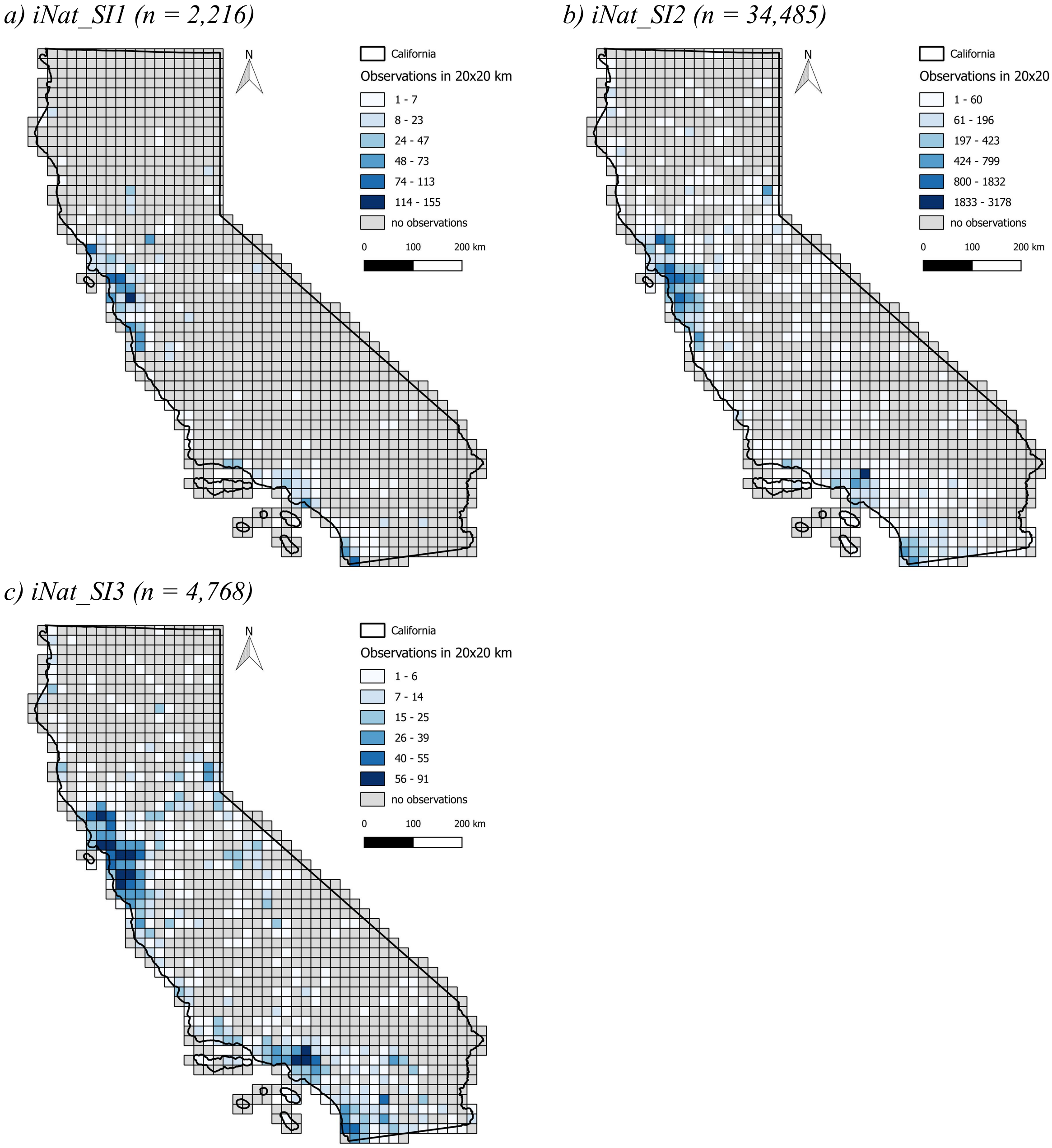
Species group	iNat_A (%)	iNat_SP (%)	iNat_SI1 (%)	Nat_SI2 (%)	iNat_SI3 (%)
plants	22.7	14.0	1.6	22.7	21.9
mammals	2.0	1.9	0.0	2.0	2.1
birds	51.3	31.7	98.4	51.3	48.9
reptiles	0.8	1.1	0.0	0.7	0.6
butterflies and moths	2.7	0.0	0.0	2.7	3.1
hymenopterans	0.1	0.0	0.0	0.1	0.1
beetles	1.3	0.0	0.0	1.3	1.5
dragonflies and damselflies	0.8	0.0	0.0	0.8	0.8
crustaceans	1.3	0.0	0.0	1.3	1.6
mollusks	10.7	43.4	0.0	10.7	12.5
other species	6.1	7.2	0.0	6.1	6.7
spiders	0.2	0.0	0.0	0.2	0.3


a) iNat_A (n = 34,821)

b) iNat_SP (n = 2,415)

Figure 4.1.5: Spatial distribution of valid candidate observations in sets of research grade and synthetic plausible iNaturalist candidates, observed communities approach. (No. of points in 20x20 km raster. Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016.)

Figure 4.1.5 and Figure 4.1.6 show that the spatial distribution remains similar in all sets of valid candidate observations to the distribution in the original dataset. It can also be seen that valid research grade candidate observations in this experiment (set iNat_A) concentrate stronger in the San Francisco Bay area, than do the original research grade candidate observations from 2016. iNat_SP shows stronger concentration of observations to the high-observation-density areas, while iNat_SI3 observations are more dispersed than in the other sets. Both effects are caused by the method of production of these two synthetic sets (see section 3.3.2). In iNat_SP, candidates are located close to high-plausibility observations from iNat_A, which are predominantly found in high-observation-density regions. On the contrary, placing iNat_SI3 candidates away from known observations of their target species pushes them away from existing clusters of observations. However, they still concentrate in regions with relatively high observation densities, because valid candidate cases need a sufficient number of context observations to produce valid candidate contexts with 10 or more species.



a) iNat_SI1 (n = 2,216)

b) iNat_SI2 (n = 34,485)

c) iNat_SI3 (n = 4,768)

Figure 4.1.6: Spatial distribution of valid candidate observations in sets of synthetic implausible iNaturalist candidates, observed communities approach. (No. of points in 20x20 km raster. Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016.)

4.2 Evaluation of the OSM Environments Approach

Evaluation of the OSM environments approach followed as closely as possible the procedure chosen for evaluation of the observed communities approach, whose results were described in detail in section 4.1. This was done to attain a maximum of comparability between the results obtained here and those obtained with the observed communities approach. Also, the basic premises for the OSM environments approach to work as a plausibility estimator for casual citizen science observations of organisms are analogous: it is a promising approach in this sense if and when plausible observations show a higher similarity of OSM tags mapped around them to the proper species' OSM environment than do implausible observations.

The same sets of plausible and implausible observations for both data use cases were used (for an overview, see Table 3.3.1). Although mostly the same parameter settings (frequency thresholds etc.) were used, the differing nature and structure of the data about geographic context (here: OSM data) led to, in part, different candidate observations passing the analysis process as valid candidates. E.g., only 35.9% of valid AF_A candidate cases in evaluation of the observed communities approach (which had 22,426 valid AF_A cases) are also valid AF_A candidate cases in OSM environments approach evaluation. The latter produces, from the same stock of accepted ArtenFinder observations of the year 2016, just 15,329 valid AF_A candidate cases, with 8,046 of them matching the former evaluation result. Results of OSM environments approach evaluation are presented in the same way as before, using the same types of charts. Results are tested with the same statistical methods, because the relevant conditions (especially concerning the statistical properties of distributions of similarity values) persist. Evaluation was thus conducted along the lines described in sections 3.2 and 3.3 with ArtenFinder and iNaturalist data, with the same similarity indices as in the observed communities approach, and with mostly analogous parameter settings:

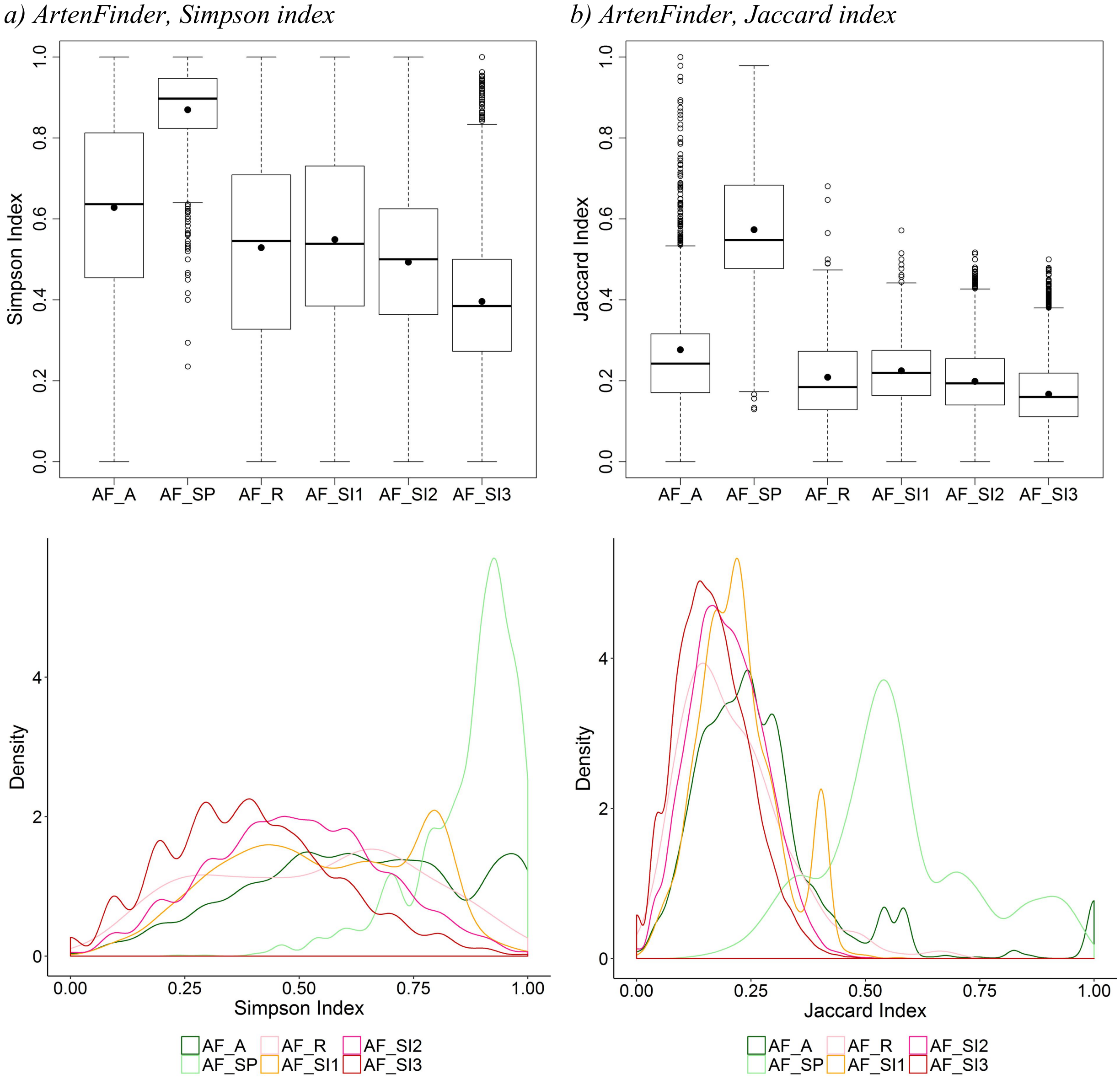
- Extraction of OSM environments:
    - Uniform search radius of 1,000 m for relevant neighborhood for all species groups,
    - target species: only species with 10 or more approved observations up to 2015 (a lower threshold is here employed than in evaluation of the observed communities approach, because it was shown in sensitivity analysis that this can be done without changing similarity distribution results to a large extent, but allows for including more target species and candidate observations in the analysis, see section 4.3.1),
    - frequency threshold for tags frequently mapped in proximity to the target species: 0.5,
    - frequency threshold for nonspecific tags (tags which are part of many OSM environments at the same time, and therefore removed from the final OSM environments): 0.5.
- Extraction of candidate context:
    - Uniform search radius of 1,000 m for relevant neighborhood for all species groups,
    - tags which are part of 50% or more of OSM environments removed from candidate tag contexts.
- Comparison of similarity values:
    - Candidate cases restricted to cases where the size of both the OSM environment and of the candidate tag context is 10 or larger,
    - candidate observations of nonspecific species (identified with the observed communities approach) not used in analysis.

As before, properties of OSM environments and of sets of valid candidate observations are described as part of evaluation results, because they are important in interpretation of evaluation results.

4.2.1 Evaluation Results, OSM Environments Approach, ArtenFinder Data

Similarity Values

Evaluation of the OSM environments approach yields results which are basically comparable to those obtained with observed communities. Accepted and synthetic plausible candidate observations mostly have higher index values with their OSM environments, than have rejected or synthetic implausible observations. However, the difference between distributions of similarity values of plausible and implausible observations is smaller, because especially the sets of synthetic implausible observations have overall higher similarity values than with the former approach. Figure 4.2.1 shows the distributions of Simpson and Jaccard similarity index values for ArtenFinder data resulting from evaluation. AF_SI2 and AF_SI3 are closer to the other distributions, than they were with the observed communities approach. Statistical tests with the Mann-Whitney-U-Test showed, however, that they are still statistically different from AF_A and AF_SP (see Table 4.2.1 and Table 4.2.2).


a) ArtenFinder, Simpson index

b) ArtenFinder, Jaccard index

Figure 4.2.1: ArtenFinder, distributions of Simpson and Jaccard similarity index values, OSM environments approach. n(AF_A) = 15,329; n(AF_SP) = 1,568; n(AF_R) = 215; n(AF_SI1 = 2,104; n(AF_SI2) = 14,964; n(AF_SI3) = 15,618.

Table 4.2.1: ArtenFinder, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with AF_A, AF_SP and the four different sets of rejected or synthetic implausible candidate observations, Simpson index, OSM environments approach.

Simpson Index, AF_A vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	0.8047	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$9.577*10^{-9}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Simpson Index, AF_SP vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Table 4.2.2: ArtenFinder, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with AF_A, AF_SP and the four different sets of rejected or synthetic implausible candidate observations, Jaccard index, OSM environments approach.

Jaccard Index, AF_A vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	0.05772	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$9.886*10^{-11}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Jaccard Index, AF_SP vs.	**AF_R**	**AF_SI1**	**AF_SI2**	**AF_SI3**
Fligner-Killeen-Test	$2.505*10^{-5}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Properties of OSM Environments

ArtenFinder observations up to 2015 are of 2,957 different species. Of these, 1061 species have 10 or more accepted observations. Extracting the OSM tags which surround these observations, reducing them to frequently associated tags and filtering out the nonspecific tags produces 415 OSM environments with 10 or more tags. There are 25 nonspecific tags which are associated to 50% or more of the target species (see Table 5.2.6, in the discussion chapter). 39 nonspecific species (resulting from the observed communities sensitivity analysis with analogous parameter settings, see section 4.3.1) were not used for evaluation purposes. 26 of these species have OSM environments with less than 10 tags anyway, so that 402 valid OSM environments remain for use in evaluation. Sizes of these OSM environments range from 10 to 105 tags (mean: 19.5). See Table 4.2.3 for an overview of these numbers for ArtenFinder data. Using a much lower number of min. 10 target species for OSM environment extraction (vs. min. 100 observations necessary for observed communities extraction in the evaluation of that approach) led to a higher number of valid OSM environments when compared to the number of valid observed communities. Valid OSM environments belong to 18 species groups. Most belong to plants (37.8%), followed by birds (16.7%) and the group of butterflies and moths (14.9%).

Table 4.2.3: ArtenFinder, key numbers describing valid OSM environments.

No. of valid OSM environments	**Mean no. of tags in OSM environments**	**No. of nonspecific tags**	**No. of nonspecific species**
402	19.5	25	39

Properties of Sets of Valid Candidate Observations

Table 4.2.4 lists some key parameters describing the sets of valid candidate observations resulting from evaluation with ArtenFinder and OSM data. Sets of valid candidates are different in size. Mean numbers of tags in candidate contexts and mean numbers of tags in OSM environments associated to these cases also differ between sets. AF_SI3 deviates strongly in mean size of its candidate contexts, and in mean no. of tags (incl. nonspecific tags) around candidates, which are both much lower due to the method of creation of this set (see section 3.3.2). Compared to AF_A, AF_R has smaller and

AF_SP larger OSM environments associated to its candidates. Numbers of valid candidate cases are lower in all sets, than they were in observed communities approach evaluation. This is the case although the number of valid OSM environments (and thus the number of evaluated target species) is higher (see above). However, less candidate cases were evaluated as valid because many are situated in locations with too few tags (< 10) within the relevant neighborhood.

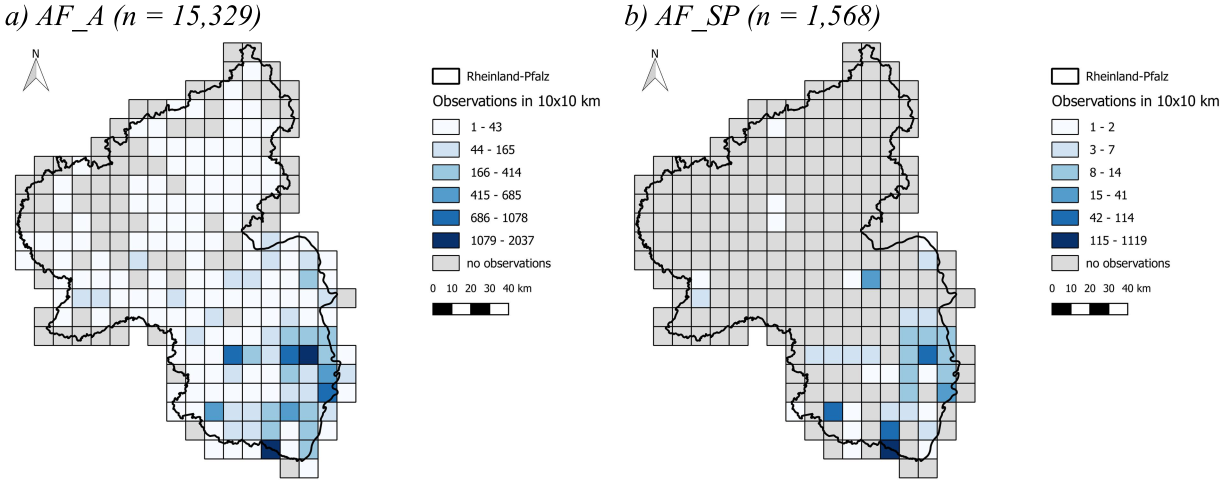
Table 4.2.4: ArtenFinder, key numbers describing sets of valid candidate observations, OSM environments approach.

Set of candidates	No. of valid candidate cases	Mean no. of tags in candidate contexts	Mean no. of tags in OSM environments, per set of candidates	Mean no. of context tags (incl. nonspecific tags)
AF_A	15,329	35.6	19.1	56.3
AF_SP	1,568	39.8	31.6	61.9
AF_R	215	32.8	16.2	52.9
AF_SI1	2,104	33.2	19.8	53.6
AF_SI2	14,964	34.0	19.1	54.3
AF_SI3	15,618	23.2	19.1	41.0

Table 4.2.5: ArtenFinder, portions of species groups in sets of candidate observations used in evaluation, OSM environments approach.

Species group	AF_A (%)	AF_SP (%)	AF_R (%)	AF_SI1 (%)	AF_SI2 (%)	AF_SI3 (%)
plants	5.7	1.1	2.4	0.0	5.6	6.0
fungi	1.4	0.6	0.0	0.0	1.3	1.4
mammals	1.9	2.0	0.5	0.0	1.9	2.0
birds	65.9	78.9	12.1	90.4	66.3	65.3
reptiles	3.3	1.0	2.8	3.9	3.3	3.3
amphibians	0.4	0.2	0.9	0.0	0.4	0.4
butterflies and moths	6.6	6.0	28.8	1.3	6.5	6.7
hymenopterans	1.8	0.8	2.3	0.0	1.7	1.8
beetles	0.7	0.4	0.9	0.0	0.7	0.7
dragonflies and damselflies	9.7	8.1	43.3	0.7	9.6	9.8
mantids	0.4	0.1	0.0	0.0	0.4	0.4
locusts	1.2	0.3	3.7	0.0	1.2	1.1
crustaceans	0.0	0.0	0.0	0.0	0.0	0.1
mollusks	0.5	0.1	0.5	3.8	0.5	0.5
true bugs	0.4	0.4	0.0	0.0	0.4	0.4
neuropterans	0.0	0.0	1.9	0.0	0.0	0.0

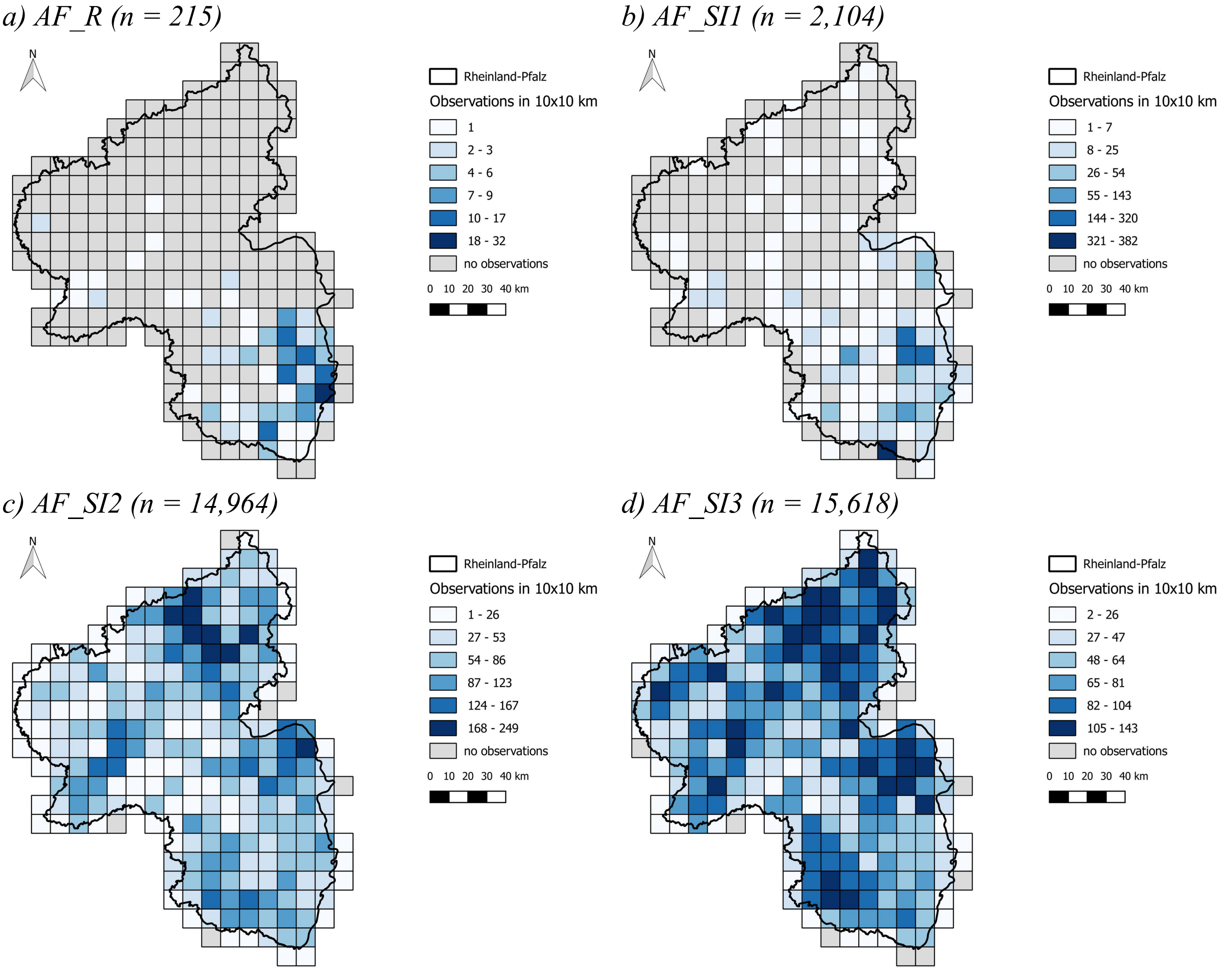
Species group compositions of the sets of valid candidate observations are shown in Table 4.2.5. Species group compositions of AF_SI2 and AF_SI3 are, expectedly, almost identical to AF_A, (AF_SI2 and AF_SI3 were created based on AF_A observations, see section 3.3.2). AF_SI1 and AF_SP have higher portions of birds. These findings go along with effects found also in observed communities approach evaluation. However, dominance of bird observations in most sets is now much stronger than in observed communities approach evaluation. The slightly higher portion of plants in candidates is an effect introduced by using a lower threshold of 10 or more (instead of 100 or more) observations of a target species available in approved observations up to 2015 for OSM environment generation. This also causes the number of species groups to be higher than in observed communities approach evaluation (16 vs. 13).


a) AF_A (n = 15,329)

b) AF_SP (n = 1,568)

Figure 4.2.2: Spatial distribution of valid candidate observations in sets of accepted and synthetic plausible ArtenFinder candidates, OSM environments approach. (No. of points in 10x10 km raster. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)

In the sets AF_A, AF_R and AF_SI1, the general northwest to southeast trend of growing observation density was not changed by the selection of valid candidate observations in evaluation, because these depend strongly on the overall distribution of ArtenFinder observation density in Rheinland-Pfalz, see Figure 4.2.2 and Figure 4.2.3. This is also true for the set AF_SP, which is based on synthetic, spatially random observations. AF_SP observations are located close to plausible AF_A observations which have their highest densities in southeastern Rheinland-Pfalz. However, the sets AF_SI2 and AF_SI3, which are also based on synthetic random points, almost invert the northwest to southeast trend of observation density, because they represent observations away from known occurrences of the species they represent. Even AF_SI2, which has synthetic observations in situations with similar OSM tag density as AF_A observations, shows this effect, because OSM tag density does not have a northwest to southeast trend. This finding confirms that OSM data in Rheinland-Pfalz are indeed capable of providing sufficient numbers of context tags for plausibility estimation in regions where observation density is insufficient for this purpose. With the observed communities approach, the northern and northwestern parts of Rheinland-Pfalz were practically devoid of valid candidate observations. This was not only the case for real observations (scarce in these regions anyway), but also for synthetic candidate observations based on random points, which did not find adequate context observation densities to provide valid numbers of context species in these regions.



a) AF_R (n = 215)

b) AF_SI1 (n = 2,104)

c) AF_SI2 (n = 14,964)

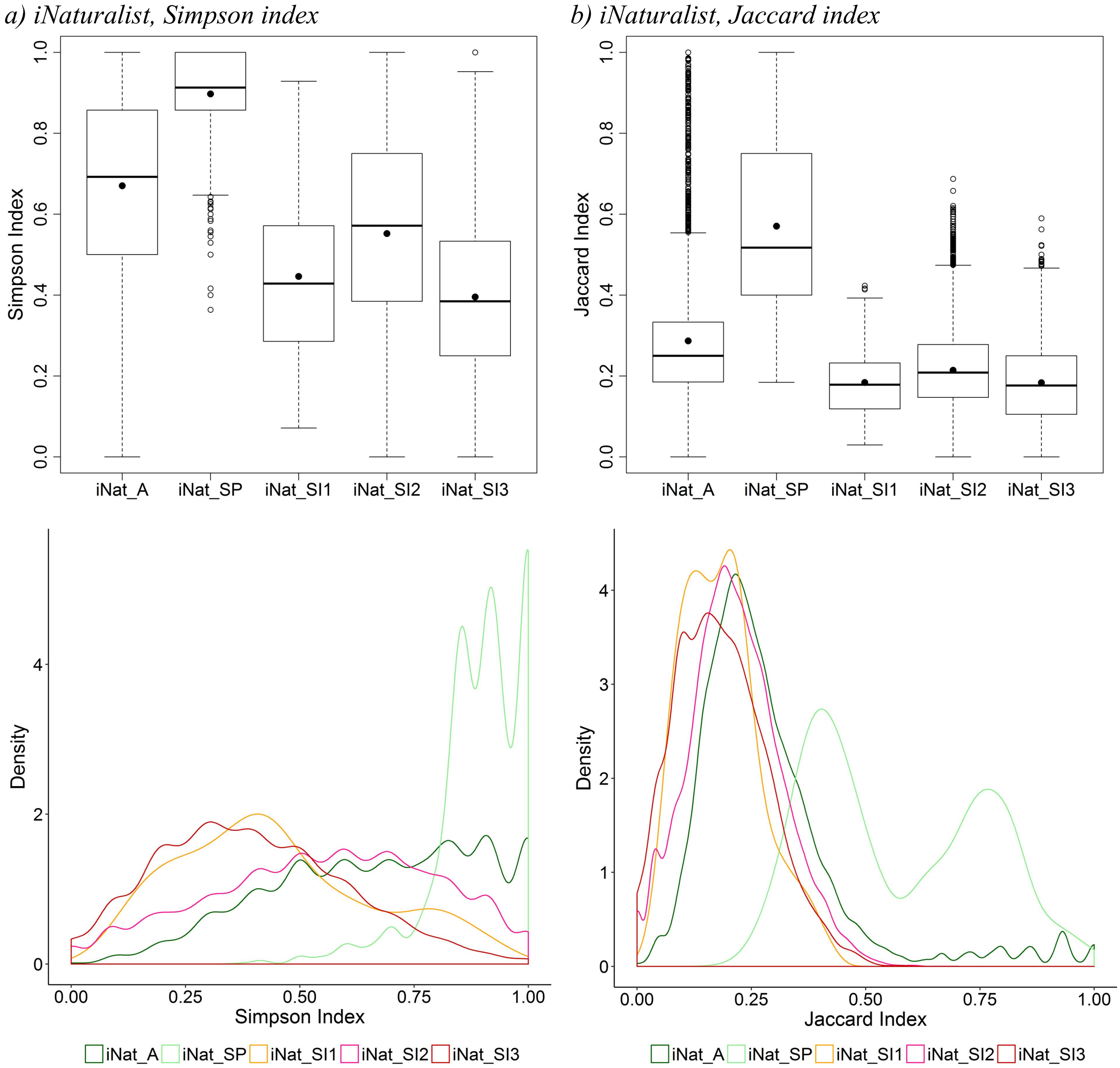
d) AF_SI3 (n = 15,618)

Figure 4.2.3: Spatial distribution of valid candidate observations in sets of rejected and synthetic implausible ArtenFinder candidates, OSM environments approach. (No. of points in 10x10 km raster. Classified by Natural Breaks. Source of Rheinland-Pfalz state line: LANIS Rheinland-Pfalz.)

4.2.2 Evaluation Results, OSM Environments Approach, iNaturalist Data

Similarity Values

Above, we found evaluation results of the OSM environments approach with ArtenFinder data to be comparable to results obtained with the observed communities approach. The same commonalities, but also differences, are evident in evaluation results with iNaturalist data. Figure 4.2.4 shows the distributions of Simpson similarity index values for iNaturalist data resulting from OSM environments approach evaluation. iNaturalist research grade candidate observations (iNat_A) and synthetic plausible candidates (iNat_SP) usually have higher Simpson index values with their observed communities, than have synthetic implausible observations. Again, distributions of similarity values of _SI2 and _SI3 sets are higher than they were with the observed communities approach, and therefore closer to the other distributions, while remaining statistically different from iNat_A and iNat_SP (see Table 4.2.6. and Table 4.2.7).


a) iNaturalist, Simpson index

b) iNaturalist, Jaccard index

Figure 4.2.4: iNaturalist, distributions of Simpson and Jaccard similarity index values, OSM environments approach. n(iNat_A) = 35,058; n(iNat_SP) = 2,131; n(iNat_SI1) = 290; n(iNat_SI2) = 34,654; n(iNat_SI3) = 4,542.

Table 4.2.6: iNaturalist, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with iNat_A, iNat_SP and the three different sets of synthetic implausible candidate observations, Simpson index, OSM environments approach.

Simpson Index, iNat_A vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	0.004178	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Simpson Index, iNat_SP vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Table 4.2.7: iNaturalist, results (p-Values) of Fligner-Killeen-Tests and of Mann-Whitney-U-Tests with iNat_A, iNat_SP and the three different sets of synthetic implausible candidate observations, Jaccard index, OSM environments approach.

Jaccard Index, iNat_A vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$7.115*10^{-7}$	$< 2.2*10^{-16}$	$1.19*10^{-14}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Jaccard Index, iNat_SP vs.	**iNat_SI1**	**iNat_SI2**	**iNat_SI3**
Fligner-Killeen-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$
Mann-Whitney-U-Test	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$	$< 2.2*10^{-16}$

Properties of OSM Environments

There are 2,641 (of 8,125) species with 10 or more research grade observations in the California iNaturalist data up to 2015. Filtering for frequently associated tags (with a threshold of 0.5) produces 1,476 OSM environments with 10 or more tags. A threshold of 0.5 for finding and eliminating nonspecific tags results in nine nonspecific tags in the OSM environments for this dataset. This means that nine tags are present in 50% or more of OSM environments. After removing these from the OSM environments, 635 valid OSM environments remain for evaluation purposes, their size ranging from 10 to 89 tags, with a mean size of 20.1 tags (Table 4.2.8). Again, the number of valid OSM environments is higher than the number of valid observed communities in the evaluation of that approach, because a lower threshold of min. 10 target species observations was employed in OSM environment extraction. Valid OSM environments are of 21 species groups, plants dominating at 26.5%, followed by mollusks (20.8%) and birds (16.1%).

Table 4.2.8: iNaturalist, key numbers describing valid OSM environments.

No. of valid OSM environments	**Mean no. of tags in OSM environments**	**No. of nonspecific tags**	**No. of nonspecific species**
635	20.1	9	0

Properties of Sets of Valid Candidate Observations

Some key parameters describing the sets of valid candidate observations resulting from evaluation of the OSM environments approach with iNaturalist data are listed in Table 4.2.9. iNat_SP has the largest OSM contexts associated to its candidates, on average. iNat_SI3 cases are situated in locations with lower tag densities than the other sets. Numbers of valid candidate cases are partly comparable to numbers obtained in observed communities approach evaluation. However, considering the higher number of target species evaluated (see above), they are relatively low.

Table 4.2.9: iNaturalist, key numbers describing sets of valid candidate observations, OSM environments approach.

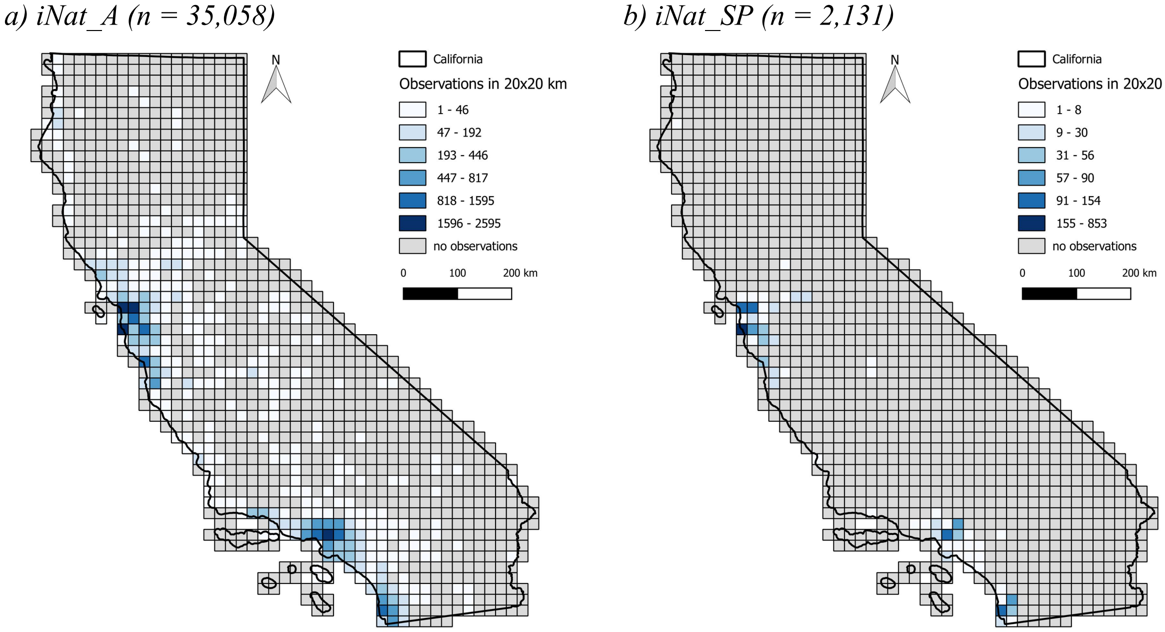
Set of candidates	No. of valid candidate cases	Mean no. of tags in candidate contexts	Mean no. of tags in OSM environments, per set of candidates	Mean no. of context tags (incl. nonspecific tags)
iNat_A	35,058	33.6	16.2	40.8
iNat_SP	2,131	29.9	19.6	36.8
iNat_SI1	290	28.9	16.8	36.2
iNat_SI2	34,654	32.5	16.2	39.2
iNat_SI3	4,542	18.5	16.1	23.5

Table 4.2.10: iNaturalist, portions of species groups in sets of valid candidate observations used in OSM environments evaluation, OSM environments approach.

Species group	iNat_A (%)	iNat_SP (%)	iNat_SI1 (%)	Nat_SI2 (%)	iNat_SI3 (%)
plants	11.6	12.6	7.6	11.7	12.2
fungi	3.6	2.7	0.0	3.6	3.5
mammals	3.0	7.0	0.0	3.0	2.8
birds	40.7	11.2	92.4	40.7	40.2
reptiles	1.3	4.3	0.0	1.3	0.9
amphibians	0.1	0.0	0.0	0.1	0.2
modern bony fishes	0.5	0.5	0.0	0.5	0.6
butterflies and moths	6.7	5.0	0.0	6.7	6.3
hymenopterans	4.0	1.2	0.0	4.0	3.8
beetles	2.6	2.3	0.0	2.7	2.7
dragonflies and damselflies	0.1	0.0	0.0	0.2	0.2
earwigs	0.5	0.0	0.0	0.5	0.3
mantids	0.1	0.0	0.0	0.1	0.1
cockroaches	0.2	0.0	0.0	0.2	0.2
locusts	0.7	0.5	0.0	0.7	0.8
crustaceans	1.8	1.5	0.0	1.8	1.6
mollusks	16.9	42.0	0.0	16.9	17.8
other species	2.4	5.9	0.0	2.4	2.6
true bugs	0.9	0.4	0.0	0.9	1.3
flies	0.3	0.4	0.0	0.4	0.3
spiders	1.7	2.6	0.0	1.7	1.8

Table 4.2.10 summarizes the species group compositions of the sets of valid candidate observations. iNat_A, iNat_SI2 and iNat_SI3 have very similar species group composition, which is due to the method used for creating iNat_SI2 and iNat_SI3. All sets except iNat_SP are clearly dominated by bird observations. Mollusks rank second in iNat_A, iNat_SI2 and iNat_SI3, and dominate iNat_SP. Plants are also prominent, mostly with portions above 10%. Set iNat_SI1, containing only species selected for their special properties (physically similar species living in different habitats) contains mostly birds as well as some plant species.

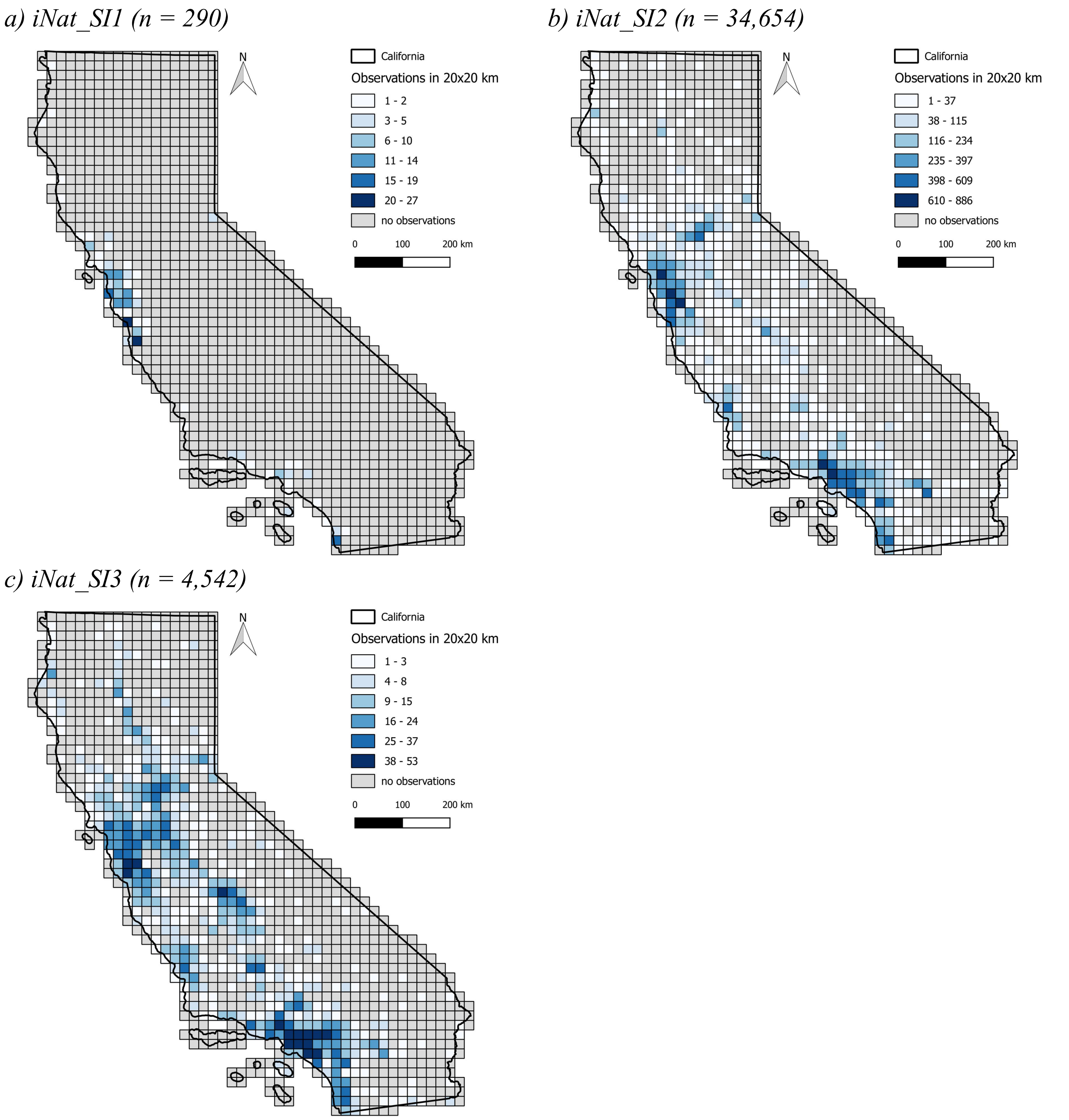
Figure 4.2.5 and Figure 4.2.6 show the spatial distribution in all sets of valid candidate observations. iNat_SP shows stronger concentration of observations to the high-observation-density areas, while iNat_SI2 and iNat_SI3 observations are more dispersed than in the other sets. iNat_SI2 and _SI3 also show higher densities in regions which had only few valid candidates in observed communities approach evaluation. However, this effect is not as pronounced as in the ArtenFinder data use case, because the spatial properties of California OSM data are more similar to iNaturalist observation distribution, than Rheinland-Pfalz OSM data are to ArtenFinder observations. Still, with OSM data as context source, iNat_SI2 and iNat_SI3 valid candidate observations can be found in considerable densities also in greater distances to the Los Angeles and San Francisco bay areas, as well as in the central valley and other places which did not produce many valid candidate cases with observed communities. Large regions in southern, eastern, and northern California remain, however, low-density regions due to a lack of adequate OSM data.



a) iNat_A (n = 35,058)

b) iNat_SP (n = 2,131)

Figure 4.2.5: Spatial distribution of valid candidate observations in sets of research grade and synthetic plausible iNaturalist candidates, OSM environments approach. (No. of points in 20x20 km raster. Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016.)



a) iNat_SI1 (n = 290)

b) iNat_SI2 (n = 34,654)

c) iNat_SI3 (n = 4,542)

Figure 4.2.6: Spatial distribution of valid candidate observations in sets of synthetic implausible iNaturalist candidates, OSM environments approach. (No. of points in 20x20 km raster. Classified by Natural Breaks. Source of state line: U.S. Geological Survey 2016.)

4.3 Results of Sensitivity Analysis

When input parameters are changed or methods are modified, evaluation results also change. Section 3.4 presented a series of such changes to input parameters and of methodological modifications which were applied mainly to the observed communities approach, but included also one specific modification to the OSM environments approach. Below, effects of each parameter change or methodological modification are described. They reveal complex interactions between output parameters such as observed community size or number of nonspecific species. For instance, an input parameter change may lead to larger observed communities, which in turn produces higher numbers of valid observed communities, because fewer of them fall below the minimum requirement of size. At the same time, larger observed communities may also yield more nonspecific species, so that more observed communities are discarded as belonging to a nonspecific species. Some factors affect the number of valid candidate observations, and thus also the methods' effectiveness. For instance, while more observed communities allow for more candidates to be evaluated, more nonspecific species also disqualify more candidate observations at the same time. Ultimately, all such effects combine in a complex way to cause changes in the numbers of valid observed communities, in numbers of candidate observations evaluated, and in distributions of similarity values of the candidate observation sets. However, the principle behavior of the observed communities approach was found to remain basically the same in all cases. Stronger effects are mostly restricted to numbers of target species and candidate observations which are actually evaluated, while distributions of similarity values do show upward or downward shifts, but mostly keep their principle relations between each other.

4.3.1 Results of Using a Lower Minimum Requirement for Target Species Observations

Here, the effects of a substantially lower threshold of min. 10 target species observations in observed community extraction were examined. Detailed numeric and graphic results can be found in the appendix, section 7.3.1. While the result with this input parameter modification has all valid target species which were part of the evaluation of the approach laid out above, the modification added a large number of target species with less than 100 observations up to 2015, for which valid observed communities were now available. In ArtenFinder data, the number of target species with valid observed communities more than quadrupled to 792 species. iNaturalist now had even more than five times the number of valid observed communities, at 1258. This also led to a substantial raise in valid candidate observation numbers, that is, in the numbers of candidate cases which could be actually evaluated with the approach. ArtenFinder sets now had, on average, 67% more valid candidate cases, and 50.1% of all AF_A candidates were actually valid cases (was 32.7%). iNaturalist sets grew by 49% on average, and 30.0% of iNat_A candidate cases were now actually used (was 20.8%). However, this also induced changes to other properties of the resulting observed communities and candidate sets for both data use cases. Interestingly, a much larger portion of 30.4% of ArtenFinder observed communities now belonged to plants, while birds, which where dominant before, now made up only 16.4% of them. However, this had only a mild effect on the species composition of valid ArtenFinder candidate sets. The same was true for iNaturalist data, where plants species now contributed 42.8% of valid observed communities, leading also to a raise but not dominance of plants in valid candidate cases.

Distributions of Simpson index values for ArtenFinder data differed only slightly from those obtained with the more restrictive requirement of target species observation numbers for observed communities extraction. With Jaccard similarity values, medians of all sets of implausible observations slightly rose, while the median for accepted observations remained constant. The AF_R distribution, close to AF_A already in evaluation, came close to being no longer significantly different from Jaccard index values for AF_A. iNaturalist index distributions also showed only very slight changes, and differences between them remained significant.

4.3.2 Results of Using Variable Search Radii

This sensitivity analysis introduced the use of different search radii for different species groups. Contrast between distributions of Simpson similarity values of plausible and implausible observations was not increased by this parameter change for most ArtenFinder sets. Detailed numeric and graphic results can be found in the appendix, section 7.3.2. While the medians of AF_A and AF_SP rose slightly, the medians of the synthetic sets of implausible observations also rose. Only the similarity values of the set AF_R dropped. The changes in distributions of Jaccard index values showed similar behavior. Differences remained, however, significant. In the iNaturalist case, relations of the distributions of Simpson index values, considering their medians, showed slightly smaller differences than before: while iNat_A and iNat_SP dropped slightly, all sets of implausible candidates went up. For distributions of Jaccard index values, one unusual effect was that the iNat_SI1 median was now even placed above the median of iNat_A.

Results with different search radii for different species groups once more revealed the complex mechanics taking effect when observed community approach parameters are changed. While numbers of valid observed communities dropped by 69.9% to 128 for ArtenFinder data, this number rose by 37.6% to 322 for iNaturalist data. With ArtenFinder data, this difference was caused by a strong increase in nonspecific species from 46 to 118, which over-compensated for the rise in valid candidate cases through larger search radii. iNaturalist now had five nonspecific species, but the positive effect of larger search radii on valid candidate case numbers was far stronger. Mean sizes of observed communities and candidate contexts increased, as was to be expected with larger search radii.

4.3.3 Results of Shifting Frequency Thresholds for Frequent Co-Observations and Nonspecific Species

Threshold for frequent co-observations

This analysis examined effects of higher or lower frequency thresholds for frequent co-observations in observed communities. See appendix, section 7.3.3.1 for graphical and numerical details. In the ArtenFinder use case, a shift towards a higher threshold value for relevant co-observations led to higher Simpson index values in all sets, while a lower threshold reduced index values. The Jaccard index's behavior for a lower threshold was more mixed, while a higher threshold reduced Jaccard index values. Differences between distributions of index values of sets of plausible and implausible observations were slightly smaller with both indices and a higher threshold, and partly larger for a reduced threshold. They remained significant. Similarity index distributions of iNaturalist data showed trends which are in part similar to results obtained with ArtenFinder data, and in part different. Raising the threshold to 0.75 led to higher similarities in sets of plausible, but lower ones in sets of implausible observations. With the lower threshold value of 0.25, which had varied effects with iNaturalist data, the distribution of iNat_SI1 Jaccard index values was no longer significantly different from iNat_A.

Through the complex interplay of the threshold values which were examined here with numbers of valid observed communities and numbers of nonspecific species, the ArtenFinder data use case had lower numbers of valid observed communities and of valid candidate cases (with smaller candidate contexts) for a lower threshold value of 0.25. Raising the threshold to 0.75 also led to fewer valid observed communities, but higher numbers of valid candidate observations. iNaturalist data reacted to the contrary, the lower threshold resulting in a higher effectiveness in terms of valid observed communities and candidate cases, and the higher threshold (in the absence of nonspecific species) strongly reducing both numbers.

Threshold for identifying nonspecific species

How did choosing a lower or higher threshold for nonspecific species affect the results? To appraise these effects, experiments were conducted which used either a threshold of 0.25 or of 0.75 for identifying nonspecific species in the process of observed communities creation. See appendix, section 7.3.3.2, for detailed results. Results with the higher value (0.75) for iNaturalist data were identical to evaluation results obtained before, because the number of nonspecific species remained 0. Lowering the threshold led to higher numbers of nonspecific species in both data use cases, and consequently to lower numbers of valid observed communities, and of valid candidate observations. With ArtenFinder data, the lower threshold value resulted in a raise of 61% to 78 nonspecific species, and the iNaturalist case now produced 8 nonspecific species. The higher threshold value reduced ArtenFinder nonspecific species to 28, a drop of 38%.

Despite the notable changes in numbers of nonspecific species, changing the threshold for identifying nonspecific species to 0.25 or to 0.75 did not change the behavior of the observed communities approach in terms of distributions of similarity values in a critical way, although differences between distributions of similarity values of plausible and implausible sets were somewhat larger for the lower threshold value for both data use cases, and slightly smaller for ArtenFinder data with the higher threshold value (especially with the Simpson index). iNaturalist results showed only very small changes with the threshold variations applied here.

4.3.4 Results of Using Auxiliary Land Cover and Ecological Land Unit Information

Detailed numeric and graphic results for this change in methodology are presented in the appendix, section 7.3.4. When CORINE land cover or ELU polygon geometries are used with ArtenFinder data to restrict relevant search areas, differences between AF_SP and AF_SI2 / AF_SI3 increase. iNaturalist results with NLCD and ELU data also mostly exhibit larger differences between iNat_SP and iNat_SI1, _SI2 and _SI3 sets, except for the ELU / Jaccard index case.

Using auxiliary land cover or ELU information did not much change the overall effectiveness of the approach in the ArtenFinder use case, although effects are different for different sets of candidate observations. Numbers of nonspecific species were markedly lower, compensating for a larger number of candidate cases with too small species contexts. The iNaturalist data use case, however, experienced a strong reduction in numbers of valid observed communities and valid candidate cases, because of the absence of nonspecific species. In the experiment with ELU data, this is also partly due to the fact that ELUs only have terrestrial units, so that all marine organisms dropped out in this analysis.

4.3.5 Results of Using a Quantitative Similarity Index

The Similarity Ratio is closely related to the Jaccard index (see section 3.1), and similarity distributions obtained with it were therefore compared to evaluation results with the Jaccard index. For detailed results in tables and graphs, see appendix, section 7.3.5. As usual, all other parameter settings of the method were kept constant with the values used in evaluation. The only change here consists in the use of a different similarity coefficient. All results which concern properties of observed communities or candidate observation sets, such as their numbers and sizes, were therefore identical (or, in the case of _SP sets, mostly close) to the results obtained in evaluation, and do not need to be reiterated.

Weighting species in observed communities and in candidate contexts by observation frequency increased differences between AF_A/AF_SP and all ArtenFinder sets of implausible observations, when compared to results obtained with the binary Jaccard index, which only uses information about the presence or absence of species in observed communities and candidate contexts. Distance weighting, on the contrary, rather decreased differences between similarity value distributions of plausible and implausible observation sets, mostly because AF_SP similarities decreased. In iNaturalist results with frequency weighting of species, all distributions of similarity shifted towards lower values. With weighting of species by distance of observations, iNat_A remained almost constant, while distributions of iNat_SP, iNat_SI1, iNat_SI2 and iNat_SI3 shifted towards lower values, reducing the differences in the process.

4.3.6 Edge Effects

In an appropriate analysis, the impact of edge effects in evaluation results was tested using a guard zone approach (see section 3.4.6 for methodological details). 8.2% of accepted ArtenFinder Observations up to 2015, and 12.2% of accepted ArtenFinder observations from 2016 were situated in the guard zone. In both data subsets, these were mostly found in the southeastern part of the area of interest, where observation density is highest. In iNaturalist data, these portions were much lower: only 0.4% of research grade observations up to 2015, and 0.2% of research grade observations from 2016 were situated within the one-km zone at the edge of the area of interest, showing that observation density is much lower here along the edge of the area of interest, than in Rheinland-Pfalz data. This led to lower numbers of valid observed communities and candidate observations, especially in the ArtenFinder case. The appendix, section 7.3.6, gives details of these changes. However, results with edge effect correction were very similar, in all respects, to those obtained without edge effect correction.

4.3.7 Results with a Variant of Simpson Index

In evaluation, the observed community was smaller than the candidate context in 71.5% of all candidate cases in ArtenFinder sets, and in 79.7% of all candidate cases in iNaturalist sets. Therefore, the Simpson index was, in the majority of cases, equal to the variant of the Simpson index suggested in section 3.4.7, and the behavior of this variant was therefore very close to the Simpson index in all respects, because most Simpson index values were calculated with the observed community size as denominator anyway. Detailed results can be found in the appendix, section 7.3.7.

4.3.8 Results of Using Date-Specific OSM Context

Results with extraction of OSM context timed to the observation date (see appendix, section 7.3.8 for detailed results) show that earlier OSM stages provide fewer tags for analysis. This was to be expected, as examination of the OSM data used in this work showed a steady growth of tag numbers over time for most keys (see section 2.1.3). Therefore, when using earlier stages of OSM which correspond with observation dates, OSM environments and candidate contexts become smaller, on average. This effect also reduced the numbers of valid OSM environments and of valid candidate cases: fewer species, and fewer candidates, can be actually evaluated. The ArtenFinder use case lost 32% of its valid OSM environments and 31% of its valid AF_A candidate cases. In iNaturalist data, this effect was less pronounced, with just 5% of valid OSM environments and 18% of valid iNat_A candidates lost.

There were no substantial shifts in distributions of similarity values which would argue for a positive effect on results in terms of contrast between sets of predominantly plausible or implausible candidate observations. AF_A and iNat_A Simpson index value distributions shifted upwards, but so did _SI2 and _SI3 distributions. Jaccard index AF_SP and iNat_SP distributions shifted slightly downwards, while _SI2 and _SI3 distributions remained more or less constant.

5 Discussion

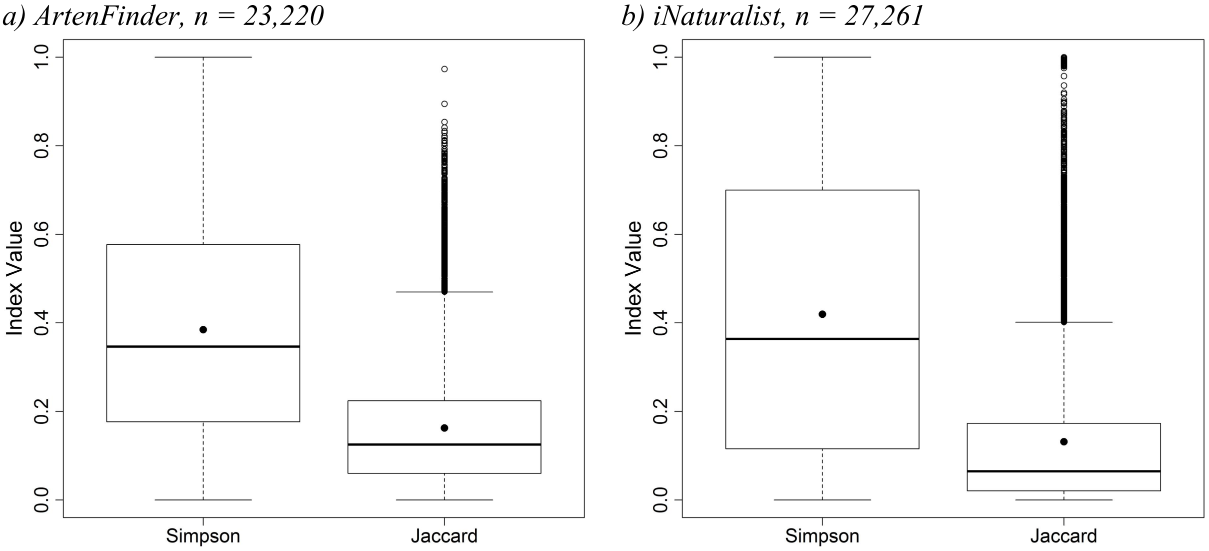
5.1 Observed Communities Approach

In both data use cases, both similarity coefficients produce distributions of values for plausible observations which are statistically different from the distributions for implausible observations. In all cases, the candidate contexts of plausible observations show significantly higher similarities to their observed communities, than do those of implausible candidate observations. Evaluation therefore was able to show that the observed communities approach is able to distinguish between sets of plausible and implausible observations.

Observed communities are the basis of the approach which was evaluated here. Discussion of evaluation results therefore starts with some important properties they show, and then proceeds to differences between the sets of candidates tested in evaluation, and the causes for these differences. Influence of the spatial properties of the data, especially of variable observation density, on similarity values is also discussed in detail, as well as the relations between Simpson and Jaccard index values and their correlation. Finally, effects of parameter changes and methodological modifications on evaluation results are discussed.

5.1.1 Similarity of Observed Communities

The observed communities approach can work only if different species have, in principle, observed communities which are different from one another. The similarity indices used for estimating plausibility of candidate observations can of course also be used to examine differences between observed communities, by calculating the Simpson and Jaccard index values for each observed community with all others. Figure 5.1.1 shows that Simpson und Jaccard similarity values between observed communities are low, for both data use cases. Medians of 0.36 for ArtenFinder Simpson index and 0.13 for ArtenFinder Jaccard index are close to the values which were found for synthetic implausible observation sets in the evaluation presented above. The same is true, to a lesser degree, for iNaturalist values.


a) ArtenFinder, n = 23,220

b) iNaturalist, n = 27,261

Figure 5.1.1: ArtenFinder and iNaturalist, distributions of Simpson and Jaccard index values, similarities of observed communities with >= 10 species among one another.

This may not come as a surprise, because observed communities come from very different species and species groups. Looking at these results in more detail, it is soon discovered that there are of course also species which have quite similar observed communities. Basically, two observed communities will be similar if the two target species are predominantly observed in places with similar contexts. This is likely if, for instance, the two target species share the same habitat preferences. This effect can be traced when looking at different species groups: groups of organisms predominantly more specialized in a certain type of habitat have, on average, higher similarities among their species' observed communities, than do groups which contain many species with more heterogeneous habitat preferences. Good examples for the former case are dragonflies and damselflies from ArtenFinder (observed often, but of course not always, close to waterbodies), with an average Simpson index value among observed communities of 0.80, and mollusks from iNaturalist (observations coming mostly from the coast) with an average Simpson index value among observed communities of 0.96. The second case of higher heterogeneity is apparent in birds both from ArtenFinder and iNaturalist: average Simpson index values among observed communities are 0.47 and 0.64, respectively.

5.1.2 Species Composition of Observed Communities

A look at the composition of observed communities, concerning the species they consist of, reveals an interesting fact: observed communities tend to be dominated by species from the species group of their target species. Table 5.1.1 and Table 5.1.2 demonstrate this for both data use cases and the most important species groups in either use case. A bird's observed community is disproportionately dominated by birds. ArtenFinder observed communities of butterflies and moths have a much larger rate of this species group than would be expected from the composition of observations they were extracted from. The same effect can be found in ArtenFinder dragonflies' and locusts' observed communities, and in iNaturalist plants' and mollusks' observed communities. In some cases, observed communities are not dominated by the target species' group, but by species of groups which are observed in close spatial proximity due to their similar ecological niches. This applies to ArtenFinder's amphibians' observed communities, which are dominated by dragonflies and damselflies also found often in or close to water bodies (but portions of amphibians themselves are also disproportionately high in these observed communities). Another case is iNaturalist's "other species" observed communities, which are of various marine organisms, and dominated by (mostly marine) mollusks (again, with "other species" also disproportionately high). ArtenFinder plants' observed communities are dominated by butterflies and moths; in this case, there is obviously a spatial coincidence of these groups in the data collection process.

Apart from this, all observed communities have relatively high mean portions of species groups which are dominant in the observation record of the respective project. E.g., ArtenFinder observed communities mostly have relatively high rates of butterflies and moths, which are mostly weak in those of iNaturalist data, while iNaturalist observed communities almost always feature a number of plants, mostly rare in ArtenFinder observed communities. Both ArtenFinder and iNaturalist observed communities mostly have more than 10% of birds (except the ArtenFinder butterflies and moths observed communities).

Table 5.1.1: ArtenFinder, mean rates of species groups in observed communities.

		Mean rates (%) of species groups in observed communities of					
Species groups	**Rates of species groups in observations up to 2015 (%)**	**plants (n = 8)**	**birds (n = 62)**	**amphibians (n = 8)**	**butterflies and moths (n = 9)**	**dragonflies and damselflies (n = 30)**	**locusts (n = 12)**
plants	6.4	19.2	1.2	1.0	1.5	0.5	0.0
fungi	1.9	0.0	0.0	0.3	0.0	0.2	0.1
mammals	2.4	1.8	4.7	1.9	1.2	2.6	0.7
birds	41.8	19.9	70.4	23.1	7.8	17.4	12.4
reptiles	1.7	1.2	3.0	7.3	4.2	3.9	5.4
amphibians	1.6	0.0	1.4	18.0	0.8	4.0	0.9
butterflies and moths	28.9	46.5	5.4	9.3	61.1	15.2	19.0
hymenopterans	0.7	1.5	0.9	1.2	4.0	1.9	3.4
beetles	0.8	0.0	0.3	0.5	0.1	0.2	0.1
dragonflies and damselflies	9.2	2.8	10.5	30.7	11.7	46.5	12.5
mantids	0.1	0.3	0.1	0.3	0.2	0.0	0.0
locusts	3.4	5.7	1.9	6.3	7.1	7.4	45.4
mollusks	0.3	1.0	0.0	0.2	0.3	0.2	0.1
true bugs	0.1	0.0	0.1	0.0	0.0	0.0	0.0

The cause of these effects can be found in the behavior of volunteers contributing to casual citizen science data collection portals: volunteers tend to specialize in certain species groups. Also, they tend to visit the same places repeatedly or even regularly. Both behaviors were discussed in section 2.2. They have a strong influence on the spatial properties of the data, leading to clustering of observations of certain species groups at certain places. The observed communities approach to plausibility estimation for casual citizen science observations of organisms reveals this effect when examining neighborhoods of a target species and aggregating them into an observed community. Also, general preferences of species groups by the participants in a project govern the overall composition of observed communities. This makes evident the fact that observed communities are different from natural species communities due to the VGI nature of the data collection process.

*Table 5.1.2: iNaturalist, mean rates of species groups in observed communities. *mollusks: mostly marine species. **"other species": various marine organisms.*

	Rates of species groups in observations up to 2015 (%)	**Mean rates (%) of species groups in observed communities of**			
Species groups		**plants (n=69)**	**birds (n=96)**	**mollusks (n=34)***	**"other species" (n=10)****
plants	33.4	66.3	0.7	16.3	11.9
fungi	2.3	0.1	0.0	0.5	0.2
mammals	4.2	1.0	1.1	1.9	2.6
birds	32.0	18.1	96.1	28.0	29.3
reptiles	5.2	5.8	0.5	0.2	0.0
amphibians	1.6	0.8	0.0	0.1	0.0
modern bony fishes	0.4	0.0	0.0	2.2	0.8
butterflies and moths	6.7	2.0	0.2	1.2	0.9
hymenopterans	1.1	2.6	0.2	0.2	0.1
beetles	1.2	0.2	0.0	0.4	0.2
dragonflies and damselflies	1.7	0.4	0.1	0.0	0.0
cockroaches	0.0	0.0	0.0	0.1	0.0
locusts	0.3	0.0	0.0	0.1	0.0
crustaceans	0.8	0.4	0.4	5.3	5.4
mollusks	5.1	1.2	0.2	29.0	32.6
„other species"	2.1	0.4	0.4	14.4	15.9
true bugs	0.6	0.5	0.0	0.0	0.0

Going deeper into the observed communities' species compositions, a pronounced influence of natural associations of species is revealed in some observed communities. It can be found in observed communities of species which are often present in certain types of habitats, and thus also predominantly observed there. For instance, observed communities of most birds which are associated to water bodies contain mostly other species associated to water. Common Kingfisher (*Alcedo atthis*) has an ArtenFinder observed community (Table 5.1.3) holding 14 species (excluding *A. atthis* itself). 50.0% of these are birds, and most of them are associated to water, such as Egyptian Goose (*Alopochen aegyptiaca*, a species recently introduced to Rheinland-Pfalz), or Great Cormorant (*Phalacrocorax carbo*), and many more. The rest of the observed community is made up of an amphibian species and several dragonflies and damselflies, and only one species, Song Thrush (*Turdus philomelos*), not (or not only) associated to the Common Kingfisher's typical habitats. Greylag Goose (*Anser anser*) can be observed mostly on lakes, ponds etc., but also foraging in adjacent fields. Its observed community (see Table 5.1.3) with 36 species (excluding *A. anser* itself) is also dominated by birds (52.8%). Of these bird species, most are water fowl or other bird species associated to water bodies, and birds also found in open fields, such as Common Pheasant (*Phasianus colchicus*) and White Stork (*Ciconia ciconia*). Additionally, most other species which are part of Greylag Goose's observed community are typically associated to aquatic habitats or open fields, such as a number of dragonflies and damselflies, Nutria (*Myocastor coypus*), Brown Hare (*Lepus europaeus*), and an amphibian species. Even one species of butterfly, Lesser Purple Emperor (*Apatura ilia*), which is associated to Aspen (*Populus tremula*)[45], a

[45] https://arteninfo.net/elearning/tagfalter/speciesportrait/1707, last accessed on 2018-11-18

tree found typically close to water bodies, is part of Greylag Goose's observed community, the only butterfly species in this list.

Such considerations are necessarily burdened with a considerable amount of uncertainty, beginning with the classification of species in certain habitat preferences, which are often quite variable even for the more specialized species, and which are made even more uncertain by the mobility of many animal species. They are therefore hard to quantify (for this reason, I abstained from giving numeric ratios which would obscure uncertainty with false precision), and they are certainly debatable. The examples cited above nevertheless illustrate the joint influence, on observed communities, of natural species associations (overall dominance of species sharing the target species' habitat preference) and VGI-related factors (disproportionately high rates of species from the target species' group within the observed communities cited). The opportunistic nature of the data collection process obviously does not blur habitat preferences in these cases. It leads, however, to observed communities whose species compositions lean towards the respective target species' group, and which are also governed by the general species group composition of the data pool from which they come.

These findings can also be traced in members of other species groups. For example, dragonflies and damselflies, although mostly quite mobile, are predominantly observed close to water bodies, leading to observed communities which favor species which are associated to water. E.g., Southern Hawker (*Aeshna cyanea*), a large dragonfly, has an ArtenFinder observed community (see Table 5.1.3) with many other dragonflies and damselflies. Their dominance in the observed community can be expected, considering the effects described above, but they also match Southern Hawker's habitat preference, as do species such as Grass Snake (*Natrix natrix*) and the amphibians in this observed community. However, the observed community also contains several species from other groups (a bird, a mammal, several butterflies, and a locust) which cannot be accounted as being especially attached to water bodies, illustrating that the quite mobile Southern Hawker is also observed in other contexts than water bodies.

The situation is in part different for species which are observed in many different habitats, either due to a lack of specific habitat preferences, or due to a high mobility, or both. For instance, European Greenfinch (*Carduelis chloris*) is widespread in Rheinland-Pfalz, occurring in a wide range of half open habitats, floodplain forest, and settled areas (Rößner et al. 2013). Its ArtenFinder observed community (see Table 5.1.4) of 26 species (not counting *C. chloris* itself) is a mix of species with unspecific as well as many different specific habitat preferences. The observed community is, however, again dominated by birds (at 69.2%). Another example is the widespread moth Silver Y (*Autographa gamma*) whose observed community (see Table 5.1.4) has similar properties, now dominated by butterflies and moths instead of birds (due to the VGI-related influence described earlier, here even with a pronounced portion of moths).

Similar examples can be found in iNaturalist data. Canada Goose (*Branta canadensis*), with similar habitat preferences as Greylag Goose, has only birds in its iNaturalist observed community (see Table 5.1.5) of 11 species (excluding *B. canadensis* itself), again showing VGI-related influences. Almost all of them are associated to water bodies or open fields (governed by natural species associations). Black Oystercatcher (*Haematopus bachmani*), living on rocky coastlines, has an iNaturalist observed community (see Table 5.1.5) of 29 species (excluding *H. bachmani*), most of which are also found predominantly or seasonally along the coast (including 14 bird species dominating the observed community at 65.6%). Seaside Daisy (*Erigeron glaucus*), also occurring mostly along California's coastline, has only about 50% species typical for the coast in its observed community (see Table 5.1.5) which is dominated by plants. Meaningful examples of valid observed communities for common or widespread species are not so easy to find in this data use case, because the larger and ecologically more diverse area of interest leads to very small observed communities for most very common species. For instance, Turkey Vulture (*Cathartes aura*), common and widespread throughout California, is associated only to itself, while no other species reaches the 50% threshold of occurrence in target observation neighborhoods necessary to become part of an observed community (see Table 5.1.6). A similar case is presented by California Poppy (*Eschscholzia californica*), associated only to itself and two other very common and frequently observed species (Western Fence-lizard (*Sceloporus occidentalis*) and Western Poison Oak (*Toxicodendron diversilobum*)) (see Table 5.1.6). This is, of course, in itself an effect of these species' wide range of observation situations, proving their own commonness. California Scrub-jay (*Aphelocoma californica*), also very common in California and occurring in a wide variety of habitats[46], has a seven-species observed community (excluding *A. californica*) dominated by birds (five out of seven species), with an overall majority of common species or species observed in many different habitats, such as Red-tailed Hawk (*Buteo jamaicensis*), Western Fence-lizard, California Poppy, and Anna's Hummingbird (*Calypte anna*) (Table 5.1.6).

It is important to stress that habitat preferences of a species will only show in observed communities if and when the data collection process propagates them into the data. A very common species with unspecific habitat preferences which is observed only in certain specific situations will have an observed community which does not represent its commonness, but will be misleading from this point of view. This is less prominent for more specialized species, which can in fact be observed predominantly in situations matching their (main) habitat preference, but individuals of more mobile species may of course also be observed elsewhere, introducing other species into the observed community (but only if the frequency of such cases is high enough). An example for the latter case might be Middle Spotted Woodpecker (*Dendrocopos medius*), which prefers forests with trees providing coarse bark, mainly oaks (Rößner et al. 2013). The Middle Spotted Woodpecker's 12-species ArtenFinder observed community (see Table 5.1.4) features a mix of species with different habitat preferences. Still, this list is a product of the underlying observation process, reflecting the properties of the data, and therefore suitable for evaluating a candidate context's match with these data.

In summary, these results show that the method developed for extracting observed communities, including a restriction to frequently associated species and the identification and elimination of nonspecific species, produces observed communities which mostly match a target species' habitat preferences, within the VGI-related limitations and modified by the VGI properties of the data pool used.

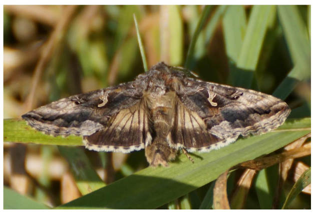
[46] http://www.audubon.org/field-guide/bird/california-scrub-jay, last accessed on 2018-11-18

Table 5.1.3: ArtenFinder, examples of observed communities: Common Kingfisher, Greylag Goose and Southern Hawker. Green: species which can be expected to be associated with the target species due to habitat preferences (source used for classification: ArtenInfo Rheinland-Pfalz[47]).

Target Species	**Common Kingfisher (*Alcedo atthis*)**	**Greylag Goose (*Anser anser*)**	**Southern Hawker (*Aeshna cyanea*)**
Photos: C. Jacobs			
Observed Community (ArtenFinder)	**Birds:** ..*Alcedo atthis* ..*Alopochen aegyptiaca* ..*Branta canadensis* ..*Casmerodius albus* ..*Cygnus olor* ..*Fulica atra* ..*Phalacrocorax carbo* ..*Turdus philomelos* **Amphibians:** ..*Rana kl. esculenta* **Dragonflies and damselflies:** ..*Aeshna cyanea* ..*Anax imperator* ..*Calopteryx splendens* ..*Ischnura elegans* ..*Orthetrum cancellatum* ..*Sympetrum striolatum*	**Mammals:** ..*Lepus europaeus* ..*Myocastor coypus* ..*Sciurus vulgaris* **Birds:** ..*Alcedo atthis* ..*Alopochen aegyptiaca* ..*Anas strepera* ..*Anser anser* ..*Aythya ferina* ..*Aythya fuligula* ..*Branta canadensis* ..*Carduelis chloris* ..*Casmerodius albus* ..*Ciconia ciconia* ..*Cuculus canorus* ..*Cygnus olor* ..*Fulica atra* ..*Gallinula chloropus* ..*Hirundo rustica* ..*Luscinia megarhynchos* ..*Phalacrocorax carbo* ..*Phasianus colchicus* ..*Podiceps cristatus* ..*Tachybaptus ruficollis* **Reptiles:** ..*Lacerta agilis* **Amphibians:** ..*Rana kl. esculenta* **Butterflies and moths:** ..*Apatura ilia* **Dragonflies and damselflies:** ..*Aeshna mixta* ..*Calopteryx splendens* ..*Coenagrion puella* ..*Crocothemis erythraea* ..*Ischnura elegans* ..*Lestes viridis* ..*Libellula fulva* ..*Orthetrum cancellatum* ..*Platycnemis pennipes* ..*Sympetrum sanguineum* ..*Sympetrum striolatum*	**Mammals:** ..*Sciurus vulgaris* **Birds:** ..*Turdus philomelos* **Reptiles:** ..*Natrix natrix* **Amphibians:** ..*Bufo bufo* ..*Rana kl. esculenta* **Butterflies and moths:** ..*Aphantopus hyperantus* ..*Argynnis paphia* ..*Colias croceus* ..*Papilio machaon* ..*Pyronia tithonus* **Hymenopterans:** ..*Vespa crabro* **Dragonflies and damselflies:** ..*Aeshna cyanea* ..*Anax imperator* ..*Calopteryx splendens* ..*Calopteryx virgo* ..*Coenagrion puella* ..*Ischnura elegans* ..*Lestes viridis* ..*Libellula depressa* ..*Libellula quadrimaculata* ..*Orthetrum cancellatum* ..*Platycnemis pennipes* ..*Pyrrhosoma nymphula* ..*Sympecma fusca* ..*Sympetrum sanguineum* ..*Sympetrum striolatum* **Locusts:** ..*Oedipoda caerulescens*

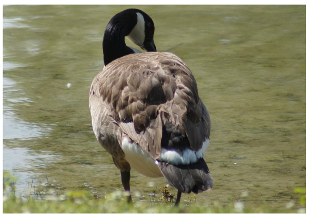
[47] https://artenfinder.rlp.de/node/15

Table 5.1.4: ArtenFinder, examples of observed communities: European Greenfinch, Silver Y and Middle Spotted Woodpecker. Green: species which can be expected to be associated with the target species due to habitat preferences (not applicable to common and widespread C. chloris and A. gamma; source used for classification: ArtenInfo Rheinland-Pfalz[48]).

Target Species	**European Greenfinch (*Carduelis chloris*)**	**Silver Y (*Autographa gamma*)**	**Middle Spotted Woodpecker (*Dendrocopos medius*)**
Photos: C. Jacobs			
Observed Community (ArtenFinder)	**Mammals:** ..Sciurus vulgaris **Birds:** ..*Accipiter nisus* ..*Alcedo atthis* ..*Carduelis chloris* ..*Carduelis spinus* ..*Coccothraustes* ..*coccothraustes* ..*Coloeus monedula* ..*Delichon urbicum* ..*Fringilla montifringilla* ..*Grus grus* ..*Hirundo rustica* ..*Milvus milvus* ..*Parus palustris* ..*Passer domesticus* ..*Prunella modularis* ..*Pyrrhula pyrrhula* ..*Serinus serinus* ..*Streptopelia decaocto* ..*Turdus philomelos* ..*Turdus pilaris* **Reptiles:** ..*Podarcis muralis* **Butterflies and moths:** ..*Lasiommata megera* ..*Macroglossum* ..*stellatarum* ..*Papilio machaon* **Hymenopterans:** ..*Xylocopa violacea* **Beetles:** ..*Lucanus cervus* **Dragonflies and damselflies:** ..*Calopteryx virgo*	**Mammals:** *Sciurus vulgaris* **Birds:** *Lanius collurio* *Milvus milvus* *Phasianus colchicus* **Reptiles:** *Lacerta agilis* **Butterflies and moths:** *Aphantopus hyperantus* *Argynnis paphia* *Aricia agestis* *Autographa gamma* *Camptogramma* *bilineata* *Chiasmia clathrata* *Colias croceus* *Colias hyale* *Diacrisia sannio* *Ematurga atomaria* *Euclidia glyphica* *Lasiommata megera* *Leptidea sinapis s.l.* *Macroglossum* ..*stellatarum* *Papilio machaon* *Pyronia tithonus* *Siona lineata* *Thymelicus lineola* *Thymelicus sylvestris* **Hymenopterans:** *Vespa crabro* **Dragonflies and damselflies:** *Aeshna cyanea* *Calopteryx splendens* *Calopteryx virgo* *Platycnemis pennipes* *Sympetrum striolatum* **Locusts:** *Oedipoda caerulescens* *Tettigonia viridissima*	**Mammals:** *Sciurus vulgaris* **Birds:** *Carduelis chloris* *Cuculus canorus* *Dendrocopos medius* *Dryocopus martius* *Lanius collurio* *Turdus philomelos* **Reptiles:** *Lacerta agilis* **Butterflies and moths:** *Papilio machaon* **Dragonflies and damselflies:** *Calopteryx splendens* *Ischnura elegans* *Sympetrum striolatum* **Locusts:** *Tettigonia viridissima*

[48] https://artenfinder.rlp.de/node/15

Table 5.1.5: iNaturalist, examples of observed communities: Canada Goose, Black Oystercatcher and Seaside Daisy. Green: species which can be expected to be associated with the target species due to habitat preferences (sources used for classification: Audubon Guide to North American Birds[49]; Jepson eFlora[50], Alden et al. 1998).

Target Species	**Canada Goose (*Branta canadensis*)**	**Black Oystercatcher (Haematopus bachmani)**	**Seaside Daisy (*Erigeron glaucus*)**
Photos: C. Jacobs			
Observed Community (iNaturalist)	***Birds:*** *..Anas platyrhynchos* *..Ardea alba* *..Ardea herodias* *..Branta canadensis* *..Buteo jamaicensis* *..Calypte anna* *..Egretta thula* *..Fulica americana* *..Nycticorax nycticorax* *..Phalacrocorax auritus* *..Sayornis nigricans* *..Zonotrichia leucophrys*	**Plants:** *..Erigeron glaucus* *..Eschscholzia californica* **Mammals:** *..Phoca vitulina* *..Zalophus californianus* **Birds:** *..Aechmophorus* *..occidentalis* *..Ardea alba* *..Ardea herodias* *..Arenaria melanocephala* *..Buteo jamaicensis* *..Calidris virgata* *..Egretta thula* *..Euphagus cyanocephalus* *..Gavia stellata* *..Haematopus bachmani* *..Larus occidentalis* *..Melanitta perspicillata* *..Melospiza melodia* *..Numenius phaeopus* *..Pelecanus occidentalis* *..Phalacrocorax auritus* *..Phalacrocorax pelagicus* *..Phalacrocorax* *..penicillatus* *..Sayornis nigricans* *..Zonotrichia leucophrys* **Crustaceans:** *..Pachygrapsus crassipes* **Mollusks:** *..Mytilus californianus* *..Tegula funebralis* **"Other species":** *..Anthopleura* *..xanthogrammica* *..Pisaster ochraceus* *..Pollicipes polymerus*	**Plants:** *..Achillea millefolium* *..Baccharis pilularis* *..Carpobrotus edulis* *..Diplacus aurantiacus* *..Dudleya farinosa* *..Erigeron glaucus* *..Eriogonum latifolium* *..Eriophyllum* *..staechadifolium* *..Eschscholzia californica* *..Fragaria chiloensis* *..Lupinus arboreus* *..Oxalis pes-caprae* *..Toxicodendron* *..diversilobum* **Birds:** *..Buteo jamaicensis* *..Calypte anna* *..Corvus corax* *..Larus occidentalis* *..Melospiza melodia* *..Pelecanus occidentalis* *..Zonotrichia leucophrys* **Mollusks:** *..Mytilus californianus*

[49] http://www.audubon.org/bird-guide

[50] Jepson Flora Project (eds.) 2018. Jepson eFlora, http://ucjeps.berkeley.edu/eflora

Table 5.1.6: iNaturalist, examples of observed communities of common and widespread species: Turkey Vulture, California Poppy and California Scrub-jay.

Target Species	**Turkey Vulture (*Cathartes aura*)**	**California Poppy (*Eschscholzia californica*)**	**California Scrub-jay (*Aphelocoma californica*)**
Photos: C. Jacobs			
Observed Community (iNaturalist)	***Birds:*** ..*Cathartes aura*	**Plants:** ..*Eschscholzia californica* ..*Toxicodendron* ..*diversilobum* **Reptiles:** .. *Sceloporus occidentalis*	**Plants:** ..*Eschscholzia californica* **Birds:** ..*Aphelocoma californica* ..*Buteo jamaicensis* ..*Calypte anna* ..*Melozone crissalis* ..*Sayornis nigricans* ..*Zonotrichia leucophrys* **Reptiles:** ..*Sceloporus occidentalis*

5.1.3 Nonspecific Species in Observed Communities

An important last step in extracting observed communities consists in identifying so-called nonspecific species, and removing them from observed communities. In the evaluation whose results are discussed here, these are species which occur in 50% or more of the observed communities and were therefore considered not to contribute to the distinctness of these species lists. Table 5.1.7 presents the 46 species identified as nonspecific species in evaluation with ArtenFinder data. The list of nonspecific species has mostly birds (54.3%) and butterflies (43.5%), and one mammal. An examination of these species reveals that they all share one or more of the following properties: they are very common, have rather unspecific habitat preferences, and are abundant and/or well detectable and identifiable. A threshold of 50% frequency in observed communities therefore seems reasonable (at least not too low, as it does not include more rare or more specialized species which would come as a surprise if identified as nonspecific species). The species listed in Table 5.1.7 are all among the top 82 of the most reported species in ArtenFinder observations from 2015 and earlier. Evaluation with iNaturalist data and the same parameters as in the ArtenFinder case does not render any nonspecific species. No species is present in 50% or more of the iNaturalist observed communities.

Table 5.1.7: ArtenFinder, species identified as nonspecific species. (Species occurring in 50% or more of observed communities at the same time). Ordered descending by frequency of occurrence in observed communities.

Species	**Species Group**	**Frequency in Observed Communities (%)**
European Peacock (*Inachis io*)	butterflies	94,5
Common Chaffinch (*Fringilla coelebs*)	birds	93,8
Small White (*Pieris rapae*)	butterflies	93,5
Red Admiral (*Vanessa atalanta*)	butterflies	93,1
Speckled Wood (*Pararge aegeria*)	butterflies	92,4
Common Buzzard (*Buteo buteo*)	birds	92,1
Eurasian Blackbird (*Turdus merula*)	birds	91,1
Great Tit (*Parus major*)	birds	90,0
Comma (*Polygonia c-album*)	butterflies	90,0
Green-veined White (*Pieris napi*)	butterflies	89,7
Common Brimstone (*Gonepteryx rhamni*)	butterflies	89,3
European Green Woodpecker (*Picus viridis*)	birds	87,6
European Robin (*Erithacus rubecula*)	birds	87,3
Common Blue (*Polyommatus icarus*)	butterflies	86,3
Eurasian Jay (*Garrulus glandarius*)	birds	84,9
Eurasian Blue Tit (*Parus caeruleus*)	birds	84,9
Small Heath (*Coenonympha pamphilus*)	butterflies	84,5
Small Tortoiseshell (*Aglais urticae*)	butterflies	83,2
Common Chiffchaff (*Phylloscopus collybita*)	birds	81,8
Meadow Brown (*Maniola jurtina*)	butterflies	81,8
Eurasian Blackcap (*Sylvia atricapilla*)	birds	80,1
Great Spotted Woodpecker (*Dendrocopos major*)	birds	79,7
Eurasian Wren (*Troglodytes troglodytes*)	birds	79,7
White Wagtail (*Motacilla alba*)	birds	77,7
European Starling (*Sturnus vulgaris*)	birds	77,7
Eurasian Kestrel (*Falco tinnunculus*)	birds	76,6
Marbled White (*Melanargia galathea*)	butterflies	76,3
Carrion Crow (*Corvus corone*)	birds	75,9
Grey Heron (*Ardea cinerea*)	birds	74,9
Map (*Araschnia levana*)	butterflies	72,9
Orange Tip (*Anthocharis cardamines*)	butterflies	71,5
Short-tailed Blue (*Cupido argiades*)	butterflies	71,5
Common Copper (*Lycaena phlaeas*)	butterflies	70,1
Eurasian Nuthatch (*Sitta europaea*)	birds	67,0
Yellowhammer (*Emberiza citrinella*)	birds	66,7
Common Wood Pigeon (*Columba palumbus*)	birds	66,3
Painted Lady (*Vanessa cardui*)	butterflies	65,3
European Goldfinch (*Carduelis carduelis*)	birds	64,3
Long-tailed Tit (*Aegithalos caudatus*)	birds	63,6
Western Roe Deer (*Capreolus capreolus*)	mammals	61,2
Eurasian Magpie (*Pica pica*)	birds	60,8
Black Redstart (*Phoenicurus ochruros*)	birds	60,5
Mallard (*Anas platyrhynchos*)	birds	60,1
Cabbage White (*Pieris brassicae*)	butterflies	59,5
Large Skipper (*Ochlodes sylvanus*)	butterflies	59,1
Holly Blue (*Celastrina argiolus*)	butterflies	58,8

5.1.4 Differences in Similarity Values Between Sets of Candidate Observations with Observed Communities

For evaluation purposes, a number of different sets of candidate observations were used to examine the observed communities approach's ability to distinguish between candidate cases which can be expected to be, for the most part, plausible, and other candidates which are, for the most part, implausible.

First of all, similarity values and their distributions are markedly different from one another for the two different similarity indices used: Simpson index values are generally much higher than Jaccard index values. This is caused by the differing structures of the indices themselves. Jaccard index values are generally low, because the size of the intersection of the two lists involved is divided by the union of the two lists. The Jaccard index is therefore sensitive to the difference between the two lists involved. In most cases, the candidate context introduces a large number of species into the calculation, which in turn causes Jaccard index values to be generally low. The opposite is true for Simpson index results: union is always put in relation to the length of the shorter of the two lists involved (which usually is the length of the observed community, in 72-80% of cases), so that index values are generally much higher.

The results of the evaluation laid out in sections 4.1.1 and 4.1.2 show great differences in plausibility values obtained with the different candidate sets. In general, sets containing candidate observations which are expected to be found plausible by the approach (that is, candidates which should obtain high similarity values between candidate context and observed community), do indeed have relatively higher values than sets containing candidates which are expected to be found implausible by the approach (that is, candidates which should obtain low similarity values between candidate contexts and observed communities). However, for both plausible and implausible candidate sets, there are pronounced differences between sets of real candidate observations and sets of synthetic candidate observations. Synthetic sets of candidates were especially designed to contain high portions of plausible or implausible observations (see section 3.3.2) while sets of real candidate observations have several issues which make them rather heterogeneous in this respect.

Real approved observations in the sets AF_A and iNat_A show high similarity values relative to all sets of expectedly implausible candidates for both indices used in evaluation. However, results also show that they also contain observations which are evaluated as implausible by the observed communities approach. In distributions of Simpson index values, frequencies drop from a peak near 1.0 towards lower index values, but frequencies are nowhere close to zero. In distributions of Jaccard index values (which are generally much lower, see above), frequencies also drop towards lower index values to the left of the frequency peak (which is not at or close to 1.0, but lower here), but also do not reach zero frequency of low index values. Based on these results, synthetic sets of plausible candidate observations were produced by placing synthetic candidate observations randomly but close to plausible observations of the same target species from the sets of approved observations, which produced the sets AF_SP and iNat_SP. Due to their spatial proximity to plausible observations, these synthetic candidate observations should also be plausible. As expected, these sets show much higher similarity values for both indices. Simpson index value frequencies of AF_SP and iNat_SP exhibit very strong peaks close to 1.0 and drop to zero (or close to zero) frequency well before similarity values of zero are reached. Frequencies of Jaccard index values in these sets peak at much higher values than before (although still not or not only close to 1.0) and drop to frequencies close to zero before reaching similarity values of zero. We can draw two conclusions from these results. One, any candidate observation from a location close to a plausible observation of the same target species will be evaluated as plausible by the observed communities approach. What is more, this can be expected also for any candidate observation located in a place which has a similar context, without necessarily having a previous observation of the target species close by. Two, many approved observations are evaluated as plausible by the observed communities approach, but some are not. This clearly shows that approval of a candidate observation may be (and often is) based on reasons other than a matching observation context. Validation procedures and reasons for approval of an observation in the two data use cases were described in section 2.1. In both projects, photo proofs play a major role in approval of species identification by experts (ArtenFinder) or by the community of observers (iNaturalist). Other reasons, such as observer experience and observation date, are also important. The location of the observation certainly plays a role, but it is not possible to judge in how far the geographic context in the form of species observed in proximity to a candidate observation is evaluated by experts or fellow observers. ArtenFinder explicitly provides this information (a list of all species observed close to the candidate) to the expert evaluating a candidate observation, but it has proved to be too complex and time-consuming to use (see section 2.1.1). iNaturalist also has a tool which provides information on observations situated close to a candidate observation, and on species generally occurring in the proper region. There is currently no information available on how much this tool is actually used by co-observers when giving a species identification, when confirming it when or disagreeing with it.

Similar conclusions as the above can be drawn for the sets of implausible observations. Two of these, the _R set (only available for ArtenFinder) and the _SI1 set, are based on real observations (see section 3.3.2). AF_R is a set of observations which were rejected by ArtenFinder experts. Its similarity values are, however, rather high, although statistically lower than both AF_A and AF_SP similarities, but still the highest of all ArtenFinder sets expected to be dominated by implausible observations. This fact points to some issues concerning the data structure and properties of the AF_R set. Some of these have their cause in the quality assurance procedure which produces rejected observations. Reasons for rejecting an observation are not always based on suspecting the species identification or the observation location to be wrong. In some cases, observations are rejected for technical reasons, most often a missing or insufficient photo proof, required especially of volunteers who are beginners, or whose reputation for the species group in question is not sufficient. Also, any observation provided with a photo proof (required or not) can only be accepted by the validating expert if the photo gives unequivocal evidence of the correctness of the species identification. In all of these cases, there is some probability that the rejected observations involved are, in fact, correct, thus adding cases to the set of rejected observations used in evaluation which would rather be expected to be plausible by the observed communities approach. However, as reasons for rejection are not recorded, these cases cannot be pruned from the AF_R set. Volunteers are provided with feedback, but this information, transmitted mostly by comments to the observation, or by email, cannot be readily used in analysis. I add here that the ArtenFinder quality assurance procedure was changed in spring 2016 by providing the possibility to put an observation on hold instead of rejecting it out-right. This change will undoubtedly lead to a reduction in rejection cases of this kind.

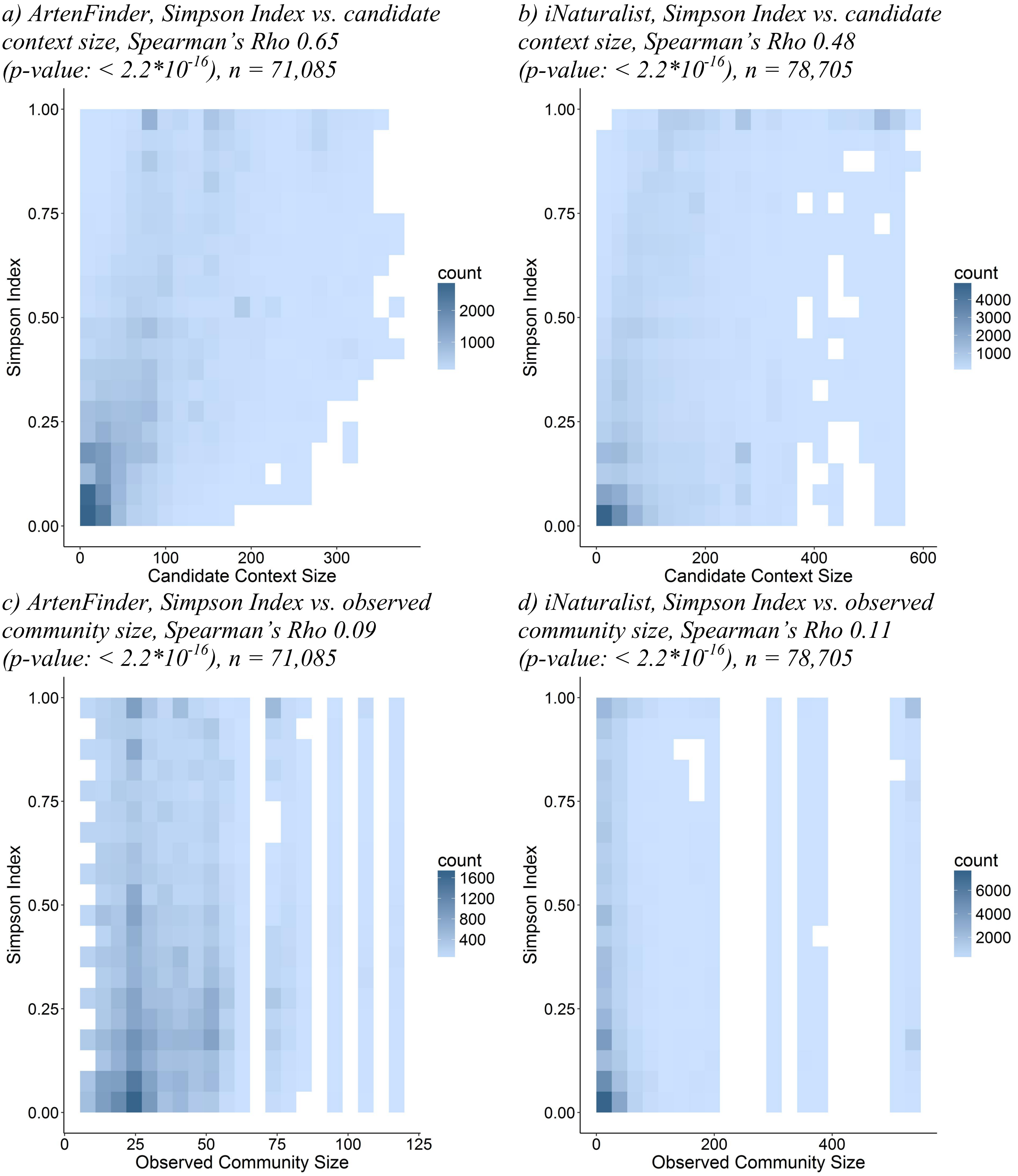
The _SI1 sets were produced by swapping species identifications between observations of real approved observations which are of physically similar species living in different habitats and/or regions. Observations synthesized in this way can, to some extent, be expected to have implausible contexts. Indeed, AF_SI1 shows mostly lower Simpson and Jaccard index values than AF_A, but still overall much higher values than AF_SI2 and AF_SI3. iNat_SI1 also has mostly lower Simpson and Jaccard index values than iNat_A, but considerably higher than iNat_SI2 and iNat_SI3. Thus it can be said that observations of a species observed in locations where its physically similar counterpart was actually observed are indeed less plausible, considering their observation context, but usually do not expose very low similarity values with their observed communities. It is, however, also true that they rarely have very high similarity values, which frequently occur in _A sets, and which dominate _SP sets.

The frequency distribution especially of the similarity values of AF_SI1 shows some peculiarities. As can be seen from Figure 4.1.1, the AF_SI1 kernel density has two distinct peaks at medium values, at 0.45 to 0.50, and at 0.60 to 0.65. Besides a higher variability in frequencies which can be expected for a smaller dataset, also typical "VGI effects" in the data play a role here. For instance, there is a body of 161 observations (almost 10% of valid AF_SI1 candidates) of the perching bird Yellowhammer (*Emberiza citrinella*), replaced in AF_SI1 by Cirl Bunting (*Emberiza cirlus*) with which it is often confused. All auf these 161 observations come from the same observer and were observed at the same location in 2016. The observer reported Yellowhammer from that location almost daily, obviously observing a certain spot very regularly over a longer period of time. At the location in question, the observations produce a Simpson index value of 0.45 with the wrong species identification of Cirl Bunting. A similar effect was produced by another observer who reported the reptile Common Wall Lizard (*Podarcis muralis*) very regularly in 2016 from a small area. For creation of the AF_SI1 set, the species for these 138 observations was replaced by Sand Lizard (*Lacerta agilis*), and the Simpson index value for these observations was 0.64. AF_A Jaccard values show a small peak between 0.9 and 1.0. It is caused by 426 candidates of two bird species, Marsh Tit (with the scientific name *Parus palustris* in ArtenFinder) and Eurasian Collared Dove (*Streptopelia decaocto*). An observer provided reports of these two species from the same location consistently over several years, so that reports up to 2015 from this location dominated in creating the observed communities for these species, and candidate observations from 2016 from this location have high similarity values. iNat_A Jaccard similarity values exhibit a similar case, here with a number of marine mollusks. These observations also cause the group of high Jaccard similarity values above ca. 0.8 in iNat_SP. Note also that the iNat_SP Jaccard values distribution's bimodality leads to a rather elongated boxplot in Figure 4.1.4, making a boxplot, strictly speaking, not a suitable way of visualizing such distributions.

Candidate observations of _SI2 and _SI3 sets were synthesized from randomly placed observations away from known observations of their target species, with the aim of producing, as far as possible, very implausible observations. While _SI2 candidates were required to mirror the spatial properties of their target species' known observations by being placed in situations with similar observation density, _SI3 observations were placed without this restriction. AF_SI3 and iNat_SI3 expose extremely low similarity values when compared to all other sets. AF_SI2 and iNat_SI2 similarity values are somewhat higher, but still much lower than _SI1 and AF_R sets (and, of course, than _A and _SP sets). Evaluation results with _SI2 and _SI3 sets tell us that the observed communities approach reliably identifies observations as implausible which come from locations away from earlier observations of the same target species, because these locations have mostly a different species context.

5.1.5 Effects of Spatial Distribution of Observation Data on Similarity Values

Due to their VGI origin and nature, casual citizen science observations of organisms show clustering (see section 2.1). Spatial density of observations is therefore variable. Candidate observations tested for plausibility, as well as target observations used for extracting observed communities, are found in different situations concerning spatial density of observations around them. Spatial observation density can be expected to have a positive influence on the number of species found around an observation. Both the Simpson and the Jaccard index use the intersection between two lists of species to measure similarity between the two lists. In this work, the two lists are the observed community of a target species, and the species context of a candidate observation. Obviously, if one of the two lists is large, this raises the probability of a larger intersection with the other list, leading to a higher similarity value. Therefore, there may be an effect of the spatial properties of the data on index values. This section examines this effect and discusses consequences for the use of the Simpson and the Jaccard index as plausibility estimators. Results are mostly presented in the form of binned scatterplots, because numbers of points are mostly very high, and true distribution of points within the scatterplots would not be well visible in classic scatterplots which draw individual points. Strengths of relevant correlations are examined with Spearman's Rho rank correlation coefficient, because the correlations involved often appear not to be linear.



a) ArtenFinder, Simpson Index vs. candidate context size, Spearman's Rho 0.65 (p-value: < $2.2*10^{-16}$), n = 71,085

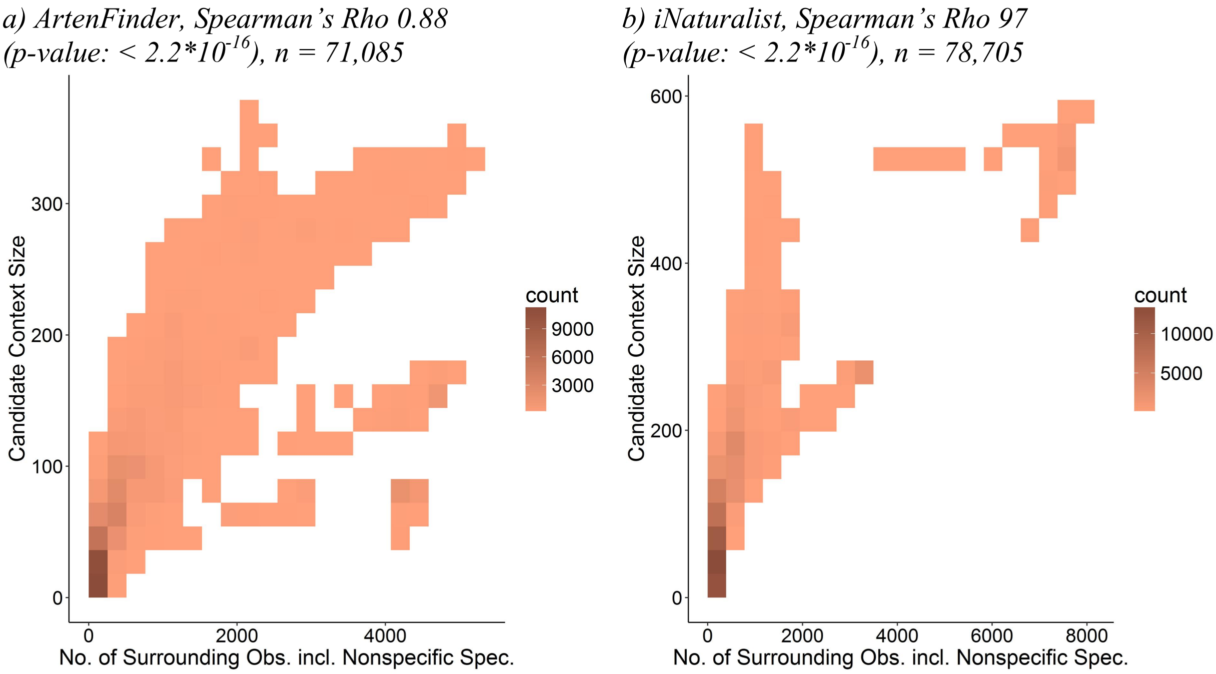
b) iNaturalist, Simpson Index vs. candidate context size, Spearman's Rho 0.48 (p-value: < $2.2*10^{-16}$), n = 78,705

c) ArtenFinder, Simpson Index vs. observed community size, Spearman's Rho 0.09 (p-value: < $2.2*10^{-16}$), n = 71,085

d) iNaturalist, Simpson Index vs. observed community size, Spearman's Rho 0.11 (p-value: < $2.2*10^{-16}$), n = 78,705

Figure 5.1.2: ArtenFinder and iNaturalist data, Simpson index values vs. candidate context size and observed community size (no. of species).

Simpson index values show a positive correlation with the numbers of species found around a candidate observation (i.e., candidate context size), which in turn depends on the number of observations found there (see Figure 5.1.2 and Figure 5.1.3). In other terms, candidate observations which are situated in areas with a higher observation density tend to have higher Simpson index values than candidate observations from regions with lower observation density. A higher number of species found around a candidate observation increases the chance of covering a larger number of observed community species. However, there is no pronounced positive correlation between Simpson index values and numbers of species in observed communities (i.e., observed community size). The Simpson index always uses the length of the shorter list as denominator, which is usually the observed community. This is the case in 71.5% of all candidate cases in ArtenFinder sets, and in 79.7% of all candidate cases for iNaturalist sets. Differences in observed community size are therefore mostly corrected for in Simpson index values, while effects of variable candidate context size remain. Note that p-values are often given here as "< 2.2*10^{-16}", this number representing the minimum value of the R function used for calculation (R package "stats", function "cor.test").


*a) ArtenFinder, Spearman's Rho 0.88 (p-value: < 2.2*10^{-16}), n = 71,085*

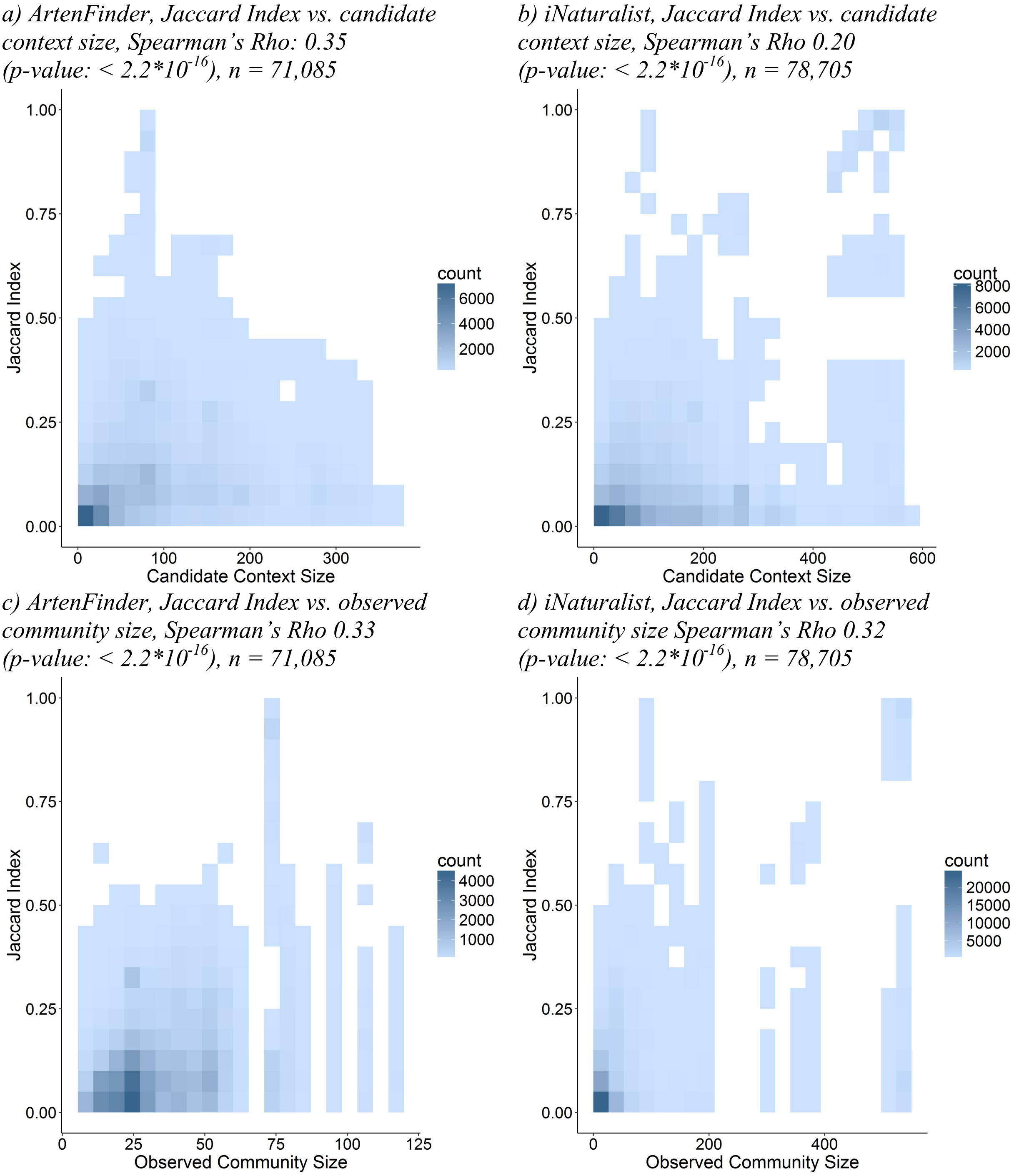
*b) iNaturalist, Spearman's Rho 97 (p-value: < 2.2*10^{-16}), n = 78,705*

Figure 5.1.3: ArtenFinder and iNaturalist data, candidate context size (no. of species) vs. no. of surrounding observations (all context observations counted).

Figure 5.1.3 illustrates that observation density has a pronounced effect on candidate context size: the more context observations can be found in a candidate observation's neighborhood, the more species will be listed in the candidate context. For the Simpson index, observation density therefore has a positive effect on similarity values and thus on the plausibility estimation: with the Simpson index as a plausibility indicator, observations from regions with higher observation density appear more plausible than observations from regions with low observation density. This may be regarded as problematic from a biological or ecological perspective: plausibility is raised merely because an observation is situated in an area with a higher observation density (e.g., because that area is visited more frequently by volunteers). From a VGI perspective, this is, however, correct: observations of organisms by volunteers tend to cluster (see section 2.1), and the probability for any observation to come from a region with high observation density is in fact higher, and thus also its actual plausibility.

The Jaccard index is different from the Simpson index in putting the intersection of the observed community and the candidate context in relation to the union of the two lists involved. Consequently, there is no imbalance between the effects of candidate context size and observed community size.

Jaccard index values show positive correlation with both (see Figure 5.1.4), but correlation coefficient values are rather low. Still, Jaccard index values do show a positive correlation with both candidate context size and observed community size. Candidate context size was found to depend on the number of context observations found around the candidate observation (see Figure 5.1.3), but what may be the reasons for observed community sizes? Two possible reasons are examined in the following.



*a) ArtenFinder, Jaccard Index vs. candidate context size, Spearman's Rho: 0.35 (p-value: < 2.2*10^{-16}), n = 71,085*

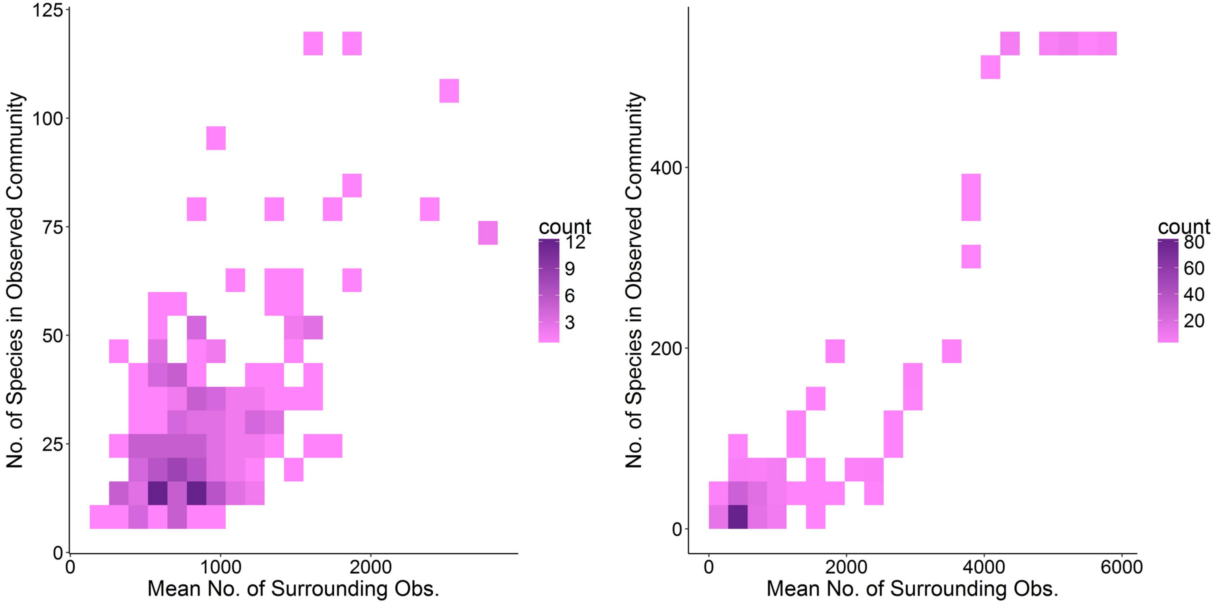
*b) iNaturalist, Jaccard Index vs. candidate context size, Spearman's Rho 0.20 (p-value: < 2.2*10^{-16}), n = 78,705*

*c) ArtenFinder, Jaccard Index vs. observed community size, Spearman's Rho 0.33 (p-value: < 2.2*10^{-16}), n = 71,085*

*d) iNaturalist, Jaccard Index vs. observed community size Spearman's Rho 0.32 (p-value: < 2.2*10^{-16}), n = 78,705*

Figure 5.1.4: ArtenFinder and iNaturalist data, Jaccard index values vs. candidate context size and observed community size (no. of species).

The first reason for differences in observed community sizes may be found in variable observation density: If observations of a target species, which are used to extract the observed community, are situated predominantly in areas with high observation density, they can be expected to have more observations around them, on average. This may lead to larger resulting observed communities, because (as was shown above) more observations usually represent more different species. This hypothesis was tested in the following way: For each target species, for which observed communities were extracted and used, the mean number of observations found around the observations used to extract the observed community was calculated. It was then tested whether the size of the resulting observed community depends to some degree on this number. The analysis found the expected positive correlations on the $p \leq 0.05$ level, see Figure 5.1.5: the sizes of observed communities are indeed positively influenced by the observation density around the observations used to extract them.



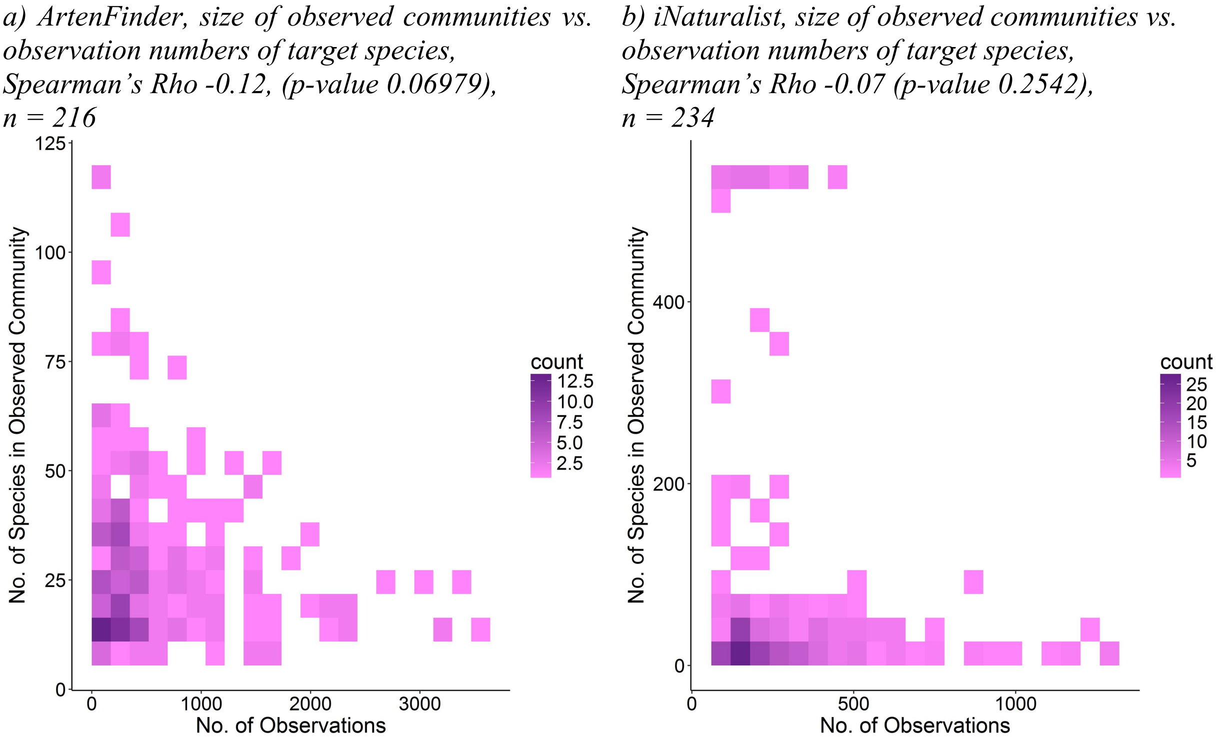
a) ArtenFinder, size of observed communities vs. mean numbers of observations surrounding observations used for observed community extraction, Spearman's Rho 0.43 (p-value $6.51*10^{-11}$), n = 216

b) iNaturalist, size of observed communities vs. mean numbers of observations surrounding observations used for observed community extraction, Spearman's Rho 0.70 (p-value < $2.2*10^{-16}$), n = 234

No. of Species in Observed Community
125
100
75
50
25
0
0
1000
2000
Mean No. of Surrounding Obs.
count
12
9
6
3
No. of Species in Observed Community
400
200
0
0
2000
4000
6000
Mean No. of Surrounding Obs.
count
80
60
40
20

Figure 5.1.5 ArtenFinder and iNaturalist data, observed community size (no. of species) vs. mean observation numbers around target species observations up to 2015 (observations used for observed community extraction).

A second cause which is not related to spatial observation density might also govern the size of observed communities: there may be a correlation with the number of observations of a target species which are available for extracting an observed community for that species, because with more observations, more different situations with potentially more different species go into observed community extraction. However, Figure 5.1.6 illustrates that the sizes of observed communities show no significant correlation (on the $p \leq 0.05$ level) with the number of observations used for extracting them. Detailed analysis of this somewhat surprising result revealed that the lengths of context species lists before restriction to frequently associated species and removal of nonspecific species indeed exhibit a pronounced correlation with the number of observations of the corresponding target species, but that the subsequent processing steps dilute this correlation (see section 3.1 for details on observed communities extraction).



a) ArtenFinder, size of observed communities vs. observation numbers of target species, Spearman's Rho -0.12, (p-value 0.06979), n = 216

b) iNaturalist, size of observed communities vs. observation numbers of target species, Spearman's Rho -0.07 (p-value 0.2542), n = 234

Figure 5.1.6: ArtenFinder and iNaturalist data, observed community size (no. of species) vs. observation numbers of target species up to 2015 (observations used for observed community extraction).

Similarity values calculated with the Simpson and Jaccard indices are governed by the species composition of candidate contexts and observed communities, and by the intersection of these two. In this section, another factor was identified and examined, which has an influence on the size of this intersection: variations in spatial observation density. With the Simpson index, this influence arises from observation density found around the candidate observation whose plausibility is estimated: candidate observations in areas with higher observation density tend to have larger candidate contexts, which may enlarge the intersection with the observed community, and consequently lead to higher Simpson index values. With the Jaccard index, the influence of this factor is smaller, but there is also moderate positive influence of higher observation density around observations used for extraction of the observed community of a species, which makes the resulting observed community larger, which in turn may raise Jaccard index values. The underlying reason for both effects is that a larger candidate context or a larger observed community both raise the chance of a larger intersection between the two lists. Due to the structure of the two indices, only candidate context size takes effect for Simpson index values in this way, while both parameters influence Jaccard index values.

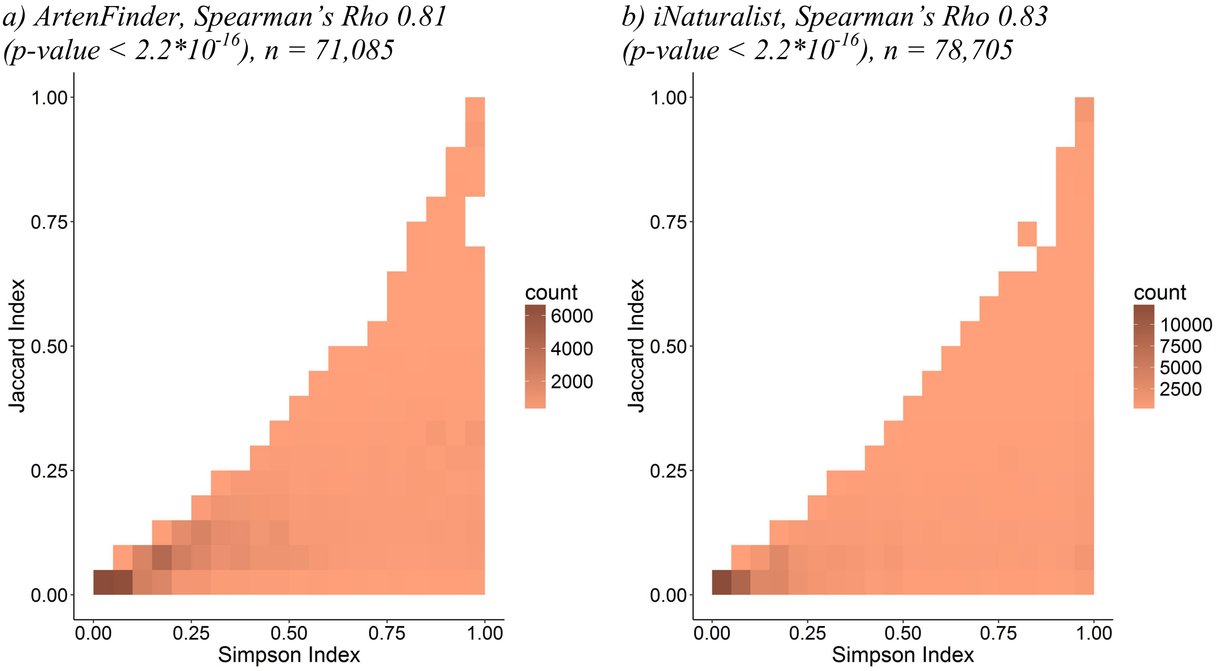
Casual citizen science observations of species have a variable spatial density and tend to cluster, which is due to their VGI nature and origin. Observations from areas with a higher observation density should appear more plausible, because they fit the spatial properties of existing data better, and this is what Simpson index values in part reflect. For Jaccard index values, VGI influence does not necessarily follow this rationale: plausibility of candidate observations of species whose earlier observations are predominantly situated in areas with higher observation density may also be raised.

How do these findings relate to evaluation results? Evaluation found significant differences between distributions of similarity index values of different sets of candidate observations. However, different sets of candidate observations also have different mean candidate context sizes as well as different mean observed community sizes (see Table 4.1.4 and Table 4.1.9). Differences in similarity values between these sets are therefore probably also due, in part, to the effects described above. For instance, AF_SI3 and iNat_SI3, with a mean candidate context size far below the average of sets in the respective data use cases, but with observed communities about average in size, probably have low Simpson similarity values also because most of their candidates are situated in (relatively) low observation density places, while _SI3 Jaccard index values, also lowest of all sets, are probably also influenced by candidate context size, but to a lesser degree. AF_SI2 and iNat_SI2 also have very low Simpson and Jaccard similarity values (if not as low as the _SI3 sets), but with about average candidate context and observed community sizes, these cases must own their low similarities for both indices manly to differences in species composition of candidate contexts and observed communities, with little or no effect of observation density. Other sets also show deviations of mean candidate context size and mean observed community size from the average in their respective data use cases, so that effects on similarity values can be expected accordingly. In some cases, such as AF_SI1 and iNat_SI1, below-average observed community sizes (with mean candidate context sizes on the average level, respectively) probably have lower Jaccard similarity values than they would have with average spatial properties in observed community source data. In AF_SP, with above-average observed community sizes (and again with mean candidate context sizes on the average level), Jaccard index values can be expected to be raised, instead. In cases where both candidate context size and observed community size deviate from the average in the same direction, such as AF_R or iNat_SP, effects on both indices are probably weak.

5.1.6 Correlation of Simpson and Jaccard Index in the Observed Communities Approach

Despite the differences in behavior between Simpson and Jaccard indices, which came to light in the evaluation results, similarity values with the two indices usually point in the same direction for the same candidate case. A candidate case with a high Simpson index value can also be expected to have a relatively high Jaccard index value (see Figure 5.1.7). Coincidences of a high Simpson index value with a low Jaccard index value for the same candidate case may also occur. Due to the structure of the two similarity indices used here, the opposite case, consisting of a coincidence of a low Simpson index value with a high Jaccard index value, cannot occur. This is easily explained. A low Simpson index value is always associated with a low intersection of observed community and candidate context. A low intersection will always also lead to a low Jaccard index value. A high intersection of observed community and candidate context always leads to a high Simpson index value, while the Jaccard index value may be low if candidate context is very large.

This also demonstrates that the two indices provide basically two different bases for plausibility estimation. The Simpson index always closely follows the intersection of observed community and candidate context. If this intersection is large, the Simpson index value will be high. Species which are part of the candidate context, but which do not match the observed community, are completely disregarded, but a large candidate context as a whole raises the chance for a large intersection and therefore, indirectly, the chance for plausibility being estimated as high. The Jaccard index includes more non-matching species in the calculation, which leads to lower Jaccard index values in general, and very high numbers of such species lead to very small index values (and thus also low plausibility estimations) in individual cases, especially when the observed community is small.



a) ArtenFinder, Spearman's Rho 0.81 (p-value < $2.2*10^{-16}$), n = 71,085

b) iNaturalist, Spearman's Rho 0.83 (p-value < $2.2*10^{-16}$), n = 78,705

Figure 5.1.7: Observed communities approach, Simpson vs. Jaccard index values.

5.2 OSM Environments Approach

In the following, evaluation results of the OSM environments approach are discussed and compared to those of the observed communities approach. Using OSM environments instead of observed communities as a source of geographic context introduces some changes into plausibility estimation of an observation. These changes are founded in the fact that OSM environments are not only based on the observation dataset involved, but also on OSM data, an extrinsic dataset (from the perspective of the citizen science observation data). Discussion starts with looking at important properties of OSM environments. Differences between the sets of candidates tested in evaluation, and their causes, are examined next. Discussion of evaluation results of the observed communities approach revealed influences of variable spatial density in the source of geographic context on similarity values. We will see that these influences are also present in the OSM environments approach. However, because the latter is an extrinsic source of geographic context, this has important consequences for the use of similarity values as plausibility indicators which differ from those with observed communities.

5.2.1 Similarity of OSM Environments

OSM environments are overall more similar to one another, than are observed communities (see Figure 5.2.1). One of the reasons for this might be that there are much less tags involved than species. For instance, the 183 valid observed communities in the evaluation with ArtenFinder data consist of 367 different species. Rheinland-Pfalz OSM environments consist of just 221 different tags. This may point to a more homogeneous structure of the geographic context information provided by OSM when compared to casual citizen science observations of organisms. This is to be expected, especially for urban, or more generally, settled areas: their inventory of real-world elements to be mapped in OSM is basically the same everywhere within each area of interest, and probably to some extent even similar between areas of interest. The OSM project is designed to work worldwide, and most of its tags are not specific to a certain country, region, or city. Mainz and Montabaur both feature residential streets tagged “highway=residential”, as do Monterey and Modesto, although their natural settings and city structures may be different. More differences can be expected for rural and natural areas, but information density in OSM is also much lower there (see section 2.1.3). An example for a difference in natural factors between the two data use cases considered here is the prominence of “intermittent=*” tags in California OSM data vs. those from Rheinland-Pfalz, speaking of seasonally dry conditions in large parts of California, which are not present in Rheinland-Pfalz.

Observed communities were found to be more similar among species of the same species group, and especially so if that species group contained predominantly species specialized in similar habitats, for instance, dragonflies and damselflies, or (mostly marine) mollusks. This effect is no longer traceable when OSM environments are used. ArtenFinder OSM environments of dragonflies and damselflies show a mean Simpson similarity among one another of 0.67, only slightly higher than the value for birds at 0.62, which are more mobile and far more varied in their habitat preferences. iNaturalist OSM environments have an average Simpson index value of 0.70 among birds, and a value of 0.56 among (mostly marine) mollusks, which is even lower although most species in the latter group share very similar habitats. This hints at a fundamental shift in factors causing OSM environments to be similar or different from one another, when compared to the factors which cause observed communities to be similar or different from one another. Basically, OSM environments can be expected to be similar if their target species are predominantly observed in places where OSM is similar. The next section examines tag composition of OSM environments in detail.

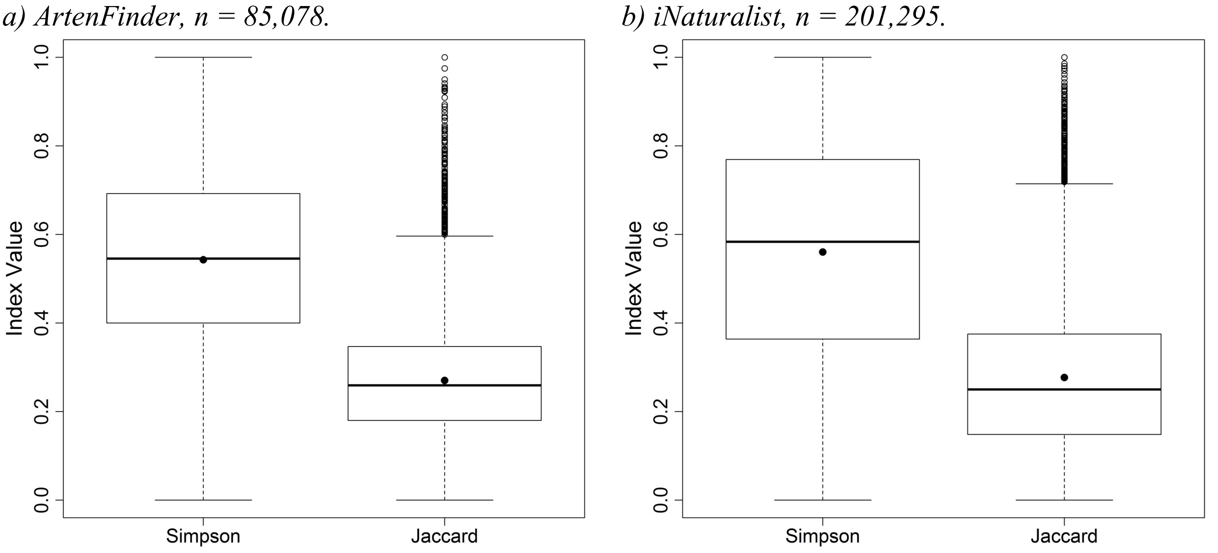


Figure 5.2.1: ArtenFinder and iNaturalist data, Simpson and Jaccard similarities of OSM environments with >= 10 tags among one another.

5.2.2 Tag Composition of OSM Environments

Species composition of observed communities was found to be determined both by the target species' habitat preferences and thus natural distribution and associations, and by VGI-related factors, especially spatial clustering of observations of the same species group due to volunteer preferences in the observation process. Natural species distribution may have some influence also on tag composition of OSM environments, because it may coincide at least to some degree with elements of geographic context represented in OSM data. Mean rates of keys in all OSM environments (Table 5.2.1 and Table 5.2.2) mirror the OSM data's properties, which were presented in section 2.1.3, to some extent: keys with high tag counts or with high area or length sums dominate. "highway=*" tags, which include all kinds of traffic routes, from large highways to footpaths, lead in both average compositions of OSM environments, showing that most observations are made in areas which are accessible by some kind of "highway". "landuse=*" and "natural=*" tags with their large area sums, and "surface=*" tags with their relatively large count and length sums, are also prominent in OSM environments of both areas of interest. In both use cases, "building=*" tags, although dominant in the OSM data used (see section 2.1.3), are not very frequent within OSM environments, on average. This is mostly because the most prominent tag with this key, "building=yes", is a nonspecific tag (see next section, Table 5.2.6 and Table 5.2.7) which was removed from OSM environments.

Table 5.2.1: Average composition of Rheinland-Pfalz OSM environments, aggregated by keys.

Key	**Mean rate (%)**
highway	28.3
landuse	14.7
surface	11.2
amenity	9.6
building	8.2
leisure	6.2
natural	5.5
barrier	5.1
sport	4.7
waterway	2.5
place	2.2
man_made	1.4
aeroway	0.2
water	0.1
military	0.1

Table 5.2.2: Average composition of California OSM environments, aggregated by keys.

Key	**Mean rate (%)**
highway	36.0
natural	16.0
surface	13.4
landuse	9.6
leisure	8.7
amenity	7.5
building	5.6
place	1.7
waterway	0.6
intermittent	0.5
water	0.3
aeroway	0.1
wetland	0.1

How far can landscape elements governing natural species distribution be traced in OSM environments? As we have seen in section 5.1.1, this question can be discussed by looking at some examples of target species which are often present, and therefore often observed, in certain geographic contexts. The example species used for discussing observed communities' species composition in section 5.1.2 mostly have too small OSM environments (numbers of tags smaller than 10) and are thus unsuitable for discussion here, because they were not actually used in evaluation. Therefore, comparable examples of other target species were selected for the present discussion. Table 5.2.3 presents the Rheinland-Pfalz OSM environments of three species associated to aquatic habitats. Eurasian Moorhen (*Gallinula chloropus*) and Small Red-eyed Damselfly (*Erythromma viridulum*) can be expected to be mostly observed on or close to ponds, small lakes or other stagnant or slowly flowing water bodies with well vegetated banks and submerged plants (Rößner et al. 2013; ArtenInfo Rheinland-Pfalz[51]). However, their OSM environments do not feature any tag fitting these habitat preferences, although tags such as "natural=lake", "water=lake", "water=pond" or "landuse=pond" are available in OSM, and

[51] https://arteninfo.net/elearning/libellen/speciesportrait/265, last accessed on 2018-08-04

were part of the specific tag list used in this work. Obviously, these and other tags which could be directly connected to the target species' habitat preferences do not occur frequently enough in OSM data in these species' observations' neighborhoods. Rather, the cited OSM environments consist of tags describing a quite unspecific context of infrastructure elements and a few natural elements such as "natural=wood" or "water=river". To derive, from these OSM environments, hints towards the two species' distribution would be speculative at best. They seem to imply that the two species are often observed close to non-automobile transportation infrastructure ("highway=bus_stop", "highway=cycleway"), and in the periphery of urban or settled areas ("landuse=allotments", "landuse=industrial", "sport=soccer", "sport=tennis"), but other interpretations are certainly possible. A somewhat similar case is presented by the OSM environment of Black-headed Gull (*Larus ridibundus*), see Table 5.2.3. Again the bulk of tags is made up of unspecific transportation infrastructure, complemented here with a body of "building=*" tags. However, in this case, the presence of the tags "waterway=river" and "waterway=riverbank" reveals the species' attachment to large water bodies such as rivers (Rößner et al. 2013).

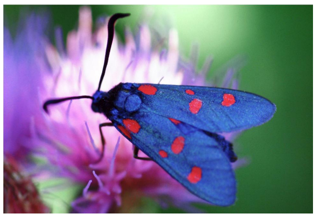
Table 5.2.3: ArtenFinder, examples of OSM environments: Eurasian moorhen, Black-headed Gull and Small Red-eyed Damselfly.

Target Species	**Eurasian Moorhen (*Gallinula chloropus*)**	**Black-headed Gull (*Larus ridibundus*)**	**Small Red-eyed Damselfly (*Erythromma viridulum*)**
Photos: C. Jacobs			

OSM environment (Rheinland-Pfalz)	barrier=fence
	highway=bus_stop
	highway=cycleway
	highway=living_street
	highway=steps
	highway=turning_circle
	highway=unclassified
	landuse=allotments
	landuse=industrial
	leisure=playground
	natural=wood
	sport=soccer
	sport=tennis
	surface=paving_stones
	tunnel=yes
		amenity=shelter
		barrier=fence
		barrier=wall
		building=garage
		building=house
		building=residential
		building=roof
		highway=bus_stop
		highway=crossing
		highway=cycleway
		highway=living_street
		highway=steps
		highway=turning_circle
		highway=unclassified
		landuse=commercial
		landuse=grass
		landuse=industrial
		leisure=park
		leisure=playground
		natural=tree
		railway=rail
		sport=soccer
		surface=concrete
		surface=paving_stones
		surface=unpaved
		tunnel=yes
		waterway=river
		waterway=riverbank
			barrier=fence
			highway=bus_stop
			highway=cycleway
			highway=steps
			highway=unclassified
			landuse=allotments
			landuse=farmyard
			landuse=industrial
			leisure=playground
			sport=soccer
			surface=unpaved
			tunnel=yes
			waterway=river

It is obvious from these examples, that OSM environments do not lend themselves as easily to be interpreted from the perspective of the target species and its properties, as do observed communities. These examples show that composition of OSM environments is often rather dominated by data properties of OSM itself: “highway=*” tags, which have relatively high tag counts (see section 2.1.3), are also the most prominent bodies of tags in many OSM environments. However, some examples exist where the OSM environment also shows visible influence of the target species’ habitat preferences. It was already glimpsed in Black-headed Gull. Another more distinct example is presented by Grayling (*Hipparchia semele*), a butterfly living in warm and dry places with sandy or rocky substrates. They can be found on arid grassland, in clearings and forest aisles in dry forests, or in quarries[52]. Due to these specific habitat preferences, the species is rare in Rheinland-Pfalz, and provides just 31 accepted observations for OSM environment extraction. Table 5.2.4 shows that Grayling’s Rheinland-Pfalz OSM environment features many elements which point directly to the species’ habitat preferences, such as “man_made=cutline” (‘cutline’ is the term used for forest aisles in OSM[53]), “natural=bare_rock”, “natural=grassland” (used only for non-cultivated areas[54]), “natural=heath”, and “sport=climbing” (the latter pointing at rock formations used for climbing). “building=*” and “highway=*” tags are almost completely missing here, which shows that this species is observed mostly away from settlement areas and the usual access routes. However, such examples of OSM environments reflecting specifically the species’ habitat preferences are hard to find. Another example may be presented by Five-spot Burnet (*Zygaena trifolii*), a small, red-and-black moth mostly occurring on wet meadows[55]. Its OSM environment (Table 5.2.4) features tags pointing to rural settings (“amenity=hunting_stand”, “amenity=shelter”, “natural=peak”, “natural=spring”, “surface=dirt”, “surface=unpaved”) and to meadows (“landuse=grass”, “natural=grassland”).

Table 5.2.4: ArtenFinder, examples of OSM environments: Grayling and Five-spot Burnet.

Target Species	**Grayling (*Hipparchia semele*)**	**Five-spot Burnet (*Zygaena trifolii*)**
	Photo: iNaturalist, © bferrero, some rights reserved (CC-BY-NC), cropped	 Photo: iNaturalist, © Dolors Bas Vall, some rights reserved (CC-BY-NC), cropped

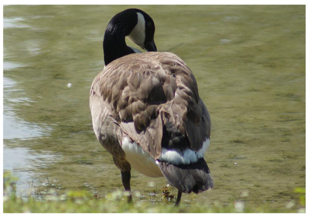
OSM environment (Rheinland-Pfalz)	highway=emergency_access_point
	man_made=cutline
	military=bunker
	natural=bare_rock
	natural=grassland
	natural=heath
	natural=peak
	natural=spring
	sport=climbing
	surface=concrete
	surface=unpaved
		amenity=hunting_stand
		amenity=shelter
		barrier=fence
		highway=emergency_access_point
		landuse=grass
		natural=grassland
		natural=peak
		natural=spring
		surface=dirt
		surface=unpaved
		tourism=picnic_site

Closer examination of California OSM environments renders comparable results. Table 5.2.5 presents some examples. Canada Goose, already used as an example in observed communities discussion (sec-

[52] https://arteninfo.net/elearning/tagfalter/speciesportrait/1696, last accessed on 2018-08-05
[53] https://wiki.openstreetmap.org/wiki/Tag:man_made=cutline, last accessed on 2018-08-04
[54] https://wiki.openstreetmap.org/wiki/Tag:natural=grassland, last accessed on 2018-08-04
[55] https://arteninfo.net/elearning/nachtfalter/speciesportrait/1566, last accessed on 2018-08-23

tion 5.1.2), has an OSM environment featuring mostly “highway=*” and “surface=*” tags, only “natural=water” hinting at a frequent association with aquatic habitats. Sea-fig or Freeway Iceplant (*Carpobrotus edulis*), a neophyte in California which is found exclusively in a number of coastal habitats[56] reflects its spatial distribution in its OSM environment by featuring the tags “natural=beach” and “natural=coastline”, but the list is still dominated by other tags. Finally, Sea Clown Triopha (*Triopha catalinae*), a marine mollusk living mostly on rocky coastlines, in tide pools and in kelp beds (Alden et al. 1998,) has a more distinct OSM environment featuring a number of tags fitting its habitat preference, such as “leisure=nature_reserve”, “man_made=pier”, “natural=coastline”, “place=islet” and maybe also “tourism=viewpoint”, while “highway=*” tags are missing. A number of other marine organisms could be cited here, which have similarly specific OSM environments.

Table 5.2.5: iNaturalist, examples of OSM environments: Canada Goose, Sea-fig and Sea Clown Triopha.

Target Species	**Canada Goose (*Branta canadensis*)**	**Sea-fig (*Carpobrotus edulis*)**	**Sea Clown Triopha (*Triopha catalinae*)**
	 Photo: C. Jacobs	 Photo: C. Jacobs	 Photo: iNaturalist, © Alison Young, some rights reserved (CC BY-NC), cropped

OSM environment (Rheinland-Pfalz)	amenity=school
	barrier=gate
	bridge=yes
	highway=cycleway
	highway=secondary
	highway=tertiary
	highway=turning_circle
	highway=unclassified
	landuse=residential
	leisure=pitch
	natural=water
	surface=asphalt
	surface=paved
		barrier=gate
		bridge=yes
		highway=bus_stop
		highway=secondary
		highway=steps
		highway=stop
		highway=tertiary
		highway=turning_circle
		highway=unclassified
		landuse=residential
		leisure=pitch
		natural=beach
		natural=coastline
		natural=water
		surface=asphalt
		surface=paved
			barrier=fence
			barrier=gate
			landuse=industrial
			landuse=military
			leisure=nature_reserve
			man_made=pier
			natural=coastline
			natural=scrub
			natural=water
			natural=wood
			place=islet
			surface=dirt
			surface=unpaved
			tourism=viewpoint

Summarizing these findings, OSM thematic data properties are the dominant factor determining the composition of OSM environments, making them more similar to one another, than are observed communities. Examples can be found where effects of a target species’ properties (especially habitat preferences) on an OSM environment’s tag composition are directly visible, but they are relatively rare. One of the reasons leading to these findings might be found in OSM’s semantic data structure. For many real-world elements, several tags are available, such as in the example of ponds and lakes used above. The available context information is therefore distributed over several tags, none of which is able to reach high enough frequencies to make it into the OSM environments. It might therefore make sense to conflate such tags semantically, and to create from them tag collections representing relevant properties of the environment. However, this would have to be done with great care, and possibly in different ways for individual species.

[56] Jepson Flora Project (eds.) 2018. Jepson eFlora, http://ucjeps.berkeley.edu/eflora

5.2.3 Nonspecific Tags

Analog to the observed communities approach, where nonspecific species were identified and removed from observed communities, so-called nonspecific tags were identified and removed from OSM environments. These were tags which occurred in 50% or more of the OSM environments and were therefore considered not to contribute to the distinctness of these lists of tags. Table 5.2.6 lists the 25 tags identified as nonspecific in Rheinland-Pfalz. The list has mostly "highway=*" (28.0%), "surface=*" (20.0), and "landuse=*" (16.0%) tags, mirroring high occurrence of these tags in the OSM dataset used. One highway tag ("highway=track") even occurred in all OSM environments. Some others, such as "highway=path" and "building=yes", came close. Even the tag "landuse=forest" reached more than 98% frequency in OSM environments from Rheinland-Pfalz OSM data.

Table 5.2.6: Rheinland-Pfalz, tags identified as nonspecific tags. (Tags occurring in 50% or more of OSM environments at the same time). Ordered bescending by frequency of occurrence in OSM environments.

Tag	**Frequency in Rheinland-Pfalz OSM environments (%)**
highway=track	100.0
highway=path	99.3
building=yes	99.1
landuse=forest	98.3
surface=asphalt	96.8
highway=service	95.9
waterway=stream	95.3
amenity=parking	92.9
landuse=meadow	90.9
landuse=residential	90.7
highway=residential	90.4
surface=gravel	90.1
bridge=yes	89.8
amenity=bench	88.7
surface=paved	88.3
surface=ground	84.1
highway=footway	84.0
landuse=farmland	81.0
surface=grass	80.6
natural=scrub	74.9
highway=secondary	74.1
natural=water	62.6
leisure=pitch	61.4
highway=tertiary	61.0
barrier=gate	53.7

In contrast to evaluation of the observed communities approach with iNaturalist (California) data, which did not render any nonspecific species, evaluation of the OSM environments approach in California resulted in some nonspecific tags, see Table 5.2.7. "highway=*" tags dominate among them (reflecting high numbers of these tags in the OSM data used here). Frequencies of individual tags in valid OSM environments start much lower than in Rheinland-Pfalz, at 85.0% ("highway=service").

However, almost all tags in this list are also found in Rheinland-Pfalz' nonspecific tags, except "leisure=park", which is unique to the California list.

Table 5.2.7: California, tags identified as nonspecific tags. (Tags occurring in 50% or more of OSM environments at the same time). Ordered by frequency of occurrence in OSM environments (desc.).

Tag	**Frequency in California OSM environments (%)**
highway=service	85.0
highway=residential	81.3
highway=path	76.6
highway=track	71.4
building=yes	65.8
amenity=parking	63.9
waterway=stream	62.5
leisure=park	58.4
highway=footway	58.0

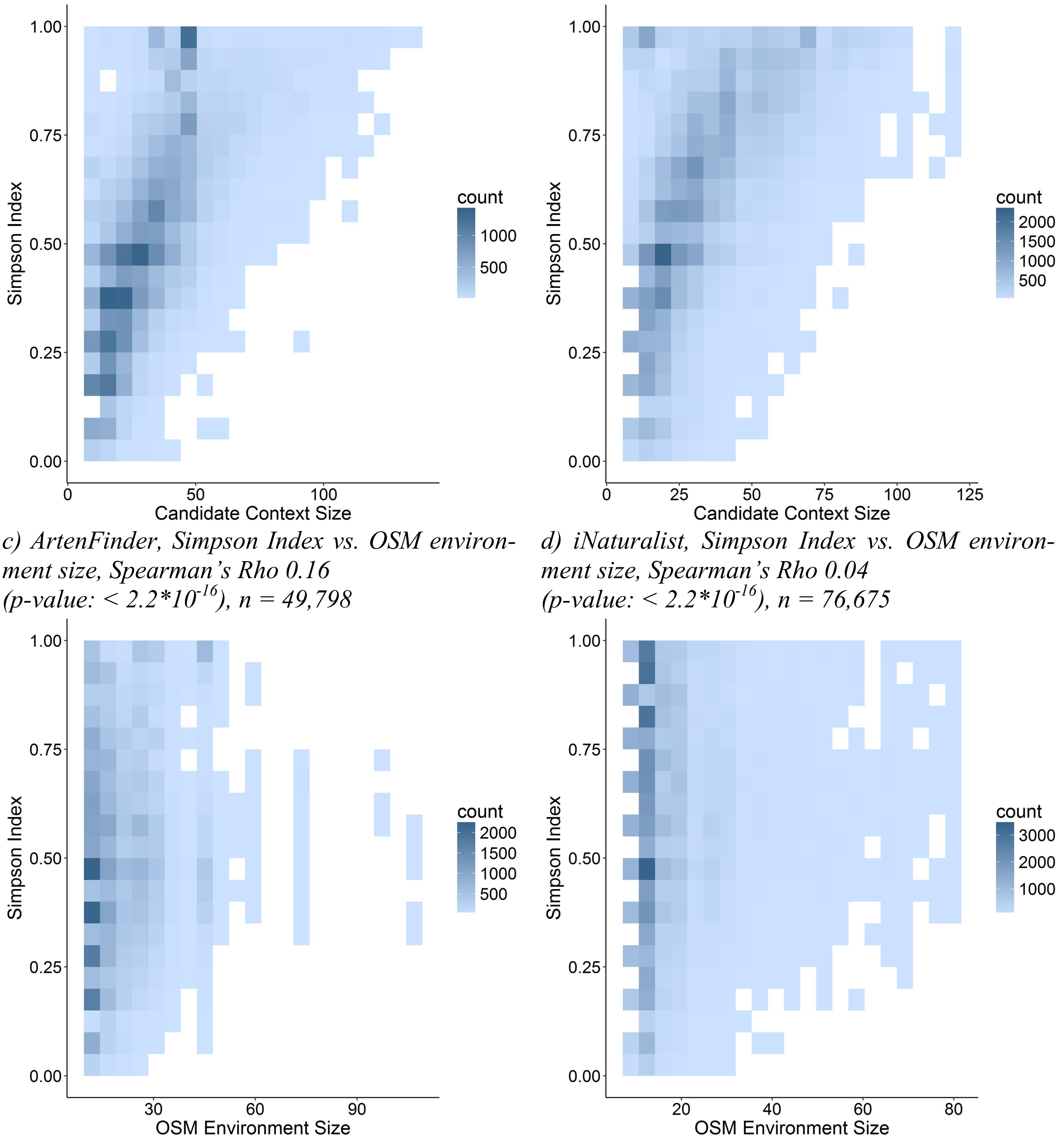
5.2.4 Differences in Similarity Values Between Sets of Candidate Observations with OSM Environments

The general findings discussed for differences in similarity values between sets of candidate observations in the observed communities approach (section 5.1.4) also hold for the OSM environments approach. Results presented in section 4.2 show, however, that differences in distributions of similarity values from sets of plausible or implausible observations (both real and synthetic) are in some cases smaller than in evaluation of the former approach. This is especially true for differences of plausible sets to synthetic implausible sets, the latter's distributions of similarity values having shifted notably towards higher values, when compared to their observed communities counterparts. In these sets (_SI2 and _SI3), synthetic observations where placed at random in places away from existing, plausible observations. With OSM data and their more homogeneous thematic properties, this procedure has a higher probability of placing a synthetic candidate observation in a position which is similar to situations in which the species is usually observed, concerning its OSM context. Thus, although these synthetic candidates are often placed in regions with low observation density (see Figure 4.2.3 and Figure 4.2.6), and therefore often further away from plausible observations than they were in the observed communities experiment, their similarity results are more similar to those of plausible sets than in the observed communities case. This is another proof of the fact that OSM data, as a context source, have different spatial properties than the observation data themselves.

Some of the distributions of similarity values show peculiarities due to their VGI origin, which was already found in observed communities evaluation results. The similarity value distributions of AF_SI1 show a peak at ca. 0.8 (Simpson similarity) and at ca. 0.4 (Jaccard similarity), caused by a body of ca. 240 candidate observations (mainly of Cirl Bunting and Marsh Tit) with relatively high Simpson and Jaccard similarity values which were contributed by a single observer, all from the same location. iNat_SP Jaccard similarity values show a secondary peak at relatively high values of ca. 0.6-0.9, corresponding to the high-similarity iNat_A candidate observations they are based on (see chapter 3.3.2). Again, this bimodality leads to a rather elongated boxplot compared to those of the other similarity distributions. A Boxplot is, strictly speaking, not a suitable way of visualizing this distribution, while the kernel density plot used in Figure 4.2.4 conveys a better picture.

5.2.5 Effects of Spatial Distribution of OSM Data on Similarity Values

Observed communities evaluation results exposed effects of the spatial properties of observation data on similarity values (section 5.1.5). As the OSM environments approach uses identical similarity indices, these effects are also present in similarity values with the OSM environments approach. However, with OSM data as an extrinsic source of geographic context, there are important consequences concerning the interpretation of these effects when using similarity values as plausibility indicators.



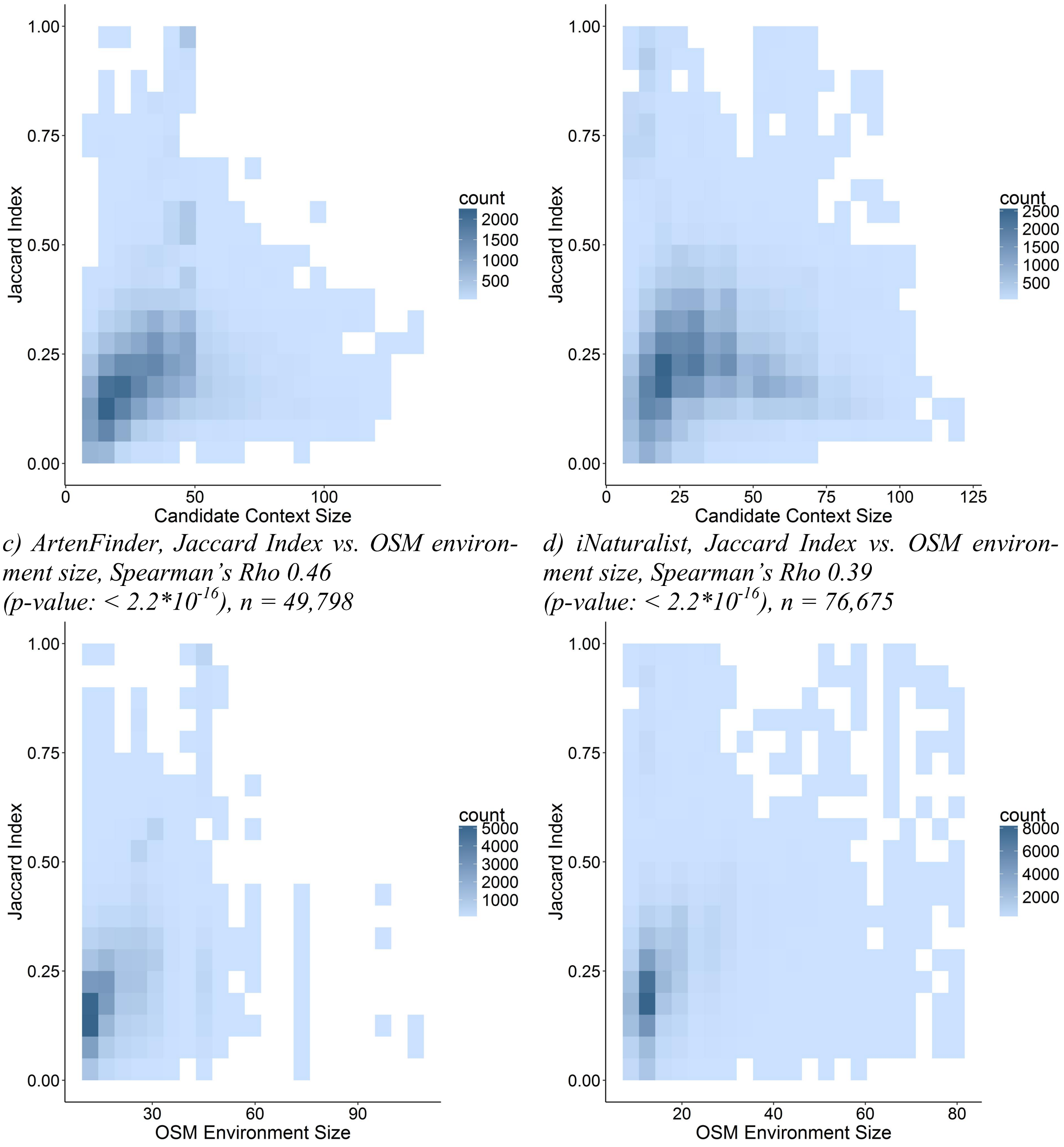
a) ArtenFinder, Simpson Index vs. candidate OSM context size, Spearman's Rho 0.73 (p-value: < $2.2*10^{-16}$), n = 49,798

b) iNaturalist, Simpson Index vs. candidate OSM context size, Spearman's Rho 0.63 (p-value: < $2.2*10^{-16}$), n = 76,675

c) ArtenFinder, Simpson Index vs. OSM environment size, Spearman's Rho 0.16 (p-value: < $2.2*10^{-16}$), n = 49,798

d) iNaturalist, Simpson Index vs. OSM environment size, Spearman's Rho 0.04 (p-value: < $2.2*10^{-16}$), n = 76,675

Figure 5.2.2: ArtenFinder and iNaturalist data, Simpson index values vs. candidate OSM context size and OSM environment size (no. of tags).



a) ArtenFinder, Jaccard Index vs. candidate OSM context size, Spearman's Rho: 0.43 (p-value: < $2.2*10^{-16}$), n = 49,798

b) iNaturalist, Jaccard Index vs. candidate OSM context size, Spearman's Rho 0.09 (p-value: < $2.2*10^{-16}$), n = 76,675

c) ArtenFinder, Jaccard Index vs. OSM environment size, Spearman's Rho 0.46 (p-value: < $2.2*10^{-16}$), n = 49,798

d) iNaturalist, Jaccard Index vs. OSM environment size, Spearman's Rho 0.39 (p-value: < $2.2*10^{-16}$), n = 76,675

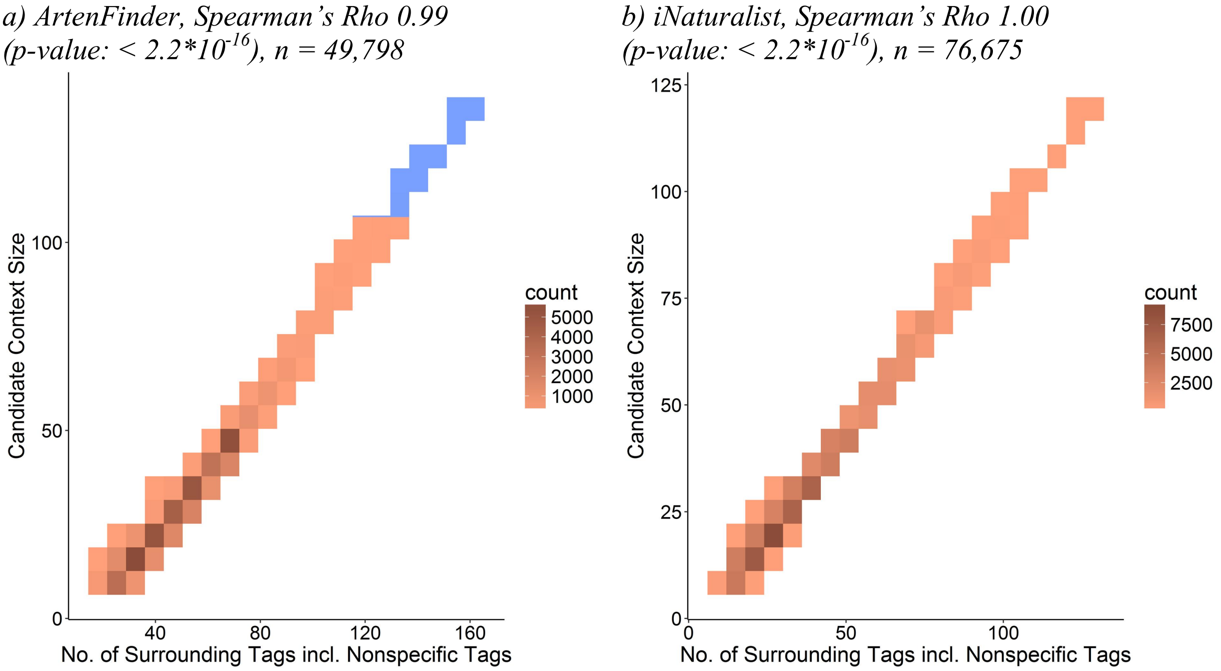
Figure 5.2.3: ArtenFinder and iNaturalist data, Jaccard index values vs. candidate OSM context size and OSM environment size (no. of tags).

Simpson similarity values show a positive correlation with numbers of tags in candidate OSM contexts (Figure 5.2.2). If a candidate observation is situated in a place where many different OSM tags can be found in the relevant neighborhood, this raises chances for a higher similarity value with the target species' OSM environment. Such observations have higher chance to appear more plausible in light of Simpson similarity. There is no pronounced correlation of Simpson index values with OSM environment size. Correlation of candidate OSM context size with ArtenFinder Jaccard index values is lower than with Simpson index values, and there is a correlation of Jaccard index values with OSM environments size on about the same level (Figure 5.2.3). Jaccard index values with iNaturalist data are only very weakly associated with candidate OSM context size. However, the findings here are basically analog to results obtained with the observed communities approach. The Simpson index correlates with candidate context size, which is due to the asymmetry of sizes of OSM environments and candidate context, the former being usually smaller, in 74.1% of ArtenFinder cases, and in 82.5% of iNaturalist cases. The Jaccard index, using the union of OSM environment and candidate context as its denominator, shows weaker correlations with both candidate context size and OSM environment size.

What do these findings mean for the use of similarity index values as indicators of plausibility of a casual citizen science observation if OSM provides geographic context, instead of observed communities? Observation data of organisms and OSM data come from completely separate projects. Although the basic factors governing data collection are similar to a high degree (as was demonstrated in section 2.2), the actual data collection processes are not connected to one another. In the observed communities case, a candidate observation which is located in a situation with a high observation density around it (which may lead to a larger candidate context, raising the probability of a higher Simpson similarity) very properly has a higher probability of a higher plausibility estimation, because it is situated in a place where it is per se more likely for an observation to come from. However, this rationale does not hold if an extrinsic source of geographic context is involved, whose spatial properties are governed by completely independent processes. If a candidate observation is placed in a situation which renders a large number of tags, this will raise the probability of a higher plausibility evaluation for the candidate, which may lead to a bias in plausibility estimation based on the Simpson index. The Jaccard index, with its weaker association between index values and candidate OSM context size, suffers less from this effect, and seems to be, from this perspective, a more suitable index for plausibility estimation for casual citizen science observations of organisms with the OSM environments approach. However, it also does show positive correlation with candidate OSM context size and with OSM environment size.

Detailed analysis of evaluation results with the observed communities approach suggested observation density around candidate observations as a factor influencing candidate context size and observed community size (see section 5.1.5). An analog factor cannot, however, be determined in the OSM environments approach. There is no proper spatial “tag density” in OSM data which would be comparable to observation density. The OSM environments approach uses the tag information attached to OSM objects to characterize geographic context. There is an n to n relationship between OSM objects and tags, which was already discussed when OSM data use cases were described in this work (see section 2.1.3): an object may (and usually does) carry several tags, and an object may also be segmented into several parts all carrying the same tag or tags. Instead of using spatial object density, a possible substitute might be OSM information density, that is, the number of different tags found around an observation (including nonspecific tags). For the reasons reiterated above, this parameter was already used to characterize the spatial properties of OSM data within the respective areas of interest (see section 2.1.3).

A candidate OSM context is, of course, basically the list of different OSM tags found around a candidate observation, from which only the nonspecific tags were removed. Unsurprisingly, candidate OSM context size therefore correlates to a high degree with OSM information density around a candidate (see Figure 5.2.4). Therefore, Simpson similarity has a higher probability of a higher value if a candidate observation is situated in a neighborhood with a high number of different OSM tags.

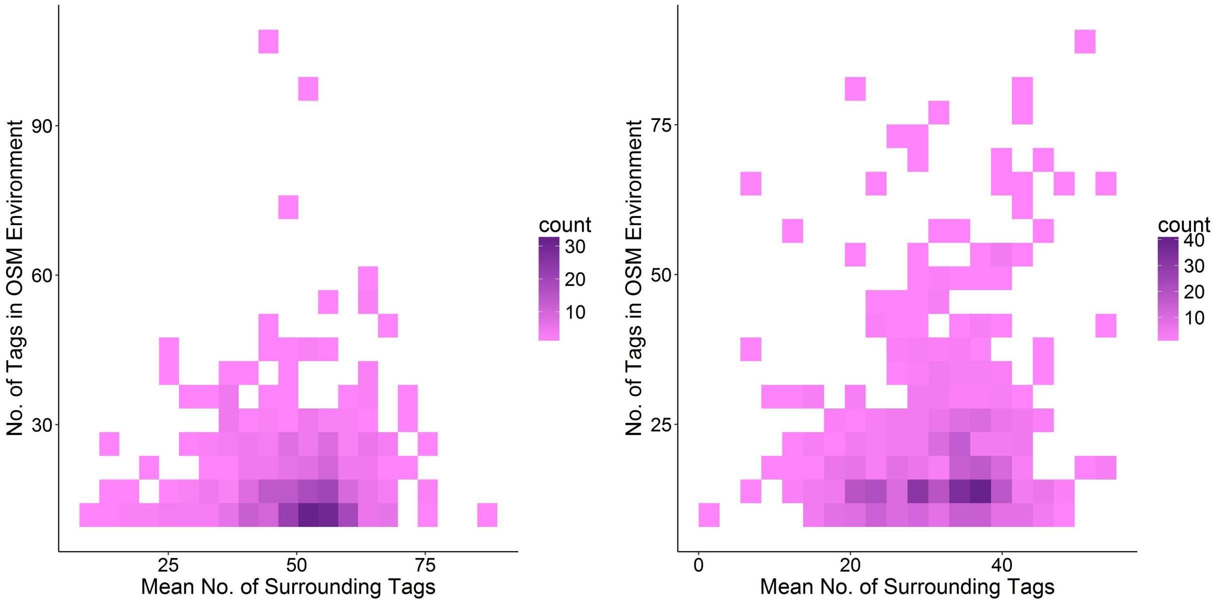


a) ArtenFinder, Spearman's Rho 0.99 (p-value: < $2.2*10^{-16}$), n = 49,798

b) iNaturalist, Spearman's Rho 1.00 (p-value: < $2.2*10^{-16}$), n = 76,675

Figure 5.2.4: ArtenFinder and iNaturalist data, candidate OSM context size (no. of tags) vs. no. of surrounding tags (OSM information density).

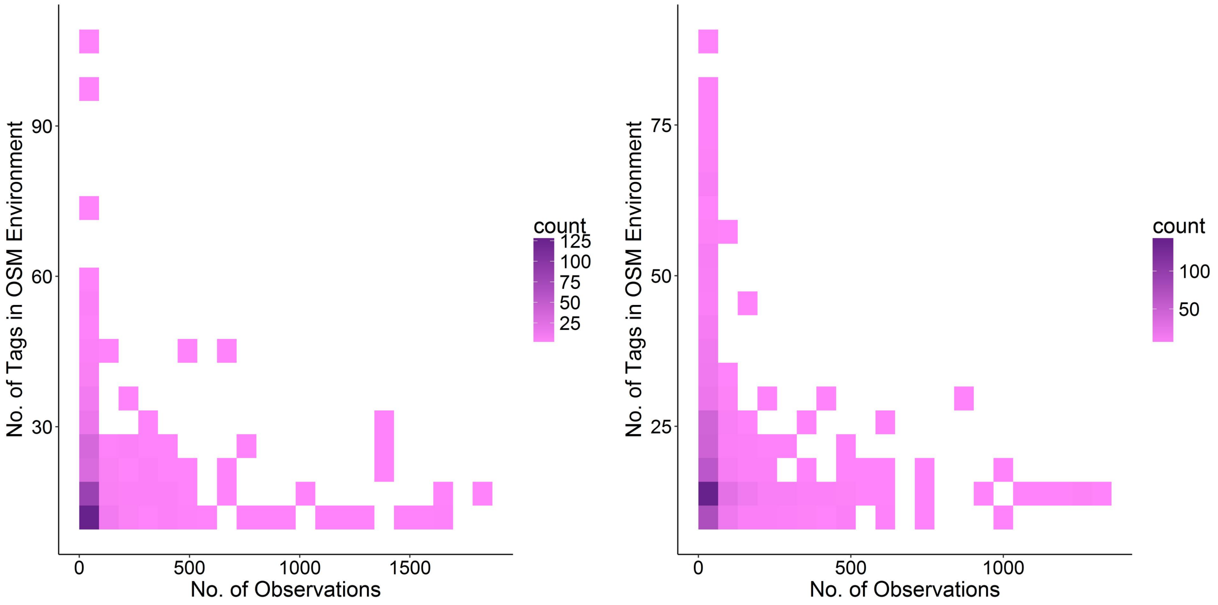
What are the factors which influence OSM environment size? Here similar parameters come to mind which were already examined for the observed communities approach. On the one hand, this is OSM information density around observations of target species: more diverse OSM contexts around target species observations (which are used for OSM environment extraction) may lead to larger OSM environments. On the other hand, there is the number of observations of target species used for OSM environment extraction: more target species observations represent more OSM context situations, which might lead to larger OSM environments. In contrast to findings with the observed communities approach, Spearman's Rho does not measure a pronounced correlation between OSM information density (expressed in the mean number of different tags found around a target species' observations) and OSM environments size, see Figure 5.2.5. Differences in OSM environment sizes are therefore probably not influenced by spatial differences in how many different tags can be found around observations of a target species used for OSM environment extraction. A candidate observation of a target species whose earlier observations (used for OSM environment extraction) are predominantly situated in places with many different tags around them is not likely to appear more plausible when the Jaccard index is used as a plausibility indicator. Analog to the observed communities approach, the sizes of OSM environments also show no pronounced correlation (on the $p \leq 0.05$ level) with the number of observations used for extracting these OSM environments (Figure 5.2.6).



a) ArtenFinder, size of OSM environments vs. mean numbers of different tags surrounding observations used for OSM environment extraction, Spearman's Rho -0.02 (p-value 0.6187), n = 402

b) iNaturalist, size of OSM environments vs. mean numbers of different tags surrounding observations used for OSM environment extraction, Spearman's Rho 0.12 (p-value 0.003051), n = 635

Figure 5.2.5: ArtenFinder and iNaturalist data, OSM environment size (no. of tags) vs. mean nos. of surrounding tags of target species observations up to 2015 (observations used for observed community extraction).



b) ArtenFinder, size of OSM environments vs. observation numbers of target species, Spearman's Rho -0.10 (p-value 0.04919), n = 402

d) iNaturalist, size of OSM environments vs. observation numbers of target species, Spearman's Rho -0.25 (p-value $1.627*10^{-10}$), n = 635

Figure 5.2.6: ArtenFinder and iNaturalist data, OSM environment size (no. of tags) vs. observation numbers of target species up to 2015 (observations used for observed community extraction).

5.2.6 Correlation of Simpson and Jaccard Index in the OSM Environments Approach

Analog to the observed communities approach, Simpson and Jaccard indices usually point in the same direction for the same candidate case (see Figure 5.2.7) also with OSM environments. Again, coincidences of a high Simpson index value with a low Jaccard index value for the same candidate case may occur, while opposite cases of a low Simpson index value with a high Jaccard index value may not. Reasons for this were explained in section 5.1.6. Association between the two indices is weaker here than in the observed communities approach, with more cases of contradicting index values for the same candidate.

a) ArtenFinder, Spearman's Rho 0.74 (p-value < $2.2*10^{-16}$), n = 49,798

b) iNaturalist, Spearman's Rho 0.60 (p-value < $2.2*10^{-16}$), n = 76,675

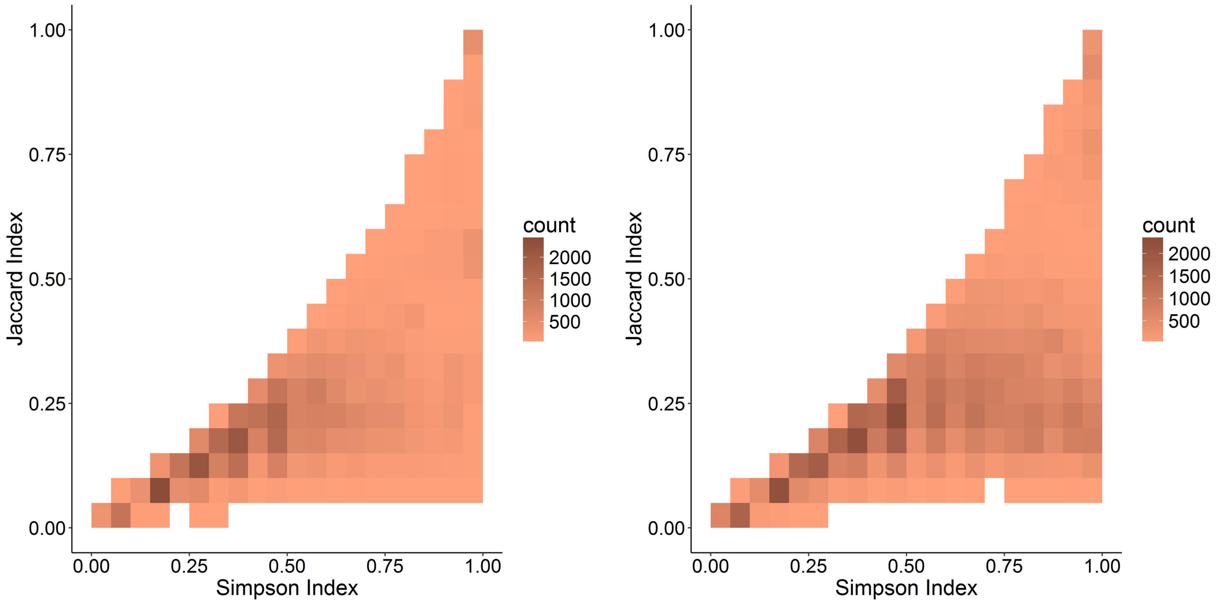


Figure 5.2.7: OSM environments approach, Simpson vs. Jaccard index values.

5.3 Discussing Results of Sensitivity Analysis

The sensitivity analysis which is discussed in this section, and which was conducted with different parameter settings and some methodological modifications, sheds light mostly on the behavior of the proposed observed communities approach to plausibility estimation for casual citizen science observations of organisms, because analysis was conducted mostly with this approach (see section 3.4). This behavior is determined by complex mechanisms taking effect if parameters are changed or methodological components added, and also by differences in data properties between the two data use cases the approach was applied to.

Most analyses rendered results where the relations between sets of plausible observations and sets of implausible observations did not critically change, and differences remained statistically significant, but there is typically a tradeoff between a raise in contrast between distributions of similarity values of plausible and implausible observations on the one hand, and a drop in effectiveness (numbers of valid observed communities and candidates) on the other hand. However, changes in differences between distributions of similarity values of plausible and implausible sets are generally small, and results are basically similar, in this aspect, for all parameters and methodological modifications. In other terms, the observed communities approach basically keeps working as a method for plausibility estimation of casual citizen science observations of organisms with many different parameter settings and methodological modifications. These settings and modifications can therefore be geared to specific properties of data use cases, or to specific knowledge-based preferences of domain experts involved. However, performance might suffer especially in terms of number of species and candidates which can be evaluated. The following subsection discusses the results obtained in sensitivity analysis in more detail.

5.3.1 Using a Lower Minimum Requirement for Approved Observations in Observed Community Extraction

Results with this parameter change show that a low minimum requirement for approved observations per target species in observed community extraction can be employed for raising the effectiveness of the observed communities approach, without deteriorating its ability to distinguish between plausible and implausible observations. Lowering this threshold allows for candidate observations of rarely observed target species to be evaluated by the approach. Among valid observed communities from ArtenFinder data, target species can now be found which are quite rarely occurring in the respective area of interest, such as Short-eared Owl (*Asio flammeus*), a rare passage migrant in Rheinland-Pfalz with just 10 accepted observations in ArtenFinder up to 2015. Their nocturnal mode of life may also be a factor in making them a rarely observed species. Also, species are now evaluated which are common and widespread, and easy to observe, but still rarely reported. For instance, the plant species Common Agrimony (*Agrimonia eupatoria*) is common and widespread in Rheinland-Pfalz, and therefore not in ArtenFinder's focus on protected and threatened species (see section 2.1.1). It was reported and approved just 10 times up to 2015. iNaturalist features similar examples. For instance, Black-capped Chickadee (*Poecile atricapillus*), a common passerine bird which, in California, occurs only in a relatively small area in the extreme northwest of the state[57], was observed and approved only 10 times in California up to 2015, but has now a valid observed community.

[57] https://www.audubon.org/field-guide/bird/black-capped-chickadee, last accessed on 2018-11-16

5.3.2 Using Variable Radii for Species Groups

Using a larger search radius for certain species groups aimed to obtain observed communities and candidate contexts which match these species' relevant neighborhoods better. Observed communities as well as candidate contexts were indeed larger on average with the radius configuration used here. Of course only observed communities of species belonging to groups with 2,000 m or 3,000 m radii changed here, while observed communities of species which still used a radius of 1,000 m remained unchanged. However, enlarging search radii for whole groups does not take into account the inherent variability in species properties within the groups. This can easily be demonstrated using two bird species examples which were already discussed above. Common Kingfisher's ArtenFinder observed community was now much larger (40 vs. 14 species), but also much more unspecific, because now only about half of the species could be accounted as being associated to water bodies, where Common Kingfisher is usually observed. Greylag Goose, usually present on lakes, ponds etc. and in adjacent fields, also had now a notably larger observed community (70 vs. 36 species). The rate of species which can be accounted as typically observable on lakes or in open fields also dropped. These two examples show that some target species which are mobile, but associated mostly to certain habitats, render observed communities with a larger search radius which are matching the target species' habitat preferences less well than before. The dragonfly species Southern Hawker's observed community was actually smaller in this experiment, than it was in the evaluation (with just 10 vs. a former 26 species), and remained dominated by species which can be expected to be observed on proximity to this target species, thus not confirming the effect found in the observed communities discussed above.

Observed communities of species with more unspecific habitat preferences already included other species with rather unspecific habitat preferences, as well as many more specialized species. This remained mostly the same. European Greenfinch's observed community (an example for a common species with relatively unspecific habitat preferences already discussed above) was now much larger as well (53 vs. 26), and still had a mix of more common and more specialized species, the latter with many different habitat preferences. The observed community of Silver Y (a common moth) was smaller with a larger search radius of 2,000m (25 vs. 31 species), but its properties remained similar.

iNaturalist showed similar trends: observed communities became generally larger, and observed communities of species with more specialized habitat preferences became more unspecific in the process. E.g., Canada Goose's observed community grew to 42 species and only about half of its species had a stronger association to water bodies or adjacent open fields. The observed community of Black Oystercatcher, grown from 29 to 93 species, also lost focus in the process, but managed to remain dominated by species associated to the coast. Seaside Daisy's observed community did not change in this analysis, because plants' search radius remained at 1,000 m. The common and widespread species Turkey Vulture now had a larger if still quite small observed community with nine species (besides Turkey vulture itself), composed of a wide variety of species, other common species as well as other ones more specialized in a variety of habitats. California Scrub-jay, with 32 species in its (now valid-size) observed community (was 7) also remained associated with a wide variety of species with different preferences.

These examples make it clear that enlarging the search radius for certain species groups does not introduce any advantage which might lead to a better performance of the approach. Species with a low mobility and/or a high specialization on certain habitats should be treated with relatively small search radii, the evaluation conducted above providing a good benchmark at 1,000 m. More common species with more unspecific habitat preferences, and/or more mobile species do not gain advantages from a larger search radius, either.

Summarizing these results, a larger search radius for certain species groups increases the diversity in ArtenFinder observed communities, which leads to a much higher number of nonspecific species causing a drop in the number of valid observed communities. Contrast between index value distributions is rather reduced. With iNaturalist data, the number of valid observed communities increases, while contrast between similarity index value distributions is also reduced. Enlarging the search radius for certain species groups in the way tested here does not overall improve the approach's ability to distinguish between plausible and implausible observations (although the change allows for including more species in the analysis in the case of iNaturalist data). Observed communities do not become more specific to the target species' properties.

5.3.3 Shifting Frequency Thresholds for Frequent Co-Observations and Nonspecific Species

Raising the frequency threshold for identifying frequently associated context species to the level tested here (0.75) has mostly negative effects on the performance of the observed communities approach, leading to lower numbers of valid species and candidate observations, and reducing differences between sets of plausible and implausible observations. Lowering this threshold has mixed effects on effectiveness in terms of species and candidate observations evaluated, as well as on evaluation results (differences between plausible and implausible sets).

Changing the threshold for identifying nonspecific species affects directly the number of these species, which bears directly on numbers of valid species and candidate observations. Results, however, remain quite stable with both a reduced and a raised threshold value.

5.3.4 Using Auxiliary Land Cover and Ecological Land Unit Information

Using auxiliary land cover geometries aimed to obtain more focused observed communities and candidate contexts in terms of the ecology of the target species. How did the observed communities react to the use of auxiliary land cover or ecological land unit information concerning their species composition? In the ArtenFinder use case, the mean size of observed communities dropped, while the number of observed communities remained constant. In the iNaturalist case, average size was raised with NLCD data, because many small valid observed communities were lost, and dropped with ELU data. Let us look again at species examples which were already discussed above. To start with ArtenFinder data combined with CORINE land cover, Common Kingfisher`s ArtenFinder observed community had now only five species left, most of which matched this target species' habitat preference. Greylag Goose was now exclusively associated to water fowl and other species associated to water, in a thus more focused observed community of just 12 species. Southern Hawker's observed community was just three species smaller than it was in evaluation of the approach, but its share of species which would be expected to be associated with this target species rather dropped. With ELU data, Common Kingfisher's four-species observed community was no longer dominated by species associated to water. Greylag Goose had 16 associated species when ELU data were used, most of which shared its habitat preferences. The ELU observed community of Southern Hawker mostly shared the properties of this target species' result with CORINE. Considering these examples, the effect of using CORINE or ELU data with ArtenFinder observations was quite mixed. However, some species did have ecologically more focused observed communities, albeit also much smaller ones, most of them dropping in size below the threshold of 10 species.

With NLCD data, Canada Goose's only remaining associated species was Mallard (*Anas platyrhynchos*), a common species of duck. With ELU data, it had six associated species, all birds and mostly water fowl. Black Oystercatcher, with NLCD data, also had a much reduced observed community with 10 species, almost all associated to the coast. With ELU data, which do not include marine habitats, its observed community was empty. Seaside Daisy was associated with just Seaside Wild Buckwheat (*Eriogonum latifolium*) when NLCD data were used; with ELU data, its observed community was also empty.

In summary, using auxiliary land cover or ecological land unit geometries mostly increases the differences between distributions of similarity values of plausible and implausible candidate sets with both ArtenFinder and iNaturalist data, and leads to more specific observed communities in some cases. However, with iNaturalist data the ensuing strong decrease in effectiveness makes this modification problematic, at least in the constellation of parameters used here.

5.3.5 Using a Quantitative Similarity Index

Looking at results obtained with the Similarity Ratio, frequency weighting might be beneficial to the performance of the observed communities approach with ArtenFinder data, while using the mean distance of context species observations did not change evaluation results. For the iNaturalist data use case, there was found no positive effect of using a quantitative similarity index with the parameters tested here on the observed communities approach's ability to distinguish between sets of plausible and implausible observations.

5.3.6 Edge Effects

Results of this experiment demonstrate that edge effects are minimal in evaluation results for both data use cases. The results suggest that using a guard zone around the state lines of Rheinland-Pfalz and California would not change evaluation results in any significant way. Working without edge effect correction holds the advantage of not losing candidate cases or target species observations at the edge of the area of interest. In practical use of the observed communities approach to plausibility estimation of casual citizen science observations of organisms, it should, however, be noted that the context of a single candidate observation may be severely influenced by its position close to the edge of the area of interest. Similarity values, or the plausibility estimations based on them, should be accompanied by a proper note in these cases.

5.3.7 Variant of the Simpson Index

The variant derived from the Simpson index which was proposed in section 3.4.7 has the advantage of providing a uniform interpretation of the index value in all candidate cases. Results with the variant were found not to differ much from evaluation results with the original Simpson index. In the context of this work, the proposed variant of the Simpson index therefore suggests an alternative to the original Simpson index.

5.3.8 Using Date-Specific OSM Context

Looking at the substantial drop in effectiveness through the loss of valid species and candidates (especially in the ArtenFinder use case) while results in terms of relations between distributions if similarity values remain basically the same, it can be said that results with date-specific OSM context do not advocate the use of this methodological modification. Taking the perspective of an individual species or candidate case, however, might change this conclusion. If, for instance, the neighborhoods of many prior observations of a target species experienced substantial changes which make an OSM environment based on the current state of OSM deviate strongly from the actual state at the dates of observation, a date-specific procedure would of course be beneficial. The same is true for the extraction of the OSM context of a candidate observation in a situation which changed (and was recorded accordingly in the OSM). Results obtained from analysis with date specific OSM context suggest, however, that this is not a common problem.

OSM history would, of course, provide the information necessary for identifying cases in which the OSM context changed. It would be possible, on this basis, to treat such cases differently from cases where no changes occurred, and to use the date-specific approach only on these. Difficulties may arise in discriminating cases with real-world changes reflected in OSM, from cases where just the OSM data changed (e.g., through addition of more tags). The former cases are the ones which should be treated with the date-specific approach, while the latter should certainly be placed in the context of the current OSM, avoiding difficulties such as a lack of completeness in older OSM data. These considerations might propose an interesting avenue for future research for improving the OSM environments approach.

5.4 Estimating the Plausibility of Candidate Observations Based on Observed Communities and OSM Environments

How can practical use be made of the approaches to plausibility estimation of casual citizen science observations of organisms presented in this work? Evaluation proved that for both similarity indices used, index values for plausible observations are usually higher than for implausible observations. Therefore, both approaches to plausibility estimation basically work. In this section, the step will be taken from looking at sets of observations and their distributions of similarity values, which was done to evaluate the approaches, towards looking at individual candidate observations and estimating their plausibility, which is the task the approaches are ultimately designed for. This is done by carefully discussing extent and limits of the approaches' indicative power concerning a candidate case's plausibility, in light of the approaches' functional properties and of evaluation results. This leads to general types of plausibility indications which may occur for certain candidate cases. These are then illustrated by looking at examples of target species and of individual candidate cases, taken from the evaluation results, which represent these case types. An approach to derive decision thresholds for plausible and implausible cases is also presented.

5.4.1 General Considerations on Indicative Power and Candidate Case Types

The observed communities approach and the OSM environments approach both estimate the plausibility of an observation by evaluating the match between the location of an observation and its species identification. They use two aspects of an observation. The first is the location, which determines the context of species observed around it, or of tags mapped around it. The second is the species identification, which determines which context species or context tags are expected, according to the species' observed community or OSM environment. In a case with a high level of similarity, the location of the candidate observation matches the species given by the observer. Its location is similar to places where the target species is usually observed, when the candidate observation's species or tag context is considered. This makes the location of the candidate observation appear plausible in light of the given species. In a case with a low level of similarity, the location does not match the species identification. Its location is not similar to places where the target species is usually observed. This makes the location of the candidate observation appear implausible in light of its species identification.

Discussion of evaluation results revealed that there are species whose observed communities or OSM environments are similar, because these species are observed in the same locations. Therefore, observed communities or OSM environments of target species observed in the same places tend to be similar. This finding is very important for how the observed communities approach and the OSM environments approach can be used to estimate the plausibility of the location of an observation in light of its species. A high similarity value shows that the candidate observation's location matches places where the target species is usually observed, but the candidate species might be any other species usually observed in the same places, or in places with similar species or tag context, and consequently having a similar observed community or OSM environment. An observation of one of these other species in the same location as the candidate observation would be just as plausible. Obviously, certainty of species identification is of no small importance here.

In a citizen science environment, certainty of species identification is determined by a number of factors. Of particular interest are species which can easily be mixed up with one another due to similar physical appearance, similar sound utterance, or otherwise similar features. Experience and training on the side of observers mitigates this problem, but casual citizen science observations of organisms are also provided by untrained and at least in part inexperienced observers. This fact was already used in this work to synthesize implausible observations by swapping species identifications of real observations of species which are often mixed up due to physical similarity, but live in different habitats which should provide different species or tag contexts. For the current discussion, species are also important which are easily mixed up, but are observed in the same places for whatever reasons, and therefore have similar observed communities or OSM environments. In candidate cases which are plausible, these are the species which are the most likely species alternatives at a given location. A high similarity value makes the candidate location appear plausible, but does not allow for ruling out that the species could be mixed up with a different species usually observed in similar places, especially a physically similar species. However, it is also implausible in this case that the species actually observed was a physically similar species usually observed in locations which have a different context. In implausible cases, either the location might be incorrect, or the species might be erroneous, and then all other species with an observed community which is similar to the candidate species' are also unlikely. It is, however, possible in these cases that the observation is in truth of a species which is physically similar to the species given in the candidate observation, but which is usually observed in different places. A test for this case could consist of calculating similarities between the candidate context at hand and the observed communities or OSM environment of the proper species. If a high similarity value would be found, it would point to a possible mix-up of these species.

All of these considerations, of course, do not apply for species which are easily and clearly identifiable, with no species physically similar enough for even an unexperienced observer to mix them up. Such species exist (examples will be discussed below) and produce cases where the species identification of a candidate is certain to a high degree. Another factor making a species identification certain (also for more difficult species) is a very experienced observer. If the observed community approach or the OSM environments approach find a high level of similarity of candidate context and observed community or OSM environment with such a candidate case, both species and location are plausible. For low similarities, the location is implausible, while the species is plausible, and if there is an error, it is more likely on the side of the location. A gross mix-up of species cannot, of course, be ruled out either, but is unlikely. An experienced observer will also lead to a high degree of certainty in species identification in an observation, at least for the species or species groups in which the observer specializes. This includes observations of species which are hard to identify or easily mixed up with physically similar species. In all cases where a low similarity value makes the candidate location appear implausible, it is always possible that the candidate observation comes from a location with a context which was so far not represented in the data, and therefore appears unusual when compared to the existing observations available for the candidate species.

Analog considerations apply for the reported location of an observation. The approaches to plausibility estimation discussed here can make a reported location appear plausible or implausible, but any similar location would result in a similar plausibility for an observation of the same species. This applies to observations with certain and uncertain species identification in the same way. Even if species identification is certain and context at the reported location matches the observed species' observed community or OSM environment, all locations with a similar species or tag context are just as plausible. With implausible observations, there are of course a very large number of alternative locations, because many different context situations may result in equally low similarity index results. This leads to the question of certainty of location in an observation. ArtenFinder and iNaturalist, like many other web-portals for reporting observations of organisms, provide map viewers to locate an observation by clicking on the map. Various types of base maps and other geographic information are usually available in these viewers. iNaturalist provides Google map and satellite images, while ArtenFinder has OpenStreetMap, aerial photographs, and also official topographic maps and relief. Uncertainty is therefore introduced by the variable ability of volunteers to correctly read such forms of geographic information, and to correctly place their observations on them. This factor of uncertainty is similar in nature to volunteer uncertainty in species identification, depending very much on the individual abilities of the volunteer. Also, when placing an observation on the map, observers can choose a scale by zooming in or out, thereby inherently selecting a level of precision in placing their observation. On a small scale, an observation will probably be placed accurately, but with relatively low precision in the correct area (e.g., in a certain part of a town, on a lake, or in a stretch of forest). On a larger scale, an observation may be placed much more precisely, e.g., in an isolated tree in which a bird was actually seen. In both ArtenFinder and iNaturalist, observers are free to choose the scale at which they place their observation in the map. This choice may or may not reflect the actual certainty of the observed position on the side of the observer. An observer may have a very precise recollection of the actual position of an observation, but may still use a small scale when actually submitting the observation, resulting in low precision. iNaturalist automatically generates an accuracy value depending on the chosen map scale, which can also be manually changed by the observer before submitting the observation. In iNaturalist also, an address search allows for placing an observation in a certain street address, introducing uncertainty of geocoding to accuracy, and uncertainty of the actual position used for an address in the geocoding process to precision. Both ArtenFinder and iNaturalist do not use user-defined observation areas, an option which is available in some other platforms (e.g., naturgucker.de), and a factor reducing precision, but usually not accuracy. Errors or inaccuracies in the maps and other sources of geographic reference used in a map viewer will, of course, also directly result in errors of observation location. With apps for reporting observations by using mobile devices (also provided by ArtenFinder and iNaturalist), uncertainty of geographic location is mostly determined by the technical properties of the Global Positioning System (GPS) sensor in the device used, as well as by the usual factors for GPS accuracy, such as satellite constellation and shadowing effects through infrastructure and vegetation (e.g., Zandbergen & Barbeau 2011).

Table 5.4.1: Considerations on plausibility of location and species identification for candidate cases.

High similarity of candidate context and observed community or OSM environment	**Low similarity of candidate context and observed community or OSM environment**
Reported location matches species identification. ➔ Reported location plausible for reported species. ➔ Reported location equally plausible for any species with a similar observed community or OSM environment (especially relevant: physically similar species). ➔ Reported location implausible for any species with a different observed community or OSM environment (including physically similar species). ➔ Any other location with matching species or tag context equally plausible.	Location does not match species identification. ➔ Reported location implausible for reported species. ➔ Reported location also implausible for any species with a similar observed community or OSM environment (including physically similar species). ➔ Large number of alternative locations equally implausible.

Table 5.4.1 summarizes the considerations described here, for the two main cases of a plausible or an implausible candidate observation. The information provided by the observed communities approach or by the OSM environments approach to plausibility estimation of casual citizen science observations of organisms is limited. Especially with plausible observations, it is important to keep in mind that there are often alternative species and locations which would render similar plausibility estimations. An implausible case may basically represent either an error in species identification or in position. The

latter is more likely with species which are usually identified with high certainty, or it may represent a correct but unusual observation. These considerations are overlaid by the influences of observation density on the similarity measures used here, which were discussed in detail earlier. They introduce another set of factors to this discussion, which are rooted in the VGI nature of the data. All of these considerations are best examined and illustrated by using examples from the data use cases at hand, first on the level of species and their observed communities, and then also on the level of individual candidate observations.

5.4.2 Observed Communities Examples: Illustrating Target Species and Candidate Cases

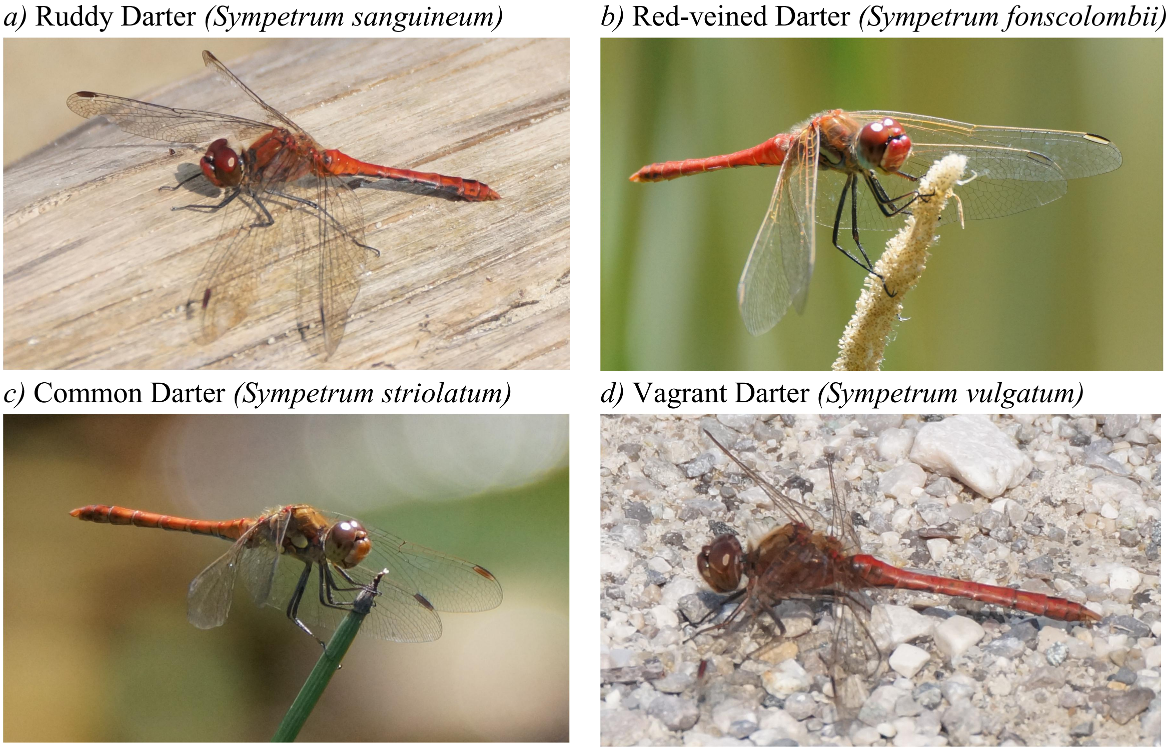
Common Kingfisher is a case of a species from the ArtenFinder data use case which is easy to identify, with no physically similar species which could easily be mixed up with it, either observable in the same locations, or elsewhere in Rheinland-Pfalz. Simpson similarities between the Common Kingfisher's observed community and those of all other species used in evaluation show that high Simpson similarities exist with species which share the Common Kingfisher's habitat, and which are therefore observed in the same places. They are mostly waterfowl, and dragonflies and damselflies. 94% of species whose observed communities have a Simpson index value of 0.8 or higher with the Common Kingfisher's belong to these two groups. Most of these also have relatively high Jaccard similarities, with some exceptions where observed communities are much larger than Common Kingfisher's. Low Simpson similarities of 0.2 or lower are found between the Common Kingfisher's observed community and observed communities of target species which are usually observed in different habitats. All of them also have low Jaccard similarities of under 0.13. Many butterfly species are found here, as well as many birds not attached to waterbodies. These two groups make up ca. 98% of species whose observed communities have a Simpson index value of 0.2 or lower with Common Kingfisher's. Some observed communities of species which share the Common Kingfisher's habitat but belong to different species groups exhibit medium Simpson similarities with the Common Kingfisher's observed community. Some dragonflies and damselflies can be counted among these cases, but also species with unspecific habitat preferences of other species groups can be found among these. There are also some cases which form exceptions from the rule. For instance, the 119-species observed community of European Hedgehog (*Erinaceus europaeus*) has a Simpson index value of 1.0 with the Common Kingfisher's 15-species observed community, although it is not especially found close to the latter's typical habitats. The large size of the European Hedgehog's observed community might be a reason for the fact that it completely covers the Common Kingfisher's much smaller observed community. However, although 87 species have inter-observed-community Simpson similarity values of 0.2 or lower with Common Kingfisher, it is hard to find a species among them which would clearly be expected to exhibit a higher similarity, so that the data used in this study do not provide an exception of this kind.

Next, we will look at some individual candidate cases of Common Kingfisher from the AF_A set. A real observation case with relatively high similarity values both for the Simpson index (at 0.87) and the Jaccard index (at 0.30) is presented by an observation (ID 54271696, observed on 2016-11-20) from a lake close to Neuburg on the Rhine in southeastern Rheinland-Pfalz. Observation density at this location is rather low: the candidate context has just 42 species from 247 observations within the 1,000 m search radius (the AF_A means for these values are 108.2 species and 1,225.4 observations). However, the candidate context still covers 13 species in the observed community: most species usually found around a Common Kingfisher observation were also observed around this candidate case. The similarity values thus both allow for evaluating this observation's location as plausible by the observed communities approach. As the candidate context is relatively small, the high Simpson index value is probably not boosted by candidate context size. The Jaccard index also reaches a relatively high value here due to the large intersection of candidate context and observed community, and because there are relatively few species in the candidate context which do not match the observed community. The species is easy to identify. There are other species with similar observed communities, but there is small danger of a mix-up even for unexperienced observers. So, in this case, place and species identification are plausible, and plausibility estimation by the observed communities approach supports the actual quality assurance decision made by the ArtenFinder experts. Of course, considerations concerning uncertainty of location apply: a small shift in the location which would not change the observed candidate context would also not change the resulting plausibility estimation, and other, more remote locations with similar context would also render similar plausibility estimations.

An example of a real observation which is implausible by the observed communities approach (although accepted by ArtenFinder experts) is found close to the village of Frankenstein, within the large forest area of the Pfälzerwald, but again on a lake (ID 54266476, observed on 2016-10-13). With a Simpson index value of 0.20 and a Jaccard index value of 0.03 both index values are low, indicating low plausibility of the location. Observation density around this location is also not high (with 260 observations within the search radius), but a higher diversity within these observations renders 102 species in the candidate context, which is about average for the AF_A set. However, only three of the species usually observed close to a Common Kingfisher observation are present in this context, making its location rather untypical in light of earlier observations of this species. Both similarity indices reflect this by giving low values. As Common Kingfishers are easy to identify, the species identification in this case is probably correct. There are various reasons for a location to appear implausible in this case. The location could have been recorded incorrectly. However, as the case at hand was accepted by ArtenFinder experts, it is more likely that it simply presents an unusual context which does not occur often enough in previous observation data to critically influence the composition of Common Kingfisher's observed community. For the same reason, the case is very properly evaluated as implausible by the observed communities approach, and thus identified as unusual. Similar plausibility estimations would result in many other possible locations with contexts equally dissimilar to Common Kingfisher's observed community.

So far, examples were discussed where both Simpson and Jaccard indices show accordant indications of plausibility, but this is not always the case. Let's look at another Common Kingfisher observation (ID 54269945, observed on 2016-10-31) on a lake close to Offenbach on the Queich, again in southwestern Rheinland-Pfalz. In this case, observation density is high, with 2,425 observations within the search radius, giving the candidate context a very high number of 308 species. They include all 15 species usually observed with Common Kingfisher. Simpson index value for this observation is therefore 1.0, the maximum possible value. However, Jaccard index is at 0.05 for this case, a low value (if not extremely low, for Jaccard index standards). Plausibility estimation is therefore divergent for the two similarity indices used. The Simpson index attaches a maximum plausibility to the location of this observation by the observed communities approach. The Jaccard index makes the location appear implausible. In this case, the high number of context observations (and consequently, high number of context species) makes a high intersection of candidate context and observed community more probable, giving this factor some weight in producing the high Simpson index value. For the Jaccard index value, considering in its calculation the very large number of species present in the candidate context which do not match the observed community, the same factor leads to a reduction of the index value. This seems to present a contradiction to the findings in section 5.1, where no negative correlation between Jaccard index values and candidate context size was found (but a moderate positive correlation). However, the Jaccard index is sensitive to the difference in size between the two lists compared, so that in the case at hand, the coincidence of a rather small observed community with a very large candidate context necessarily leads to a low Jaccard index value.

The above example of Common Kingfisher is of a species which cannot easily be mixed up with other species. Species identification can therefore be considered to be certain in most cases. If the observed communities approach to plausibility estimation indicates a high plausibility of a candidate's location, the combination of species and location appears plausible. With a contrarious plausibility indication, the location appears implausible while the species identification can still be regarded as certain. Now, what would be the interpretation of a plausibility indication with the observed community approach for candidate observations with uncertain species identification? Table 5.4.1 describes the consequences in general terms. Let's make an ArtenFinder example of such a case with the dragonfly species Ruddy Darter (*Sympetrum sanguineum*), see Figure 5.4.1. Three more species of the genus *Sympetrum* occur in Rheinland-Pfalz and produced valid observed communities in the evaluation of the observed communities approach: Red-veined Darter (*Sympetrum fonscolombii*), Common Darter (*Sympetrum striolatum*) and Vagrant Darter (*Sympetrum vulgatum*). These species are all similar in appearance to Rudy Darter. Speaking of similar appearance, it is important to stress that in a citizen science environment this concept has to be used in a wider sense than it would be used for professional experts. Of course, experienced citizen observers often develop skills in species identification which are on a professional level. However, a citizen science project will always have a certain number of less skilled observers due to a lack of experience or other factors. Also, in nature there is always a certain degree of variability in the appearance of individuals of a species, making it sometimes difficult even for professionals to distinguish similar species with certainty. Different *Sympetrum* species have a different phenology, which sometimes allows for excluding species because they usually fly earlier or later in the year, but seasons overlap.



a) Ruddy Darter *(Sympetrum sanguineum)*

b) Red-veined Darter *(Sympetrum fonscolombii)*

c) Common Darter *(Sympetrum striolatum)*

d) Vagrant Darter *(Sympetrum vulgatum)*

Figure 5.4.1: Four Sympetrum species occurring in Rheinland-Pfalz. (All males; photos: C. Jacobs).

Observed communities of Ruddy Darter and the other *Sympetrum* species are similar, with Simpson similarity values between 0.78 to 0.84, and Jaccard similarities ranging from 0.52 to 0.68. This shows that all of these *Sympetrum* species are usually observed in places with a similar observation context. They share a preference for standing or only slowly flowing water bodies, although there are differences in vegetation structure in their preferred habitats. High similarities of Ruddy Darter's observed community can also be found with observed communities of many other dragonfly and damselfly species (e.g. a Simpson index value of 1.0 with Southern Hawker), and with those of other species living in or close to water, such as Common Kingfisher, Grass Snake, and some amphibian species. All in all, ca. 77% of species with a Simpson index value of 0.8 or higher match Ruddy Darter's habitat preference, at least to some degree. Exceptions exist, such as high Simpson similarity with Middle Spotted Woodpecker's observed community, which is somewhat surprising, but might be explained by the latter's preference (among others) of riparian forests rich in tree species with coarse bark[58]. However, Jaccard similarity is not high in this case, at 0.19, because Middle Spotted Woodpecker's observed community is rather small, with just 13 species. Ruddy Darter's observed community is relatively large, with 55 species, and this results in a relatively low Jaccard index value (although not extremely low, because 11 species are part of both observed communities). Other cases of medium to low Jaccard similarity among high Simpson similarity cases occur for the same reasons. Ruddy Darter's observed community is rather different from species not associated to waterbodies, and especially of other species groups, such as many butterflies and birds. Among the 12 species whose observed communities have a Simpson index value of 0.2 or lower with Ruddy Darter's, only Western Marsh Harrier (*Circus aeruginosus*) can be expected to be observed often close to certain kinds of waterbodies. For reasons explained above, Jaccard similarity values are low for these cases, too. The current example species' observed community has a larger size of 55 species (above average in the AF_A set, and among all valid observed communities) than Common Kingfisher's 15-species observed community. Simpson similarities of Ruddy Darter's observed community to other observed communities are therefore higher, on average, than Common Kingfisher's, because larger intersections with other observed communities are more probable.

For similarity values in individual observations of Ruddy Darter, the same mechanics take effect which were already discussed in detail for Common Kingfisher observation examples above take effect. One AF_A example observation (ID 54251632, observed on 2016-08-14) from the fields near the village of Gommersheim in southeastern Rheinland-Pfalz has high similarity values (Simpson index 0.85, Jaccard index 0.40) with a candidate context of 110 species (about average in the AF_A set), covering 47 of the 55 species in the observed community. The observed communities approach therefore indicates a plausible location, which matches typical previous observations of Ruddy Darter. The observation is not located directly on a water body, but there are several ponds within the search radius of 1,000 m (the closest at a distance of ca. 320 m). Thus, although dragonflies have a certain mobility and will also be observed away from their reproduction habitats (as in the current case), the observed community approach evaluated the location at hand as plausible. However, would the species identification for this observation have been one of the other *Sympetrum* species discussed above, plausibility estimation of the observation's location would not have been much different. For instance, a Red-veined Darter observation in the same place would have reached an only slightly lower Simpson index value of 0.79 and an identical Jaccard index value of 0.40. Index values for a Common Darter observation would be at 0.78 and a somewhat lower 0.33, respectively. Vagrant Darter would have resulted in a lower Simpson index value of 0.60 and a Jaccard index value of 0.33. In this example case, therefore, the observed communities approach gives us a high plausibility of the location for a Ruddy Darter observation, but it cannot be ruled out, on just these grounds, that the species might have been mixed up with one of the other *Sympetrum* species, especially Red-veined Darter or Common Darter. Further evidence, such as a good photo proof, or high experience on the side of the observer in identifying dragonflies, is needed here to support the species identification, and was probably present, as the observation was approved by ArtenFinder experts. As already stated in Table 5.4.1, the observed communities plausibility indication in this case makes mix-ups with species with different observed

[58] http://arteninfo.net/elearning/libellen/speciesportrait/310, 2018-05-06

communities implausible, including physically similar species (not present in the dataset in this case, but Vagrant Darter presenting a less plausible case).

An AF_A observation of Ruddy Darter with an untypical and therefore implausible location by the observed communities approach can be found east of Kaiserslautern in south-central Rheinland-Pfalz (ID 54257534, observed on 2016-08-30). Its candidate context is of average size (107 species), but covers only eight of the 55 species in Ruddy Darter's observed community. Although the candidate observation itself is situated on a pond, most observations in its 1,000 m spatial context area are located in other habitats. Consequently, both the Simpson index (at 0.15) and the Jaccard index (at 0.05) indicate low plausibility of the location. For reasons discussed above, this result renders the location also implausible for other *Sympetrum* species with observed communities similar to Rudy Darter's. Plausibility indication for the location would probably have been alike if the observer would have identified the species as one of these in this case. Red-veined Darter would have resulted in almost identical index values of 0.15 and 0.06, as well as Common Darter with 0.16 and 0.5; Vagrant Darter would have reached slightly higher index values at 0.22 and 0.10. Again, other evidence was obviously present which led the expert in charge of this species and area to accept the observation as correct, a decision which would not have been supported by the observed communities approach to plausibility estimation.

Table 5.4.2: Properties of example candidate observations discussed in section 5.4.2.

Observation ID	**Observation date**	**Species**	**No. of species in observed community**	**No. of species in candidate context**	**No. of species in both lists**	**Simpson index**	**Jaccard index**
54271696	2016-11-20	Common Kingfisher (*Alcedo atthis*)	15	42	13	0.87	0.30
54266476	2016-10-13	Common Kingfisher (*Alcedo atthis*)	15	102	3	0.20	0.03
54269945	2016-10-31	Common Kingfisher (*Alcedo atthis*)	15	308	15	1.00	0.05
54251632	2016-08-14	Ruddy Darter (*Sympetrum sanguineum*)	55	110	47	0.85	0.40
54257534	2016-08-30	Ruddy Darter (*Sympetrum sanguineum*)	55	107	8	0.15	0.05

In this discussion, real ArtenFinder observations were used. Table 5.4.2 lists their relevant properties. Analog examples could be found in iNaturalist, but the general mechanisms laid out in Table 5.4.1 and illustrated by the above examples are the same. These examples, all coming from the AF_A set, that is, real observations from the year 2016 which were accepted as correct by experts in the project's validation process, illustrate that similarity values range from low to high values in this set. Clearly, the observed communities approach indicates low plausibility in light of their species identification and observation context for locations of many approved observations. It is also possible that evaluation of the approach with ArtenFinder data found some cases which were erroneously approved. The same findings are true for the set of observations which were rejected by experts in the ArtenFinder project (AF_R): some of these (although a smaller portion than in AF_A) appear plausible in observed communities approach evaluation. In both cases, there are many different reasons for acceptance or rejection, as explained in section 2.1.1.The above examples illustrate that the observed communities approach to plausibility estimation can support such decisions, but that its indications are not always in accordance with actual decisions. This finding is furher discussed below in section 5.4.4, including an attempt at quantifying it. The discussion above takes great care of explaining the exact extent of the observed communities approach's indicative power and its inherent limits.

5.4.3 OSM Environments Examples: Illustrating Target Species and Candidate Cases

While the general mechanics taking effect in estimating the plausibility of a candidate observation with OSM environments are the same as with observed communities, it became clear from the results discussed above that relations between species concerning similarity of their OSM environments are different. Observed communities are similar among species with similar habitat preferences, often from the same species group. This is no longer true with OSM environments. The following examples were chosen to illustrate this, and to illustrate the consequences for plausibility estimation of individual candidate observations with OSM environments.

Eurasian Moorhen was portrayed above as a species which can be expected to be predominantly observed on ponds, small lakes or other stagnant or slowly flowing water bodies. It is also a species which is rather easy to identify. Even the somewhat similar Eurasian Coot (*Fulica atra*) is easily distinguishable as an adult bird by distinct coloring (Rößner et al. 2013) for untrained observers. The above discussion also revealed that Eurasian Moorhen's Rheinland-Pfalz OSM environment does not feature tags which point to its habitat preference, which has been proven to be typical for OSM environments. Consequently, similarities with other OSM environments include species from all species groups and with many different properties. Only ca. 22% of the species whose OSM environments have a Simpson index value of 0.8 or higher with Eurasian Moorhen are other waterfowl, dragonflies or damselflies, or other species which can also be expected to be predominantly observed close to ponds etc. Among species with a relatively high Jaccard similarity of 0.4 or higher are ca. 28% of such species, a comparable value. However, all species with high Simpson or Jaccard index values between their OSM environment and the Eurasian Moorhen's are often observed in places with an OSM context which is similar to the one often found around observations of the Eurasian Moorhen. Else, their OSM environments would not be similar to the Eurasian Moorhen's. In contrast to the observed community approach, however, high or low similarity of OSM environments is obviously not critically influenced by biological or ecological species properties such as habitat preference, but simply by the fact that species are predominantly observed in OSM context situations which are similar. In some cases, high index values (especially of the Simpson index) may be boosted by a large OSM environment size, because Eurasian Moorhen's OSM environment, with 15 tags, is relatively small. OSM environments with high Simpson similarities with Common Moorhen's have a mean of 34.5 tags, which is above the average of 19.5 for all ArtenFinder OSM environments. For OSM environments with Simpson similarities of 0.2 and lower with Common Moorhen's this average is just 13.9%. For OSM environments with high (0.4 or higher) or low (0.125 or lower) Jaccard index values, the corresponding mean numbers of tags are both much closer to average, at 21.2 tags (for high similarity OSM environments) and 17.2 tags (for low similarity OSM environments). These numbers support the finding discussed earlier, that Simpson similarity is biased by OSM information density in the OSM environments approach to plausibility estimation of casual citizen science observations of organisms. OSM environments of other species have high or low Simpson similarities to Eurasian Moorhen's OSM environment at least in part because they are predominantly large or small. With Jaccard similarity, OSM environment size is much less influential.

Some examples of individual candidate observations are now used to illustrate how the OSM environments approach works as an indicator of plausibility. The AF_A set of accepted ArtenFinder observations from the year 2016 has Eurasian Moorhen candidate cases which have relatively high index values both for the Simpson and the Jaccard index. One of these (ID 54274304, observed on 2016-12-18) can be found on a small lake close to the village of Neuburg am Rhein which is situated on the river Rhine in the extreme southeast of Rheinland-Pfalz. The location has an average-size OSM candidate context of 36 tags, which covers 13 of the Eurasian moorhen's 15 OSM environment tags, resulting in a Simpson index value of 0.87 and a Jaccard index value of 0.34. The location therefore appears plausible for a Eurasian Moorhen observation by the OSM environments approach to plausibility estimation, supporting the positive verification decision made by the ArtenFinder experts who accepted this observation as probably correct.

A different situation is presented by a Eurasian Moorhen observation (ID 54273201, observed on 2016-12-05) on a pond in Kaiserslautern's "Volkspark", a public park in a city in south-central Rheinland-Pfalz. Here, the location's OSM context is much larger: 72 different tags are present in its neighborhood, covering all 15 tags in Eurasian Moorhen's OSM environment. Consequently, the Simpson index value is 1.0. However, the Jaccard index reaches only a value of 0.21. The high OSM information density provided by the urban setting of the observation certainly plays a role in creating a very high Simpson similarity, while the Jaccard index, taking the large number of candidate context tags into account which are not present in the species' OSM environment, does not necessarily justify evaluation of the observation as plausible, although ArtenFinder experts accepted this observation as correct, which seems well justified by the location.

A third instructive example of a Eurasian Moorhen observation, this time with low index values of 0.15 (Simpson index) and 0.08 (Jaccard index) and therefore appearing implausible (but again taken from the AF_A set of accepted ArtenFinder observations), can be found ca. 18 km southwest of Kaiserslautern in a rural setting away from settled areas. The outskirts of the closest village, Waldfischbach-Burgalben, are ca. one km away, placing the village mostly outside the 1,000 m radius employed in candidate OSM context extraction. The observation (ID 54223799, observed on 2016-05-21) is situated in a pond which is part of a golf course. This location's OSM context has only 13 tags, covering just two of the tags in Eurasian Moorhen's OSM environment, leading to low similarity values. This candidate observation therefore appears unusual by its OSM context. It is also unusual in that the candidate context, with 13 tags, is smaller than the OSM environment involved, and therefore used as the denominator of the Simpson index in this case, while in most cases the candidate context exceeds the target species' OSM environment.

These examples highlight the effect of spatial properties of OSM data, with their contrast between rural and urban areas, on the way the OSM environments approach to plausibility estimations works. High OSM information density in urban settings lead to observations being evaluated as plausible (at least by the Simpson index value), while rural settings with low numbers of tags rather produce low plausibility results. Of course, in the examples of such settings used above, the fact remains that the intersection between candidate context and the target species' OSM environment simply was high in the urban setting and low in the rural one, which is not necessarily the case in all candidate cases. Despite its large size of 72 tags, the urban OSM context cited above might not have covered Eurasian Moorhen's OSM environment to such a high extent, and the mere 13 tags found in the rural setting example used above might have covered a much larger portion of Eurasian Moorhen's OSM environment. However, Eurasian Moorhen does not present such examples in the data used.

Still, examples exist. An observation (ID 54238227, 2016-07-12) of Grayling taken from the AF_A set features 16 surrounding tags all of which cover the target species' 11-tags OSM environment, producing high similarity values of 1.0 (Simpson index) and 0.69 (Jaccard index). The candidate observation comes from a former military area within a large forested region (the Pfälzerwald) in southern Rheinland-Pfalz, with 23 of the 31 prior observations of the species (from which the OSM environment was extracted) nearby. The OSM environments approach identified the location as plausible because the candidate observation's OSM context necessarily shares many tags with these prior observations.

An example observation (ID 54033011, 2014-06-17) of Five-spot Burnet, a species of moth already presented above, taken from the ArtenFinder candidates set AF_R (set of rejected observations), features a 58 tags candidate observation OSM context, which covers only one of the 11 tags in this target species' OSM environment. Consequently, the OSM environments approach evaluates this candidate observation as implausible, with a Simpson index value of 0.09 and a Jaccard index value of 0.01. The candidate observation is situated at the edge of a settled area right next to a highway. Prior observations (which produced the species' OSM environment) are dispersed over Rheinland-Pfalz, and are situated predominantly in rural, mostly forested areas with meadows. Therefore, the OSM environments approach confirms the ArtenFinder experts' decision of rejecting this candidate observation. The genus *Zygaena* has many species occurring in Rheinland-Pfalz, which are physically very similar among one another[59]. A mix-up with a different *Zygaena* species is therefore probable in this case.

Table 5.4.3: Properties of example candidate observations discussed in section 5.4.3.

Observation ID	**Observation date**	**Species**	**No. of species in observed community**	**No. of species in candidate context**	**No. of species in both lists**	**Simpson index**	**Jaccard index**
54274304	2016-12-18	Eurasian Moorhen (*Gallinula chloropus*)	15	36	13	0.87	0.34
54273201	2016-12-05	Eurasian Moorhen (*Gallinula chloropus*)	15	72	15	1.00	0.21
54223799	2016-05-21	Eurasian Moorhen (*Gallinula chloropus*)	15	13	2	0.15	0.08
54238227	2016-07-12	Grayling (*Hipparchia semele*)	11	16	11	1.00	0.69
54033011	2014-06-17	Five-spot Burnet (*Zygaena trifolii*)	11	58	1	0.09	0.01

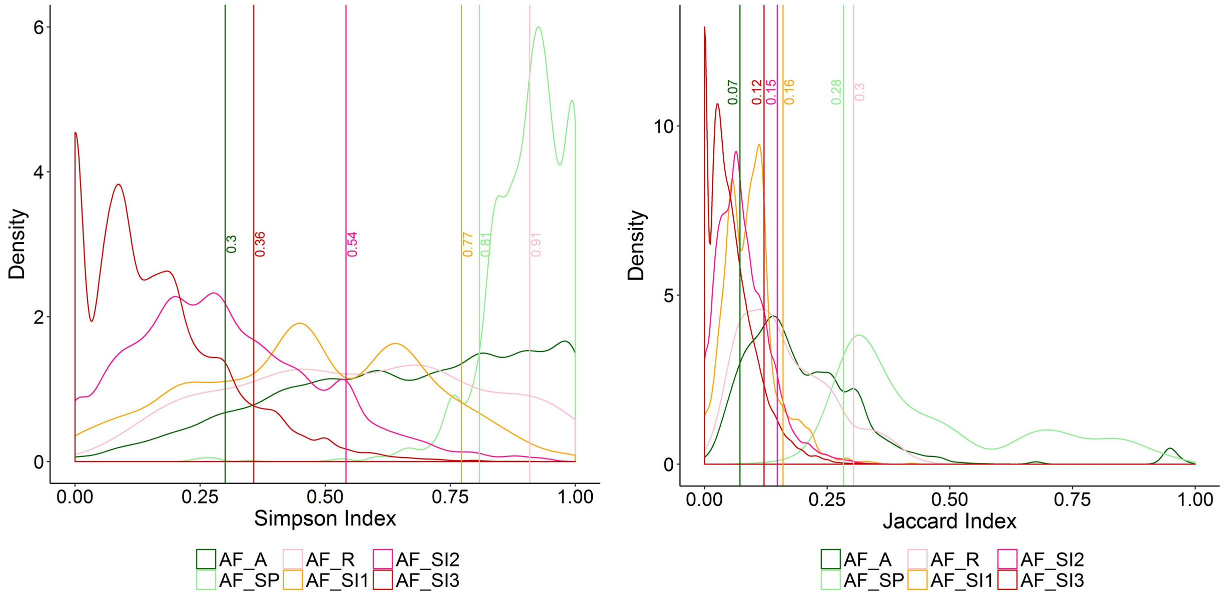
In all of the above examples, the usual considerations concerning certainty of species identification and location apply in the same way which was already discussed for the candidate observation examples illustrating the observed communities approach to plausibility estimation. They are described in general terms in Table 5.4.1. In any candidate case, all species with a similar OSM environment would lead to similar plausibility estimations in the same location, which is especially relevant if there are physically similar species with this property. A mix-up with a physically similar species which has a different OSM environment is, however, unlikely in such a case. Any candidate observation with a plausible location would be estimated as equally plausible in a different location with a similar OSM context. Table 5.4.3 summarizes the example observations' properties.

[59] https://arteninfo.net/elearning/nachtfalter/select_species#latZ, last accessed on 2018-08-24

5.4.4 Deriving Decision Thresholds for Plausible and Implausible Cases

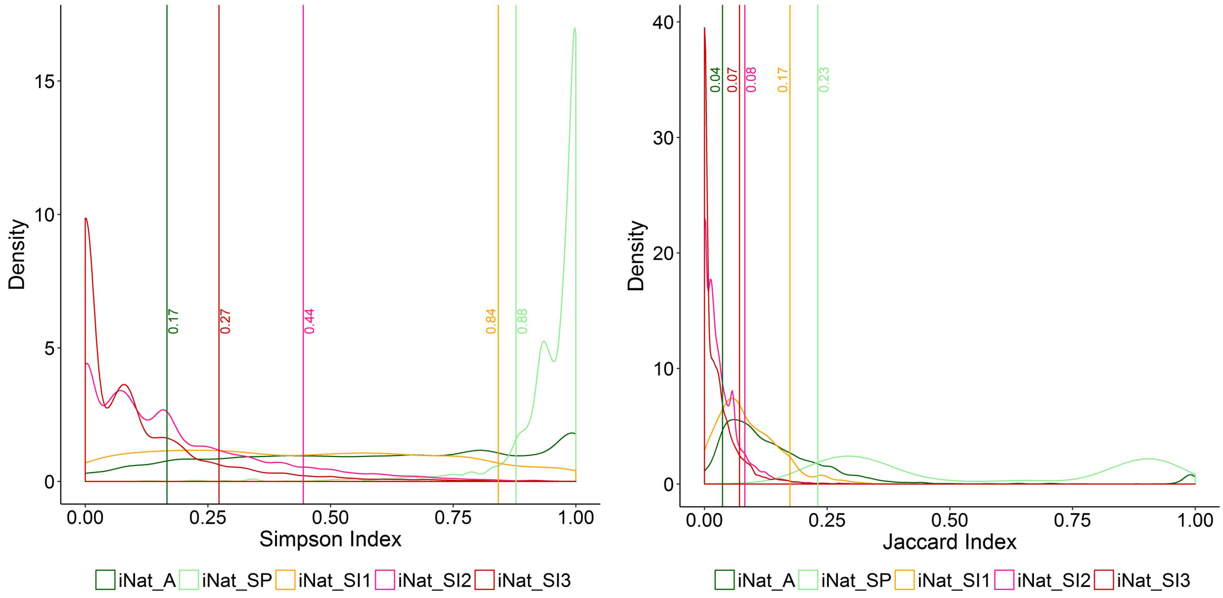
The examples discussed above present cases with rather high or rather low index values, clearly indicating plausible or implausible cases. But when exactly do Simpson or Jaccard index values indicate a plausible case, and when do they indicate implausibility? In quality assurance regimes where a decision between approval and flagging as unusual is used (e.g., in the ArtenFinder project), some way of defining decision thresholds is needed. The kernel density estimations derived from the distributions of index values used in evaluation may provide some guidance for defining decision thresholds, which is demonstrated here using similarity values from the observed communities approach.

Quantiles of probabilities can be derived from kernel density estimations by integrating the area under the density graphs. Let's look at the Simpson index first. For instance, 90% of synthetic plausible ArtenFinder candidate observations (AF_SP) have Simpson index values of 0.81 or higher. Therefore, an observation could be expected to belong to the set of plausible observations in light of its observed community, if it has a Simpson index value of 0.81 or higher, with 90% certainty. With the AF_A set as a benchmark, the critical index value would be much lower, at 0.3. However, it was already discussed that AF_A has many observations which appear implausible by the observed communities approach, because the decision for acceptance is based also on other criteria. Therefore, this set is unsuitable for the purpose of defining a decision threshold for observations which are plausible only by the observed communities criterion. For the decision threshold of observations' locations to be considered implausible on the same level of certainty, several possibilities are to be considered. In the synthetic set of implausible observations AF_SI3, 90% of observations have a Simpson index value of 0.36 or less. An observation could be expected to belong to the set of implausible observations in light of its observed community, if it has a Simpson index value of less than 0.36, with 90% certainty. Other sets suggest other decision thresholds, with a maximum at 0.91 when real rejected observations (AF_R) are used, see Figure 5.4.2. Issues with the AF_R and AF_SI1 sets have already been discussed, rendering their results useless for these considerations. The critical Simpson index value of 0.54 for the AF_SI2 set is produced with synthetic candidate observations with spatial properties (concerning observation density around them) comparable to real observation data (other than AF_SI3 candidates, which are, on average, situated in locations with lower observation density) and therefore presents a realistic alternative. Between the upper and lower thresholds, there is a range of Simpson index values in which a decision is not possible, on the given level of certainty. Thresholds derived in this way for ArtenFinder Jaccard index values are, of course, much lower, but the approach of finding them can be applied to this index in the same way (results see Figure 5.4.2). iNaturalist data render comparable results (see Figure 5.4.3) with a larger difference between iNat_SP and iNat_SI2/_SI3 for the Simpson index, and overall lower thresholds for the Jaccard index.


a) ArtenFinder, Simpson index

b) ArtenFinder, Jaccard index

Figure 5.4.2: ArtenFinder, Kernel Density Estimations for Simpson and Jaccard Index Values, with 10% (AF_A, AF_SP) and 90% (AF_R, AF_SI1, AF_SI2, AF_SI3) probability quantiles.


a) iNaturalist, Simpson index

b) iNaturalist, Jaccard index

Figure 5.4.3: iNaturalist, Kernel Density Estimations for Simpson and Jaccard Index Values, with 10% (iNat_A) and 90% (iNat_SI1, iNat_SI2, iNat_SI3) probability quantiles.

What if this technique of finding critical index values for finding plausible or implausible cases would have been applied to the real approved or rejected observations in the data use cases employed in this work? In how many candidate cases would it have supported or contradicted the actual decision made by the experts? The discussion of example candidate observations in section 5.4.2 already revealed that plausibility estimations with the observed communities approach do not always confirm actual quality assurance decisions. The above considerations on similarity value thresholds for plausible or implausible observations allow for attempting to quantify this finding. Table 5.4.4 shows results for the _SP and _SI2 thresholds and the parameters used above. 35.0% of accepted ArtenFinder cases from 2016 (AF_A) had Simpson index values of 0.81 and higher. In these cases, the observed communities approach supports the actual quality assurance decision of accepting these candidate observations as correct, because their locations appear plausible. However, 33.3% of AF_A candidates with Simpson index values at or below 0.54 have implausible locations by the observed communities criterion (if the critical index value rendered by the AF_SI2 set is used). About half of the rejected ArtenFinder candidates (AF_R) are evaluated as implausible (47.5% of candidates) with Simpson index values at or below 0.54. A much lower portion of just 19.9% of these cases is evaluated as plausible by the observed communities criterion, with a Simpson index above 0.81. Jaccard index results differ in that less accepted ArtenFinder candidates are evaluated as plausible (22.1%) while more appear implausible (43.0%). Rejected ArtenFinder candidates show the same effect: more than half of them (55.2%) have Jaccard index values which make them appear implausible and only very few (12.7%) appear plausible. With approved (research grade) iNaturalist candidates (iNat_A), findings are quite similar with both Simpson and Jaccard index values (see Table 5.4.4). Portions of candidates that are evaluated as plausible or implausible are rather lower than with ArtenFinder data. A set of actually rejected candidate observations does not exist in the iNaturalist data use case (see section 3.3.2).

Table 5.4.4: Portions of candidate observations evaluated as plausible or implausible, using critical index values.

ArtenFinder	**Simpson index**		**Jaccard Index**		**iNaturalist**	**Simpson index**		**Jaccard Index**	
	≥ 0.81	**≤ 0.54**	**≥ 0.28**	**≤ 0.15**		**≥ 0.88**	**≤ 0.44**	**≥ 0.23**	**≤ 0.08**
AF_A	35.0%	33.3%	22.1%	43.0%	**iNat_A**	22.9%	34.5%	18.2%	37.0%
AF_R	19.9%	47.5%	12.7%	55.2%	-	-	-	-	-

These findings show that the observed communities approach is rather conservative when it comes to identifying implausible observations, at least with the parameter settings used in this discussion. Many candidate observations that were actually approved would have been evaluated as implausible by this indicator. This is a benefit in the sense that the probability to miss an erroneous observation becomes smaller. However, it is also a disadvantage because it raises the number of candidate observations that appear implausible and have to be checked. This problem is amplified by the fact that in practice, the actual portion of erroneous reports in all submitted observations is quite small. Domain experts usually give overall estimations of under 10% of erroneous observations submitted to a citizen science project collecting casual observations of organisms. However, in quality assurance regimes such as that of ArtenFinder each and every observation has to be checked by experts so far. Using a technique such as the observed communities approach to flag one third to half of the candidate observations and check these, while one fifth to one third of the candidate observations almost certainly do not have to be checked, is a great advantage in terms of absolute observation numbers, and thus in terms of work load for experts who do the checking. Of course, it is also true that only a part of all candidate observations can be evaluated at all (for various methodological reasons, and depending on parameter settings, see chapter 4). It is also very important to stress here again that the observed communities indicator evaluates only one aspect of plausibility, which is a candidate observation’s location in light of its species ID and surrounding observations. Actual quality assurance decisions such as those made in ArtenFinder are based on many more criteria, some even not related to plausibility (see section 2.1), so that a certain degree of mismatch between actual approval and rejection of observations on the one hand, and plausibility estimation results with the observed communities approach on the other hand, is to be expected.

6 Summary, Conclusions and Future Research

6.1 Summary and Conclusions

This thesis set out to answer a number of research questions. To do so, it explored novel approaches to quality assessment of casual citizen science observations of organisms based on geographic context and plausibility. This context was gained either intrinsically from the stock of observations available in citizen science projects, or extrinsically from OSM, a VGI data source providing a heterogeneous, but often very detailed representation of the environment. The former approach casts geographic context into observed communities which mirror a target species' typical observation context within a citizen science dataset of casual observations of various organisms. The latter casts the VGI context provided by OSM into a target species' OSM environment, holding elements which are frequently mapped in places where the target species is observed. Both approaches were evaluated and thoroughly examined with two data use cases of casual citizen science observations. Evaluation proved that both approaches are indeed useful for plausibility assessment of casual citizen science observations of organisms, but that they also present a number of difficulties. Detailed examination and discussions provided insights into the approaches' behavior and revealed the extent and limits of their indicative power. The most important findings for each research question can be summarized as follows:

1. **Principle research questions:**

a) How can geographic context be used for intrinsic assessment of the plausibility of casual citizen science observations of organisms?

a1) How can casual observations be turned into an intrinsic source of geographic context?

Observed communities propose a way of using casual observations of organisms as an intrinsic source of geographic context. An observed community of a target species represents a typical observation situation of that target species within the respectice dataset of observation data. It is produced by extracting all species frequently observed in close spatial association to available approved observations of the target species, and therefor carries characteristics caused both by natural species distributions and by VGI-related influences on the observation data used. In considering all available context observations regardless of species, the approach uses the full context potential of a multi-species dataset. Context is therefore grounded on a broad basis of observations of different species.

a2) How can this intrinsic context information be used to estimate plausibility of a candidate observation?

Similarity indices were successfully used in this work to measure how well the geographic context of a candidate observation fits the observed community of the species which was reported in that candidate observation. This work compared results with two similarity indices, the Simpson and the Jaccard index. These indices are based on the rate of species which both the observed community and the candidate context have in common. Evaluation was able to show that contexts of plausible candidate observations usually have higher similarity values with their observed communities, than have contexts of implausible observations. Similarity values can therefore be used to estimate plausibility of a candidate observation. In this way, the approach is able to show whether an observation comes from a typical location, or from an unusual location, considering the species identification given by the observer. As no external data are used in this approach, it is completely intrinsic.

b) How can extrinsic VGI data be used for assessing the plausibility of casual citizen science observations of organisms?

b1) How can OSM data be used as an extrinsic source of geographic context?

OSM environments describe the typical geographic context of a target species in terms of OSM tags frequently found in close spatial proximity to observations of that target species. In a procedure which is analogous to the extraction of observed communities, an OSM environment of a target species is produced by collecting all tags which are frequently found in close spatial association with approved observations of that target species. OSM data are an extrinsic source of geographic VGI context.

b2) How can this extrinsic context information be used to estimate plausibility of a candidate observation?

The same similarity indices as used in the observed communities approach were evaluated and proved basically successful in serving as plausibility indicators with OSM environments as well. This approach is able to evaluate candidate observations which did not provide valid cases in the observed communities approach, because the extrinsic geographic context of OSM provides context in places where observation data do not. However, using this extrinsic source of geographic context may lead to spatial bias in plausibility estimation caused by spatially variable information density in OSM. Also, evaluation of the approach showed differences between sets of plausible or implausible candidate observations to be smaller than with the observed communities approach. Specifically, similarity values of implausible observations are generally higher.

2. In-depth research questions:

a) What are the effects of the spatial properties of geographic context data on plausibility estimation?

Both casual citizen science observations of organisms and OSM data are VGI. Such data are spatially heterogeneous, with spatially variable density of context observations, or of OSM information density. Both higher context observation density and higher OSM information density may lead to higher plausibility of candidate observations, especially with the Simpson index. In the intrinsic observed communities approach, this effect can be regarded as a legitimate contribution of the VGI nature of the data collection process to plausibility estimation, because a candidate observation from a location with higher observation intensity is indeed more plausible. In the extrinsic OSM environments approach, however, a higher OSM information density at a location does not justify a higher plausibility of a candidate observation, because mapping of OSM data and collection of species observation data are not connected in any way.

b) How do species properties affect results?

Observed communities usually exhibit, in their species composition, distinct characteristics governed by properties of their target species, especially preference of certain habitats.

OSM environments rarely do so, but are rather shaped in their tag composition by general properties of the thematic structure of the underlying OSM data, with generally frequent tags also dominating in OSM environments, regardless of the ecological features of the target species. Results with both approaches are also influenced by more general VGI-related factors: datasets of casual citizen science observations are always biased towards certain popular species groups, towards species which can easily be photographed, etc. These effects also determine to a large degree which species can be evaluated by the approaches proposed here.

c) How do changes to parameters and methodological modifications affect results?

Both approaches to plausibility estimation of casual citizen science observations proposed in this work have a number of parameters, such as the size of the search radius for relevant geographic context, or frequency parameters for associations of species or tags, and for identifying nonspecific species or tags. Effects of parameter changes were examined in detail for the observed communities approach. The approach was found to basically keep its ability to distinguish between plausible and implausible observations with many different parameter settings. However, effectiveness, in terms of numbers of species and candidate observations evaluated, can be quite different for different settings. Some modifications to the basic methodology of the approaches, such as using a quantified similarity calculation with observation frequency, or introducing polygon geometries which are related to habitat structures for focusing relevant geographic context, showed potential of improving results. Other modifications which were also tested, such as species-group-specific search radii or date-specific OSM context adjusted to observation dates, did not prove to introduce any advantage.

d) What are the extent and limits of indicative power of the obtained approaches to plausibility estimation?

Both the observed communities approach and the OSM environments approach can be used to estimate the plausibility of a candidate observation's location in light of the species identification given by the observer. With a high level of similarity of candidate context and observed community or OSM environment, the reported location is plausible for the given species, while a low level of similarity identifies an implausible location for the reported species. All other species with a similar observed community or OSM environment would produce similar plausibility estimations at the same location. In a casual citizen science setting, this is especially relevant for species which can be easily mixed up because they are physically similar. Also, all other locations with a similar geographic context would render similar plausibility estimations for the reported species.

More conclusions can be drawn when looking at the contributions of this work in more detail. Some of them present weaknesses and drawbacks of the approaches presented here, while others highlight their strengths and advantages.

The number of species and of candidate observations which can actually be used with the approaches evaluated in this work is limited by several parameters:

- minimum observation numbers of a target species required for extraction of observed communities or OSM environments,
- threshold for association frequency,
- frequency threshold for the elimination of nonspecific species or tags, and

- requirement of at least 10 species or tags per observed community, OSM environment and candidate context for similarity index calculation.

With medium frequency thresholds at 0.5, and a minimum number of 10 observations of a target species for observed communities extraction, the ArtenFinder use case produced valid observed communities for 26.8% of all species observed up to 2015, and the iNaturalist case did so for 15.5%. These numbers seem low, but because they represent the most frequently observed species in each dataset, they still allowed for estimating the plausibility of 50.1% of accepted ArtenFinder observations of the year 2016, and of 30.0% of 2016 iNaturalist research grade observations from California. The OSM environments approach was less effective with the same parameters, because OSM data have lower information content as a geographic context source, than have context observations. OSM environments and candidate OSM contexts are usually smaller than observed communities and candidate species contexts, and therefore drop more often below the requirement of at least 10 tags per OSM environment and candidate context for similarity calculation, which was set in this work to avoid erratic similarity results. The bottom line of these considerations is to keep in mind that both approaches cannot be used on all candidate observations. In the form developed and evaluated here, they can only be used on species with a certain minimum stock of previous observations, and on candidates from locations which provide adequate information density in the form of context observations or of OSM tags. A citizen science project must run for some time to acquire enough data for both approaches to be feasible at least for some species, and in some locations or regions. This problem is mitigated, for observed communities, by the fact that observations are usually clustered, supporting adequate context observation numbers within clusters. The OSM environments approach is, of course, dependent on an adequate information density in the form of tags.

When selecting observation data for use with the observed communities or OSM environments approach, it is important to carefully consider the nature of the location information these observation data provide. Both approaches are suitable only for observations which are equipped with individual coordinates of the location where they were actually made. Observations which have displaced coordinates, such as the central point of an arbitrary area or of a map quadrant (both also widespread in casual biodiversity observation datasets) cannot be used with these approaches in their form presented here, because observations would appear associated with other observations or with OSM tags which may actually not have been observed or mapped close to the true observation location. Of course, observation coordinates always have issues of precision, caused by technical properties of GPS receivers, and by uncertainty inherent in the process of volunteers placing an observation on different kinds of base maps. These uncertainties may erroneously place a context observation or OSM tag within the search radius around a target or candidate observation, although actually observed or mapped outside of this neighborhood, and vice versa. Mobility of many species also adds an element of randomness to observation coordinates. However, results of the studies conducted in this work show that these factors do not prevent the observed communities approach and the OSM environments approach from working, while certainly adding a certain amount of noise to results. Another important conclusion in this context is that both approaches, although evaluated here with casual citizen science observations of organisms, have the potential to be used with any data source providing observation data whose location information fulfills the condition described above.

Example cases taken from real, approved observations demonstrated that plausibility estimation by the observed communities approach or the OSM environments approach does not always support actual approval of a candidate case by the quality assurance mechanism in the respective project, although it does in the larger part of such cases. This underpins the important fact that these approaches can be used as indicators for identifying cases with unusual locations in light of existing data, but that validation decisions should always be based on more information about the candidate case, e.g., observation date, observer experience, or a photo proof, none of which, alone, allow for a well-founded validation decision.

Using VGI as a source of geographic context, as was done in this work, also holds numerous advantages. The observed communities approach, on the one hand, is completely intrinsic. It does not need any data other than the observation data already available in the project in question. It also uses the full context potential of a multi-species dataset. Context is therefore grounded on a broad basis of different species. In doing so, this approach avoids problems of local data scarcity as far as possible, and will work also for a candidate observation which has no previous observations of the same species close by, if there are enough observations of other species around its location. The OSM environments approach, on the other hand, builds on a freely and globally available source of geographic context which is growing more and more complete in many regions. In fact, all data sources used in this work were shown to grow in terms of data content, steadily improving the data basis. Typical for VGI data however, spatial inhomogeneity in the data persists, with some regions becoming more and more complete, while others remain relatively underrepresented. Still, taking the VGI-related properties inherent in the data into account is also a strength of the approaches presented here. In particular, they integrate important factors governing the thematic and spatial properties of the specific data they use, and return a plausibility estimation which is based both on the natural distribution of the target species, and on patterns created by the special data acquisition process which is specific to the citizen science project in question. What is more, both approaches inherently reflect regional characteristics of species distribution and observer behavior, reproducing in their results the data characteristics specific to the region whose data they use.

The observed communities approach and the OSM environments approaches are spatially explicit in the sense that their plausibility estimations are specific to the location aspect of the candidate observation. This distinguishes them from most other approaches to plausibility estimation of observation data. Examination of the extent and limits of indicative power of the approaches presented in this work (chapter 5.4) rendered an intuitive way of demonstrating this fact, simply asking if and how the plausibility estimation with an indicator would change if the location of the candidate observation changed. With the observed communities and the OSM environments approaches, plausibility of a candidate observation usually changes when the location of the candidate observation is changed to some extent, except changes to certain alternative locations with similar geographic context. Other plausibility indicators used in VGI quality assessment do not show this property. In some cases, this is quite obvious. For instance, a quite common technique of estimating the plausibility of a candidate observation is to compare its observation date to known annual periods of occurrence of the species observed. This indicator obviously does not hold any information on the plausibility of a candidate observation's location: changing location of a candidate observation would not change the plausibility estimation given by such an indicator. Other approaches to estimating plausibility of a candidate observation present less obvious cases. For instance, a plausibility estimation based on the reputation of a volunteer considers a candidate observation to be plausible if the volunteer's reputation is high. This plausibility estimation usually includes all aspects of the candidate observation, such as species identification, place, and date, all of which can be considered to be accurate if the volunteer is considered to be reliable. However, at a different location, the same candidate observation would be just as plausible, because changing the location does not change the volunteer's reputation. In fact, the candidate observation would be just as plausible at any other location (and, of course, with any other date). Another well-established plausibility indicator for observations of organisms, which actually assesses an observations location, is comparison of observed location with the known spatial range of the species observed. In this case, plausibility of a candidate observation is the same everywhere within this range and candidate observations outside the range are implausible, giving this indicator's information a binary character. Plausibility information of the observed communities and the OSM environments approaches is different: it can be expected to show some form of continuous decay with growing distance to a candidate observation's location, and to rise again when approaching a location which has a geographic context similar to the original location. This simple exercise demonstrates that established plausibility indicators, as well as the indicators presented in this work, have limits to what they actually convey. However, these limits are rarely communicated.

Finally, the similarity indices used in this work bring to mind another consideration which might have some weight in a citizen science setting. Indices such as the Simpson or the Jaccard index have a straightforward meaning, which renders the approaches' underlying principle comprehensible and transparent to a high degree. In the observed communities approach, Simpson index values express, in most cases, the rate of species in the observed community covered by the candidate context. The modiefied Simpson index tested in this work (see section 3.4.7) unifies this interpretation of the similarity value in all candidate cases. Jaccard index values express the rate of species which are in common in observed community and candidate context, in the union of both species lists. Meanings of the indices are analogous in the OSM environments approach, where context species are replaced by OSM tags. Their straightforward meaning makes these indices especially suitable for use in citizen science settings where projects face the challenge of communicating their data processing procedures, including quality control methods, to participants and to potential data users who are non-experts in the domain of biodiversity research.

6.2 Future Research

The research, methods, and results presented in this thesis open up a large number of avenues for further scientific investigation. This work did certainly not achieve an absolutely exhaustive investigation of the proposed methods' properties, behavior and potential. Especially, more work on examining effects of properties of individual species, and of individual candidate cases, is certainly needed. Future work should extend the sensitivity analysis conducted in this work to explore effects for individual candidate cases with certain properties in more detail. Performing a more extensive sensitivity analysis also on the OSM environments approach would reveal where this approach's reaction to parameter changes and methodological modifications is different from the observed communities approach. Also, changes to parameters or modifications to the method might be combined to obtain certain desired effects (e.g., raising numbers of species which can be evaluated) while counteracting undesired ones (e.g., a parallel raise in numbers of nonspecific species).

The concept of species groups used here, borrowed mainly from the way the ArtenFinder project organizes its information, proved to be inadequate in some respects, because it aggregates species in part by criteria which do not go along with species properties relevant for the problems at hand. Regrouping species across these groups by criteria such as mobility, habitat preferences, detectability, etc., might render more focused insights. A related problem was identified in the use of individual OSM tags as context information source: fragmentation of certain environmental information or elements into many different individual tags causes problems in attaching relevant properties of the environment to target species. A meaningful grouping of tags might help to solve this problem. Approaches to use OSM as a source of land cover information (Schultz et al. 2017) might lead the way here, but need adjustment to a more habitat-centered perspective. Also, applying the methods from this work on more data use cases will certainly reveal more problems, and trigger insights into possible improvements. What is also missing so far is a thorough assessment of the approaches' usefulness, usability, and benefit in data validation practice. This requires implementation of the approaches in the quality control workflow of a citizen science project, and the development and employment of suitable evaluation procedures.

Future research should also be aimed at overcoming the major shortcomings of the methods which identified in this work. Although the perpetual observation process in casual citizen science observation projects, and of the mapping process in OSM, brings a steady improvement of these data's suitability as sources of geographic context, local or regional context data scarcity is one of the most important obstacles for the approaches to work on many candidate observations. A promising avenue towards overcoming this problem for the observed communities approach is to combine observation data of various sources into a potentially much more content-rich geographic context which would have the potential to possess a better geographic coverage, and therefore to provide a geographically more complete geographic context. However, such an endeavor is also fraught with difficulties, some of which were already hinted at. For instance, even highly unified data sources such as GBIF, which integrate observation data from many different sources in a homogeneous, easily accessible data source, still need careful consideration of data properties concerning the nature, accuracy, and precision of location information of data from different original sources. Data with individual locations should not be mixed with displaced locations such as map quadrant or user-defined area center points. Of course, the latter could be treated with modified methodological approaches, e.g., generating observed communities from all observations of a user-defined area, if said area is suitably delimited (i.e., not too large or internally too heterogeneous). Use of data sources not-yet integrated or unified may present difficulties connected to taxonomy, which, however, can now be overcome with the help of appropriate online services. Also, combining different data sources will be most effective if they provide spatially complementary data on different scales, locally as well as regionally. This implies that such an approach should aim to use data from sources with different data acquisition processes which result in disparate spatial distribution of their data. For instance, it might not suffice to choose several different casual citizen science observation repositories, as the volunteers involved in these projects, following the same basic rules of low-protocol data collection, would probably tend to visit the same locations or areas. A larger data basis might also allow for a more fine-grained treatment of observation data in terms of species' behavior and life cycles. In this work, observed communities and OSM environments were extracted using all observations of a target species, although these observations may represent very different life stages or behaviors (e.g., feeding or breeding) which may be associated with very different contexts in terms of other species or OSM tags, because they are taking place in different locations. One could think of several different observed communities, or OSM environments, for the same target species, e.g., representing resting or feeding behavior. Tapping into these differences would add a new dimension to plausibility estimation, but would certainly also require large amounts of observation data. It might, however, already be feasible for some species. This is especially relevant in large regions where species migrate internally between different parts of that region, e.g. water fowl in California breeding on inland freshwater while passing the winter on the coast.

Ideally, overcoming local or regional data scarcity will also work towards mitigating the influence of heterogeneity in spatial density of context information (context observations or OSM tags) on plausibility estimation results, because a higher overall spatial density of observations might also lead to a reduction in density contrast between regions or locations. However, this heterogeneity can certainly be expected to persist for some more time, and to do so perpetually in certain regions, because it is an inherent property of opportunistic data collection processes. Therefore, future research must seek to weaken or even eliminate this factor, especially in the extrinsic OSM environments approach, where it definitely introduces an undesirable bias into plausibility estimation.

An important conclusion of this work is that its approaches cannot alone determine a validation decision on a candidate observation, because their indicative power is limited. This is true for all plausibility indicators, and they are therefore often combined. Future research must examine how the indicators from this work can be integrated with others, which are based on other information, such as observation date, or trustworthiness of the volunteer, to play their part in a robust, objective plausibility estimation of candidate observations which is based on many proven approaches.

References

Alden, P., Heath, F., Keen, R., Leventer, A., Zomlefer, W. (Eds.), 1998. National Audubon Society field guide to California. Knopf, New York.

Ali, A.L., Schmid, F., Al-Salman, R., Kauppinen, T., 2014. Ambiguity and plausibility: managing classification quality in volunteered geographic information, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14. ACM Press, Dallas, Texas, pp. 143–152. DOI: 10.1145/2666310.2666392

Allaby, M., 2004. A dictionary of ecology. Oxford University Press, New York.

Amano, T., Lamming, J.D.L., Sutherland, W.J., 2016. Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. BioScience 66, 393–400. DOI: 10.1093/biosci/biw022

Arsanjani, J., Mooney, P., Zipf, A., Schauss, A., 2015. Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets, in: Arsanjani, J.J., Zipf, A., Mooney, P., Helbich, M. (Eds.), OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography. Springer, Switzerland. DOI: 10.1007/978-3-319-14280-7_3

August, T., Harvey, M., Lightfoot, P., Kilbey, D., Papadopoulos, T., Jepson, P., 2015. Emerging Technologies for Biological Recording. Biological Journal of the Linnean Society 115, 731–749. DOI: 10.1111/bij.12534

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial Point Patterns: Methodology and Applications with R, Chapman & Hall/CRC Interdisciplinary Statistics. Chapman and Hall/CRC, New York.

Ballatore, A., Zipf, A., 2015. A Conceptual Quality Framework for Volunteered Geographic Information, in: Fabrikant, S.I., Raubal, M., Bertolotto, M., Davies, C., Freundschuh, S., Bell, S. (Eds.), Spatial Information Theory. Springer International Publishing, Cham, pp. 89–107. DOI: 10.1007/978-3-319-23374-1_5

Balmford, A., Crane, P., Dobson, A., Green, R.E., Mace, G.M., 2005. The 2010 challenge: data availability, information needs and extraterrestrial insights. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 221–228. DOI: 10.1098/rstb.2004.1599

Barrington-Leigh, C., Millard-Ball, A., 2017. The world's user-generated road map is more than 80% complete. PLOS ONE 12, e0180698. DOI: 10.1371/journal.pone.0180698

Barron, C., Neis, P., Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Transactions in GIS 18, 877–895. DOI: 10.1111/tgis.12073

Baselga, A., Jimenez-Valverde, A., Niccolini, G., 2007. A multiple-site similarity measure independent of richness. Biology Letters 3, 642–645. DOI: 10.1098/rsbl.2007.0449

Bennett, B., 2001. What is a Forest? On the Vagueness of Certain Geographic Concepts. Topoi 20, 189–201.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173, 144–154. DOI: 10.1016/j.biocon.2013.07.037

Bishr, M., Kuhn, W., 2007. Geospatial Information Bottom-Up: A Matter of Trust and Semantics, in: Fabrikant, S.I., Wachowicz, M. (Eds.), The European Information Society, Lecture Notes in Geoinformation and Cartography. Springer, Berlin, Heidelberg, pp. 365–387.

Bishr, M., Mantelas, L., 2008. A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal 72, 229–237. DOI: 10.1007/s10708-008-9182-4

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 Biology 8, e1000385. DOI: 10.1371/journal.pbio.1000385

Bonn, A., Richter, A., Vohland, K., Pettibone, L., Brandt, M., Feldmann, R., Goebel, C., Grefe, C., Hecker, S., Hennen, L., Hofer, H., Kiefer, S., Klotz, S., Kluttig, T., Krause, J., Küsel, K., Liedtke, C., Mahla, A., Neumeier, V., Premke-Kraus, M., Rillig, M.C., Röller, O., Schäffler, L., Schmalzbauer, B., Schneidewind, U., Schumann, A., Settele, J., Tochtermann, K., Tockner, K., Vogel, J., Volkmann, W., von Unger, H., Walter, D., Weisskopf, M., Wirth, C., Witt, T., Wolst, D., Ziegler, D., 2016. Grünbuch Citizen Science Strategie 2020 für Deutschland.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., Wilderman, C., 2009. Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Center for Advancement of Informal Science Education, Washington, DC.

Bonter, D.N., Cooper, C.B., 2012. Data validation in citizen science: a case study from Project FeederWatch. Frontiers in Ecology and the Environment 10, 305–307. DOI: 10.1890/110273

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., Rampini, A., 2016a. On predicting and improving the quality of Volunteer Geographic Information projects. International Journal of Digital Earth 9, 134–155. DOI: 10.1080/17538947.2014.976774

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., Rampini, A., 2014. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. Information Sciences 258, 312–327. DOI: 10.1016/j.ins.2013.07.013

Bordogna, G., Frigerio, L., Kliment, T., Brivio, P., Hossard, L., Manfron, G., Sterlacchini, S., 2016b. “Contextualized VGI” Creation and Management to Cope with Uncertainty and Imprecision. ISPRS International Journal of Geo-Information 5, 234. DOI: 10.3390/ijgi5120234

Brenton, P., von Gavel, S., Vogel, E., Lecoq, M.-E., 2018. Technology infrastructure for citizen science, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 63–80.

Budhathoki, N.R., Bruce, B. (Chip), Nedovic-Budic, Z., 2008. Reconceptualizing the role of the user of spatial data infrastructure. GeoJournal 72, 149–160. DOI: 10.1007/s10708-008-9189-x

Budhathoki, N.R., Haythornthwaite, C., 2013. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. American Behavioral Scientist 57, 548–575. DOI: 10.1177/0002764212469364

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.-F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Morcillo, M.H., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vie, J.-C., Watson, R., 2010. Global Biodiversity: Indicators of Recent Declines. Science 328, 1164–1168. DOI: 10.1126/science.1187512

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. DOI: 10.1126/sciadv.1400253

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. DOI: 10.1016/j.biocon.2016.09.004

Chapman, A.D., 2005. Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Colwell, R.K., Mao, C.X., Chang, J., 2004. Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves. Ecology 85, 2717–2727. DOI: 10.1890/03-0557

Connors, J.P., Lei, S., Kelly, M., 2012. Citizen Science in the Age of Neogeography: Utilizing Volunteered Geographic Information for Environmental Monitoring. Annals of the Association of American Geographers 102, 1267–1289. DOI: 10.1080/00045608.2011.627058

Conrad, C.C., Hilchey, K.G., 2011. A review of citizen science and community-based environmental monitoring: issues and opportunities. Environmental Monitoring and Assessment 176, 273–291. DOI: 10.1007/s10661-010-1582-5

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J., Waller, D.M., 2011. Assessing citizen science data quality: an invasive species case study. Conservation Letters 4, 433–442. DOI: 10.1111/j.1755-263X.2011.00196.x

D'Antonio, F., Fogliaroni, P., Kauppinen, T., 2014. VGI Edit History Reveals Data Trustworthiness and User Reputation, in: Connecting a Digital Europe Through Location and Place. Proceedings of the 17th AGILE International Conference on Geographic Information Science, Castellón, June, 3-6, 2014.

Dennis, R.L.H., Sparks, T.H., Hardy, P.B., 1999. Bias in butterfly distribution maps: the effects of sampling effort. Journal of Insect Conservation 3, 33–42.

Dennis, R.L.H., Thomas, C.D., 2000. Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. Journal of Insect Conservation 4, 73–77. DOI: 10.1023/A:1009690919835

Dickinson, J.L., Bonney, R. (Eds.), 2012. Citizen science: public participation in environmental research. Comstock Pub. Associates, Ithaca.

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. DOI: 10.1890/110236

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annual Review of Ecology, Evolution, and Systematics 41, 149–172. DOI: 10.1146/annurev-ecolsys-102209-144636

Ebert, G., Rennwald, E. (Eds.), 1991. Die Schmetterlinge Baden-Württembergs. Verlag Eugen Ulmer, Stuttgart.

Eckle, M., Porto de Albuquerque, J., 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes, in: Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24-27.

ECSA, European Citize Science Association, 2015. Ten principles of citizen science. London.

Ehrlich, P.R., 1988. The loss if biodiversity: causes and consequences, in: Wilson, E.O., Peter, F.M., National Academy of Sciences (U.S.), Smithsonian Institution (Eds.), Biodiversity. National Academy Press, Washington, D.C, pp. 21–27.

Elwood, S., Goodchild, M.F., Sui, D.Z., 2012. Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. Annals of the Association of American Geographers 102, 571–590. DOI: 10.1080/00045608.2011.595657

Encarnaçao, J.A., Nöding, J., Reiners, T.E. und N.I.B., 2012. Ehrenamtlich erhobene Daten verbessern hessenweite Verbreitungsmodelle der FFH-relevanten Haselmaus (Muscardinus avellanarius). Natur und Landschaft 87, 208–214.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. International Journal of Geographical Information Science 28, 700–719. DOI: 10.1080/13658816.2013.867495

Finke, P., 2014. Citizen Science: Das unterschätzte Wissen der Laien. Oekom, München.

Flanagin, A.J., Metzger, M.J., 2008. The credibility of volunteered geographic information. GeoJournal 72, 137–148. DOI: 10.1007/s10708-008-9188-y

Fligner, M.A., Killeen, T.J., 1976. Distribution-Free Two-Sample Tests for Scale. Journal of the American Statistical Association 71, 210–213.

Foody, G., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C., Boyd, D.S., 2013. Assessing the Accuracy of Volunteered Geographic Information arising from Multiple Contributors to an Internet Based Collaborative Project. Transactions in GIS 17, 847–860. DOI: 10.1111/tgis.12033

Franke, N., Eissing, H., 2010. Naturschutz und Ehrenamt im 21. Jahrundert: Potenziale, Optionen und Strategien. Natur und Landschaft 85, 24–26.

Freitag, A., Meyer, R., Whiteman, L., 2016. Strategies Employed by Citizen Science Programs to Increase the Credibility of Their Data. Citizen Science: Theory and Practice. 1, 2. DOI: 10.5334/cstp.6

Goodchild, M.F., 2013. The quality of big (geo)data. Dialogues in Human Geography 3, 280–284. DOI: 10.1177/2043820613513392

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. DOI: 10.1007/s10708-007-9111-y

Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spatial Statistics 1, 110–120. DOI: 10.1016/j.spasta.2012.03.002

Gouveia, C., Fonseca, A., Câmara, A., Ferreira, F., 2004. Promoting the use of environmental data collected by concerned citizens through information and communication technologies. Journal of Environmental Management 71, 135–154. DOI: 10.1016/j.jenvman.2004.01.009

Grey, F., 2009. Viewpoint: The age of citizen cyberscience. CERN Courier 29.

Haklay, M., 2013. Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation, in: Sui, D., Elwood, S., Goodchild, M. (Eds.), Crowdsourcing Geographic Knowledge. Springer Netherlands, Dordrecht, pp. 105–122. DOI: 10.1007/978-94-007-4587-2_7

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design 37, 682–703. DOI: 10.1068/b35097

Haklay, M., Singleton, A., Parker, C., 2008. Web Mapping 2.0: The Neogeography of the GeoWeb. Geography Compass 2, 2011–2039. DOI: 10.1111/j.1749-8198.2008.00167.x

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 & Evolution 27, 130–137. DOI: 10.1016/j.tree.2011.11.006

Howe, J., 2006. The rise of crowdsourcing. Wired 14.

Idris, N.H., Jackson, M.J., Ishak, M.H.I., 2014. A conceptual model of the automated credibility assessment of the volunteered geographic information. 8th International Symposium of the Digital Earth (ISDE8), IOP Conf. Series: Earth and Environmental Science 18. DOI: 10.1088/1755-1315/18/1/012070

Irwin, A., 1995. Citizen Science: A study of people, expertise and sustainable development. Routledge, London.

Isaac, N.J.B., Pocock, M.J.O., 2015. Bias and information in biological records. Biological Journal of the Linnean Society 115, 522–531. DOI: 10.1111/bij.12532

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 5, 1052–1060. DOI: 10.1111/2041-210X.12254

Jackson, M.M., Gergel, S.E., Martin, K., 2015. Citizen science and field survey observations provide comparable results for mapping Vancouver Island White-tailed Ptarmigan (Lagopus leucura saxatilis) distributions. Biological Conservation 181, 162–172. DOI: 10.1016/j.biocon.2014.11.010

Jacobs, C., 2016. Data quality in crowdsourcing for biodiversity research: issues and examples, in: Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., Purves, R. (Eds.), European Handbook of Crowdsourced Geographic Information. Ubiquity Press, London, pp. 75–86.

Jacobs, C., Schotthöfer, A., 2015. Citizen-Science-Daten zur Biodiversität - Methoden zur Unterstützung der Qualitätssicherung, in: Strobl, J., Zagel, B., Griesebner, G., Blaschke, T. (Eds.), AGIT - Journal Für Angewandte Geoinformatik. pp. 470–479.

Jacobs, C., Zipf, A., 2017. Completeness of citizen science biodiversity data from a volunteered geographic information perspective. Geo-spatial Information Science 20, 3–13. DOI: 10.1080/10095020.2017.1288424

Johnson, C.N., Balmford, A., Brook, B.W., Buettel, J.C., Galetti, M., Guangchun, L., Wilmshurst, J.M., 2017. Biodiversity losses and conservation responses in the Anthropocene. Science 356, 270–275. DOI: 10.1126/science.aam9317

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 and Distributions 22, 1024–1035. DOI: 10.1111/ddi.12463

Kampen, H., Medlock, J.M., Vaux, A., Koenraadt, C., van Vliet, A., Bartumeus, F., Oltra, A., Sousa, C.A., Chouin, S., Werner, D., 2015. Approaches to passive mosquito surveillance in the EU. Parasites & Vectors 8, 9. DOI: 10.1186/s13071-014-0604-5

Keßler, C., de Groot, R.T.A., 2013. Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap, in: Vandenbroucke, D., Bucher, B., Crompvoets, J. (Eds.), Geographic Information Science at the Heart of Europe. Springer International Publishing, Cham, pp. 21–37. DOI: 10.1007/978-3-319-00615-4_2

Keßler, C., Maué, P., Heuer, J.T., Bartoschek, T., 2009. Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags, in: Janowicz, K., Raubal, M., Levashkin, S. (Eds.), GeoSpatial Semantics. Springer, Berlin, Heidelberg, pp. 83–102. DOI: 10.1007/978-3-642-10436-7_6

Kettunen, J., Silander, J., Lindholm, M., Lehtiniemi, M., Setälä, O., Kaitala, S., 2016. Changing role of citizens in national environmental monitoring, in: Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., Purves, R. (Eds.), European Handbook of Crowdsourced Geographic Information. Ubiquity Press, pp. 257–267. DOI: 10.5334/bax.s

Koleff, P., Gaston, K.J., Lennon, J.J., 2003. Measuring beta diversity for presence-absence data. Journal of Animal Ecology 72, 367–382. DOI: 10.1046/j.1365-2656.2003.00710.x

Koperski, K., Han, J., 1995. Discovery of spatial association rules in geographic information databases, in: Egenhofer, M.J., Herring, J.R. (Eds.), Advances in Spatial Databases. Springer, Berlin, Heidelberg, pp. 47–66. DOI: 10.1007/3-540-60159-7_4

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14, 551–560. DOI: 10.1002/fee.1436

Legendre, P., Legendre, L., 1998. Numerical ecology, 2nd Engl. ed, Developments in environmental modelling. Elsevier, Amsterdam, New York.

Lennon, J.J., Koleff, P., GreenwooD, J.J.D., Gaston, K.J., 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. Journal of Animal Ecology 70, 966–979. DOI: 10.1046/j.0021-8790.2001.00563.x

Liu, H.-Y., Grossberndt, S., Kobernus, M., 2017. Citizen Science and Citizens' Observatories: Trends, Roles, Challenges and Development Needs for Science and Environmental Governance, in: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V. (Eds.), Mapping and the Citizen Sensor. Ubiquity Press, pp. 351–376. DOI: 10.5334/bbf.o

MacFarland, T.W., Yates, J.M., 2016. Introduction to Nonparametric Statistics for the Biological Sciences Using R. Springer International Publishing, Cham. DOI: 10.1007/978-3-319-30634-6

Mann, H.B., Whitney, D.R., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics 18, 50–60. DOI: 10.1214/aoms/1177730491

Mazumdar, S., Ceccaroni, L., Piera, J., Hölker, F., Berre, A.J., Arlinghaus, R., Bowser, A., 2018. Citizen science technologies and new opportunities for participation, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 303–320.

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. DOI: 10.1890/110278

Minghini, M., Antoniou, V., Fonte, C.C., Estima, J., Olteanu-Raimond, A.-M., See, L., Laakso, M., Skopeliti, A., Mooney, P., Arsanjani, J.J., Lupia, F., 2017. The Relevance of Protocols for VGI Collection, in: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V. (Eds.), Mapping and the Citizen Sensor. Ubiquity Press, pp. 223–247. DOI: 10.5334/bbf.j

Mocnik, F.-B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., Klonner, C., 2018. A grounding-based ontology of data quality measures. Journal of Spatial Information Science. DOI: 10.5311/JOSIS.2018.16.360

Mooney, P., Corcoran, P., 2012a. How social is OpenStreetMap?, in: Bridging the Geographic Information Science. Proceedings of the 15th AGILE International Conference on Geographic Information Science, Avignon, April, 24-27, 2012.

Mooney, P., Corcoran, P., 2012b. Characteristics of Heavily Edited Objects in OpenStreetMap. Future Internet 4, 285–305. DOI: 10.3390/fi4010285

Mülligann, C., Janowicz, K., Ye, M., Lee, W.-C., 2011. Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information, in: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (Eds.), Spatial Information Theory. Springer, Berlin, Heidelberg, pp. 350–370. DOI: 10.1007/978-3-642-23196-4_19

Munzinger, S., Ott, J., Schulemann-Maier, G., Strub, O., 2017. Citizen-Science-Beobachtungsdaten, Teil 2: Theorie der Plausibilisierung. Naturschutz und Landschaftsplanung 49, 229–235.

Neis, P., Zielstra, D., 2014. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. Future Internet 6, 76–106. DOI: 10.3390/fi6010076

Neis, P., Zielstra, D., Zipf, A., 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet 4, 1–21. DOI: 10.3390/fi4010001

Neis, P., Zipf, A., 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. ISPRS International Journal of Geo-Information 1, 146–165. DOI: 10.3390/ijgi1020146

Netzwerk Phytodiversität Deutschland, Bundesamt für Naturschutz (Eds.), 2013. Verbreitungsatlas der Farn- und Blütenpflanzen Deutschlands. Landwirtschaftsverlag, Münster.

Newell, D.A., Pembroke, M.M., Boyd, W.E., 2012. Crowd Sourcing for Conservation: Web 2.0 a Powerful Tool for Biologists. Future Internet 4, 551–562. DOI: 10.3390/fi4020551

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. DOI: 10.1890/110294

Newman, G., Zimmerman, D., Crall, A., Laituri, M., Graham, J., Stapel, L., 2010. User-friendly web mapping: lessons from a citizen science website. International Journal of Geographical Information Science 24, 1851–1869. DOI: 10.1080/13658816.2010.490532

Novacek, M.J., 2008. Engaging the public in biodiversity issues. Proceedings of the National Academy of Sciences 105, 11571–11578. DOI: 10.1073/pnas.0802599105

Novacek, M.J., Cleland, E.E., 2001. The current biodiversity extinction event: Scenarios for mitigation and recovery. Proceedings of the National Academy of Sciences 98, 5466–5470. DOI: 10.1073/pnas.091093698

Ostermann, F.O., Spinsanti, L., 2011. A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management, in: Advancing Geoinformation Science for a Changing World. Proceedings of the 14th AGILE International Conference on Geographic Information Science, Utrecht, April, 18-21, 2011.

O'Sullivan, D., Unwin, D., 2010. Geographic information analysis, 2nd ed. John Wiley & Sons, Hoboken, N.J.

Ott, D., Frank, D., Schotthöfer, A., Willigalla, C., 2017. Libellen in Rheinland-Pfalz - beobachten und erkennen. Pollichia (Eigenverlag), Neustadt an der Weinstraße.

Owen, R., Parker, A.J., 2018. Citizen science in environmental protection agencies, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 284–300.

Pereira, H.M., Leadley, P.W., Proença, V., Alkemade, R., Scharlemann, J.P.W., Fernandez-Manjarrés, J.F., Araújo, M.B., Balvanera, P., Biggs, R., Cheung, W.W.L., Chini, L., Cooper, H.D., Gilman, E.L., Guénette, S., Hurtt, G.C., Huntington, H.P., Mace, G.M., Oberdorff, T., Revenga, C., Rodrigues, P., Scholes, R.J., Sumaila, U.R., Walpole, M., 2010. Scenarios for Global Biodiversity in the 21st Century. Science 330, 1496–1501. DOI: 10.1126/science.1196624

Pettorelli, N., Laurance, W.F., O'Brien, T.G., Wegmann, M., Nagendra, H., Turner, W., 2014. Satellite remote sensing for applied ecologists: opportunities and challenges. Journal of Applied Ecology 51, 839–848. DOI: 10.1111/1365-2664.12261

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software 79, 214–232. DOI: 10.1016/j.envsoft.2016.02.008

Pocock, M.J.O., Chandler, M., Bonney, R., Thornhill, I., Albin, A., August, T., Bachman, S., Brown, P.M.J., Cunha, D.G.F., Grez, A., Jackson, C., Peters, M., Rabarijaon, N.R., Roy, H.E., Zaviezo, T., Danielsen, F., 2018. A Vision for Global Biodiversity Monitoring With Citizen Science, in: Advances in Ecological Research. Elsevier, pp. 169–223. DOI: 10.1016/bs.aecr.2018.06.003

Pocock, M.J.O., Tweddle, J.C., Savage, J., Robinson, L.D., Roy, H.E., 2017. The diversity and evolution of ecological and environmental citizen science. PLOS ONE 12, e0172579. DOI: 10.1371/journal.pone.0172579

Powney, G.D., Isaac, N.J.B., 2015. Beyond maps: a review of the applications of biological records. Biological Journal of the Linnean Society 115, 532–542. DOI: 10.1111/bij.12517

Proença, V., Martin, L.J., Pereira, H.M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R.D., Honrado, J.P., Jürgens, N., Opige, M., Schmeller, D.S., Tiago, P., van Swaay, C.A.M., 2017. Global biodiversity monitoring: From data sources to Essential Biodiversity Variables. Biological Conservation 213, 256–263. DOI: 10.1016/j.biocon.2016.07.014

Resch, B., 2013. People as Sensors and Collective Sensing - Contextual Observations Complementing Geo-Sensor Network Measurements, in: Krisp, J.M. (Ed.), Progress in Location-Based Services. Springer, Berlin, Heidelberg, pp. 391–406. DOI: 10.1007/978-3-642-34203-5_22

Richter, A., Dörler, D., Hecker, S., Heigl, F., Pettibone, L., Sanz, F.S., Vohland, K., Bonn, A., 2018. Capacity building in citizen science, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 269–283.

Ripley, B.D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability 13, 255–266. DOI: 10.2307/3212829

Robinson, L.D., Cawthray, J.L., West, S.E., Bonn, A., Ansine, J., 2018. Ten principles of citizen science, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 27–40.

Robinson, O.J., Ruiz-Gutierrez, V., Fink, D., 2018. Correcting for bias in distribution modelling for rare species using citizen science data. Diversity and Distributions 24, 460–472. DOI: 10.1111/ddi.12698

Roick, O., Hagenauer, J., Zipf, A., 2011. OSMatrix - Grid-based Analysis and Visualization of OpenStreetMap, in: Proceedings of the State of the Map Conference 2011. Denver, CO.

Röller, O., 2015. Citizen Science - Neue Möglichkeiten für Naturforschung und Naturschutz in Deutschland. Pollichia (Eigenverlag), Neustadt an der Weinstraße.

Rossiter, D.G., Liu, J., Carlisle, S., Zhu, A.-X., 2015. Can citizen science assist digital soil mapping? Geoderma 259–260, 71–80. DOI: 10.1016/j.geoderma.2015.05.006

Rößner, R., Helb, H.-W., Schotthöfer, A., Röller, O., 2013. Vögel in Rheinland-Pfalz - beobachten und erkennen. Pollichia (Eigenverlag), Neustadt an der Weinstraße.

Samy, G., 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. DOI: 10.17161/bi.v8i2.4124

Sayre, R., Dangermond, J., Frye, C., Vaughan, R., Aniello, P., Breyer, S., Cribbs, D., Hopkins, D., Nauman, R., Derrenbacher, W., Wright, D., Brown, C., Convis, C., Smith, J., Benson, L., Paco VanSistine, D., Warner, H., Cress, J., Danielson, J., Hamann, S., Cecere, T., Reddy, A., Burton, D., Grosse, A., True, D., Metzger, M., Hartmann, J., Moosdorf, N., Dürr, H., Paganini, M., DeFourny, P., Arino, O., Maynard, S., Anderson, M., Comer, P. (Eds.), 2014. A new map of global ecological land units - an ecophysiographic stratification approach. Association of American Geographers, Washington, DC.

Schotthöfer, A., Scheydt, N., Blum, E., Röller, O., 2014. Tagfalter in Rheinland-Pfalz - beobachten und erkennen. Pollichia (Eigenverlag), Neustadt an der Weinstraße.

Schrauth, F., Wink, M., 2018. Changes in Species Composition of Birds and Declining Number of Breeding Territories over 40 Years in a Nature Conservation Area in Southwest Germany. Diversity 10, 97. DOI: 10.3390/d10030097

Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017. Open land cover from OpenStreetMap and remote sensing. International Journal of Applied Earth Observation and Geoinformation 63, 206–213. DOI: 10.1016/j.jag.2017.07.014

See, L., Estima, J., Podör, A., Arsanjani, J.J., Laso Bayas, J.-C., Vatseva, R., 2017. Sources of VGI for Mapping, in: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C.C., Antoniou, V. (Eds.), Mapping and the Citizen Sensor. Ubiquity Press, pp. 13–35. DOI: 10.5334/bbf.b

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., Rutzinger, M., 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS International Journal of Geo-Information 5, 55. DOI: 10.3390/ijgi5050055

Shirk, J.L., Ballard, H.L., Wilderman, C.C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B.V., Krasny, M.E., Bonney, R., 2012. Public Participation in Scientific Research: a Framework for Deliberate Design. Ecology and Society 17. DOI: 10.5751/ES-04705-170229

Silvertown, J., 2009. A new dawn for citizen science. Trends in Ecology and Evolution 24, 467–471.

Simpson, E.H., 1949. Measurement of Diversity. Nature 163, 688–688. DOI: 10.1038/163688a0

Simpson, G.G., 1943. Mammals and the nature of continents. American Journal of Science 241, 1–31. DOI: 10.2475/ajs.241.1.1

Snäll, T., Forslund, P., Jeppsson, T., Lindhe, A., O'Hara, R.B., 2014. Evaluating temporal variation in Citizen Science Data against temporal variation in the environment. Ecography 37, 293–300. DOI: 10.1111/j.1600-0587.2011.00544.x

Spyratos, S., Lutz, M., Pantisano, F., 2014. Characteristics of Citizen-contributed Geographic Information, in: Connecting a Digital Europe Through Location and Place. Proceedings of the 17th AGILE International Conference on Geographic Information Science, Castellón, June, 3-6, 2014.

Steinbauer, M.J., Grytnes, J.-A., Jurasinski, G., Kulonen, A., Lenoir, J., Pauli, H., Rixen, C., Winkler, M., Bardy-Durchhalter, M., Barni, E., Bjorkman, A.D., Breiner, F.T., Burg, S., Czortek, P., Dawes, M.A., Delimat, A., Dullinger, S., Erschbamer, B., Felde, V.A., Fernández-Arberas, O., Fossheim, K.F., Gómez-García, D., Georges, D., Grindrud, E.T., Haider, S., Haugum, S.V., Henriksen, H., Herreros, M.J., Jaroszewicz, B., Jaroszynska, F., Kanka, R., Kapfer, J., Klanderud, K., Kühn, I., Lamprecht, A., Matteodo, M., di Cella, U.M., Normand, S., Odland, A., Olsen, S.L., Palacio, S., Petey, M., Piscová, V., Sedlakova, B., Steinbauer, K., Stöckli, V., Svenning, J.-C., Teppa, G., Theurillat, J.-P., Vittoz, P., Woodin, S.J., Zimmermann, N.E., Wipf, S., 2018. Accelerated increase in plant species richness on mountain summits is linked to warming. Nature 556, 231–234. DOI: 10.1038/s41586-018-0005-6

Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J.W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W.M., Iliff, M.J., Lagoze, C., La Sorte, F.A., Merrifield, M., Morris, W., Phillips, T.B., Reynolds, M., Rodewald, A.D., Rosenberg, K.V., Trautmann, N.M., Wiggins, A., Winkler, D.W., Wong, W.-K., Wood, C.L., Yu, J., Kelling, S., 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation 169, 31–40. DOI: 10.1016/j.biocon.2013.11.003

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. DOI: 10.1016/j.biocon.2009.05.006

Tittensor, D.P., Walpole, M., Hill, S.L.L., Boyce, D.G., Britten, G.L., Burgess, N.D., Butchart, S.H.M., Leadley, P.W., Regan, E.C., Alkemade, R., Baumung, R., Bellard, C., Bouwman, L., Bowles-Newark, N.J., Chenery, A.M., Cheung, W.W.L., Christensen, V., Cooper, H.D., Crowther, A.R., Dixon, M.J.R., Galli, A., Gaveau, V., Gregory, R.D., Gutierrez, N.L., Hirsch, T.L., Hoft, R., Januchowski-Hartley, S.R., Karmann, M., Krug, C.B., Leverington, F.J., Loh, J., Lojenga, R.K., Malsch, K., Marques, A., Morgan, D.H.W., Mumby, P.J., Newbold, T., Noonan-Mooney, K., Pagad, S.N., Parks, B.C., Pereira, H.M., Robertson, T., Rondinini, C., Santini, L., Scharlemann, J.P.W., Schindler, S., Sumaila, U.R., Teh, L.S.L., van Kolck, J., Visconti, P., Ye, Y., 2014. A mid-term analysis of progress toward international biodiversity targets. Science 346, 241–244. DOI: 10.1126/science.1257484

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013. Realising the full potential of citizen science monitoring programs. Biological Conservation 165, 128–138. DOI: 10.1016/j.biocon.2013.05.025

Turner, A., 2006. Introduction to Neogeography. O'Reilly.

U.S. Census Bureau, 2012. 2010 Census of Population and Housing, Population and Housing Unit Counts, CPH-2-6, California. U.S. Government Printing Office, Washington, DC.

U.S. Geological Survey, 2016. National Boundary Dataset (NBD).

Vahidi, H., Klinkenberg, B., Yan, W., 2018. Trust as a proxy indicator for intrinsic quality of Volunteered Geographic Information in biodiversity monitoring programs. GIScience & Remote Sensing 55, 502–538. DOI: 10.1080/15481603.2017.1413794

van Strien, A.J., van Swaay, C.A.M., Termaat, T., 2013. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Journal of Applied Ecology 50, 1450–1458. DOI: 10.1111/1365-2664.12158

Walz, U., Bastian, O., Kästner, A., Wende, W., Schwarze, H., 2013. Situation des Ehrenamtes im Naturschutz: Ergebnisse einer Studie in Sachsen. Naturschutz und Landschaftsplanung 45, 233–240.

Whittaker, R.H., 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30, 279–338. DOI: 10.2307/1943563

Wiersma, Y.F., 2010. Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. Avian Conservation and Ecology 5. DOI: 10.5751/ACE-00427-050213

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., LeBuhn, G., Littauer, R., Lotts, K., Michener, W., Newman, G., Russel, E., Stevenson, R., Weltzin, J., 2013. Data Management Guide for Public Participation in Scientific Research. DataOne Public Participation in Scientific Research Working Group.

Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K., 2011. Mechanisms for Data Quality and Validation in Citizen Science. Presented at the "Computing for Citizen Science" workshop at the IEEE eScience Conference, Stockholm.

Williams, L., Chapman, C., Leibovici, D.G., Lois, G., Matheus, A., Oggioni, A., Schade, S., See, L., van Genuchten, P.P.L., 2018. Maximising the impact and reuse of citizen science data, in: Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (Eds.), Citizen Science: Innovation in Open Science, Society and Policy. UCL Press, London, pp. 321–336.

Williams, P.H., Margules, C.R., Hilbert, D.W., 2002. Data requirements and data sources for biodiversity priority area selection. Journal of Biosciences 27, 327–338. DOI: 10.1007/BF02704963

Wilmanns, O., 1998. Ökologische Pflanzensoziologie: eine Einführung in die Vegetation Mitteleuropas, UTB für Wissenschaft Uni-Taschenbücher Botanik/Ökologie. Quelle & Meyer, Wiesbaden.

Wilson, E.O. (Ed.), 1988. Biodiversity. National Acad. Press, Washington, DC.

Wink, M., 2017. Citizen Science in der Biologie - Schwerpunkt Ornithologie. Heidelberger Jahrbücher Online, Bd. 2 (2017): Wissenschaft für alle: Citizen Science. DOI: 10.17885/heiup.hdjbo.2017.0.23695

Wink, M., 1987. Die Vögel des Rheinlandes. Band 3: Atlas der Brutvogelverbreitung im Rheinland, Beiträge zur Avifauna des Rheinlandes 25-26. Kilda-Verlag, Düsseldorf.

Wittig, R., Niekisch, M., 2014. Biodiversität: Grundlagen, Gefährdung, Schutz. Springer Spektrum, Berlin Heidelberg.

World Meteorological Organisation, 2001. Volunteers for Weather, Climate and Water (No. 919). World Meteorological Organisation (WMO), Geneva, Switzerland.

Yan, Y., Feng, C.-C., Wang, Y.-C., 2017. Utilizing fuzzy set theory to assure the quality of volunteered geographic information. GeoJournal 82, 517–532. DOI: 10.1007/s10708-016-9699-x

Zielstra, D., Hochmair, H.H., 2011. Comparative Study of Pedestrian Accessibility to Transit Stations Using Free and Proprietary Network Data. Transportation Research Record: Journal of the Transportation Research Board 2217, 145–152. DOI: 10.3141/2217-18

Zielstra, D., Hochmair, H.H., Neis, P., 2013. Assessing the Effect of Data Imports on the Completeness of OpenStreetMap - A United States Case Study. Transactions in GIS 17, 315–334. DOI: 10.1111/tgis.12037

Zielstra, D., Zipf, A., 2010. A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany, in: Geospatial Thinking. Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimaraes, May, 11-14, 2010.

Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analysing ecological data, Statistics for biology and health. Springer, New York ; London.

7 Appendix

Contents of the Appendix

7.1 Observed Communities Approach: Species Pairs for _SI1 Sets of Synthetic Implausible Observations

7.1.1 ArtenFinder

Table 7.1.1: ArtenFinder, species pairs used for candidate set AF_SI1. (Physically similar species living in different habitats used for creating synthetic implausible observations by swapping species identifications between accepted observations.)

Species 1	**Species 2**
Alauda arvensis	*Lullula arborea*
Anthus pratensis	*Anthus trivialis*
Apatura ilia	*Apatura iris*
Athene noctua	*Strix aluco*
Boloria dia	*Brenthis ino*
Boloria eunomia	*Boloria selene*
Boloria euphrosyne	*Boloria selene*
Brintesia circe	*Limenitis camilla*
Certhia familiaris	*Certhia brachydactyla*
Coenagrion puella	*Coenagrion scitulum*
Conocephalus dorsalis	*Metrioptera bicolor*
Cornu aspersum	*Helix pomatia*
Corvus corax	*Corvus corone*
Emberiza cirlus	*Emberiza citrinella*
Galerida cristata	*Lullula arborea*
Lacerta agilis	*Podarcis muralis*
Libellula fulva	*Orthetrum coerulescens*
Lycaena alciphron	*Lycaena hippothoe*
Lycaena dispar	*Lycaena virgaureae*
Melitaea athalia	*Melitaea aurelia*
Motacilla cinerea	*Motacilla flava*
Orthetrum brunneum	*Orthetrum coerulescens*
Parus montanus	*Parus palustris*
Phoenicurus ochruros	*Phoenicurus phoenicurus*
Podarcis muralis	*Zootoca vivipara*
Satyrium ilicis	*Satyrium acaciae*
Satyrium ilicis	*Thecla betulae*

7.1.2 iNaturalist

Table 7.1.2: iNaturalist, species pairs used for candidate set iNat_SI1. (Physically similar species living in different habitats used for creating synthetic implausible observations by swapping species identifications between accepted observations.)

Species 1	**Species 2**
Allium campanulatum	*Allium unifolium*
Arenaria melanocephala	*Tringa semipalmata*
Branta bernicla	*Branta hutchinsii*
Cistothorus palustris	*Thryomanes bewickii*
Clarkia gracilis	*Clarkia rubicunda*
Clarkia purpurea quadrivulnera	*Clarkia rubicunda*
Delphinium cardinale	*Delphinium nudicaule*
Fragaria chiloensis	*Fragaria virginiana*
Keckiella cordifolia	*Keckiella corymbosa*
Melospiza lincolnii	*Passerculus sandwichensis*
Melospiza melodia	*Passerculus sandwichensis*
Passerculus sandwichensis	*Zonotrichia atricapilla*
Passerculus sandwichensis	*Passerella iliaca*
Poecile atricapillus	*Poecile rufescens*
Tringa incana	*Tringa semipalmata*
Tringa incana	*Tringa melanoleuca*

7.2 OSM environments Approach: Tags

Table 7.2.1: OSM tags (key-value pairs) used in the OSM environments approach. (Continued next page.)

key	values
aeroway	aerodrome, airstrip, apron, fuel, gate, hangar, helipad, heliport, holding_position, launchpad, marking, navigationaid, parking_position, runway, spaceport, taxilane, taxiway, terminal, windsock
amenity	animal_breeding, animal_shelter, bbq, bench, biergarten, boat_rental, boat_sharing, boat_storage, bus_station, canoe_hire, car_pooling, cinema, coast_guard, coast_radar_station, college, community_centre, courthouse, feeding_place, ferry_terminal, festival_grounds, fire_station, flight_school, fountain, game_feeding, garages, grave_yard, hospital, hunting_stand, kindergarten, lavoir, library, life_ring, lifeboat_station, marketplace, monastery, park, parking, public_building, ranger_station, research_institute, sanatorium, school, shelter, stables, table, theatre, townhall, university, waste_basket, waste_disposal, waste_transfer_station, water, watering_place, winery, yacht_club
attraction	animal
barrier	bump_gate, cable_barrier, cattle_grid, city_wall, ditch, fence, gate, hampshire_gate, hedge, hedge_bank, horse_jump, horse_stile, kissing_gate, log, retaining_wall, sally_port, stile, tank_trap, wall, wire_fence, wood_fence
basin	detention, infiltration, retention
bridge	aqueduct, boardwalk, cantilever, covered, movable, pontoon, simple_brunnel, swing, trestle, viaduct, yes
building	allotment_house, apartments, barn, barrack, boathouse, brewery, bridge, bungalow, bunker, cabin, carport, cathedral, chapel, church, civic, college, commercial, condominium, conservatory, construction, cowshed, detached, dormitory, farm, farm_auxiliary, garage, garages, greenhouse, hangar, hospital, house, houseboat, hut, industrial, kindergarten, manufacture, mosque, office, parking, pavilion, public, residential, riding_hall, roof, ruins, school, semi, service, shed, shrine, slurry_tank, sports_hall, stable, stadium, static_caravan, stilt_house, sty, supermarket, synagogue, tech_cab, temple, terrace, train_station, transformer_tower, transportation, university, warehouse, yes
camp_site	basic, deluxe, standard
construction	bridleway, cycleway, footpath, footway, light_rail, living_street, minor, motorway, motorway_link, pedestrian, preserved, primary, primary_link, rail, residential, road, secondary, service, steps, tertiary, track, tram, trunk, trunk_link, unclassified, yes
crop	asparagus, bananas, barley, cassava, coffee, corn, flowers, grape, grass, hay, hop, maize, rape, rice, rye, strawberry, sugar, tea, wheat, yes
cutting	left, right, yes
embankment	yes
emergency	fire_water_pond, life_ring, lifeboat_station, lifeguard_place, lifeguard_platform, slipway, suction_point, water_rescue_station
ford	yes
geology	moraine, outcrop
golf	bunker, driving_range, fairway, hole, lateral_water_hazard, rough, tee, water_hazard
grassland	dehesa, dune, grey_dune, moor, pampas, prairie, puszta, savanna, steppe, veld

Table 7.2.1: Continued, OSM tags (key-value pairs) used in the OSM environments approach. (Continued next page.)

key	values
harbour	yes
highway	abandoned, bridleway, bus_guideway, bus_stop, byway, corridor, crossing, cycleway, elevator, emergency_access_point, emergency_bay, escape, footway, ford, give_way, incline, living_street, milestone, mini_roundabout, motorway, motorway_junction, motorway_link, no, passing_place, path, pedestrian, platform, primary, proposed, raceway, razed, residential, rest_area, road, secondary, service, services, speed_camera, speed_display, steps, stile, stop, street_lamp, tertiary, track, traffic_mirror, traffic_signals, trail, trunk, trunk_link, turning_circle, turning_loop, unclassified, unsurfaced, via_ferrata
historic	aqueduct, battlefield, bomb_crater, building, castle, city_gate, city_wall, citywalls, farm, fort, manor, mine, monastery, ruins, shieling, wreck
industrial	mine, port, salt_pond, scrap_yard, shipyard
intermittent	yes
landmark	cairn, cemetery, chimney, large_rock, rock_pinnacle, tower, windmotor
landuse	agriculture, allotments, apiary, aquaculture, basin, brownfield, building_site, cemetery, churchyard, civic_admin, commercial, construction, depot, farm, farmland, farmyard, field, fishfarm, flowerbed, forest, garages, grass, grave_yard, greenfield, greenhouse_horticulture, harbour, hop_garden, industrial, institutional, landfill, logging, meadow, military, observatory, orchard, pasture, peat_cutting, piste, plant_nursery, plantation, plot, pond, port, prison, quarry, railway, recreation ground, religious, reservoir, residential, retail, salt_pond, street, traffic_island, trees, turbary, utility, village_green, vineyard, wasteland, well, wellsite, winter_sports, wood
leisure	bandstand, bathing_place, bbq, beach_resort, bird_hide, common, dog_park, firepit, fishing, garden, golf_course, horse_riding, hot_spring, landscape_reserve, marina, maze, miniature_golf, nature_reserve, park, picnic, picnic_site, picnic_table, pitch, playground, recreation_ground, resort, sailing_club, shooting_ground, stadium, summer_camp, swimming_area, table_tennis_table, wildlife_hide
lock	yes
man_made	adit, antenna, beehive, breakwater, bridge, bunker_silo, buoy, cairn, campanile, cellar_entrance, chimney, clearcut, communications_tower, cooling_tower, cross, cutline, dike, dolphin, dovecote, dyke, embankment, frost_fan, gasometer, goods_conveyor, groyne, insect_hotel, levee, lighthouse, mast, nesting_site, offshore_platform, petroleum_well, pier, pillar, pipeline, pumping_station, quay, reservoir_covered, silo, snow_cannon, snow_fence, snow_net, spoil_heap, storage_tank, tell, tower, tunnel, utility_pole, wastewater_plant, water_tank, water_tower, water_well, water_works, watermill, wildlife_crossing, wildlife_opening, windmill, windpump, works
military	airfield, barracks, bunker, danger_area, exclusion_zone, naval_base, obstacle_course, range, training_area

Table 7.2.1: Continued, OSM tags (key-value pairs) used in the OSM environments approach.

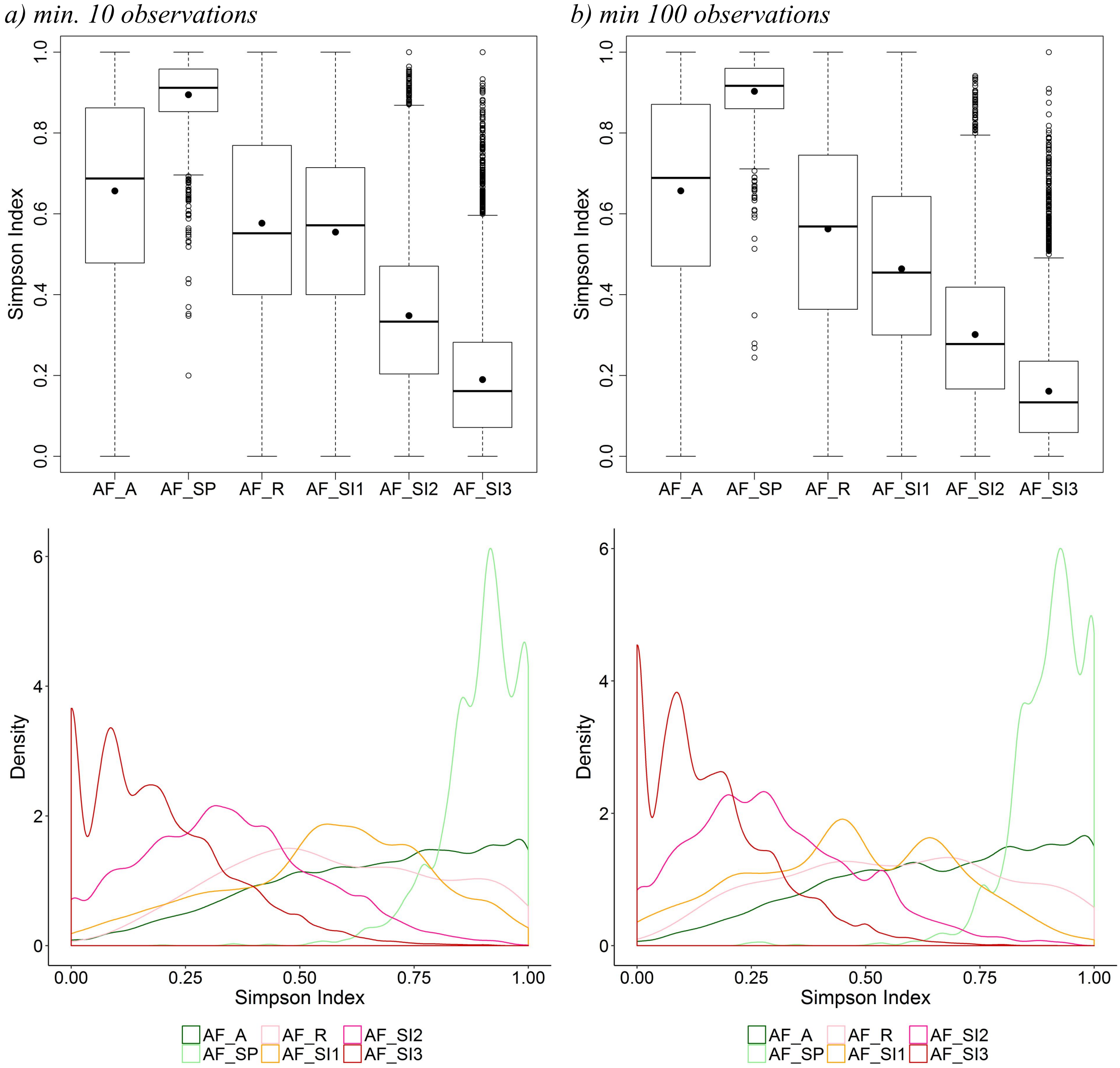
key	values
natural	arete, , avalanche_dam, bare_rock, bay, beach, bedrock, breaker, cape, cave, cave_entrance, cliff, coastline, continental_shelf, crater, crevasse, desert, dune, earth_bank, esker, fell, fjord, geothermal, geothermal_area, geothermal_field, geyser, glacier, grassland, gully, heath, hot_spring, lake, land, landform, landslide, lava, marsh, moor, moraine, mud, naled, peak, plant, reef, ridge, river_terrace, riverbed, rock, saddle, sand, scree, scrub, shingle, shoal, shrub, sinkhole, spring, stone, strait, tidal, tree, tree_group, tree_row, tundra, valley, volcano, wallow, water, wetland, wood
place	allotments, archipelago, atoll, city, city_block, farm, hamlet, island, islet, isolated_dwelling, sea, square, suburb, town, village
railway	abandoned, construction, dismantled, disused, preserved, rail
residential	rural, university, urban
route	bicycle, canal, ferry, fitness_trail, foot, hiking, horse, inline_skates, mtb, mudflat_hiking, nordic_walking, pipeline, piste, railway, road, running, ski, tracks, train, tram
sport	archery, baseball, beachvolleyball, bmx, canoe, chess, cliff_diving, climbing, climbing_adventure, cricket, croquet, cycling, diving, dog_racing, equestrian, free_flying, gaelic_games, golf, high_rope_course, horse_racing, karting, kitesurfing, model_aerodrome, motocross, motor, multi, orienteering, paintball, parachuting, paragliding, roller_skating, rowing, sailing, scuba_diving, shooting_range, skating, skiing, soccer, surfing, tennis, water_ski
surface	asphalt, chipseal, cobblestone, compacted, concrete, dirt, earth, fine_gravel, grass, grass_paver, gravel, ground, ice_road, metal, mud, paved, paving_stones, pebblestone, roman_paving, sand, sett, unpaved, wood
tidal	rocks, yes
tourism	alpine_hut, attraction, camp_site, caravan_site, chalet, picnic_site, resort, theme_park, trail_riding_station, viewpoint, wilderness_hut, zoo
tunnel	yes
wall	noise_barrier
water	canal, intermittent, lake, lock, pond, reservoir, river, salt_pool, tidal
waterway	brook, canal, construction, dam, derelict_canal, ditch, drain, drystream, fish_pass, mooring, rapids, river, riverbank, seaway, stream, stream_end, wadi, waterfall, weir
wetland	bog, fen, mangrove, marsh, mud, reedbed, saltern, saltmarsh, string_bog, swamp, tidalflat, wet_meadow
wood	coniferous, deciduous, eucalypt, evergreen, mixed, palm
zoo	aviary, birds, enclosure, falconry, petting_zoo, safari_park, wildlife_park

7.3 Sensitivity Analysis: Graphs and Tables

Results concerning effects of parameter changes and methodological modifications were described in section 4.3, and discussed in section 5.3. The following figures and tables present detailed graphic and numeric results and put them in contrast to evaluation results with basic parameter settings and basic methodology.

In almost all comparisons between distributions of similarity values, statistical differences determined with the Mann-Whitney-U-Test (see section 3.3.1) were significant on the $p \leq 0.05$ level, and variances determined with the Fligner-Killeen-Test (see also section 3.3.1) were not homogeneous. P-values are therefore not listed here in detail.

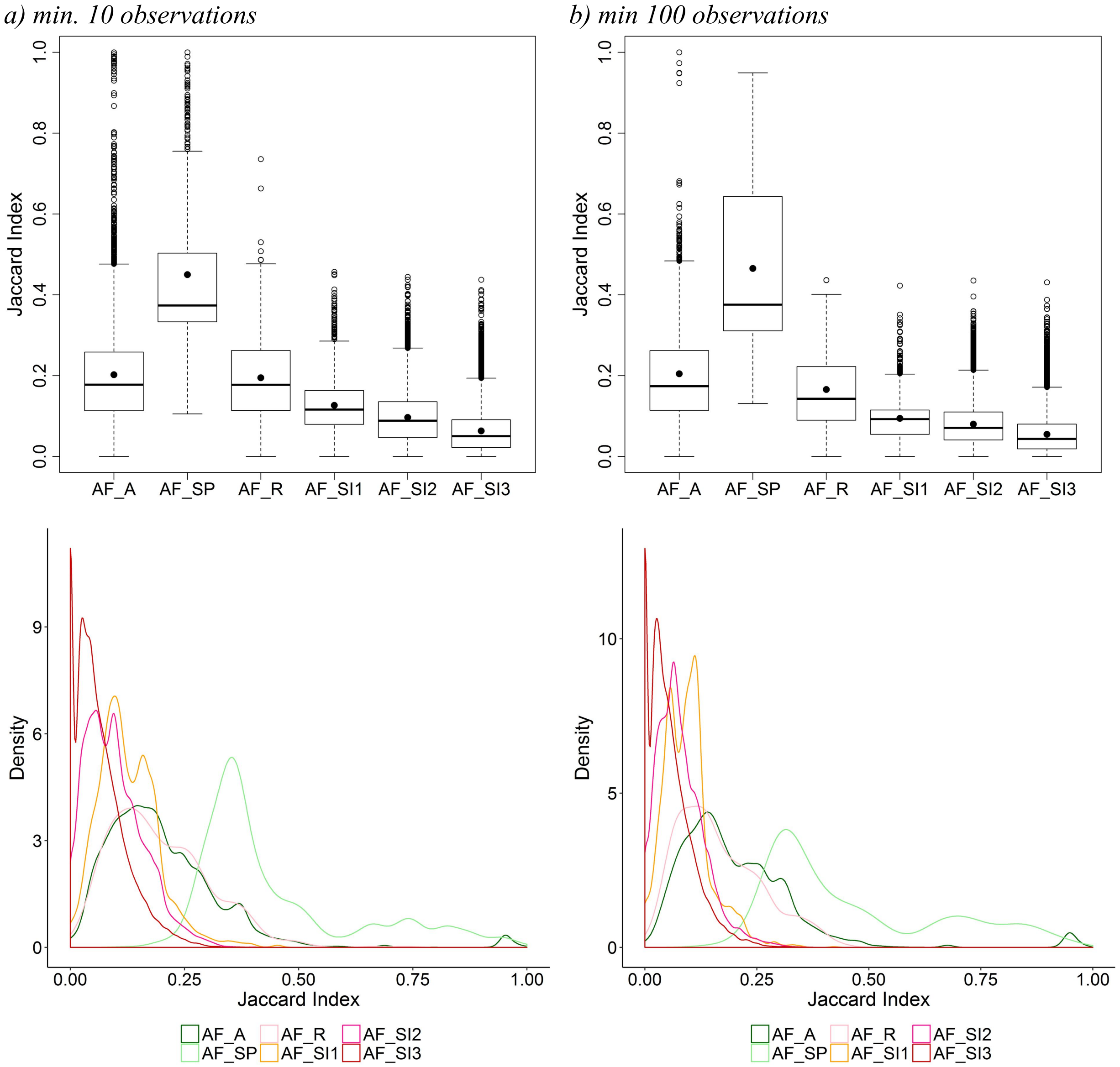
7.3.1 Using a Lower Minimum Requirement for Approved Observations in Observed Community extraction



a) min. 10 observations

b) min 100 observations

Figure 7.3.1: ArtenFinder, distributions of Simpson similarity index values, analysis with smaller minimum number of 10 target species observations in observed community extraction (a,) and more conservative minimum number of 100 observations (b).



a) min. 10 observations

b) min 100 observations

Figure 7.3.2: ArtenFinder, distributions of Jaccard similarity index values, analysis with smaller minimum number of 10 target species observations in observed community extraction (a), and more conservative minimum number of 100 observations (b).

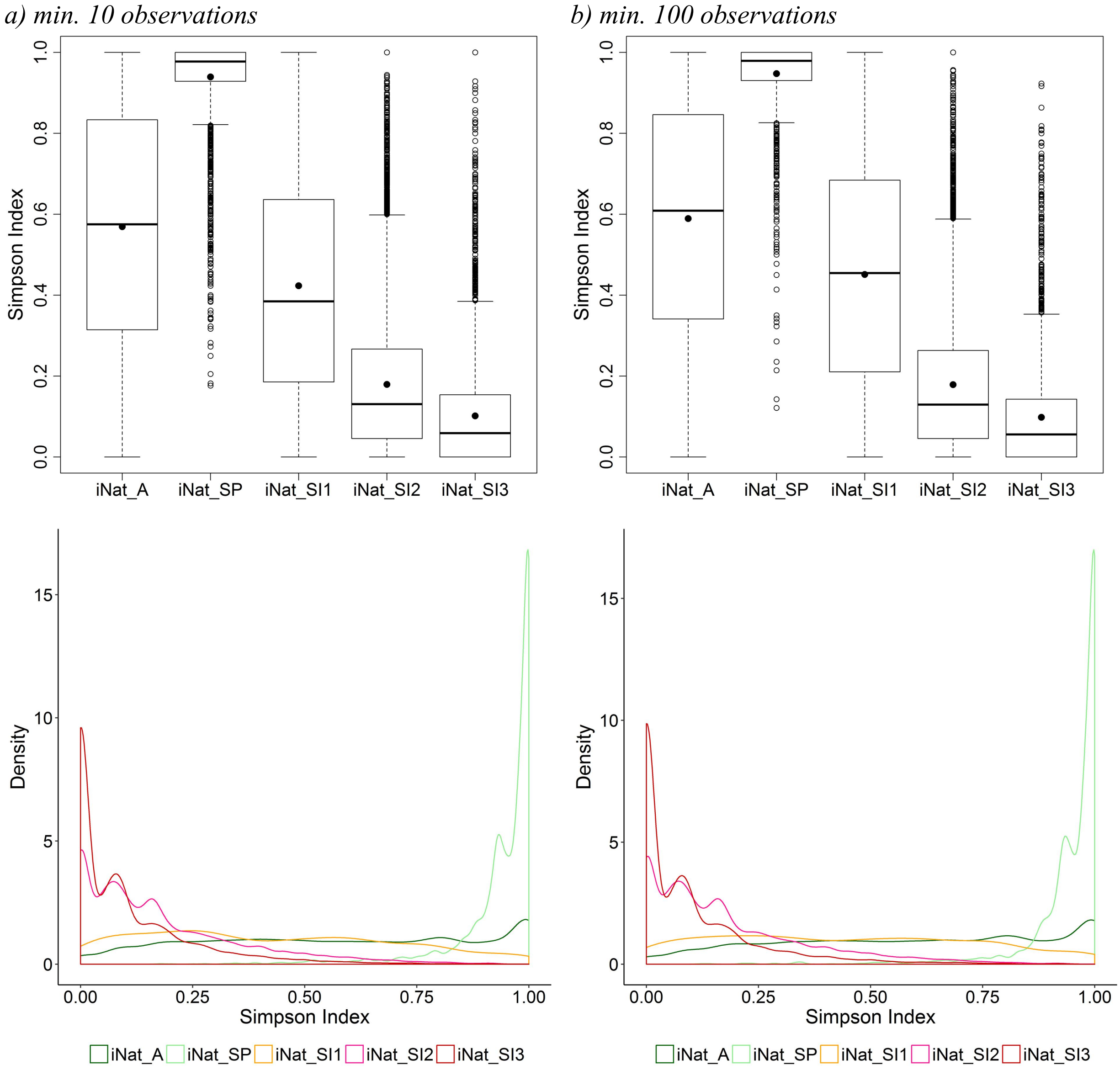
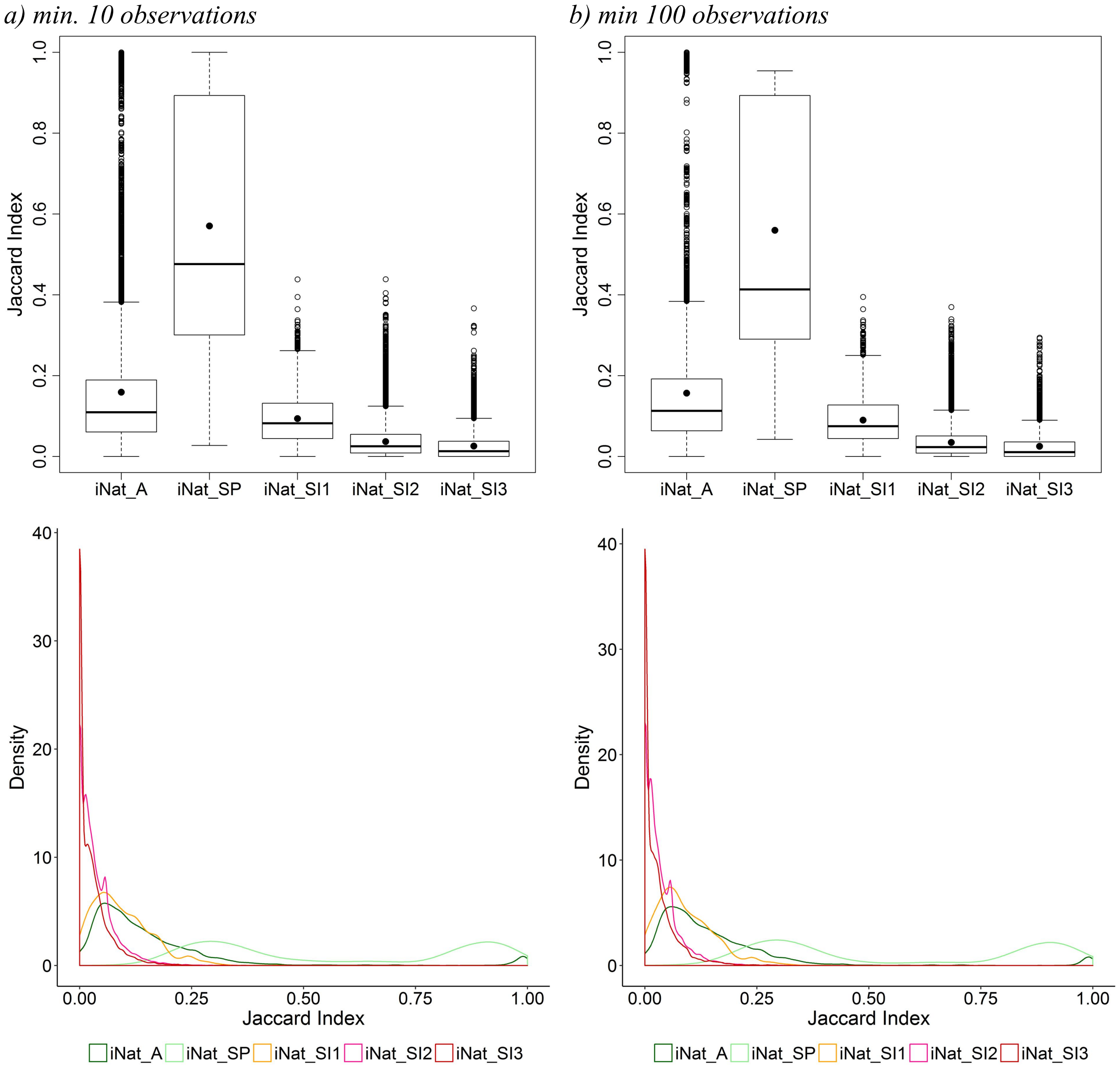


Figure 7.3.3: iNaturalist, distributions of Simpson similarity index values, analysis with smaller minimum number of 10 target species observations in observed community extraction (a), and more conservative minimum number of 100 observations (b).



a) min. 10 observations

b) min 100 observations

Figure 7.3.4: iNaturalist, distributions of Jaccard similarity index values, analysis with smaller minimum number of 10 target species observations in observed community extraction (a,) and more conservative minimum number of 100 observations (b).

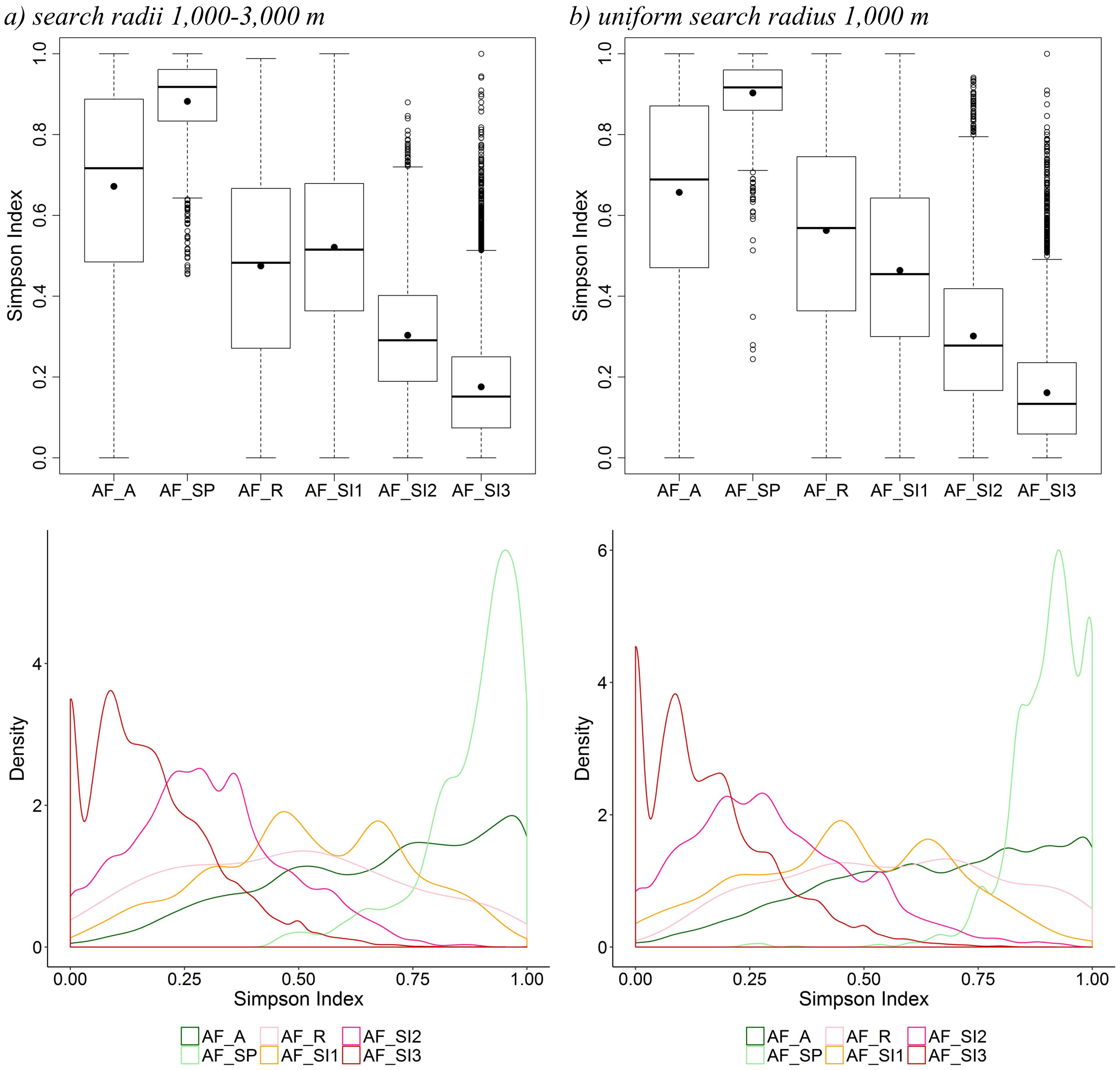
Table 7.3.1: Analysis with smaller minimum number of target species observations in observed community extraction (10), key numbers describing valid observed communities.

Data use case	No. of valid observed communities		Mean no. of species in observed communities		No. of nonspecific species	
	with min. no. of target observations:					
	100	10	100	10	100	10
ArtenFinder	183	792	31.7	43.6	46	39
iNaturalist	234	1258	86.9	85.9	0	0

Table 7.3.2: Analysis with smaller minimum number of target species observations in observed community extraction (10), key numbers describing sets of valid candidate observations.

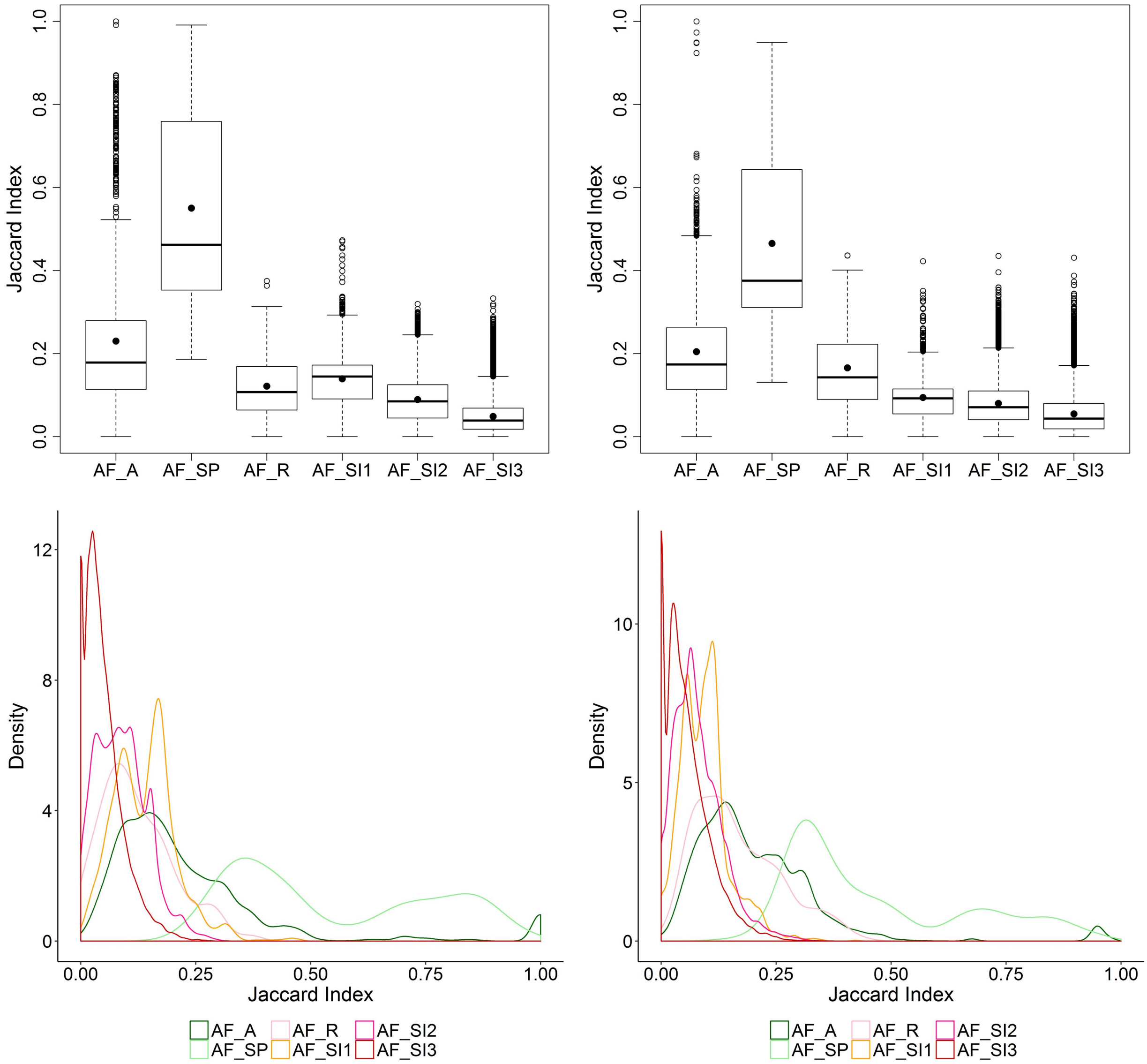
Set of candidates	No. of valid candidate cases		Mean no. of species in candidate contexts		Set of candidates	No. of valid candidate cases		Mean no. of species in candidate contexts	
	with min. no. of target obs.:					with min. no. of target obs.:			
	100	10	100	10		100	10	100	10
AF_A	22,426	34,719	108.2	111.9	**iNat_A**	34,821	50,147	132.3	132.9
AF_SP	1,486	2,357	96.5	108.9	**iNat_SP**	2,415	3,868	316.9	309.2
AF_R	362	608	115.1	125.8					
AF_SI1	1,718	3,486	99.5	105.5	**iNat_SI1**	2,216	3,116	106.4	106.7
AF_SI2	22,179	34,303	102.5	104.7	**iNat_SI2**	34,485	49,662	127.6	125.2
AF_SI3	22,896	36,720	31.2	32.6	**iNat_SI3**	4,768	7,374	34.4	35.5

7.3.2 Using Variable Search Radii


a) search radii 1,000-3,000 m

b) uniform search radius 1,000 m

Figure 7.3.5: ArtenFinder, distributions of Simpson similarity index values, analysis with search radius of 1,000-3,000 m for different species groups (a), and uniform search radius of 1,000 m (b).



a) search radii 1,000-3,000 m

b) uniform search radius 1,000 m

Figure 7.3.6: ArtenFinder, distributions of Jaccard similarity index values, analysis with search radius of 1,000-3,000 m for different species groups (a), and uniform search radius of 1,000 m (b).

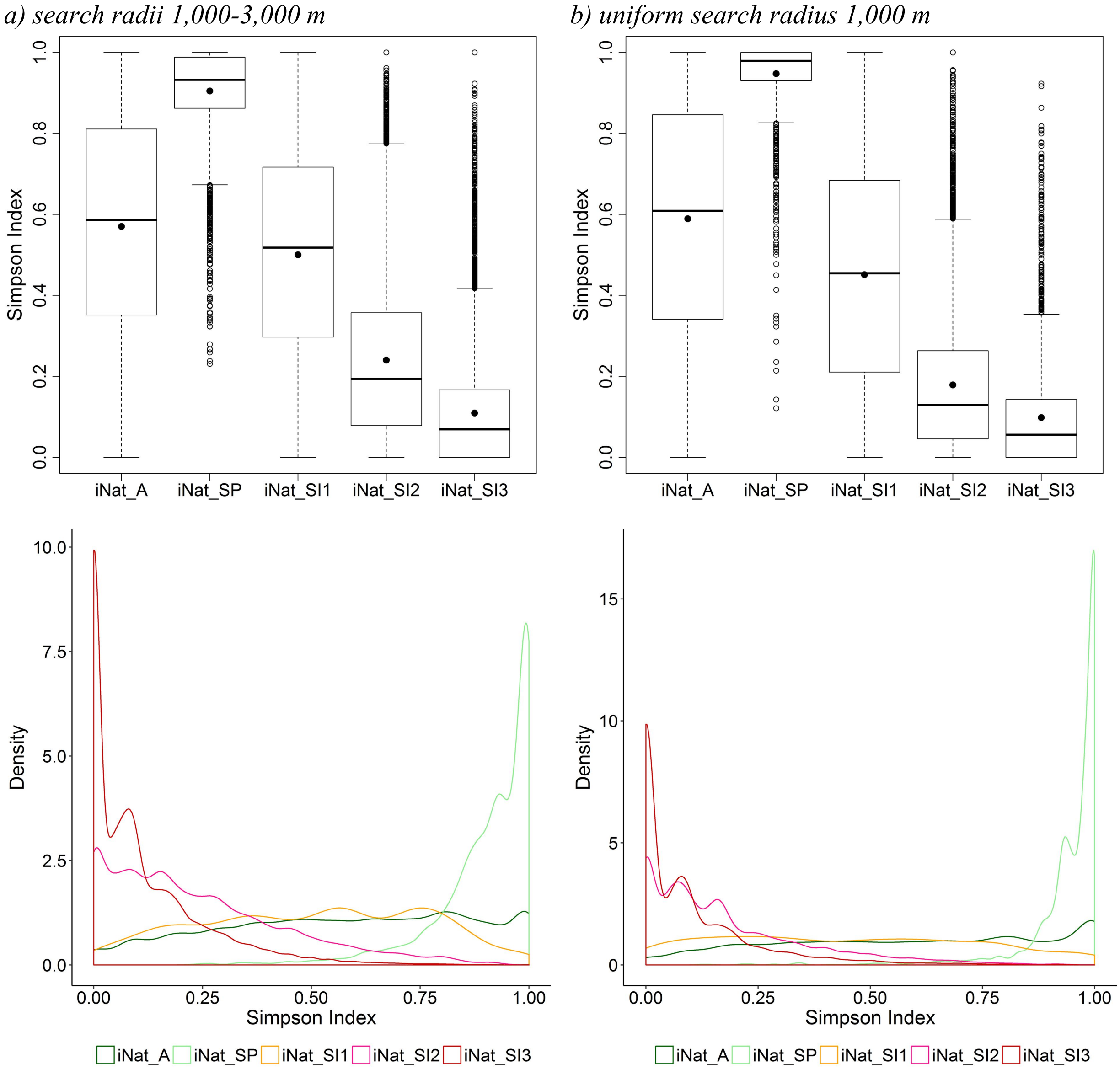
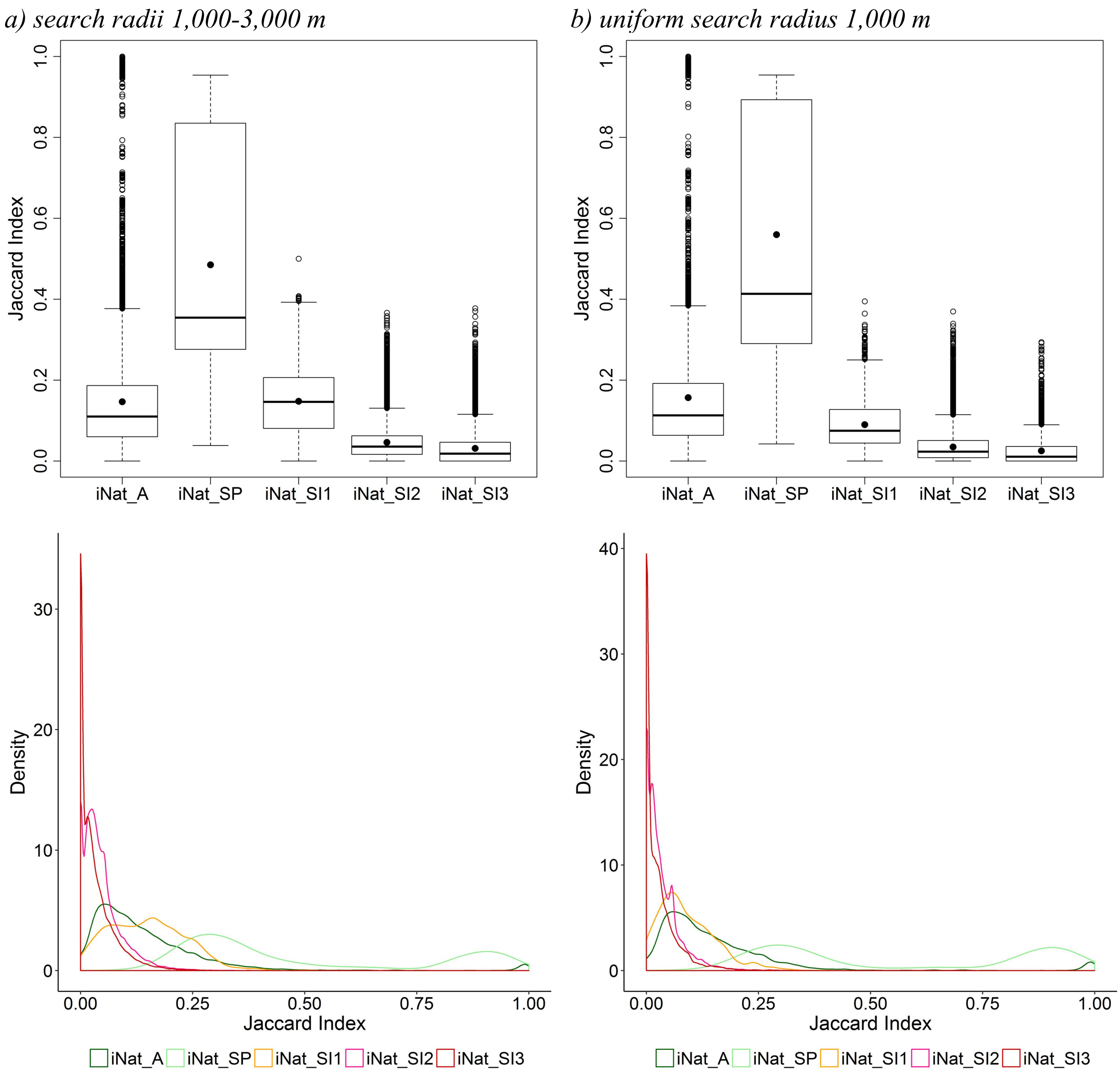


Figure 7.3.7: iNaturalist, distributions of Simpson similarity index values, analysis with search radius of 1,000-3,000 m for different species groups (a), and uniform search radius of 1,000 m (b).



a) *search radii 1,000-3,000 m*

b) *uniform search radius 1,000 m*

Figure 7.3.8: iNaturalist, distributions of Jaccard similarity index values, analysis with search radius of 1,000-3,000 m for different species groups (a), and uniform search radius of 1,000 m (b).

Table 7.3.3: Analysis with search radius of 1,000-3,000 m for different species groups, key numbers describing valid observed communities.

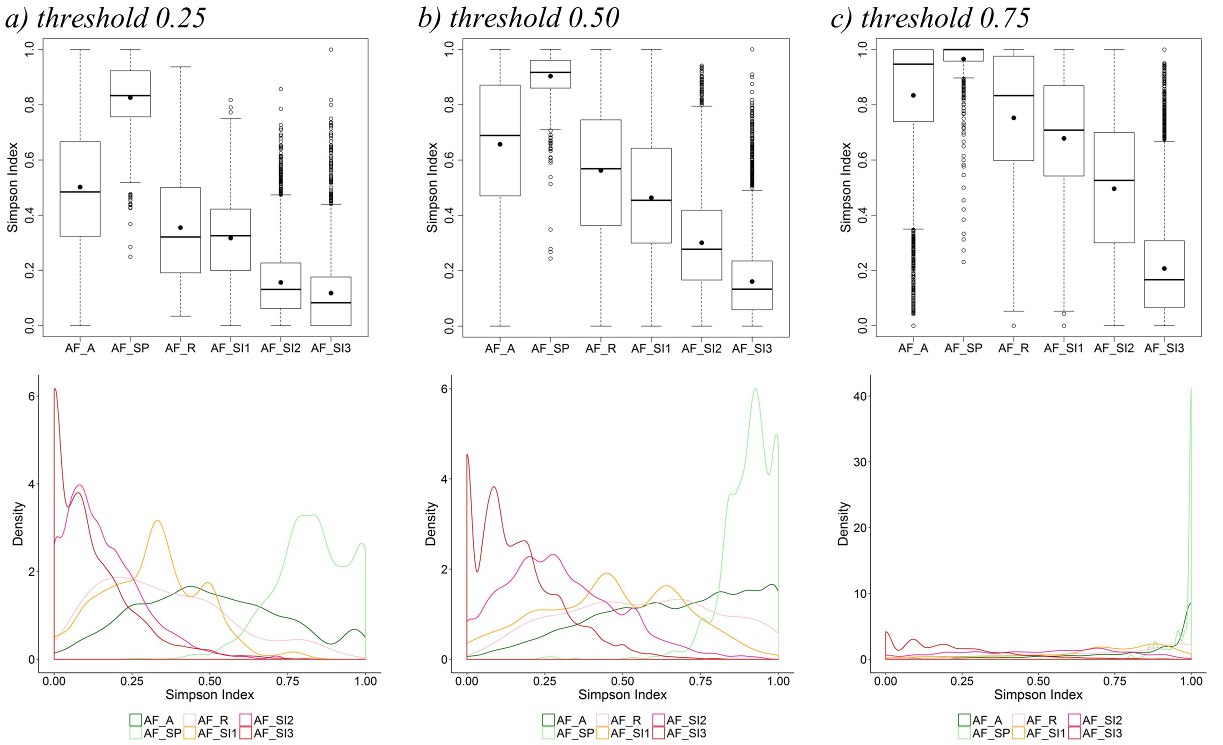
Data use case	No. of valid observed communities		Mean no. of species in observed communities		No. of nonspecific species	
	with radius configuration:					
	1,000	1,000-3,000	1,000	1,000-3,000	1,000	1,000-3,000
ArtenFinder	183	128	31.7	52.2	46	118
iNaturalist	234	322	86.9	85.0	0	5

Table 7.3.4: Analysis with search radius of 1,000-3,000 m for different species groups, key numbers describing sets of valid candidate observations.

Set of candidates	No. of valid candidate cases		Mean no. of species in candidate contexts		Set of candidates	No. of valid candidate cases		Mean no. of species in candidate contexts	
	with radius configuration:					with radius configuration:			
	1,000	1,000-3,000	1,000	1,000-3,000		1,000	1,000-3,000	1,000	1,000-3,000
AF_A	22,426	12,071	108.2	181.8	**iNat_A**	34,821	58,131	132.3	180.9
AF_SP	1,486	1,062	96.5	207.6	**iNat_SP**	2,415	3,560	316.9	300.2
AF_R	362	188	115.1	155.3					
AF_SI1	1,718	2,074	99.5	195.8	**iNat_SI1**	2,216	2,447	106.4	201.4
AF_SI2	22,179	11,935	102.5	166.0	**iNat_SI2**	34,485	57,664	127.6	175.6
AF_SI3	22,896	32,791	31.2	42.7	**iNat_SI3**	4,768	29,857	34.4	42.1

7.3.3 Modifying Thresholds for Frequently Associated Specie sand Nonspecific species

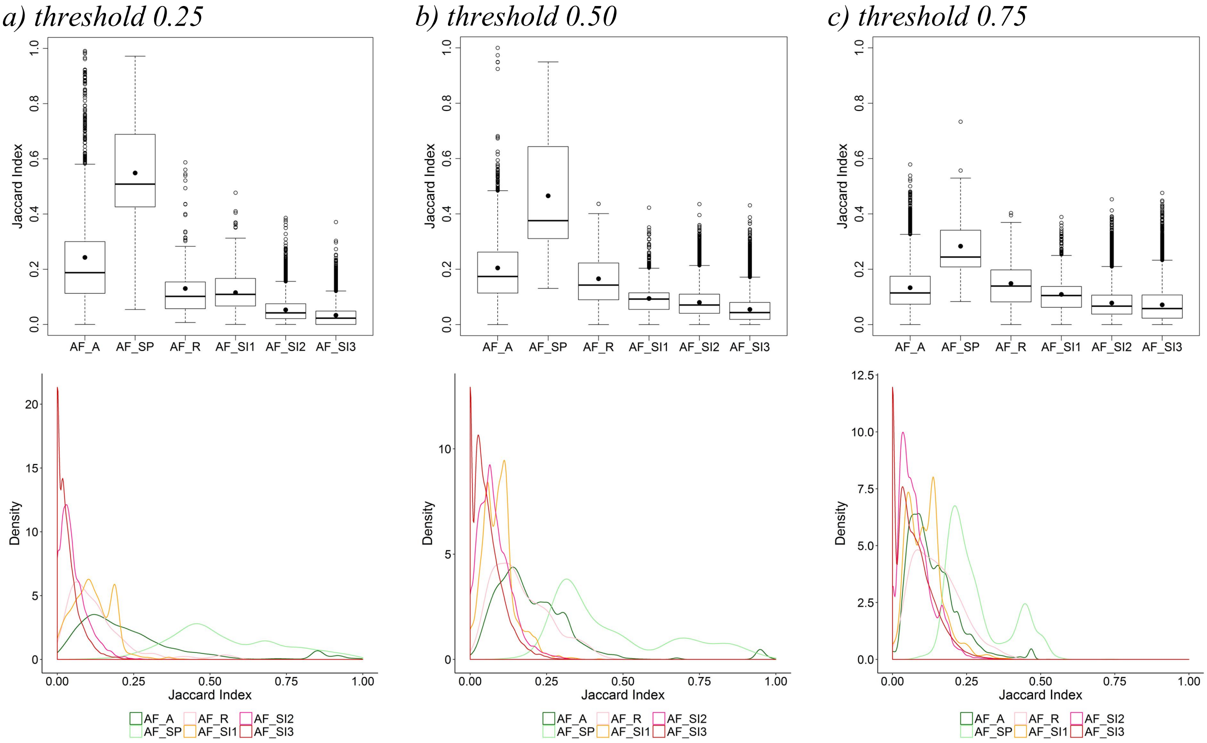
7.3.3.1 Modified Threshold for Identifying Frequent Associations Between a Target Species and its Context Species


a) threshold 0.25

b) threshold 0.50

c) threshold 0.75

Figure 7.3.9: ArtenFinder, distributions of Simpson similarity index values. a) Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75.

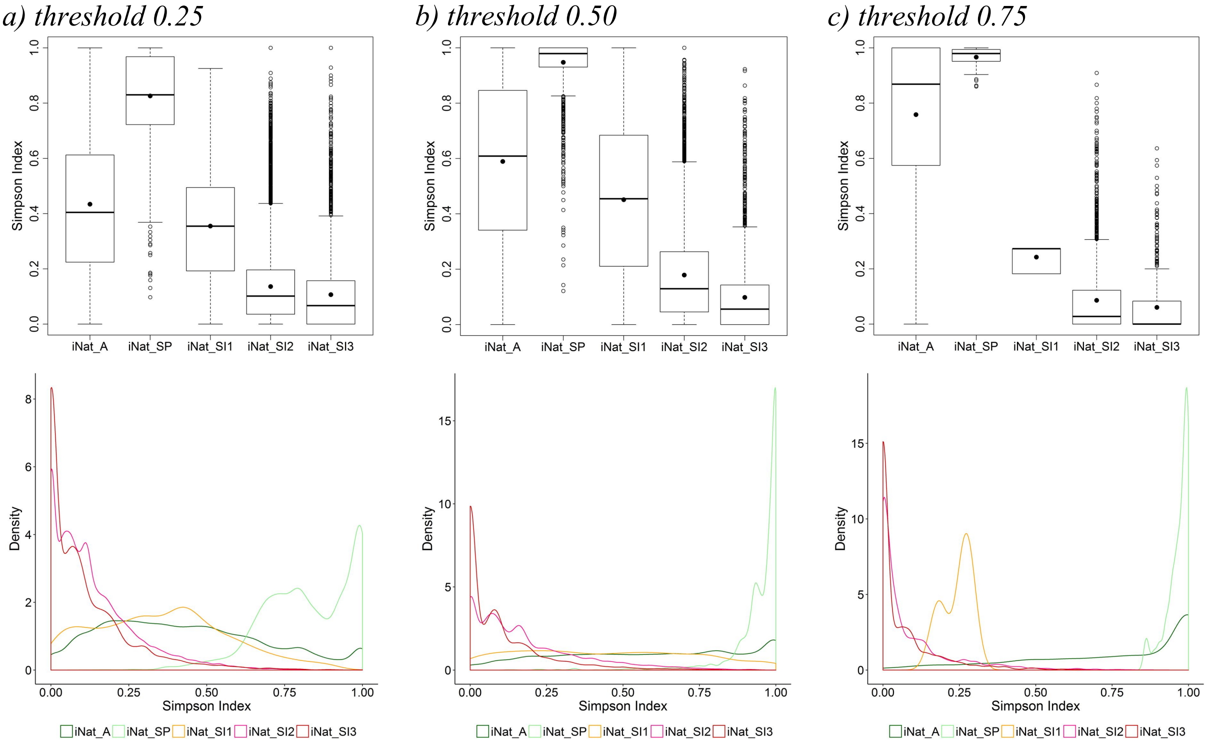


a) threshold 0.25

b) threshold 0.50

c) threshold 0.75

Figure 7.3.10: ArtenFinder, distributions of Jaccard similarity index values. a) Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75.

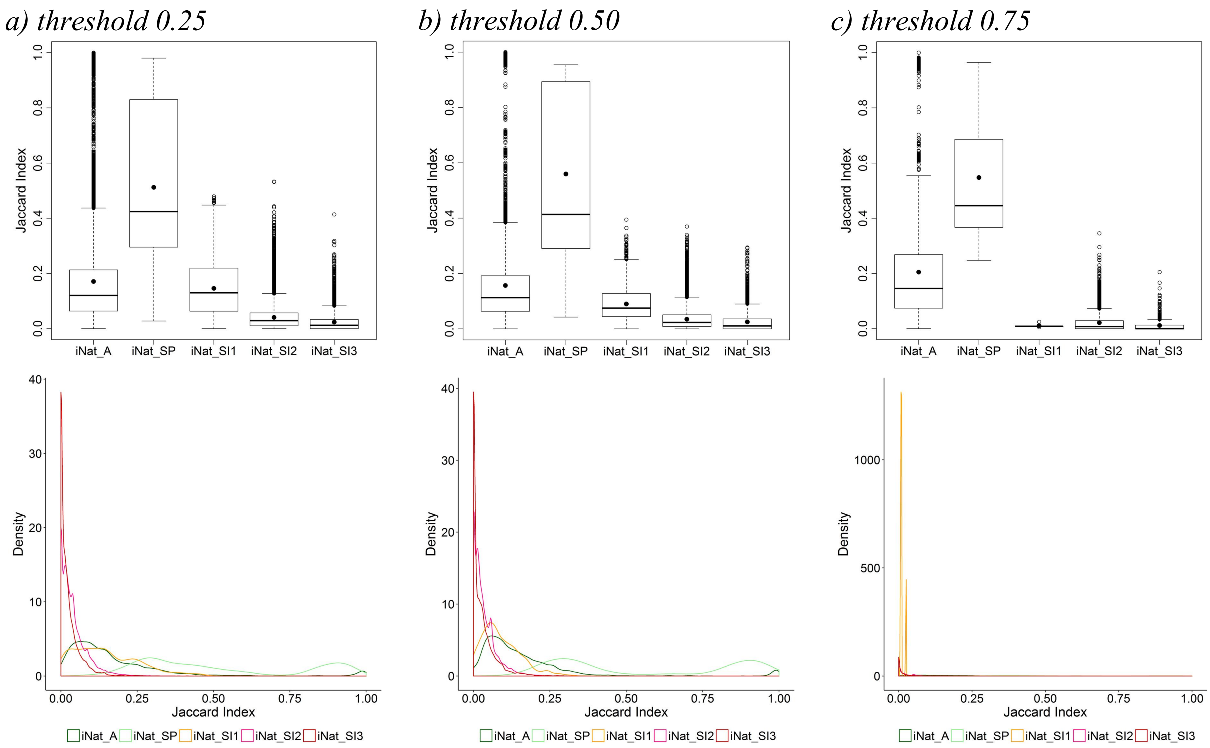


a) threshold 0.25

b) threshold 0.50

c) threshold 0.75

Figure 7.3.11: iNaturalist, distributions of Simpson similarity index values. a) Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75 (iNat_SI1: n = 6).



a) threshold 0.25

b) threshold 0.50

c) threshold 0.75

Figure 7.3.12: iNaturalist, distributions of Jaccard similarity index values. a) Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75 (iNat_SI1: n = 6).

Table 7.3.5: Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25 or raised to 0.75, key numbers describing valid observed communities.

	No. of valid observed communities			**Mean no. of species in observed communities with threshold:**			**No. of nonspecific species**		
Data use case	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
ArtenFinder	143	183	159	52.8	31.7	20.9	131	46	5
iNaturalist	474	234	47	105.3	86.9	100.0	31	0	0

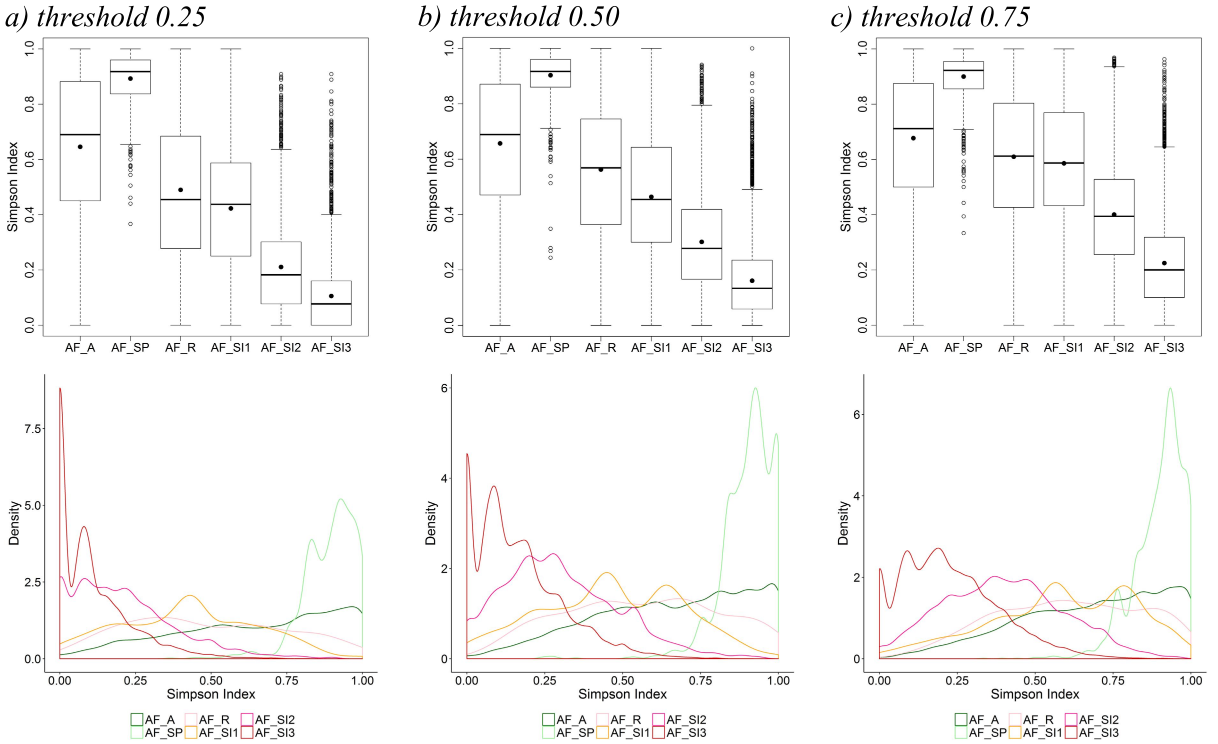
Table 7.3.6: Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25 or raised to 0.75, key numbers describing sets of valid candidate observations, ArtenFinder data.

Set of candidates	No. of valid candidate cases			Mean no. of species in candidate contexts		
	with threshold:					
	0.25	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
AF_A	9,967	22,426	35,024	67.9	108.2	136.5
AF_SP	1,303	1,486	2,000	71.3	96.5	105.6
AF_R	221	362	331	68.8	115.1	142.2
AF_SI1	1,891	1,718	1,740	66.6	99.5	125.4
AF_SI2	9,640	22,179	34,757	74.8	102.5	121.4
AF_SI3	7,651	22,896	38,878	28.7	31.2	32.3

Table 7.3.7: Analysis with threshold for identifying frequent associations between a target species and its context species reduced to 0.25 or raised to 0.75, key numbers describing sets of valid candidate observations, iNaturalist data.

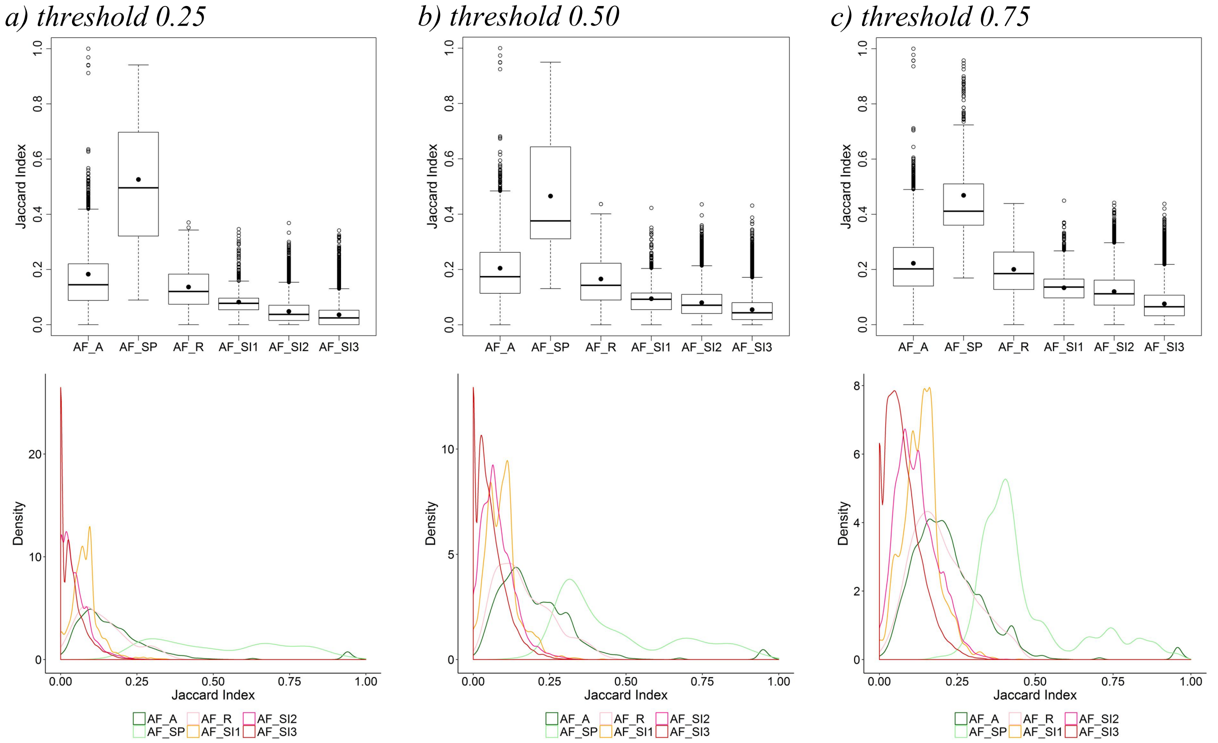
Set of candidates	No. of valid candidate cases			Mean no. of species in candidate contexts		
	with threshold:					
	0.25	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
iNat_A	59,207	34,821	4,575	108.2	132.3	243.2
iNat_SP	6,435	2,415	465	226.3	316.9	443.7
iNat_SI1	2,040	2,216	6	97.2	106.4	295.0
iNat_SI2	58,297	34,485	4,554	103.4	127.6	170.4
iNat_SI3	7,621	4,768	684	33.2	34.4	34.7

7.3.3.2 Modified Threshold for Identifying Nonspecific Species



a) threshold 0.25 *b) threshold 0.50* *c) threshold 0.75*

Figure 7.3.13: ArtenFinder, distributions of Simpson similarity index values. a) Analysis with threshold for identifying nonspecific species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75.



a) threshold 0.25 *b) threshold 0.50* *c) threshold 0.75*

Figure 7.3.14: ArtenFinder, distributions of Jaccard similarity index values. a) Analysis with threshold for identifying nonspecific species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75.

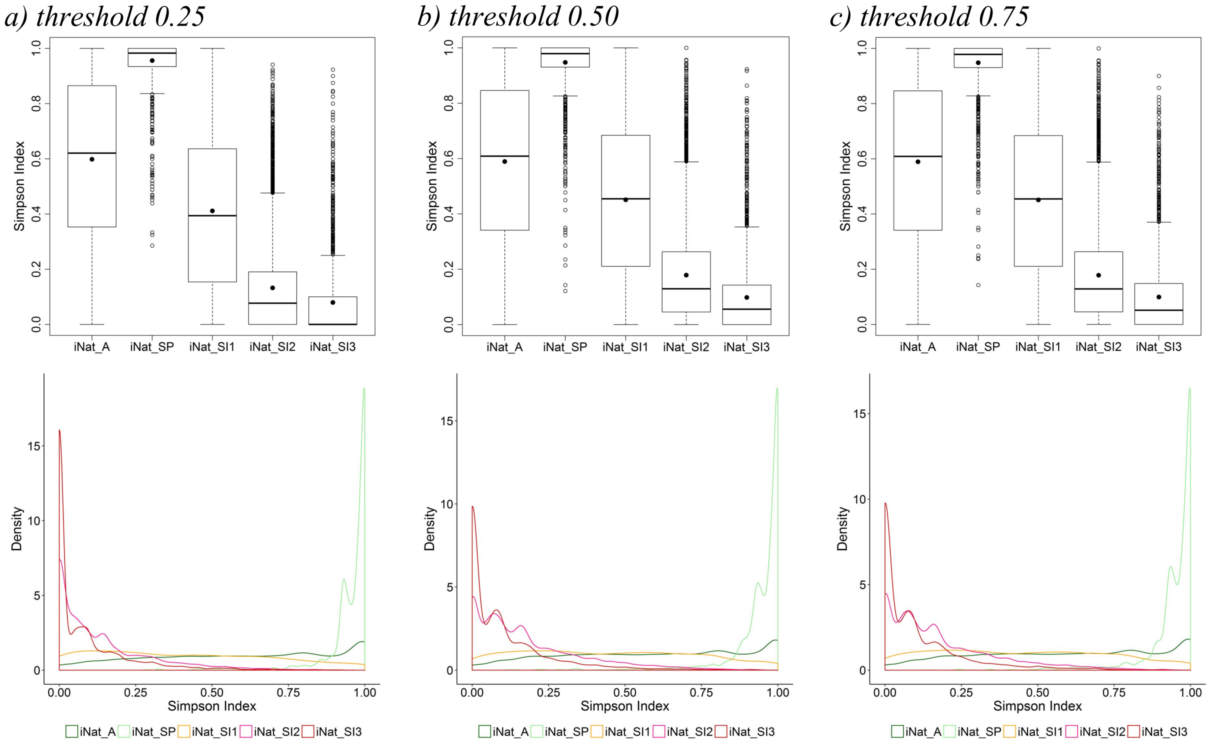

a) threshold 0.25
b) threshold 0.50
c) threshold 0.75

Figure 7.3.15: iNaturalist, distributions of Simpson similarity index values. a) Analysis with threshold for identifying nonspecific species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75.

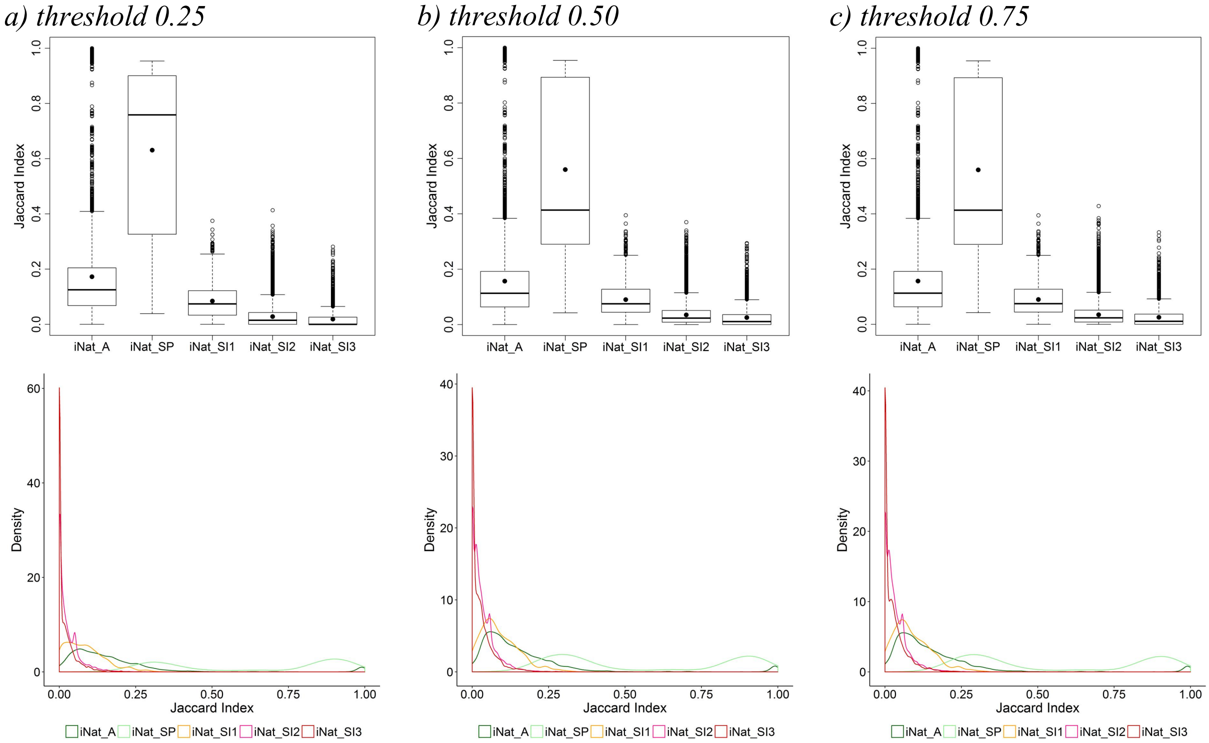

a) threshold 0.25
b) threshold 0.50
c) threshold 0.75

Figure 7.3.16: iNaturalist, distributions of Jaccard similarity index values. a) Analysis with threshold for identifying nonspecific species reduced to 0.25. b) Threshold 0.50. c) Threshold raised to 0.75

Table 7.3.8: Analysis with threshold for identifying nonspecific species reduced to 0.25 or raised to 0.75, key numbers describing valid observed communities.

Data use case	**No. of valid observed communities**			**Mean no. of species in observed communities**			**No. of nonspecific species**		
	with threshold:								
	0.25	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
ArtenFinder	122	183	192	24.9	31.7	40.3	74	46	28
iNaturalist	186	234	234	101.8	86.9	86.9	8	0	0

Table 7.3.9: Analysis with threshold for identifying nonspecific species reduced to 0.25 or raised to 0.75, key numbers describing sets of valid candidate observations, ArtenFinder data.

Set of candidates	**No. of valid candidate cases**			**Mean no. of species in candidate contexts**		
	with threshold:					
	0.25	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
AF_A	14,009	22,426	34,552	93.3	108.2	117.1
AF_SP	879	1,486	2,102	81.1	96.5	115.1
AF_R	217	362	422	99.7	115.1	126.1
AF_SI1	1,192	1,718	2,813	84.8	99.5	108.1
AF_SI2	13,822	22,179	34,181	93.0	102.5	108.9
AF_SI3	13,127	22,896	35,885	30.2	31.2	31.7

Table 7.3.10: Analysis with threshold for identifying nonspecific species reduced to 0.25 or raised to 0.75, key numbers describing sets of valid candidate observations, iNaturalist data.

Set of candidates	**No. of valid candidate cases**			**Mean no. of species in candidate contexts**		
	with threshold:					
	0.25	**0.50**	**0.75**	**0.25**	**0.50**	**0.75**
iNat_A	26,025	34,821	34,821	137.8	132.3	132.3
iNat_SP	1,923	2,415	2,412	356.5	316.9	319.3
iNat_SI1	1,556	2,216	2,216	111.3	106.4	106.4
iNat_SI2	25,811	34,485	34,524	132.2	127.6	127.4
iNat_SI3	3,640	4,768	4,841	35.4	34.4	33.9

7.3.4 Using Auxiliary Land Cover Information

Table 7.3.11: CORINE land cover classes occurring in Rheinland-Pfalz.

Code	**Land cover**
1	*Artificial surfaces*
1.1.	Urban fabric
1.1.1.	Continuous urban fabric
1.1.2.	Discontinuous urban fabric
1.2.	Industrial, commercial and transport units
1.2.1.	Industrial or commercial units
1.2.2.	Road and rail networks and associated land
1.2.3.	Port areas
1.2.4.	Airports
1.3.	Mine, dump and construction sites
1.3.1.	Mineral extraction sites
1.3.2.	Dump sites
1.3.3.	Construction sites
1.4.	Artificial non-agricultural vegetated areas
1.4.1.	Green urban areas
1.4.2.	Sport and leisure facilities
2	*Agricultural areas*
2.1.	Arable land
2.1.1.	Non-irrigated arable land
2.2.	Permanent crops
2.2.1.	Vineyards
2.2.2.	Fruit trees and berry plantations
2.3.	Pastures
2.3.1.	Pastures
2.4.	Heterogeneous agricultural areas
2.4.2.	Complex cultivation
2.4.3.	Land principally occupied by agriculture, with significant areas of natural vegetation
3	*Forests and semi-natural areas*
3.1.	Forests
3.1.1.	Broad-leaved forest
3.1.2.	Coniferous forest
3.1.3.	Mixed forest
3.2.	Shrub and/or herbaceous vegetation association
3.2.1.	Natural grassland
3.2.2.	Moors and heathland
3.2.4.	Transitional woodland shrub
3.3.	Open spaces with little or no vegetation
3.3.2.	Bare rock
4	*Wetlands*
4.1.	Inland wetlands
4.1.1.	Inland marshes
5	*Water bodies*
5.1.	Inland waters
5.1.1.	Water courses
5.1.2.	Water bodies

Table 7.3.12: NLCD land cover classes occurring in California.

Code	**Land cover**
Water	
11	Open Water
12	Perennial Ice/Snow
Developed	
21	Developed, Open Space
22	Developed, Low Intensity
23	Developed, Medium Intensity
24	Developed High Intensity
Barren	
31	Barren Land (Rock/Sand/Clay)
Forest	
41	Deciduous Forest
42	Evergreen Forest
43	Mixed Forest
Shrubland	
52	Shrub/Scrub
Herbaceous	
71	Grassland/Herbaceous
Planted/Cultivated	
81	Pasture/Hay
82	Cultivated Crops.
Wetlands	
90	Woody Wetlands
95	Emergent Herbaceous Wetlands

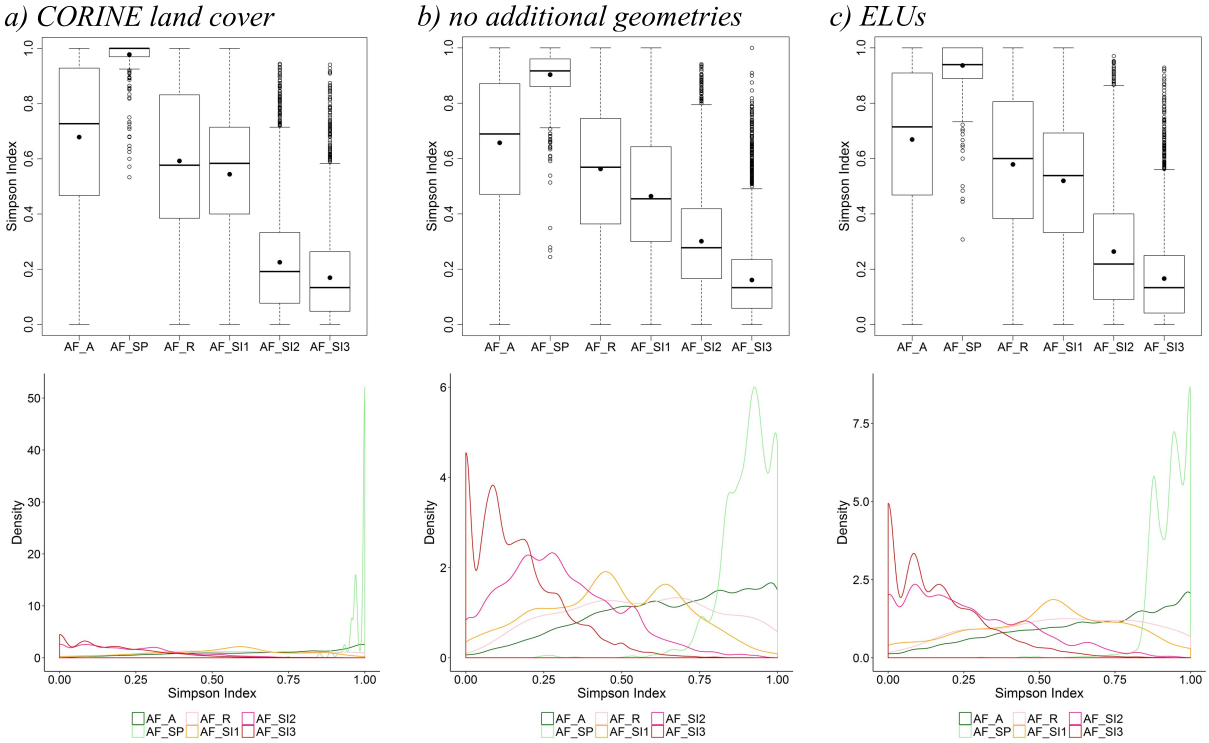


Figure 7.3.17: ArtenFinder, distributions of Simpson similarity index values. a) Analysis using CORINE land cover geometries. b) No additional geometries used. c) ELU geometries used.

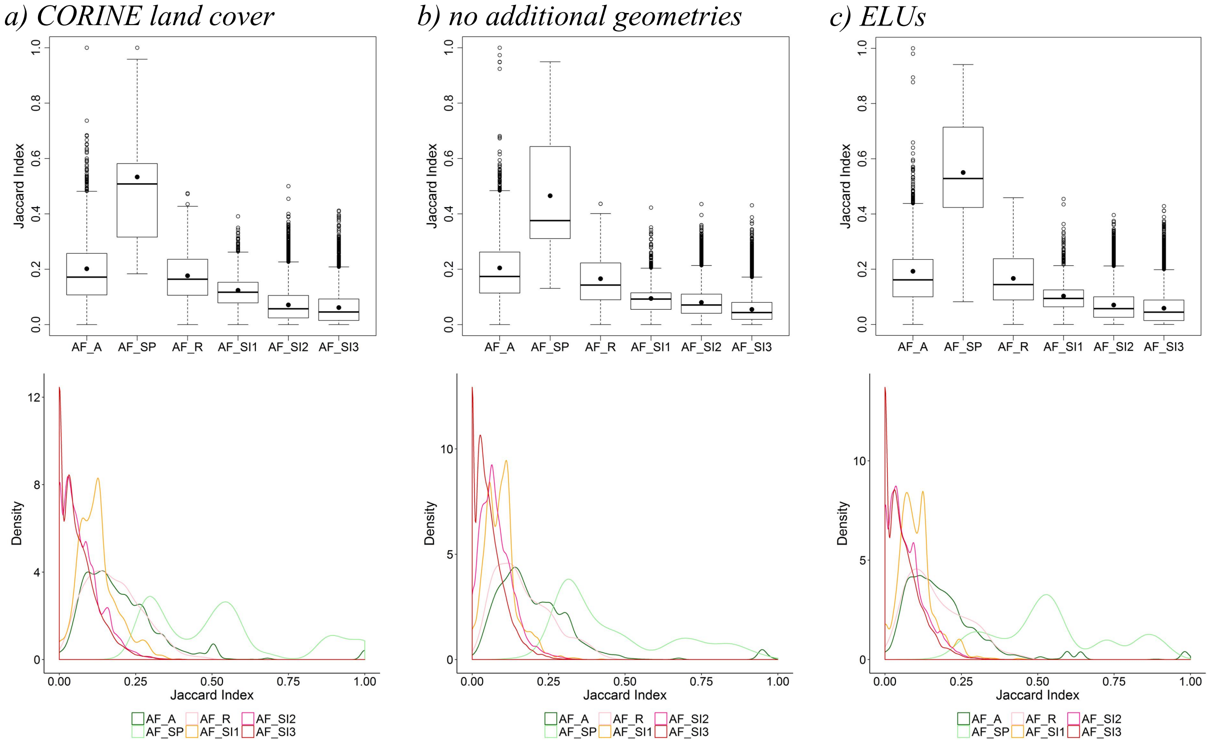


Figure 7.3.18: ArtenFinder, distributions of Jaccard similarity index values. a) Analysis using CORINE land cover geometries. b) No additional geometries used. c) ELU geometries used.

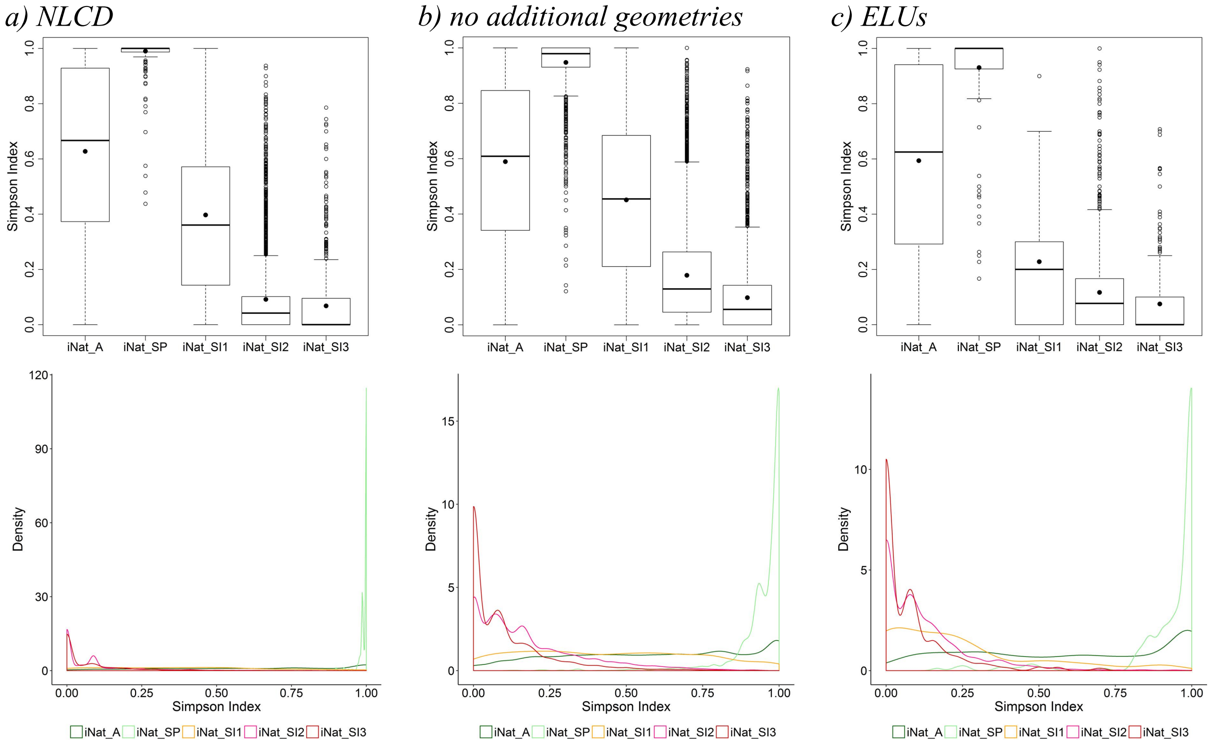


Figure 7.3.19: iNaturalist, distributions of Simpson similarity index values. a) Analysis using NLCD geometries. b) No additional geometries used. c) ELU geometries used. (iNat_SI1 ELU: three species in 136 valid candidate cases).

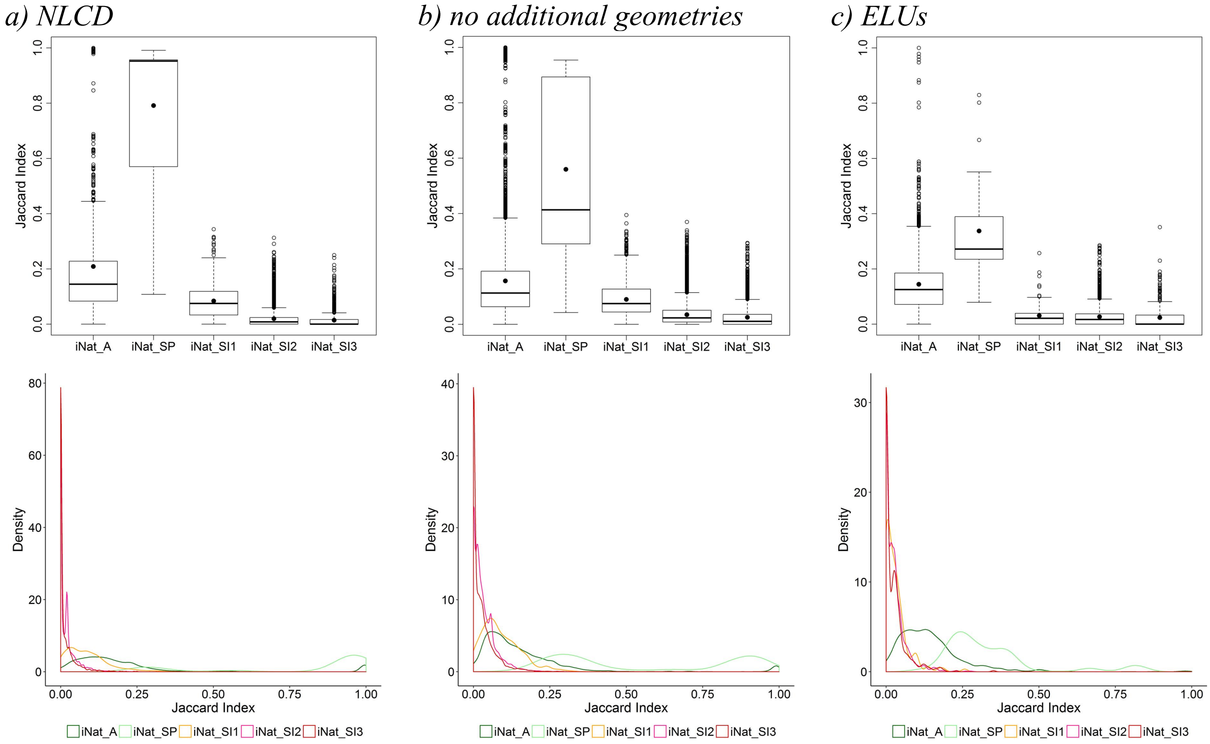


Figure 7.3.20: iNaturalist, distributions of Jaccard similarity index values. a) Analysis using NLCD geometries. b) No additional geometries used. c) ELU geometries used. (iNat_SI1 ELU: three species in 136 valid candidate cases).

Table 7.3.13: Analysis using CORINE land cover (CLC), NLCD or ELU data in defining relevant search areas, key numbers describing valid observed communities.

	No. of valid observed communities			Mean no. of species in observed communities			No. of nonspecific species		
				with geometries from:					
Data use case	**CLC/ NLCD**	**none**	**ELUs**	**CLC/ NLCD**	**none**	**ELUs**	**CLC/ NLCD**	**none**	**ELUs**
ArtenFinder	184	183	186	25.2	31.7	24.5	13	46	23
iNaturalist	120	234	50	118.5	86.9	28.8	0	0	0

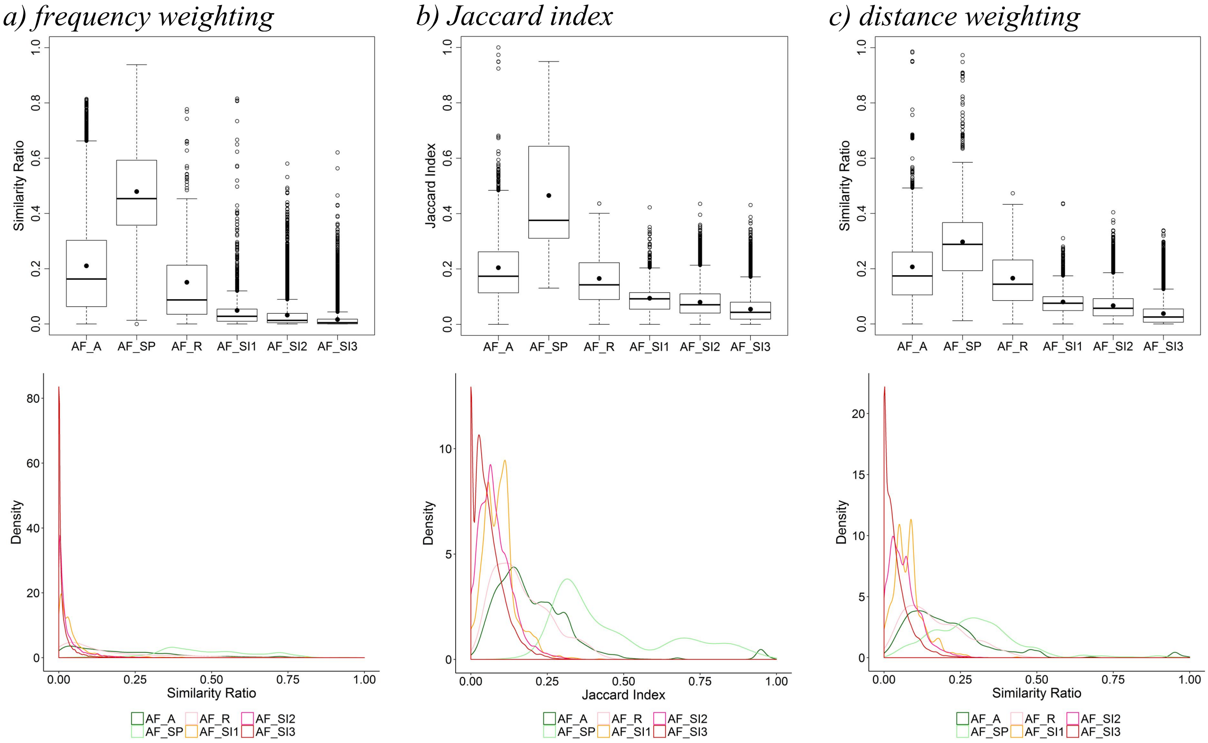
Table 7.3.14: Analysis using CORINE land cover (CLC) or ELU data in defining relevant search areas, key numbers describing sets of valid candidate observations, ArtenFinder data.

	No. of valid candidate cases			Mean no. of species in candidate contexts		
			with geometries from:			
Set of candidates	**CLC**	**none**	**ELUs**	**NLCD**	**none**	**ELUs**
AF_A	31,570	22,426	29,800	82.3	108.2	85.4
AF_SP	2,065	1,486	1,606	86.2	96.5	71.8
AF_R	379	362	369	87.8	115.1	88.6
AF_SI1	1,466	1,718	1,735	74.1	99.5	76.5
AF_SI2	21,041	22,179	23,447	52.2	102.5	67.4
AF_SI3	13,490	22,896	17,857	26.7	31.2	29.0

Table 7.3.15: Analysis using NLCD or ELU data in defining relevant search areas, key numbers describing sets of valid candidate observations, iNaturalist data.

	No. of valid candidate cases			Mean no. of species in candidate contexts		
			with geometries from:			
Set of candidates	**NLCD**	**none**	**ELUs**	**NLCD**	**none**	**ELUs**
iNat_A	13,681	34,821	5,649	122.6	132.3	83.1
iNat_SP	1,275	2,415	208	408.4	316.9	122.1
iNat_SI1	606	2,216	136	79.2	106.4	76.9
iNat_SI2	10,079	34,485	3,890	73.7	127.6	75.8
iNat_SI3	1,124	4,768	462	29.1	34.4	31.6

7.3.5 Using a Quantitative Similarity Index

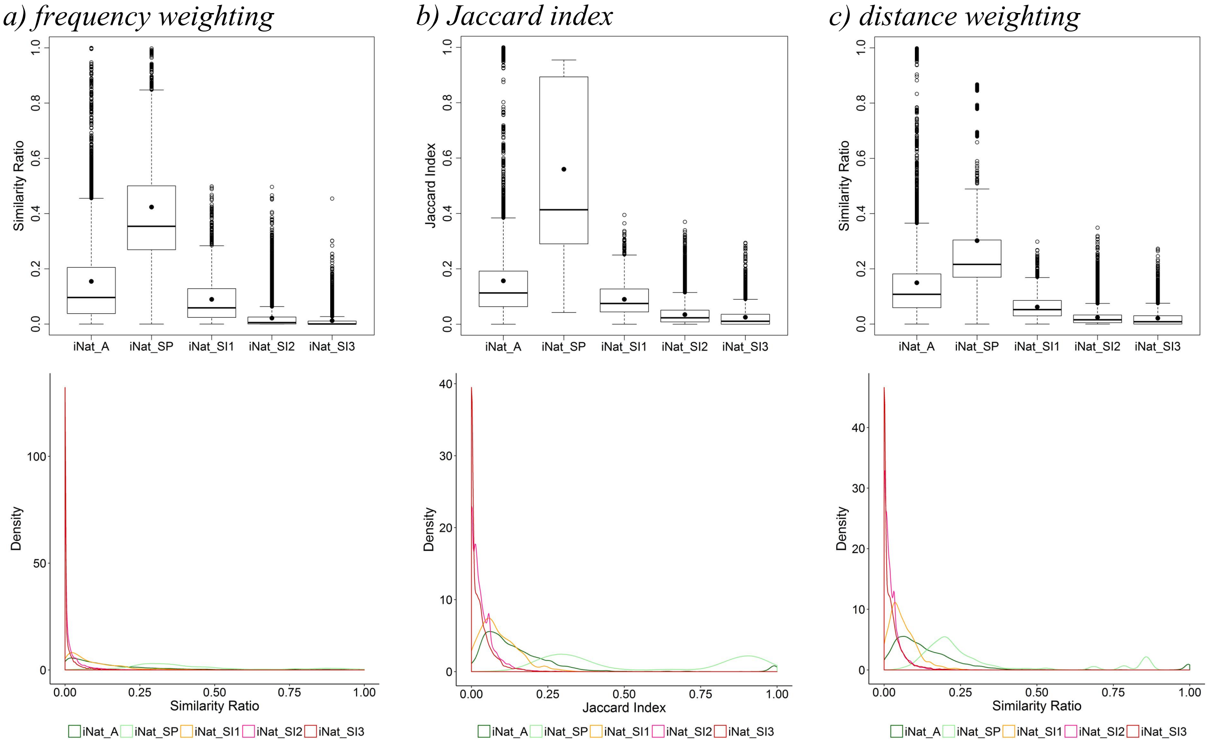


a) frequency weighting

b) Jaccard index

c) distance weighting

Figure 7.3.21: ArtenFinder, distributions of Similarity Ratio values. a) Analysis using observation frequency for weighting. b) Jaccard index. c) Distance weighting used.



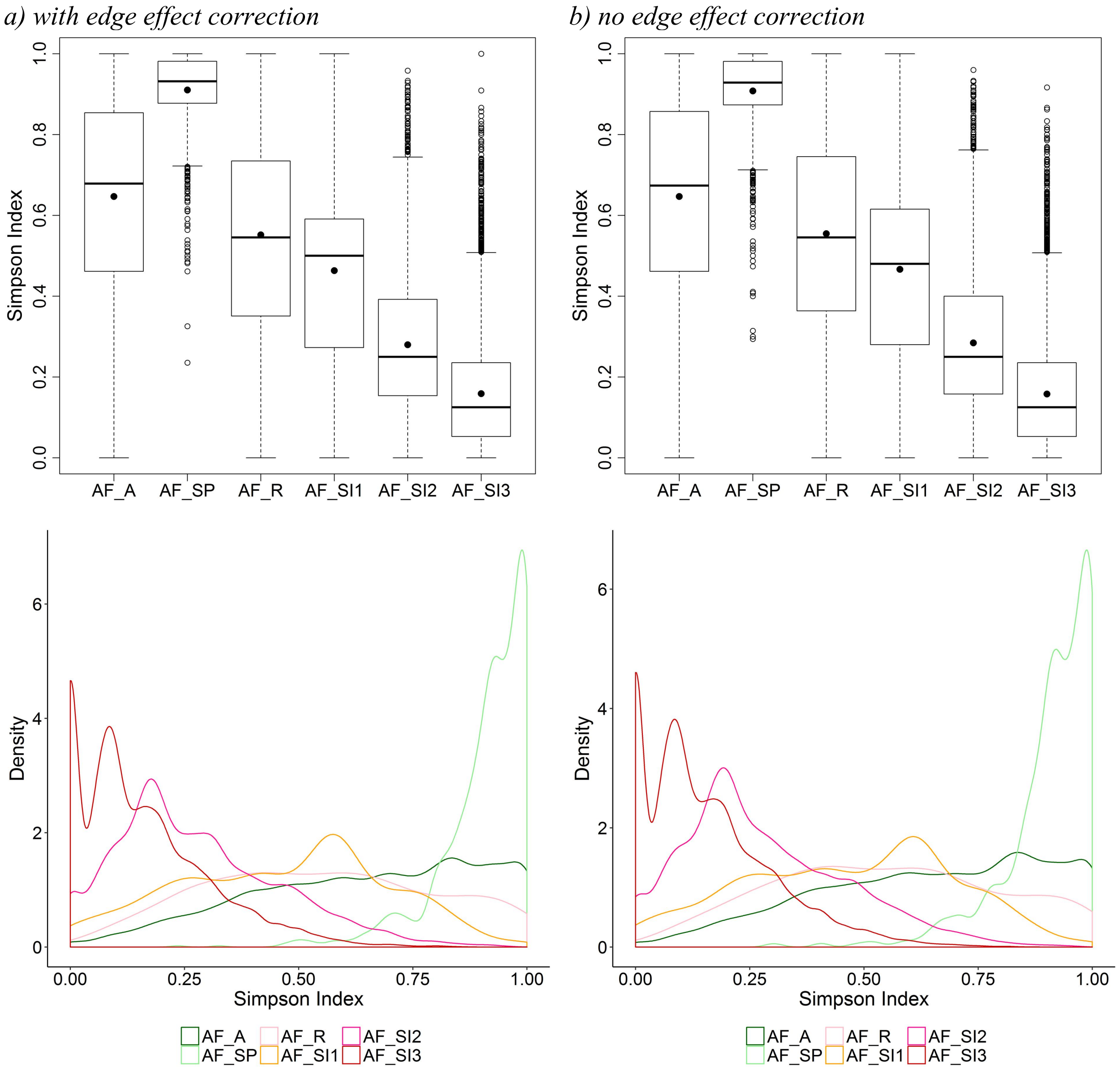
a) frequency weighting

b) Jaccard index

c) distance weighting

Figure 7.3.22: iNaturalist, distributions of Similarity Ratio values. a) Analysis using observation frequency for weighting. b) Jaccard index. c) Distance weighting used.

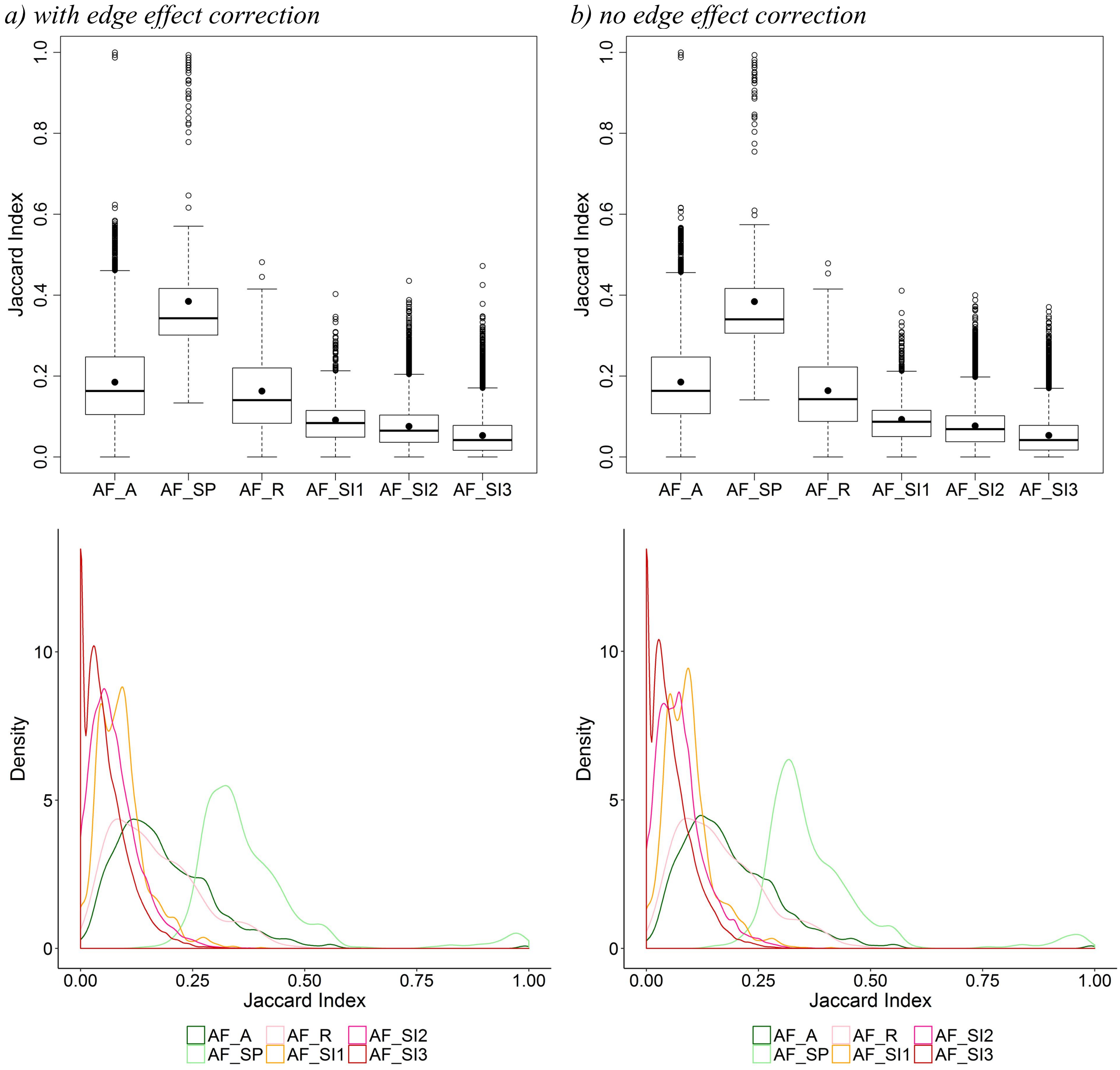
7.3.6 Examining Edge Effects



a) with edge effect correction

b) no edge effect correction

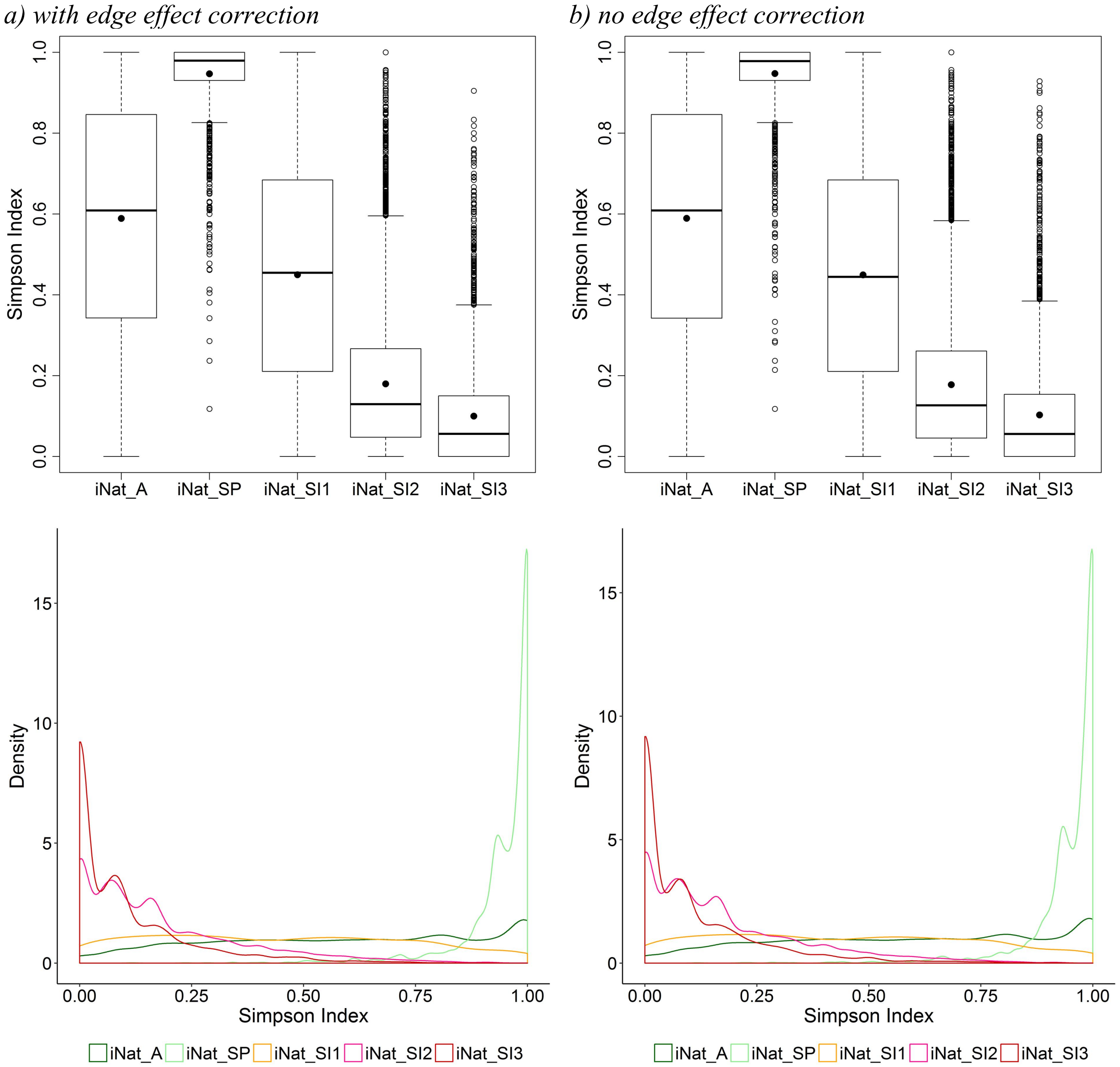
Figure 7.3.23: ArtenFinder, distributions of Simpson similarity index values, analysis of edge effects.



a) with edge effect correction

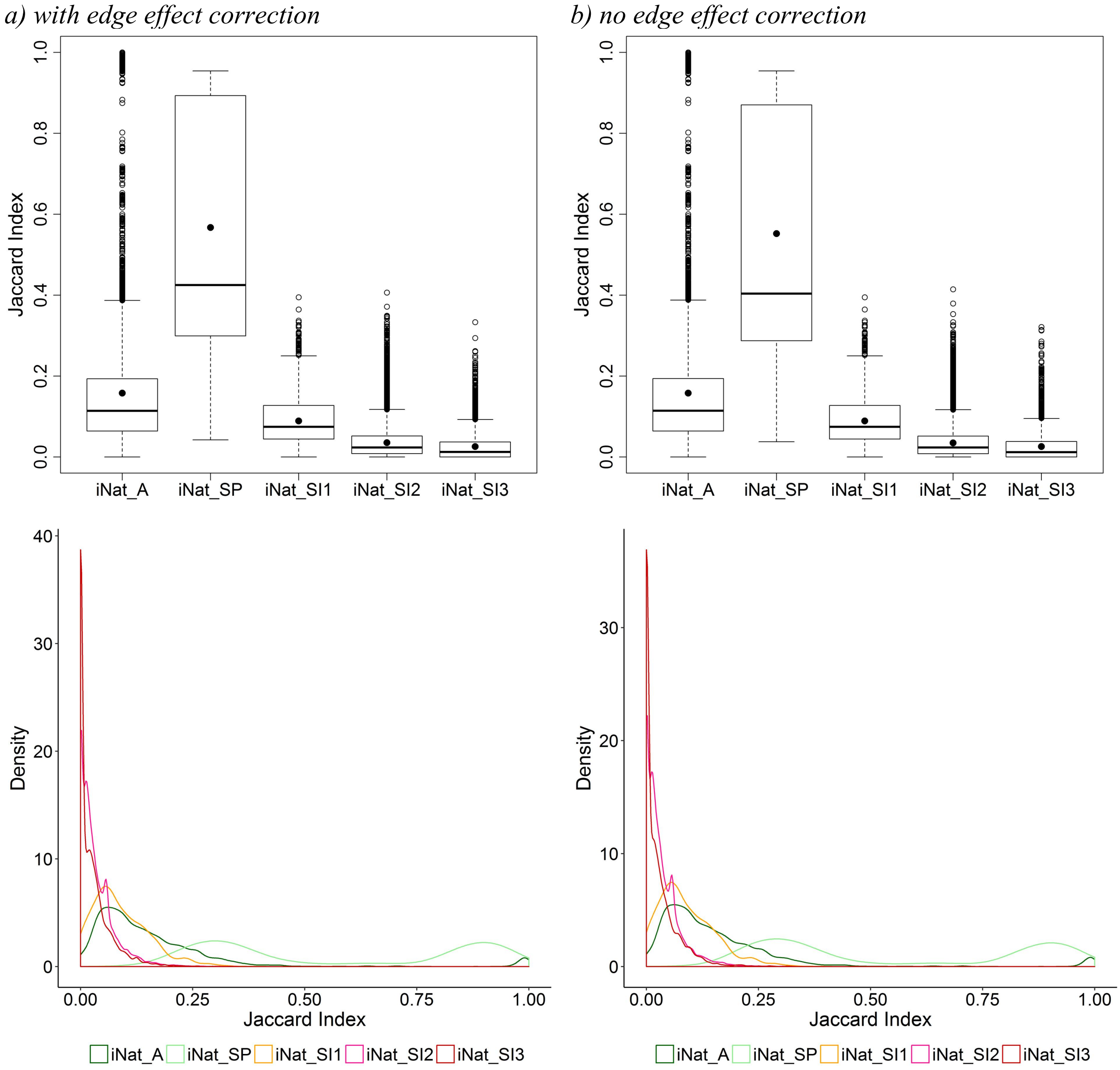
b) no edge effect correction

Figure 7.3.24: ArtenFinder, distributions of Jaccard similarity index values, analysis of edge effects.


a) with edge effect correction

b) no edge effect correction

Figure 7.3.25: iNaturalist, distributions of Simpson similarity index values, analysis of edge effects.



a) with edge effect correction

b) no edge effect correction

Figure 7.3.26: iNaturalist, distributions of Jaccard similarity index values, analysis of edge effects.

Table 7.3.16: Analysis of edge effects, key numbers describing valid observed communities.

Data use case	**No. of valid observed communities**		**Mean no. of species in observed communities**		**No. of nonspecific species**	
	with (right) or without (left) edge effect correction:					
	No corr.	**With corr.**	**No corr.**	**With corr.**	**No corr.**	**With corr.**
ArtenFinder	168	168	32.5	32.3	49	50
iNaturalist	232	233	87.7	87.4	0	0

Table 7.3.17: Analysis of edge effects, key numbers describing sets of valid candidate observations.

Set of candidates	**No. of valid candidate cases**		**Mean no. of species in candidate contexts**		**Set of candidates**	**No. of valid candidate cases**		**Mean no. of species in candidate contexts**	
	with (right) or without (left) edge effect correction:					**with (right) or without (left) edge effect correction:**			
	No corr.	**With corr.**	**No corr.**	**With corr.**		**No corr.**	**With corr.**	**No corr.**	**With corr.**
AF_A	18,374	18,475	112.2	112.1	**iNat_A**	34,491	34,589	132.7	132.6
AF_SP	1,063	1,041	126.3	127.3	**iNat_SP**	2,464	2,344	314.7	318.1
AF_R	314	314	115.5	115.5					
AF_SI1	1,391	1,391	102.0	101.6	**iNat_SI1**	2,214	2,214	106.4	106.4
AF_SI2	18,171	18,271	97.5	96.8	**iNat_SI2**	34,160	34,268	127.7	127.8
AF_SI3	18,035	18,108	31.2	31.3	**iNat_SI3**	4,802	4,836	34.8	34.4

7.3.7 Variant of Simpson Index

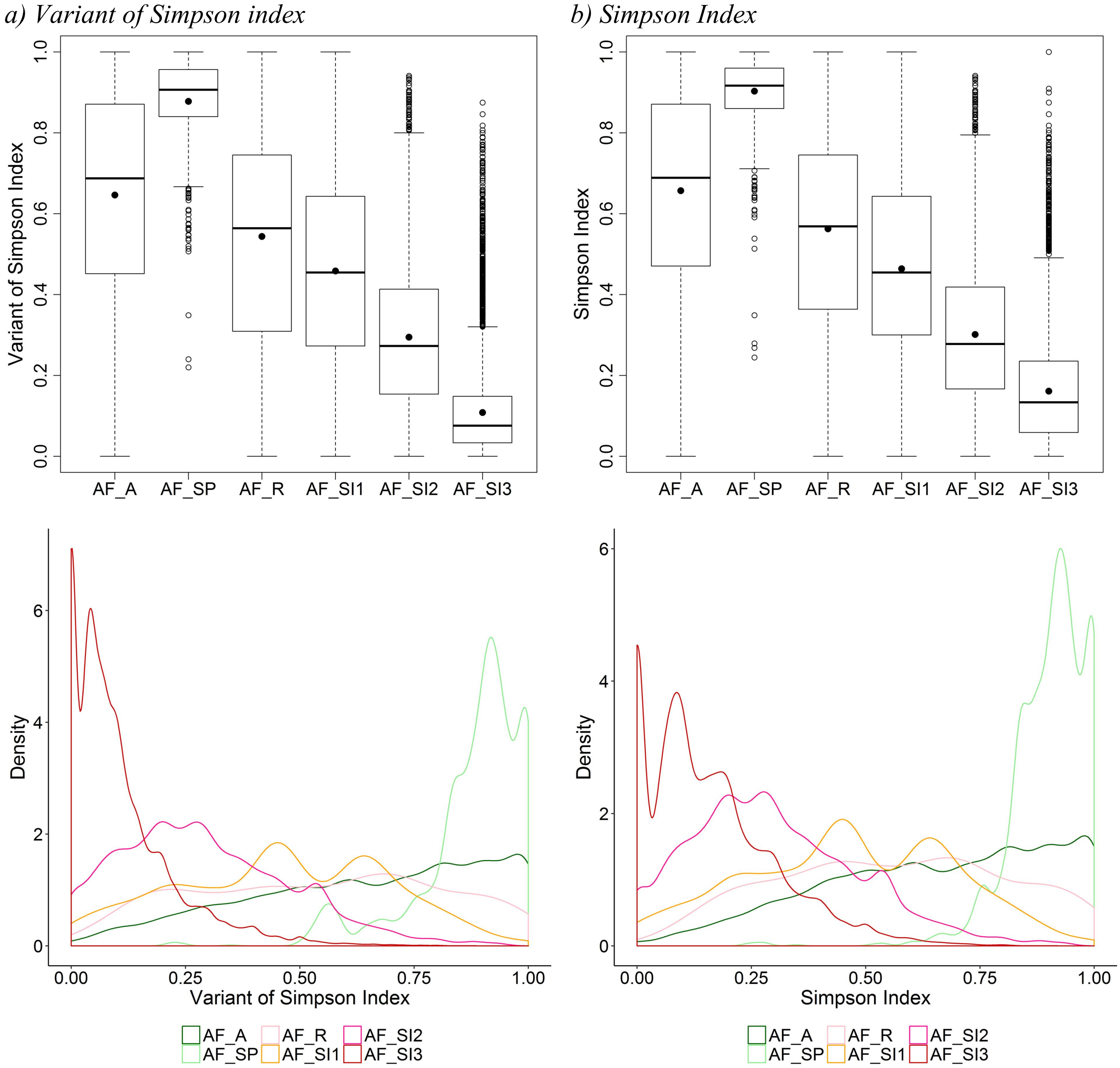
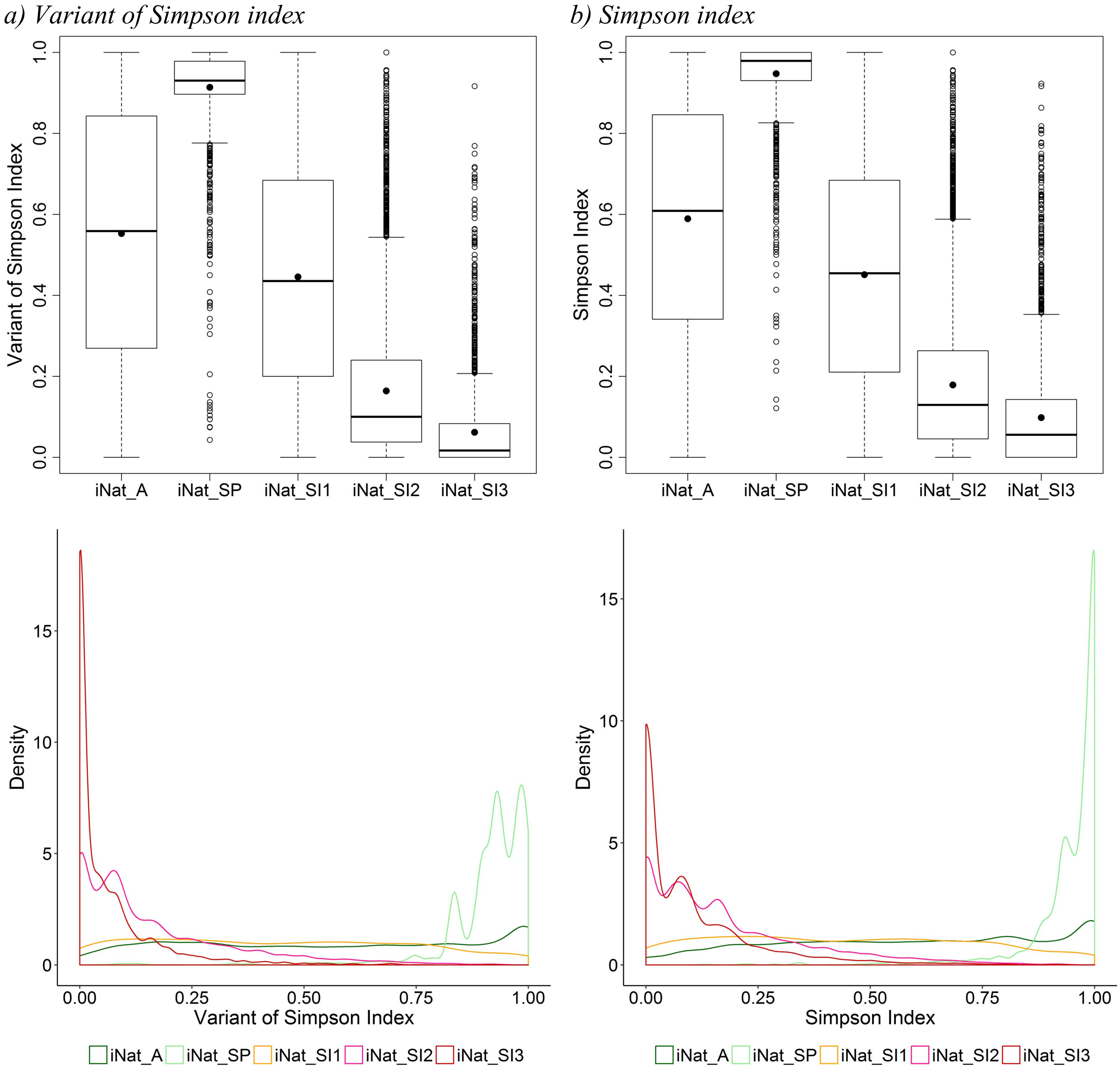


Figure 7.3.27: ArtenFinder, distributions of variant of Simpson index and original Simpson index values.

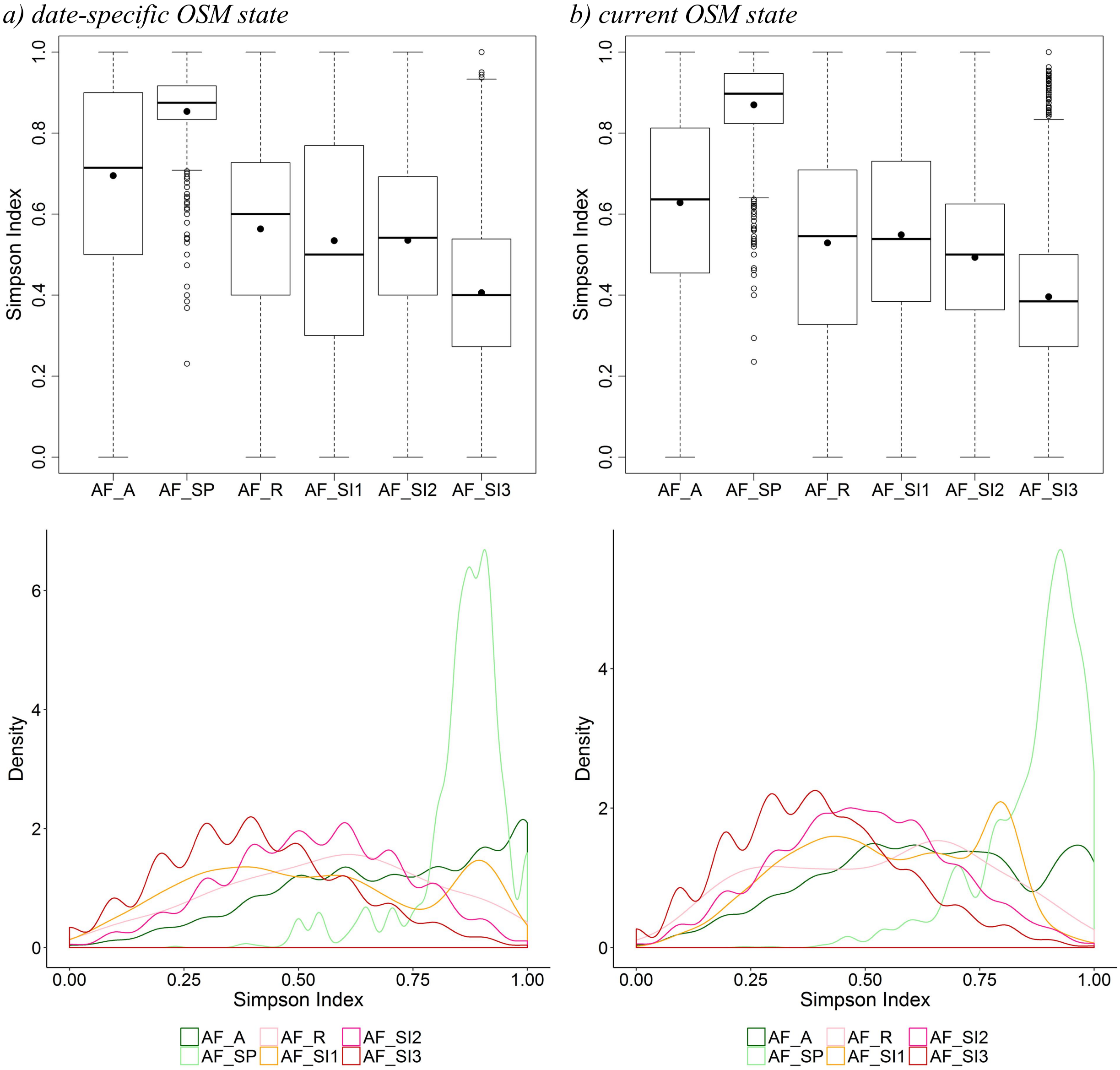


a) Variant of Simpson index

b) Simpson index

Figure 7.3.28: iNaturalist, distributions of variant of Simpson index and original Simpson index values.

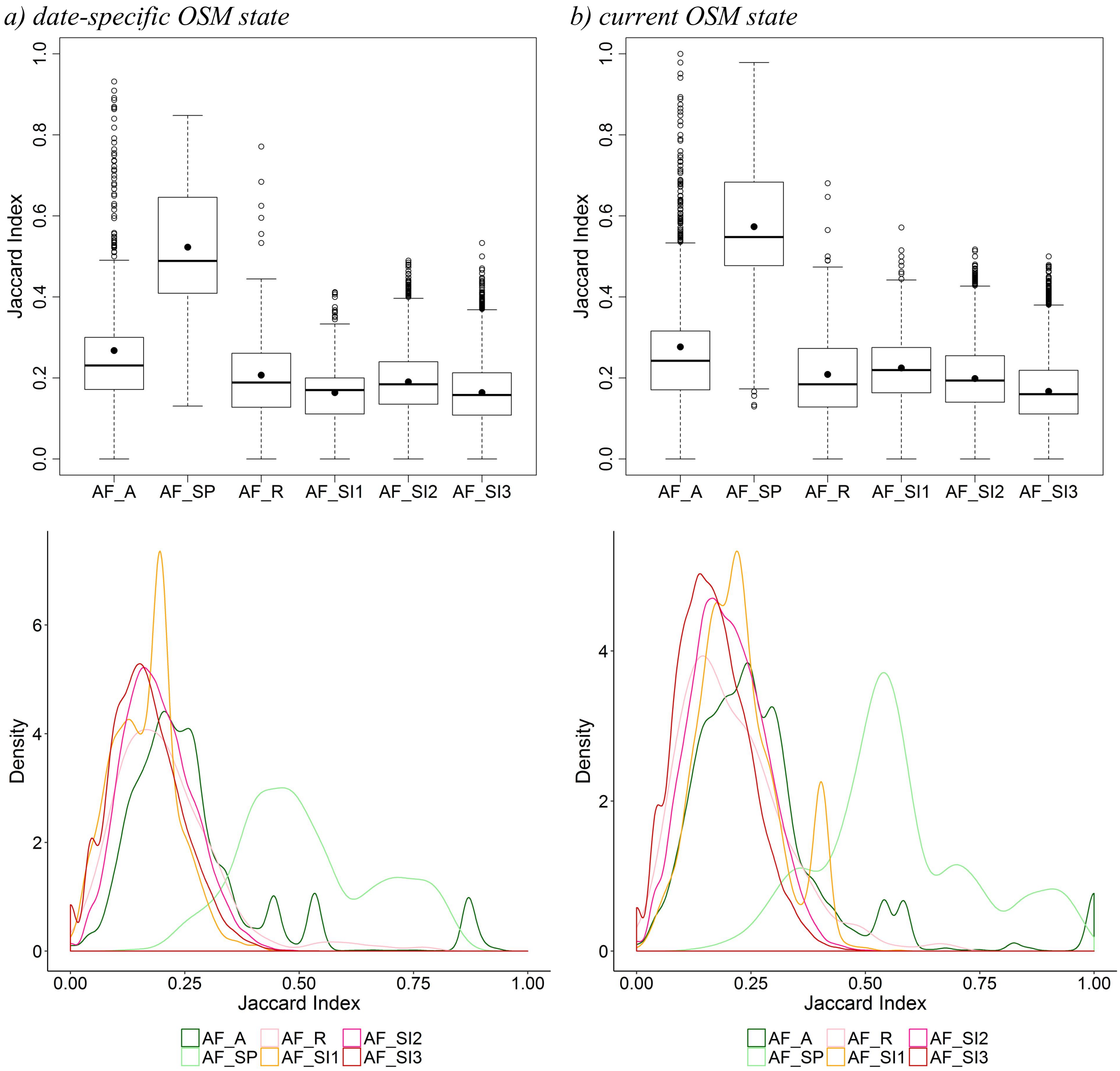
7.3.8 Using Date-Specific OSM Context



a) date-specific OSM state

b) current OSM state

Figure 7.3.29: ArtenFinder, distributions of Simpson similarity index values, analysis with date-specific OSM state (a), and with current OSM state (b).


a) date-specific OSM state

b) current OSM state

Figure 7.3.30: ArtenFinder, distributions of Jaccard similarity index values, analysis with date-specific OSM state (a), and with current OSM state (b).

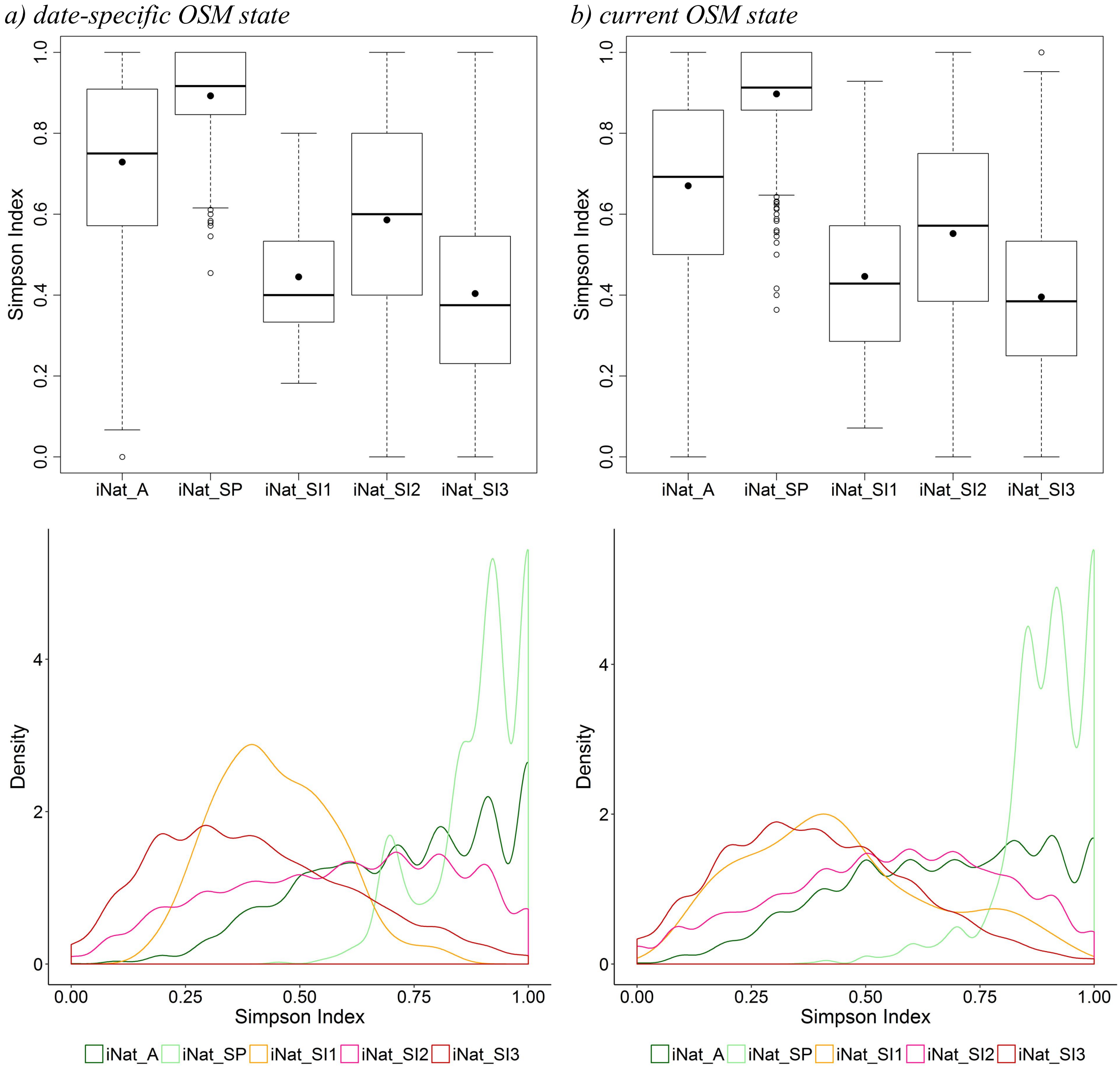


Figure 7.3.31: iNaturalist, distributions of Simpson similarity index values, analysis with date-specific OSM state (a), and with current OSM state (b).

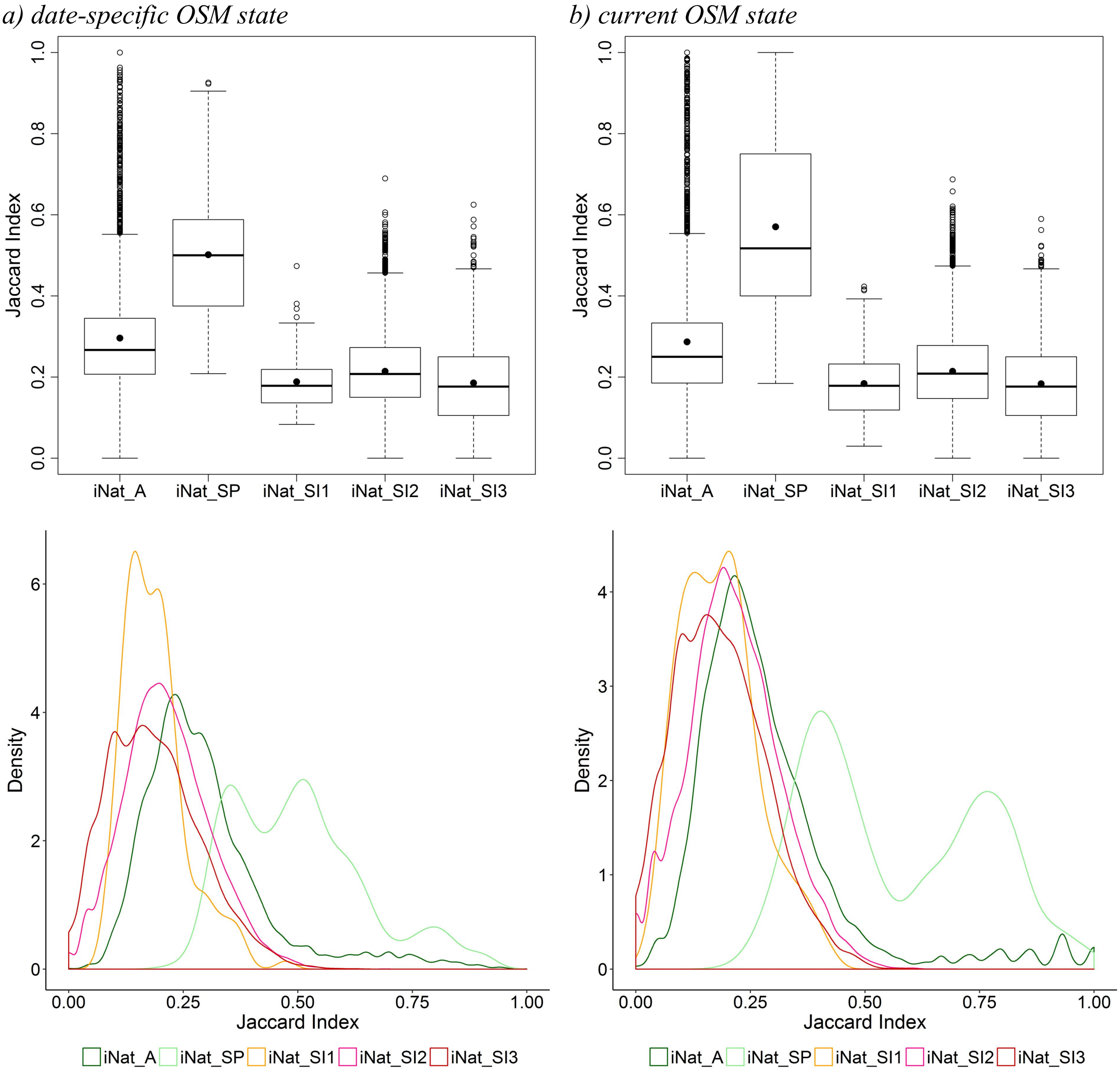


Figure 7.3.32: iNaturalist, distributions of Jaccard similarity index values, analysis with date-specific OSM state (a), and with current OSM state (b).

Table 7.3.18: Analysis with date-specific OSM state, key numbers describing valid observed communities.

	No. of valid OSM environments		Mean no. of tags in OSM environments		No. of nonspecific tags		No. of nonspecific species	
	with OSM state:							
Data use case	current	date-specific	current	date-specific	current	date-specific	current	date-specific
ArtenFinder	402	275	19.5	16.5	25	20	39	39
iNaturalist	635	606	20.1	18.3	9	3	0	0

Table 7.3.19: Analysis with date-specific OSM state, key numbers describing sets of valid candidate observations.

	No. of valid candidate cases		Mean no. of tags in candidate contexts			No. of valid candidate cases		Mean no. of tags in candidate contexts	
	with OSM state:					with OSM state:			
Set of candidates	current	date-specific	current	date-specific	Set of candidates	current	date-specific	current	date-specific
AF_A	15,329	10,647	35.6	37.3	**iNat_A**	35,058	28,873	33.6	33.0
AF_SP	1,568	1,317	39.8	38.2	**iNat_SP**	2,131	1,627	29.9	28.5
AF_R	215	177	32.8	30.3					
AF_SI1	2,104	996	33.2	31.7	**iNat_SI1**	290	161	28.9	27.8
AF_SI2	14,964	10,473	34.0	35.0	**iNat_SI2**	34,654	28,588	32.5	32.0
AF_SI3	15,618	11,114	23.2	23.1	**iNat_SI3**	4,542	4,408	18.5	18.2

Eidesstattliche Versicherung gemäß § 8 der Promotionsordnung der Naturwissenschaftlich-Mathematischen Gesamtfakultät der Universität Heidelberg

1. Bei der eingereichten Dissertation zu dem Thema „Data Quality of Citizen Science Observations of Organisms: Plausibility Estimation Based on Volunteered Geographic Information Context“ handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum

Clemens Jacobs