UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Biodiversity informatics – Entomological data processing, analysis and visualization

Leonor Fernanda Venceslau Azeredo Pontes

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Doutora Alexandra Marçal Correia
Doutor Luis Filipe Lopes

2019

Acknowledgements

First of all, a big thank you to my advisors, Dr. Luis Filipe Lopes and Dr. Alexandra Marçal Correia. They gave me the opportunity of working in a fascinating project and they were always available whenever I had doubts or needed help.

I would also like to thank all the colleagues and friends at the Zoology department of the *Museu Nacional de História Natural e da Ciência*, including, but not limited to: Isabel Queirós Neves, Leonor Brites Soares, Diogo Parrinha, Dr. Paula Souto, Dr. Alexandra Cartaxana, Dr. Cristiane Bastos-Silveira and Dr. Judite Alves. In different ways, they all contributed to help me with this work, either during the first months or later when I was working with a research fellowship at the Museum while finishing this project.

To Professor Luis Mendes, for help with finding the Tabanidae specimens of the collections of the *Instituto de Investigação Científica Tropical* and their records; the digitization process would have been much harder without his help. To Dr. Hélcio Gil, for the valuable information about Tabanidae species, and for verifying taxonomic identifications of the specimens to add more information to the Tabanidae dataset.

To everyone at the *Instituto de Higiene e Medicina Tropical* who welcomed me and provided insight during the process of digitization of the Tabanidae collection of the IICT.

I would like to thank my family for all the love and support without which it wouldn't have been possible to complete this work. To Diogo Simões, for being there for me everytime I needed, listening and helping with every problem along the way. To my mother, my father, my brother, my sister and my grandparents. They all believed in me unconditionally and helped me believe in myself.

And finally, to everyone not mentioned here, but who directly or indirectly helped me during this work, thank you!

Abstract

This work is based on data records associated with the insect Collections from the *Museu Nacional de História Natural e da Ciência* (MNHNC) and *Instituto de Investigação Científica Tropical* (IICT), *Universidade de Lisboa*. In 2014 a dataset with 30 535 records was published in the Global Biodiversity Information Facility (GBIF). Since then data has been improved and new records acquired. Currently, the collection catalogue includes 39 139 validated records, corresponding to 79 885 specimens, with much more to be added from collections donated by private collectors or unprocessed samples. The data for these specimens was cleaned, formatted and geocoded and published on the GBIF.

During this work, different APIs were tested to allow automated geocoding of sampling locations. Google Maps achieved the best results, with 57.6% of results within 1000 m of the correct location. A citizen science project was developed and tested to accelerate the digitization process, including two workflows with different objectives. One was focused on the transcription of specimen label data, which resulted in the data for 130 specimens being successfully transcribed. The other was focused on the taxonomic identification of specimens from photographs, directed to specialists in the respective group's taxonomy, which resulted in 61 new identifications and the verification of identifications for the remaining 69 specimens.

The MNHNC and IICT collections contain collections of horseflies (Order Diptera, Family Tabanidae) which are of particular importance due to its size and completeness of associated data. Horseflies are widely distributed worldwide and are important vectors in transmission of diseases to humans and cattle. The IICT collection includes a sub-collection which was compiled and studied by J. A. Travassos Santos Dias, a prominent specialist in this group. The specimens in these collections were photographed, all the associated data were transcribed, taxonomic identifications were verified and records were geocoded, resulting in a dataset of 1666 specimens. These specimens were collected between 1899 and 2018, mainly in Portugal, but also in São Tomé and Príncipe, Guinea-Bissau, Mozambique, Spain and other countries. To better understand the distribution of this group, distribution maps were made for the most well-represented species in the collections.

Keywords

Natural history collections; data digitization; data cleaning; geocoding; citizen science

Resumo alargado

Este trabalho foca-se na digitalização, tratamento e análise de dados de colecções de história natural fazendo uso de ferramentas da informática da biodiversidade. Foram usados dados das colecções de insectos do Museu Nacional de História Natural e da Ciência (MNHNC) e do Instituto de Investigação Científica Tropical (IICT), Universidade de Lisboa. Em 2014, um *dataset* com 30 535 registos da colecção de insectos do MNHNC foi publicado no *Global Biodiversity Information Facility* (GBIF). Desde então, novos registos foram digitalizados e foram adicionados novos dados, tais como novas identificações taxonómicas, entre outros. Actualmente, o catálogo da colecção de insectos do MNHNC inclui 39 139 registos validados, que correspondem a cerca de 98% do total, referentes a 79 885 espécimes. Para que este dataset actualizado pudesse ser publicado, foram aplicadas ferramentas de limpeza de dados para detecção e correcção de erros, bem como a georreferenciação de registos, de forma a que os dados possam ser localizados num mapa a partir das coordenadas. Relativamente à limpeza e homogeneização de dados, todos os campos foram limpos e formatados de acordo com as normas do modelo de metadados DarwinCore. Este processo incluiu a verificação de identificações taxonómicas para detectar sinonímias e erros ortográficos, alteração do formato de datas e horas, e aplicação de um vocabulário controlado para os restantes campos.

Paralelamente a este processo, foram testadas ferramentas para acelerar a digitalização em duas fases diferentes: transcrição e georreferenciação de dados a partir de etiquetas de espécimes.

Foram testadas cinco ferramentas de georreferenciação que disponibilizam *Application Programming Interfaces* (APIs), que podem ser usadas para georreferenciar registos automaticamente a partir de nomes de localidades: *Google Maps*, *MapQuest*, *GeoNames*, *OpenStreetMap* e *GEOLocate*. Destes, a ferramenta *Google Maps* foi a que produziu melhores resultados, com 57.6% dos resultados a uma distância de 1000 m ou menos do local correcto.

Foi também desenvolvido e testado um projecto de ciência cidadã na plataforma *Zooniverse*, que contemplou duas vertentes: uma de transcrição de dados a partir de fotografias de espécimes com etiquetas, direccionada ao público geral, e uma de identificação taxonómica de espécimes a partir de fotografias, direcionada a especialistas na taxonomia do respectivo grupo. A primeira vertente resultou na transcrição com sucesso dos dados de todos os 130 espécimes disponibilizados. A segunda resultou na identificação dos 61 espécimes que não tinham identificação prévia, e na verificação das identificações dos restantes 69 espécimes. Conclui-se, portanto, que os projectos de ciência cidadã serão uma boa maneira de acelerar o projecto de digitalização, desde que sejam implementados métodos de verificação e correcção de erros adequados.

Por forma a focar todos os passos do processo de digitalização de uma forma mais completa, foram selecionadas as colecções de tabanídeos (Diptera: Tabanidae) do IICT e do MNHNC. Este grupo é de especial interesse por incluir importantes vectores de transmissão de doenças a humanos e gado, e por ter uma distribuição ampla em todo o Mundo. A colecção de tabanídeos do IICT é particularmente importante por ter sido, na sua maioria, compilada e estudada por J. A. Travassos Santos Dias, um especialista neste grupo que publicou extensos trabalhos com base nos espécimes da colecção. Ambas as colecções incluem espécimes tipo de espécies descritas por Travassos Santos Dias e outros autores. Apesar da sua importância, a informação associada aos espécimes das colecções do IICT/MNHNC ainda não estava digitalizada. Neste trabalho, foram fotografados todos os espécimes e transcritos os seus dados, resultando num dataset com 1 666 exemplares. Foi feita a georreferenciação dos registos sempre que possível. Os espécimes da colecção foram recolhidos entre 1899 e 2018, maioritariamente em Portugal, mas também em São

Tomé e Príncipe, Guiné-Bissau, Moçambique, Espanha e outros países. Para uma melhor visualização da distribuição geográfica dos espécimes, foram criados mapas de distribuição, recorrendo a R, para as espécies mais bem representadas nas colecções. A publicação deste dataset na plataforma GBIF será uma mais-valia para o estudo da distribuição deste grupo, devido à sua ampla cobertura geográfica e temporal, bem como ao facto da maioria dos espécimes (85.1%) estarem identificados até à espécie ou subespécie.

Palavras-chave

Colecções de história natural; digitalização de dados; limpeza de dados; georreferenciação; ciência cidadã

Contents

List of Figures

List of Tables

1. Introduction

Natural history collections (NHC) are a powerful source of information for research with many possible applications. They are essential for systematic studies, since type specimens used to describe species are stored there and constitute an important reference. Moreover, many new species are identified from specimens in these collections [1]. Often, specimens stored in museums represent species that would be difficult or impossible to collect or study in the present time, due to population reduction or extinction, costs of revisiting field sites or restrictive export policies [2]. These collections can be used for training taxonomic specialists [3, 4] and the information contained in these records can be used for population distribution analysis and modelling, for molecular biology and morphology studies [5–7], as well as to define strategies to deal with ecosystem changes due to issues such as climate change, deforestation, overfishing, infectious diseases, and invasive species [8–10]. According to Drew (2011), conservation biology benefits from natural history institutions in three main ways: collection-based research projects and taxonomic expertise, collections data digitization and public engagement. Because NHC usually contain specimens collected over several decades or even centuries, they hold data that can be used for studies across long periods of time [3]. By combining data from several collections, especially if it is available online and in a standardized format, studies can be conducted within wide geographic and temporal ranges, allowing large scale analyses of changes in biodiversity and species distribution.

In 2002, the Convention on Biological Diversity, ratified by 196 Countries, set a target to significantly reduce the worldwide rate of biodiversity loss by 2010, which was incorporated in the United Nations Millennium Development Goals. Not only this target was missed, but also some of the factors contributing to biodiversity loss have increased in this period. The five main factors identified by member countries were habitat loss, the unsustainable use and overexploitation of resources, climate change, invasive alien species, and pollution. Some of the main obstacles pointed out to achieving a reduction on biodiversity loss were the lack of scientific information and lack of awareness among the public and decision makers regarding this issue [11]. Making biodiversity data freely accessible, not only to the scientific community but also to Governments and the general public, will be an important step to raise awareness and generate the necessary detailed information to drive conservation measures to counter biodiversity loss [12]. This illustrates the importance of making biodiversity data publicly accessible, in particular that of NHC, as they represent a rich source of information. As technology advances, NHC specimens and record data will have even more applications [9]. In fact, there have been concerted global efforts to digitize these collections and to make that information public [9, 13]. Thousands of biodiversity datasets, related to NHC, observation data, and others, have been published on digital repositories, where they can be freely accessed and retrieved for analysis. For instance, the Global Biodiversity Information Facility (GBIF, www.gbif.org) currently includes over 45 000 datasets, including data from NHC, field observations, among others; the Integrated Digitized Biocollections portal (iDigBio, www.idigbio.org/) contains over 1 600 NHC datasets. Despite this effort and the importance and usefulness of NHC data, only a small part is published and available online. The most recent estimates indicate that a total number of 1.2 to 3 x 10^9 specimens are stored in museums worldwide [9, 14, 15]. There are currently 164 x 10^6 records based on preserved specimens stored on GBIF, corresponding to 5.3% of the estimated total of 3 x 10^9. The percentage of digitized information for insect collection specimens is estimated to be lower than 2% [2].

One of the main obstacles to publishing biodiversity data has to do with the backlog of data yet to be digitized, particularly in NHC. In some cases, a significant part of the specimens hasn't been screened and catalogued yet. Additionally, for many of the specimens already processed data have not been digitized due to the lack of resources, mainly staff. Therefore, major efforts are being made in natural history museums worldwide to digitize data on their collections [3], and to develop methods to automate

digitization, either for specimen-level data capture or in bulk – for example, using whole-drawer imaging methods [16–18]. Wheeler *et al*. (2012) proposed a set of guidelines to describe and redescribe 10 million animal and plant species within 50 years. One of the steps proposed is to digitize all NHC data, with a special focus on type specimens, including storing recorded data and photographs on biodiversity databases – ultimately creating a catalogue of all museum specimens worldwide. Taxonomy literature should also be digitized and freely accessible to complement this data. The ultimate goal of this work is to create a knowledge base of life on Earth, including morphology data, distribution, genomic sequences and automated classification tools based on morphology and sequence data [9].

In order to publish a dataset of NHC records, it is first necessary to digitize, normalize and validate the data, as most of the original records consist of index cards, labels or registry books and do not follow standard metadata models. The five main stages of the digitization of NHC are pre-digitization curation and staging, specimen image capture, specimen image processing, electronic data capture, and geocoding locality descriptions [19]. According to Guralnick *et al.* (2006), the three main challenges of creating such a dataset are: i) transcribing the data to a computer database, ii) geocoding the records, and iii) publishing it online [20]. Several tools have been developed and optimized to address these challenges, which are discussed throughout this work.

1.1. Objectives

The main objective of this work was to foster the digitization, verification/validation and online publishing of biodiversity data from two of the largest entomological collections in Portugal, held by the IICT and MNHNC, from the University of Lisbon, therefore increasing the collections accessibility and visibility to scientists and the civil society.

For this end several methods were tested and implemented with the objective of a faster and more accurate processing of large datasets, with specific objectives presented and developed in four different chapters:

1) Comparison of methods for automated geocoding;

2) Cleaning, enrichment and publication of the MNHNC insect collection data;

3) Implementation of a project on the Zooniverse platform to assess the possible contribution of citizen science to digitization of NHC data;

4) Digitization, publication and analysis of a tabanid fly dataset.

2. Evaluation of automated geocoding tools

2.1. Introduction

Geocoding occurrence records is an important step of the digitization and data enrichment process of NHC records. Geocoding, or georeferencing, consists of obtaining geographic coordinates from a location description, for example, the name of a city or an address [21]. For the vast majority of NHC specimens, the coordinates of the sampling location are not recorded in the field, but rather a textual description of the sampling locality. Accurate geocoding of sampling locations is very important, as it will affect the quality of the data and the accuracy of any distribution models derived from it [22]. Therefore, in order to conduct any kind of geographic analysis, such as species distributions, records must be geocoded. This is usually done during the digitization process. The most common practice for geocoding is to record the coordinates for the point that most closely matches the description of the sampling locality. Since in most cases it is impossible to pinpoint the exact location retrospectively, in addition to latitude and longitude, it is also necessary to estimate an uncertainty value, usually in the form of a radius, in meters, around the estimated sampling point [22, 23]. The uncertainty radius might define, for example, the approximate boundaries of the locality where the sampling took place.

Geocoding is a difficult and time-consuming task, and for some records it is almost impossible to obtain coordinates, due to missing or incomplete information registered at the time of collection. Of the 164 million records of preserved specimens currently available in GBIF, only 87 million (53%) include coordinates. Regardless of the methods used for geocoding, the results always have to be individually verified. There can be errors, for example, if there are different locations with the same name, or if a certain location had its name changed or ceased to exist. For geocoding, as for transcription of record data, citizen science has been proposed as a way to make the process faster, especially if the volunteers are given records of specimens collected in countries or areas they are familiar with [24].

Since geocoding has many applications and is used in many fields of research, several tools have been developed to offer an Application Programming Interface (API) for automated/batch geocoding. Examples of these are Google Maps (https://www.google.com/maps), Mapquest (https://www.mapquest.com), Geonames (https://www.geonames.org) and OpenStreetMap (https://www.openstreetmap.org). Because geocoding is such a significant part of the effort to digitize biodiversity data, GEOLocate (http://www.geo-locate.org) has been developed as a specific tool for this purpose. Its website includes a user interface that allows the user to select a point from a list of results and to define an uncertainty radius or polygon on a map. It also allows matching water bodies or highway/river crossings for aquatic specimens.

Of the geocoding services mentioned, only Google Maps has usage restrictions – it currently allows around 40 000 free geocoding requests per month, beyond which the service is paid. This might be a factor to consider for very large databases.

In 2004, Murphey *et al*. published a comparative review of the geocoding tools available for museum collections data and concluded that the fastest and most accurate method for geocoding collection records was manual geocoding [25]. This was mainly due to the time required to pre-process data and to validate results obtained with the existing automated tools. Since then, new tools for geocoding have been developed and optimized.

2.1.1. Objectives

The objective of this work was to test different geocoding APIs using R with records of the MNHNC insect collection catalogue, to assess which one presents better results in terms of total number of results and their accuracy.

2.2. Methods

Five geocoding services were selected for comparison: Google Maps, MapQuest, GeoNames, OpenStreetMap and GEOLocate.

A test dataset of 100 records from the catalogue of the MNHNC insect collection was selected, including sampling locations from 14 countries. Of the records selected, 50 were from Portugal, due to the high prevalence of specimens from this country in the collection; the records were selected to include simple locations, such as names of cities or villages. These records were individually geocoded, to be used as a standard (reference location) for comparison against the automated method results. A csv file was created with columns for the sampling country, state/province, island, county, municipality and locality, all in the DarwinCore format. For each record, all available data was included in this file. Geocoding was done according to the protocol by Chapman and Wieczorek (2006) [23]. Each location was geocoded using the GEOLocate Web Application (https://www.geo-locate.org/web/WebGeoref.aspx), and when necessary, coordinates were confirmed using the Google Maps website (https://www.google.com/maps). The results were saved in a csv file containing the latitude, longitude and uncertainty in meters for each location. All coordinates used were in the WGS84 standard.

An R [26] script was created to iterate through all the locations, concatenate all data available for each record in a string and call the API for each of the 5 services to obtain coordinates. In the cases where more than one result was returned by the API, only the first one in the list was considered. The results were saved in csv files, containing the latitude, longitude and uncertainty in meters in the case of GEOLocate, and only the latitude and longitude for the other four services. The script used is available in Annex A.

For a second test, a subset of records of specimens collected in Portugal and with complex location descriptions, such as *between x and y* or *5 km North of x* was manually selected and geocoded using the same method as before. In order to ensure optimal results for each tool, location descriptions were translated into English prior to geocoding.

Two factors were considered in evaluating accuracy: 1) total number of results obtained and 2) distance to the reference location. For each result, the distance to the reference location was calculated using the geosphere R package, with the 'Vincenty' (ellipsoid) great circle distance function, corresponding to the function distVincentyEllipsoid [27].

For the second test using complex location descriptions, because Google Maps and GEOLocate yielded results with very similar average distances to the reference location, a test was performed in R to compare the distance values obtained. The distances to reference locations obtained with Google Maps and GEOLocate, for locations that yielded results in both tools, were saved in a dataframe. The function shapiro.test() was used to test normality of the differences between the values, resulting in a p-value < 2.2^{-16}, meaning the hypothesis of normality is rejected. Therefore, a paired samples Wilcoxon test was performed to compare whether or not the two vectors of distances were significantly different, using the function wilcox.test(), with the null hypothesis that difference between the distances obtained with the two tools was equal to 0.

For each tool, four indicators were calculated:

1. Total number of results for which coordinates were returned;
2. Average distance to the reference location, considering all results returned;
3. Total number of results that were at a distance of 1000 m or less from the reference location;

4. Total number of results with a distance to the reference location smaller than the reference uncertainty radius.

Two different measures of distance (number of results less than 1000 m from the reference location and number of results within the uncertainty radius) were considered due to the fact that the size of the sampling location can vary by several orders of magnitude; e.g. for large cities it can be several kilometres, for small villages it can be hundreds of meters. In either case, any result within the given area should be considered acceptable.

2.3. Results

A first test was conducted with a random set of 100 locations from the whole insect database. The results of this test are summarized in Table 2.1. Of the five programs tested, Google Maps yielded the most results (99) and was the most accurate with 57 results within 1000 m from the reference location and 79 within the reference uncertainty radius. GEOLocate provided results for 87 locations, of which 47 were within 1000 m of the correct location, and 57 were within the uncertainty radius. The other 3 services tested had less than 35 results within 1000 m from the reference location, and less than 50 results within the uncertainty radius. Regarding the average distance from the reference location, Google Maps and OpenStreetMap were the most accurate with comparable results, both around 5 500 m. The average distance for the results from GEOLocate was much higher (111 277 m), because 13 results were over 150 km off the reference location; of these, the maximum distance from the reference location was of 1 719 km.

Table 2.1 Results of geocoding 100 insect records with the 5 APIs tested. Coordinates and uncertainty radius for the reference location of each record were obtained by geocoding using the GEOLocate Web Application, confirmed by Google Maps where necessary. Distance to the reference location for each result was calculated using the 'Vincenty' (ellipsoid) great circle distance function.

	GEOLocate	GeoNames	Google Maps	MapQuest	OpenStreetMap
Total number of results	87	56	99	39	61
Average distance from reference location (m)	111 277	12 665	5 544	858 870	5 582
Number of results within 1000 m (% of total results)	47 (54.0)	27 (48.2)	57 (57.6)	12 (30.8)	34 (55.7)
Number of results within uncertainty radius (% of total results)	57 (65.5)	38 (67.9)	79 (79.8)	17 (43.6)	47 (77)

For a second test, a subset of 100 records of the insect collection was selected, consisting only of locations in Portugal with especially complex descriptions. The results are summarized in Table 2.2.

For the complex location descriptions, GeoNames returned no results, and OpenStreetMap returned only 2. As in the previous test, GEOLocate and Google Maps presented the best results, with 97 and 100 results respectively, and an average distance from the reference point over 100 000 m. GEOLocate generated 7 results less than 1 km from the reference location, and 20 were within the uncertainty radius. Google Maps produced 10 results within 1 km of the reference point, and 25 within the uncertainty radius. MapQuest returned a total of 54 results, with an average distance 3 750 248 m from the reference location. Only 3 of the results were less than 1 km off the reference location and within the uncertainty radius.

Table 2.2 Results of geocoding 100 insect records with complex location descriptions, with the 5 APIs tested. Coordinates and uncertainty radius for the reference location of each record were obtained by geocoding using the GEOLocate Web Application, confirmed by Google Maps where necessary. Distance to the reference location for each result was calculated using the 'Vincenty' (ellipsoid) great circle distance function.

	GEOLocate	GeoNames	Google Maps	MapQuest	OpenStreetMap
Total number of results	97	0	100	54	2
Average distance from reference location (m)	104 197	-	108 109	3 750 248	4 377
Number of results within 1000 m (% of total results)	7 (7.2)	0	10 (10)	3 (5.6)	1 (50)
Number of results within uncertainty radius (% of total results)	20 (20.6)	0	25 (25)	3 (5.6)	1 (50)

In order to assess if accuracy of the results obtained with Google Maps and with GEOLocate for complex location descriptions was significantly different, a paired samples Wilcoxon test was performed on the vectors of distances to the reference locations obtained with each tool. This resulted in a p-value of 0.0002219, so the null hypothesis that the distances obtained by both tools were not significantly different was rejected at any reasonable significance level. Although the average distance obtained with GoogleMaps is higher (approximately 108 km against approximately 104 km for GEOLocate), this happens due to outliers, as is suggested by the fact that Google Maps produced more results within 1000 m and within the uncertainty radius than GEOLocate.

2.4. Discussion

In order to evaluate the tools tested, all factors considered should be taken into account. In terms of number of results, Google Maps and GEOLocate clearly stand out, with 99% / 87% of simple locations, returning results, and 100% / 97% for complex locations, respectively. Regarding complex location descriptions, results were expected to be less in number and less accurate than for simple ones. Interestingly, both tools, as well as MapQuest, returned more results for complex locations than for simple ones. This may be related to the fact that often, complex locality descriptions include two or more place names, sometimes with references to roads and rivers (e.g. "Rio Ponsul, E.N.332 near Idanha-a-Velha", one of the locations used for testing); it is more likely for the service to find one of these locations than in cases where only one location name is provided as input. Therefore, for complex locations more results were returned, but with very little accuracy. For instance, for the previous example, GEOLocate returned the coordinates for the center of Idanha-a-Velha, since it is not able to interpret the road and river used as references for the sampling location. For the same description, Google Maps returned a different point of the river Ponsul, 27 km from the actual sampling location. For locations of the type *x km from y*, the geocoding services will likely return the coordinates for one locality, without considering the offset. Of the tools tested, GEOLocate is the only one that was developed to handle these cases. Because it was developed specifically for NHC data, it can identify displacements. However, this service is not available in Portuguese yet, so to use this functionality it is necessary to first translate all locality descriptions into English.

As for the accuracy of the results for simple locations Google Maps and GEOLocate again delivered the best results, both returning over 40% of results within 1000 m of the reference location and over 50% within the uncertainty radius. When considering the average distance from the reference location, Google Maps and OpenStreetMap had similar results, both around 5500 m. However, OpenStreetMap yielded much less results in total (61% vs. 99% for Google Maps). The results generated were, however,

more accurate than the results returned by GEOLocate, with 77% of the total 61 within the uncertainty radius, against 65.5% of a total of 87 for GEOLocate.

Overall, both in terms of total number of results and their accuracy, Google Maps seems to be the most suitable API for geocoding both simple and complex sampling locations. Although the results obtained for complex location descriptions seem to be similar for Google Maps and GEOLocate, a paired samples Wilcoxon test showed that they are significantly different. Furthermore, in the case of geocoding of sampling locations of NHC collections, the results always have to be verified manually, meaning that it is preferable to have a higher number of results close to the reference location (considered to be the correct one), than to have a lower average distance, but also a lower number of results close to the reference location, as was the case with GEOLocate.

For the case of biodiversity data and NHC, other factors besides the geocoding accuracy need to be taken into account. For instance, because GEOLocate was developed specifically to use with biodiversity data, it has a series of functionalities that are not available on Google Maps. It can detect displacements from a locality, and it provides an option to snap the coordinates to the nearest water body, which is useful for geocoding records of aquatic species. Another advantage of GEOLocate is the Collaborative Georeferencing Web Portal (http://www.geo-locate.org/web/webcomgeoref.aspx). This platform allows users to create datasets that are available to other users to be geocoded. In this way, different staff members or volunteers can collaborate to geocode the sampling locations of a collection.

Another factor to take into account is usage limits. Although all tools tested are free to use to some extent, Google Maps has some restrictions. It only allows around 40 000 free geocoding requests per month, beyond which the service is paid. This could be a limitation for very large collections. However, considering that the number of locations to be geocoded does not exceed the limit of 40 000 per month, a possible approach would be to automatically geocode all records using Google Maps, and then validate the results and determine an uncertainty radius using the GEOLocate Collaborative Georeferencing Web Portal. This would be a way to take advantage of both the accuracy and automation provided by the Google Maps API, and the possibility of manual confirmation (required for all records) and defining the uncertainty radius, offered by the GEOLocate web portal.

3. Data cleaning and enrichment of the MNHNC insect collection catalogue

3.1. Introduction

The primary goal of digitizing NHC data is to make it accessible to the scientific community, through online repositories or biodiversity databases. Costello *et al.* (2013) argued that all biodiversity data should be made available online with high quality, otherwise it cannot be used efficiently [28]. The increasing recognition of the importance of publishing datasets lead to the outset of data papers, i.e. papers describing datasets that can be cited to acknowledge authors who publish them [29]. Data papers were proposed as a way to increase the amount of primary biodiversity data available for analysis and citation. They consist of a description of a dataset that is accessible online in a specific database. This description includes information such as the taxonomic, spatial and temporal coverage of the dataset, by whom it was created and which tools were used [30]. It can also contain some preliminary analysis, depending on the data and the goals of its description.

Several repositories have been established over the years to store biodiversity data, some focusing on specific groups – e.g. Tropicos (http://www.tropicos.org) for plant specimens, AntWeb (https://www.antweb.org) for ant specimens, Avibase (https://avibase.bsc-eoc.org) for bird taxonomy and distribution - or locations – e.g. Georgian Biodiversity Database (http://www.biodiversity-georgia.net) for animals, plants, and fungi observations in Georgia, Naturdata (https://naturdata.com) for biodiversity data in Portugal, and others [10]. Two of the larger, worldwide biodiversity repositories are the Global Biodiversity Information Facility (GBIF; https://www.gbif.org), and the Integrated Digitized Biocollections database (iDIgBio; https://www.idigbio.org). GBIF was created in 2001 to store global biodiversity records of preserved specimens, observations and material samples, among others, and currently contains over 1.3×10^9 occurrence records. iDigBio was created as an infrastructure for the digitization and publishing of NHC records in the United States, and its database currently contains over 119 million specimen records.

In order for data to be shared, interoperable and easily used for analysis, stardardization procedures are necessary [10]. The GBIF uses the Darwin Core format, a metadata standard specific for biodiversity data derived from the Dublin Core metadata format for digital resources. Darwin Core was established in 2009 to define a standard to organize and format biodiversity data. It includes terms to describe an occurrence event, its location, geological context, taxonomy, among other details [31].

There are several ways to digitize the data from a NHC. The first step of this process mainly consists of transcribing information in labels, field notebooks, record books or cards. Sometimes, in the lack of human resources for these tasks, institutions turn to volunteers, who can either work at the museum or through citizen science digitization projects, which can be created in platforms such as Zooniverse (https://www.zooniverse.org) or SciStarter (https://scistarter.org). The second step is to compile all the information in the form of a database using a suitable metadata standard. After that data can be enriched, e.g. by geocoding sampling locations or adding taxonomic identifications.

After data are digitized they need to be cleaned and checked for inconsistencies and errors, especially very large databases. This is a laborious process that must be completed before data analysis and/or publishing on public repositories [19]. The most common errors in biodiversity data are geocoding errors and taxonomic identification errors [21]. When cleaning biodiversity data for publication and analysis, it is recommended that all species names are valid and in accordance to accepted taxonomic databases or checklists [32]. Another issue is format incongruence, for example, the collection dates for different records may be written in different ways (for instance, “2 December of 1989”, “2-12-1989” and ”1989-12-2” may all be used for the same date if recorded or transcribed by different people and a specific format has not been initially defined). Some tools have been developed to address these issues, such as

automated workflows for data cleaning and homogenization, R packages for taxonomic data verification and reconciliation services for use with Open Refine [33]. The use of these tools and other resources generally requires expert knowledge.

OpenRefine (http://openrefine.org) is a powerful tool for data cleaning that allows the use of General Refine Expression Language (GREL), a language specifically developed for cleaning and manipulating data, including making use of regular expressions. Reconciliation services can be added for specific cleaning steps; these allow for terms in a column of data to be searched across a database and matched to similar terms [34]. For example, a column containing the country where a specimen was collected may be reconciled against a database of countries, to check for misspellings or other errors.

The taxize R package was created to aid in the verification of taxonomic classifications of biodiversity data [35]. It uses several taxonomy databases, such as GBIF Backbone Taxonomy, Encyclopedia of Life, ITIS and NCBI Taxonomy, to resolve taxonomic names, returning for each a list of the closest matches in the selected databases. The function to resolve names uses fuzzy matching, allowing the detection of spelling errors, and retrieves the most up to date names in cases of synonymy. It can also return higher level taxonomic information and taxon authorship, among other details.

3.1.1. Objectives

In 2014, a dataset of the MNHNC insect collection of was published on GBIF, containing 30 535 records corresponding to approximately 66 000 specimens [36] and 26% geocoded records. This dataset did not include data from all the specimens held in the collection, as a large part still remains to be sorted, prepared and catalogued. Furthermore, since then the collection has grown significantly with the integration of two private collections and many more specimens were catalogued and their associated data digitized, which warranted the publication of an updated version of the dataset.

The MNHNC insect collection catalogue currently includes over 39 000 records, corresponding to over 79 000 specimens. In order to publish the complete dataset on GBIF, it was first necessary to clean and format the data. The objective of this work was to clean and enrich the collection database and increase the number of records, through several steps:

1. Formatting existing data according to the DarwinCore standard;
2. Geocoding as many sampling localities as possible for the remaining records;
3. Publishing the insect collection catalogue on GBIF.

Publication of an updated version of the dataset is a way to increase the accessibility of the MNHNC insect collection data, both in terms of the total number of records and the quality of data, which now includes a higher number and percentage of geocoded records and taxonomic determinations.

3.2. Methods

The dataset of the MNHNC insect collection contains 39 139 records of specimens and is stored in an Excel spreadsheet. Each record includes data related to the collection event, geographic location, taxonomic identification and location in the museum, among others.

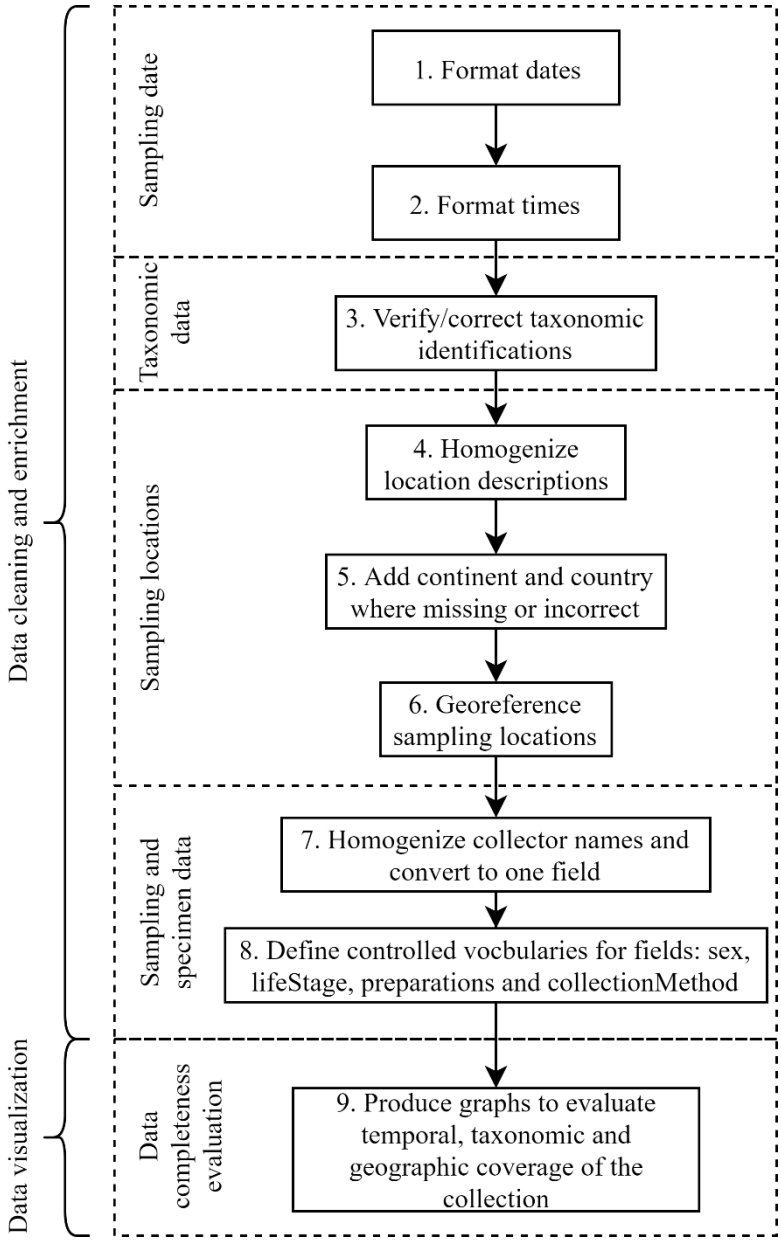

Figure 3.1 Steps employed in data cleaning and visualization for the MNHNC insect collection catalogue.

The methods used here (Figure 3.1) included a data cleaning process and then data completeness evaluation (i.e., information and taxonomic, temporal and geographic coverage of the collection) through visualization. The methods were applied as follows:

The first step of the data cleaning process was to format dates. The DarwinCore standard format best practice is defined to conform to the ISO 8601 international standard to represent date and time [37]. This format was used for the fields startDate, endDate and eventDate, and corresponds to "YYYY-MM-DD" for dates and "YYYY-MM-DD/YYYY-MM-DD" for date intervals. The dataset included dates in various formats, such as "DD/MM/YY" and "(DD-DD)-MM-YYYY". All of these formats were

identified and the Transform function of OpenRefine was used to change the dates to the correct format. To do this, regular expressions were constructed for each individual format using GREL. An example is shown below, to transform a string of format "(DD-DD)-MM-YYYY" to the format "YYYY-MM-DD/DD".

```
if(isNotNull(value.match(/\([0-9]{2}-[0-9]{2}\)-[0-9]{2}-[0-
9]{4}/)), slice(value,11,15) + "-" + slice(value,8,10) + "-" +
slice(value,1,3) + "/" + slice(value,4,6), value)
```

For the field eventTime, the same method was used to format all entries according to the DarwinCore standard, which is "hh:mm".

In order to confirm and correct taxonomic data, several steps were necessary. First, a list of canonical names was imported to OpenRefine and reconciled using the NCBI taxonomy standard service [38]. The reconciled names were verified individually to check for errors or incorrect matches, and the genera and/or specific epithets were updated where necessary. As a second verification step, the list of canonical names was imported into R [26], and the gnr_resolve function of the taxize package [35] was used to produce a list of corrected canonical names. For each canonical name, the corresponding Family and Order was obtained using the upstream function. The names that were different from the original were verified to check for errors. The resulting list was saved as a csv file and imported into Excel, where the VLOOKUP function was used to add the correct canonical name, Family and Order to matching records.

For the geocoding process, the first step was to homogenize location descriptions. For this, the Cluster feature of OpenRefine was used. Key collision methods were used first for clustering, followed by the nearest neighbour method. For each method, all clusters corresponding to the same location described in different manners, with differences in punctuation, letter case, etc. were merged. The columns continent and country were reconciled against the Wikidata knowledge base to correct misspellings.

After that, a csv file was created containing all individual locations that hadn't been geocoded yet. For each individual location, columns for the country, state/province, island, county, municipality and locality were included. This file was imported into R using the read.csv function; for each location a string was created containing all existing information, and these were geocoded using the Google Maps API to obtain latitude and longitude. The resulting coordinates were then individually verified through the GEOLocate web application, to check if they corresponded to the correct location, and an uncertainty radius was attributed to each location. The csv file including location descriptions, coordinates and uncertainty radii was imported to Excel, where the VLOOKUP function was used to add coordinates and uncertainty to all the records sampled from each location.

The last cleaning step consisted in homogenizing the remaining fields and defining controlled vocabularies where possible. The collector field was represented by several columns, one for each collector name; the Cluster feature of OpenRefine was used to homogenize names known to be different representations but corresponding to the same collector throughout all columns, which were then concatenated using "|" as separator, to create the column recordedBy. For the fields sex, lifeStage, preparations and collectionMethod, controlled vocabularies were defined and used to replace all values.

After cleaning and enriching the data, plots were produced using R to evaluate the completeness of the information and for a better visualization of the taxonomic, temporal and geographic coverage of the collection. All graphs were produced using the R graphics package [26], except the barplot with a gap produced to represent the countries where specimens were collected, for which the gap.barplot function of the plotrix package [39] was used.

3.3. Results

The dataset published on GBIF is available at https://www.gbif.org/dataset/79673413-746f-48f2-bd8a-7cf27807317e.The insect collection catalogue includes a total of 39 139 validated records, corresponding to a total of 79 885 specimens. Each record corresponds either to a single specimen or to a sample containing several specimens, collected at the same date and time, at the same location and by the same collector. The number of specimens for each record varies between 1 and 353.

A significant part of the collection was donated by private collectors. Only a small part of these donated collections was already catalogued. Of those that have been digitized and data integrated in the database, the Mendoça collection is the most well-represented in the collection catalogue, with 12 812 specimens recorded (16.0% of the total number of specimens). Other significant contributions are the collection donated by Teresa Pité, with 9897 specimens (12.4%), and the specimens collected in the EB network Biodiversity Stations (901 specimens, 1.1%) and by the Tagis – Butterfly Conservation Center (464 specimens, 0.6%).

The specimens in the MNHNC collection were collected between 1905 and 2018. The number of specimens collected per decade is shown in Figure 3.2. The decades when most specimens were collected were 1970-1979 (20 517, 25.7% of total) and 2000-2009 (30 405, 38.1% of total).

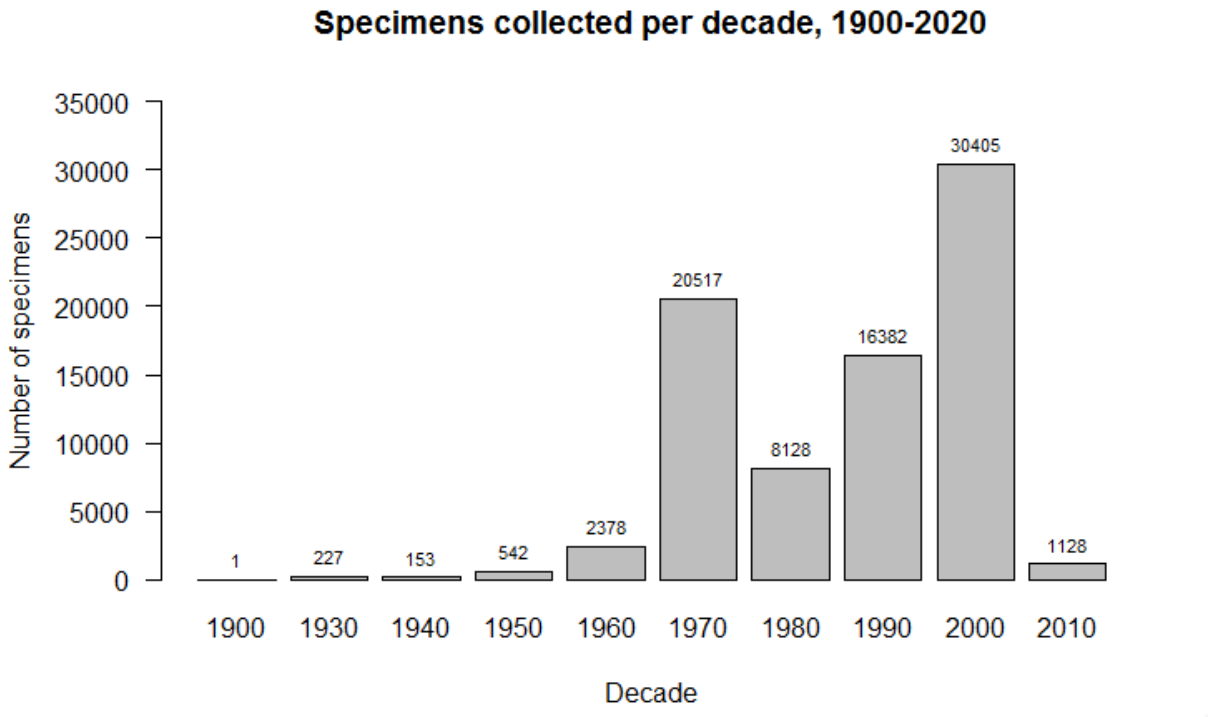

Figure 3.2 Histogram of specimens of the MNHNC insect collection by decade of collection.

The number of specimens collected per country, for the countries represented by over 100 specimens in the collection, is shown in Figure 3.3. The large majority of the specimens were collected in Portugal (65 838, corresponding to 82.4% of all specimens). For other countries, the most represented ones are Guinea-Bissau, Angola, São Tomé and Príncipe and Mozambique (all ranging between 1000 and 2000 specimens). Of the 39 139 records, 27 746 (70.9% of total) are geocoded. Of these, 9 252 were geocoded during this work.

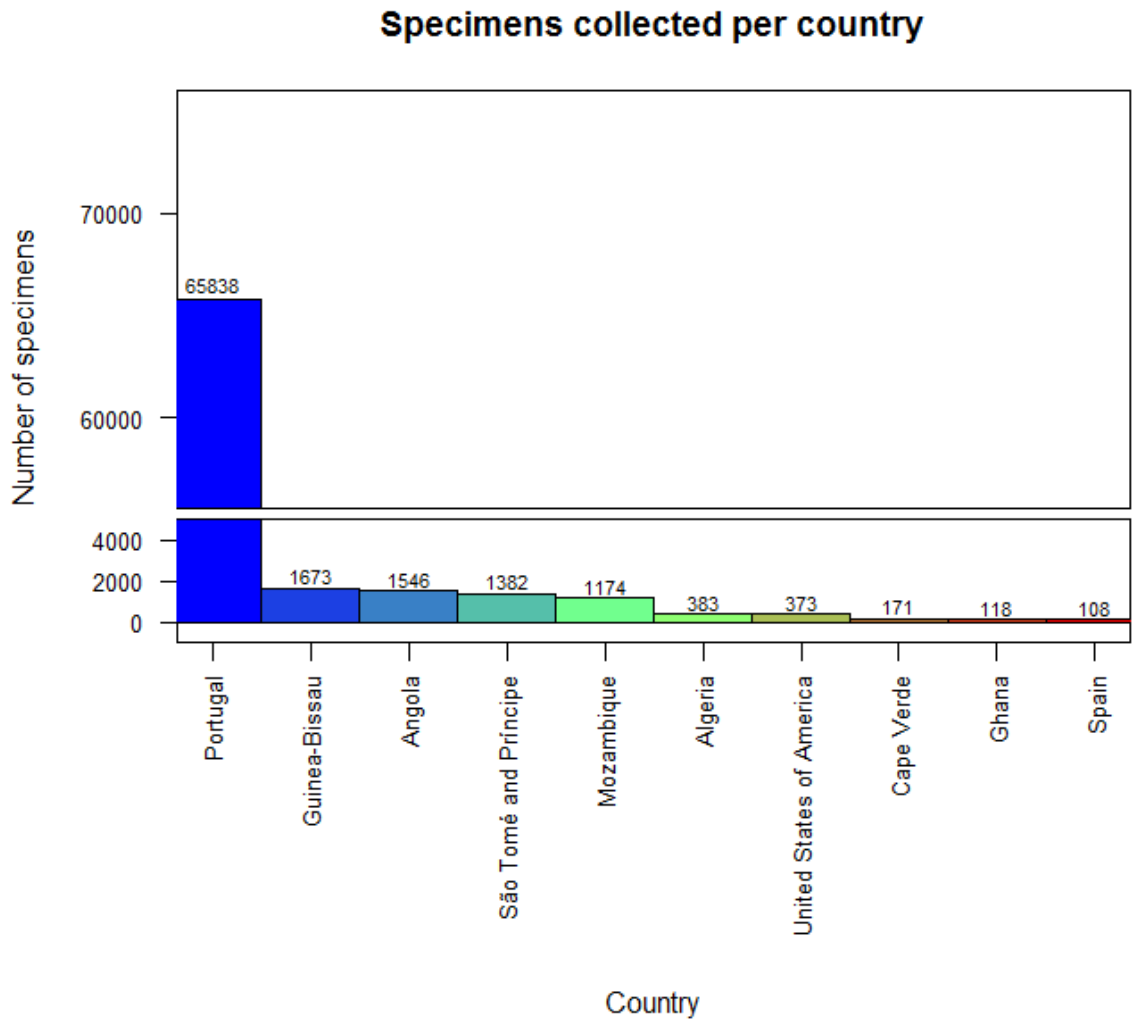

Figure 3.3 Bar graph of specimens of the MNHNC insect collection by sampling country. The plot includes the countries where 100 or more specimens were sampled.

Regarding the taxon rank to which specimens are identified (Figure 3.4 B), the majority are classified to the Species level (21 471, 26.9% of total) or to the Order level (49 824, 62.4% of total). Of the remaining specimens, 5313 (6.7%) are classified to the Family, 2757 (3.5%) to the Genus and 520 (0.7%) to the Subspecies.

The previous version of the MNHNC insect collection dataset was published on GBIF in 2014. The current version is more complete, includes more records and more detailed validated data. The version published in 2014 included a total of 30 535 records, corresponding to a total of 64 008 specimens. Of all records, 7916 (25.9%) were geocoded. In the new version of the dataset, the percentage of geocoded records increased to 70.9%. The percentage of specimens identified to each taxon rank in the dataset published in 2014 is shown in Figure 3.4 A. Thanks to the contribution of specialists, there was a significant increase in the percentage of specimens classified to the Species level (5.4% to 26.9%), along with a decrease in the percentage of specimens classified to the Order level (83.6% to 62.4%), meaning the taxonomic characterization of the collection is more complete in the current dataset.

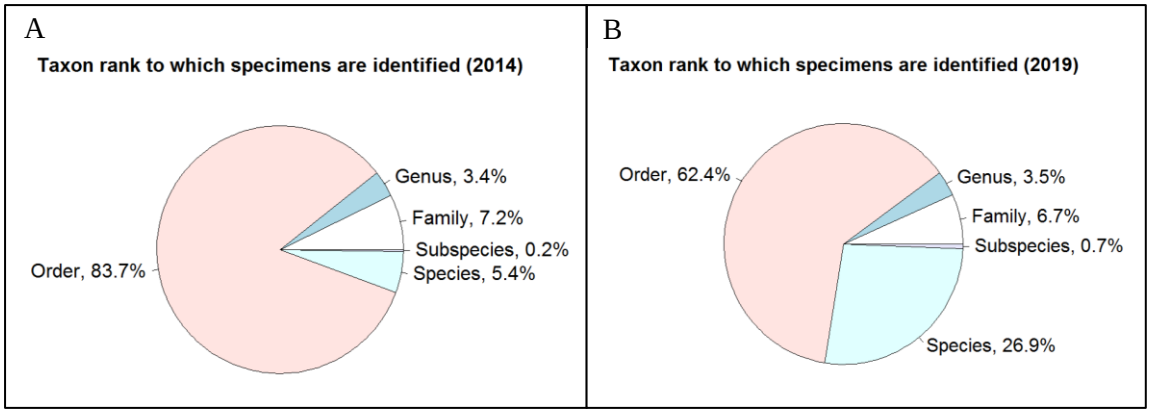

Figure 3.4 Percentage of specimens of the insect collection catalogue identified to each taxon rank, in the dataset published on GBIF in 2014 (A) and in the dataset published in 2019 (B). In (A), Class is omitted for clarity, accounting for 0.4% of specimens.

Figure 3.5 illustrates the taxonomic coverage of the insect collection, represented as the number of specimens contained in the collection for the most well represented Orders (over 100 specimens). Diptera is the most represented Order in the collection, with 23 270 specimens (29.1% of total), followed by Coleoptera (16 886 specimens, 21.1% of total) and Hemiptera (14 327 specimens, 17.9% of total). Specimens belonging to the Class Entognatha, which includes the Orders Collembola, Diplura and Protura, are included in the insect collection, even though they do not belong to the Class Insecta. In fact the MNHNC collection includes specimens from the subphylum Hexapoda, which comprises both classes.

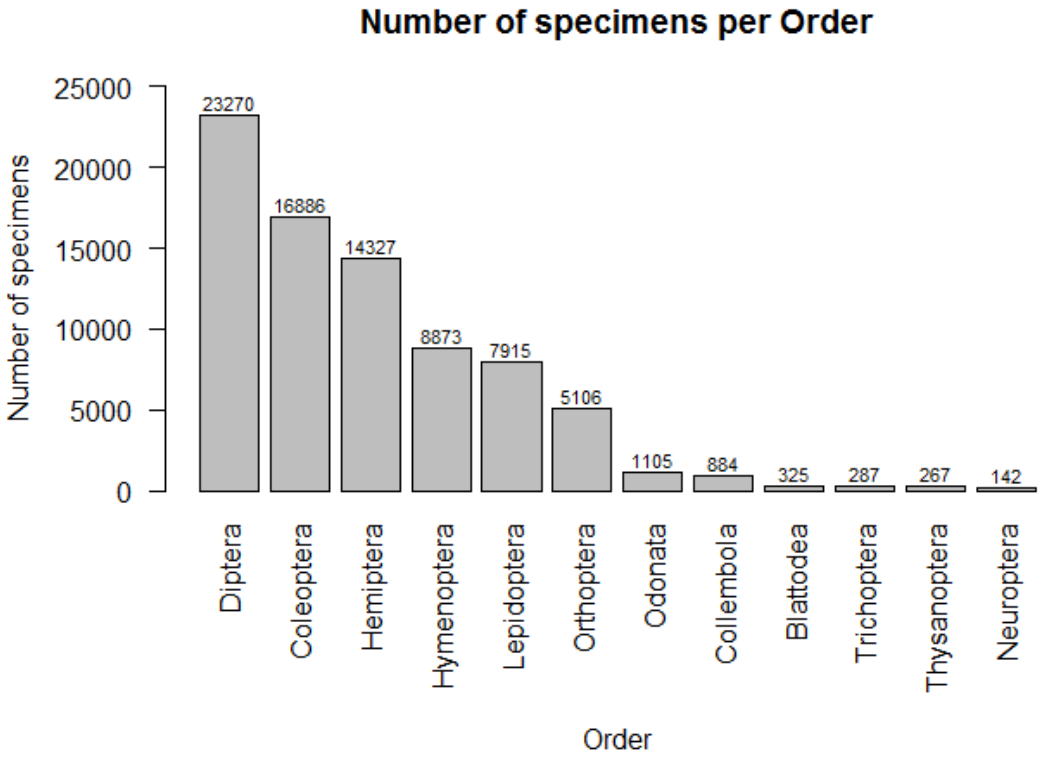

Figure 3.5 Bar graph of specimens of the MNHNC insect collection by Order. The plot includes the Orders represented by more than 100 specimens in the collection.

The number of specimens of each Family, for the Families represented by over 200 specimens in the collection, is shown in Figure 3.6. The most common Families in the collection are Drosophilidae (Order

Diptera; 9985 specimens, 12.5% of total), Nymphalidae (Order Lepidoptera; 2166 specimens, 2.7%), Chrysomelidae (Order Coleoptera; 1533 specimens, 1.9%), Chironomidae (Order Diptera; 1453 specimens, 1.8%) and Pieridae (Order Lepidoptera; 1016 specimens, 1.3%).

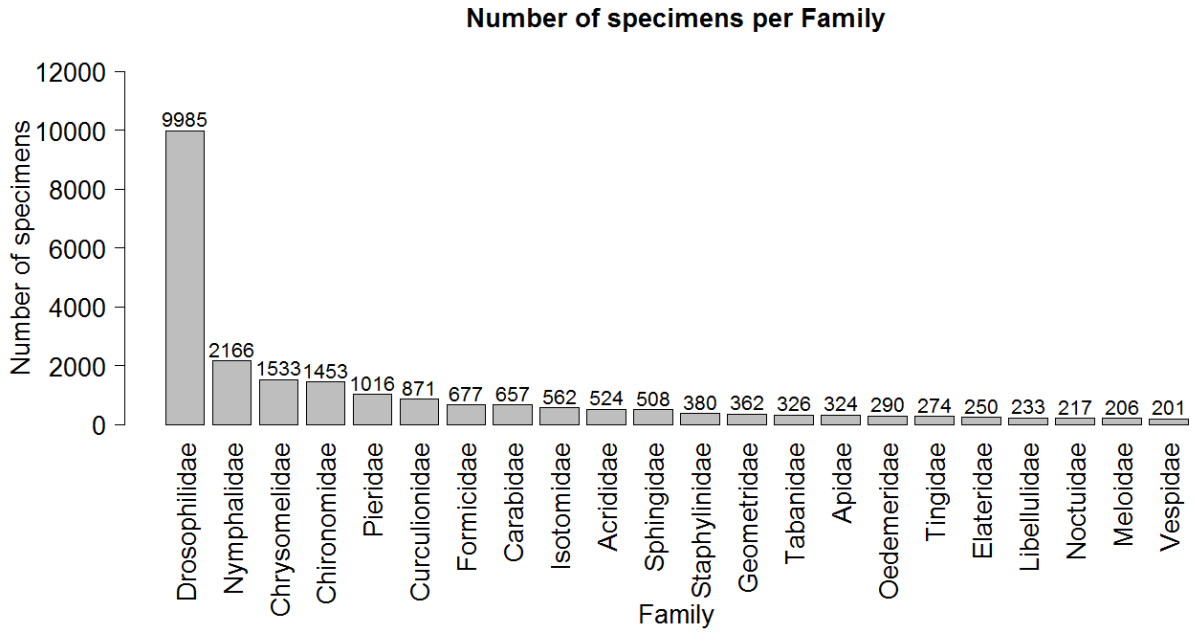

Figure 3.6 Bar graph of specimens of the MNHNC insect collection by Family. The plot includes the Families represented by more than 200 specimens in the collection.

The MNHNC collection includes a total of 67 type specimens representing 42 species. Of these, 32 are holotypes, 5 are allotypes and 45 are paratypes.

3.4. Discussion

The main contribution of this work has been the overall quality improvement of the insect collection catalogue, as new data have been produced, such as many new geocoded records, improvement of existing data quality by data standardization and the publishing on GBIF of the improved dataset. In the current collection catalogue, all dates and times are standardized to the DarwinCore format. Fields such as collector name, sampling locality, life stage and preparation method have been standardized to a controlled vocabulary.

The new version of the dataset is more complete, both in terms of number of specimens (64 008 in 2014 vs. 79 885 in the current dataset) and in terms of geographic data available (25.9% of records geocoded in 2014 vs. 70.9% in the current dataset). During this work, taxonomic identifications were also verified and updated in cases of synonymy, meaning this data is more accurate. The use of specific tools for data cleaning, such as R and OpenRefine, allowed the data cleaning process to be done quickly and efficiently. After the dataset was complete, these tools were also used to visualize the data in terms of temporal, geographic and taxonomic coverage.

It is important to note that some of the donated collections already stored in the museum include more specimens, but a significant part of the data associated with them hasn't been digitized or included in the catalogue yet, and so it wasn't accounted for in the results presented here. Of these collections, the number of specimens yet to be digitized is estimated to be almost as much as the total number of specimens currently in the collection catalogue. The process of adding specimen data to the collection catalogue may be accomplished more quickly with the help of volunteers through a citizen science project, discussed in Section 4 of this work. There are also many specimens and samples in the museum that haven't been screened, prepared or digitized yet. An important future work will be to continue these tasks as it is calculated that over 50% of the specimens in the MNHNC insect collection remain to be catalogued (L. F. Lopes personal communication). This will be an important contribution to biodiversity knowledge, making a significantly larger and more complete dataset available that can be used for research projects in different areas, such as species distribution modelling and studies of invasive species. As previously mentioned, NHC can represent species that are no longer possible to collect, for instance due to population reduction or extinction [2]. Therefore, it is, important to have these data, and preserved specimens, available for use when studying changes in biodiversity over time.

The MNHNC collection includes type specimens, which are of particular importance because they were used to describe a species, and therefore can be used as a comparison to identify other specimens. An important future work will be to exhaustively verify the data for these specimens in the dataset and document the type specimens in the collection.

4. Zooniverse project for data digitization

4.1. Introduction

Citizen science, consisting of involving the general public in research, has been used for centuries in fields such as astronomy and meteorology. However, it has become more widespread in the last decades, with the development of online tools that facilitate the participation of volunteers. Online citizen science projects are most helpful for analysing or interpreting large amounts of data that require detection of patterns and anomalies unrecognized by computers [40]. They have been successfully used for purposes as diverse as designing self-assembling RNA molecules [41] and transcribing historical documents [42]. Currently, the most used tool for developing citizen science projects is Zooniverse (www.zooniverse.org), which has 1.7 million registered volunteers worldwide. It includes a free project builder platform which allows researchers to create projects quickly and at no cost [40].

One of the most time-consuming steps of NHC data digitization is the transcription of specimen data. Sampling information for most specimens is registered in labels, index cards and/or field notebooks, requiring transcription to a digital database. Several methods can be used to achieve this. One is having associated staff or volunteers transcribing each record individually from labels or paper records, which is time-consuming. An alternative is to use optical character recognition (OCR) software to transcribe data from photographs or scanned documents. This has proven useful for typewritten labels, but it is still not efficient for handwritten text [43, 44], which is the most common in NHC labels. Moreover, OCR can't interpret or infer information, e.g. from abbreviations, and it doesn't categorize the text into separate fields for collector, sampling location, sampling date, etc.

Citizen science projects have proven very useful for NHC specimen data transcription. They can be hosted on accessible online platforms where volunteers can transcribe data remotely from digitized representations of the labels or other paper records. Several institutions have developed platforms specifically for this purpose, such as DigiVol, developed by the Australian Museum and the Atlas of Living Australia [45], Les Herbonautes, of the French National Museum of Natural History [46], Notes from Nature, developed by the Natural History Museum in London, the Southeast Regional Network of Expertise and Collections organization, Calbug and the University of Colorado Museum [44], and the Smithsonian Transcription Center [47].

Besides being useful to reduce the time necessary to digitize NHC data, these platforms are also important tools for engaging scientific communication and education, since the volunteers learn more about the subjects of the projects as they participate. Citizen science projects involving data collection, such as species monitoring, increase the participant's knowledge in the field [48]. For example, volunteers gathering data to monitor invasive plant species, showed increased awareness of invasive plants and their effects on ecosystems [49]. Data processing projects, such as transcription or image classification, are thought to raise public awareness to previously unknown fields of scientific research [48].

When provided with clear instructions or training, e.g. a tutorial on the structure of the records, volunteers can separate the different elements in the label (e.g. sampling date, location, the name of the collector, and taxonomic identification) in a way that cannot yet be done automatically. When the volunteers are fluent in the language the labels are written in, they can correct spelling errors and update location names that have changed since the sampling took place, given that information is easily available. Citizen science projects can also be used to enrich the information associated with the specimens, e.g. regarding damage to the specimen or morphological features, and even taxonomic determination data can be acquired from volunteers with expertise on the taxonomy of the target group [24].

One of the problems of volunteer work, either locally or through citizen science platforms, is the enhanced possibility of error, especially if volunteers are not familiar with the data. In fact, the resulting datasets may have errors or inconsistencies which need to be corrected, and data must be validated. Two main strategies have been proposed to deal with this. One is to have one volunteer transcribing the data and another validating and/or correcting it, another is to allow transcription of the same data by multiple volunteers independently, and then reconcile the data – if all fields match, the record is considered correct; if there are differences between transcriptions, the record is marked for revision by museum staff [24, 44].

4.1.1. Objectives

The objective of this part of the work was to create and optimize a citizen science project on the Zooniverse platform, to aid in specimen label transcription and taxonomic identification of the MNHNC insect collection. Once implemented the project will continue, allowing volunteers to participate in the digitization process and learn more about the MNHNC insect collection.

4.2. Methods

To develop an effective citizen science project, a set of photographs (N = 130 subjects/specimens) was used with the objective of testing the project and evaluating the participation by volunteers and the accuracy of their contributions, in order to optimize the project structure (workflow organization and supporting information) and define the best method to process and validate the contributions. These preliminary analyses provided useful information to develop the final project to be published on the Zooniverse platform.

A citizen science project with the title "MB-07 - The Insects of the Museu Nacional de História Natural e da Ciência" was created on the Zoniverse platform (www.zooniverse.org), and was made available for beta testing (https://www.zooniverse.org/projects/lfilipevsl/mb07-the-insects-of-the-museu-nacional-de-historia-natural-e-da-ciencia). The project included two workflows designed for different audiences. One was delineated for the general public and is a simple text transcription activity (hereinafter referred to as transcription workflow). The other was conceived for expert volunteers on insect taxonomy able to contribute with accurate taxonomic information (hereinafter referred to as taxonomic workflow). Both workflows guide the volunteers through a series of tasks that need to be completed. These tasks are of two types: 1) multiple choice questions; 2) free text input.

As part of all projects created on Zooniverse, a workflow is considered to be a set of tasks that the volunteer is asked to complete. Each workflow is applied to a set of subjects (in this work, the subjects were photographs of specimens). A task is a single step, e.g. to transcribe the sampling date in a label. When a volunteer completes all tasks in a workflow, he/she is asked to review the responses. The set of responses provided by a volunteer for each subject, is termed as a classification. The set of classifications can be downloaded as a csv file by the creators of the project.

The tasks for the transcription workflow are shown in the flowchart below (Figure 4.1).

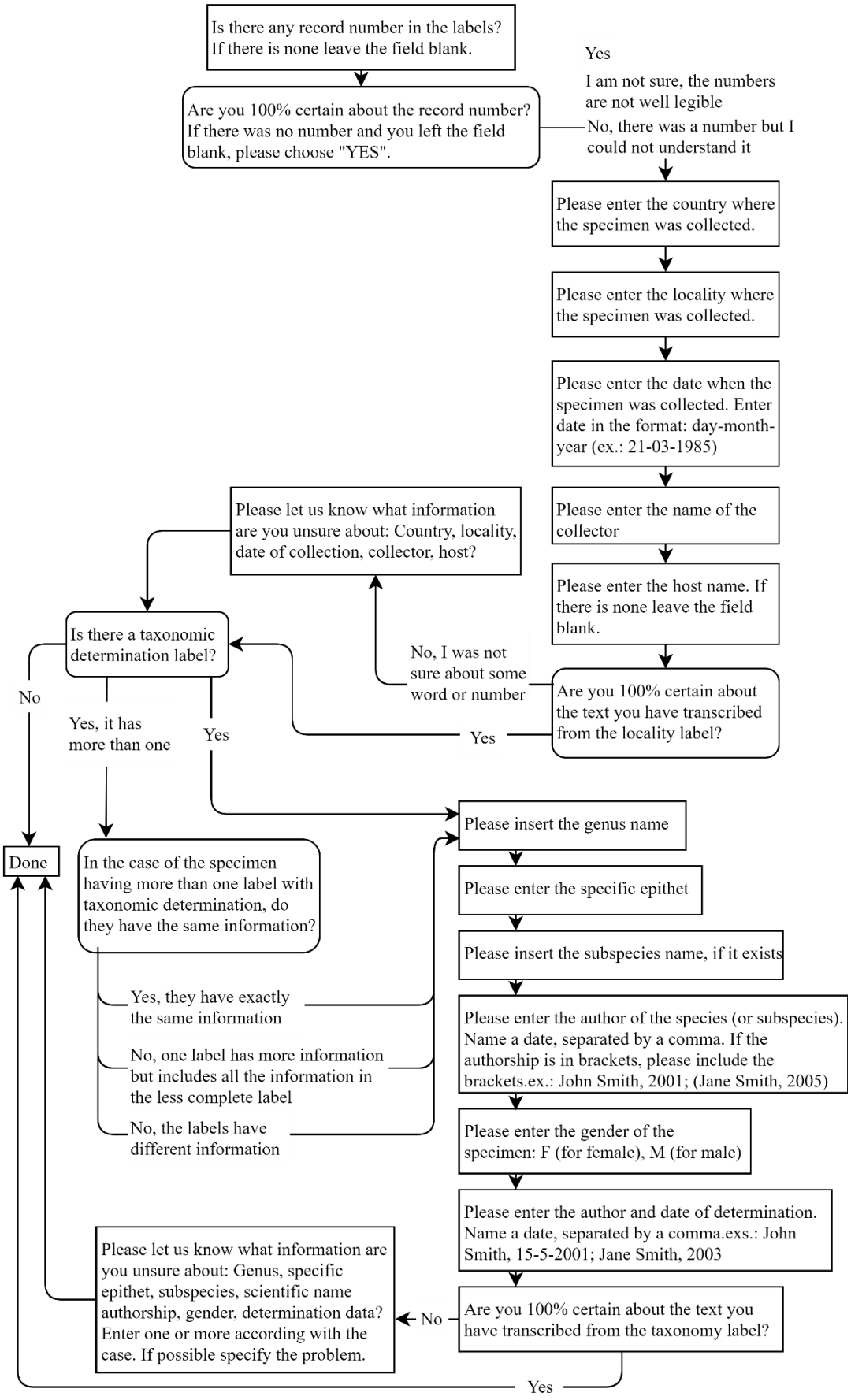

Figure 4.1 Flowchart representing the tasks in the transcription workflow of the project developed on the Zooniverse platform. Rectangles represent tasks where the user is prompted for text input, squared rectangles represent multiple answer questions.

The tasks for the taxonomic workflow are shown in the next flowchart (Figure 4.2).

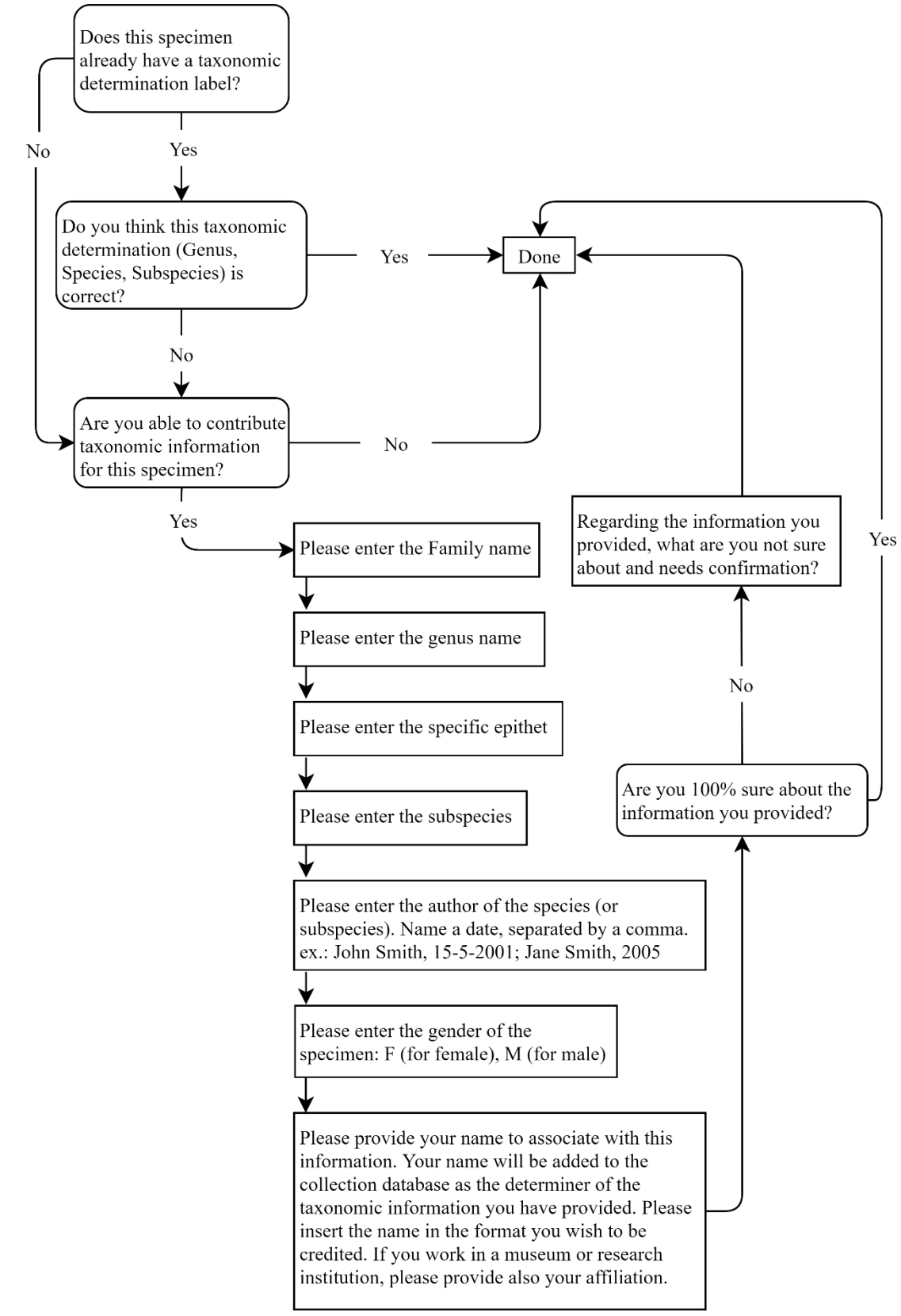

Figure 4.2 Flowchart representing the tasks in the taxonomic identification workflow of the project developed on the Zooniverse platform. Rectangles represent tasks where the user is prompted for text input, squared rectangles represent multiple answer questions.

A set of 130 photos of moth specimens (Order Lepidoptera, Family Sphingidae) of the collection donated by José Passos de Carvalho were included in the project, and were made available for both workflows (Figure 4.1, Figure 4.2). Each photograph depicts a specimen, a scale bar and all the specimen label(s) associated with it.

An example of a task (a step in the workflow), presented to volunteers as part of the transcription workflow is shown in Figure 4.3. The instructions, included in a "Tutorial" section, and help text were provided both in English and Portuguese, to allow volunteers to choose the language they were most comfortable with and enabling Portuguese and English speakers to participate in the project. Each subject was withdrawn after the submission of 5 different classifications from different users.

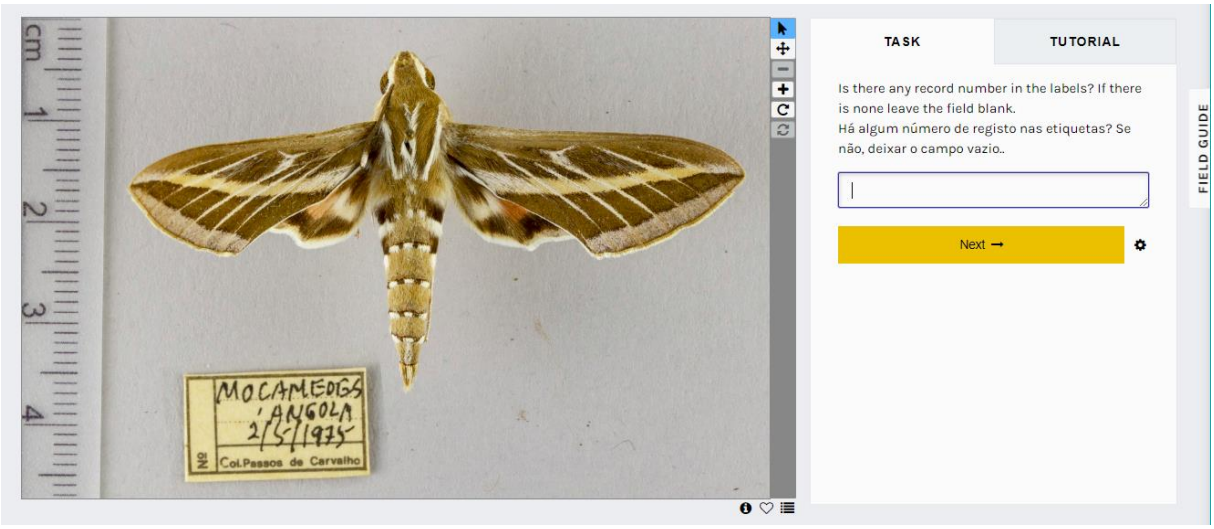

Figure 4.3 Image of the first task that volunteers were asked to complete in the transcription workflow of the project developed on the Zooniverse platform.

The classification data was downloaded as a set of CSV files on May 16, 2019. The classification file contained data processed by volunteers, the workflow data file contained all the tasks included in each workflow, and the subject data file contained a list of subjects, with the filename of the photograph and the corresponding Zooniverse ID.

Panoptes_aggregation, a software created specifically for extracting data from the csv files exported from Zooniverse [50], was used to extract the data from these files. As a first step, two extractor configurator files were created, one for each workflow. The extractor configurator files listed all tasks in the workflow, including the type of each task, so that the responses could be extracted correctly. The files were created using the panoptes_aggregation graphical user interface (GUI), using the workflow data csv file as input, and then edited manually to assign the appropriate task type to each task. These files were used to extract data from the classification data file, also using the panoptes_aggregation GUI. This resulted in two csv files with all classifications made by volunteers.

An R script was written to format these files, creating a csv file for each workflow, organized by specimen, with classification information by each volunteer for a single subject in one row, with one column for each task. This script is available in Annex B.

The final csv files were used to evaluate volunteer classification data. Each was assigned one of three values: correct (all answers to tasks correspond to the data in the subject photograph); incorrect (with one or more incorrect values); or blank (only blank values). These values were then counted to evaluate the number of correct, incorrect and blank responses for each workflow.

4.3. Results

Although the project is still in a testing phase and it is not yet available on the main page of Zooniverse, there was a call, from the platform, for beta testers. Furthermore, it was advertised through the MNHNC social media and has received contributions by volunteers. In May 16, 2019, data was extracted from the project to evaluate participation and the quality of contributions.

For the transcription project, a total of 582 classifications were made by 104 individual volunteers, with an average of 5.7 classifications per volunteer. One classification corresponds to a volunteer completing all tasks for a single subject photograph. Of these classifications, 78 (13.4%) contained only blank values, 129 (22.2%) contained incorrect transcriptions, and 375 (64.4%) contained correct transcriptions. Classifications were considered correct when all responses to tasks corresponded to the data in the label of the subject. For each subject, there was an average of 4.5 classifications. On average, 2.9 of them were correct, 1.0 was incorrect and 0.6 were blank values.

The number of classifications made per day is shown in Figure 4.4.

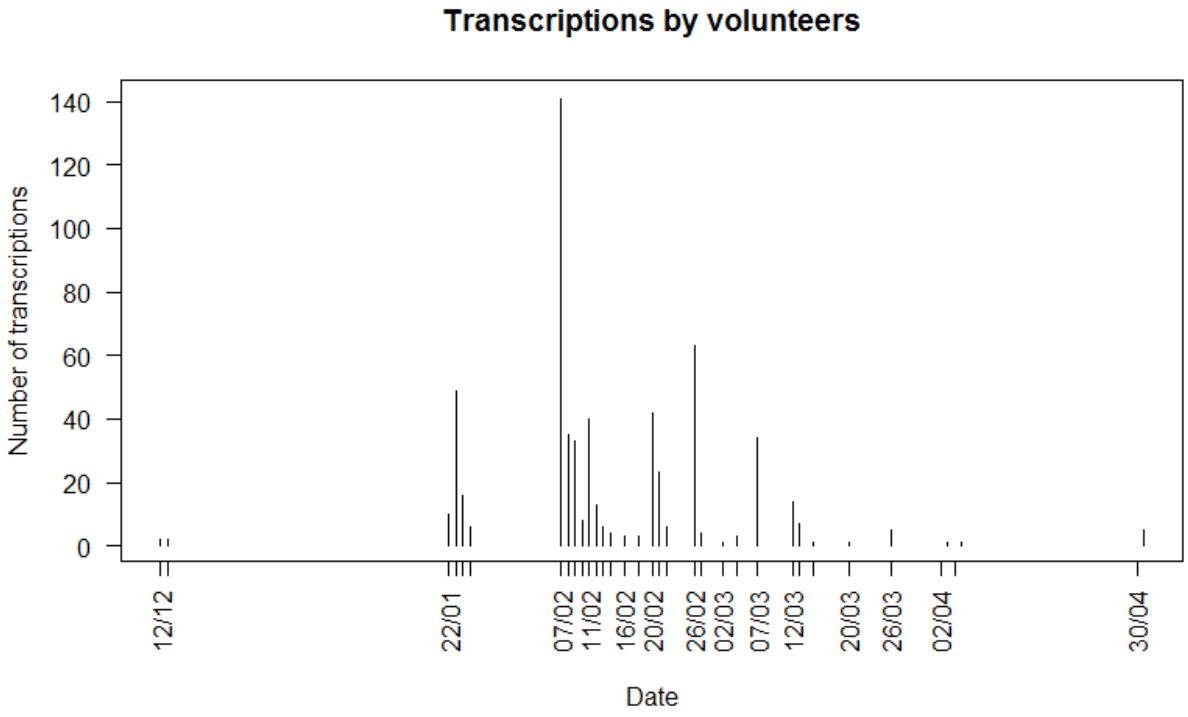

Figure 4.4 Transcriptions made by volunteers per day, between December 2018 and April 2019. Each transcription corresponds to completing all the tasks by one volunteer. Each image is transcribed by more than one volunteer.

In the taxonomic identification workflow, 194 classifications were made by 13 volunteers. In order to evaluate the classifications, values were considered expected when they matched the corresponding fields, since only an expert can verify if taxonomic identifications are correct or not. Of these

classifications, 65 (33.5%) were blank, and 129 (66.5%) had expected values. None of the classifications had unexpected values. The number of classifications made per day is shown in Figure 4.5.

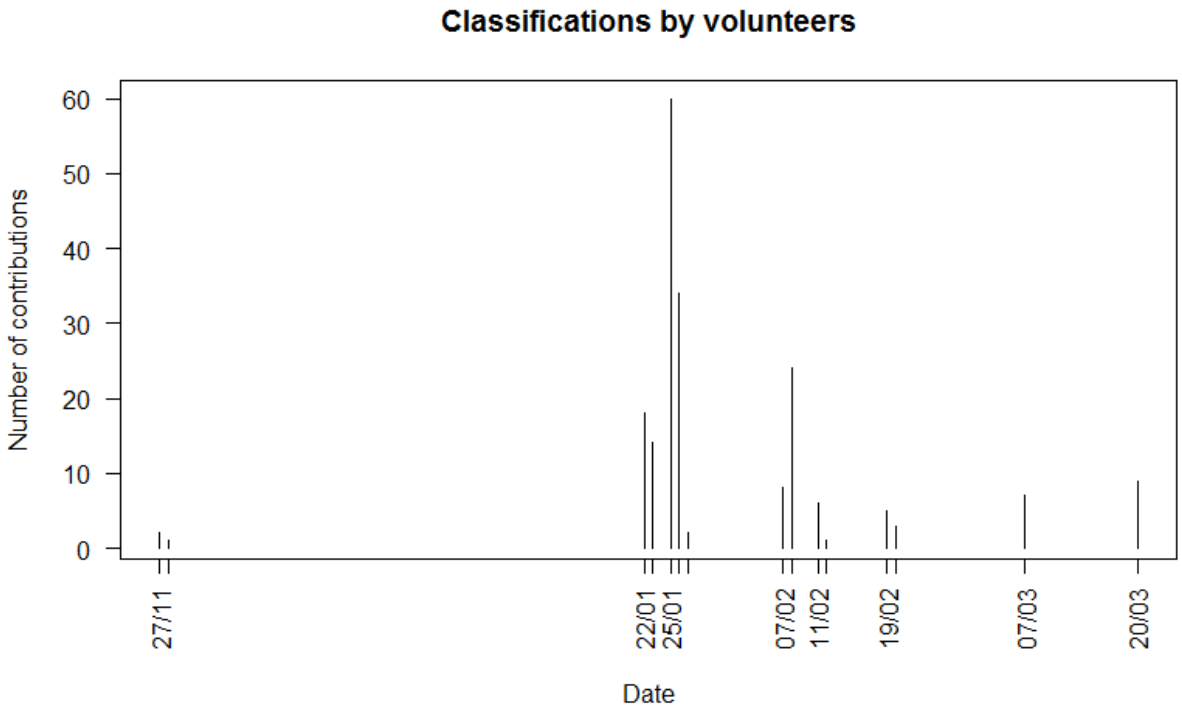

Figure 4.5 Contributions to the taxonomic workflow made by volunteers per day, between December 2018 and March 2019. Each contribution corresponds to either a transcription and confirmation of taxonomic identification labels in one specimen, or a new taxonomic identification of a specimen, done by a volunteer. Each image can be classified by more than one volunteer.

For each subject, there was on average 1 classification containing expected data, and 0.5 classifications with only blank values.

Regarding the taxonomic identification of specimens, a total of 61 identifications were provided for specimens with no taxonomic identification. The remaining 69 specimens already had taxonomic identifications in the labels; the identifications for those specimens were verified.

As a part of the Zooniverse testing process, volunteers who made classifications were asked by the platform to fill out a feedback survey. This allows the Zooniverse staff and the project creators to evaluate volunteer interest in the project, and the likelihood of volunteers participating once it is advertised on the projects section of Zooniverse (www.zooniverse.org/projects). The answers to this survey were not separated by workflow, thus they are assumed to apply to both workflows.

The questions in the survey were mostly intended to evaluate the difficulty of the project, if the help and tutorials provided were suitable, and the interest of volunteers in the project. A total of 32 volunteers answered the survey. Regarding difficulty of the tasks, 34.4% of volunteers found the tasks moderately easy, 9.4% found them very easy, 43.8% found them somewhat hard, and 9.4% found them very hard (Figure 4.6). The main reasons pointed out by volunteers who found the tasks somewhat hard or very hard were difficulty in reading the labels, and not being sure of which term corresponded to which field (especially for taxonomic identifications, and for volunteers who were not familiar with binomial nomenclature).

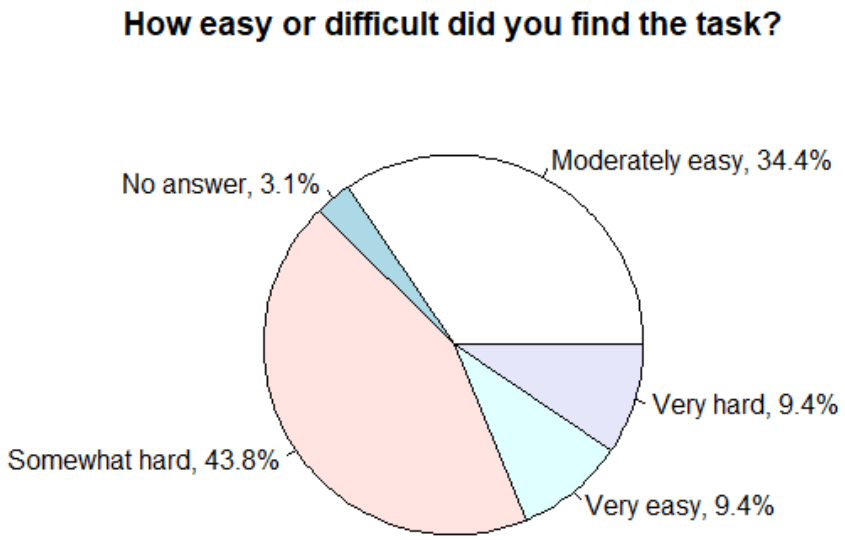

Figure 4.6 Answers by volunteers to the question "How easy or difficult did you find the task?" in the feedback survey for the Zooniverse project. The survey was filled out by 32 volunteers who contributed with a classification.

The volunteers were asked whether or not the project was suitable for Zooniverse (Figure 4.7). To this question, 28 of the participants (87.5%) answered "Yes", 1 participant (3.1%) answered "No", and 3 volunteers (9.4%) provided no response.

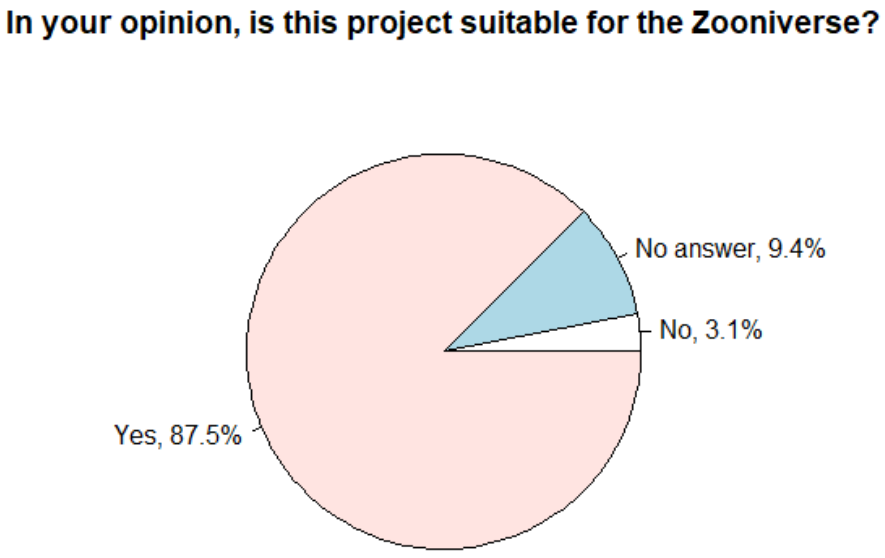

Figure 4.7 Answers by volunteers to the question "In your opinion, is this project suitable for the Zooniverse?" in the feedback survey for the Zooniverse project. The survey was filled out by 32 volunteers who contributed with a classification.

The final question in the survey evaluated the willingness of volunteers to participate in the project once it becomes an official Zooniverse project and it is available through the website homepage (Figure 4.8). To this question, 10 volunteers (31.2%) replied "Yes" and 4 (12.5%) replied "Yes and I'll bring friends!".

Of the remaining answers, 13 volunteers (40.6%) replied “Not sure”. Only 3 of the volunteers (9.4%) answered “No”, and 2 participants (6.2%) provided no answer.

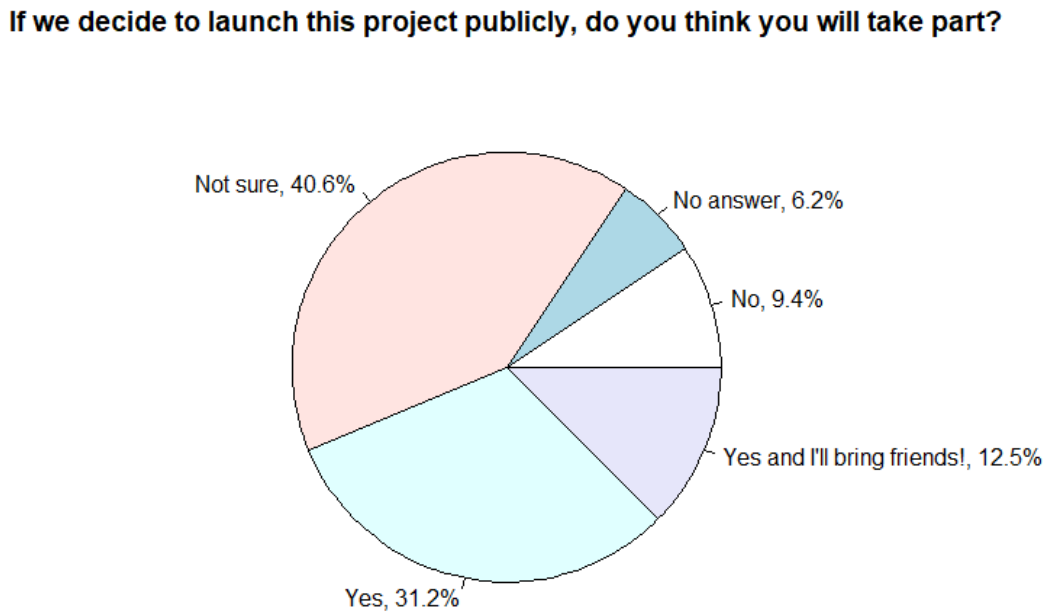

Figure 4.8 Answers by volunteers to the question “If we decide to launch this project publicly, do you think you will take part?” in the feedback survey for the Zooniverse project. The survey was filled out by 32 volunteers who contributed with a classification.

4.4. Discussion

Citizen science has been shown to be an effective way of speeding up the digitization of biodiversity data labels and field notebooks. It has also been used to geocode sampling locations and annotate images of specimens [24]. The objective of this work was to evaluate if it would be a suitable tool to speed up the process of specimen data digitization.

The transcription workflow yielded an average of 2.9 correct classifications per subject, versus 1 incorrect classification. Based on this result, an error correction method can be proposed, in which the majority of similar answers to each task is accepted as the correct value. Another possible verification method is to define some volunteers as advanced users, who can verify others´ classifications. It is also expected that, as the project continues, volunteers will become more experienced and provide more reliable classifications, since most classifications are made by a small number of volunteers – in a study of seven different citizen science projects, it was concluded that the top 10% of contributors were responsible for 71% to 88% of classifications [51].

For the taxonomic identification workflow, 61 new identifications were provided for the total of 130 subjects included. All but one of the new taxonomic identifications were provided by the curator of the collection. This demonstrates that the Zooniverse project is useful not only to allow participation by volunteers, but also to facilitate the work of museum staff. After the specimens are photographed and uploaded as subjects, both volunteers and museum staff working in the digitization of collections can contribute with transcriptions and taxonomic identifications.

In the future, this project will be used to expedite the MNHNC digitization process for specimens that have labels but weren't included in the collection catalogue yet, such as the remaining specimens of the Passos de Carvalho collection. It will first be necessary to photograph the specimens of the collection with the accompanying labels; after that, the photographs can be uploaded to the website, where more classifications can be done by volunteers. Currently, the project is under review, meaning it is not an official Zooniverse project yet; it is expected that, once the project becomes an official Zooniverse project, there will be more participation by Zooniverse volunteers, as it will be featured on the Zooniverse Projects Page (www.zooniverse.org/projects), where 1.7 million volunteers registered in Zooniverse will be able to access it and make their contribution [52]. Furthermore, the museum should also use its own communication channels to disseminate the project.

Results of the feedback survey suggest a positive response of volunteers to the project. Although it is not possible to distinguish between the two workflows in the volunteers' responses, it can be hypothesized that the volunteers who considered the project very hard were mainly referring to the taxonomic identification workflow, which is not intended for the general public. There is also an added level of difficulty for volunteers who are not Portuguese speakers, since most of the labels available to transcribe are in Portuguese and refer to sampling locations and collectors from Portugal or Portuguese-speaking countries.

Citizen science projects similar to the transcription workflow created here have achieved astounding results; for example, in February 2015, the Atlas of Living Australia's DigiVol had obtained 130 816 transcriptions from 860 contributors, and Zooniverse's Notes from Nature had 1 042 592 transcriptions from 6 833 contributors [24]. Both these projects used a reward system to encourage volunteer participation, with badges attributed to volunteers that make the most contributions, or upon reaching a certain number of contributions [24, 44].

The resulting data will need further processing in order to be included in the database, mostly for error correction and converting data to the DarwinCore format. However, even with this required data

cleaning step, it will make the digitization process faster, contributing to the MNHNC insect collection data being digitized, accessible and searchable.

As an additional benefit, this citizen science project will give more visibility to the MNHNC insect collection, and it will be a way to increase interest in entomology and natural history among the general public, as citizen science is not only a tool for research, but also for education and science dissemination [24].

5. Tabanid collection data digitization

5.1. Introduction

Tabanid flies (order Diptera, family Tabanidae), which include the commonly called horse and deer flies, are vectors of mechanical and biological infectious diseases transmitted to animals. Examples of disease agents transmitted to livestock and other animals include protozoan parasites, such as *Trypanosoma theileri* [53] and *Trypanosoma vivax* [54], the haemoplasma *Mycoplasma (E.) wenyonii* [55] and viruses such as the *Equine infectious anemia virus* and *Indiana vesiculovirus* [56]. Tabanids can also carry agents that transmit infectious diseases to humans, such as the nematode *Loa loa* (cyclically transmitted by *Chrysops silaceus* and *C. dimidiatus*) [57], the bacteria causing anthrax (*Bacillus anthracis*) [56] and tularemia (*Francisella tularensis*) [54]. Cases of anaphylaxis as a result of tabanid bites have been also reported [58].

There are currently 4300 described species of tabanids, comprised in 133 genera [54]. Female tabanids of most species take a blood meal during egg development, while males usually feed exclusively on nectar [59]. For most species, females are more active than males, especially in bright sunlight and during the spring and summer [54]. Tabanids are widely distributed worldwide and are present in a wide range of habitats and climates [60]. Therefore, acquiring scientific knowledge on tabanid species and their distribution is of special interest for disease control and livestock protection [61]. Studies have been conducted to characterize the distribution of Tabanids in different locations [62, 63] and to create checklists of Tabanidae in NHC [64], but more information is necessary.

The entomological collections of the *Instituto de Investigação Científica Tropical* (IICT) include a collection of tabanid flies, compiled and classified by the researcher J. A. Travassos Santos Dias, but its data had not yet been digitized, and its reduced visibility hampered its use in scientific studies. This collection includes specimens collected in the decades of 1920-1990, mainly in Portugal, Spain, Mozambique, Angola and São Tomé and Príncipe, among other countries worldwide. It includes 1051 specimens. In addition to this collection, other tabanids were collected more recently by Luis Mendes and other researchers, which are also stored at the IICT. The IICT collections have recently become part of the *University of Lisbon* patrimony, and are under the care of the MNHNC, which also holds tabanid specimens in its insect collection. All of these specimens were included in the dataset developed during this work. Most of these specimens were determined to the species level by Travassos Santos Dias and other specialists. Some of the specimens correspond to species described by Travassos Santos Dias himself, and both collections include type specimens.

5.1.1. Objectives

The objective of this part of the work was to digitize and analyse the data of the IICT and MNHNC tabanid collections, including several steps:

1. Photographing all specimens in the collection and associated labels;
2. Transcribing label data to create a dataset;
3. Geocoding the collection locations;
4. Publishing the resulting dataset on GBIF;
5. Analysing the dataset in terms of geographic, temporal and taxonomic coverage, by comparing it to existing data published on GBIF and in the literature, and by producing distribution maps for the more well-represented species.

Studying this collection will increase the existing knowledge about this group, especially regarding its geographical distribution through time, as the dataset has an extensive temporal and geographic coverage.

5.2. Methods

As a first step, all the specimens in the collections that hadn't been digitized yet were photographed with the accompanying labels and a scale. Specimens were photographed using a Canon EOS 7D Camera with a macro lens. Photographs were then edited using Adobe Photoshop CS5 version 12.0 to add the complete species name and to edit white balance, color levels, contrast and brightness. An example of an edited photograph is shown in Figure 5.1.

A taxonomic review was done, using the GBIF database as reference, to verify the current validity of the taxonomic classifications and to add synonyms for species where necessary. All the photographs were sent to a tabanid taxonomy specialist, Hélcio Gil, of *Instituto Oswaldo Cruz*, to confirm the taxonomic identifications and further cases of synonymy.

Figure 5.1 - Example of a photograph of a specimen of the Tabanid collection, with collection and classification labels.

Data was compiled in an Excel spreadsheet, by transcribing information contained in the labels for each individual specimen. These data include the collector, sampling date and location, taxonomic classification (updated if necessary, but maintaining a record of previous data) and sex. Some specimens included other additional information, such as previous collection number or the host from which the specimen was collected, which was also transcribed and included in the dataset.

After compiling all data, records were geocoded using the same method used in section 3.2.

In order to evaluate the geographic, temporal and taxonomic coverage of this dataset, graphs were produced using R [26]. A distribution map for the countries represented in the collection and another for all specimens sampled in Portugal, the most represented country, were produced using the tmap package [65]. Distribution maps for the species represented by more than 30 specimens in the collections with geocoded sampling locations were produced using the mapview package [66].

5.3. Results

The dataset contains a total of 1666 records, each corresponding to one pinned specimen. Taxonomic coverage of the dataset is very complete, with 1415 specimens (84.9% of the total) classified to the Species level. Of the remaining specimens, 179 (10.7%) are classified to the Family, 68 (4.1%) to the Genus and 4 (0.2%) to the Subspecies (Figure 5.2). 1065 specimens were identified by Travassos Santos Dias, and 377 are of Species described by Travassos Santos Dias.

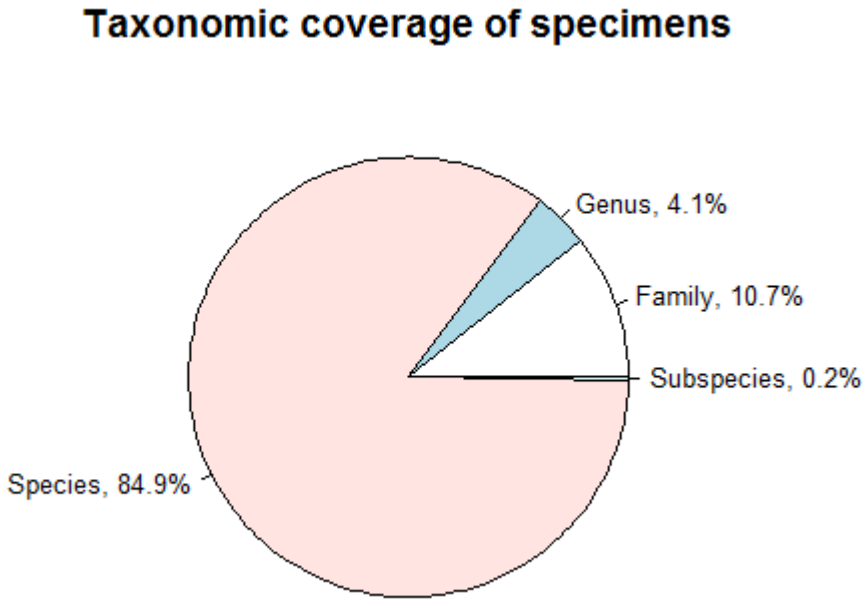

Figure 5.2 Percentage of tabanid specimens in the IICT/MNHNC collections identified to each taxon rank.

The most represented Genus is *Tabanus* (956 specimens), followed by *Haematopota* (241), *Chrysops* (52), *Atylotus* (47) and *Pangonius* (46). The most represented species are *Tabanus monocallosus* Travassos Dias, 1955 (260 specimens), *Tabanus eggeri* Schiner, 1868 (120), *Haematopota italica* Meigen, 1804 (84), *Tabanus autumnalis* Linnaeus, 1760 (68), *Tabanus bromius* Linnaeus, 1758 (43) and *Tabanus sudeticus* Zeller, 1842 (43). Figure 5.3 shows the Genera represented by 10 or more specimens in the collections; the species represented by 20 or more specimens in the collections are shown in Figure 5.4.

In what concerns gender, 1177 specimens (70.6%) are females, 118 (7.1%) are males and 371 (22.3%) are of undetermined sex. The higher number of female than male specimens was expected, since females are more active and thus more likely to be sampled.

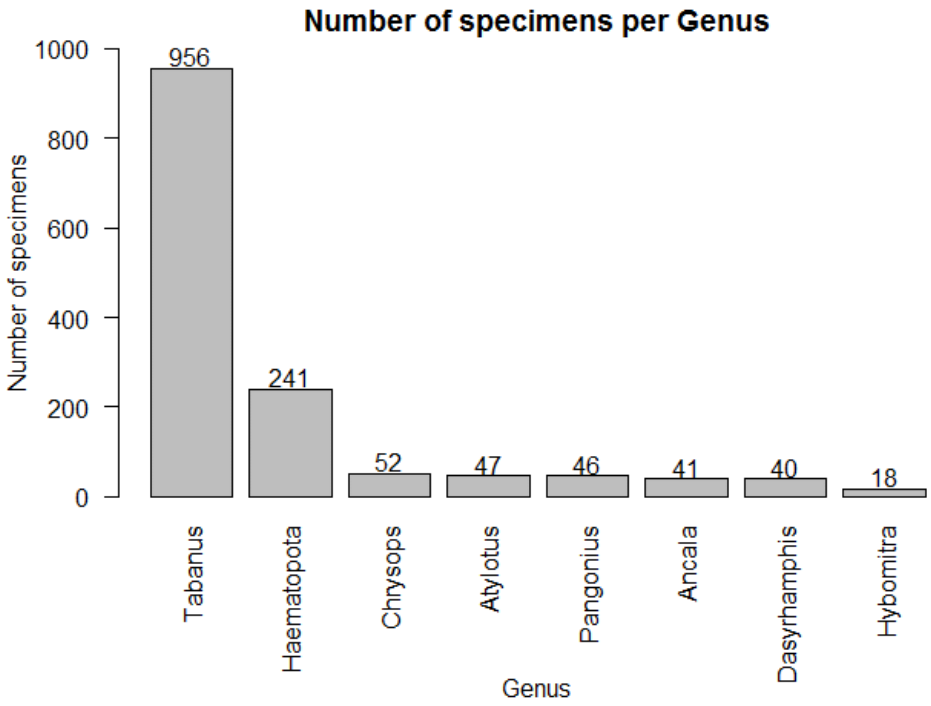

Figure 5.3 Bar graph of tabanid specimens in the IICT/MNHNC collections by Genus, for Genera represented by 10 or more specimens in the collections.

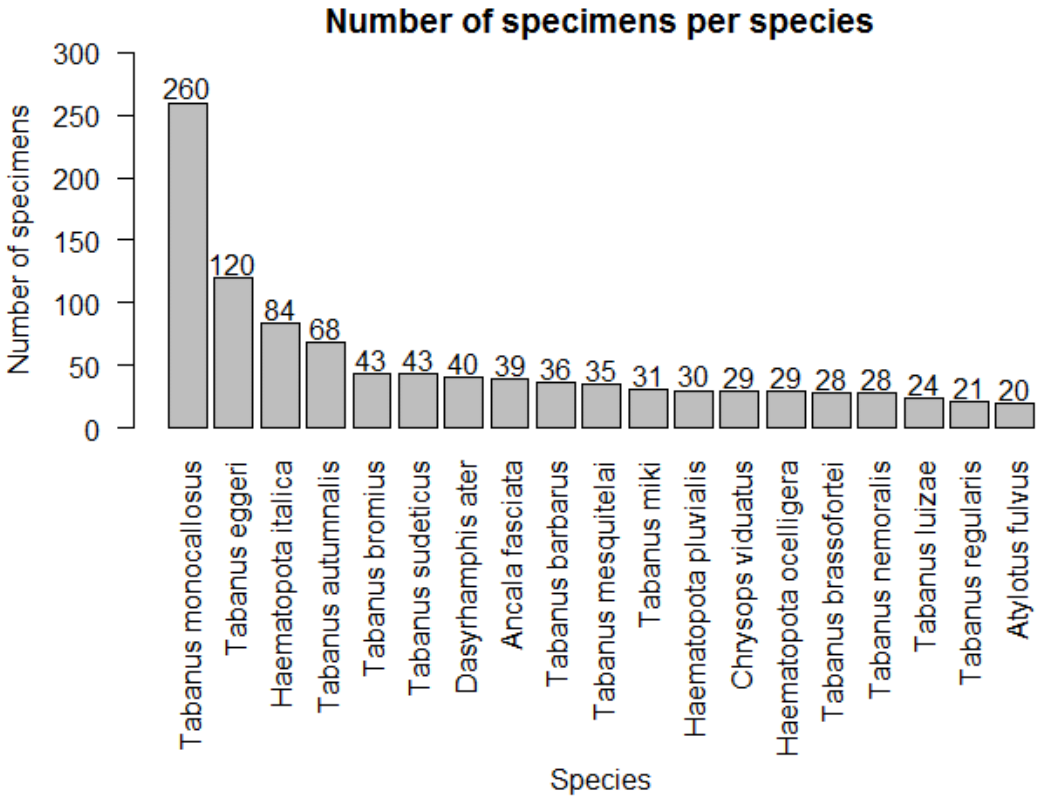

Figure 5.4 Bar graph of tabanid specimens in the IICT/MNHNC collections by species. Species represented by 20 or more specimens in the collection are shown.

The collections contain a total of 88 type specimens corresponding to 53 species described by Travassos Santos Dias, Portillo, Portillo and Schacht, and Coscarón and Fairchild. These include 41 holotypes, 1 allotype and 46 paratypes. Of the 53 species, 22 have since been considered junior synonyms of other ones. Mendes *et al.* (1988) [67] published an annotated list of all insect type specimens stored in the IICT, where the sampling location and conservation state of each type specimen were described in detail. A list of type specimens stored in the IICT collection is presented in Table 5.1, including the specimens listed by Mendes *et al.* and species described by Travassos Dias after 1988. Tabanidae type specimens stored in the MNHNC collections are also listed in

Table 5.2. For 2 of the species described by Travassos Santos Dias (*Philoliche luderitzi* and *Philoliche penrithi*), it was not possible to find the work where the species description was published; therefore, these are not included in the list of type specimens.

Table 5.1 Type specimens of the IICT tabanid collection. The number of holotypes, allotypes and paratypes in the collection is accounted for each species. Scientific name refers to the species described by the author. For the cases where the species was later considered a synonym of another the updated name is given under Current name.

Genus	Scientific name	Current name	Holotypes	Allotypes	Paratypes	Ref.
Atylotus	*Atylotus fairchildi* Travassos, 1983	*Atylotus agrestis* (Wiedemann, 1828)	1			[68]
	Atylotus olsufjevi Travassos, 1983	*Atylotus latistriatus* (Brauer, 1880)	1			[68]
Bartolomeu-diasiella	*Bartolomeudiasiella atlanticus* Travassos Santos Dias, 1987				2	[69]
Chrysops	*Chrysops piresi* Travassos, 1985	*Chrysops caecutiens* (Linnaeus, 1758)	1			[70]
Haematopota	*Haematopota amicoi* Travassos Santos Dias, 1996				1	[71]
	Haematopota angolensis Travassos Santos Dias, 1989		1			[72]
	Haematopota arrabidaensis Travassos, 1985	*Haematopota ocelligera* (Krober, 1922)	1		1	[70]
	Haematopota chongoroiensis Travassos Santos Dias, 1989		1		2	[73]
	Haematopota conninckae Dias, 1993				1	[74]
	Haematopota eugeniae Portillo & Schacht, 1984				1	[75]
	Haematopota gamae Travassos, 1991	*Haematopota lambi* Villeneuve, 1921	1			[76]
	Haematopota intrincata Travassos, 1985	*Haematopota ocelligera* (Krober, 1922)	1			[70]
	Haematopota quartaui Travassos, 1985	*Haematopota lambi* Villeneuve, 1921	1			[77]
	Haematopota ribeirorum Travassos, 1984	*Haematopota enriquei* Leclerq, 1971	1			[78]

	Haematopota salomae Travassos, 1990	*Haematopota ocelligera* (Krober, 1922)	1		[76]
	Haematopota serranoi Travassos, 1984	*Haematopota eugeniae* Portillo & Schacht, 1984	1	2	[79]
Hybomitra	*Hybomitra alegrei* Travassos, 1984	*Hybomitra tamujosoi* Schacht & Portillo, 1982	1		[80]
	Hybomitra medeirosi Travassos Santos Dias, 1989		1		[81]
	Hybomitra mendesi Travassos Santos Dias, 1989		1		[82]
	Hybomitra tamujosoi Schacht & Portillo, 1982			1	[83]
	Hybomitra zaballosi Portillo, 1991			2	[84]
Pango-nius	*Pangonius brancoi* Travassos, 1984	*Pangonius hermanni* Krober, 1921	1		[85]
Philoliche	*Philoliche dubiosa* Travassos Santos Dias, 1991			1	[86]
	Philoliche pamelae Travassos Santos Dias, 1991			2	[86]
	Philoliche penrithi			2	
Poecilo-deras	*Poeciloderas pampeanus* (Coscarón and Fairchild, 1976)			1	[87]
Tabanus	*Tabanus brancoi* Travassos Santos Dias, 1989		1		[81]
	Tabanus capelai Travassos, 1992	*Tabanus flavofemoratus* Strobl, 1908	1		[88]
	Tabanus cruzesilvai Dias, 1980	*Tabanus bromius* Linnaeus, 1758	1		[89]
	Tabanus ilharcoi Travassos, 1990	*Tabanus nemoralis* Meigen, 1820	1		[90]
	Tabanus luizae Travassos Santos Dias, 1979		1		[91]
	Tabanus mateusi Travassos Santos Dias, 1980	*Tabanus brassofortei* Travassos, 1980	1		[91]
	Tabanus mesquitelai Travassos Santos Dias, 1991		1	15	[92]

Tabanus pseudolunatus Dias, 1980		1			[91]
Tabanus pseudothoracinus Travassos Santos Dias, 1996				4	[71]
Tabanus rosarioi Dias, 1994		1	1		[93]
Tabanus rubioi Travassos, 1987	*Tabanus nemoralis* Meigen, 1820	1			[94]
Tabanus tendeiroi Travassos, 1980	*Tabanus barbarus* Coquebert, 1804	1			[89]
Tabanus varelai Dias, 1980		1			[89]

Table 5.2 Type specimens of the MNHNC tabanid collections. The number of holotypes, allotypes and paratypes in the collections is accounted for each species. Scientific name refers to the species described by the author. For the cases where the species was later considered a synonym of another the updated name is given under Current name.

Genus	Scientific name	Current name	Holotypes	Allotypes	Paratypes	Ref.
Chrysops	*Chrysops angolensis* Travassos Dias, 1974		1			[95]
	Chrysops passosi Travassos, 1980	*Chrysops caecutiens* (Linnaeus, 1758)	1		1	[89]
Haematopota	*Haematopota grandvauxi* Travassos Dias, 1973		1			[96]
	Haematopota mendossaorum Travassos Santos Dias, 1992	*Haematopota ocelligera* (Krober, 1922)	1			[97]
	Haematopota passosi Travassos Dias, 1973		1			[96]
	Haematopota teixeirai Travassos Dias, 1974		1			[95]
Hybomitra	*Hybomitra portucalensis* Travassos Santos Dias, 1985	*Hybomitra ciureai* (Séguy, 1937)	1			[98]
Tabanus	*Tabanus bivari* Travassos Santos Dias, 1985		1			[98]
	Tabanus brassofortei Travassos, 1980		1		2	[89]
	Tabanus maiombensis Travassos Dias, 1973		1		2	[96]
	Tabanus mendossai Travassos, 1992	*Philipomyia aprica* (Meigen, 1820)	1			[97]
	Tabanus mossambicensis Travassos Santos Dias, 1985		1			[99]
	Tabanus passosi Travassos Dias, 1974		1			[95]

The countries where 10 or more specimens of the IICT/MNHNC collections were sampled are shown in Figure 5.5. The majority of the specimens were collected in Portugal (938, 57.2% of total), São Tomé and Príncipe (270, 16.5%), Guinea-Bissau (96, 5.9%), Mozambique (74, 4.5%), Spain (71, 4.3%) and Angola (53, 3.2%). Geographic distribution of the specimens sampling locations is further detailed in section 5.3.1, including distribution maps for some species.

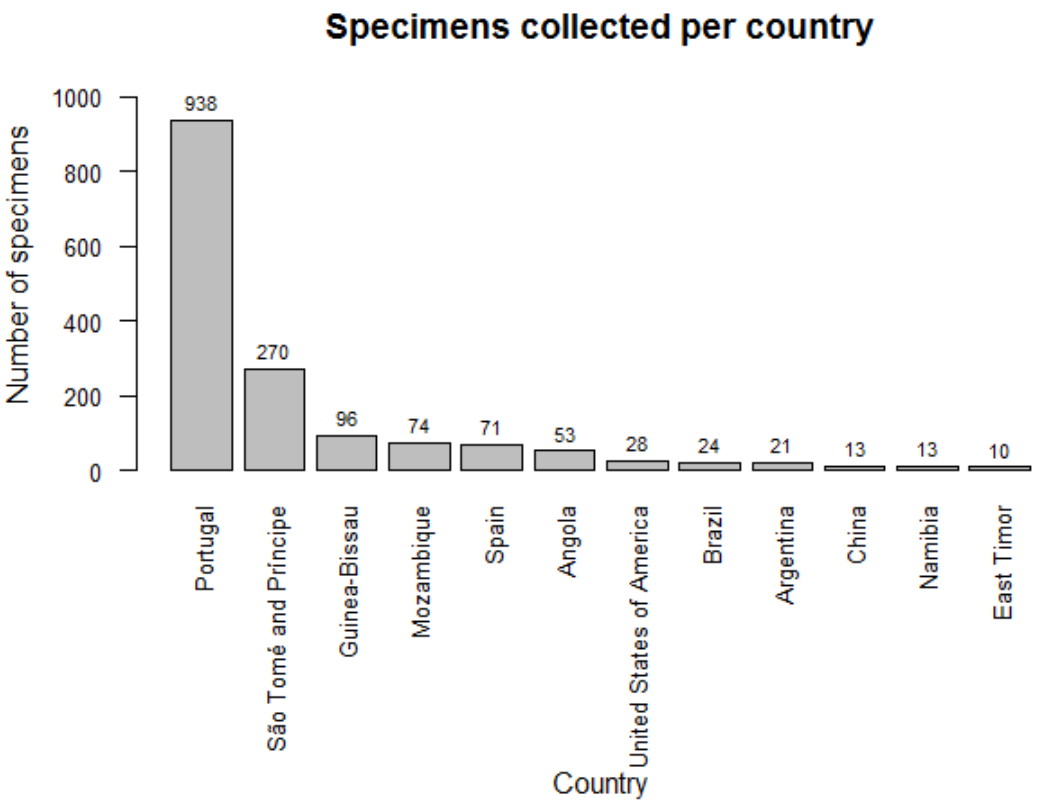

Figure 5.5 Bar graph of Tabanidae specimens in the IICT/MNHNC collections by sampling country. Countries represented by 10 or more specimens in the collections are shown.

Regarding temporal coverage, specimens in the collections were sampled between 1899 and 2018. Only one specimen was sampled in 1899, the remaining specimens were sampled after 1920. The majority of the specimens were sampled after 1970, and 674 (40.5%) of them were sampled between 1980 and 1990 (Figure 5.6).

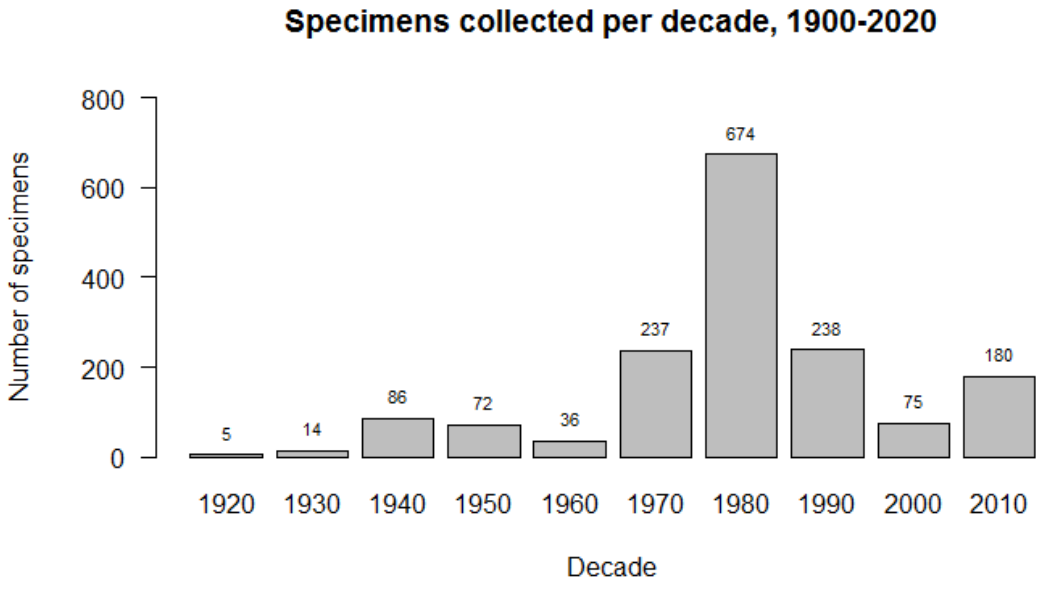

Figure 5.6 Histogram of tabanid specimens in the IICT/MNHNC collections by sampling year aggregated per decade.

5.3.1. Distribution maps

In order to provide a more complete characterization of the geographic distribution of the IICT/MNHNC Tabanidae collections, distribution maps are presented for the 9 species with over 30 specimens with geocoded sampling locations in the dataset. Figure 5.7 shows sampling countries for all specimens in the collections. The collections have a wide geographic distribution. Portugal is the country where most specimens were collected, being the continental territory well represented in the collection (Figure 5.8).

For species that also have occurrences published on GBIF, maps of countries represented on GBIF and in the IICT/MNHNC collections are shown. The datasets used for GBIF occurrences are referenced in Annex C.

In general, the sampling locations of specimens are in accordance with the known distributions of the species, with a few exceptions that should be noted. For *Haematopota csikii*, distribution is reported by Portillo (2002), in Portugal, for Alentejo [100], while one of the specimens in the IICT/MNHNC collections was collected further North, in Gerês. *Hybomitra solstitialis* has been reported for Minho, while in the IICT/MNHNC collections there is one specimen collected in Beja and one in Serra da Estrela. Both of these specimens were identified as *Hybomitra bimaculata*, which was considered a junior synonym of *Hybomitra solstitialis* [100]. *Pangonius haustellatus* hasn't been reported for Portugal, only for Spain [100]; one specimen of *P. haustellatus* is included in the collections, sampled in Santarém, Portugal.

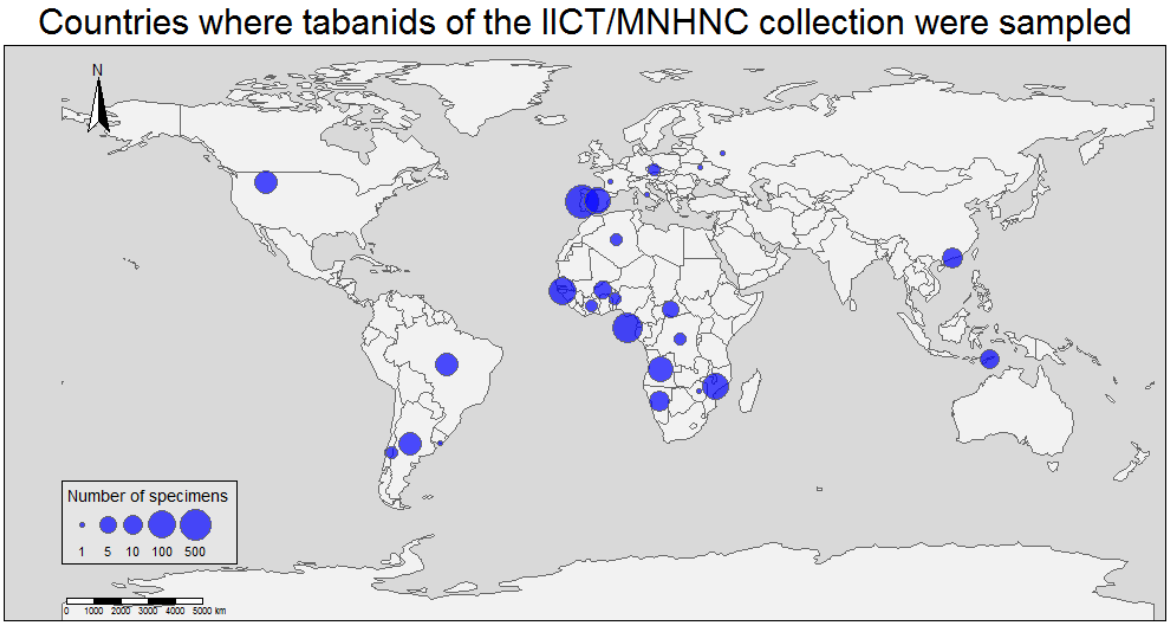

Figure 5.7 World map representing the countries where Tabanidae specimens of the IICT/MNHNC collections were collected. Circle size represents the number of specimens collected in each country.

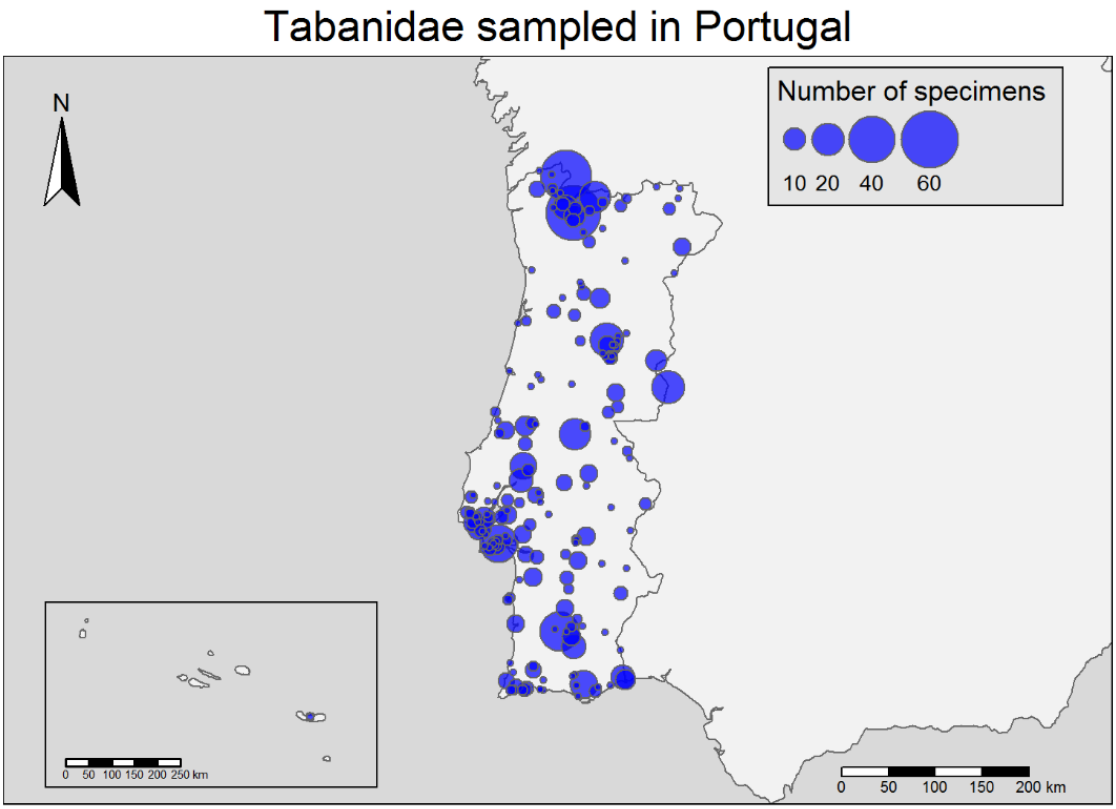

Figure 5.8 Map of sampling locations of Tabanidae of the IICT/MNHNC collections in Portugal. Circle size represents the number of specimens sampled at each location. Inset shows the Azores islands.

Tabanus monocallosus Travassos Dias, 1955

The IICT collection includes 260 specimens of *Tabanus monocallosus*, collected between 1972 and 2018. Of these, 77 were collected in São Tomé island, 174 in Príncipe island, and the other 9 don´t include information about the collection location other than the country. The sampling locations that were possible to geocode are shown in Figure 5.9. GBIF currently contains only one record of this species, collected in São Tomé and Príncipe, without further information about collection location or date.

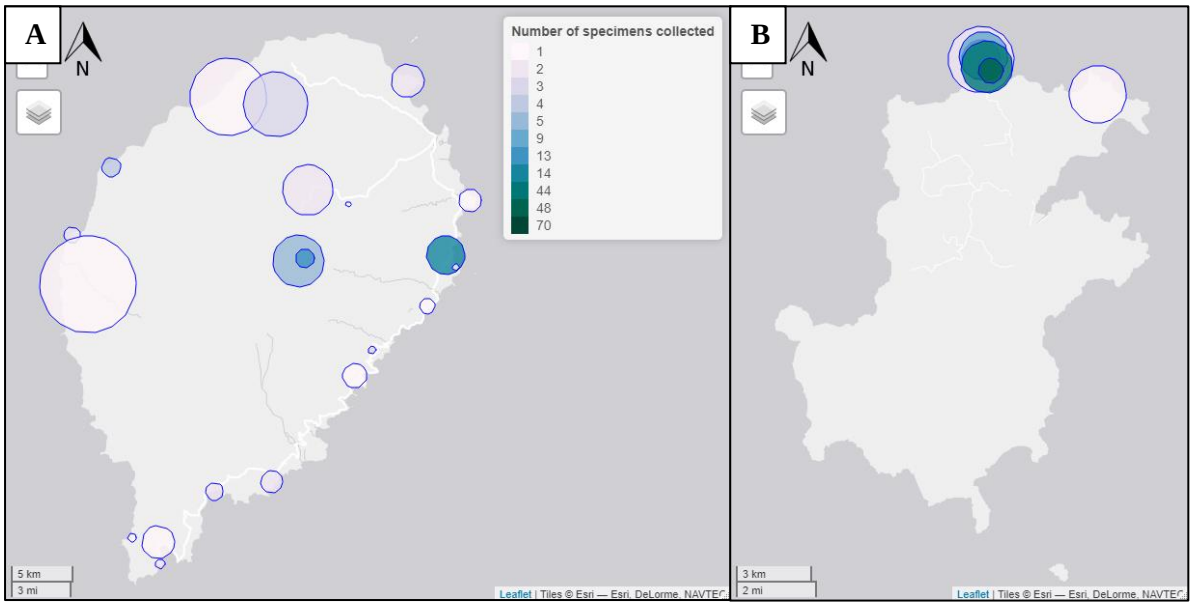

Figure 5.9 Collection location for specimens of *Tabanus monocallosus* of the IICT/MNHNC collection collected in São Tomé and Príncipe, in the islands of (A) São Tomé and (B) Príncipe. Circle size corresponds to the uncertainty area for each sampling location, fill colour represents the number of specimens collected.

Tabanus eggeri Schiner, 1868

This species is widely distributed throughout the Mediterranean region and is very common in the Iberian Peninsula [100]. GBIF lists occurrences in Portugal, France, Greece, Czech Republic, Turkey and Lebanon (Figure 5.10). Considering the IICT/MNHNC collections, 3 specimens were collected in Spain and 116 in Portugal across the continental territory (Figure 5.11).

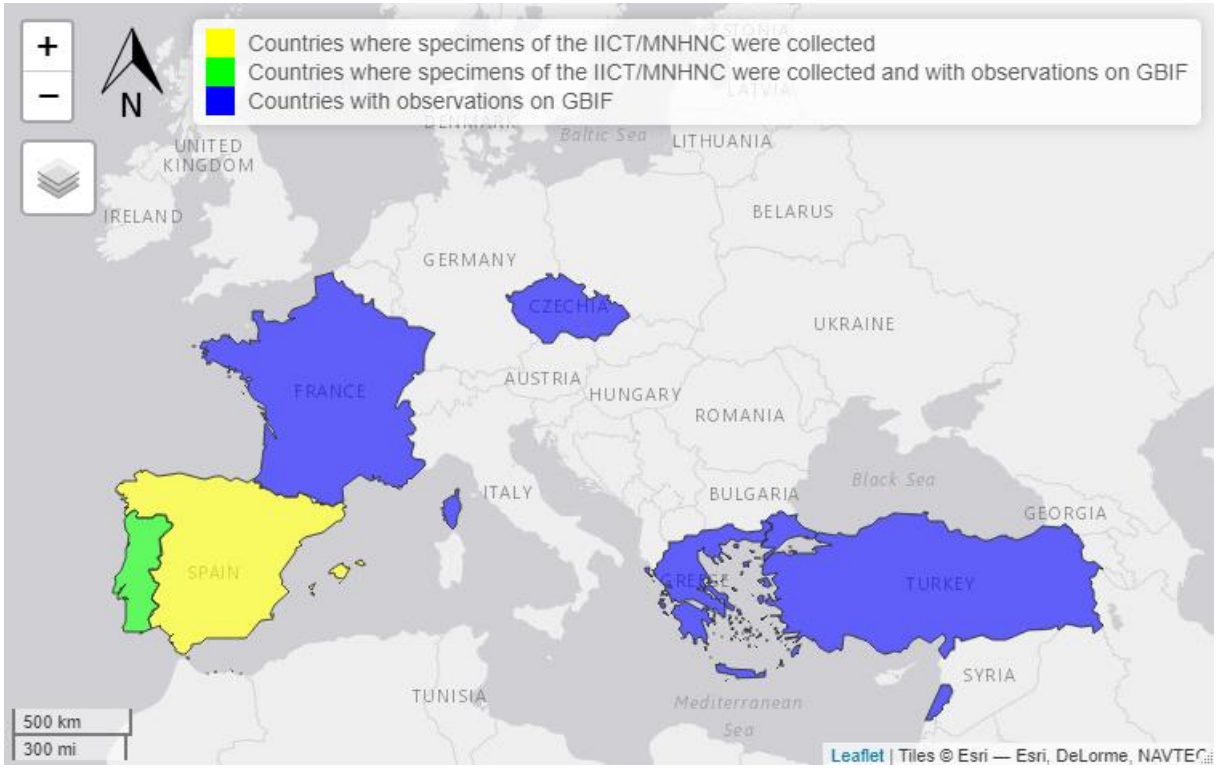

Figure 5.10 Countries where sampling of specimens of *Tabanus eggeri* has been registered on GBIF (blue) and where *T. eggeri* specimens of the IICT/MNHNC collections were sampled (yellow). Portugal, which has occurrences of *T. eggeri* registered on GBIF and is also represented in the MNHNC/IICT collections is shown in green.

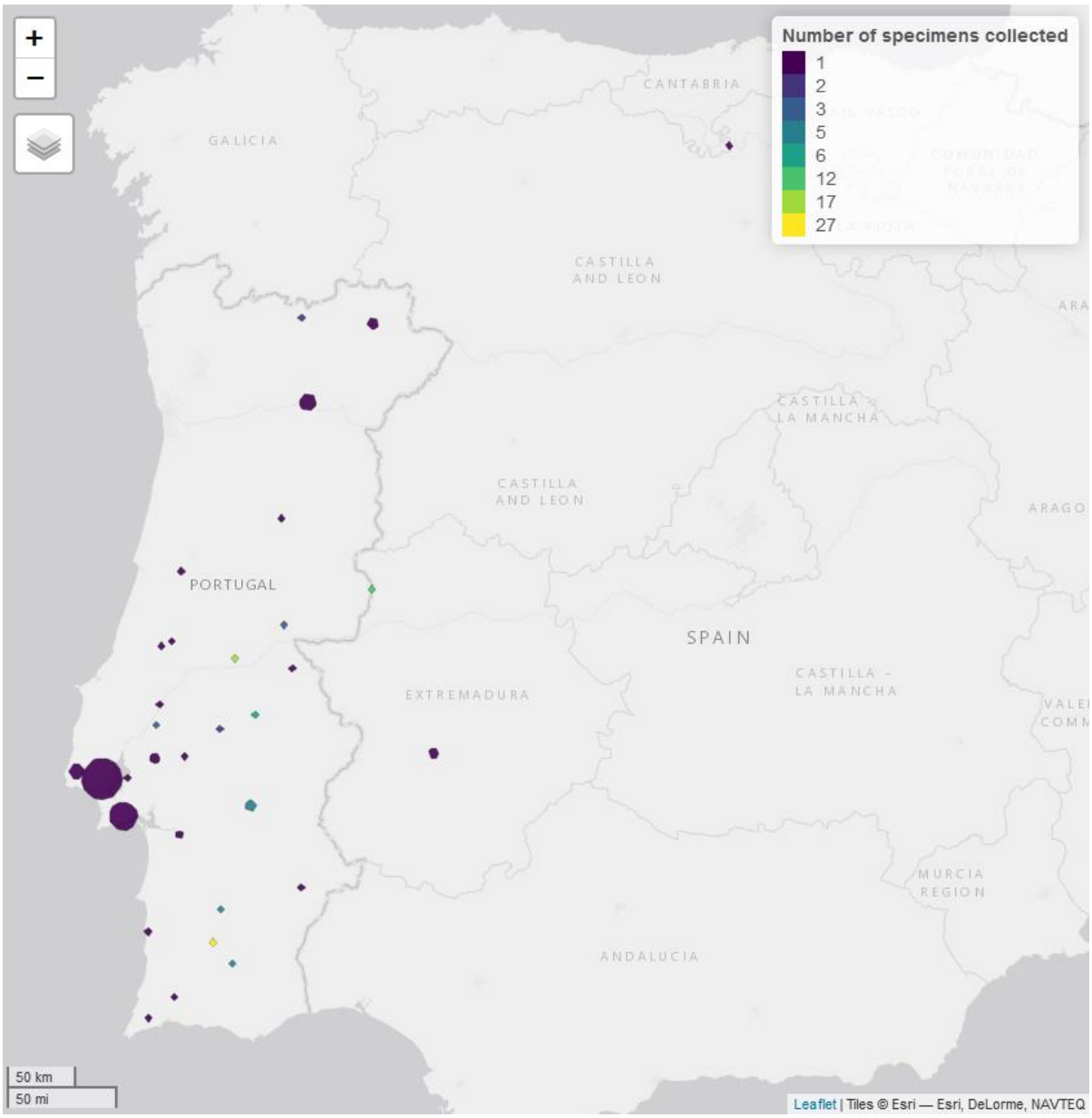

Figure 5.11 Locations where specimens of *Tabanus eggeri* in the IICT/MNHNC collections were collected. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 3000 m. Locations with geocoding uncertainty smaller than 3000 m are represented by lozenges. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 27).

Haematopota italica Meigen, 1804

According to Portillo (2002) this species occurs in all of Europe except Finland and Norway, and in North Africa [100]. GBIF also lists occurrences of this species throughout Europe, including Finland and Norway (Figure 5.12). 84 specimens included in the IICT/MNHNC collections were sampled in Portugal (82) and Spain (2). Of these, it was possible to geocode the sampling locality of 77 specimens from Portugal. Specimens sampled in Portugal were mostly from the North of the country, although there were some specimens collected in Santarém (5), Setúbal (6), Faro (1) and Portalegre (1), as shown in Figure 5.13.

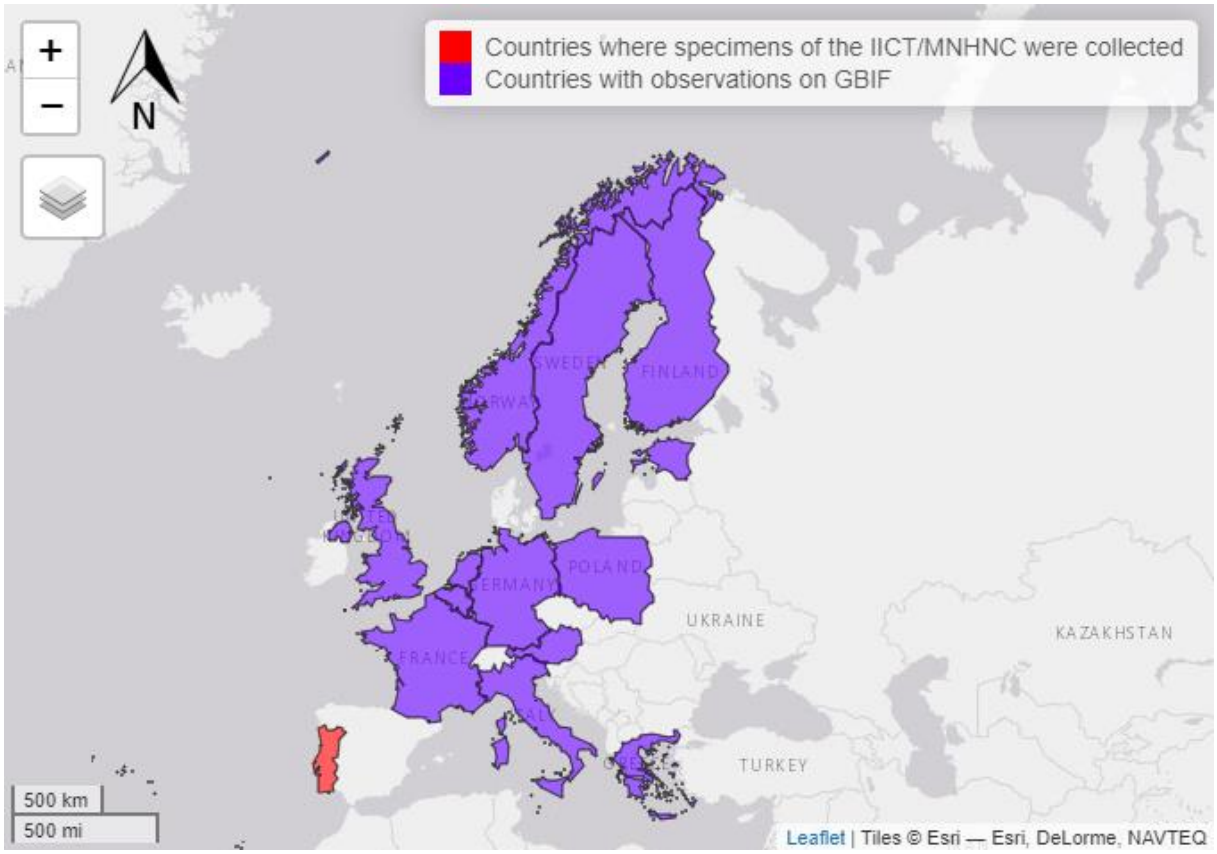

Figure 5.12 Countries where sampling of specimens of *Haematopota italica* has been registered on GBIF (blue) and where *H. italica* specimens of the IICT/MNHNC collections were sampled (red).

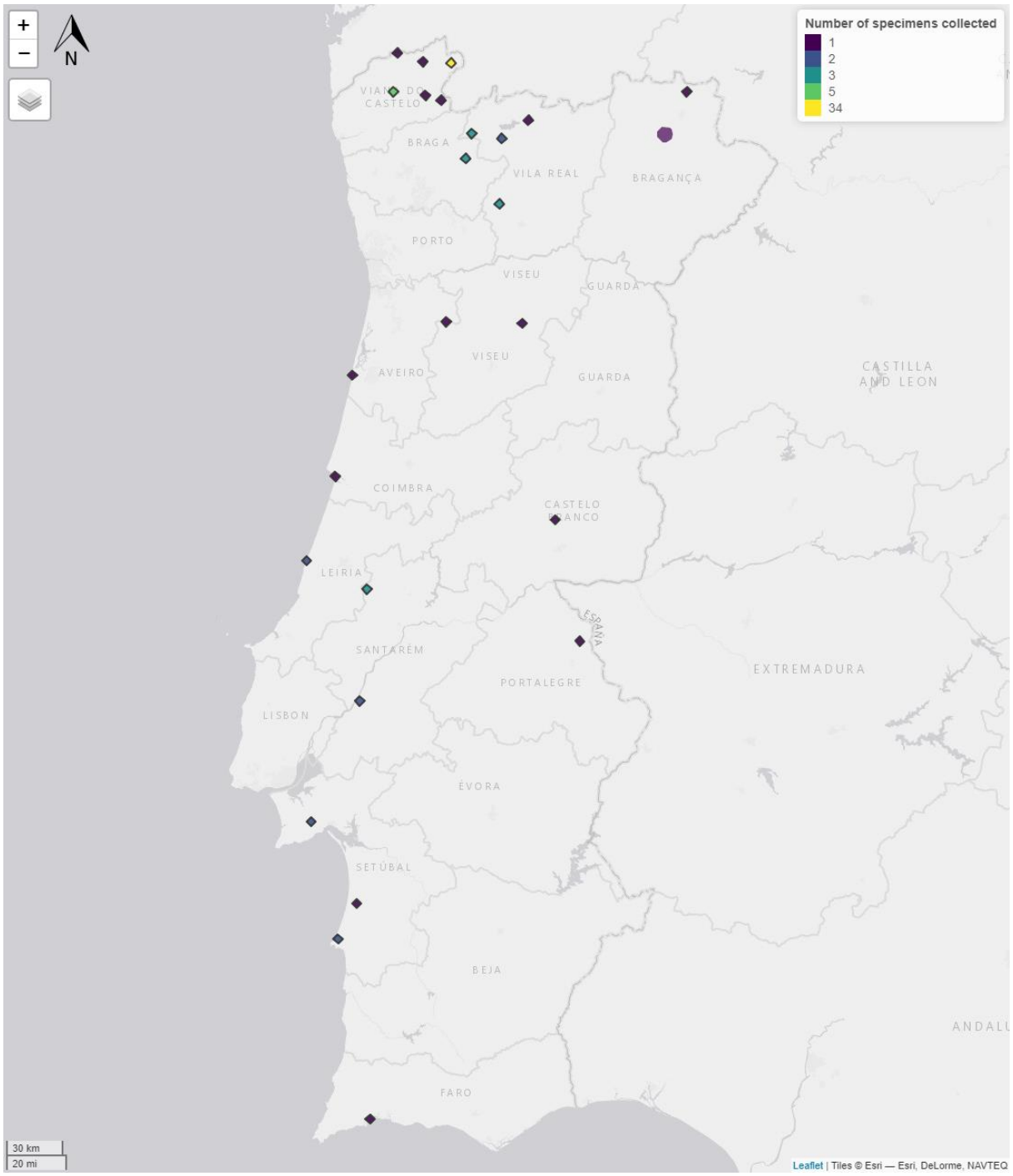

Figure 5.13 Locations where specimens of *Haematopota italica* in the IICT/MNHNC collections were collected. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 2500 m. Locations with geocoding uncertainty smaller than 2500 m are represented by lozenges with black trim. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 34).

Tabanus autumnalis Linnaeus, 1761

This species is widely distributed, but most common in low to medium altitude areas; it occurs in the western half of the Palearctic region [100]. Occurrences reported on GBIF are listed all over Europe (but none in Portugal and only 3 in Spain, in the Balearic Islands and in Barcelona), as well as in Algeria and Pakistan (Figure 5.14). In the IICT/MNHNC collections, 65 specimens are from Portugal, and one from Spain (Figure 5.15). The specimens from Portugal were collected mainly in the South (Lisbon, Santarém, Setúbal, Beja, Faro), although 3 were collected further North, and one was collected in the Azores (Figure 5.16).

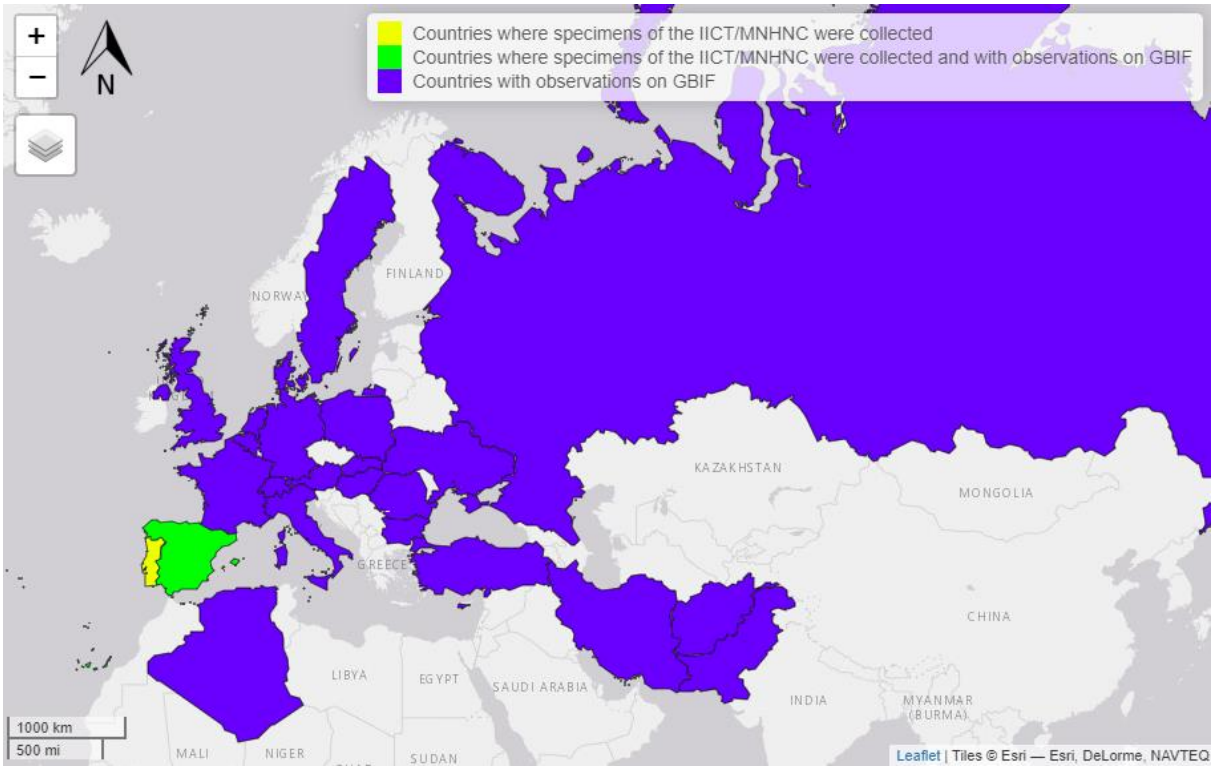

Figure 5.14 Countries where *Tabanus autumnalis* specimens have been registered on GBIF (blue) and in the IICT/MNHNC collections (yellow). The only country where specimens of *T. autumnalis* of the IICT/MNHNC collections were sampled and where there are occurrences of this Species registered on GBIF is Spain, shown in green.

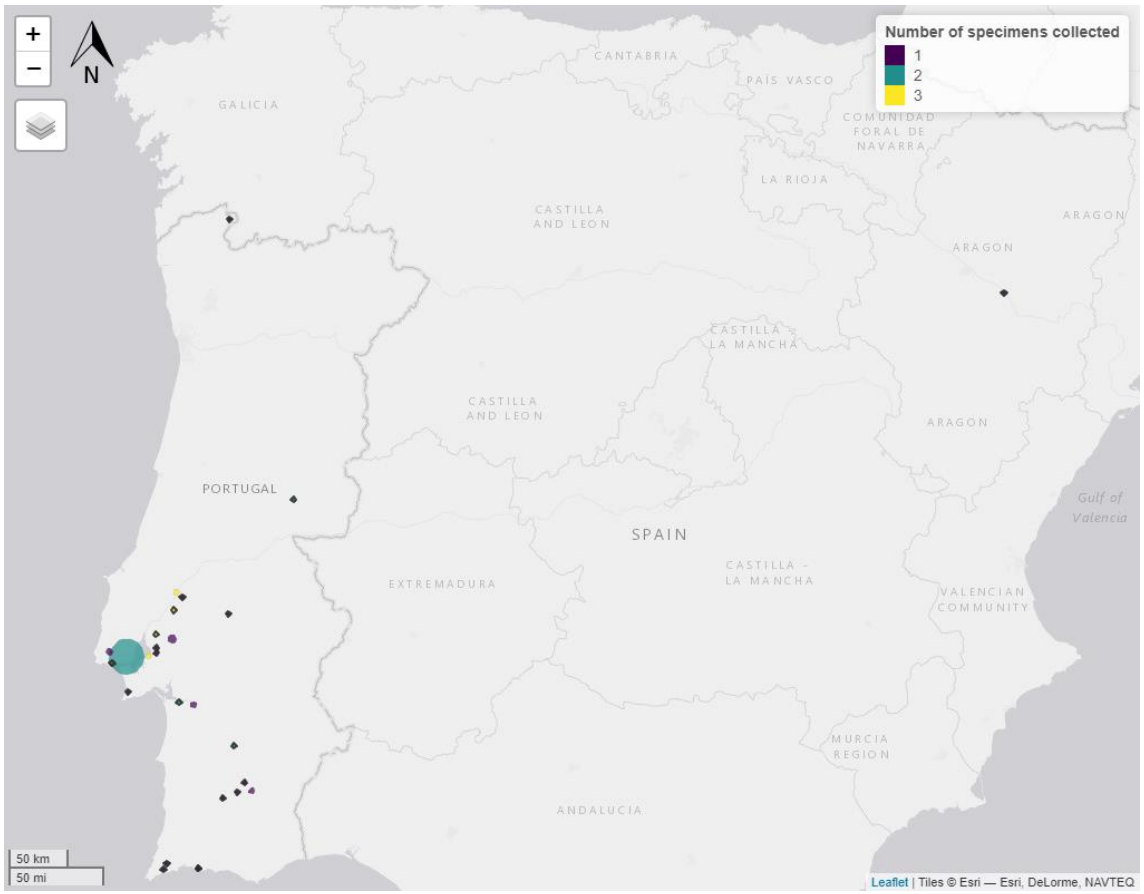

Figure 5.15 Locations where specimens of *Tabanus autumnalis* in the IICT/MNHNC collections were sampled. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 2500 m. Locations with geocoding uncertainty smaller than 2500 m are represented by lozenges with black trim. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 3).

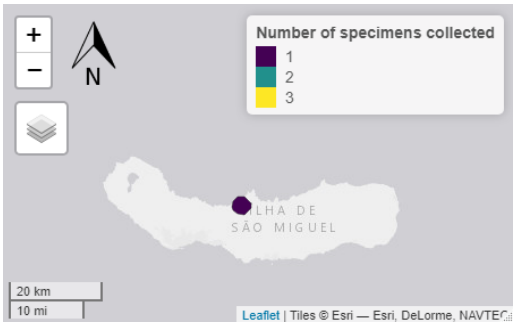

Figure 5.16 Sampling location for the specimen of *Tabanus autumnalis* in the IICT/MNHNC collections sampled in São Miguel island, Azores.

Tabanus sudeticus Zeller, 1842

This species is more commonly found in hills and mountains and has a known distribution through Europe, Asia Minor and North Africa [100]. GBIF contains occurrences in almost all of Europe, but none in Portugal (Figure 5.17). The IICT/MNHNC collection contains 43 specimens of this species. These were collected in Spain (10), Portugal (32) and France (1), with collection locations shown in Figure 5.18 and Figure 5.19. Almost all of them were collected in mountainous regions: Sierra Nevada (10), Serra da Cabreira (21), Serra da Estrela (10) and Serra do Gerês (1).

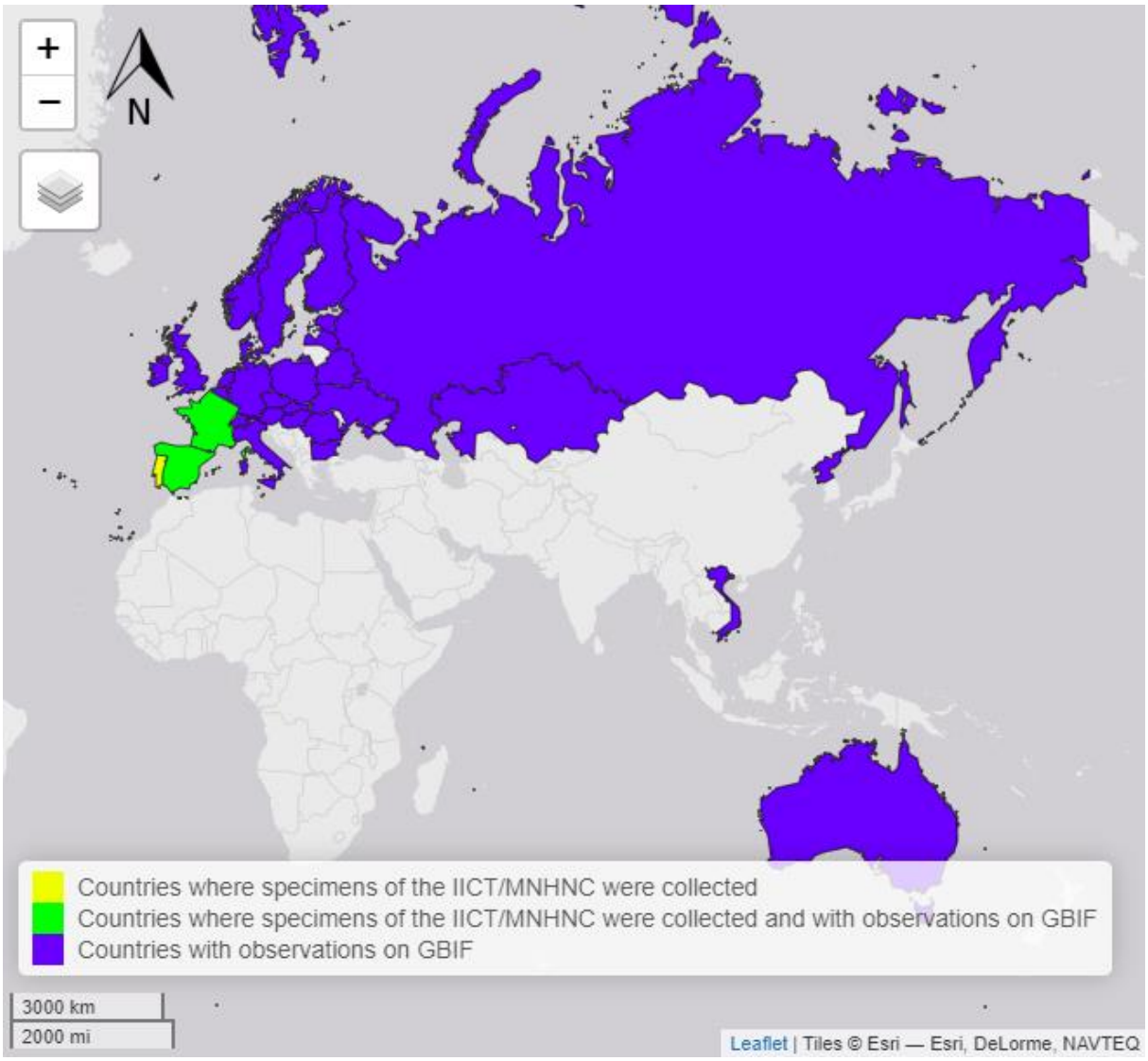

Figure 5.17 Countries where *Tabanus sudeticus* specimens have been registered on GBIF (blue) and in the IICT/MNHNC collections (yellow). The countries where specimens of *T. sudeticus* of the IICT/MNHNC collections were sampled and where there are occurrences of this Species registered on GBIF are Spain and France, shown in green.

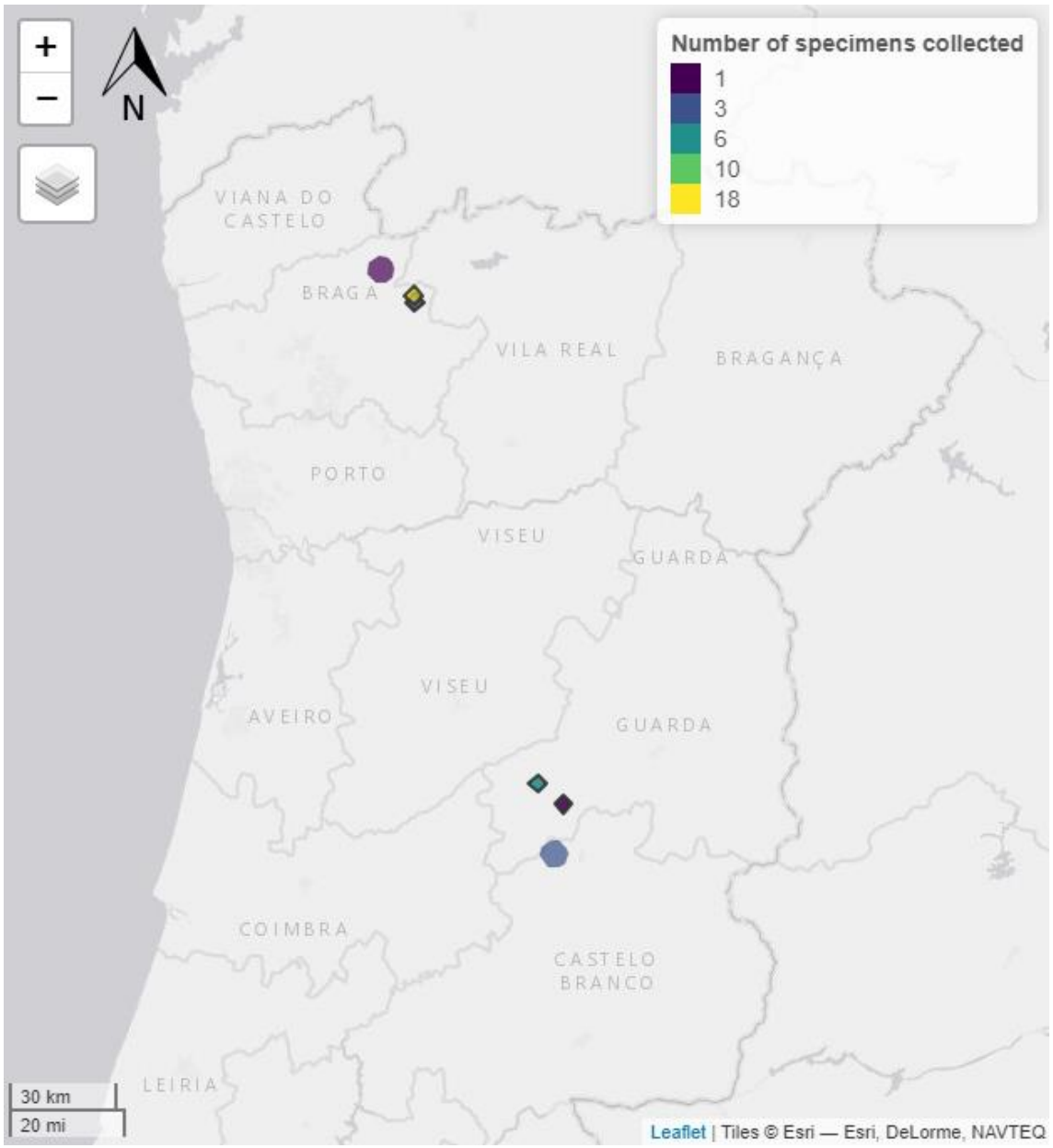

Figure 5.18 Sampling locations of *Tabanus sudeticus* specimens from Portugal found in the IICT/MNHNC collections. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 2500 m. Locations with geocoding uncertainty smaller than 2500 m are represented by lozenges with black trim. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 18).

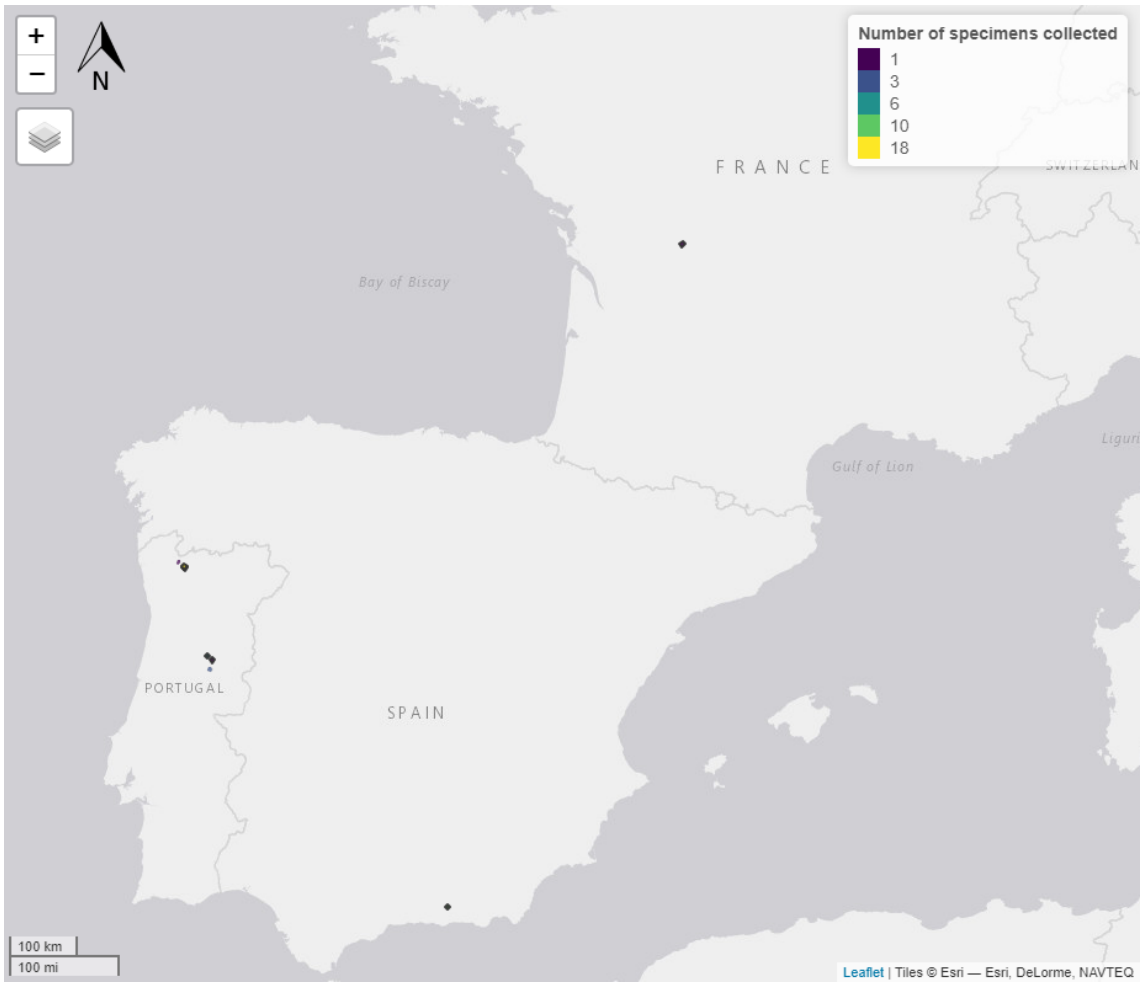

Figure 5.19 Sampling locations of *Tabanus sudeticus* specimens in the IICT/MNHNC collections sampled in Portugal (32 specimens), Spain (10 specimens, all in the same location in Andalucia) and France (1 specimen).

Tabanus bromius Linnaeus, 1758

This species is found all over Europe, the Middle East and North Africa [100]. GBIF contains occurrences located throughout Europe, in Northern Africa, the United States and Australia (Figure 5.20). The IICT collection contains 43 specimens collected in Portugal, including 2 specimens classified as *Tabanus cruzesilvai* (one of them a holotype), a species described by Travassos Dias in 1980 [89] which was considered a synonym of *T. bromius* by Portillo (2002) [100]. These specimens were collected across continental Portugal (Figure 5.21).

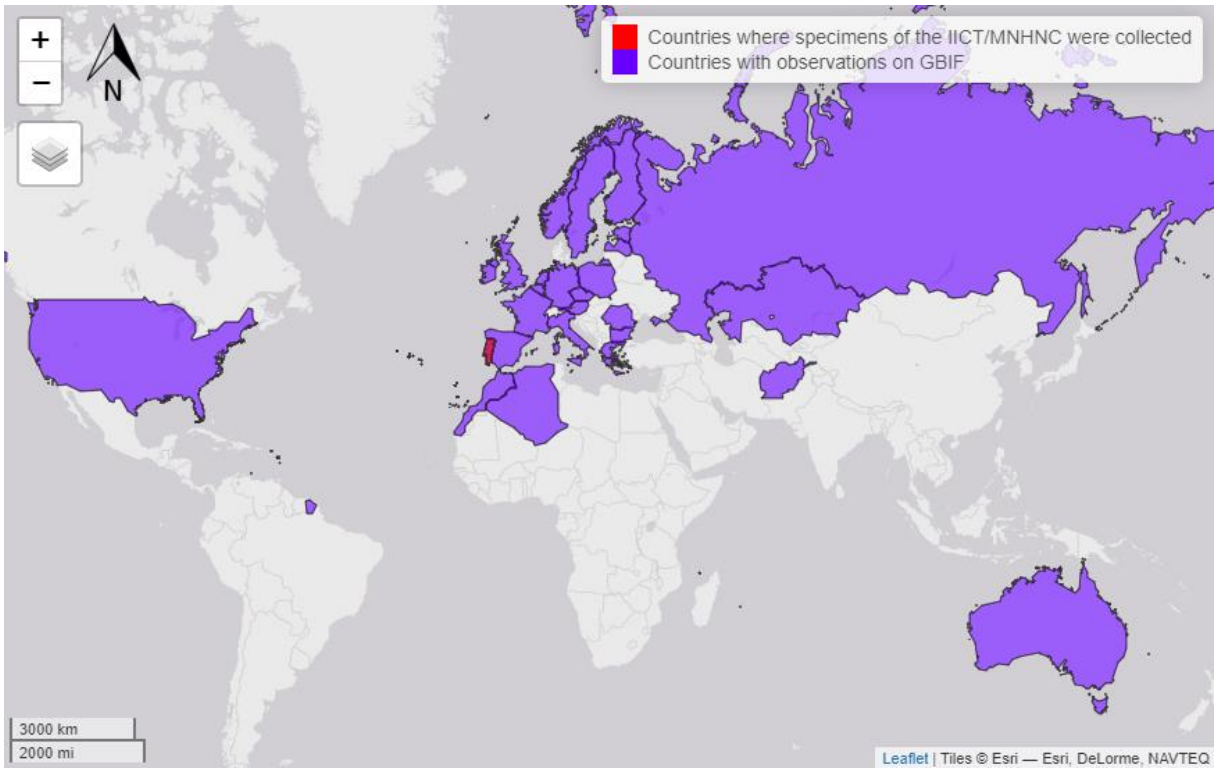

Figure 5.20 Countries where sampling of specimens of *Tabanus bromius* have been registered on GBIF (blue) and where *T. bromius* specimens of the IICT/MNHNC collections were sampled (red).

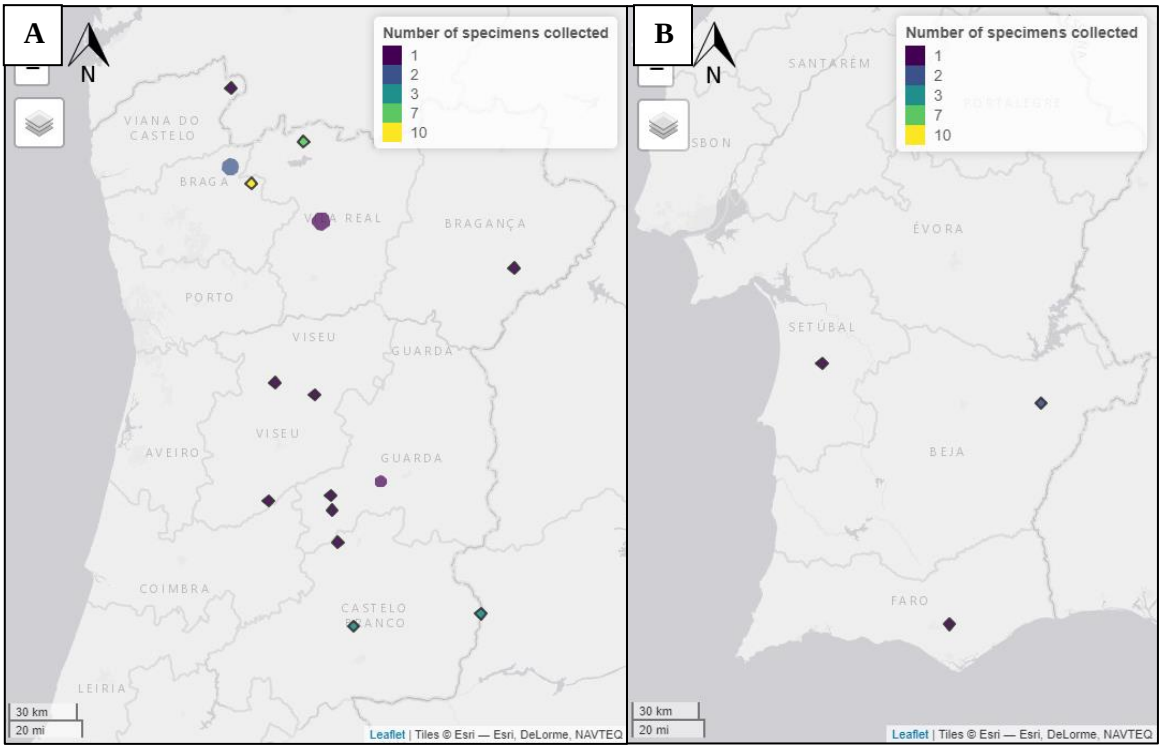

Figure 5.21 Locations where specimens of *Tabanus bromius* in the IICT/MNHNC collections were collected, in the (A) North and (B) South of Portugal. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 2500 m. Locations with geocoding uncertainty smaller than 2500 m are represented by lozenges with black trim. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 10).

Tabanus barbarus Coquebert, 1804

Portillo (2002) cites this species as occurring in the Southwest of Europe and Northern Africa [100]. 2 occurrences sampled in Tunisia are published on GBIF. 33 specimens of IICT/MNHNC collections were collected in Portugal, of which it was possible to geocode the sampling locations of 31 (Figure 5.22). All of the specimens were sampled in the region of Lisboa, Setúbal, Santarém and Leiria, since in Portugal, this species occurs in center-west region [100].

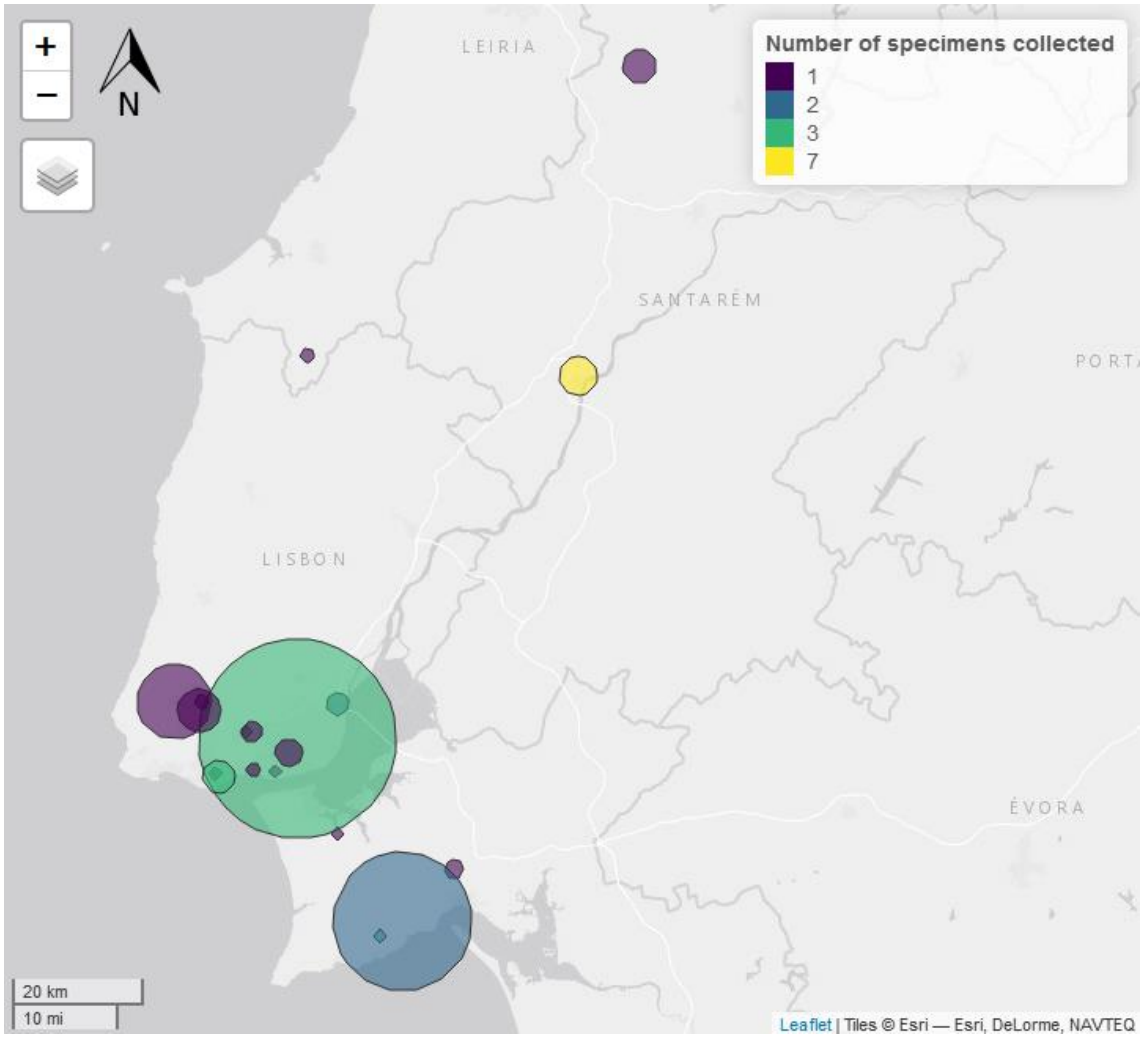

Figure 5.22 Locations where specimens of *Tabanus barbarus* in the IICT/MNHNC collection were collected. Circle size represents the geocoding uncertainty of the sampling locality, for localities with uncertainty greater than 1000 m. Locations with geocoding uncertainty smaller than 1000 m are represented by lozenges with black outlines. In both cases, fill color represents the number of specimens collected in each location (minimum = 1, maximum = 7).

Ancala fasciata (Fabricius, 1775)

GBIF lists occurrences of this species in Cameroon, Central African Republic, Gabon, Ghana, Nigeria and Senegal (Figure 5.23). Considering the specimens in the IICT/MNHNC collections, 39 were found in Guinea-Bissau, 31 of which were possible to geocode (Figure 5.24).

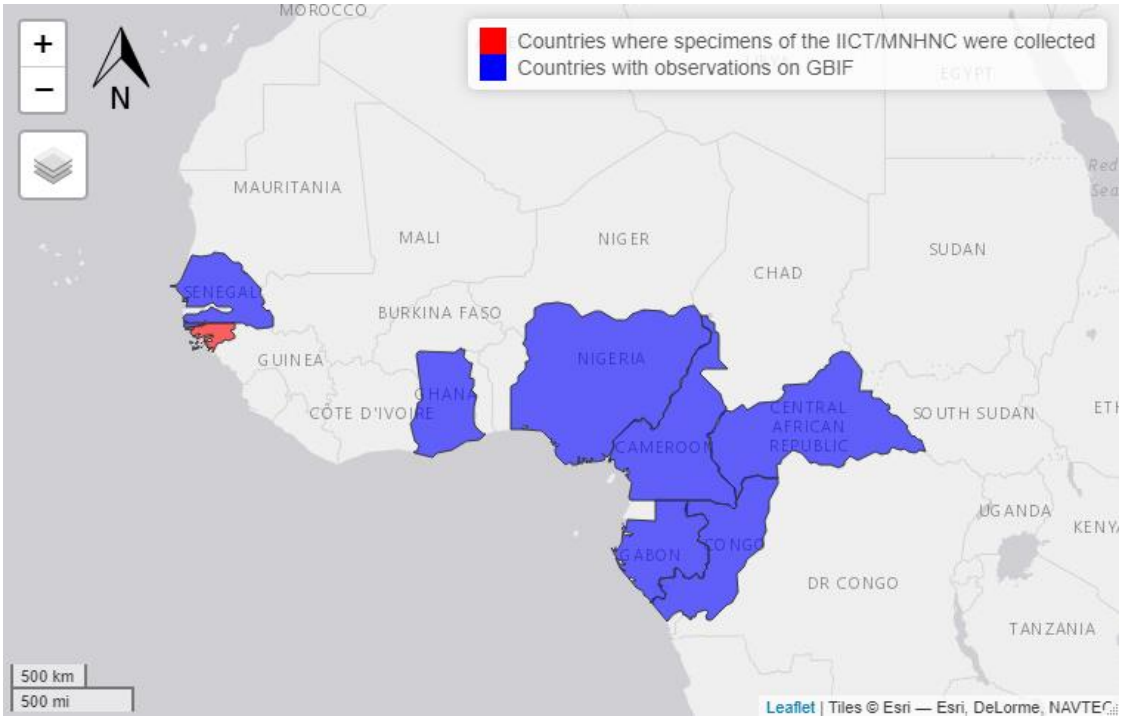

Figure 5.23 Countries where occurrences of *Ancala fasciata* have been registered on GBIF (blue) and in the IICT/MNHNC collections (red).

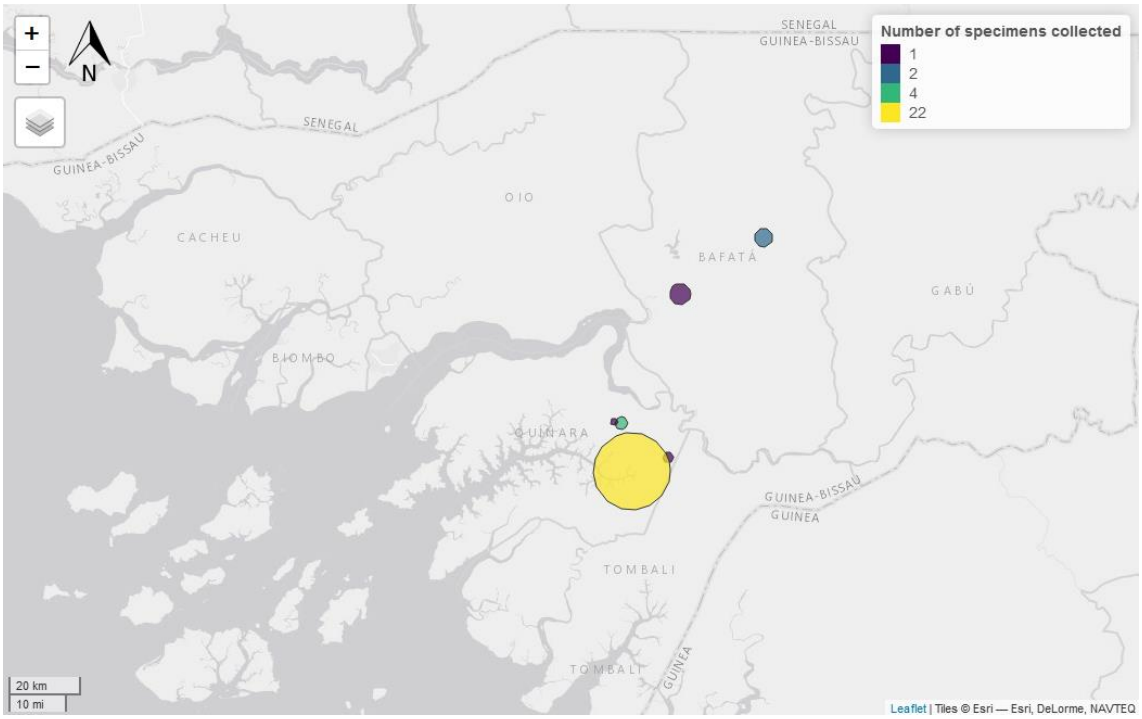

Figure 5.24 Locations where specimens of *Ancala fasciata* in the IICT/MNHNC collections were sampled. Circle size represents the uncertainty of the sampling locality. Circle color represents the number of specimens sampled in each location (minimum = 1, maximum = 22).

Tabanus mesquitelai Travassos Santos Dias, 1991

35 specimens of this species are contained in the IICT/MNHNC collections, all sampled in Guinea-Bissau, including 1 holotype and 15 paratypes. Of these, it was possible to geocode the sampling locations of 31 (Figure 5.25). There are currently no occurrences of this species published on GBIF.

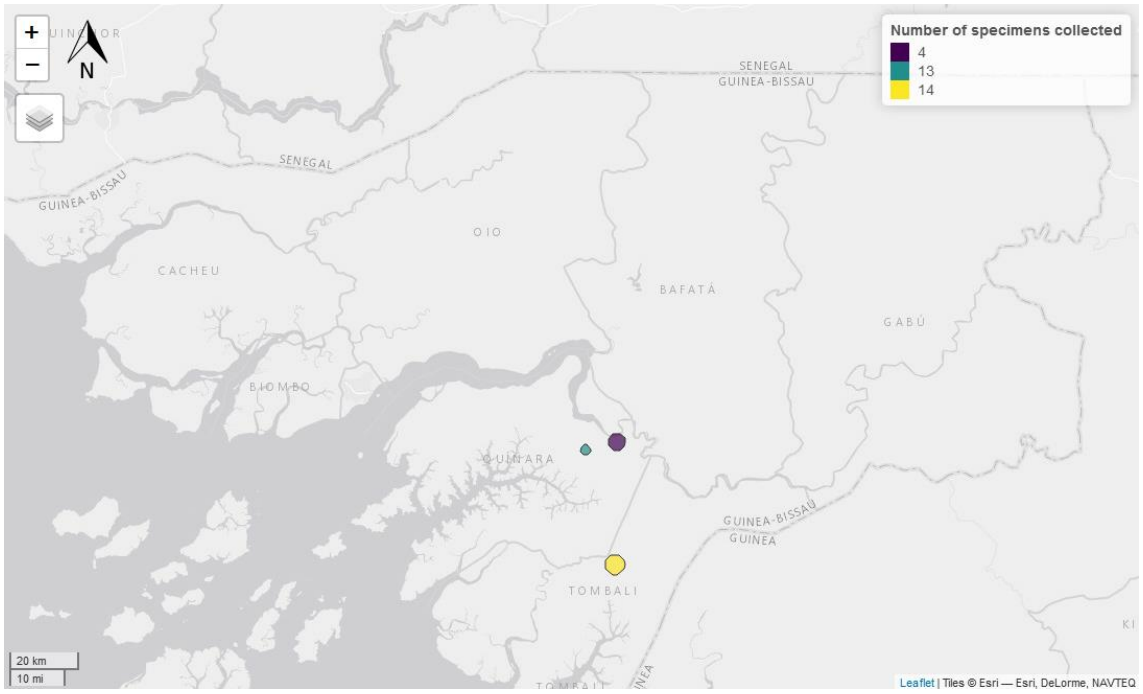

Figure 5.25 Locations where specimens of *Tabanus mesquitelai* in the IICT/MNHNC collections were sampled. Circle size represents the uncertainty of the sampling locality. Circle color represents the number of specimens sampled in each location (minimum = 4, maximum = 14).

5.4. Discussion

In this work, a dataset representing 1666 specimens of the Tabanidae (Diptera) Family was digitized and characterized. Publication of this dataset on GBIF will significantly increase the data available about this group. The collections represented in the dataset are of special importance considering that most of the specimens were compiled and/or identified and studied by Travassos Santos Dias, a specialist who described many species in this family.

Several of the species in this list have few or no occurrences published on GBIF. For example, *Tabanus monocallosus*, a species known to occur only in São Tomé and Príncipe, has only one occurrence published on GBIF, but with no associated coordinates for the sampling location. The publication of this dataset will add 260 occurrences to São Tomé and Príncipe, with specimens sampled in both islands. 238 of these occurrences are geocoded. *Tabanus eggeri*, a widespread species in Europe, has just 18 registered occurrences on GBIF, and only 10 include coordinates. The publication of this dataset will add 120 more occurrences to this species, 108 of which are geocoded.

This work also provides a characterization of the distribution of 9 species of tabanids through distribution maps. For these species, mostly sampled in Portugal, São Tomé and Príncipe and Guinea-Bissau, a wider understanding of their distribution is gained. Although many of these occurrences were reported by Travassos Dias individually, a comprehensive geographic analysis has not yet been done for these collections. Therefore, this work provides a better understanding of the locations where *Tabanus monocallosus*, *Tabanus eggeri*, *Haematopota italica*, *Tabanus autumnalis*, *Tabanus sudeticus*, *Tabanus bromius*, *Tabanus barbarus*, *Ancala fasciata* and *Tabanus mesquitelai* occur.

The sampling countries represented in the IICT/MNHNC collections were compared with the countries where occurrences of the same species were previously published on GBIF. This leads to the conclusion that for most of the species occurring in Portugal, no occurrences in this country were previously published in that platform. In fact, Portugal is underrepresented in terms of occurrences of Tabanidae published on GBIF; of 156 651 occurrences of the Tabanidae family published on GBIF prior to this work, only 42 were sampled in Portugal.

The publication of the dataset described here on GBIF, and the characterization of the IICT/MNHNC collections done in this work, will be available for further scientific studies of significant medical and veterinary importance, and on the distribution of this group, respectively.

6. Conclusions

In this work, focusing on the digitization of NHC data, the main steps of the digitization process were covered. These included enrichment of specimen data through automatic geocoding of sampling locations, the cleaning, enrichment and publication on GBIF of the MNHNC insect collection catalogue, the creation of a citizen science project in the Zooniverse platform, and the digitization of the data pertaining to a significant collection of specimens of the Tabanidae family, stored at the IICT and the MNHNC.

Different tools were tested to geocode sampling locations in an automated way through the use of APIs. This led to the conclusion that Google Maps presents the most accurate results, but the only tool that provides uncertainty radii, necessary for the gocoding of NHC sampling locations, is GEOLocate. These findings were published in [101]. A method combining both tools was chosen to geocode sampling locations in the MNHNC insect collection catalogue: the Google Maps API was first used to obtain coordinates, and the GEOLocate web interface was then used to confirm the coordinates and obtain uncertainty radii.

The insect collection catalogue was enriched by geocoding sampling locations, homogenizing and formatting data according to the DarwinCore standard. This resulted in a more complete, enriched dataset to be published on GBIF and available for consultation.

As a way to make the digitization process faster, a citizen science project was created in the Zooniverse platform and tested with specimens of the MNHNC insect collection. The results showed that citizen science is a very effective tool for the digitization process, as long as effective error checking methodologies are implemented. It also proved to be a tool that can be used by museum staff as a practical digitization platform, allowing the easy simultaneous view of the specimen together with a form to submit specimen data. A description of the citizen science project and the results achieved so far have been published in [102].

The digitization and publication of the tabanid collections of the IICT/MNHNC on GBIF resulted in an important source of information becoming available, providing a large number of new online records of wide geographic and temporal coverage. Moreover, distribution maps were produced for the most represented species in the collection to better illustrate their sampling locations. These data can be used in the future for studies of this group of insects of great medical and veterinary importance.

Since all steps of a digitization process were covered in this work, it is possible to revisit the conclusions of Guralnick *et al.* (2006) [20], proposing three main challenges to create a NHC dataset: i) transcribing the data to a computer database, ii) geocoding the records, and iii) publishing it online [20]. Based on the findings of the present work, a new structure for these challenges can be proposed: i) specimen imaging and management of resulting files; ii) transcribing the data to a computer database, iii) geocoding the records, iv) adding and verifying taxonomic identifications, v) cleaning and standardizing data; vi) publishing it online. The challenge of publishing the dataset online has been greatly facilitated by the development of specific databases for biodiversity data, such as GBIF and iDigBio, but it is still relevant.

While digitization of NHC always presents challenges, insect collections have specificities that make them different from collections of other taxonomic groups. In general, the number of specimens in these collections is much larger than in collections of other groups [2], meaning that there are more records to digitize. Citizen science projects are an excellent way to deal with this issue, and they are versatile enough to allow transcription of both handwritten and typewritten data. This can also be seen as an advantage to insect collections, due to the sheer amount of information contained in them. Additionally,

insects are an especially diverse group [103], leading to challenges in taxonomic data cleaning. These can be countered with the use of taxonomic checklists to validate data. Fuzzy matching is important to detect transcription errors, but it has to be used with caution to avoid swapping very similar species names.

These challenges show the importance of applying biodiversity informatics tools to the digitization of NHC data. While significant efforts are being taken to digitize this information, there is still much work left to be done; automation of tasks will be key to achieving the goal of having all NHC data digitized and available online. While some methods were tested in this work, others can be proposed for other steps of the digitization process. For instance, advances in optical character recognition (OCR) technology may in the future allow automatic transcription of handwritten data. Machine learning might be helpful in the taxonomic identification of species; deep learning has been tested to identify herbarium specimens to the species level with an accuracy between 58.5% and 79.6% [104]. With the development of deep learning models and with sufficiently large training datasets it may be possible to identify some species of insects from photographs.

References

1. Kemp C. Museums: The endangered dead. Nature. 2015;518:292–4. doi:10.1038/518292a.

2. Short AEZ, Dikow T, Moreau CS. Entomological collections in the age of big data. Annu Rev Entomol. 2018;63:513–30. doi:10.1146/annurev-ento-031616-035536.

3. Drew J. The role of natural history institutions and bioinformatics in conservation biology. Conserv Biol. 2011;25:1250–2. doi:10.1111/j.1523-1739.2011.01725.x.

4. Smith GF, Figueiredo E. Capacity building in taxonomy and systematics. Taxon. 2009;58:697–9.

5. McLean BS, Bell KC, Dunnum JL, Abrahamson B, Colella JP, Deardorff ER, et al. Natural history collections-based research: progress, promise, and best practices. J Mammal. 2016;97:287–97. doi:10.1093/jmammal/gyv178.

6. Holmes MW, Hammond TT, Wogan GOU, Walsh RE, LaBarbera K, Wommack EA, et al. Natural history collections as windows on evolutionary processes. Mol Ecol. 2016;25:864–81. doi:10.1111/mec.13529.

7. Schindel DE, Cook JA. The next generation of natural history collections. PLOS Biol. 2018;16:e2006125. doi:10.1371/journal.pbio.2006125.

8. Costello MJ, May RM, Stork NE. Can we name Earth's species before they go extinct? Science. 2013;339:413–6. doi:10.1126/science.1230318.

9. Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, et al. Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Syst Biodivers. 2012;10:1–20. doi:10.1080/14772000.2012.665095.

10. Thessen A, Patterson D. Data issues in the life sciences. Zookeys. 2011;150:15–51. doi:10.3897/zookeys.150.1766.

11. Secretariat of the Convention on Biological Diversity. Global biodiversity outlook 3. Montréal; 2010. www.cbd.int/GBO3.

12. Hobern D, Apostolico A, Arnaud E, Bello J, Canhos D, Dubois G, et al. Global biodiversity informatics outlook: Delivering biodiversity knowledge in the information age. Copenhagen: Global Biodiversity Information Facility; 2012. https://doi.org/10.15468/6jxa-yb44.

13. Beaman R, Cellinese N. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. Zookeys. 2012;209:7–17. doi:10.3897/zookeys.209.3313.

14. Butler D, Gee H, Macilwain C. Museum research comes off list of endangered species. Nature. 1998;394:115–7. doi:10.1038/28009.

15. Ariño AH. Approaches to estimating the universe of natural history collections data. Biodivers Informatics. 2010;7:81–92. doi:10.17161/bi.v7i2.3991.

16. Bertone M, Blinn R, Stanfield T, Dew K, Seltmann K, Deans A. Results and insights from the NCSU Insect Museum GigaPan project. Zookeys. 2012;209:115–32. doi:10.3897/zookeys.209.3083.

17. Hudson LN, Blagoderov V, Heaton A, Holtzhausen P, Livermore L, Price BW, et al. Inselect: Automating the digitization of natural history collections. PLoS One. 2015;10:e0143402. doi:10.1371/journal.pone.0143402.

18. Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS. No specimen left behind: industrial scale digitization of natural history collections. Zookeys. 2012;209:133–46.

doi:10.3897/zookeys.209.3178.

19. Nelson G, Paul D, Riccardi G, Mast A. Five task clusters that enable efficient and effective digitization of biological collections. Zookeys. 2012;209:19–45. doi:10.3897/zookeys.209.3135.

20. Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ. BioGeomancer: Automated georeferencing to map the world's biodiversity data. PLoS Biol. 2006;4:e381. doi:10.1371/journal.pbio.0040381.

21. Chapman A. Principles and methods of data cleaning: Primary species and species-occurrence data, version 1.0. Copenhagen: Global Biodiversity Information Facility; 2005.

22. Wieczorek J, Guo Q, Hijmans R. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. Int J Geogr Inf Sci. 2004;18:745–67. doi:10.1080/13658810412331280211.

23. Chapman A, Wieczorek J. Guide to best practices for georeferencing. Copenhagen: Global Biodiversity Information Facility; 2006.

24. Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, et al. Accelerating the digitization of biodiversity research specimens through online public participation. Bioscience. 2015;65:383–96. doi:10.1093/biosci/biv005.

25. Murphy PC, Guralnick RP, Glaubitz R, Neufeld D, Ryan JA. Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). Phyloinformatics. 2004;3:1–29.

26. R Core Team. R: A language and environment for statistical computing. 2018. https://www.r-project.org/.

27. Hijmans RJ. geosphere: Spherical trigonometry. 2017. https://cran.r-project.org/package=geosphere.

28. Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne PE. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol Evol. 2013;28:454–61. doi:10.1016/J.TREE.2013.05.002.

29. Escribano N, Galicia D, Ariño AH. The tragedy of the biodiversity data commons: a data impediment creeping nigher? Database. 2018;2018. doi:10.1093/database/bay033.

30. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics. 2011;12 Suppl 15 Suppl 15:S2. doi:10.1186/1471-2105-12-S15-S2.

31. Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. Darwin Core: An evolving community-developed biodiversity data standard. PLoS One. 2012;7:e29715. doi:10.1371/journal.pone.0029715.

32. Costello MJ, Wieczorek J. Best practice for biodiversity data management and publication. Biol Conserv. 2014;173:68–73. doi:10.1016/J.BIOCON.2013.10.018.

33. Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, et al. A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodivers Data J. 2014;2:e4221. doi:10.3897/BDJ.2.e4221.

34. Verborgh R, De Wilde M. Using OpenRefine: The essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web. 1st edition. Birmingham: Packt Publishing; 2013. https://www.packtpub.com/eu/big-data-and-business-intelligence/using-openrefine.

35. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Research. 2013;2:191.

doi:10.12688/f1000research.2-191.v2.

36. Lopes LF. Insect collection from the Museu Nacional de História Natural e da Ciência, Universidade de Lisboa, Portugal. National Museum of Natural History and Science, University of Lisbon. Occurrence dataset. GBIF.org. 2014. https://doi.org/10.15468/en7jvk. Accessed 19 Jun 2019.

37. Biodiversity Information Standards (TDWG). Darwin Core quick reference guide. dwc.tdwg.org/terms/. Accessed 22 Aug 2019.

38. Page RDM. iPhylo: Using Google Refine and taxonomic databases (EOL, NCBI, uBio, WORMS) to clean messy data. iPhylo. 2012. http://iphylo.blogspot.com/2012/02/using-google-refine-and-taxonomic.html. Accessed 30 Oct 2018.

39. Lemon J. Plotrix: a package in the red light district of R. 2006;:8–12.

40. Trouille L, Lintott CJ, Fortson LF. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. Proc Natl Acad Sci. 2019;116:1902–9. doi:10.1073/PNAS.1807190116.

41. Lee J, Kladwang W, Lee M, Cantu D, Azizyan M, Kim H, et al. RNA design rules from a massive open laboratory. Proc Natl Acad Sci U S A. 2014;111:2122–7. doi:10.1073/pnas.1313039111.

42. Causer T, Grint K, Sichani A-M, Terras M. 'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. Digit Scholarsh Humanit. 2018;33:467–87. doi:10.1093/llc/fqx064.

43. Barber A, Lafferty D, Landrum LR. The SALIX Method: A semi-automated workflow for herbarium specimen digitization. Taxon. 2013;62:581–90. doi:10.2307/taxon.62.3.581.

44. Hill A, Guralnick R, Smith A, Sallans A, Rosemary Gillespie, Denslow M, et al. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. Zookeys. 2012;209:219–33. doi:10.3897/zookeys.209.3472.

45. Australian Museum, The Atlas of Living Australia. DigiVol. https://volunteer.ala.org.au/. Accessed 3 May 2019.

46. Muséum national d'Histoire naturelle. Les herbonautes. http://lesherbonautes.mnhn.fr/. Accessed 3 May 2019.

47. Smithsonian. Smithsonian transcription center. 2019. https://transcription.si.edu/. Accessed 3 May 2019.

48. Bonney R, Phillips TB, Ballard HL, Enck JW. Can citizen science enhance public understanding of science? Public Underst Sci. 2016;25:2–16. doi:10.1177/0963662515607406.

49. JORDAN RC, GRAY SA, HOWE D V., BROOKS WR, EHRENFELD JG. Knowledge Gain and Behavioral Change in Citizen-Science Programs. Conserv Biol. 2011;25:1148–54. doi:10.1111/j.1523-1739.2011.01745.x.

50. Krawczyk C. panoptes_aggregation. 2019. https://aggregation-caesar.zooniverse.org/index.html.

51. Sauermann H, Franzoni C. Crowd science user contribution patterns and their implications. Proc Natl Acad Sci U S A. 2015;112:679–84. doi:10.1073/pnas.1408907112.

52. Tang V, Trouille L, Lintott C. Unlocking Data through Zooniverse: Science with 1.7 Million Volunteers. AGU Fall Meet Abstr. 2018;2018:ED11B-11. https://ui.adsabs.harvard.edu/abs/2018AGUFMED11B..11T/abstract.

53. Böse R, Friedhoff KT, Olbrich S, Büscher G, Domeyer I. Transmission of Trypanosoma theileri to

cattle by Tabanidae. Parasitol Res. 1987;73:421–4. doi:10.1007/BF00538199.

54. Service M. Horse flies (Tabanidae). In: Medical Entomology for Students. 5th edition. Cambridge: Cambridge University Press; 2012. p. 116–24.

55. Hornok S, Micsutka A, Meli ML, Lutz H, Hofmann-Lehmann R. Molecular investigation of transplacental and vector-borne transmission of bovine haemoplasmas. Vet Microbiol. 2011;152:411–4. doi:10.1016/j.vetmic.2011.04.031.

56. Krinsky WL. Animal disease agents transmitted by horse flies and deer flies (Diptera: Tabanidae). J Med Entomol. 1976;13:225–75.

57. Connal A, Connal SLM. The development of Loa Loa (Guyot) in Chrysops Silacea (Austen) and in Chrysops Dimidiata (van der Wulp). Trans R Soc Trop Med Hyg. 1922;16:64–89. doi:10.1016/S0035-9203(22)90984-8.

58. Hemmer W, Focke M, Vieluf D, Berg-Drewniok B, Götz M, Jarisch R. Anaphylaxis induced by horsefly bites: identification of a 69 kd IgE-binding salivary gland protein from Chrysops spp. (Diptera, Tabanidae) by western blot analysis. J Allergy Clin Immunol. 1998;101 1 Pt 1:134–6. doi:10.1016/S0091-6749(98)70208-8.

59. Strother S. Genus Tabanus. Tabanids (horseflies). What is this insect and how does it affect man? Dermatol Online J. 1999;5:6.

60. Squitier JM. Deer flies, yellow flies and horse flies - Chrysops, Diachlorus and Tabanus spp. Featured Creatures. University of Florida. 2014. http://entnemdept.ufl.edu/creatures/livestock/deer_fly.htm. Accessed 2 Nov 2018.

61. Baldacchino F, Desquesnes M, Mihok S, Foil LD, Duvallet G, Jittapalapong S. Tabanids: Neglected subjects of research, but important vectors of disease agents! Infect Genet Evol. 2014;28:596–615. doi:10.1016/j.meegid.2014.03.029.

62. Taioe MO, Motloang MY, Namangala B, Chota A, Molefe NI, Musinguzi SP, et al. Characterization of tabanid flies (Diptera: Tabanidae) in South Africa and Zambia and detection of protozoan parasites they are harbouring. Parasitology. 2017;144:1162–78. doi:10.1017/S0031182017000440.

63. Baldacchino F, Porciani A, Bernard C, Jay-Robert P. Spatial and temporal distribution of Tabanidae in the Pyrenees Mountains: the influence of altitude and landscape structure. Bull Entomol Res. 2014;104:1–11. doi:10.1017/S0007485313000254.

64. Henriques AL. Tabanidae (Diptera) of the American Museum of Natural History collection. Zootaxa. 2016;4137:151–86.

65. Tennekes M. tmap : Thematic maps in R. J Stat Softw. 2018;84:1–39. doi:10.18637/jss.v084.i06.

66. Appelhans T, Detsch F, Reudenbach C, Woellauer S. mapview: Interactive viewing of spatial data in R. 2019. https://cran.r-project.org/package=mapview.

67. Mendes LF, Fernandes IM, Paulos L. Lista anotada dos espécimes-tipo depositados nas colecções do Centro de Zoologia do Instituto de Investigação Científica Tropical. II. Insectos. Garcia Orta, Série Zool. 1988;15:45–62.

68. Travassos JA. Mais alguns dados sobre os tabanídeos (Diptera-Tabanidae) de Portugal. Garcia Orta, Série Zool. 1983;11:119–28.

69. Travassos Santos Dias JA. Descrição de um novo género e de uma nova espécie de tabanídeo (Diptera - Tabanidae) do Sudoeste Africano Bartolomeudiasiella atlanticus n. gen. n. sp. Bol da Soc Port Entomol. 1987;3:1–8.

70. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XI. Acerca de um lote proveniente da Arrábida. Bol da Soc Port Entomol. 1985;4:123–36.

71. Travassos Santos Dias JA. Contribuição para o conhecimento dos Tabanídeos (Diptera-Tabanidae) da República Centro-Africana. Garcia Orta, Série Zool. 1996;21:67–80.

72. Tabanid PEET project, Smithsonian Entomology Diptera Diversity Digitization Team. Haematopota angolensis Travassos Santos Dias, 1989. http://tabanidae.myspecies.info/taxonomy/term/1518/specimens. Accessed 23 Sep 2019.

73. Tabanid PEET project, Smithsonian Entomology Diptera Diversity Digitization Team. Haematopota chongoroiensis Travassos Santos Dias, 1989. http://tabanidae.myspecies.info/taxonomy/term/1590. Accessed 23 Sep 2019.

74. Travassos Santos Dias JA. Contribuição para o conhecimento dos tabanídeos (Diptera-Tabanidae) da Montanha do Ruwenzori. Garcia Orta, Série Zool. 1991;18.

75. Portillo M, Schacht W. Descripción de Haematopota eugeniae n. sp. (Diptera, Tabanidae). Doriana. 1984;6:1–7.

76. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XVI. Descrição de novas espécies do género Haematopota Meigen, 1803. Bol da Soc Port Entomol. 1990;4–24:290–300.

77. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XII. Alguns tavões coligidos pelo Prof. Doutor J. A. Quartau. Bol da Soc Port Entomol. 1985;4:137–48.

78. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, I. Estudo de uma colecção proveniente do Instituto de Higiene e Medicina Tropical de Lisboa. Bol da Soc Port Entomol. 1984;2–18:197–204.

79. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, II. Acerca de uma pequena colecção remetida pelo Dr. Artur Serrano. Bol da Soc Port Entomol. 1984;2–30:389–96.

80. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, IX. A propósito de uma colecção remetida pelo Dr. F. Colaço Alegre. Bol da Soc Port Entomol. 1984;2–35:457–66.

81. Travassos Santos Dias JA. Mais alguns novos tabanídeos (Diptera-Tabanidae) para a Fauna de Portugal. Garcia Orta, Série Zool. 1989;14:79–87.

82. Travassos Santos Dias JA. Descrição de uma nova espécie de tabanídeo (Diptera-Tabanidae) de Itália. Hybomitra mendesi n. sp. Bol da Soc Port Entomol. 1989;4 9,111:101–8.

83. Schacht W, Portillo M. Hybomitra (Mouchaemyia) tamujosoi sp. n., eine neue Bremsenart aus Spanien, nebst einem. Anhang zu Stonemyia hispanica (Krober, 1921) und Tabanus bromius var. flavofemuratus Strobl, 1909 (Diptera, Tabanidae). Entomofauna. 1982;3:161–76.

84. Portillo M. Descripción de Hybomytra zaballosi n. sp. (Diptera, Tabanidae). Nouv Rev d'Entomologie. 1988;5:383–7.

85. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, III. Estudo de um lote organizado pelo Engº. Tristão Branco. Bol da Soc Port Entomol. 1984;2–31:397–406.

86. Travassos Santos Dias JA. Contribuição para o conhecimento dos tabanídeos (Diptera-Tabanidae) da África meridional. Garcia Orta, Série Zool. 1988;15:97–124.

87. Coscarón S, Fairchild GB. El género Poeciloderas Lutz em Argentina (Tabanidae, Diptera, Insecta). Physis. 1976;35:95–103.

88. Travassos Santos Dias JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XVII.

Descrição de uma nova espécie, registo de um novo achado e de um segundo encontro de uma entidade ainda pouco conhecida. Bol da Soc Port Entomol. 1992;Supl.3:7–13.

89. Travassos JA. Contribuição para o estudo dos tabanídeos (Diptera-Tabanidae) de Portugal. Garcia Orta, Série Zool. 1980;9:105–28.

90. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XV. Descrição de uma nova espécie do género Tabanus Linnaeus, 1758. Bol da Soc Port Entomol. 1990;4–23:282–8.

91. Travassos JA. Nova contribuição para o conhecimento dos tabanídeos (Diptera-Tabanidae) de Portugal. An da Fac Ciências da Univ do Porto. 1979;61:211–32.

92. Travassos JA. Uma nova espécie de tabanídeo (Diptera, Tabanidae), da Guiné-Bissau: Tabanus mesquitelai n. sp. Garcia Orta, Série Zool. 1990;17:27–30.

93. Travassos Santos Dias JA. Descoberta de uma nova espécie do género Tabanus Linnaeus, 1758 (Diptera-Tabanidae) para a fauna de Angola. Garcia Orta, Série Zool. 1994;20:69–76.

94. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, VI. Acerca de algumas espécies provenientes da Reserva Natural Parcial da Serra da Malcata. Garcia Orta, Série Zool. 1985;12:95–100.

95. Travassos Santos Dias JA. Mais alguns novos dados sobre a fauna tabanidológica (Diptera-Tabanidae) de Angola. Inst Investig Agronómica Angola Série Científica. 1974;35:1–26.

96. Travassos Santos Dias JA. Nova contribuição para o conhecimento dos tabanídeos (Diptera-Tabanidae) de Angola. Sep da Rev Ciências Veterinárias. 1973;6 A:137–85.

97. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, XVIII. Estudo de uma colecção remetida pela família do Sr. Nuno Mendoça. Bol da Soc Port Entomol. 1992;3:15–30.

98. Travassos JA. Notas sobre os tabanídeos (Diptera-Tabanidae) de Portugal, X. Acerca de um lote organizado pelo Dr. A. Bivar de Sousa. Bol da Soc Port Entomol. 1985;3–5:1–19.

99. Travassos Santos Dias JA. Une nouvelle espèce de Tabanide (Diptera, Tabanidae) du Mozambique : Tabanus mossambicensis n. sp. Cah ORSTOM Série Entomol Médicale Parasitol. 1985;23:31–3. http://horizon.documentation.ird.fr/exl-doc/pleins_textes/cahiers/entomo/28394.pdf.

100. Portillo M. Diptera, Tabanidae. In: Ramos MA, Tercedor JA, Ros XB, Noguera JG, Sierra AG, Mayol EM, et al., editors. Fauna Iberica. 1st edition. Madrid: Museo Nacional de Ciencias Naturales. CSIC; 2002.

101. Venceslau L, Lopes L. Comparison of Automated Georeferencing Tools Using Insect Collection Data. Biodivers Inf Sci Stand. 2019;3:e37345. doi:10.3897/biss.3.37345.

102. Lopes L, Venceslau L, da Costa L. Citizen (Science) Involvement in Data Digitization and Enrichment of the Insect Collection of the Museu Nacional de História Natural e da Ciência (Lisboa, Portugal). Biodivers Inf Sci Stand. 2019;3:e36346. doi:10.3897/biss.3.36346.

103. Stork NE. How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annu Rev Entomol. 2018;63:31–45. doi:10.1146/annurev-ento-020117-043348.

104. Carranza-Rojas J, Goeau H, Bonnet P, Mata-Montero E, Joly A. Going deeper in the automated identification of Herbarium specimens. BMC Evol Biol. 2017;17:181. doi:10.1186/s12862-017-1014-z.

7. Annex A. Script used for geocoding with APIs

For APIs that require a username or key, these are not provided in this script.

```
library(httr)

#Import list of locations to test

locais <- read.csv("Locais_teste_BDinsecta.csv", header = T, stringsAsFactors = F)

lista_locais <- c()
for (n in 1:nrow(locais)){
  local <- paste(gsub(" ", "+", locais[n,8]), gsub(" ", "+", locais[n,7]), gsub(" ", "+", locais[n,6]), gsub("
", "+", locais[n,5]),
                 gsub(" ", "+", locais[n,4]), gsub(" ", "+", locais[n,3]), sep="+")
  local <- gsub(" ", "", local, fixed = TRUE)
  local <- gsub("++", "+", local, fixed = TRUE)
  print(local)
  lista_locais <- c(lista_locais, local)
}

########################
##Geonames
########################

locais_geonames <- locais
user <- ""

for(n in 1:length(lista_locais)){
  URL=paste0("http://api.geonames.org/searchJSON?q=", lista_locais[n],
"&maxRows=1&username=", user)
  response=GET(url=URL)
  if(content(response)$totalResultsCount!=0){
    locais_geonames$DwC.Decimal.Latitude[n] <- content(response)$geonames[[1]]$lat
    locais_geonames$DwC.Decimal.Longitude[n] <- content(response)$geonames[[1]]$lng

  }
}

#locais_geonames
write.csv(locais_geonames, "geonames-results.csv", row.names=FALSE)

########################
##Mapquest
########################

MapquestKey <- ""

locais_mapquest <- locais
```

```
for(n in 1:length(lista_locais)){
  URL=paste0("http://www.mapquestapi.com/geocoding/v1/address?key=", MapquestKey,
"&location=", lista_locais[n], "&outFormat=json")
  response=GET(url=URL)
  if(response$status_code!=400){
    locais_mapquest$DwC.Decimal.Latitude[n] <-
content(response)$results[[1]]$locations[[1]]$latLng$lat
    locais_mapquest$DwC.Decimal.Longitude[n] <-
content(response)$results[[1]]$locations[[1]]$latLng$lng
  }
}

#locais_mapquest
write.csv(locais_mapquest, "mapquest-results.csv", row.names = F)

########################
##Geolocate
########################

locais_geolocate <- locais

for (i in 1:nrow(locais_geolocate)){
  url <- paste0("http://geo-locate.org/webservices/geolocatesvcv2/glcwrap.aspx?Country=", gsub(" ",
"+", locais$DwC.Country[i]),
                "&Locality=", gsub(" ", "+", locais$DwC.Locality[i]),
                "&State=", gsub(" ", "+", locais$DwC.State.Province[i]),
                "&County=", gsub(" ", "+", locais$DwC.County[i]))
  response <- GET(url=url)
  if(content(response)$numResults!=0){
    print(paste("RESULT NUMBER:", i))
    locais_geolocate$DwC.Decimal.Latitude[i] <-
content(response)$resultSet$features[[1]]$geometry$coordinates[[2]]
    locais_geolocate$DwC.Decimal.Longitude[i] <-
content(response)$resultSet$features[[1]]$geometry$coordinates[[1]]
    locais_geolocate$DwC.coordinateUncertaintyInMeters <-
content(response)$resultSet$features[[1]]$properties$uncertaintyRadiusMeters

  }
}
write.csv(locais_geolocate, "geolocate-results.csv", row.names = F)

########################
##Google Maps
########################

locais_google <- locais

GoogleKey <- ""

for (i in 1:nrow(locais_geolocate)){
```

```
  url <- paste0("https://maps.googleapis.com/maps/api/geocode/json?address=", lista_locais[i],
"&key=", GoogleKey)
  response <- GET(url=url)
  if(content(response)$status!="ZERO_RESULTS"){
    locais_google$DwC.Decimal.Latitude[i] <- content(response)$results[[1]]$geometry$location$lat
    locais_google$DwC.Decimal.Longitude[i] <-
content(response)$results[[1]]$geometry$location$lng
  }

}

write.csv(locais_google, "googlemaps-results.csv", row.names = F)

######################
##OpenStreetMap/Nominatim
######################

locais_OSM <- locais

for (i in 1:nrow(locais_OSM)){
  url <- paste0("https://nominatim.openstreetmap.org/search/", gsub("\\+", "%20", lista_locais[i]),
"?format=json")
  response <- GET(url=url)
  if(length(content(response))!=0){
    locais_OSM$DwC.Decimal.Latitude[i] <- content(response)[[1]]$lat
    locais_OSM$DwC.Decimal.Longitude[i] <- content(response)[[1]]$lon
  }

}

write.csv(locais_OSM, "openstreetmaps-results.csv", row.names = F)
```

8. Annex B. Script used to clean csv files exported from Zooniverse after panoptes_aggregation

```
#File names
input_file <- "slider_extractor_lepidoptera-classification-extractions.csv"
output_file <- "dados_lepidoptera.csv"

#Read input file
dados <- read.csv(input_file, header = T, stringsAsFactors = F)
dados <- data.frame (user_name = dados$user_name, task = dados$task, task_description = dados$task_description, date = dados$created_at,
                     file = dados$filename, result = dados$data.slider_value, stringsAsFactors = F)

#Dimensions
nrows <- nrow(unique(data.frame(dados$file, dados$user_name, dados$date)))
ntasks <- length(unique(dados$task))
tasks <- dados$task_description[1:ntasks]

#Create ataframe to save extracted data
dados_ext = data.frame(matrix(nrow = nrows, ncol = ntasks+3))
colnames(dados_ext) <- c("file", "user", "date", tasks)

#Extract data
frow <- 1
for (i in 1:nrows){
  lrow <- frow+ntasks-1
  data <- dados[frow:lrow,]
  dados_ext$file[i] <- data$file[1]
  dados_ext$user[i] <- data$user[1]
  dados_ext$date[i] <- data$date[1]
  dados_ext[i,4:ncol(dados_ext)] <- t(data$result)
  frow <- lrow+1
}

#Save extracted data as csv
 write.csv(dados_ext, output_file,row.names = F)
```

9. Annex C. List of GBIF Tabanidae datasets per species

Tabanus monocallosus Travassos Dias, 1955
GBIF.org (04 September 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.eifena

Tabanus eggeri Schiner, 1868
GBIF.org (21 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.1wcwrs

Haematopota italica Meigen, 1804
GBIF.org (22 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.55jmqm

Tabanus autumnalis Linnaeus, 1761
GBIF.org (22 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.b1gomk

Tabanus sudeticus Zeller, 1842
GBIF.org (23 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.devdih

Tabanus bromius Linnaeus, 1758
GBIF.org (23 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.s5o8m5

Tabanus barbarus Coquebert, 1804
GBIF.org (04 August 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.d5kupq

Ancala fasciata (Fabricius, 1775)
GBIF.org (16 May 2019). GBIF Occurrence Download. https://doi.org/10.15468/dl.s8ilpo