Providing means for a better understanding of biodiversity: improving primary data and using it for threat assessment and in situ conservation planning in South America

Progress report No. 1

Submitted by the Decision and Policy Analysis (DAPA) program
International Centre for Tropical Agriculture (CIAT), Cali, Colombia

Andy Jarvis, Julián Ramírez, Louis Reymondin, Daniel Amariles, Hector Tobón, Jorge Camacho and Jhon Jairo Tello

Summary

The Inter-American Biodiversity Information Network (IABIN) has the mandate of providing a networking information infrastructure as well as primary biodiversity data on a number of topics in order to improve decision-making, particularly for issues at the interface of human development and biodiversity conservation. Attempts to correct geographic errors in IABIN Thematic Networks (TNs) were done and software packages based on the Java programming language were developed and tested. Documentation was also provided so that partners are able to use the tools we provide. Using these automated tools, we found that **14**% (~500,000 records) of the Species and Specimens Thematic Network (SSTN) has either not reliable coordinates or have no coordinates at all; whereas for the I3N database (for which we could only assess **14.3**% of the data owing to difficulties in the interpretation of coordinates), we found that **30.3**% have unreliable coordinates. All these data can be potentially georeferenced, though particular attention needs to be paid to the location data needed to retrieve coordinates. The Pollinators Thematic Network (PTN) database has not been yet assessed as it has not been delivered by TN partners. We were able to properly implement the georeferencing algorithm and apply it over various sets of test data. Our results indicate that in the vast majority of cases a reliable geographic reference can be retrieved from the biogeomancer service. In addition, we set up several trials to test the algorithm on its accuracy, but were not able to perform them yet given the fact that the service has been down during the last 2-3 weks. Particular attention needs to be paid to the processing time given the condition of the biogeomancer service as an online platform and optimisation pathways are currently under investigation.

Contents

1. Introduction

The Inter-American Biodiversity Information Network (IABIN) has the mandate of providing a networking information infrastructure as well as primary biodiversity data on a number of topics in order to improve decision-making, particularly for issues at the interface of human development and biodiversity conservation. IABIN currently provides access to scientific information existent throughout the world in different institutions. Currently, under a grant from the Global Environment Facility (GEF), IABIN is improving all the quality, the impact and the access of its data.

Currently, IABIN holds five thematic networks (TNs, hereafter), from which three provide point-based primary biodiversity data (occurrences). These thematic networks, namely:

1. The Species and Specimens Thematic Network (SSTN);
2. The Invasive Species Thematic Network (I3N); and
3. The Pollinators Thematic Network (PTN)

Hold a considerable amount of data. However, the three different thematic networks have different data structures, different standards, as well as different providers. In addition, the data has not been checked or verified in either quality or reliability, and more important, no attempt to correct the possible errors, homogenise the databases, or use them to in fact inform or guide decision-making processes has been done.

The International Centre for Tropical Agriculture (CIAT) has been funded within the GEF grant to **"Improve primary data and use it for threat assessment and in situ conservation planning in South America"**, which means, in a first instance, that the quality of the data within the different thematic networks will be assessed using a scientifically rigorous and mostly automated approach. Attempts to correct errors are also to be developed, and finally, the data will be used as input to a variety of modelling approaches intended to improve the knowledge on the vulnerability and conservation of biodiversity in South America. In a greater detail, as a whole the project will deliver:

1. Automatic cross-checking scripts
2. Automatic georeferencing scripts
3. Results on both the cross-checking and georeferencing of the TNs with suitable databases for such purposes (SSTN, I3N and PTN)
4. Automatic species distribution models training scripts
5. An assessment of level of both anthropogenic threats and conservation status for a number of species (for which geographic distributions can be modelled)
6. A Google-maps based navigation tool for all the modelling results (from [5])

In this report we provide a progress on deliverables (1), (2) and (3), and provide a detailed workplan for the next deliverables.

2. Developed scripts and documentation for data cleansing

One of the most relevant issues regarding the analyses and latter conclusions derived from the usage of primary biodiversity data is the reliance on its quality. Poor quality biodiversity data could lead to incorrect and biased conclusions as well as cause inefficient and/or wrong investment of the available resources and inadequate policy development. Checking of biodiversity data quality as well as using it adequately is a key issue in order to aid decision-making processes

We have built automated algorithms (Figure 1) developed in the Java programming language that allow a thorough coordinate verification process (error detection) and georeferencing process (error correction). Through this process we intend to develop an automated platform for IABIN's TNs to assess their own data whenever more data is incorporated on any of their databases.

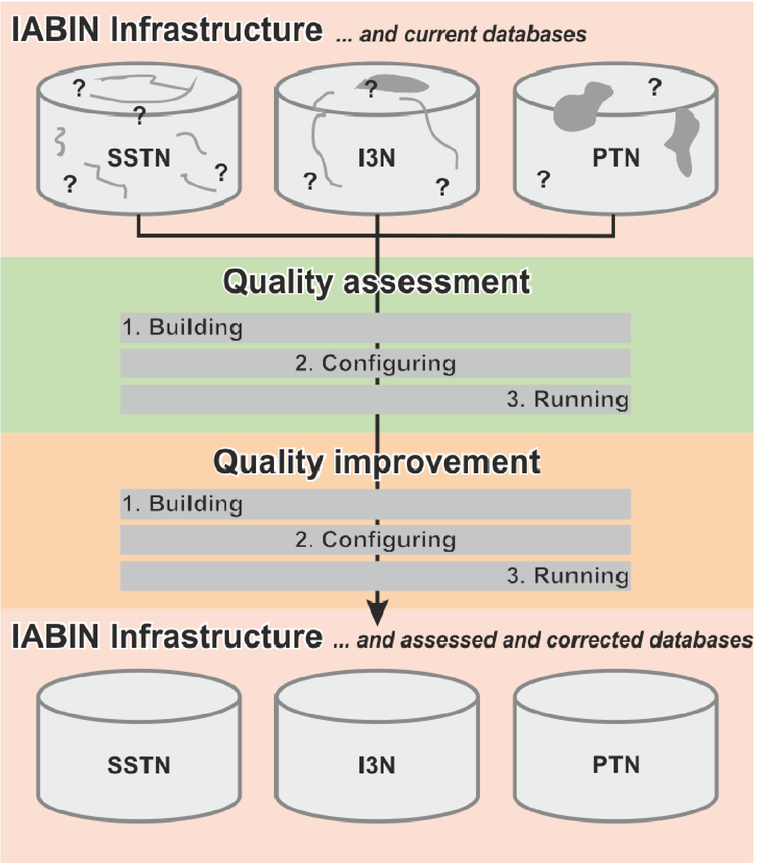


Figure 1 Assessment and improvement of IABIN TNs databases using automated algorithms

Adequate documentation on how to build, configure and run the tools is also provided in order to better manage knowledge transfer between and within TNs, and between TNs and project developers. A ***googlecode*** project has been therefore created with this aim (http://code.google.com/p/iabin-threats/), where all the source code and documentation is stored, and where bugs and problems can be raised as issues and discussed with developers.

2.1. Data-filtering and error cross-checking

Description

Common biodiversity data sources are observations from researchers, herbarium vouchers and genebank accessions, although there are also some additional observations such as archaeological findings and zoo living specimens. When these observations are done, detailed information is often recorded on the collection site, the type of observation (taxonomic identification), the surrounding habitat, and even the climatic conditions at the particular place of collection. With the development of Geographic Information Systems (GIS), satellites and Geographic Positioning Systems (GPS), currently, even the precise numeric location (i.e. coordinates) of the collection site can be retrieved. All this information is stored in primary biodiversity databases. However, due to the differences in data formats, institutional organization of the data, and due to common human errors in spelling of species names, mistyping information, big databases with primary biodiversity data often also have big errors. These errors turn very relevant when researchers intend to use the data for any scientific purpose. Here we are concerned about the geographic errors. Common errors in primary biodiversity databases are (1) misspellings of country, state, county and locality names, (2) swapping of latitude or longitude, (3) assignation of the value zero when missing data is found, (4) coordinates in different systems or unknown systems, (5) wrong usage of decimal places or truncation of all decimals, (6) usage of different coordinate formats (e.g. degrees-minutes-seconds vs. decimal degrees) without proper documenting, among others.

In view of that we have developed and implemented automated Java software that runs in batch mode and assesses primary biodiversity data from large databases automatically, and provides statistics on the quality of the data. Our software verifies geographic references (coordinates) at three different levels (Figure 2):

a. Continental level
b. Country level
c. Environmental level

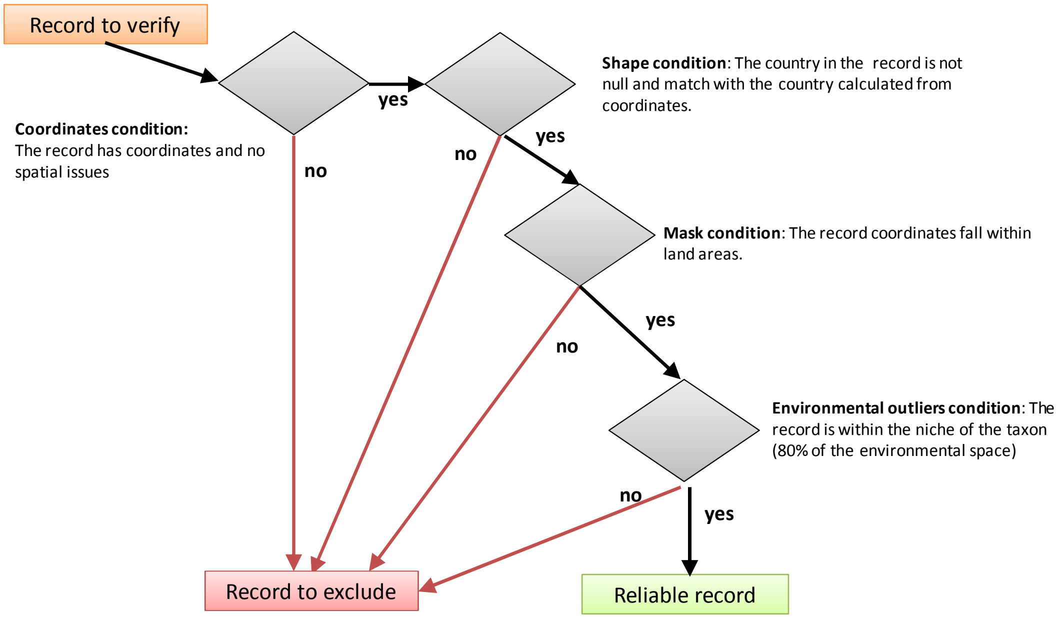

Record to verify
Coordinates condition:
The record has coordinates and no spatial issues
yes
no
Shape condition: The country in the record is not null and match with the country calculated from coordinates.
yes
no
Mask condition: The record coordinates fall within land areas.
yes
no
Environmental outliers condition: The record is within the niche of the taxon (80% of the environmental space)
no
yes
Record to exclude
Reliable record

Figure 2 Coordinate verification process

To verify at the continental level, we use a high resolution land areas mask from the SRTM Digital Elevation Model coastlines (Jarvis *et al.*, 2008); to verify at the country level, we use the data from the Global Administrative Areas (Hijmans, 2010); and to verify at the environmental level, we use the Tukey outlier test (Tukey, 1977) in a twenty-dimensional space given by 19 bioclimatic indices

(Ramirez & Bueno-Cabrera, 2009) derived from the WorldClim dataset (Hijmans *et al.*, 2005) and the elevation (Jarvis *et al.*, 2008). We flag a record as reliable only if it:

a. Is located in the country where it is reported to have been collected and/or observed
b. Being a record from a terrestrial species, falls within land areas, and
c. Is not flagged as an outlier for a given species in less than 80% of the 20 environmental variables used to describe the environment

Code and software provided

The software was coded in the Java programming language, and the **source code** is provided at http://code.google.com/p/iabin-threats/source/browse/#svn/trunk/data-filtering. A final working version to be compiled is provided at http://code.google.com/p/iabin-threats/downloads/detail?name=ita-0.1Beta-SNAPSHOT-src.zip&can=2&q= and at http://code.google.com/p/iabin-threats/source/browse/#svn/tags/ita%20v0.2_RC1_beta, an API compiled version is provided at http://code.google.com/p/iabin-threats/downloads/detail?name=ita-0.1Beta-SNAPSHOT-api.zip&can=2&q=, and a jar version is provided at http://code.google.com/p/iabin-threats/downloads/detail?name=ita-0.1-SNAPSHOT.jar&can=2&q=

Usage documentation

All the needed documentation in order to build the project is available as "wiki" pages, and are indicated as follows:

a. How to build the software (http://code.google.com/p/iabin-threats/wiki/HowToBuildITA): Contains all the information required to build the code from the source code (required libraries, configuration of the database, checking out of the source code, etc)
b. How to configure the software (http://code.google.com/p/iabin-threats/wiki/Configuration): Contains all the information regarding how to set up the database for data extraction, how to harvest the environmental and geographic data required, and how to configure the computer that will be used for running the software.
c. How to run the software (http://code.google.com/p/iabin-threats/wiki/RunningProcedure): Contains all the information required in order to perform the analysis of a particular database.

If any bugs are found, they can be issued through the ***googlecode*** project via http://code.google.com/p/iabin-threats/issues/list, by just clicking on the link "New issue" and then filling all the information requested.

2.2. Coordinate correction (georeferencing)

Description

Even when a coordinate verification process is implemented, it does not directly imply that the best-bet is being done from the available data. A quality improvement process is still required in order to be able to use the best-shaped data in further analyses. Geographically speaking, this means the retrieval of coordinates when they are either unavailable or erroneous. Using biogeomancer (Guralnick & Hill, 2009, Guralnick *et al.*, 2006, Hill *et al.*, 2009), one can retrieve the coordinates of a particular location by means of the collecting place information provided in the database (e.g. country, state, county, and locality name). Though the location information is not available in all the cases when also the coordinates are lacking, it is available in a number of cases and can be retrieved with a considerable degree of confidence (Hill *et al.*, 2009)

In this particular case, we have developed software that uses all the records flagged as not reliable (from the automatic data filtering and cross-checking), is capable of **(1) identifying the records** that have enough location information to retrieve a coordinate**, (2) querying the biogeomancer service** at http://bg.berkeley.edu/latest/, **(3) interpreting the result** from the biogeomancer service, and (**4) adding the retrieved coordinate** to the database.

Code and software provided

As the data filtering and cross-checking software, we have developed the georeferencing project in the Java language. The complete **source code** is at http://code.google.com/p/iabin-threats/source/browse/#svn/trunk/Biogeomancer, while the final working version is at http://code.google.com/p/iabin-threats/source/browse/#svn/tags/ita-bg%20v0.1_RC1_alpha, and a final compiled version ready-to-use as a ".jar" file is provided at http://code.google.com/p/iabin-threats/downloads/detail?name=ita-bg-1.0-SNAPSHOT.zip&can=2&q=.

Usage documentation

Again, all the documentation is provided online as "wiki" pages, which are also part of the whole ***googlecode*** project used to develop all the software. There are three wiki pages explaining the building (i.e. compilation), the configuration and the execution (running), as detailed below:

a. How to build the software (http://code.google.com/p/iabin-threats/wiki/HowToBuildITABG?wl=en): Contains all the information required to build the code from the source code (required libraries, checking out of the source code, etc)
b. How to configure the software (http://code.google.com/p/iabin-threats/wiki/Configurationbg?wl=en): Contains all the information regarding how to set uo the internet connection, the server, and any other requirements that make possible the usage of the software.
c. How to run the software (http://code.google.com/p/iabin-threats/wiki/RunningProcedureBG?wl=en): Contains all the information required in order to perform the analysis of a particular database.

In the same way as for the data filtering software, any bug (i.e. error or problem) can be reported via the ***googlecode*** repository (http://code.google.com/p/iabin-threats/issues/list)

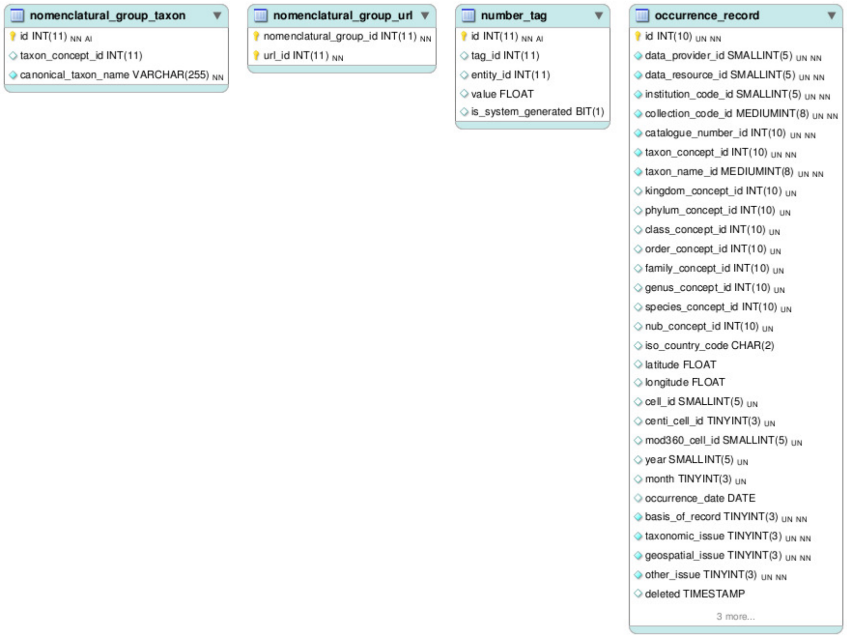
3. Data

As stated before, IABIN's TNs databases contain a bunch of data that requires a practical assessment and improvement in terms of its geographic quality. However, often a way one database is built (structured) does not work for other databases. In this section we provide an overview of the databases we have received and assessed.

3.1. Species and Specimens Thematic Network (SSTN)

The SSTN database has the same structure as the GBIF database and it was provided by the TN technical staff Maria Auxiliadora Mora on November 12th 2010. From the Entity-Relationship

Diagram it was clear that the main table (as with GBIF) was the "occurrence_record" table (Figure 3), from which we were able to extract all the required information in order to filter the data.



```
nomenclatural_group_taxon
id INT(11) NN AI
taxon_concept_id INT(11)
canonical_taxon_name VARCHAR(255) NN

nomenclatural_group_url
nomenclatural_group_id INT(11) NN
url_id INT(11) NN

number_tag
id INT(11) NN AI
tag_id INT(11)
entity_id INT(11)
value FLOAT
is_system_generated BIT(1)

occurrence_record
id INT(10) UN NN
data_provider_id SMALLINT(5) UN NN
data_resource_id SMALLINT(5) UN NN
institution_code_id SMALLINT(5) UN NN
collection_code_id MEDIUMINT(8) UN NN
catalogue_number_id INT(10) UN NN
taxon_concept_id INT(10) UN NN
taxon_name_id MEDIUMINT(8) UN NN
kingdom_concept_id INT(10) UN
phylum_concept_id INT(10) UN
class_concept_id INT(10) UN
order_concept_id INT(10) UN
family_concept_id INT(10) UN
genus_concept_id INT(10) UN
species_concept_id INT(10) UN
nub_concept_id INT(10) UN
iso_country_code CHAR(2)
latitude FLOAT
longitude FLOAT
cell_id SMALLINT(5) UN
centi_cell_id TINYINT(3) UN
mod360_cell_id SMALLINT(5) UN
year SMALLINT(5) UN
month TINYINT(3) UN
occurrence_date DATE
basis_of_record TINYINT(3) UN NN
taxonomic_issue TINYINT(3) UN NN
geospatial_issue TINYINT(3) UN NN
other_issue TINYINT(3) UN NN
deleted TIMESTAMP
3 more...
```

Figure 3 Main table of the SSTN database

The fields we were interested in were the id (of the record), the latitude and longitude, the iso_country_code (link to the country name), and the nub_concept_id (link to the taxonomy). In addition, we applied a basic filtering using the field geospatial_issue, as it is commonly used to report any problems with the data. We also used all the data in the table "taxon_concept", since it provides all the taxonomical information, mostly useful for modelling purposes (Figure 4).

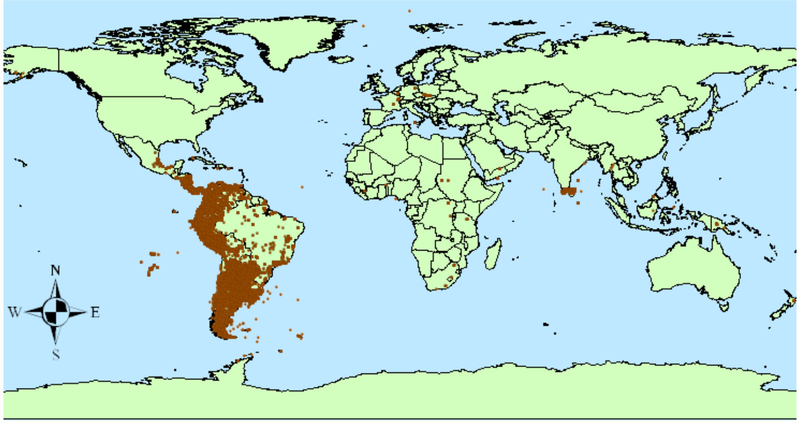


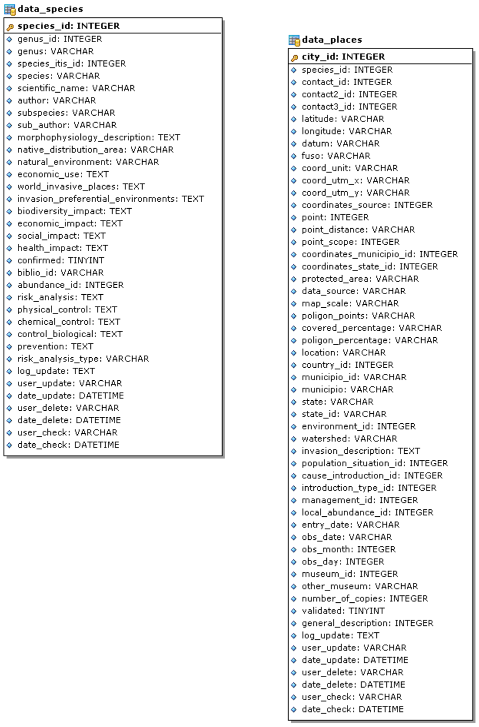
Figure 4 Occurrences to be analysed

Even if all the data points are supposed to be located in the Americas, there are some of them that appear to be located in Africa, Europe and India. These need to be further analysed and, if necessary, corrected.

3.2. Invasive Species Thematic Network (I3N)

On the 12th November 2010 a tele-conference was held with Alejandro Moreno, part of the technical staff from the I3N. On the 29th of November, he and the technical staff at I3N, provided us with access to the data. The I3N staff provided us with access to 8 different database, one for each of the following countries: Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador and Guatemala. The databases contain all the data in a different structure as compared to either SSTN or GBIF.

Alas, an Entity-Relationship model was not provided by the I3N, it was quite clear to us that two main tables contain the information we might require for our analyses (Figure 5). The table "data_places" containing all the occurrences (locations) where particular species have been observed, and the table "data_species", in which all the species descriptions are provided. There are also several tables referring to the different taxonomic levels related to a particular occurrence in the database.



```
data_species
species_id: INTEGER
genus_id: INTEGER
genus: VARCHAR
species_itis_id: INTEGER
species: VARCHAR
scientific_name: VARCHAR
author: VARCHAR
subspecies: VARCHAR
sub_author: VARCHAR
morphophysiology_description: TEXT
native_distribution_area: VARCHAR
natural_environment: VARCHAR
economic_use: TEXT
world_invasive_places: TEXT
invasion_preferential_environments: TEXT
biodiversity_impact: TEXT
economic_impact: TEXT
social_impact: TEXT
health_impact: TEXT
confirmed: TINYINT
biblio_id: VARCHAR
abundance_id: INTEGER
risk_analysis: TEXT
physical_control: TEXT
chemical_control: TEXT
control_biological: TEXT
prevention: TEXT
risk_analysis_type: VARCHAR
log_update: TEXT
user_update: VARCHAR
date_update: DATETIME
user_delete: VARCHAR
date_delete: DATETIME
user_check: VARCHAR
date_check: DATETIME
```

```
data_places
city_id: INTEGER
species_id: INTEGER
contact_id: INTEGER
contact2_id: INTEGER
contact3_id: INTEGER
latitude: VARCHAR
longitude: VARCHAR
datum: VARCHAR
fuso: VARCHAR
coord_unit: VARCHAR
coord_utm_x: VARCHAR
coord_utm_y: VARCHAR
coordinates_source: INTEGER
point: INTEGER
point_distance: VARCHAR
point_scope: INTEGER
coordinates_municipio_id: INTEGER
coordinates_state_id: INTEGER
protected_area: VARCHAR
data_source: VARCHAR
map_scale: VARCHAR
poligon_points: VARCHAR
covered_percentage: VARCHAR
poligon_percentage: VARCHAR
location: VARCHAR
country_id: INTEGER
municipio_id: VARCHAR
municipio: VARCHAR
state: VARCHAR
state_id: VARCHAR
environment_id: INTEGER
watershed: VARCHAR
invasion_description: TEXT
population_situation_id: INTEGER
cause_introduction_id: INTEGER
introduction_type_id: INTEGER
management_id: INTEGER
local_abundance_id: INTEGER
entry_date: VARCHAR
obs_date: VARCHAR
obs_month: INTEGER
obs_day: INTEGER
museum_id: INTEGER
other_museum: VARCHAR
number_of_copies: INTEGER
validated: TINYINT
general_description: INTEGER
log_update: TEXT
user_update: VARCHAR
date_update: DATETIME
user_delete: VARCHAR
date_delete: DATETIME
user_check: VARCHAR
date_check: DATETIME
```

Figure 5 Main tables of the I3N database

Details on how we treated these data in order to suit our needs for the data filtering software is provided in the sect. 4.1.2. We found that a considerable amount of the data from the I3N database has problems with the reliability, precision and/or accuracy of its coordinates (Figure 6)

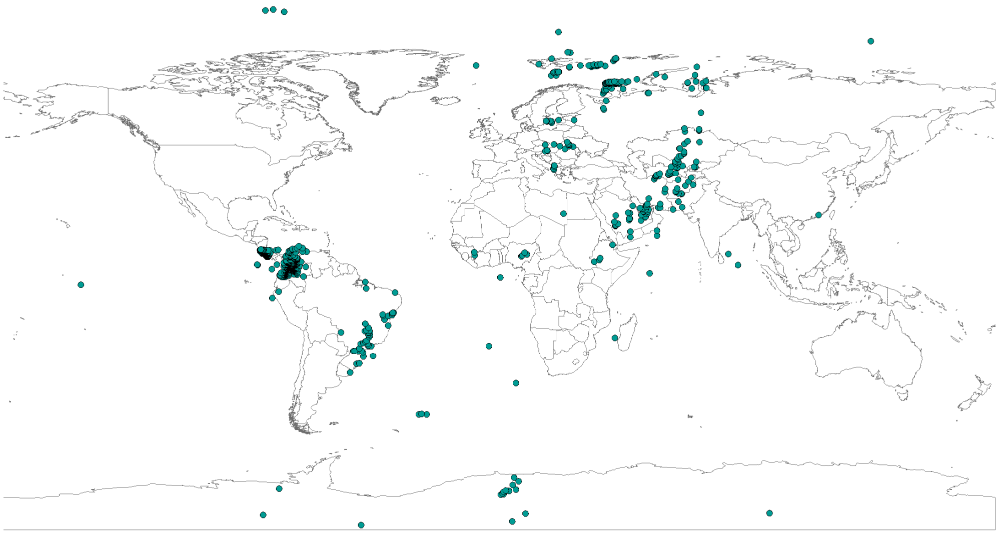


Figure 6 Geographic locations of all I3N occurrences

There are several records that appear to be wrongly located around the world, and these comprise about 50% of the data we were able to map (Figure 6). A greater proportion of the data remains unexplored, since a decent interpretation of the coordinates was not possible.

3.3. Pollinators Thematic Network (PTN)

Pedro Correa and Antonio Mauro Saraiva are part of the technical staff at IABIN in charge of the database management. Although we made several attempts to contact them, they seemed to be on travel and unable to properly respond emails, have any tele-conference, or provide us with the data. We still look forward to contacting them in order to gather and analyse the data.

4. Results

This section provides the main results of the data filtering with both the SSTN and the I3N databases, and also provides a proof of concept and validation of the biogeomancer services and analysis of their suitability for the improvement of the data stored both in SSTN and I3N.

4.1. Data-filtering and error cross-checking

We extracted the data from both the SSTN and the I3N databases and applied the coordinate cross-checking scripts over the whole datasets.

4.1.1. Species and Specimens Thematic Network (SSTN)

The SSTN database contains 3,866,145 occurrences, belonging to 87,600 taxonomies. These data are all in the “occurrence_table”. From these records, we selected all records with coordinates and with no geospatial issues (“geospatial_issue” field equals to 0), which gives us a total of 3,452,938 records.

Since our cross-checking algorithms can only be applied over occurrences that have been in fact recorded in terrestrial areas (i.e. all our spatial datasets are developed for land areas), we need to carefully select the set of species from the complete database that corresponds to terrestrial plants. However, since reviewing the taxonomical information can take a considerable amount of time and does imply a manual review of the actual data, we define a terrestrial species as a species for which 90% or more of the records have been observed (or are reported to have been observed) on terrestrial areas. After this basic filter, 3,441,589 records were available for the cross-checking (89%, Figure 7). **A cleansed version of the database was provided on November 23rd to our colleagues at the Conservation Biology Institute (CBI) for the integration in the Data Integration and Analysis Centre (DIAC).**

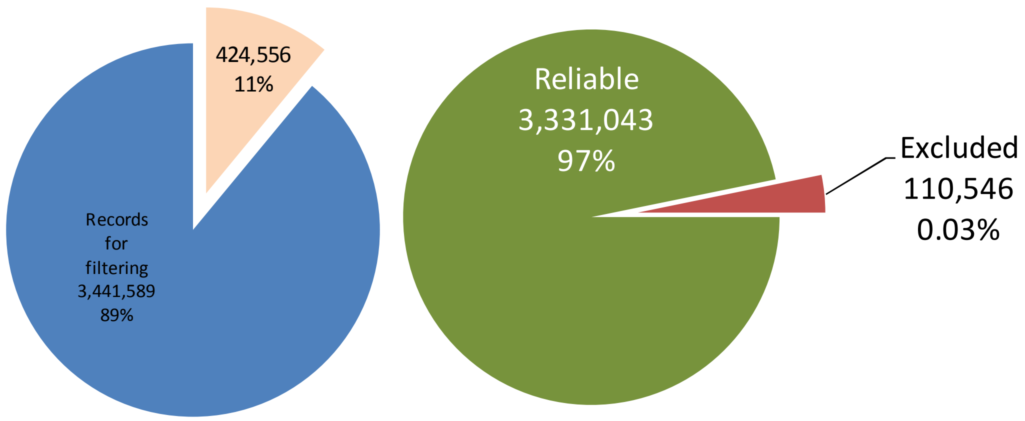


Figure 7 Records available for the filtering (left) and records considered as reliable from the total available for filtering (right)

We then applied our algorithms over these records and found that 97% of the records are accurate at the three cross-checking levels implemented. Yet more than a hundred thousand records were flagged as unreliable (Figure 7, right) due to different problems (Figure 8). The most common error was found when records were supposed to be located within land areas but they were found outside any continental boundary (though within a 5 km range). At least 45% of the data were attributed to this type of error, and it is very likely that this error is due to a deficient coordinate precision.

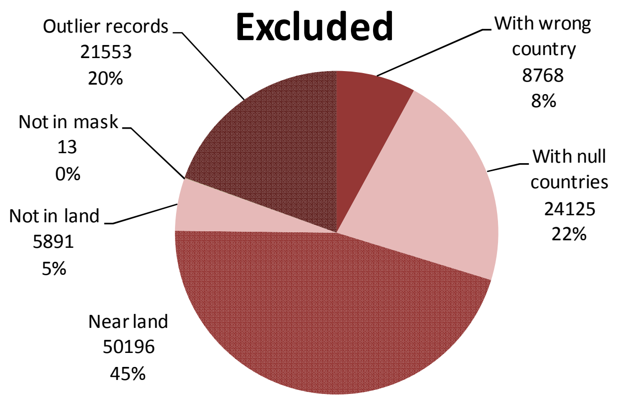


Figure 8 Proportions of records with different problems during the filtering

In addition, 42% of the errors in the data were attributed to not-reported (or not properly interpreted) country names (22%) and records outside the environmental niche of the species (20%). These three errors sum up to 87% of the total records that were flagged as not reliable. The least common error was found to occur when a record falls outside any logical latitude/longitude boundary (longitude below -180 or above 180 degrees, and latitude below -60 or above 90 degrees). Results were also mapped (Figure 9).

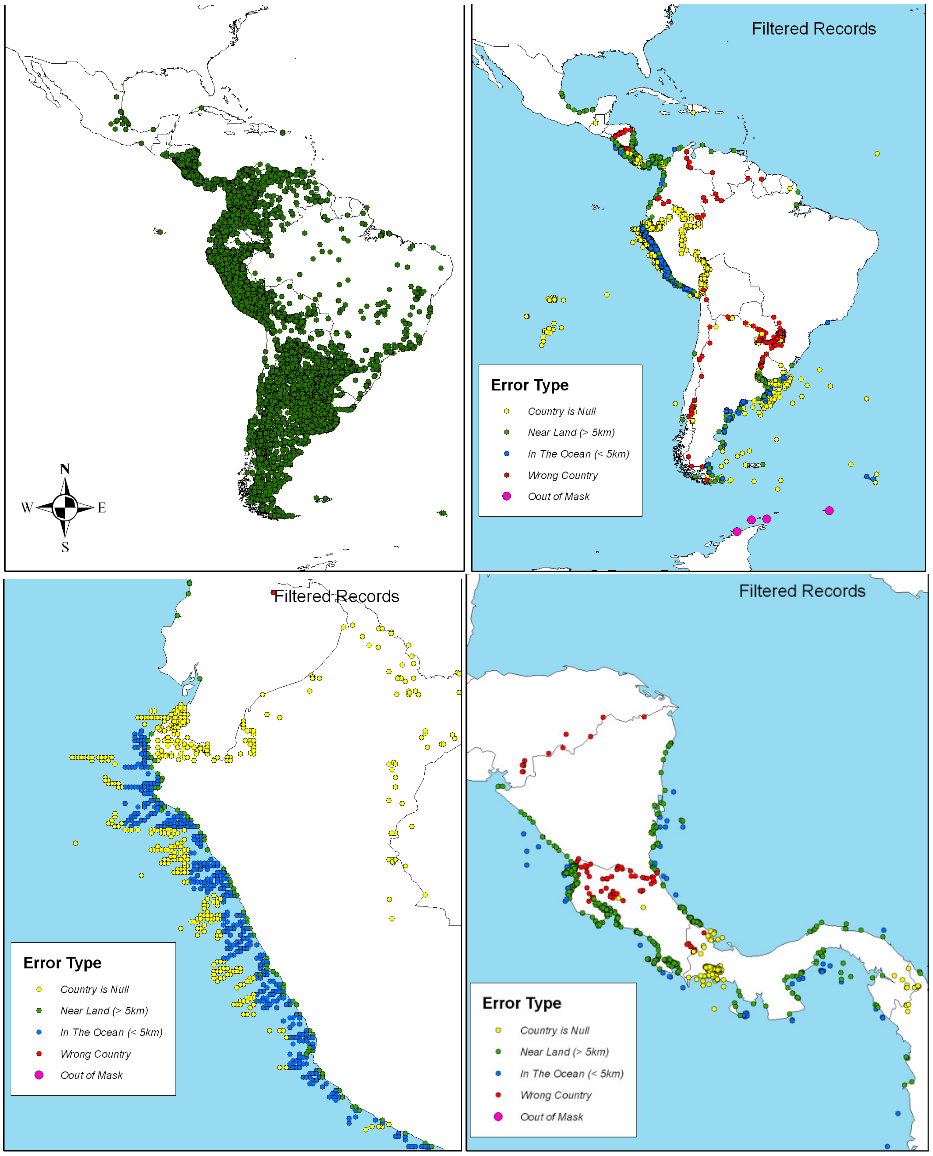


Figure 9 Errors found in the data analysed in the SSTN database. A) All records classified as reliable, B) all records presenting errors coloured by the type of error, C) an example of errors over Peru, D) an example of errors over Central America.

In summary, the SSTN database holds 433,207 records that have either a reported (i.e. known) geospatial issue or that do not have any coordinate reported, plus a total of 110,546 records that were found to have very poor quality according to our algorithm. This sums up some 14% of the total number of occurrences in the database for which coordinates can be retrieved and added to the database (although the availability of location data also needs to be assessed).

4.1.2. Invasive Species Thematic Network (I3N)

The I3N database contains **19,421** occurrences. These data are all in the "data_places" table.

We queried the information needed for the filtering; however, we realised that several issues were present, particularly in the fields referring to the coordinates. To start with, **there were two different fields for latitude** (or Y location) **and two different fields for longitude** (or X location). The original latitude and longitude fields (which we assumed refer to "interpreted latitude" and "interpreted longitude") were all empty, whereas the fields "coord_utm_x" and "coord_utm_y" were filled with a mix of data in different formats (Table 1). Some of the data were in decimal degrees, others were in degrees, minutes and seconds (without indicating if the record was North [N] or South [S] in the case of latitude, or East [E] or West [W] in the case of longitude), others had numbers that appear to be in meters (somehow in UTM projection, though the projection information is not available), and there's also a mixture between characters and numbers within these fields.

Table 1 Various data included in the fields latitude (coord_utm_y) and longitude (coord_utm_x)

id	nub_concept_id	coord_utm_y	coord_utm_x	iso_country_code
675	2	24Ã,Âº 06' 36''	47Ã,Âº 56' 11''	BR
3818	20	7532979	221878	BR
5594	41	6515987 - 6568941	482259 - 523511	BR
4507	3	S 12Ã,Âº 31' 44"	W 41Ã,Âº 33' 32"	BR
11587	140	343668	7072984	BR
5593	41	6515987 - 6568941	482259 - 523511	BR
10429	287	77 70 00	37 20 00	BR
403	20	7205466	754352	BR
907	74	1984300	4277700	BR
10640	348	7756179	361322	BR
10530	72	77 03 549	26 03 53	BR
501	98	1971200	4273200	BR
8975	44	6763175 - 6780370	580298 - 595840	BR
222	154	23Ã,Âº 55'	23Ã,Âº 55'	BR
10884	383	20 40 S	46 19 W	BR
10071	307	25 30	53 25	BR
10076	95	25 32	53 29	BR
10363	157	7744150 - 7740650	360292 - 362792	BR

Interpreting these types of data a hard task, since even the names of the fields do not give a concrete clue of what kind of information they might contain. We suggest the I3N staff to better work in a proper interpretation of these data before a better cleansing can be done (we have done it at the extent possible). If the data do correspond to latitude and longitude values in degrees, minutes and seconds, these must be separated in different fields, or properly and uniformly formatted for easy interpretation of any algorithm. If the data are in UTM (Universal Transverse Mercator) system, then the projection (zone and datum) needs to be specified for each of the data points or at least for groups of data points.

However, we used the following rules:

1. We used all occurrences having both an "S"/"N" (for latitude) and an "E"/"W" for longitude
2. Within these occurrences, we selected those having at least two different numbers separated either by an space or by a degree sign ("º")
3. For these, we calculated the latitude and longitude as:

$$Lat\text{¦}Lon = degree + \frac{minutes}{60} + \frac{seconds}{3600}$$

Being negative for latitude when South [S] is indicated and negative for longitude when West [W] is indicated, otherwise positive.

Using this approach, we were able to properly interpret 2,776 records out of the 2,860 that had coordinates (14.3%, Figure 7, left). We assessed all these records for their geographic consistency and found that 10% (1,934) of the data in the I3N network is reliable (Figure 10, right).

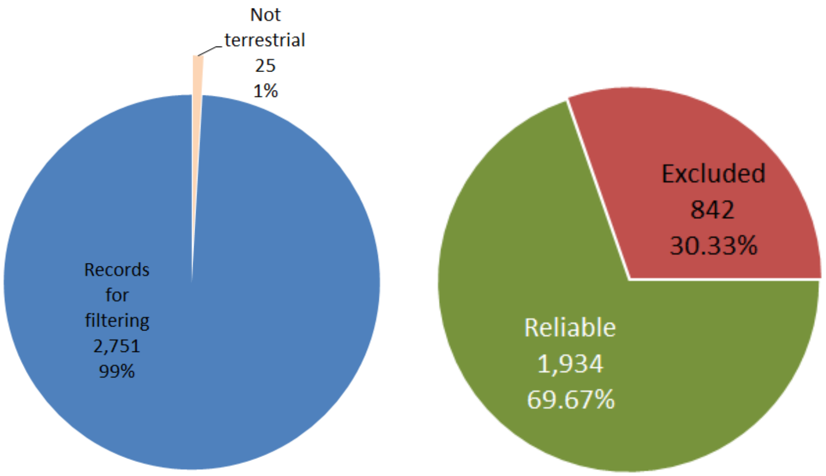


Figure 10 Records available for the filtering (left) and records considered as reliable from the total available for filtering (right) for the I3N database

The most common error was when an occurrence was located in a wrong country and when an occurrence was falling outside land areas (Figure 11), which accounted to up to 96% of the data errors, whereas the least common error was when a record was found outside the geographically analysable domain (1%). Results were also mapped (Figure 12)

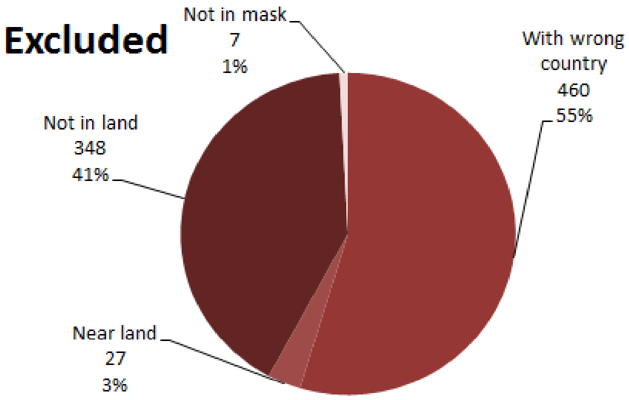


Figure 11 Proportions of records with different problems during the filtering

The I3N database showed overall less quality, but this might be due to the fact that the interpretation of the coordinates was somehow tricky. Better coordination within the I3N needs to be fostered in order to improve the interpretation of latitude and longitude fields.

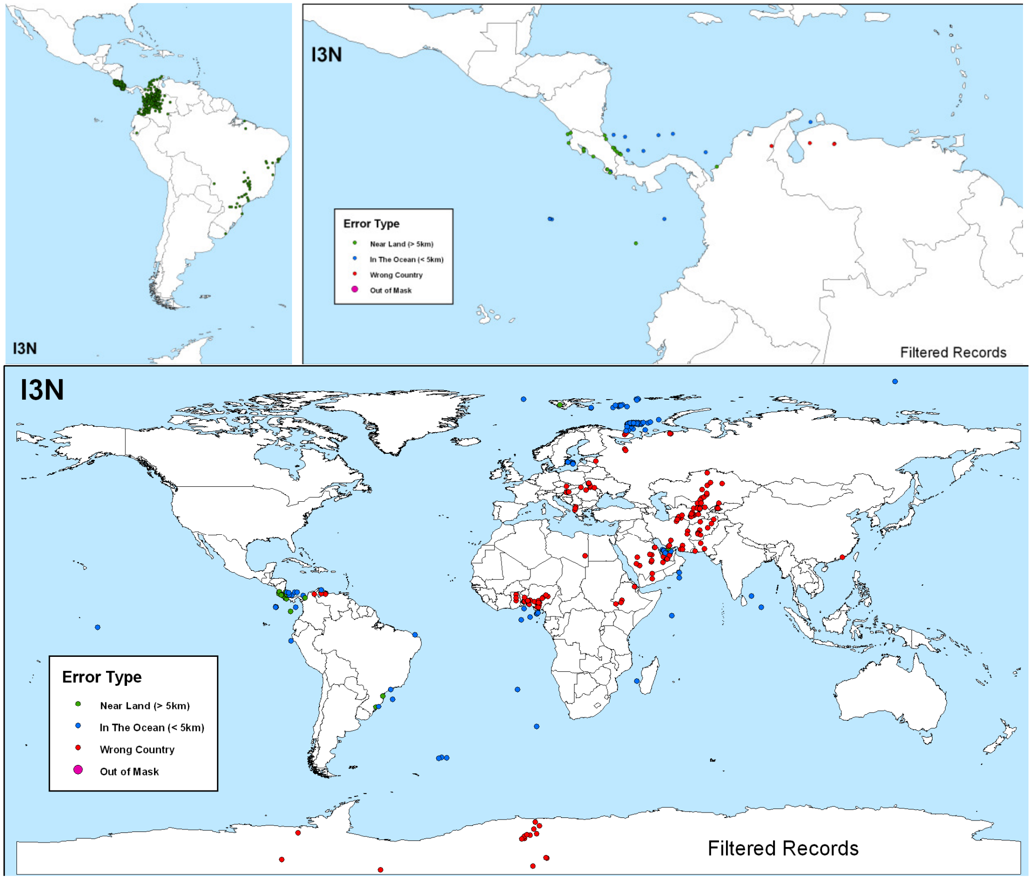


Figure 12 Errors found in the data analysed in the I3N database. A) All records classified as reliable, B) an example of errors over Central America and Northern South America, and C) all records presenting errors coloured by the type of error

In summary, the I3N database holds **16,561** (85.2%) records that couldn't be assessed due to the difficulties on interpreting something meaningful (in an automated way) from the coordinates provided. An additional **842** records were flagged as not geographically reliable, and are thus potentially useful for georeferencing using biogeomancer. The proportion of not reliable records in relation to the total analysed was higher than the SSTN (**30.3**%), but is in fact the interpretability of the geographic coordinates the most important issue with the I3N database.

4.1.3. Pollinators Thematic Network (PTN)

Since we have not yet received any data from the PTN staff, we were not able to obtain any result, and worse than that, we still do not know which kind of issues we might have to deal with.

4.1.4. Summary

The SSTN database, with a pretty much similar structure to the GBIF database (with which he have plenty of experience) holds up to **3,866,145** records, from which **433,207** records have either a reported (i.e. known) geospatial issue or that do not have any coordinate reported, and a total of **110,546** records were found to have very poor quality according to our algorithm. We estimate that around 14% of the dataset can be georeferenced, although this might depend on the availability of location information.

The I3N database (split in 8 different country-specific databases) has a completely different structure. Latitude and longitude fields in the database were somehow tricky formatted and they were difficult to interpret. We were able to assess **14.3**% of the I3N data and found that **30.3**% of those have quality issues, being locations in wrong countries the main problem.

The PTN database has not still been received from the TN technicians, and therefore, was not assessed at all.

4.2. Coordinate correction (georeferencing)

We developed an automated script for the correction and retrieval of new coordinates in order to improve the quality of the current IABIN TNs databases. Since the databases arrived somehow late and several issues arose while trying to interpret and manage the databases themselves, we were not able to apply the georeferencing scripts over the bunch of data flagged as not-reliable. However, we're currently on the way of doing it. Processes are running locally and results will be obtained and analysed within the next 15 working days.

In spite of that, one very key thing with something as an automated georeferencing is the processing time, the reliability of the retrieved information and our ability to properly track the problems and in fact improve the data (rather than only adding meaningless information to it). We made several tests in order to test (1) the processing time, and (2) the reliability of the algorithm at three different levels (a) location of a record, (b) correction of a record flagged with mask-type error, and (c) correction of a record with a country-type error.

4.2.1. Processing time, main issues and solutions

Since the biogeomancer service is an online platform, it is considerably sensitive to several affections such as the amount of information asked in a single query and the time it takes to process and retrieve one single record. We found that, due to the time the remote biogeomancer server takes to respond a single query, it might process around 4 records per minute (Figure 13)

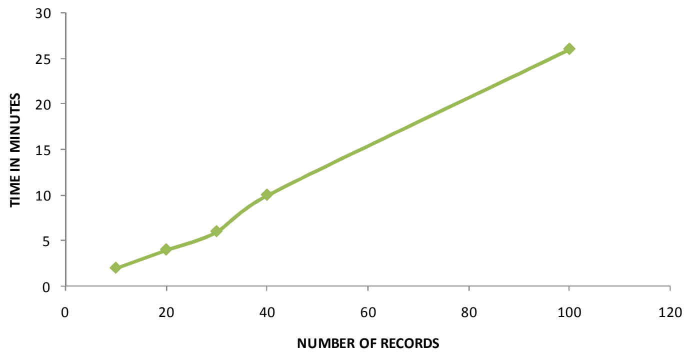


Figure 13 Time required to process different amounts of records as measured

For the processing of around 1,000 records, we need to currently spend some 4.3 hours, whilst for the processing of 100,000 records we might need around 18 days. However, these are only taking into account a single query, and several optimisation schemes can be achieved by finding the optimum number of records that can be queried at a time, and also using parallel processing schemes. All these are currently being explored.

4.2.2. Proof of concept: reliability of the algorithm

In order to test the reliability of the algorithm, we conducted a general background test and three specific experiments. The general background experiment intended to show how many records might be actually retrieved with a coordinate from a set of records randomly selected. Records with enough locality information were selected for this purpose (Figure 14).

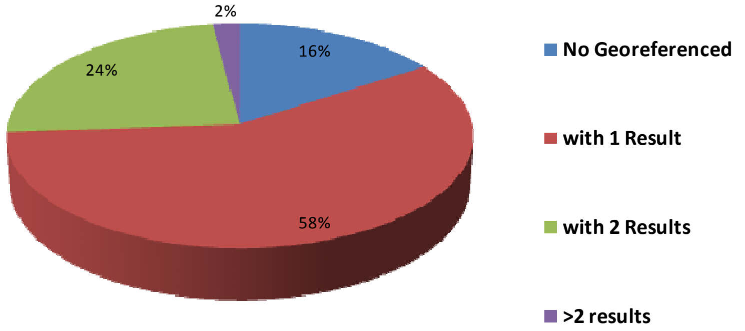


Figure 14 Proportion of times a result is retrieved from the biogeomancer service

We found that 16% of the times, the service might fail to locate the occurrence, even when providing enough locality information to locate the record. This might happen because the locality does not exist in the biogeomancer database, or because the biogeomancer service was not able to properly interpret the locality names, or the state or county names. In some cases more than one result was provided (26%), and this might happen because several localities with the same name might occur in the same country or area. In these cases, further investigation and review of the georeferences retrieved by our algorithm needs to be done by database curators.

Uncertainties in locations are also an issue that arises from the usage of a service such as biogeomancer (Figure 15). We found that some 50% of the times, uncertainty in the coordinates retrieved is considerably high (>10 km), and that in 13% of the times the coordinate was not found. In all the other cases, however, uncertainties remain below the 10km limit, which is a decent limit that allows performing further geographical or modelling approaches with the data.

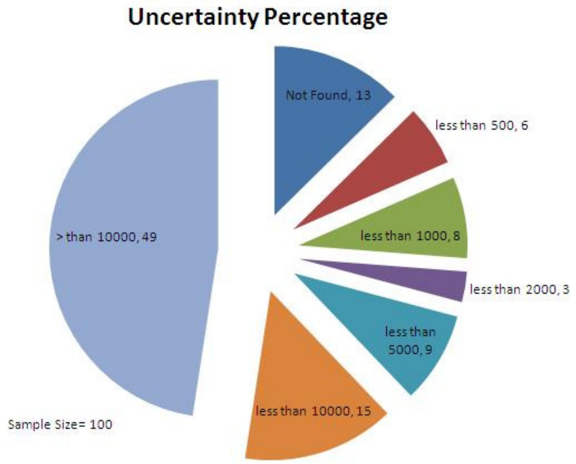


Figure 15 Uncertainties in the georeferencing process

4.2.2.1. Are we able to correctly locate a record?

In order to test the validity of the georeferences provided by the service, we conducted a simple test. We first selected a group of data (1) flagged as geographically reliable, (2) evenly geographically distributed (50 points distributed along different countries), (3) with coordinates, and (4) with enough location data in order to retrieve georeferences from biogeomancer (country, state, county and locality names). We queried the biogeomancer service with the locality information and gathered the coordinates to then compare the results between the original and retrieved coordinates.

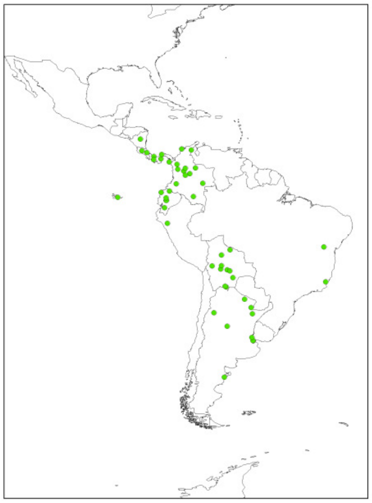


Figure 16 Original locations of the data to be used in the accuracy test

4.2.2.2. Are we able to correct mask-type errors?

We tested the ability of the service to correct records that we flagged as errors during our coordinate cross-checking processes. To this purpose, we selected records that were previously flagged as not reliable because they were found to be located outside land areas (i.e. mask-type errors) that (1) had enough location information for retrieving coordinates and (2) were evenly distributed among the countries. We corrected the geographic references and drew maps of "before" and "after" the georeferencing process.

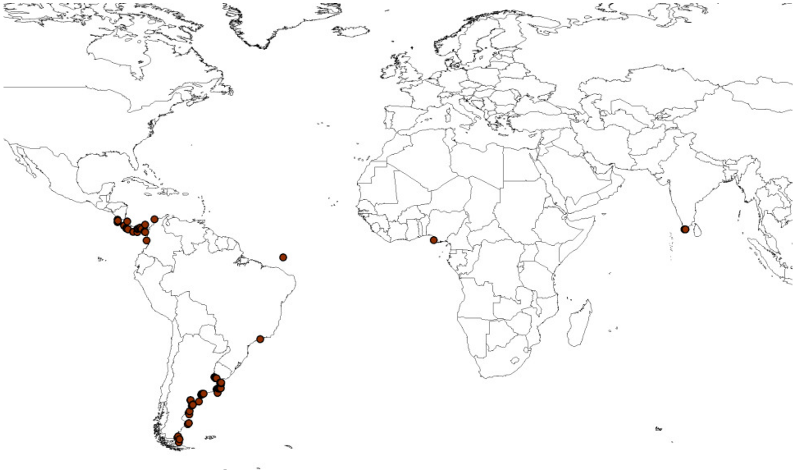


Figure 17 Original locations of the data to be used in the mask-type error correction test

4.2.2.3. Are we able to correct country-type errors?

We tested the ability of the service to correct records that we flagged as errors during our coordinate cross-checking processes. To this purpose, we selected records that were previously flagged as not reliable because they were found to be located in the wrong country (i.e. country-type error) that (1) had enough location information for retrieving coordinates and (2) were evenly distributed among the countries. We corrected the geographic references and drew maps of "before" and "after" the georeferencing process.

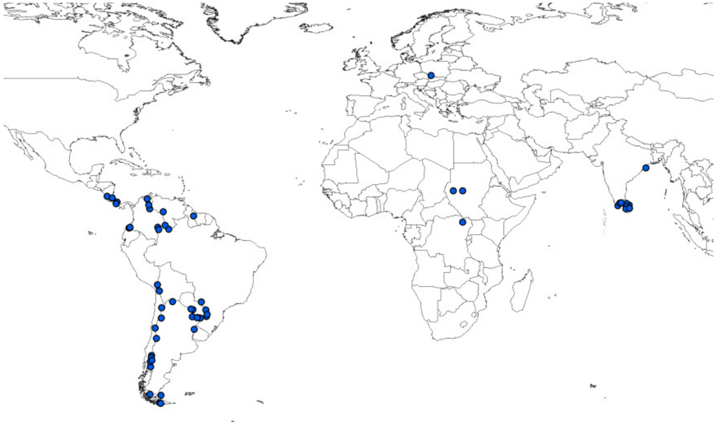


Figure 18 Original locations of the data to be used in the country-type error correction test

4.2.3. Summary

We were able to properly implement the georeferencing algorithm and apply it over various sets of test data. Our results indicate that in the vast majority of cases a reliable geographic reference can be retrieved from the biogeomancer service, though attention must be paid in those cases where more than one georeference is obtained from the online platform. In addition, we set up several trials to test the algorithm for consistency, though we were not able to perform them due to unavailability of the service during the last 2-3 weeks. Particular attention needs to be paid to the processing time given the condition of the biogeomancer service as an online platform.

5. Conclusions

Several issues arose from the gathering of the primary biodiversity data from IABIN TNs. First of all, we were not able to gather all the data from all TNs: PTN data is still on its way. On the other hand, I3N data was in a completely different format as compared to SSTN data (which fairly easy to assess), and even more important than that, proper documentation was not available for the I3N database, plus the latitude and longitude fields contained information that was fairly complicated to interpret and use. Although, we made several attempts to interpret and use these data and found that in general the data had less quality in comparison with SSTN data.

We also tested and implemented an automated geographic referencing approach and were successfully able to apply it over reduced datasets in order to test its reliability and ability to retrieve correct geographic references. There are still a number of issues accounted to the processing times required, and to the amount of data that can be sent at a single time as a query to the server, but several improvements and optimisations are being implemented to cope with these problems.

6. Brief summary on next steps

Additional tasks have arisen from the interactions with the Conservation Biology Institute (CBI). In addition to the current appointed deliverables, which include delivering the filtered and georeferenced databases to CBI, we have committed to merge these databases with the filtered data from GBIF corresponding to the Americas. This is planned to be delivered before the end of the year.

Significant improvement is suggested to the I3N database so that a better interpretation of the coordinates can be done. Up to now, only a limited percentage of the data could be assessed, and even these data need to be reviewed for inconsistencies in our interpretation (which sometimes remain as a bare guess). Some additional interpretation algorithms can be investigated and implemented; however, this might require additional time.

For the time being, we provide a working code and ready to use tools as well as installation, configuration and execution manuals. We suggest TN partners to start reading the manuals and provide us with the necessary feedback for us to improve both the tools and the documentation.

As an additional note, we are now applying the whole modelling procedure over IABIN terrestrial records, whilst at the same time performing several tests in order to set up the web interface required for the visualisation of modelling results. We plan to have all the modelling work finished by the end of January, whilst the first version (beta) of the web interface will be ready over the first week of March. Follow up on these activities will be done accordingly.

References

Guralnick R, Hill A (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics,* **25**, 421-428.

Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ, The Biogeomancer Working G (2006) BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. *PLoS Biol,* **4**, e381.

Hijmans RJ (2010) Global Administrative Areas (GADM). (ed Zoology MOV) pp Page, Berkeley.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* **25**, 1965-1978.

Hill A, Guralnick R, Flemons P *et al.* (2009) Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics,* **10**, 1-9.

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled seamless SRTM data V4. (ed (Ciat) ICFTA) pp Page, International Center for Tropical Agriculture (CIAT).

Ramirez J, Bueno-Cabrera A (2009) Working with climate data and niche modeling I. Creation of bioclimatic variables. pp Page, Cali, Colombia, International Center for Tropical Agriculture (CIAT).

Tukey JW (1977) *Exploratory Data Analysis*, Addison-Wesley.