Natureza & Conservação
Brazilian Journal of Nature Conservation

Research Letters

Natureza & Conservação 10(2):139-144, December 2012

Handling Editor: Paulo De Marco Jr
http://dx.doi.org/10.4322/natcon.2012.018

The Real Task of Selecting Records for Ecological Niche Modelling

Renato De Giovanni[1]*, Luís Carlos Bernacci[2], Marinez Ferreira de Siqueira[3] & Flávia Souza Rocha[4]

[1] *Centro de Referência em Informação Ambiental – CRIA, Campinas, SP, Brazil*
[2] *Instituto Agronômico de Campinas – IAC, Campinas, SP, Brazil*
[3] *Instituto de Pesquisas do Jardim Botânico do Rio de Janeiro – JBRJ, Rio de Janeiro, RJ, Brazil*
[4] *Universidade Federal Rural do Rio de Janeiro – UFRRJ, RJ, Brazil*

Abstract

Biological collections evoke contrasting feelings for being such a vast source of biodiversity data which is prone to all sorts of errors and uncertainties. The situation is not different for Brazilian herbaria, currently sharing more than two million easily accessible records on the Web. Properly dealing with this reality is a crucial task when using this kind of data for ecological niche modelling (ENM), so that errors and uncertainties do not generate misleading results in conservation. Here we investigate some of the issues that can be found in herbarium specimen data, describing a set of automatic procedures that can be used for a prior selection of records for ENM. In total, 11531 records for 135 species of Passifloraceae that natively occur in Brazil were analyzed considering different spatial resolutions, ranging from 30 arc-seconds to 10 arc-minutes. After applying the procedures, the proportion of spatially unique records was 9.3% for the highest resolution considering all species, with an average number of 8 records selected per species. These numbers increased to 17% and 16, respectively, for all other resolutions. This scenario highlights the importance of using data quality filters and further developing ENM presence-only methods that can work with a low number of records per species. Automatic procedures still cannot discard expert review, but they can greatly facilitate it by drawing attention to a much smaller number of records potentially useful for ENM. Most of the data quality procedures described here can also be applied to other taxonomic groups, regions and specimen data sources.

Key words: Biological Collections, Data Quality, Species Distribution Modelling.

Introduction

Ecological niche modelling (ENM) has attracted substantial interest during the last years, producing a remarkable growth in the number of published papers (Lobo *et al*. 2010). This can be mainly accounted to its broad applicability and the increasing number of related data and tools (Peterson *et al*. 2011). The most common method used in ENM is known as the correlative approach (Soberón & Peterson 2005), where spatially explicit species occurrence data and environmental data are combined so that specific algorithms can build a mathematical representation of the species' niche. Such models can be used to understand and predict species occurrences under different scenarios, which is why they have been extensively used in conservation planning and biodiversity management (Peterson *et al*. 2011).

***Send correspondence to:** Renato De Giovanni
Centro de Referência em Informação Ambiental – CRIA,
Av. Dr. Romeu Tórtima, 388, CEP 13084-791,
Campinas, SP, Brasil
E-mail: renato@cria.org.br

Biological collections are one of the main sources for species occurrence data. Current estimates indicate that they may contain between 1.2 and 2.1 billion specimens collected from nature during the last centuries (Ariño 2010). Access to this information is being improved as more records are digitized and shared with biodiversity data networks. In Brazil, the *species*Link network (Canhos *et al*. 2004) has recently reached 4 million records from biological collections and is steadily growing at half million records per year. Considering only plant specimens from Brazilian collections, *species*Link currently provides access to ~2.3 million records, corresponding to approximately 38% of all holdings stored in Brazilian herbaria (assumed be around 6 million according to Egler & Santos 2006).

Nevertheless, there are many practical issues that need to be carefully addressed when ENM is performed with data from biological collections. Such issues may include spatial errors and uncertainties, nomenclatural issues, identification conflicts, among others (Soberón & Peterson 2004; Chapman 2005). High positional uncertainties can

degrade model performance (Fernandez *et al*. 2009), while specimen identifications in biological collections are subject to many nomenclatural and taxonomic variations (Soberón & Peterson 2004) making it difficult to find all existing records for a certain species and to correctly deal with those that can easily be found. For plant species, in particular, additional care must be taken to distinguish between specimens collected from cultivated individuals and those collected from the wild. By overlooking these issues, models may be trained with incorrect environmental data or, even worse, with data from different species, generating misleading results.

Here we 1) investigate some of the taxonomic and spatial uncertainties that can be associated with specimen records in Brazil; 2) describe a series of data quality filters and additional strategies that can be used to select herbarium records for ENM and; 3) examine the result of these filters for different spatial resolutions by testing them in records of Passifloraceae that natively occur in Brazil. Although we used a single botanical family and a specific country to test the procedures described here, most principles that guided record selection for ENM are general enough to be used with other taxonomic groups and geographical regions. Therefore, this work can also be seen as a more general contribution to disseminating best practices for using biological collections' data in ENM.

Material and Methods

Dealing with species names

We started by using the currently accepted scientific names for all species of Passifloraceae according to the official Checklist of the Brazilian Flora (Cervi *et al*. 2010). First, each accepted name was searched in the Tropicos database (Tropicos 2011) to find possibly related:

- Incompatible homonyms: Names published with the same binomial of an accepted name but referring to a different species. For example, *Passiflora rubra* L. is currently an accepted name, but *Passiflora rubra* Vell. corresponds to *Passiflora organensis* Gardner;
- Incompatible varieties: Varieties sharing the same binomial with an accepted name but referring to another species after taxonomic revision. For example, *Passiflora amethystina* J.C.Mikan is currently an accepted name, but *Passiflora amethystina* var. *bolosii* Cervi corresponds to *Passiflora loefgrenii* Vitta;
- Unambiguous synonyms: Different binomials that can be used to search for specimen records of the same species, as long as they do not have any incompatible homonyms.

Additional synonyms were identified after examining other sources such as the checklist of the Brazilian Flora itself and literature (Killip 1938; Cervi 1997; Bernacci 2003).

Finally, name misspellings were discovered by interacting directly with the source of specimen data – the *species*Link network. This was achieved by browsing all distinct specific epithets associated with specimen records for each genus involved (*Ancistrothyrsus*, *Dilkea*, *Mitostemma* and *Passiflora*). Each detected typo or misspelling was followed by confirmation in the Tropicos database that it did not correspond to a published name.

Filtering specimen data

Occurrence data was retrieved from *species*Link using the genus and specific epithet from the accepted name and from alternative names. Records from species that had incompatible homonyms were automatically tagged if no author was provided in the identification or if the author did not match the expected content according to the Checklist of the Brazilian Flora. Specimen records named with incompatible varieties were associated with the current species they refer to. All other records identified at the infraspecific level were considered valid records for the species with the corresponding binomial.

The following sources of positional uncertainty were determined: precision error, datum error and whether retrospective georeferencing by municipality was used. Precision errors were calculated according to Wieczorek *et al*. (2004) using the original verbatim coordinates whenever possible instead of the decimal values provided by *speciesLink*, as the latter may not reflect the original precision. Datum errors were calculated since none of the collections seemed to store this information. Three options for horizontal datum were considered: WGS84, SAD69 and "Córrego Alegre". The overall location uncertainty was calculated as the precision error added to the datum error following Wieczorek *et al*. (2004). In the few occasions when a positional uncertainty was already specified by the data provider, this value was used instead of the calculated one.

Retrospective georeferencing by municipality was detected by determining the distance between the record coordinates and the municipality coordinates provided by a standard gazetteer (IBGE 2003). If the distance was less than 50 m, the record was tagged as being georeferenced by municipality. The same gazetteer also provided georeferenced shapes for each municipality, from where uncertainties associated with retrospective georeferencing were calculated by finding the maximum distance between the municipality coordinates and its borders. The shapes were also used to verify whether the coordinates of each record were within the boundaries of the specified municipality. If they were more than 2 km away outside the borders (added to the positional uncertainty), the record was tagged as having

conflicting georeferencing information. All distances were calculated in Spherical Mercator projection.

The following conditions were also checked: records without coordinates, records with coordinates outside the valid range or when both values are equal to zero, records collected or observed outside Brazil, observation or living collection records (only vouchered specimens were used) and records identified as not being collected from native habitat. This last condition was automatically detected by searching for sequences of characters ("cultiv", "procedente" or "procedência") in the observations field as an initial approximation. Additionally, records with the same collector name and collector number were considered duplicates, in which case only one was used in the final selection based on its location quality (prioritizing records with no data conflicts and with the lowest uncertainty). Upon divergence of identification among duplicates, all of them were tagged with an identification conflict.

The final number of records potentially useful for ENM was calculated for each spatial resolution currently available in WorldClim environmental data (Hijmans *et al*. 2005), which is extensively used by researchers: 30 arc-seconds, 2.5 arc-minutes, 5 arc-minutes and 10 arc-minutes. For this specific purpose, records without coordinates but indicating the municipality of the collecting event were associated with the standard municipality coordinates and its corresponding positional uncertainty. A maximum acceptable uncertainty was used for each resolution: 500 m, 2500 m, 5000 m and 10000 m, respectively, which corresponds to approximately half the environmental resolution at the Equator.

Results

A total number of 135 accepted names for species of Passifloraceae were found in the checklist of the Brazilian Flora. Only one of them was not found in the Tropicos database (*Passiflora boticarioana* Cervi). For all other names, 5 incompatible homonyms, 14 incompatible varieties and 23 synonyms were found in Tropicos (see Tables S1, S2 and S3 in the additional supporting information available at www.abeco.org.br). Additionally, 45 misspellings and orthographical variants were identified on the *species*Link network.

Using accepted and alternative names, a total number of 11531 records were retrieved from *species*Link from more than 30 institutions (see list in Table S4 of additional supporting information). From this number, 545 records (~5%) could only be found by means of synonyms, misspellings and orthographical variants, 11210 (97%) explicitly indicated Brazil as the country of origin, 5487 (47%) contained valid coordinates, 979 (8%) were associated with at least one duplicate, 71 (0.6%) had an identification conflict, 587 (5%) were automatically tagged as coming from cultivated individuals, 7 (0.06%) contained taxonomic conflict (ambiguity due to incompatible homonyms) and 33 (0.3%) were observation or living collection records.

Among the records with coordinates in Brazil (5367), 288 (5%) contained georeferencing conflict and 2080 (39%) were likely georeferenced by municipality. The average uncertainty found for the standard gazetteer coordinates in Brazil (IBGE, 2003) considering all municipalities was 38 km (n = 5507, σ = 46), including disjunct territories.

Only 12 records explicitly provided coordinate uncertainty, reinforcing the need to estimate this value. On the other hand, 1595 records (29% of the records with valid coordinates) included verbatim coordinates that were different from the decimal coordinates provided by the *species*Link Web interface, allowing precision errors to be calculated based on the original degrees, minutes and seconds format. Datum errors were found to be negligible in the study area when the alternatives were restricted to WGS84 and SAD69. However, when "Córrego Alegre" is included, errors can be greater than 100 m for points in the south or northwest of the country (Figure 1). The average positional uncertainty for records with valid coordinates in Brazil (discarding those that were likely georeferenced by municipality) was 2 km (n = 3287, σ = 10).

After applying all filters, the final number of records potentially useful in ENM ranged from 1263 (11%) for the 30-arc seconds resolution to 2758 records (24%) for the 10 arc-minutes resolution (Figure 2). Only one record georeferenced by municipality was present in the final selection for the 5 arc-minutes resolution, and nine records for the 10 arc-minutes resolution. Considering only spatially unique records for each resolution, the final proportion started in 9.3% for 30-arc seconds and stabilized around 17% for the others (Figure 2). Five species (*Passiflora balbis* Feuillet, *Passiflora cryptopetala* Hoehne, *Passiflora margaritae* Sacco, *Passiflora rufa* Feuillet & J.M.MacDougal and *Passiflora reitzii* Sacco) had no records at all in *species*Link. On the opposite side, only two species (*Passiflora cincinnata* Mast. and *Passiflora foetida* L.) had more than 100 selected records for ENM. Discarding the 5 species without records, the average number of unique pixels per species was 8, 16, 16 and 15 for the 30-arc seconds, 2.5 arc-minutes, 5 arc-minutes and 10 arc-minutes resolutions, respectively.

Discussion

The large proportion of discarded records corroborates the importance of data quality filters when using data from biological collections in ENM. Among the records with coordinates, special attention is drawn to the high number of records that were likely georeferenced by municipality. In Brazil, the average value for this kind of uncertainty is too high for all spatial resolutions considered in this study. In some cases uncertainties can reach more than 500 km in municipalities from the Amazon region. Such records would therefore be suitable only to ENM experiments in a much coarser resolution than the lowest resolution currently available for WorldClim data.

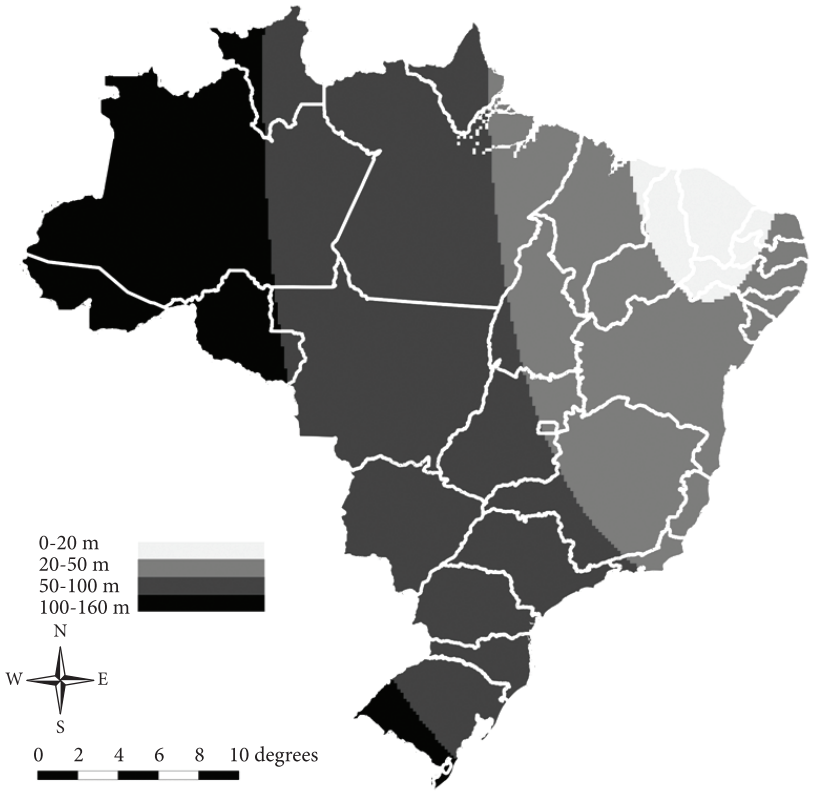

Figure 1. Uncertainty ranges (in meters) when no datum is indicated for coordinates in Brazil. Values were calculated as the maximum distance between points with the same coordinates but a different datum (Wieczorek *et al.* 2004), including WGS84, SAD69 and "Córrego Alegre".

The use of synonyms and misspellings for data retrieval resulted in a relatively low number of additional records. However, when looking at the selected records per species, this can make an important difference, including cases where the single record selected came from an orthographic variant or where most records selected came from synonyms. Nevertheless, if we consider the limitations of scientific names as identifiers for biological taxa (Kennedy *et al.* 2005), this approach does not include all possible taxonomic verifications that should ideally be performed. To mention one example, *Passiflora contracta* Vitta was described as a new species after reviewing existing specimens of *Passiflora ovalis* Vell. ex M.Roem (Vitta & Bernacci 2004). According to this revision, all records identified as *P. ovalis* before 2004 – especially those collected in the northeastern coast of Brazil – should ideally be re-examined. Since scientific names are still the main way of accessing primary biodiversity data (Chapman 2005), such important details are currently difficult to be detected automatically, requiring expert revision.

Even being more restrictive than usual procedures to select records from biological collections, the set of steps described here still do not encompass other data cleaning techniques, such as checking for a possible conflict between the provided altitude and the corresponding value from a Digital Elevation Model at the given coordinates (Chapman 2005), looking for inconsistencies in collectors' itineraries (Peterson *et al.* 2004) and verifying the species kingdom (Heap & Culham 2010) as there can be homonyms across different kingdoms. Additionally, automatic filters may not detect problems in all situations. Digitization errors in coordinates or retrospective georeferencing based on another nearby location, such as a farm or a village, may not be detected with the described procedures. In such cases, algorithms may be trained with inaccurate environmental conditions depending on the resolution and uncertainty. It is also possible for records not collected from native habitat to pass the corresponding filter, although the simple filter described here was very efficient in avoiding false detections (all tagged records were reviewed and only one seemed to be collected from the wild). For these reasons, after performing automatic filters it is still important to include human intervention, allowing experts to indicate possible misidentifications, geographic errors and whether specimens were collected from cultivated areas. This task is immensely facilitated by the filters, allowing attention to be focused on a much smaller number of records with a greater potential to be used in ENM.

The need of certain filters and subsequent expert review is also related to the fact that collections' management software do not always offer the possibility to indicate more detailed positional data, such as coordinate uncertainty and horizontal datum for GPS measurements, as well as

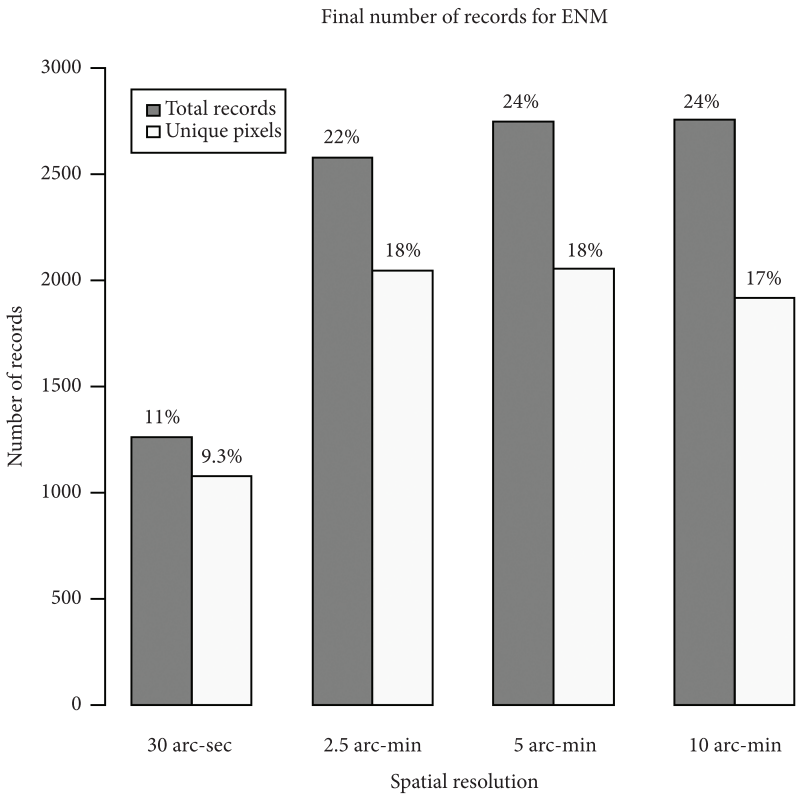

Figure 2. Final number of records to be used in ecological niche modelling for different spatial resolutions considering all species. Dark grey indicates the number of records, while light grey indicates the number of spatially unique records. Percentages were calculated considering the total number of records (11531) retrieved from *speciesLink* (15/02/2011).

to indicate more details about the collecting event in a structured way. Existing data exchange standards already include specific terms for all these purposes, but they can only be used if collectors include the corresponding data when depositing new materials and if collections can store it in a standard format.

Interestingly, despite the methodological differences and the use of more restrictive filters here, the final result did not produce a smaller number of records comparing with a recent study for tropical plants. Feeley & Silman (2011) also found an average of ~8 usable spatially unique points for ENM per species in South America for the 30 arc-seconds resolution. Similarities could also be observed in the number of species per range of points for the same resolution: 33%, 9.4%, 2.7% and 0.7% of the species had 5, 20, 50, 100 or more points respectively (Feeley & Silman 2011), while here we found 36%, 14%, 3% and 1.5% (discarding species with no records in *species*Link). By including more spatial resolutions here, we found that the final number of usable records can double for 2.5 arc-minutes, remaining approximately the same for the next resolutions. These values still stress the importance of further developing ENM presence-only methods that can work with a low number of records per species.

Although data from biological collections may contain errors and uncertainties, they are an invaluable source of biogeographic information, with the important advantage that records are backed by specimens that can be examined and have their data updated whenever necessary. This work provides additional resources that can be used to better explore and stimulate the proper use of herbarium data in ENM, hopefully also serving as a more general guide for selecting biological collections' records before generating ecological niche models.

Acknowledgements

We would like to thank Flávia Santos Pinto and A. Townsend Peterson for all their comments and suggestions on a former version of the manuscript, and to all collections that contributed with plant specimen data.

References

Ariño AH, 2010. Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, 7(2):81-92.

Bernacci LC, 2003. Passifloraceae. In: Wanderley MGL *et al.* (coords.). *Flora Fanerogâmica do Estado de São Paulo*. São Paulo: RiMa, FAPESP. v. 3, p. 247-248

Canhos VP *et al.*, 2004. Global Biodiversity Informatics: Setting the scene for a "New World" of ecological modelling. *Biodiversity Informatics*, 1:1-13.

Cervi AC, 1997. Passifloraceae do Brasil: estudo do gênero *Passiflora* L., subgênero *Passiflora*. *Fontqueria*, 45:1-92.

Cervi AC *et al.*, 2010. Passifloraceae. In: Lista de Espécies da Flora do Brasil. Rio de Janeiro: Jardim Botânico do Rio de Janeiro. Available from: <http://floradobrasil.jbrj.gov.br/2010/FB000182>.

Chapman AD, 2005. *Principles and methods of data cleaning - Primary species and species occurrence data.* version 1.0. Copenhagen: Report for the Global Biodiversity Information Facility.

Egler I & Santos MM (coords.), 2006. *Diretrizes e estratégias para a modernização de coleções biológicas brasileiras e a consolidação de sistemas integrados de informação sobre biodiversidade*. Brasília: MCT/CGEE.

Feeley KJ & Silman MR, 2011. The data void in modelling current and future distributions of tropical species. *Global Change Biology*, 17:626-630. http://dx.doi.org/10.1111/j.1365-2486.2010.02239.x

Fernandez M *et al.*, 2009. Locality uncertainty and the differential performance of four common niche-based modelling techniques. *Biodiversity Informatics*, 6:36-52.

Heap MJ & Culham A, 2010. Automated Pre-processing Strategies for Species Occurrence Data Used in Biodiversity Modelling. In: Setchi R *et al.* (eds.). *Knowledge-Based and Intelligent Information and Engineering Systems*: Lecture Notes in Computer Science. Springer. v. 6279, p. 517-526. http://dx.doi.org/10.1007/978-3-642-15384-6

Hijmans RJ *et al.*, 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965-1978. http://dx.doi.org/10.1002/joc.1276

Instituto Brasileiro de Geografia e Estatística - IBGE, 2003. *Base cartográfica integrada digital do Brasil ao milionésimo*. IBGE.

Kennedy J *et al.*, 2005. Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. In: Ludäscher B & Raschid L (eds.). *Data Integration in the Life Sciences*: Proceedings of the Second International Workshop. San Diego: SpringerLink. v. 3615, p. 80-95. http://dx.doi.org/10.1007/11530084_32

Killip EP, 1938. The American species of Passifloraceae. *Field Museum of Natural History, Botanical Series,* 49:1-613.

Lobo, JM *et al.*, 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33:103-114. http://dx.doi.org/10.111/j.1600-0587.2009.06039.x

Peterson AT *et al.*, 2004. Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British Ornithologists' Club,* 124:143-151.

Peterson AT *et al.*, 2011. *Ecological Niches and Geographical Distributions*. Princeton: Princeton University Press.

Soberón J & Peterson AT, 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B*: Biological Sciences, 359:689-98. http://dx.doi.org/10.1098/rstb.2003.1439

Soberón J & Peterson AT, 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, 2:1-10.

Tropicos, 2011. Missouri Botanical Garden. Available from: <http://www.tropicos.org>.

Vitta FA & Bernacci LC, 2004. A new species of Passiflora in section Tetrastylis (Passifloraceae) and two overlooked species of Passiflora from Brazil. *Brittonia*, 56(1):89-95. http://dx.doi.org/10.1663/0007-196X(2004)056[0089:ANSOPI]2.0.CO;2

Wieczorek J *et al.*, 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science,* 18(8):745-767.

Received: December 2011
First Decision: February 2012
Accepted: April 2012