Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

Spatial data quality of herbarium datasets and implications for decision-making on biodiversity conservation in Brazil

Barros, F.S.[1], Fernandes, R.A.[1], Moraes, M.A.[1], Pougy, N.M.[1], Caram, J.S.[1], Dalcin, E.C.[2] and Martinelli, G.[2]

[1] Centro Nacional de Conservação da Flora/Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão 915, 22460-030, Rio de Janeiro – Brazil.
[2]Instituto de Pesquisa Jardim Botânico do Rio de Janeiro, Rua Pacheco Leão 915, 22460-030, Rio de Janeiro – Brazil.

Abstract

The present level of biodiversity depletion and loss makes quality datasets important for biodiversity conservation. However poor data quality is still critical and limits the usefulness of these datasets. Thereby, data quality assessments are important to ensure a responsible use of those datasets. The Brazilian National Centre for Flora Conservation was created with the objective of assessing the extinction risk of plant species, enabling conservation action planning. In this context, a dataset was created after the compilation of occurrence records of threatened species. The present study aims to assess quality of the dataset and records, and to test quality improvement after data cleaning efforts. We have used the five-component scheme for assessing dataset quality. Significance of the differences between expected and observed proportions were tested using the degree of confidence between them. The Mann-Whitney test was used to compare errors between the original dataset and the cleaned out one. Results indicate poor quality, not only for dataset ($p<0.10$) but also for records ($p<0,10$). Only 54,306 records (22.30%) were considered of good quality. Logical inconsistencies in the dataset were present in 8,237 records (3.37%).

Keywords: biodiversity conservation, data cleaning, data quality, spatial accuracy.

1. Introduction

Present levels of biodiversity depletion are alarming and impose urgency to understand the processes that drive species towards extinction (Buchart *et al.*, 2009). To understand global scale phenomena and their importance for humans and the environment, scientists and policy makers require accurate information (Shortridge and Goodchild, 1999). Poor quality datasets implies that decision based on their content will also be poor (Godchild, 2002). As information is intended to reduce uncertainty in decision-making, significant errors may have practical,

financial even legal implications (Kumi-Boateng and Yakubu, 2010) affecting our capacity to manage biodiversity information and to transform knowledge in conservation actions. Government capability for taking action relies on the capacity to manage biodiversity information, especially in mega diverse countries like Brazil, where scientific data for most species and ecosystems are scarce or absent.

During the last two decades several important datasets like TROPICOS, JSTOR, JABOT among others, have become available. On the other hand, the proliferation of easy to use desktop mapping systems, has allowed non GIS professionals to easily visualize and analyse spatial relationships from their data, but this is often done using inappropriate scales (Chapman *et al.,*2005), and without regard to the spatial error and uncertainty inherent in the data (Chapman, 1999). This can lead to erroneous results, misleading information, unwise environmental decisions and increased costs (Chapman, 2005).

Preserved botanical collections records are basically compose by taxonomic and nomenclatural data, spatial data, collection data and associated descriptive data. The rapid increase in taxonomic and species occurrence data sharing has now made the considerations of data quality principles an important concern (Chrisman, 1983;Goodchild and Gopal, 1989; Chapman, 2005). Since primary plant species data have often been collected without the broader user community in mind, users of datasets need to know something of the inherent quality of dataset's content in order to assess the fitness of the data set for specific purposes, and to determine the quality of products derived from them (Goodchild and Clark, 2002). Traditionally, herbarium data have been collected with the aim of providing information for taxonomic or biogeography research. On the other hand, governments are looking to use the data for improved environmental decision-making, environmental management and conservation planning (Chapman and Busby, 1994).

Appropriate documentation of data is also neglected. Metadata is a description of the characteristics of data that has been collected for a specific purpose (ANZLIC, 1996a). They allow users to assess the fitness of data for a particular application, particularly with respect to quality. Lack of metadata can also contribute to lack of quality, if a user makes the wrong assumptions about the data's meaning (Goodchild and Clarke, 2002). However, good documentation occurs not only at the dataset level, but also at the data record level.

Recently, the publication of the last official version of the Threatened Species List of the Brazilian Flora generated disagreement between scientists and politicians (Scarano and Martinelli, 2010). Despite the careful work conducted by scientists, the lack of proper documentation was pointed as a limiting factor. And the final decisions ended up being taken based in different aspects other than the biological ones. According to the national regulation in Brazil (MMA, 2008), a new version of the red list must be published in 2012, four years after the last publication. The Brazilian National Centre for Flora Conservation - CNCFlora was created in 2008, at the Rio de Janeiro Botanical Garden -JBRJ with the objective of coordinating national efforts for plant conservation, assessing the extinction risk of plant species and producing the new version of the red list for plants. In this context, a dataset was created after the compilation of occurrence records (248,837) of 4,711 threatened species, obtained from 70 herbaria. This dataset is intended to be used as evidence for species occurrences, in order to prove spatial information to be provided with the extinction risk assessments.

Nevertheless, to be able to use this dataset for this purpose a process of validation and data cleaning, including a retrospective georeferencing process, was

conducted. The present study aims to test quality of this dataset and to quantify quality improvements after the extensive checks and corrections carried out. The fitness of the dataset created is discussed for this particular application and adequacy of methods for consistency checks and data cleaning are also addressed.

2. Materials and Methods

A thorough survey was conducted in order to compile digital occurrence records corresponding to testimonies of 4,711 species, a resulting list of every species officially considered as threatened, since 1968, in Brazil. A list of 10,451 synonyms was also used in order to identify testimonies of these biological entities in herbaria datasets. Therefore a total of 15,162 names were used for compilation of records.

Quality as applied to data has various definitions, but for the purposes of the present study we will adopt the largely accepted definition of "fitness for use" (Chrisman, 1983) or even the less restrictive version "fitness for future or potential use" (Chapman, 2005). For quality check procedures we followed the five-component scheme devised in the mid 1980s by the Spatial Data Transfer Standards (www.fgdc.gov) and discussed by several authors (Guptill and Morison, 1995; Goodchild and Clarke, 2002; Dalcin, 2004). The scheme identifies five dimensions of data quality that can be measured providing useful information on the dataset. These dimensions are: a) lineage; b) completeness; c) logical consistency; d) attribute accuracy; and e) positional accuracy. Lineage is understood as the information on the process of creation or data acquisition that specifies collection methods, devices used, and vertical datum among other. Completeness refers to the degree to which the dataset captures all of the expected data. Completeness can also refer to spatial extent, the number of available attributes actually included and to known missing data. For plant specimen datasets, the "Darwin Core" standard is normally used for data sharing and allows structured and logical organization of the attributes. However, completeness is a complex concept that may include several other components and may vary accordingly the intended use of the dataset. Logical consistency refers to the internal consistency of data, and the dataset adherence to its own defined rules. In spatial datasets, logical inconsistency between the geometric content of a dataset and the topological content can exist. But if the rules are well defined, it is possible to detect errors and correct them. Attribute accuracy refers to the accuracy of the recorded attributes associated with each object. While positional accuracy is related to the accuracy of the actual position of an object on the Earth's surface. In some cases it may be impossible to separate positional accuracy from attribute accuracy.

In order to make data quality assessment possible it was necessary to identify and correct all kinds of errors in the datasets. Since our intent was to provide evidence for species occurrence, data quality dimensions were evaluated considering only attributes that contained spatial references such as state, municipality, longitude and latitude. Since spatial references are mandatory for including data in herbarium collections, and considering a confidence interval of 10%, we expected that at least 90% of dataset would comply with completeness, attribute accuracy and logical consistency. Therefore, if the results showed a higher proportion of compliance, we would consider the dataset of good quality. Significance of differences between expected and observed proportions of

compliance were tested using the confidence interval between proportions by the R language and environment for statistical computing and graphics (R Development Core Team, 2011). Completeness was accounted considering the fields mentioned above. Missing data and inconsistencies on any of these fields were considered as inaccuracy of attributes. For checking dataset quality, logical consistency between municipality and state was considered. For quality check of the records we used positional accuracy and logical consistency of the geo references with the attribute field municipality, only of those records that were considered as of good quality in the previous analysis (65,252). Since analyzed data fields have hierarchical relations, and considering a confidence interval of 10%, we expected that at least 90% of the records would have positional accuracy and logical consistency concerning geo references with municipality. A higher proportion of compliance would indicate good quality of records. Significance of differences between expected and observed proportions of compliance were also tested using the confidence interval between proportions by the R language and environment for statistical computing and graphics (R Development Core Team, 2011). In order to evaluate if there was significant quality improvement on data after the data cleaning processes we used the Mann-Whitney test to compare geo references errors between the original dataset and the resulting one (Zar, 2010).

3. Results and Discussion

A dataset composed of 243,923 records was compiled. Nevertheless, the dataset did not comply with the five quality dimensions and was considered of poor quality (73.24%; Confidence Interval: 0,28%; Figure 1). None of the records had proper data on lineage, which is considered an important quality component (Goodchild and Clarke, 2002). However, lack of lineage might be a consequence of database structure, since "Darwin Core" framework does not provide sufficient attribute fields for recording details about the process of data creation. By knowing these details it is possible to make inferences about the quality of data and its utility for certain purposes. It is the data lineage that provides documentation, which allows repeatability of experimental results, and therefore the independent confirmation of findings necessary to the scientific method (Goodchild and Clarke, 2002). Without proper data on: a) instruments used to make measurements; b) the institutions responsible for data creation; c) adopted methods and associated errors; any analytical conclusions may be questioned, since the degree of uncertainty associated to data is not evident. In this way, if historical datasets of herbaria and museums are going to be used to support the decision making processes on biodiversity a thorough historical analysis must be made, in order to track back details of data collection and transformations.

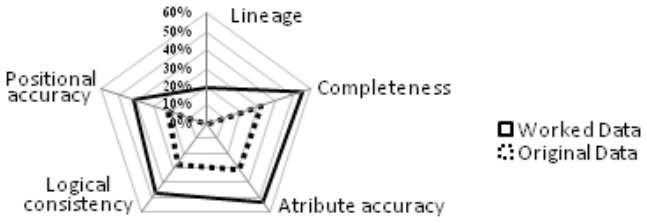

Figure 1: Quality comparison between the before and after CNCFlora data cleaning.

Considering the attribute fields pointed before, completeness was quite low. Only 76,149 records (31.20%) complied with completeness (Figure 1). Herbarium datasets are composed by legacy collections, many of them created in the 19th century. European naturalists used to come to Brazil and spend months, or even years, describing hundreds of new species. Field expeditions would take naturalists deep into the country interior, and political boundaries were not accurately defined at that time. Most records (241,045 - 98.82%) have at least the state attribute properly filled. Many type specimens were deposited in international herbaria, with duplicates in local institutions. For each duplicate, when lineage data was lost, it would become an independent record. Despite institutional control of origin, if metadata was not available, every attribute field could be altered, reducing fidelity to the original data. Furthermore, to consider duplicates as one single sample may increase logical inconsistencies in the dataset. In these cases, correction is the only option for quality improvement (Dalcin, 2004; Chapman, 2005).

Attribute accuracy was considered a major problem, since 170,434 records (69.87%) had missing data, or typo. Only 73,489 records (30.10%; Figure 1) had the analyzed attribute fields properly filled. Logical inconsistencies in the dataset were present in 8,237records (3.37%). This type of error is normally associated to mistakes during the processes of data collection. Sometimes political boundaries are hard to identify on field and the collector only realizes the correct location when a spatial projection is made. Positional inaccuracy was registered in 10,946 records (4.48%). But this result was expected to be even higher. Historical collections of Brazilian herbaria are composed by different datasets, which were incorporated gradually over time without proper metadata. Without data on the vertical datum, for example, is almost impossible to achieve high positional accuracy.

On regard of the quality of records, our results indicate that spatial data also has poor quality, even when complying with the five-component scheme. Positional accuracy and logistical consistency between the geo references, expressed by the attribute field latitude and longitude, and the attribute field municipality was low. Only 54,306 records (83.23% out of 65,252 records; Figure 1) were considered of good quality, while 10,946 records were considered of poor quality (16.77%; Confidence Interval: 0.24%). But it is important to highlight that quality is strictly related to the intent of use, and therefore these records might show better fitness for other purposes, like taxonomic revisions.

Results of the Mann-Whitney test indicates that significant data quality improvement was achieved after CNCFlora's data cleaning process ($U = 64448306$, d.f. = 1; $p < 0.0001$; Figures 1). Organizing metadata on records showed to be indispensable, since it allows you to evaluate the degree of uncertainty associated to each record, understanding it and considering during further analysis helping to reduce commission mistakes (Chapman, 2005; Lemes *et al.*, 2011).

4. Conclusion

Historical herbarium collections are scattered among several herbaria across the country and factors like poor data quality, limited access to databases, absence of records in databases, lack of appropriate metadata, and restricted budgets for data cleaning and quality control still represent a barrier that narrows utility of datasets. Therefore it is important to manage carefully this datasets, promoting constant

quality improvements and always maintaining the original data as backup. Any analysis must consider associated metadata in order to avoid wrong assumptions that may lead to bad decisions with serious consequences for biodiversity. Improvements in the adopted protocols for specimen data collection are needed.

References

Butchart, S.H.M., Walpole, M., Collen, B. *et al*. (2010), Global biodiversity: Indicators of Recent Declines. Science, 328:1164-1168.

Chapman.A.D. (1999), Quality Control and Validation of Point-Sourced Environmental Resource Data.In: Lowell, K., Jaton, A. (eds.). Spatial accuracy assessment: Land information uncertainty in natural resources, Chelsea, pp. 409-418.

Chapman, A.D., Busby, J.R. (1994), Linking plant species information to continental biodiversity inventory, climate and environmental monitoring. In:Miller, R.I. (ed.). Mapping the Diversity of Nature, London: Chapman and Hall, London. pp. 177-195.

Chapman, A.D (2005), Principles of Data Quality, version1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Chrisman, N.R (1983), The Role of Quality Information in the Long-term Functioning of a geographic information system. Cartographica, 21:78-87.

Dalcin, E.C (2004), Data quality Concepts and Techniques Applied to Taxonomic Databases. Thesis for de degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton.

Darwin Core Task Group (2011). Darwin core terms: a aquick reference guide. Available in <http://rs.tdwg.org/dwc/terms/>. Access in Apr 2011.

Goodchild, M.F., Clarke, K.C. (2002), Data Quality in Massive Data Sets. In: Abello, J., Pardalos, P.M., Resende M.G.C. (eds.). Handbook of massive datasets. Dordrecht: Kluwer, pp. 643–660.

Goodchild, M.F., Gopal, S., editors, (1989), The Accuracy of Spatial Databases.Basingstoke: Taylor and Francis, pp. 290.

Guptill, S.C., Morrison, J.L. (1995), Elements of Spatial Data Quality, Elsevier Sci., U.K. 78p.

JABOT – Banco de dados da Flora Brasileira, Instituto de Pesquisas Jardim Botânico do Rio de Janeiro.Avaiable in <http://www.jbrj.gov.br/jabot>.Acess in Apr 2011.

Kumi-Boateng, B., Yakabu, I. (2010), Assessing the Quality of Spatial Data. EuropeanJournalofScientific Research, Vol. 43(2):507-515.

Lemes, P., Faleiro, F.A.M.V., Tessarolo, G., and Loyola, R.D. (2011) Refinando Dados Espaciais para a Conservação da Biodiversidade. Natureza & Conservação 9(2):240-243.

MMA –Ministério do Meio Ambiente (2008), Instrução Normativa Nº 6 de 23/09/2008.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URLhttp://www.R-project.org/.

Shortridge, A.M., Goodchild, M.F. (1999), Communicating Uncertainty for Global Data Sets. In: Shi, W., Goodchild, M.F., Fisher, P.F. (eds). Proceedings of International Symposium on Spatial Data Quality, Hong Kong. Hong Kong: Hong Kong Polytechnic University, pp. 59–65.

Zar, H. J. (2010). Biostatistical Analysis.5Th Ed. Prentice Hall, New Jersey, 944p.