Report No. 6 **TOOLS**

Environmental Data Quality
— b. Data Cleaning Tools

Arthur D. Chapman

June 2003
Modified and updated 26 January 2004

1. Contents:

2. Aim:

To

a. examine existing tools and methodologies for testing, cleaning and validating species data;
b. examine the feasibility of producing a tool kit for data cleaning;
c. prepare guidelines and possible tool kit on data cleaning for provision to institutions
d. give talks to institutions on data cleaning and validation.

3. Background:

This report follows on from Report No. 5 – Environmental Data Quality - Discussion paper – where there is a more detailed discussion of data quality issues and methodologies for cleaning and validating spatial data. This report concentrates solely on species data.

Museums and herbaria throughout the State of São Paulo and elsewhere in Brazil are beginning to database their collections. Some of these, especially in the State of São Paulo are being carried out as part of the FAPESP/Biota *specie*sLink project being managed through CRIA (CRIA 2002).

The main goal of the *species*Link project is to implement a distributed information system to retrieve primary biodiversity data from collections throughout the State. Twelve collections (3 herbaria, 2 acari, 3 fish, 1 algae and 3 microorganism collections) are already engaged in the first phase of the project. Others will join the project from time to time.

This report concentrates on both web systems and software that include within their packages tools that have been designed specifically for data cleaning, or which may be used to help identify errors in species data and thus lead to cleaning and validation of those data. It also examines the feasibility of using these software within an integrated data-cleaning tool kit, or whether they may be modified in some way to be so included. The report also makes a number of recommendations for future data cleaning research areas.

4. Species Data:

See Report 5 (Chapman 2004) for a discussion of general data quality issues, and how errors in species data may arise.

5. Data Quality:

Errors in data are common and are to be expected. The usual view of errors and uncertainties is that they are bad, but a good understanding of errors and error propagation can lead to active quality control and managed improvement in the overall data quality (Burrough and McDonnell 1998). Errors in species' data are particularly common and need to be catered for. Errors in spatial position (geocoding) and in taxonomic circumscription are two of the most common errors found in specimen databases, and these errors can cause major problems in modelling and biogeographic studies. Assessment of the accuracy of input data is essential otherwise the results of any modelling will be meaningless. Correcting errors in data and weeding out bad records can be a time consuming and tedious process (Williams *et al.* 2002) but should not be ignored.

In determining the quality of species data from a spatial viewpoint, there are a number of issues that need to be examined. These include the identity of the collection – a wrong identification can be the cause of major spatial error, errors in the geocoding (latitude and longitude), and spatial bias in the collection of the data.

6. Error Checking Methods:
Methods have been developed to identify georeferencing errors in species' data. These include the use of climate models to identify outliers in climate space (Chapman 1992, 1999, Chapman and Busby 1994) and the use of automated georeferencing tools (Beaman 2002, Wieczorek and Beaman 2002). Most collection institutions do not have a high level of expertise in data management techniques or in Geographic Information Systems (GIS). What is needed in these institutions is a simple, inexpensive set of tools to assist in the input of data and information, including geocoding information, and similar simple and inexpensive tools for data validation that can be used without the necessary incorporation of expensive GIS software. Some tools have already been developed to assist with the first of these – tools such as Biota (Colwell 2002), BRAHMS (University of Oxford 2003), Specify (University of Kansas 2003a), BioLink (Shattuck and Fitzsimmons 2000) and others that provide database management and associated data entry tools; eGaz (Shattuck 1997), geoLoc-CRIA (CRIA 2004) and BioGeoMancer (Beaman *et al.* 2003), that assist in the georeferencing of collections; and Diva-GIS (Hijmans *et al.* 2003) and FloraMap (Jones and Gladkov 2001) that use basic GIS tools and various outlier detection methodologies for detecting possible errors. There are also a number of documented guidelines available on the internet that can assist institutions in setting up and managing their databasing programs. Examples include the MaNIS Georeferencing Guidelines (Wieczorek 2001a), the MaPSTeDI Guide to Georeferencing (University of Colorado 2003) and HISPID (Conn 1996, 2000). For a link to these and other tools, see under Item 9 (below) - Links to software tools.

Because of the very nature of natural history collections, it is not possible that all geocode information be highly precise, or that there be a consistent level of precision within a database. Data with a very low precision, however, are not necessarily of low quality. Quality only comes into being once the data are being used and is not a character of the data *per se*. For example, a collection record that has just the locality information: "São Paulo State" may well be of "poor quality" if one wants to know where in the State it may have been collected, but if all one wants to know is if the species occurs in South America or not, then that record is of high quality. Quality is merely a factor of fitness for use and is a relative term. What is important is for users of the data to be able to determine from the data itself, if the data is likely to be fit for the purpose for which they want to put it. The level of accuracy of each given geocode should therefore be recorded within the database. I prefer this to be in non-categorical form, recorded in meters, however many databases have developed categorical codes for this purpose. When this information is available, a user can request, for example, only those data that are better than a certain metric value – e.g. better than 5,000 meters. There are a number of ways of determining accuracy of geocoded records. The point-radius method (Wieczorek *et al.* in press) is, I believe, the easiest and most practical method, and is one I have previously recommended (Chapman and Busby 1994). It is also important that automated georeferencing tools include calculated accuracy as a field in the output. I understand that the

BioGeoMancer program (Beaman *et al.* 2003), which is still under development, intends to include this feature (Beaman *pers. com.* 2002), and the geoLoc-CRIA (CRIA 2004) developed as a result of an earlier draft of this paper, already includes such a feature.

Over time, it is hoped that species collection data resources will improve as institutions move to more precise instrumentation (such as GPS) for recording the location of new records and as historic records are corrected and improved.

7. Data Cleaning Methods:

As mentioned above, errors in spatial position (geocoding) and in taxonomic circumscription are two of the major causes of error in modelling and biogeographic analysis. Other errors, such as the misspelling of names etc., can also not be ignored. There are any number of methods and techniques that can aid in cleaning up these types of errors. They range from methods that have been operating in museums and herbaria for hundreds of years, to automated methods that are still largely untested.

7.1 Names and non-spatial data

Species names provide the most important key to most taxonomic databases. Names, whether they are scientific binomials or common names, provide the first point of entry for most databases. Errors in names may arise in a number of ways: the identification (i.e. its taxonomic circumscription) may be wrong, the name may be misspelt, or the format may be wrong. The first of these is not easy to check or rectify without a lot of tedious effort, and requires the services of a taxonomic expert. The others though, are more easily catered for, and methods can and have been developed to assist data entry so that these errors do not occur or are rare. A separate paper on Guidelines to Nomenclature have been prepared as an adjunct to this report (Chapman 2003a), and more details can often be found there.

7.1.1 Taxonomic circumscription of names

Traditionally, museums and herbaria have had a determinavit system in operation whereby experts working in taxonomic groups from time to time examine the specimens and determine their circumscription or identification. This may be done as part of a larger revisionary study, or by an expert who happens to be visiting an institution and checks the collections while there. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be anyway around this, however, as automated computer identification is unlikely to be an option in the near or even long-term future.

One option may be the incorporation of a field in databases that provides some indication of the certainty of the identification when made. This would be a code field, and may be along the lines of:

- identified by World expert in the taxa with high certainty
- identified by World expert in the taxa with reasonable certainty
- identified by World expert in the taxa with some doubts
- identified by regional expert in the taxa with high certainty
- identified by regional expert in the taxa with reasonable certainty
- identified by regional expert in the taxa with some doubts
- identified by non-expert in the taxa with high certainty
- identified by non-expert in the taxa with reasonable certainty

- identified by non-expert in the taxa with some doubt
- identified by the collector with some doubt

How one might rank these would be open to some discussion, and likewise whether these were the best categories or not. I understand that there are some institutions that do have a field of this nature, but at this stage, I have not been able to find an example. The HISPID Standard Version 4 (Conn 2000) does include a simplified version – the Verification Level Flag with five codes, viz:

0	The name of the record has not been checked by any authority
1	The name of the record determined by comparison with other named plants
2	The name of the record determined by a taxonomist or by other competent persons using herbarium and/or library and/or documented living material
3	The name of the plant determined by taxonomist engaged in systematic revision of the group
4	The record is part of type gathering or propagated from type material by asexual methods

Table 1. Verification Level Flag in HISPID (Conn 200).

Many institutions already have a form of certainty recording with the use of terms such as: “aff.”, “cf.”, “*s. lat.*”, “*s. str.*”, “?”, etc., however, just what is meant by these terms from one institution to another is not always clear.

Geocode checking methods (see 7.2 below) can often help identify errors in taxonomic circumscription through the identification of outliers in geographic or environmental space. Although generally an error picked up through geocode checking will be an error in either the latitude or longitude, occasionally it indicates that the specimen has been given the wrong name and because of this falls outside the normal climate or environmental range of the species. See under 7.2 below for a more detailed discussion.

7.1.2 Spelling of names

7.1.2.1 Scientific names

The correct spelling of a scientific name is generally governed by one of the various Codes of Nomenclature (see References). However, errors can still occur through typing errors, ambiguities in the Nomenclatural Code, etc. The easiest method to ensure such errors are kept to a minimum is to use an ‘Authority File” during input of data. Most databases can be set up to incorporate either an unchangeable authority file, or an authority file that can be updated during input.

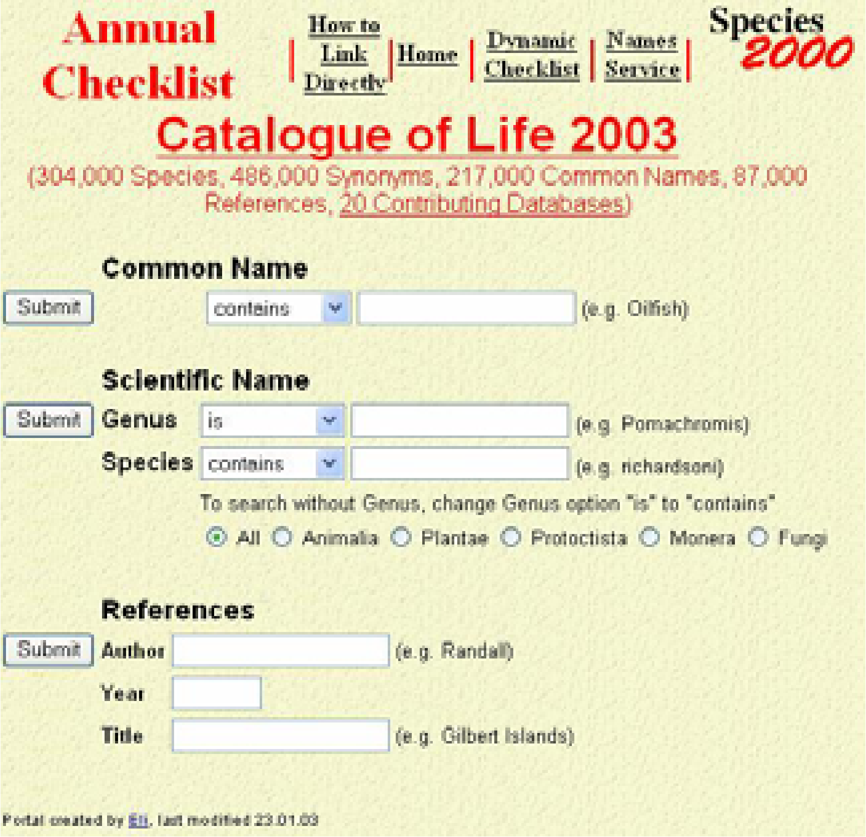
Authority files exist for a number of taxonomic groups, and are being developed by a range of agencies. It is unlikely that a detailed authority file for all of Brazil’s taxa will be produced in the near future, however, existing authority files (see Species2000) can be used as a beginning, and the databases set up in such a way that new names can be added from time to time. For example, assume a database has had an authority file added with a pull down list, or fills in the field as one types (for example as happens in an EXCEL spreadsheet if one starts to type a name in a field where that name may already be in an earlier row).

1. Use the pull down list to search for the name
2. It is not there
3. Click on the button – “New name”
4. Add the New Name
5. The database may come back and say “This name is similar to <name>” do you want to continue?
6. Yes
7. The name is added to the list, and the next time you wish to add a name, that name will now appear in the pull-down list.

In this way, you are gradually adding to and improving the authority file.

As an extra check, these names may then go into a secondary list that a supervisor looks at from time to time and either approves or discards. Depending on the level of sophistication of the database, the list may include synonyms and if you begin to type in a name, it may ask you if you really wish to add this name as it is listed in the authority file as a synonym of <name>.

I recommend that Authority files be used wherever possible, and that over time, an authority file for many Brazilian taxa may be built up in this manner between collaborating institutions. A good start is the Species2000 Catalogue of Life (Species2000 2002), available on CD as an Annual Checklist, although the format of this document needs improving to make it easier to incorporate into databases.



Catalogue of Life 2003

(304,000 Species, 486,000 Synonyms, 217,000 Common Names, 87,000 References, 20 Contributing Databases)

Common Name

contains (e.g. Oilfish)

Scientific Name

Genus is (e.g. Pomachromis)

Species contains (e.g. richardsoni)

To search without Genus, change Genus option "is" to "contains"

● All ○ Animalia ○ Plantae ○ Protoctista ○ Monera ○ Fungi

References

Author (e.g. Randall)

Year

Title (e.g. Gilbert Islands)

Portal created by Eli, last modified 23.01.03

Fig. 1. Search page from the Species2000 – Catalogue of Life – online version.

The Species2000 Annual checklist is also available electronically for checking individual names and is in addition to a regularly updated checklist, which is also available on-line (Fig. 1).

The Global Biodiversity Information Facility (GBIF) is also planning the development (in conjunction with Species 2000 and others), a global names catalogue called the Electronic Catalogue of the Names of Known Organisms (ECAT) (GBIF 2003). Once developed, this will be a major source of names for most of the World's biota, and aims to include 90% of all scientific names available by 2013 (GBIF 2003).

7.1.2.2 Common names

There are no hard and fast rules for 'common' names, be they in Portuguese, English or regionally-based indigenous names. In some groups, for example birds (see Christidis & Boles 1994), agreed conventions and recommended English names have been developed. In many groups, and especially plants, one taxon may have a number of common names with these often being region specific. A good example is the species *Echium plantagineum* which is known variously as 'Paterson's Curse' in one Australia State and 'Salvation Jane' in another. Many Brazilian examples can be seen at http://www.recor.org.br/publicacoes/plantas-nativas.html.

Often what are called 'common' names are in reality colloquial names (especially in botany) and may have just been coined from a translation of the latin scientific name.

It is recommended that when databasing common names, that some form of consistency in construction be followed. For English and Spanish common names, I am recommending that a similar convention to that developed for use in Environment Australia (Chapman *et al.* 2002) and modified here for use in Brazil and South America (Chapman 2003a) be followed. An explanation on the use of common names in Portuguese may be found at http://www.afarmacia.hpg.ig.com.br/index.html.

As common names are generally tied to the scientific name, checks can be carried out from time to time to check for consistency within the database. This can be a tedious procedure, but only need be carried out at irregular intervals. Checks can be done by extracting all unique occurrences and checking for inconsistencies, e.g. missing hyphens etc.

7.1.2.3 Infraspecific Rank

The use of an infraspecific rank field is a lot more of a problem in databases of plants than in databases of animals. Animal taxonomists general only use the one rank, that of subspecies, and even this is not usually cited, with the name treated as a trinomial.

Stipiturus malachurus parimeda

With plants, there are several levels below species that may be used. These infraspecific ranks are ***subspecies, variety, subvariety, forma*** and ***subforma***. The last three are seldom used, but do need to be catered for in plant databases. Again, a pick-list should be set up with a limited number of choices. If this is not done, then errors begin to creep in, and you will invariably see subspecies given as: subspecies, subsp., ssp., subspp., etc. This can then be a nightmare for anyone trying to extract data. It is better to restrict the options at the time of input, than have to cater for a full range at the time of data extraction, or attempt data cleaning to enforce consistency at a later date. I recommend the use of:

subsp.	subspecies
var.	variety
subvar.	subvariety
f.	form/forma
subf.	subform
cv.	cultivar

In collection databases, I don't recommend the inclusion of a hierarchy where more than one level may exist, because this just adds an extra layer of confusion, and under the International Code for Botanical Nomenclature (2000), the hierarchy is unnecessary to unambiguously define the taxon. For more details, see the Guidelines to Nomenclature (Chapman 2003a).

7.1.2.4 Unpublished names

Not all collections placed in a collections database are going to belong to a validly published name. To be able to retrieve these collections from the database it is necessary to provide a 'temporary' name for that collection. If unpublished names can be incorporated into a database in a standard format, it makes it a lot easier to keep track of them, and to be able to retrieve them at a later date.

In the 1980s in Australia, botanists agreed on a formula (Croft 1989, Conn 1996, 2000) for use with unpublished names. This was to avoid confusion arising through the use of such things as "*Verticordia* sp.1", "*Verticordia* sp.2" etc. Once databases begin to be combined, for example through the Australian Virtual Herbarium (CHAH 2002) or *species*Link (CRIA 2002), names like these can cause even more confusion as there is no guarantee that what was called "sp.1" in one institution is identical to "sp.1" in a second. One way to keep these databases clean and consistent, and enable the smooth transfer of data from one to another, is through the use of a formula similar to that adopted in Australia. See also the Guidelines on Nomenclature (Chapman *et al.* 2002, Chapman 2003a).

The agreed formula is in the form of: "<Genus> sp. <colloquial name or description> (<Voucher>)":

> *Prostanthera* sp. Somersbey (B.J.Conn 4024)

Later, when the taxon is formally described and named, the formula-name can be treated as a synonym, just like any other synonym.

I recommend, where possible, that a similar form be adopted for use in databases linked through *species*Link, and elsewhere in Brazil.

7.1.2.5 Author Names

The authors of species names may be included in some specimen databases, but more often than not, their inclusion can lead to error as they are seldom thoroughly checked before inclusion. They are only really necessary where the same name may have inadvertently have been given to two different taxa (homonyms) within the same genus or where the database attempts to include different concepts. The inclusion of the author's name following the species (or infraspecies) name can then distinguish the two names. If databases do include authors of species names, then these should definitely be included in fields separate from the species' names themselves. Fortunately, this is usually the case.

With animal names the author name is always followed by a year; with plants, the author name or abbreviation is given alone. For details, see the Guidelines on Nomenclature (Chapman *et al.* 2002, Chapman 2003).

7.1.2.5 Collector's names

Collector's names are generally not standardised in collection databases, although an attempt at standardisation of plant collector's names is being attempted for plant names in the *species*Link project (Koch, 2003). Without such a standard list, very little can easily be done with data cleaning. It is recommended, however, that names be included in collection database in a standard format. The HISPID Standard (Conn 2000) recommends the following:

> "Primary collector's family name (surname) followed by comma and space (,) then initials (all in uppercase and each separated by fullstops). All initials and first letter of the collector's family name in uppercase. For example, Chambers, P.F."

It is recommended that secondary collectors be placed in a second field. If this is not the case, then it is recommended that they be cited with a comma and space used to separate the multiple collectors. For example:

> Tan, F., Jeffreys, R.S.

Where there is a chance of confusion, other given names should be spelt out. For example, to distinguish between Wilson, Paul G. and Wilson, Peter G. (with a space after the given name; no punctuation, except as separator between two names, as described above).

Titles should be omitted.

If the family name (surname) consists of a preposition and a substantive, as in many European names (e.g. C.G.G.J. van Steenis), then the preposition is in lower case and the substantive has an initial capital letter. For example:

> Steenis, C.G.G.J. van

Other names of similar form include de la Salle, d'Entrecasteaux, van Royen etc. It should be noted, however, that many of these names have been anglicised, particularly in America, such that both parts of the family name are treated as substantive. In such cases, these names can be transferred as follows:

> De Nardi, J.C.

The prefixed O', Mac', Mc' and M' (e.g. MacDougal, McKenzie, O'Donnell) should all be treated as part of the substantive and hence transferred as part of the family name. For example:

McKenzie, V.

Hyphenated given names should be transferred as all uppercase, with the first and last initial separated by a hyphen (without spaces), and only the last terminated by a fullstop. For example:

Quirico, A-L.
Peng, C-I.

If the collector of the record is unknown, then the term "Anonymous" should be used.

Interpreted information should be enclosed in square brackets, eg.

Anonymous [? Mueller, F.]

The use of a personal collection is admissible: For example:

Anonymous (Herb. J.M. Black).

7.2 Geocodes

As previously mentioned, a number of programs do exist that can aid in checking and testing for errors in geocodes attached to specimen records. Other tools are available to assist in the original assignment of geocodes to the data from the location information (such as distance and direction from a named location).

The testing of errors in already assigned geocodes involves

- checking against other information internal to the record itself, for example, State, named district, etc.;
- checking against an external reference using a database – is the record consistent with the collecting localities of the collector, for example.
- checking against an external reference using a GIS, etc. – that the record falls on land rather than at sea, for example;
- checking for outliers in geographic space; or
- checking outliers in environmental space.

7.2.1 Geocode assignment

Traditionally, geocodes have been assigned to specimen data using maps of varying quality and scale, and has involved a rather tedious, and lengthy procedure. It has been variously estimated that it costs as much, and takes as long to add the geocode to a specimen record using traditional methods as it takes to database all the rest of the information on the label (Armstrong 1992, Chapman 1991). Various automated and semi-automated methods have been, and are being, developed to speed up this process and to make the process more transparent with documented accuracy.

Most automated methods of geocode assignment rely on the use of a Gazetteer of named places with associated latitude and longitude, although the database itself, using previously databased information, can also be used in a semi-automated way with little modification.

One important part of assigning geocodes is the simultaneous process of assigning an accuracy figure. As mentioned elsewhere, it is important that databases record the accuracy of the geocoding in a separate field. I recommend that this be recorded in meters; however, others prefer the use of a code.

7.2.1.1 Semi-automated geocode assignment

As data is built up within a specimen database, the information already held in the database can be used to quickly assign geocodes to specimens being added. A simple report procedure can be incorporated from the database that allows for a search of a known place to see if a specimen from the same locality as being added has already been databased and assigned a geocode.

For example, you are about to database a collection that has the location information “10 km NW of Campinas”. You can search the database for “Campinas” and look through the collections already databased to see if a geocode has already been assigned to another collection from “10 km NW of Campinas”. This process can be made a lot simpler if the database structure includes fields for “Nearest Named Place”, “Distance” and “Direction” or similar, in addition to the traditional free text locality description.

This methodology has the drawback that if the first geocode had been assigned with an error, then that error will be perpetuated throughout the database. It does, however, allow for a global correction if such an error is found in any one of the collections so databased.

With linked databases, such as the Australian Virtual Herbarium (CHAH 2002) or *species*Link (CRIA 2002), on-line procedures could be set up to allow for a collaborative geocoding history to be developed and used in a similar way. Of course, one drawback of this is that there is a certain amount of loss of control within your database, where an error in another database can be inadvertently copied through to your own database. Good feed back mechanisms need to be developed between institutions to ensure that firstly, errors are not perpetuated inadvertently, and secondly, that information on errors that are detected are fed back to the originating database custodians as well as other dependent databases.

Many plant collections are distributed as ‘duplicates’ to other collection institutions. Traditionally this has been done prior to geocoding, and one can often find exactly the same collection in a number of herbaria, all with different geocodes. To circumvent this problem, geocodes either need to be added before distribution, or a collaborative arrangement entered into between institutions. As explained earlier, it costs a lot in both time and money to add geocodes, it is an extremely wasteful exercise if several institutions individually spend time and resources geocoding the same collections. The waste is further compounded if different geocodes are given to the same collection in different institutions.

In their paper on the point-radius method of georeferencing locality descriptions, Wieczorek and others (in press), provide a table of nine types of locality descriptions found in natural history collections. The first three of these they recommend should not be georeferenced, but an annotation be given as to why it was not georeferenced. Previously (Chapman and Busby 1994) recommended that a general geocode be given with an accuracy figure of 100 000 000 meters be given. This latter method has the drawback of users extracting the data only using the geocode and not the associated accuracy field and ending up with what looks like a point without its associated huge radius. The Wieczorek method overcomes this drawback by not providing such a misleading geocode. The nine categories listed by Wieczorek *et al.* (in press) are:

1. ***dubious*** (e.g. 'Isla Boca Brava?')
2. ***cannot be located*** (e.g. Mexico', 'locality not recorded')
3. ***demonstrably inaccurate*** (e.g. contains contradictory statements)
4. ***coordinates*** (e.g. with latitude or longitude, UTM coordinates)
5. ***named place*** (e.g. 'Alice Springs'
6. ***offset*** (e.g. '5 km outside Calgary')
7. ***offset along a path*** (e.g. '24 km N of Toowoomba along Darling Downs Hwy')
8. ***offset in orthogonal directions*** (e.g. '6 km N and 4 km W of Welna')
9. ***offset at a heading*** (e.g. 50 km NE of 'Mombasa')

Each of these would require a different method of calculation of the accuracy as discussed in the paper (Wieczorek *et al.* in press).

7.2.1.2 Automated geocode assignment

Automated geocode tools are based on determining a latitude and longitude from the textual locality information using a distance and direction from a known location. Ideally, databases include at least a "Nearest Named Place", "Distance" and "Direction", or better still, "Named Place 1", "Dist 1", "Dir. 1", "Named Place 2", "Dist 2", "Dir 2". Thus "5 km E of Smithtown, 20 km NNW of Jonestown" would be appropriately parsed into the six fields cited above.

As most databases are not so structured, attempts are being made to develop automated parsing software to parse free text locality descriptions into basic "Nearest Named Place", "Distance" and "Direction" fields, and then using these fields, in association with appropriate Gazetteers to determine the Geocode (see BioGeoMancer below).

At the same time as a geocode is determined in this way, it is important that an extra field, that of "Geocode Accuracy" (see Conn 2000 and earlier discussion, this paper) be included to give an idea of the accuracy of the determined geocode.

Drawbacks of this methodology include possible errors in the Gazetteers (most publicly available gazetteers I have examined lately have a considerable number of errors – see for example, figure 2), many location fields are not as straight forward as those cited above, often historic place names are used, and many distances on collection labels are "by road" distances rather than direct, although this is seldom stated on the label itself. Accuracy fields need to take into consideration these issues as well as the error inherent in vector distances – does "South West" mean between "South" and "West" or between SSW and WSW. As this distance from the source increases, the inherent error in these will also rapidly increase (see discussion in Wieczorek *at al.* in press). A combination of this method and a simple GIS methodology (see below) would provide the greatest accuracy. For example, where the automated method cited here produced a point on a map that the user can “grab and drag” to a more appropriate place – for example to the nearest road.

7.2.2 Geocode checking and validation

There are four main methods that can be used for checking and validating geocodes on specimen records. These are the use of databases for checking internal inconsistencies, the use of geographic information systems, the use of environmental space to check for outliers and the use of statistics to check for outliers in geographic or environmental space.

7.2.2.1 Databases

Databases can be used to check for inconsistencies within the data itself. This often involves checking the data in one field with data in another field to make sure it is not inconsistent. For example, checking that towns are in the correct district or State where both are cited. Most of these checks are simple, and generally can only be used to check textual information. More sophisticated use of databases can, however, be used to check the accuracy of the altitude fields by comparing the altitude cited with that of a databased Digital Elevation Model (DEM). It is important that the DEM used be at an appropriate scale, and due to the varying accuracy of most specimen data, can lead to false or misleading errors if not used critically. Such a technique has been used successfully in ERIN (Environmental Resources Information Network) in Australia for over 10 years (Chapman unpublished). The process uses batch processing using an ORACLE database and can check (or assign) altitude records to over 1000 records every 20 seconds.

More recently, sophisticated spatial databases have been developed such as ESRI’s Spatial Database Engine (ArcSDE) (ESRI 2003) that allows for more complicated database searching using the geocodes themselves. This type of software, however, is very expensive, and very few museums or herbaria are likely to afford them or have the need for them and for that reason, these methods are not be outlined further in this report.

7.2.2.2 GIS Checks

Geographic Information Systems (GIS) are very powerful tools that have become much more user friendly in recent times. GISs range from expensive, high functionality systems to free, off-the-shelf products with more limited functionality. Many of the free GISs are powerful enough, however, to provide much of the functionality required by a herbarium or museum, and can be easily adapted to provide a range of data-checking and data cleaning routines.

The use of a simple GIS to plot points (specimen records) against polygons (regions, States, Countries, etc.) can aid in detecting mismatches in the data (either geographic or altitudinal). One of the most important tests a GIS can perform is to check that records that are supposed to be on the land, actually are on land, and those that are supposed to be in the ocean, are. It is obvious, when one first loads a large data set into a GIS, that many records are obviously in the wrong place just from this simple check. Checks for misplaced records using a GIS can range from simple visual inspection to more automated checking. Visual inspection using a GIS can also be valuable in determining if records fall in the correct country, for example. If you have a database of records from Brazil, by using a GIS you can quickly identify records that are misplaced in such a way that they are outside of Brazil. For example, in Fig. 2, records from a publicly available Gazetteer of Brazilian place names have some obvious errors. Errors in specimen records can similarly be identified using this methodology.

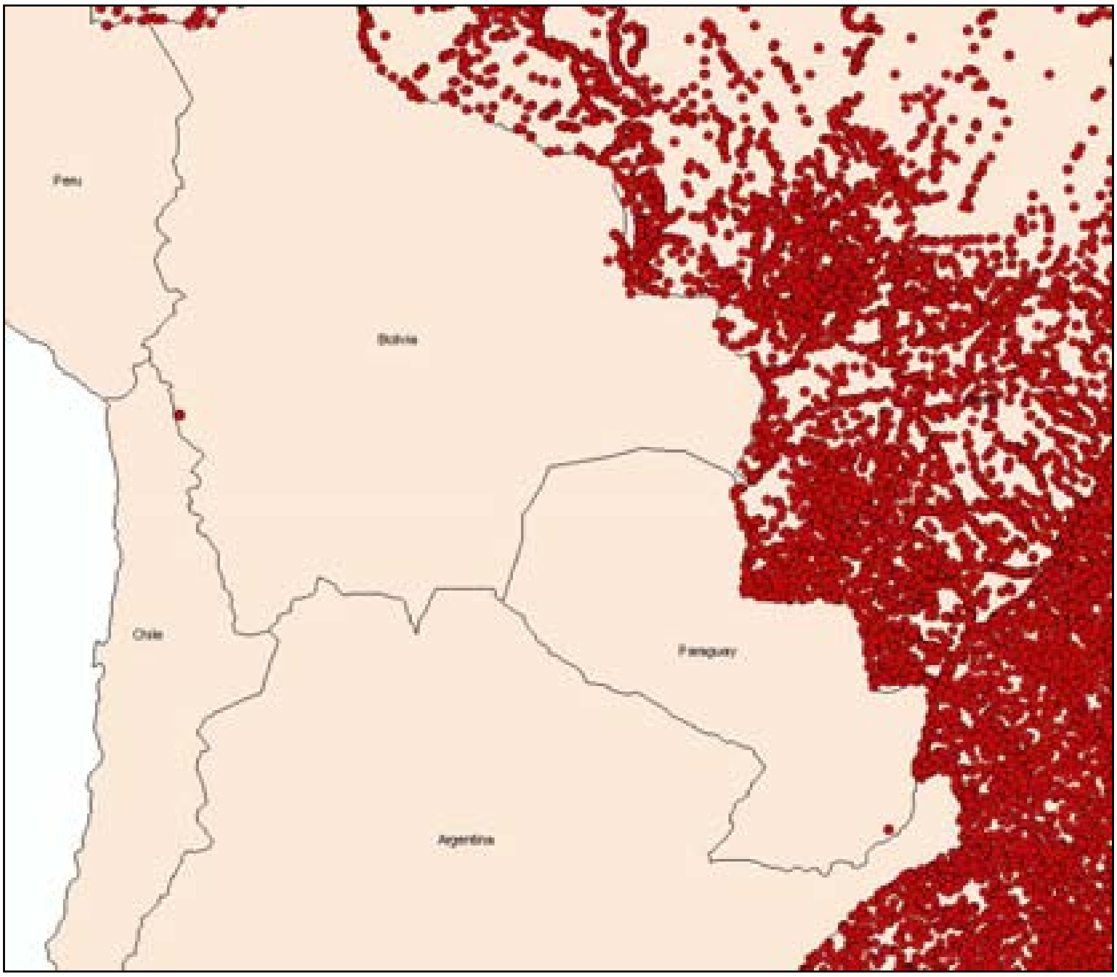


Fig. 2. *Records from a Gazetteer of Brazilian place names showing a number of errors, with one obvious error sitting on the Chile-Bolivian border and another in southern Paraguay.*

A number of the tools mentioned in the next section (for example Diva-GIS) have routines that assist in identifying such errors.

Other uses of GISs include overlaying the specimen points on layers such as soils, vegetation, roads and rivers. If you know information about the records and where they occur, there is no end to how a GIS may be used to help detect errors. For example, if you are working on a fish collection, the buffering of streams may be a method of restricting possible distributions. Some of these methods, however, may be quite time-consuming, and it is often worth examining methods of using automated detection to help find errors in large numbers of records.

Yet another use is to track collectors' itineraries. This can be particularly useful with 18th and 19th Century collectors before the days of collecting by helicopter and being able to cover large distances in a short period by motor vehicle or aeroplane (see figure 28, for example). As the collections database is built up, it is possible to use it to map a collector's itinerary and then use this to check whether other collections by the same collector are likely. An extension of this idea, without the use of a GIS, has been proposed by Peterson *et al.* (in press) (see discussion under 7.2.2.5, below).

7.2.2.3 Using Environments

Methods for checking for geocode outliers in specimen data using environmental layers have been around for around 20 years. As early as the mid 1980s early versions of the program BIOCLIM (Nix 1986, Busby 1991) were used to detect possible outliers by excluding records that fall outside the 90 percentile range of the climate profile for the taxon (Busby 1991), or by using cumulative frequency curves (Busby 1991, Lindemeyer *et al.* 1991). Although these techniques are still in use and are very valuable (Houlder *et al.* 2000, Hijmans *et al.* 2004) they do not allow for taxa that may not include any genuine outliers, or that include many outliers. They are also suspect for very small sample sizes (Chapman and Busby 1994, Chapman 1999).

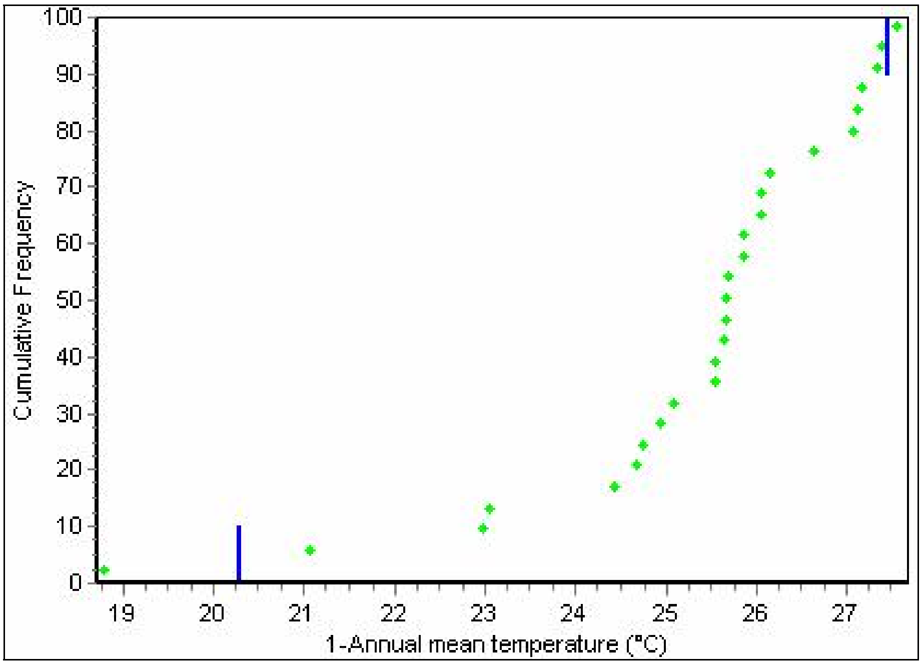


Fig. 3. *Cumulative frequency curve used to detect outliers in climate space using Annual Mean Temperature. The Blue lines represent the 97.5 percentile, the point on the bottom left (or even the two to the bottom left), may be regarded as a possible outlier worth checking for error in the geocode.*

In the early 1990s an automated methodology was developed using reverse jackknifing in conjunction with BIOCLIM to detect possible outliers using batch processing and reporting (Chapman 1992, 1999, Chapman and Busby 1994). This was used in conjunction with a GIS to detect terrestrial records whose errors in geocoding placed them out to sea.

Since the mid 1990s, several new techniques have been developed using a range of outlier detection methods. In addition to the increased use of a simple GIS to plot points (specimen records) against polygons as mentioned above, simple statistic, for example principal components analysis and/or cluster analysis, have been used in software such as Diva-GIS (Hijmans *et al.* 2004) and FloraMap (Jones and Gladkov 2001). More details of these methods are given under the software section, below.

An increasing number of modelling programs are also being used to check for specimen outliers. These vary from looking for records that fall outside of a previously modelled prediction, for example with the use of GARP (Scachetti-Pereira 2002) in LifeMapper (University of Kansas 2003b) and BIOCLIM (Chapman *et al.* in prep.) to included methods that look for outliers in climate profiles in programs such as BIOCLIM (Houlder *et al.* 2000, Chapman 1999), Diva-GIS (Hijmans *et al.* 2004), FloraMap (Jones and Gladkov 2001), and others.

7.2.2.4 Expert validation

The use of expert assessment for validation for species distributions has been around for some years, and is largely the basis of Gap Analysis used in particular in the United States of America as part of the USGS Gap Analysis Program (Jennings and Scott 1997). Informally, it has been used by many organizations to detect errors in specimen data. In 1999, Environment Australia, developed a number of formalised approaches to expert validation (Chapman *et al.* 2001 and in prep.) as part of the Environment Protection and Biodiversity Conservation Act (EPBC) Decision Support System. This involved modelling species using BIOCLIM (Houlder *et al.* 2000), and taking the mapped models, along with specimen point records to meetings of experts from a range of disciplines (taxonomic, ecology, forestry and amateur) to discuss not only the models, but also probably more importantly, the specimen records themselves.

It is important that with any form of expert validation of specimen records, the process be fully documented, and that the experts be required to document their reasons for suggesting a record may be in error. Without such documentation, records could be altered in error, further compounding the errors in the database, and without good documentation as to why a change may have taken place and when, not easy to reverse.

7.2.2.5 Other methods

Peterson *et al.* (in press) have recently suggested a novel statistical method associated with collectors for detecting errors in specimen collections. Using the birds of Mexico as an example, they order the collections of a particular collector in temporal order and for each day (or group of days) impose a maximum radius of likely movement. Using a formula-based approach in EXCEL, they identified possible errors in specimens that fall outside the calculated range. Similar methods to this could be carried out in the database itself, and could even possibly be added for checking at time of entry. Such a method will only work, however, if there is a detailed record of the collector’s itinerary, or if the databased collections from that collector are large enough to create such an itinerary on the fly.

As a result of an earlier draft of this paper, have developed an on-line outlier detection tool (spOutlier-CRIA) for identifying outliers in latitude, longitude and altitude (CRIA 2004b), and includes (February 2004) an additional routine for identifying records that are wrongly located either in the sea for terrestrial species, or on land for oceanic species. Further details are given below.

8. Data Cleaning Tools

As mentioned above, there are a number of data cleaning tools already in existence. Some of these are available without charge; others are available at a cost. Some may be suitable for use in the FAPESP/Biota projects as is, others may need some modification or adaptation. Below, I provide information on each, and provide comments on the usefulness, as I see them, for CRIA and the Biota projects, and especially for those institutions involved in the *species*Link project.

8.1 Web-based tools

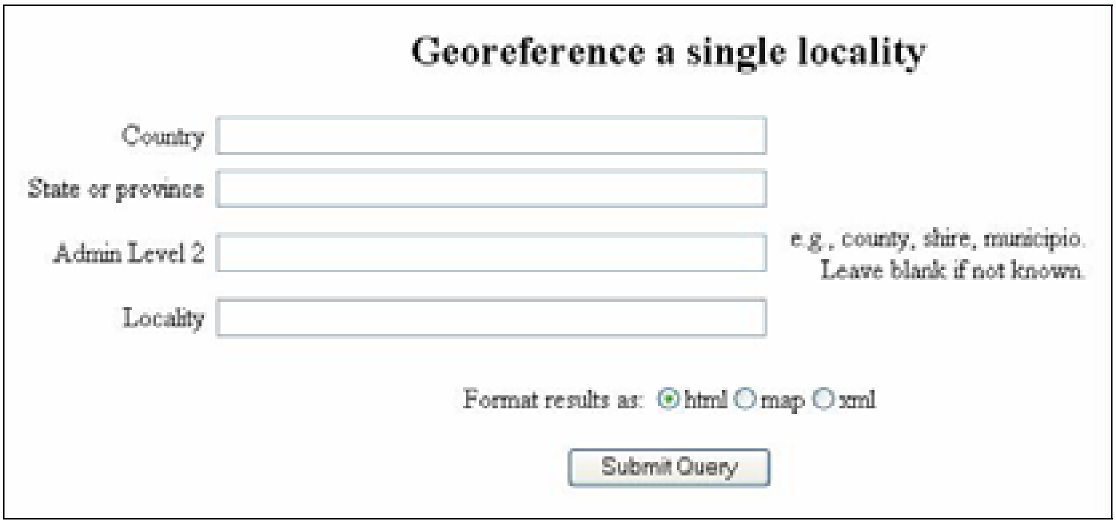
8.1.1 BioGeoMancer

BioGeoMancer is an automated georeferencing system for natural history collections (Wieczorek and Beaman 2002). In its present state, BioGeoMancer can parse English language place name descriptions and provide a set of latitude and longitude coordinates associated with that description. The parsing of free-text, English language locality data provides an output of nearest named place, distance and direction, in the format (Wieczorek 2001a):

- 2.4 km WNW of Pandemonium
- Springfield, 22 miles E
- Springfield, 0.5 mi. E of Pandemonium

The BioGeoMancer is a prototype system at this stage, and the comments below do not take into account planned enhancements that are sure to improve its useability. It is reported that a greatly enhanced version will be available in the next few months.

Like a number of other programs (e.g. Diva-GIS, eGaz) it takes the parsed information and in conjunction with an appropriate gazetteer, calculates a set of latitude and longitude coordinates. BioGeoMancer has the advantage over other geocoding programs in that is provides the parsing of the text. It is the first such geoparsing program available to the public and researchers over the internet.



Georeference a single locality

Country

State or province

Admin Level 2 — e.g., county, shire, municipio. Leave blank if not known.

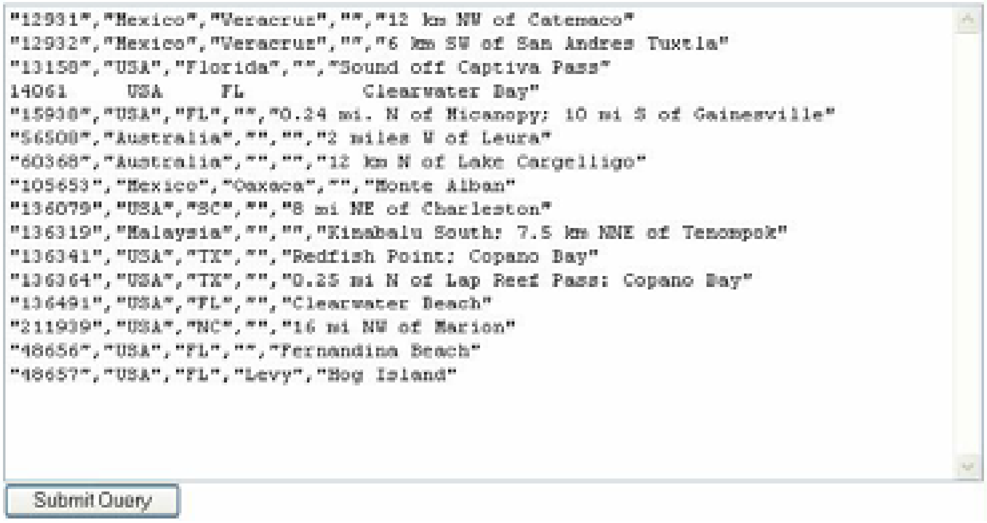
Locality

Format results as: ● html ○ map ○ xml

Fig. 4. *Single locality BioGeoMancer query form http://biogeomancer.org/. (University of Kansas 2003c)*

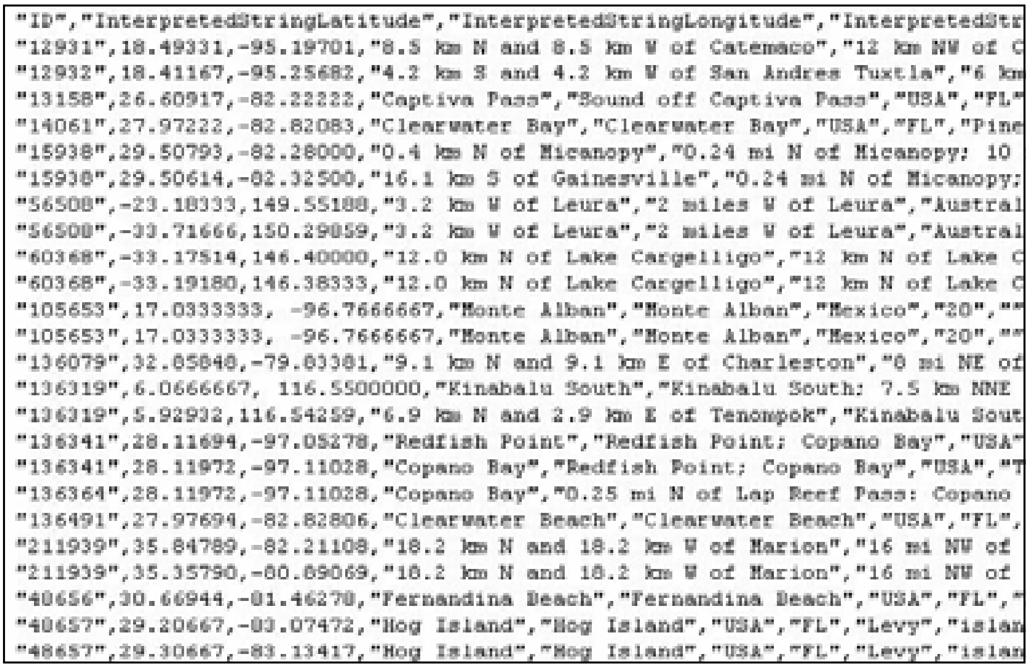
The BioGeoMancer program exists in two forms. The first is a single specimen Web query form (Fig.4) that allows the user to type in a locality and have a georeference returned.

The second form, a batch process, accepts data through either an HTTP/CGI interface in a comma-delimited version (Fig.5) or in a SOAP/XML version and provides a return file with delimited georeferenced records (Fig.6).



```
"12931","Mexico","Veracruz","","12 km NW of Catemaco"
"12932","Mexico","Veracruz","","6 km SW of San Andres Tuxtla"
"13158","USA","Florida","","Sound off Captiva Pass"
14061      USA      FL            Clearwater Bay"
"15938","USA","FL","","0.24 mi. N of Micanopy; 10 mi S of Gainesville"
"56508","Australia","","","2 miles W of Leura"
"60368","Australia","","","12 km N of Lake Cargelligo"
"105653","Mexico","Oaxaca","","Monte Alban"
"136079","USA","SC","","8 mi NE of Charleston"
"136319","Malaysia","","","Kinabalu South: 7.5 km NNE of Tenompok"
"136341","USA","TX","","Redfish Point: Copano Bay"
"136364","USA","TX","","0.25 mi N of Lap Reef Pass: Copano Bay"
"136491","USA","FL","","Clearwater Beach"
"211939","USA","NC","","16 mi NW of Marion"
"48656","USA","FL","","Fernandina Beach"
"48657","USA","FL","Levy","Hog Island"
```

Fig. 5. *Input format for the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections http://biogeomancer.org/bgm-forms/batch-int.htm (University of Kansas 2003c).*



```
"ID","InterpretedStringLatitude","InterpretedStringLongitude","InterpretedStr
"12931",18.49331,-95.19701,"8.5 km N and 8.5 km W of Catemaco","12 km NW of C
"12932",18.41167,-95.25682,"4.2 km S and 4.2 km W of San Andres Tuxtla","6 km
"13158",26.60917,-82.22222,"Captiva Pass","Sound off Captiva Pass","USA","FL"
"14061",27.97222,-82.82083,"Clearwater Bay","Clearwater Bay","USA","FL","Pine
"15938",29.50793,-82.28000,"0.4 km N of Micanopy","0.24 mi N of Micanopy; 10
"15938",29.50614,-82.32500,"16.1 km S of Gainesville","0.24 mi N of Micanopy;
"56508",-23.18333,149.55188,"3.2 km W of Leura","2 miles W of Leura","Austral
"56508",-33.71666,150.29859,"3.2 km W of Leura","2 miles W of Leura","Austral
"60368",-33.17514,146.40000,"12.0 km N of Lake Cargelligo","12 km N of Lake C
"60368",-33.19180,146.38333,"12.0 km N of Lake Cargelligo","12 km N of Lake C
"105653",17.0333333, -96.7666667,"Monte Alban","Monte Alban","Mexico","20",""
"105653",17.0333333, -96.7666667,"Monte Alban","Monte Alban","Mexico","20",""
"136079",32.85848,-79.83381,"9.1 km N and 9.1 km E of Charleston","8 mi NE of
"136319",6.0666667, 116.5500000,"Kinabalu South","Kinabalu South: 7.5 km NNE
"136319",5.92932,116.54259,"6.9 km N and 2.9 km E of Tenompok","Kinabalu Sout
"136341",28.11694,-97.05278,"Redfish Point","Redfish Point; Copano Bay","USA"
"136341",28.11972,-97.11028,"Copano Bay","Redfish Point; Copano Bay","USA","T
"136364",28.11972,-97.11028,"Copano Bay","0.25 mi N of Lap Reef Pass: Copano
"136491",27.97694,-82.82806,"Clearwater Beach","Clearwater Beach","USA","FL",
"211939",35.84789,-82.21108,"18.2 km N and 18.2 km W of Marion","16 mi NW of
"211939",35.35790,-80.89069,"18.2 km N and 18.2 km W of Marion","16 mi NW of
"48656",30.66944,-81.46278,"Fernandina Beach","Fernandina Beach","USA","FL","
"48657",29.20667,-83.07472,"Hog Island","Hog Island","USA","FL","Levy","islan
"48657",29.30667,-83.13417,"Hog Island","Hog Island","USA","FL","Levy","islan
```

Fig. 6. *Sample partial output from the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections http://biogeomancer.org/bgm-forms/batch-int.htm (University of Kansas 2003c).*

Where more than one option is possible, then all are reported under that ID. Where no options are obvious, then the record is not returned.

The system works well for a lot of data, but does have some problems with text that is not easily parsed into the above named place, distance and direction. It needs significant enhancement before it could be regarded as of major value to the general herbarium or museum community as significant pre-processing is required to put the data into a form acceptable by the program.

Other noted problems include:

- It is restricted to English-language descriptions.
- Accuracy is not reported in the present version, and this would be a major enhancement. I understand (Beaman *pers. com.*) that future enhancements are likely to include such a feature. Already, a related program developed by John Wieczorek (2001b) – the Georeferencing Calculator - can supply this information http://elib.cs.berkeley.edu/manis/gc.html and this is likely to be linked to BioGeoMancer at a later date. Already work has begun on a method of assigning accuracy automatically through what has been termed the "point-radius method" for georeferencing and calculating associated uncertainty (Wieczorek *et al*. in press)
- The use of written direction rather than abbreviated direction ("south west" instead of "SW") causes a lack of a return
- Two named localities (e.g. "10 km W of Toowoomba toward Dalby") produces a null result.

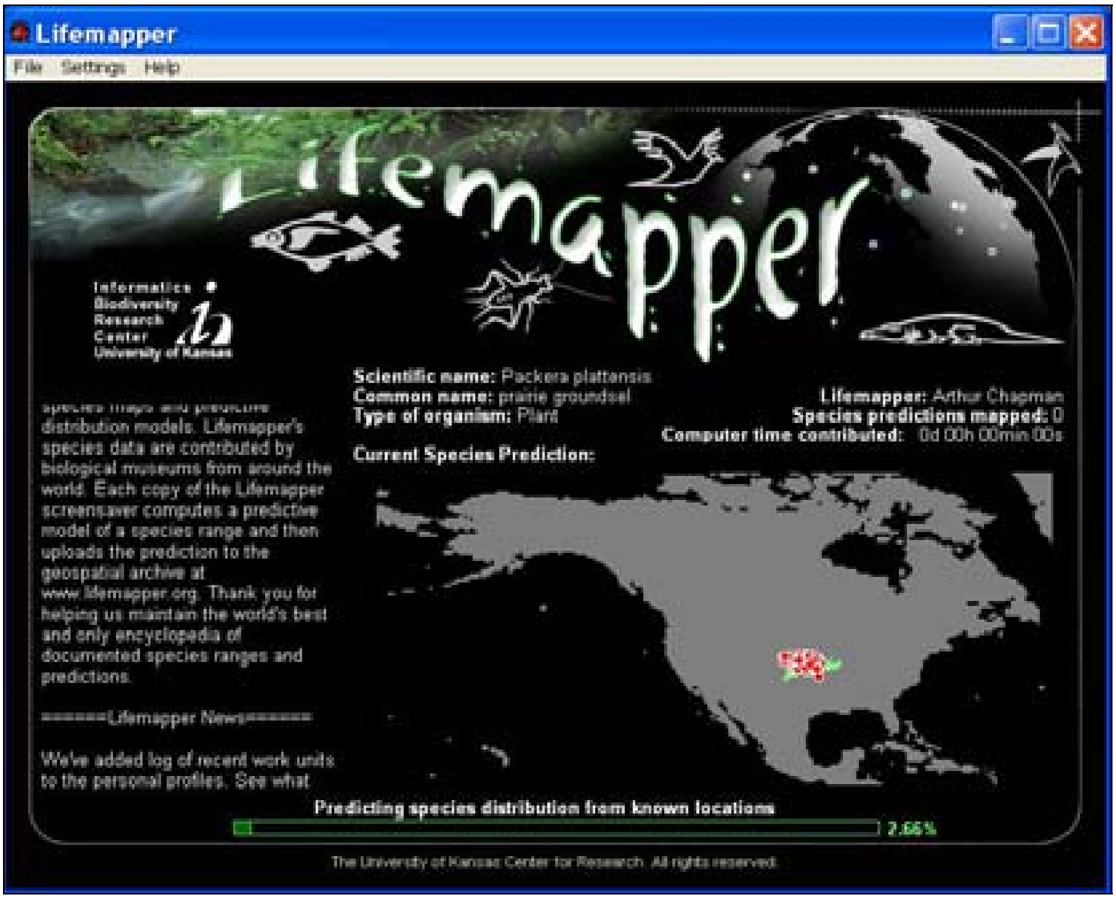
Another parsing program, RapidMap Geocoder (Specht 1997) was developed in 1993 by the US National Museum of Natural History and the Bernice P. Bishop Museum in Hawaii, however it does not seem to have been continued with. Some useful information on the parsing methodologies used, however, is available on the internet at: http://users.ca.astound.net/specht/rm/tr_place.htm.

8.1.2 LifeMapper

LifeMapper (University of Kansas 2003b) is a project from the University of Kansas' Informatics Biodiversity Research Center. Lifemapper's primary goal is to provide maps and predictive models of the World's biodiversity by harvesting the CPU of desktop computers of registered individuals (Fig.7). Lifemapper uses the Internet and leading-edge information technology to retrieve records of millions of plants and animals from the world's collaborating natural history museums, analyse the data, compute an ecological profile of each species using environmental layers such as climate, map the known locations of the species and produce a modelled potential distribution for each species.

At this stage, Lifemapper does not include any data validation or cleaning tools, but it is possible (and is planned - Scachetti-Pereira *pers. com.*) to use the thousands of modelled potential-distribution maps to allow for the checking of new records for possible error. This could be done using the Internet. An institution (or individual) could submit one or more geo-referenced and named point specimen records to Lifemapper, and if the species to which the specimens belong has already been mapped and modelled, the new records could be checked to see if any lie outside the mapped potential distribution. A report could then be returned to the submitter with the results, flagging possible errors (Beaman *pers. com*.).

Although probably some time off being in operation, this could prove a valuable tool for use by museum and herbaria to check collections. For it to work effectively, however, the species to which the specimens being checked belong, must have been previously modelled, and modelled with a significant number of records in order to provide a stable and meaningful predicted potential-distribution. It is unlikely to work for rare species or species for which few records have previously been databased and made available. I also have some concerns at the scale of modelling and the environmental layers being used at present in Lifemapper (Chapman 2003b) using the GARP software (Scachetti-Pereira 2002), however this is likely to improve over time. For checking for possible errors, the small-scale prediction surfaces being used are likely to lead to identification of fewer errors rather than the identification of too many good records as being in error. Also, there would need to be a more representative coverage of taxonomic range for species (in particular plant species) than is the case at the moment (Chapman 2003b).



Scientific name: Packera plattensis
Common name: prairie groundsel
Type of organism: Plant

Current Species Prediction:

Lifemapper: Arthur Chapman
Species predictions mapped: 0
Computer time contributed: 0d 00h 00min 00s

species maps and predictive distribution models. Lifemapper's species data are contributed by biological museums from around the world. Each copy of the Lifemapper screensaver computes a predictive model of a species range and then uploads the prediction to the geospatial archive at www.lifemapper.org. Thank you for helping us maintain the world's best and only encyclopedia of documented species ranges and predictions.

=======Lifemapper News=======

We've added log of recent work units to the personal profiles. See what

Predicting species distribution from known locations

2.66%

The University of Kansas Center for Research. All rights reserved.

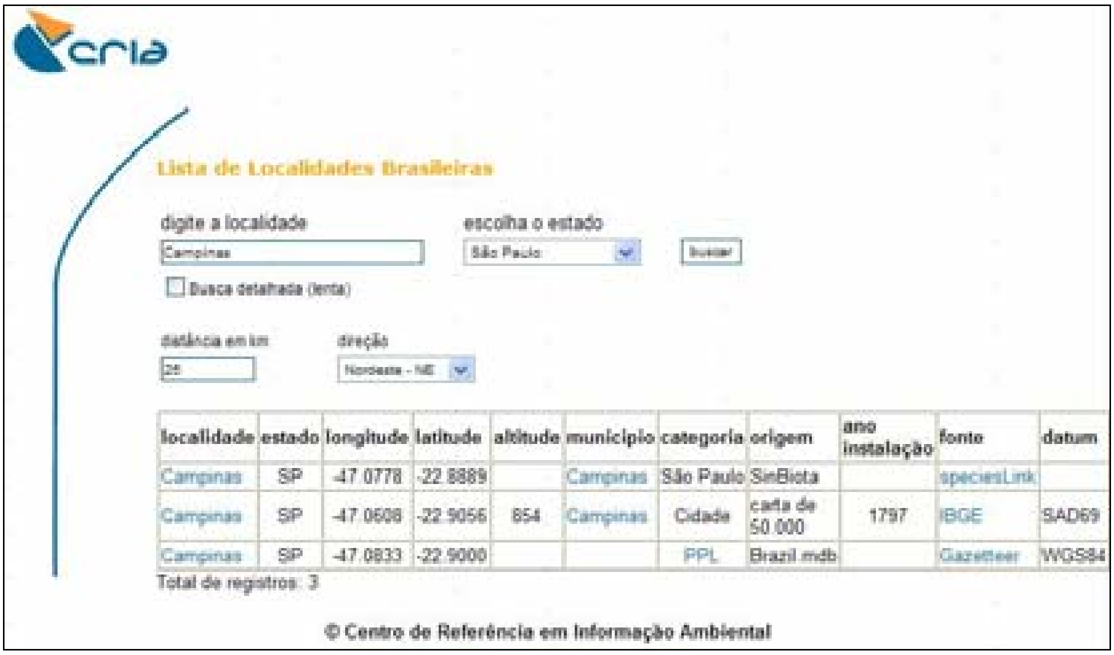
Fig. 7. *An example of a Lifemapper application being run as a Screensaver or desktop application.*

8.1.3 GeoLoc-CRIA and spOutlier-CRIA

Following an earlier draft of this report, simple web-based programs to find a locality in Brazil, a known distance and direction from a gazetted locality has been developed at CRIA, along with simple outlier routines for detecting outliers in latitude, longitude and altitude and for testing off-shore versus on-shore records (Marino et al. in prep.). These two programs are still in the Beta testing phase.

The first, “GeoLoc-CRIA”, works in a similar way to the EGaz program described under 8.2.4 below. A prototype can be found at http://splink.cria.org.br/tools/ (CRIA 2004a). The prototype includes a number of gazetteers and provides the user with the potential to select which gazetteer if more than one is available for an area, and also provides a calculated error value.

An example can be seen in figure 8, where the latitude and longitude of a locality of 25 km NE of Campinas is sought.



Lista de Localidades Brasileiras

digite a localidade: Campinas

escolha o estado: São Paulo

☐ Busca detalhada (lenta)

distância em km: 25

direção: Nordeste - NE

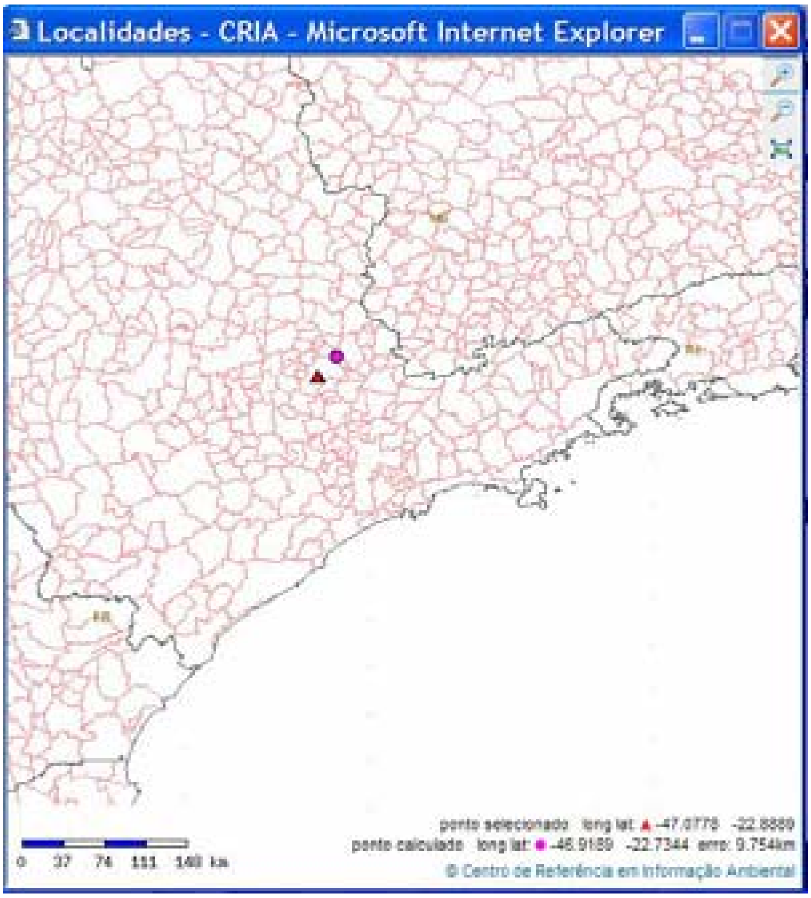
localidade	estado	longitude	latitude	altitude	município	categoria	origem	ano instalação	fonte	datum
Campinas	SP	-47.0778	-22.8889		Campinas	São Paulo	SinBiota		speciesLink	
Campinas	SP	-47.0608	-22.9056	854	Campinas	Cidade	carta de 50.000	1797	IBGE	SAD69
Campinas	SP	-47.0833	-22.9000			PPL	Brazil.mdb		Gazetteer	WGS84

Total de registros: 3

© Centro de Referência em Informação Ambiental

Fig. 8. *Using CRIA’s ‘Localidade – CRIA’ program to find the geocode for a locality 25 km NE of Campinas, SP.*

The results are supplied on an associated pop-up map as shown in figure 9. The geocode is given as -46.9189, -22.7344 with an error of 9.754 km.

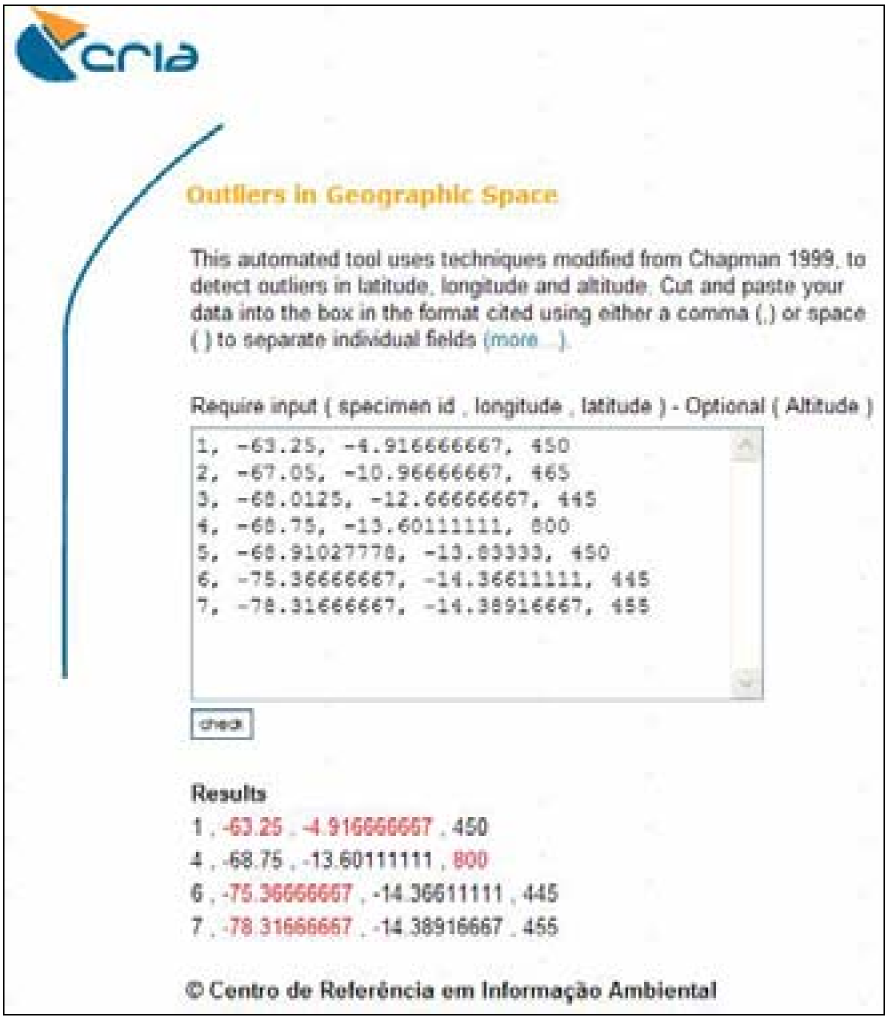

Localidades - CRIA - Microsoft Internet Explorer
0 37 74 111 148 km
© Centro de Referência em Informação Ambiental

ponto selecionado long lat ▲ -47.0778 -22.8889
ponto calculado long lat ● -46.9189 -22.7344 erro: 9.754km

Fig. 9. *Results of the above selection showing the location of "Campinas" and the point 25 km NE of Campinas, with associated geocode information and error.*

The second program (spOutlier-CRIA) allows the user to type (or cut and paste) specimen records into a box on the internet in the form of "id, latitude, longitude, altitude" and the program returns information on likely errors, both in textual form and on a map interface. It also allows the user to identify their data set as either an on-shore (terrestrial) or off-shore (marine) dataset and again the program will return a list of mismatches. This is a unique program, and one that will prove very useful to biologists around the world as it is developed and refined. It is hoped that it will soon be possible for users to submit a document on-line and have it returned, annotated with information on possible errors. A prototype can be seen at http://splink.cria.org.br/tools/ (CRIA 2004b).

In figure 10, the list of localities have returned four records with possible errors, 3 with possible errors in latitude, one with a possible error in longitude and one with a possible error in altitude. These points are then shown on the associated map with the records with possible errors identified in red.



Outliers in Geographic Space

This automated tool uses techniques modified from Chapman 1999, to detect outliers in latitude, longitude and altitude. Cut and paste your data into the box in the format cited using either a comma (,) or space ( ) to separate individual fields (more...)

Require input (specimen id , longitude , latitude) - Optional (Altitude)

```
1, -63.25, -4.916666667, 450
2, -67.05, -10.96666667, 465
3, -68.0125, -12.66666667, 445
4, -68.75, -13.60111111, 800
5, -68.91027778, -13.83333, 450
6, -75.36666667, -14.36611111, 445
7, -78.31666667, -14.38916667, 455
```

Results

1 , -63.25 , -4.916666667 , 450
4 , -68.75 , -13.60111111 , 800
6 , -75.36666667 , -14.36611111 , 445
7 , -78.31666667 , -14.38916667 , 455

© Centro de Referência em Informação Ambiental

Fig. 10. *Shows the prototype Outliers in Geographic Space system at CRIA identifying records 1, 4, 6 and 7 as having possible errors in geocoding.*

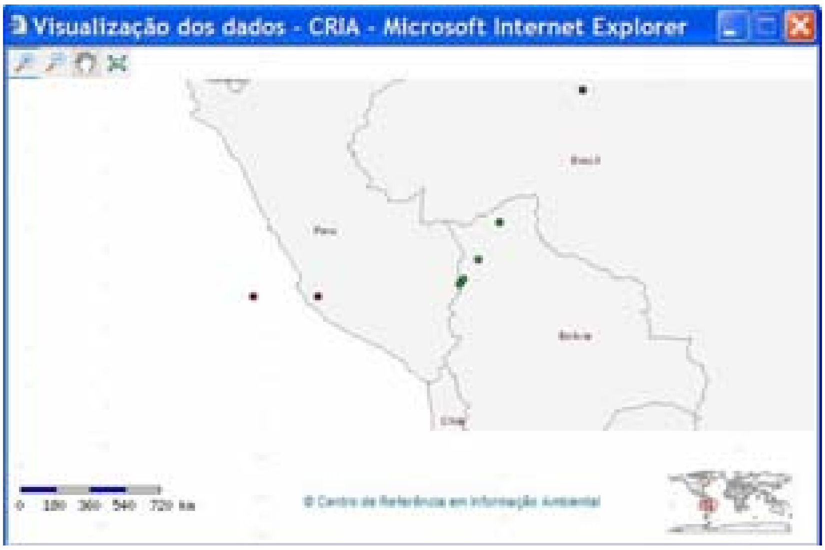

Visualização dos dados - CRIA - Microsoft Internet Explorer
Brazil
Peru
Bolivia
Chile
0 180 360 540 720 km
© Centro de Referência em Informação Ambiental

Fig. 11. *Map output associated showing identified suspect records from fig. 10.*

8.2 Stand-alone Software Tools

There are several stand-alone software programs already developed that can be used to assist in attaching geocodes to specimen records or in identifying possible geocoding errors in specimen data. Most of these are part of larger packages, for example modelling, GIS or data-entry, but can be valuable tools when used for data validation or cleaning.

8.2.1 FloraMap

FloraMap (Fig. 12) is a “computer tool for predicting the distribution of plants and other organisms in the wild” (Jones and Gladkov 2001). It is largely based on the use of Principal Components Analysis to link specimen distributions with climate grids to produce potential distribution maps. The program is available from CIAT (Centro Internacional de Agricultura Tropical) in Cali, Colombia for $US100, and can be ordered on-line (http://www.floramap-ciat.org/inicio.htm).

As part of the package, the program incorporates several tools to help with data cleaning and validation.

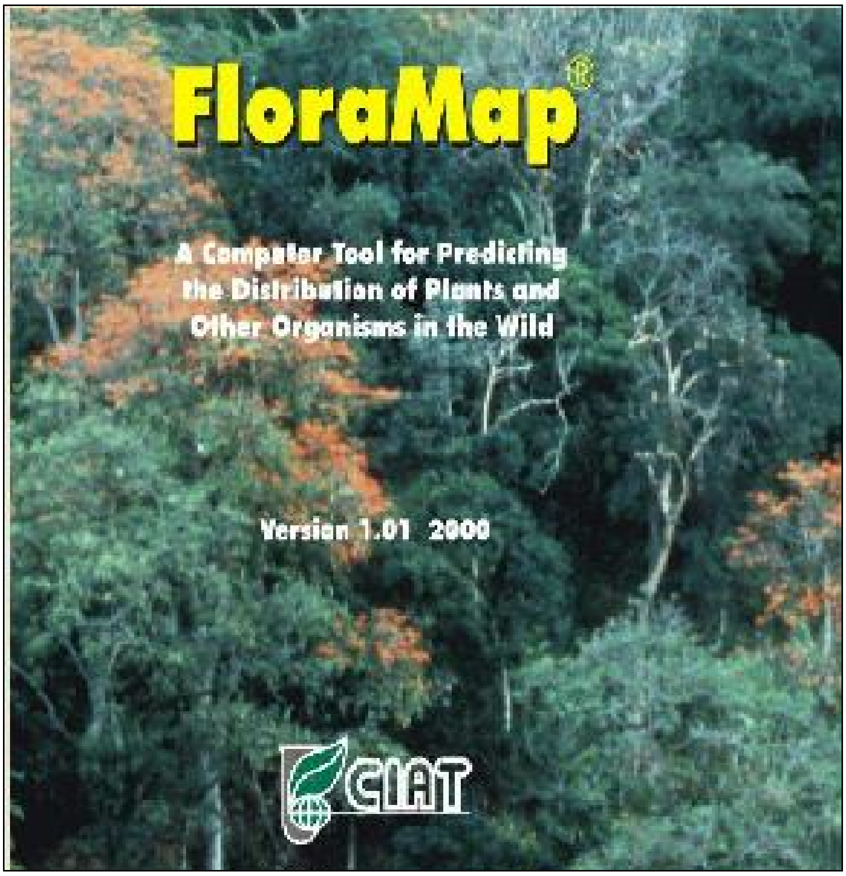

FloraMap®
A Computer Tool for Predicting
the Distribution of Plants and
Other Organisms in the Wild
Version 1.01 2000
CIAT

Fig. 12. *FloraMap – a Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild (Jones and Gladkov 2001).*

8.2.1.1 Use of Climate Grids

The first of these looks for outliers that fall outside of the included 10-arc minute climate grids. Once the data is loaded (via a .dbf file created from EXCEL with a minimum of latitude and longitude fields), and several options set in the Tool box, two files are produced. A file of points that fall within the climate grids, and a second file (mismatch.dbf) of points that fall outside the climate grid (fig.13).

The mapping of *Rauvolfia littoralis* (fig. 13) shows that some records that look perfectly acceptable (i.e. occur on the mainland) are identified by FloraMap as being possible errors. This arises because of the scale of the climate grid (10-arc minutes) being used in the program. The identification of possible errors in this way, however, allows the user to check them and either accept them as being correct, or modify them. The use of climate grids at this scale is a major draw back of the program as can be seen in figure 14 which shows a 10–arc minute climate grid overlayed over a portion of the Carribean.

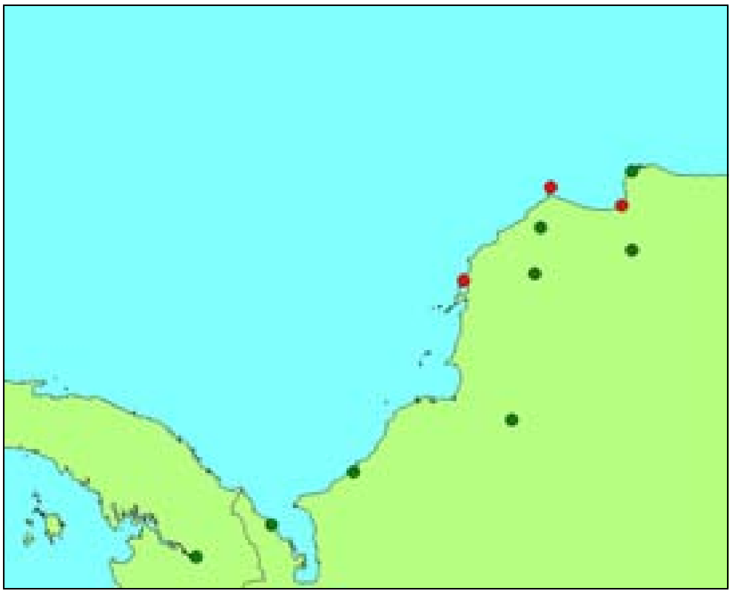


Fig 13. *Accepted (green) and mismatched (red) points for* Rauvolfia littoralis *derived from FloraMap.*

The program has a second option that allows for the automatic movement of the mismatched points to the centre of the nearest 10-minute climate grid. This is of value if the user wishes to use the records in modelling the potential distribution of the species, but has little value for museums and/or herbaria in correcting possible data errors. A similar methodology could, however, be used in a GIS to automatically move records that are "just" off-shore to the nearest land mass (see under Future research ideas below).

8.2.1.2 Principal Components Analysis

A second part of FloraMap that can be used for error checking is through the use of outliers in the Principal Components Analysis (PCA) created in the modelling process.

An example of the use of PCA to identify an outlier can be seen in Figure 15 with *Rauvolfia littoralis*. A possible outlier can be identified from the PCA graph (**A**). When circled using the program's selection tool, the mapped specimen (**C**) flashes on the screen. At the same time, the record from the database (**B**) pops up along with the climate profile (**D**). In this case, the database record (**B**) gives the habitat as "littoral" where as the point (**C**) is shown a considerable distance inland. At the same time, the climate profile (**D**) gives the altitude as 1432 meters. This is an obvious error that requires checking. The use of PCA in this way provides quite a powerful data validation tool, in that it uses climate layers to detect outliers, and can identify errors in altitude as well as geocode positioning or misidentifications. One can also rotate through different PCA graphs to find other outliers in the different combinations of climate components.

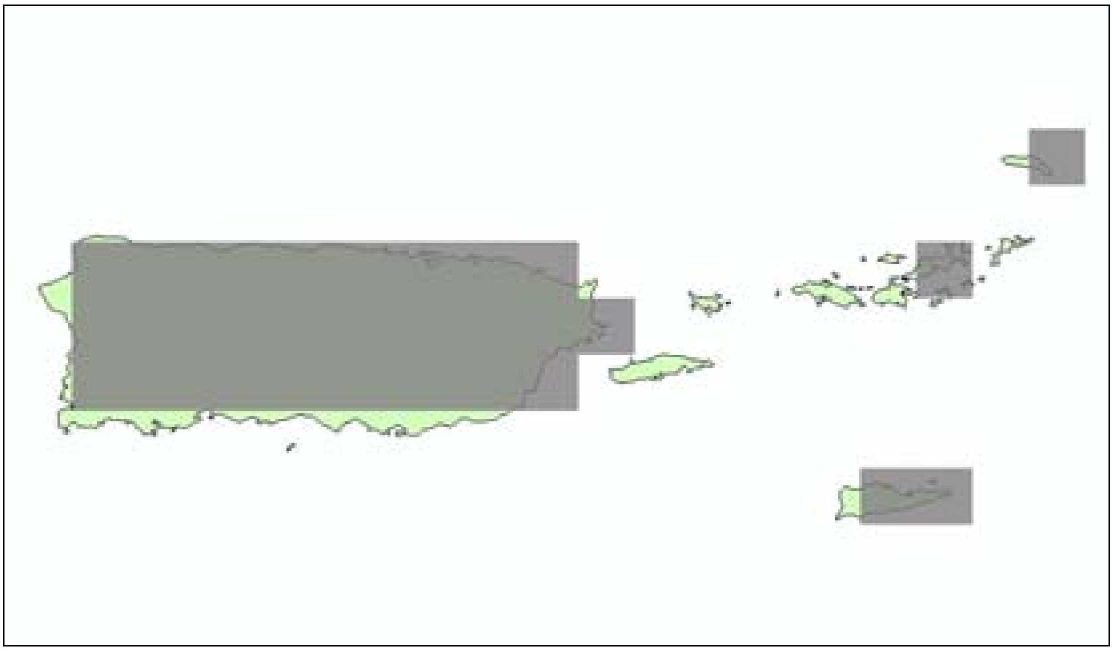


Fig. 14. *Shows problems with using a 10-min climate grid (grey) as a surrogate for islands (green) for determination of errors in specimen data. Some land areas do not fall within a grid square, while parts of the grid cover oceanic areas.*

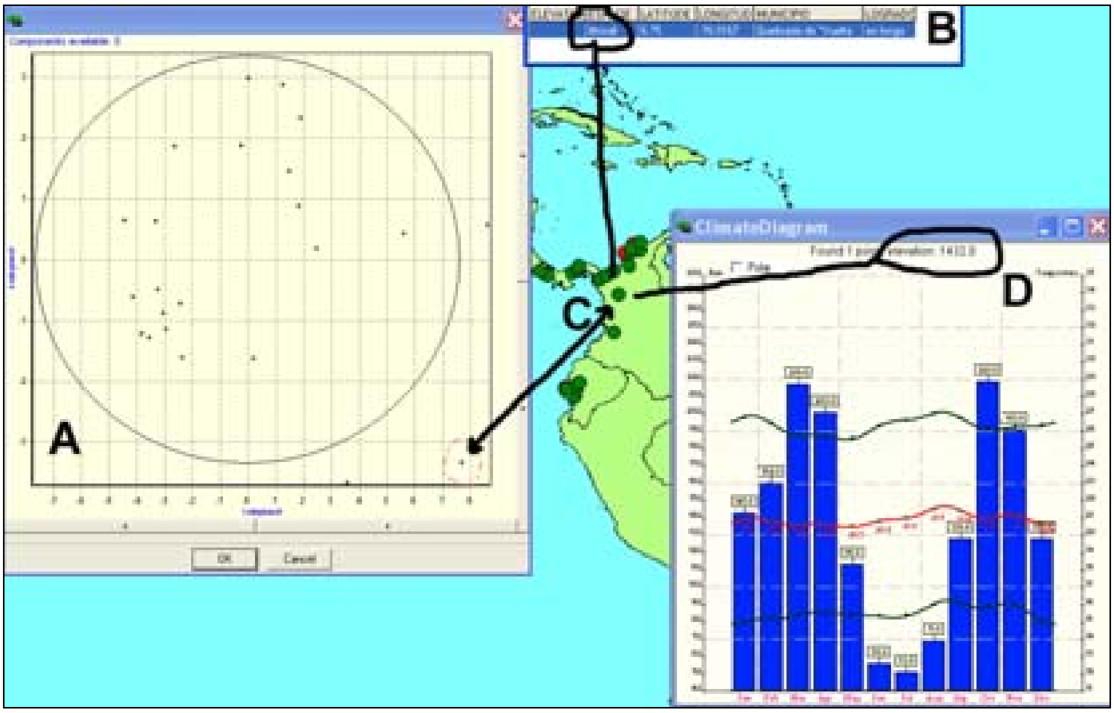


Fig. 15. *Image from FloraMap showing use of Principal Components Analysis to identify an outlier in* Rauvolfia littoralis *specimen data*. **A**. *Principal Components Analysis graph* **B**. *Specimen record.* **C**. *Mapped specimen*. **D.** *Climate profile.*

8.2.1.3 Cluster Analysis

A third part of FloraMap that can be used for error checking is the Cluster Analyis function. The FloraMap program includes the cluster analysis function to help identify possible multiple populations within the data (Jones and Gladkov 2001). The program provides 7 different cluster methods as options and this can be valuable in helping the identification of outliers and possible errors.

Again, using *Rauvolfia littoralis* as an example, one can see how the Cluster Analysis diagrams in FloraMap can identify an error (fig 16). In the example, the Cluster (**A**) identifies several records that are isolated from the other records. In this I have highlighted record No. 8. That record is marked on the PCA (**B**) as blue (as opposed to green for other records). When circled using the program's capture or selection tool, the record (**C**) flashes and, as above, the database record (**E**) and climate profile (**D**) pop up. In this case, the record again appears quite a distance inland and is shown on the climate profile (**D**) as having an altitude of 975 meters. On checking the database record (**E**), the location is given as "Cali". A check of the gazetteer shows there to be several possible localities named "Cali", with one on the coast. As the species is obviously a coastal species, and this it is likely that this was a simple misidentification of locality when the geocode was originally added.

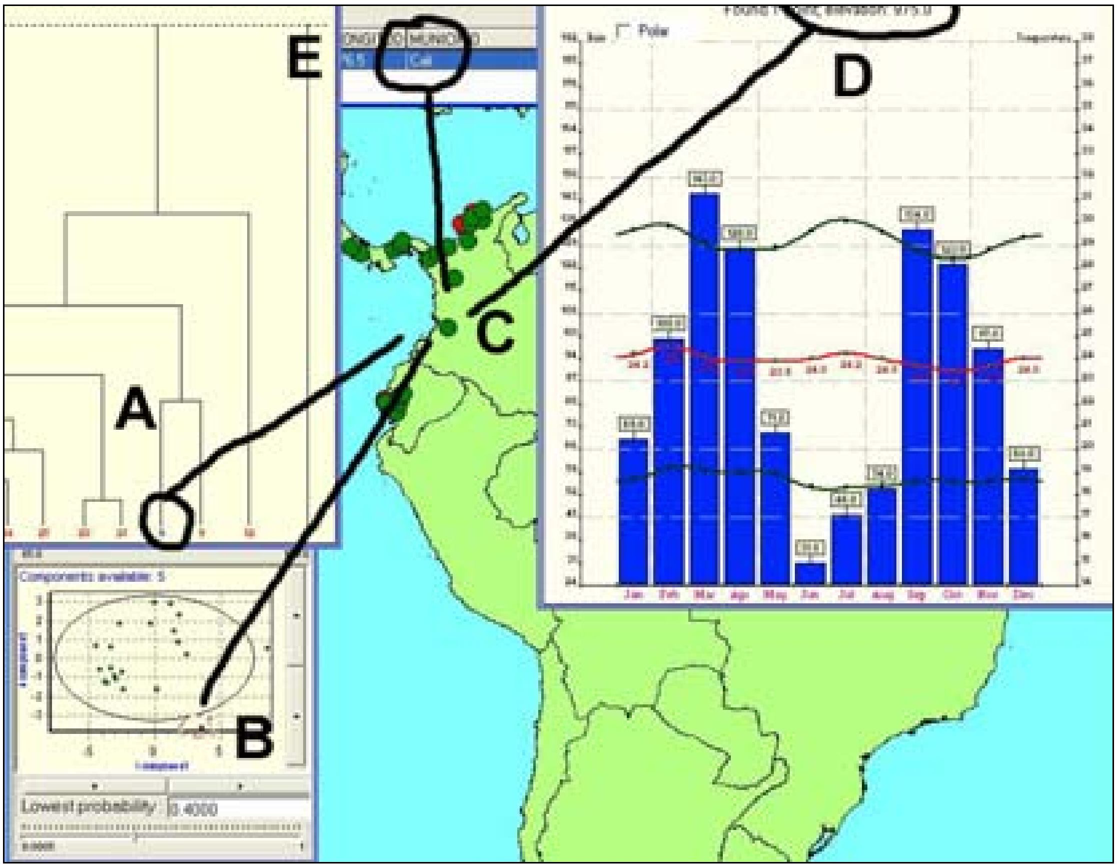


Fig. 16. *Image from FloraMap showing use of Cluster Analysis to identify an outlier in* Rauvolfia littoralis *specimen data.* **A**.*Cluster Analysis* **B**. *Principal Components Analysis.* **C**. *Mapped specimen.* **D.** *Climate profile.* **E.** *Specimen record*

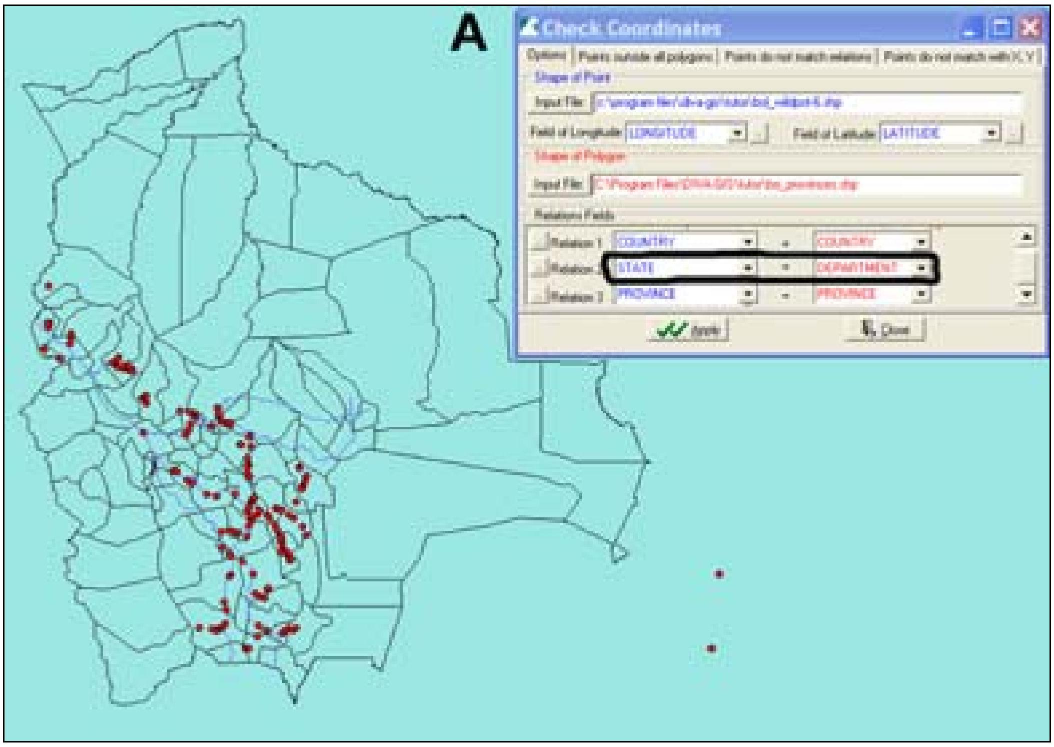
8.2.2 Diva-GIS

Diva-GIS (Hijmans *et al.* 2004) is a geographic information system developed for the analysis of biodiversity data. The software is distributed freely on an “as is” basis. It is MS-Windows based, developed on an ESRI MapObjects® base. It is available for free download from http://www.diva-gis.org. The program was originally developed to support genebank and herbarium databases to elucidate ecological and geographic patterns in plant species data (Hijmans *et al*. 2004). In addition to its basic GIS capabilities, the program also includes several simple modelling algorithms - an early version of BIOCLIM (Busby 1991), Domain (Carpenter *et al.* 1993) and EcoCrop (FAO 2000).

The Diva-GIS package includes a number of data quality checking algorithms worth noting.

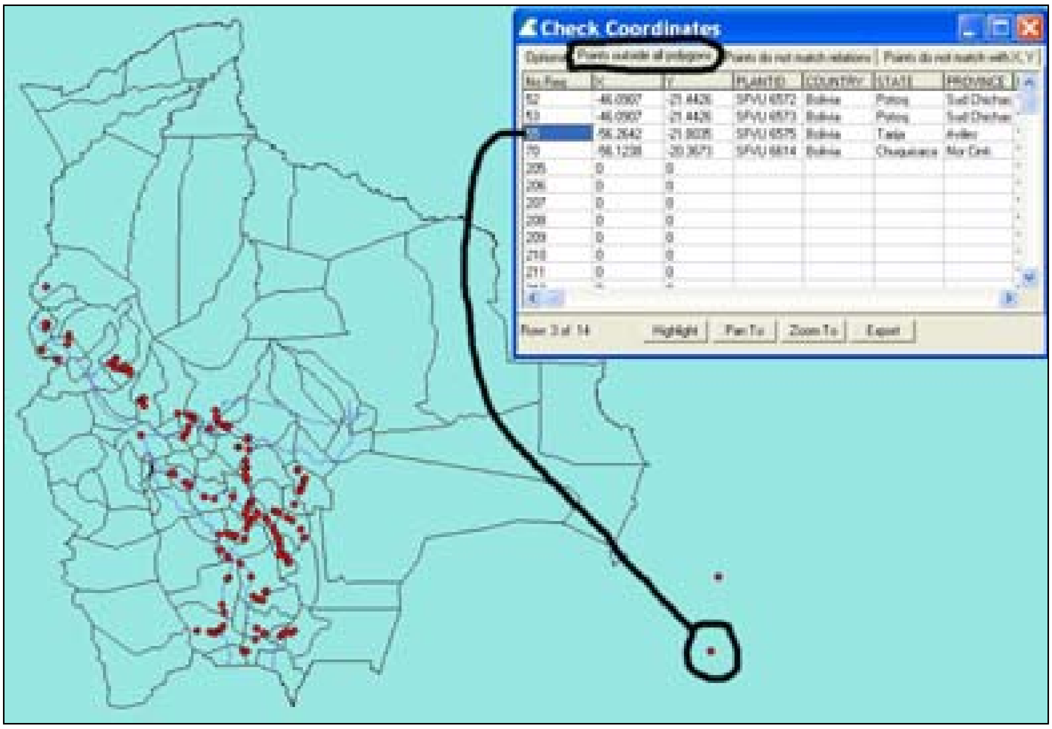
8.2.2.1 Check Coordinates Algorithm

The Check Coordinates Algorithm allows the checking of a file of point specimen records that includes fields such as “State”, “Province”, etc. against a polygon that includes similar attributes (State, Province, etc.). The methodology is described in Hijmans *et al.* (1999). The mismatch-results are mapped on the GIS, as well as being shown in a spreadsheet format that can be exported as a tab-delimited text file. The interface is simple to use, with drop down menus allowing the user to select the equivalents between the two files – for example, one file may call a regional classification “state” while another calls it “department” etc. (see fig.17). The example shown here, uses records from Bolivia of a wild potato species which are supplied with the tutorial for the program (Hijmans *et al*. 2004).



Field of Longitude: LONGITUDE
Field of Latitude: LATITUDE
Relation 1: COUNTRY = COUNTRY
Relation 2: STATE = DEPARTMENT
Relation 3: PROVINCE = PROVINCE

Fig 17. *Using Diva-GIS to check coordinates by comparing a file of point specimen records (red) against a polygon of Bolivian provinces. Input dialogue box is shown at* **A**, *where it can be seen that “STATE” in the point file has been set to the equivalent “DEPARTMENT” in the polygon file.*



Check Coordinates

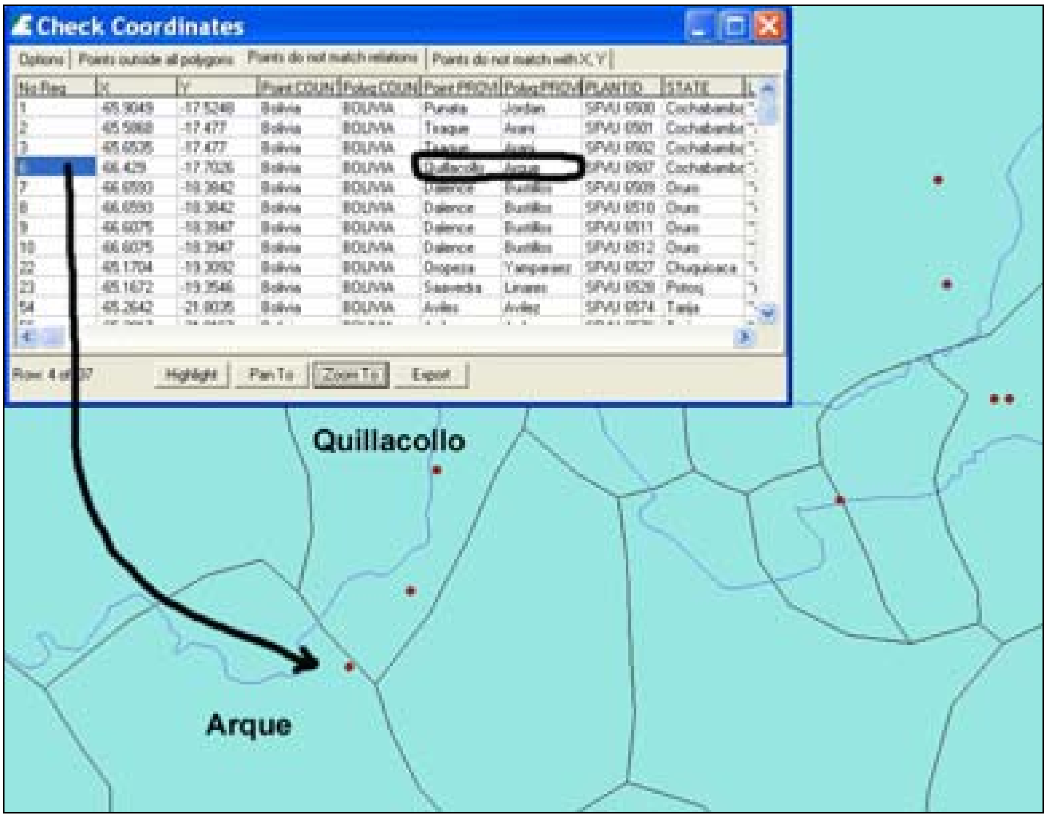
Options
Points outside all polygons
Points do not match relations
Points do not match with X, Y

No Reg	X	Y	PLANTID	COUNTRY	STATE	PROVINCE
52	-66.0907	-21.4426	SPVU 6572	Bolivia	Potosi	Sud Chichas
53	-66.0907	-21.4426	SPVU 6573	Bolivia	Potosi	Sud Chichas
55	-56.2642	-21.8035	SPVU 6575	Bolivia	Tarija	Aviles
70	-58.1238	-20.3671	SPVU 6614	Bolivia	Chuquisaca	Nor Cinti
205	0	0				
206	0	0				
207	0	0				
208	0	0				
209	0	0				
210	0	0				
211	0	0				

Row: 3 of 14

Highlight
Pan To
Zoom To
Export

Fig 18. *Results from Diva-GIS showing point records that fall outside all polygons in the Bolivian provinces polygon file. The highlighted record shows the linking between the results dialogue box and the mapped record.*



Check Coordinates

Options
Points outside all polygons
Points do not match relations
Points do not match with X, Y

No Reg	X	Y	Point COUN	Polyg COUN	Point PROVI	Polyg PROVI	PLANTID	STATE
1	-65.9049	-17.5248	Bolivia	BOLIVIA	Punata	Jordan	SPVU 6500	Cochabamba
2	-65.5868	-17.477	Bolivia	BOLIVIA	Tiraque	Arani	SPVU 6501	Cochabamba
3	-65.6535	-17.477	Bolivia	BOLIVIA	Tiraque	Arani	SPVU 6502	Cochabamba
6	-66.429	-17.7026	Bolivia	BOLIVIA	Quillacollo	Arque	SPVU 6507	Cochabamba
7	-66.6593	-18.3842	Bolivia	BOLIVIA	Dalence	Bustillos	SPVU 6509	Oruro
8	-66.6593	-18.3842	Bolivia	BOLIVIA	Dalence	Bustillos	SPVU 6510	Oruro
9	-66.6075	-18.3947	Bolivia	BOLIVIA	Dalence	Bustillos	SPVU 6511	Oruro
10	-66.6075	-18.3947	Bolivia	BOLIVIA	Dalence	Bustillos	SPVU 6512	Oruro
22	-65.1704	-19.3092	Bolivia	BOLIVIA	Oropeza	Yamparaez	SPVU 6527	Chuquisaca
23	-65.1672	-19.3546	Bolivia	BOLIVIA	Saavedra	Linares	SPVU 6528	Potosi
54	-65.2642	-21.8035	Bolivia	BOLIVIA	Aviles	Aviles	SPVU 6574	Tarija

Row: 4 of 37

Fig 19. *Results from Diva-GIS showing point records that do not match set relationships between the specimen point file and the polygon of Bolivian provinces. The highlighted record where the geocoding on the specimen record causes it to fall in the wrong province.*

When the algorithm is run, several sets of results are produced. The first identifies records that fall completely outside the polygon – in this case, Bolivia (fig.18). Once a record is highlighted in the Check Coordinates Box (in this case record 55), it flashes on the map.

The second output from the algorithm are records that do not match the relations set in the input dialogue box (fig.17), ie. “province = province”, etc. In this example (fig.19), the highlighted record (no. 6) is given in the Point file as being in province “Quillacolla”, but when mapped, the geocode places it in the province “Arque”. The point can to be seen to be very close to the border between the two provinces and perhaps indicates a minor error in the specimen geocoding.

A third output are records that do not match with X and Y. This allows you to check if the coordinates of a point match those in the table (R.J.Hijmans *pers. com*.). This compares mapped locations back to a table, and has less use for checking collections records than the other two.

8.2.2.2 Assign Coordinates algorithm

As well as being able to check coordinates already provided, DIVA-GIS has an algorithm that assists in assigning coordinates to specimen data where this is lacking. Some pre-processing is necessary to organise the data into a format acceptable to the program, but a number of databases are already beginning to structure their data in this way. The input file requires the textual location data to be parsed into a number of specialised fields. These are “Named Place1”, “Distance 1”, “Direction 1” and “Named Place2”, “Distance 2”, “Direction 2”. For example the locality record:

> “growing at a local place called Ulta, 25.2 km E of Chilla”

would be parsed to:

Named place 1:	Ulta
Distance 1:	
Direction 1:	
Named Place 2:	Chilla
Distance 2:	25.2km
Direction 2:	E

and

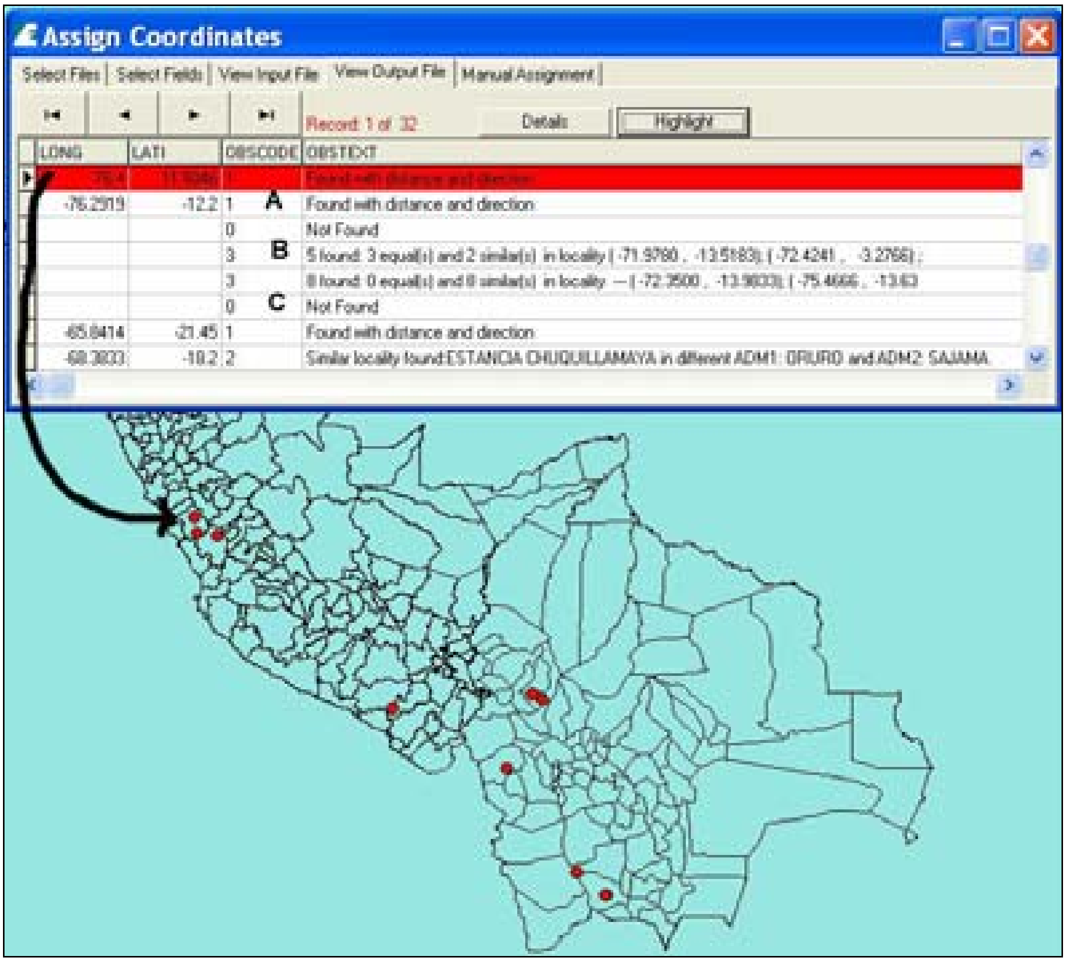
> “14 km ESE of Sucre on road to Zudanez, 1:250,000-scale

would parse to:

Named place 1:	Sucre
Distance 1:	14 km
Direction 1:	ESE
Named Place 2:	Zudanez
Distance 2:	
Direction 2:	

Just one set of “Named Place”, “Distance” and “Direction”, however, will be able to provide the geocoding for a lot of records, and this is all the information most institutions will have. The authors of the Diva-GIS (Hijmans *et al.* 2004) recommend rounding the distance down to whole numbers to account for inaccuracies in the data, and to cater for cases where 25 km North of a place, really means 25 km North by road and not in a direct line. I would recommend to the contrary, and would record the most accurate figure given, and place an accuracy figure in an “Accuracy” field in meters (see discussion under Item 6, Error Checking Methods, above).

Once an input file has been selected, an output file named, and the appropriate field names selected from a pull-down list, the algorithm is run and produces an output file (fig.20). The algorithm uses an appropriate Gazetteer and uses that to assign appropriate coordinates.



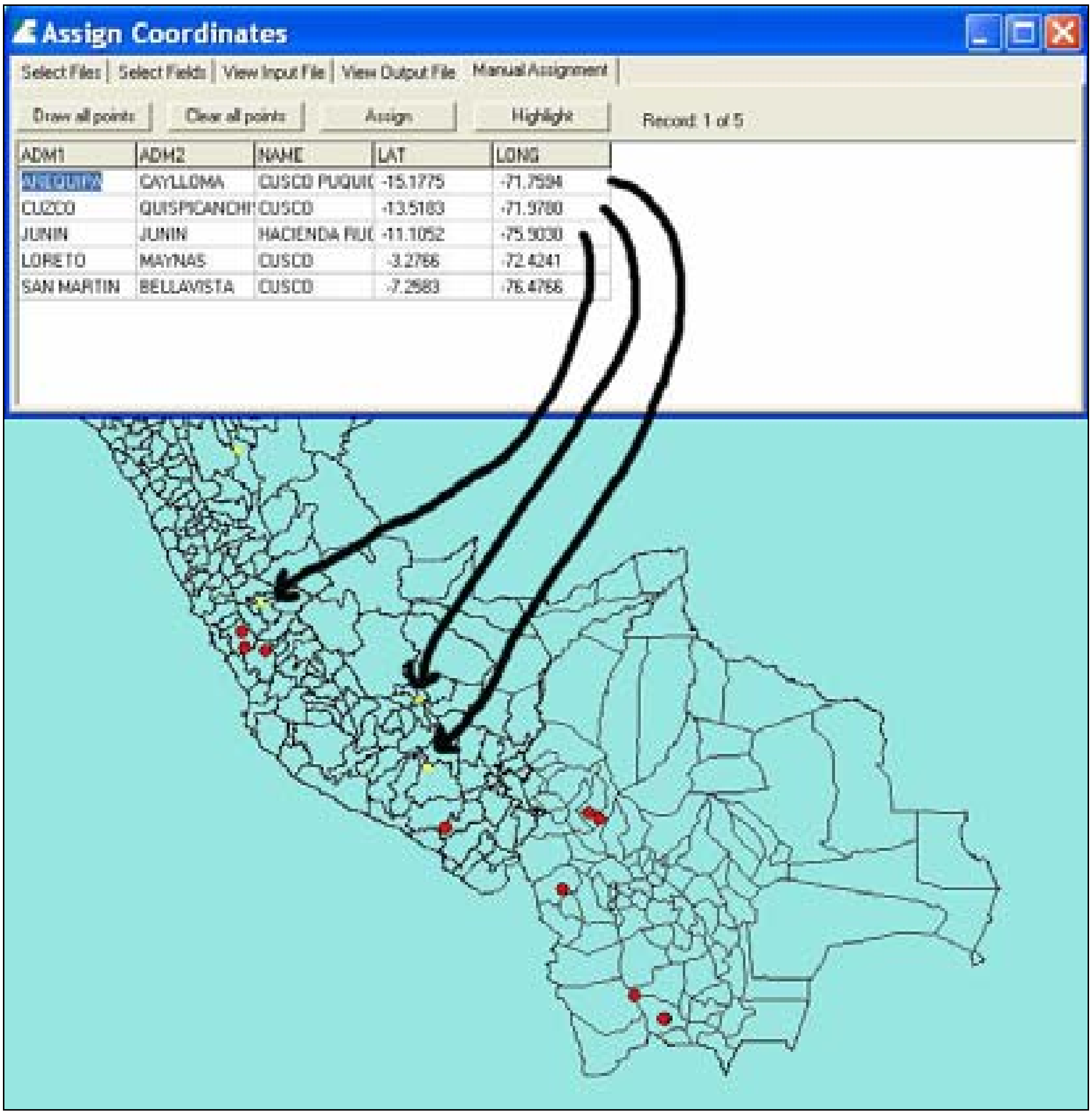
Record: 1 of 32

LONG	LATI	OBSCODE	OBSTEXT
-75.4	-11.9346	1	Found with distance and direction
-76.2919	-12.2	1	Found with distance and direction
		0	Not Found
		3	5 found: 3 equal(s) and 2 similar(s) in locality (-71.9780 , -13.5183); (-72.4241 , -3.2766) ;
		3	8 found: 0 equal(s) and 8 similar(s) in locality -- (-72.3500 , -13.9833); (-75.4666 , -13.63
		0	Not Found
-65.8414	-21.45	1	Found with distance and direction
-68.3833	-18.2	2	Similar locality found:ESTANCIA CHUQUILLAMAYA in different ADM1: ORURO and ADM2: SAJAMA

Fig 20. *Results from Diva-GIS showing point records with geocodes automatically assigned.* ***A.*** *Unambiguous geocodes found by the program and assigned.* ***B.*** *Ambiguous geocodes identified.* ***C.*** *Appropriate geocodes not found.*

As shown in the example (fig. 20), the program has found unambiguous matches in the Gazetteer(s) for a number or records using the “Named Place” field in the input file and assigned those records an appropriately calculated geocode (**A**). Once the output file has been loaded and a shape file created, each of these records can be highlighted to produce a flashing point on the map. In a number of other cases, the program has found several possible matches in the Gazetteer(s) for the “Named Place” and reported on that appropriately (**B**). In yet other cases (**C**) the program has been unable to find a match in the Gazetteer.

In the case of records where a number of possible matches were found, one can go to the next stage by double clicking on one of the (**B**) records and producing another output file (fig.21).



Record: 1 of 5

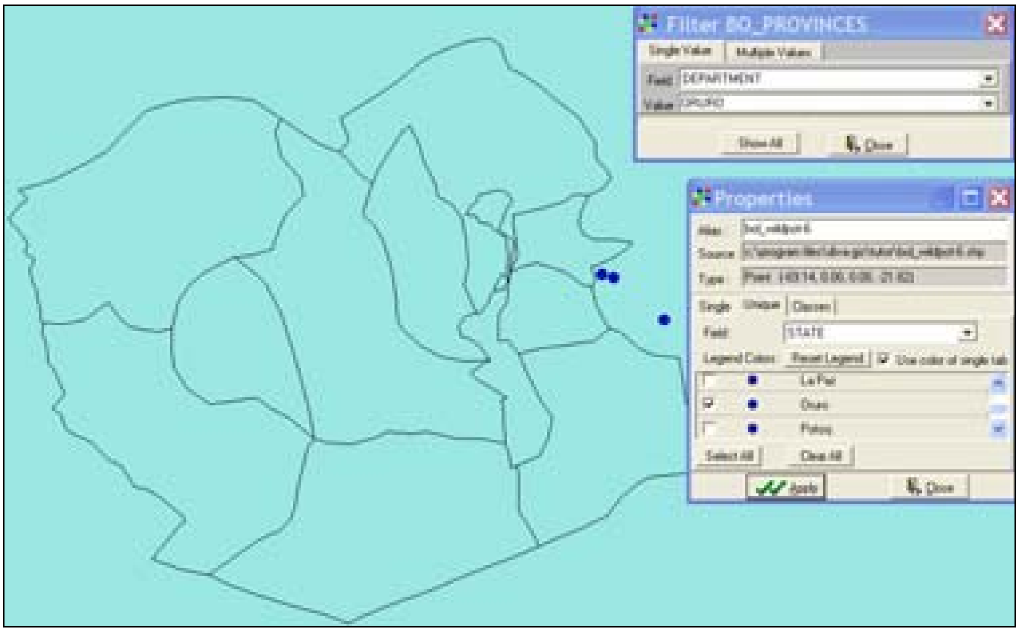
ADM1	ADM2	NAME	LAT	LONG
AREQUIPA	CAYLLOMA	CUSCO PUQUI(	-15.1775	-71.7594
CUZCO	QUISPICANCHI	CUSCO	-13.5183	-71.9780
JUNIN	JUNIN	HACIENDA RUI	-11.1052	-75.9030
LORETO	MAYNAS	CUSCO	-3.2766	-72.4241
SAN MARTIN	BELLAVISTA	CUSCO	-7.2583	-76.4766

Fig 21. *Results from Diva-GIS showing alternate geocodes for a record where use of the Gazetteer has produced a number of credible alternatives.*

In the case of the record shown in figure 21, the program has identified five possible alternative locations from the Gazetteer(s) and presents these alternatives on the GIS for the user to choose. When one is chosen, it is just a matter of clicking on the “Assign” button for that to be assigned to the output file. Alternatively, one can decide on another location altogether and use the “Manual Assignment” to add a geocode or modify one of the assigned ones.

8.2.2.3 Filter option

The Filter option is a new feature of Version 4 released in early 2004 and not available in earlier versions. This option allows the checking of records against polygons such as regions by easily mapping subgroups of the data (regions or species, etc.) For example in fig. 22, the polygon layer has been set to just map the Oruro Department, and the point layer to just show points identified as being in the State of Oruro. From this, it can be seen that the three points lay outside the polygon, and may thus be in error. This is a particularly powerful tool, and one that is quick and easy to use to visualise the data in subsets and to identify likely errors.



Field: DEPARTMENT
Value: ORURO

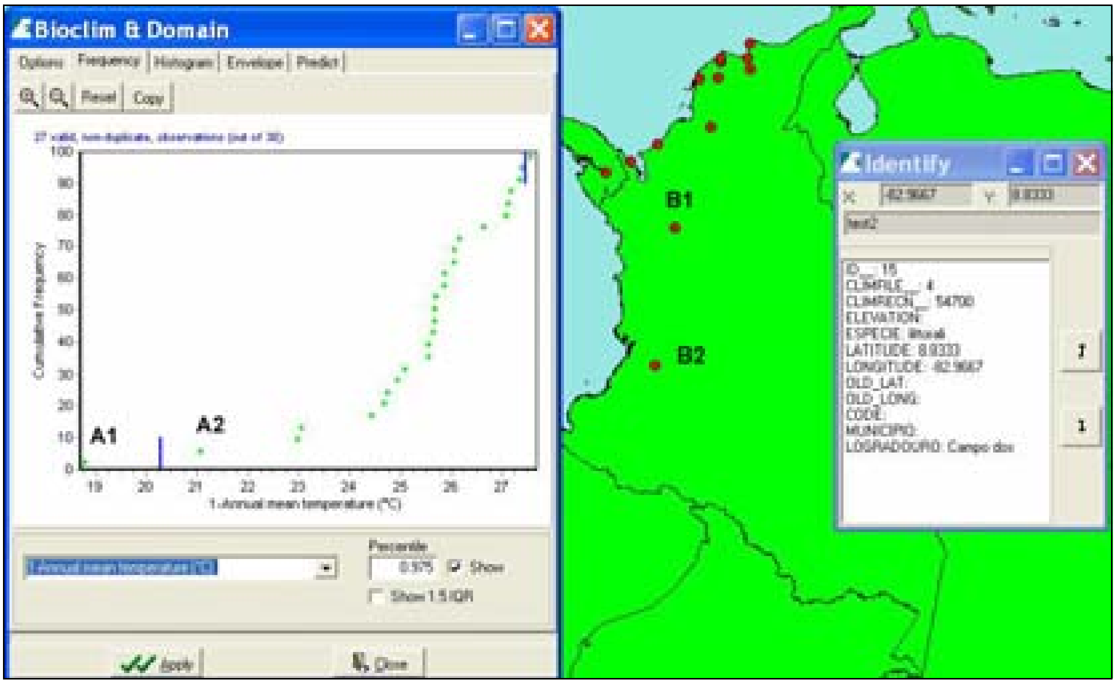
Field: STATE
☑ Use color of single tab

☐ ● La Paz
☑ ● Oruro
☐ ● Potosi

Fig. 22. *Shows the use of the Filter in Diva-GIS to possible outliers in the data and thus possible errors. The three points are identified in the point file as being in the State of Oruro, but the points fall outside of the mapped polygon boundaries for Oruru.*

8.2.2.4 BIOCLIM Cumulative Frequency

Diva-GIS has incorporated a number of modelling algorithms into its structure. One of these is an early version of BIOCLIM (Busby 1991). It is possible, using a number of BIOCLIM features to use the program to identify outliers in various climate parameters (annual mean temperature, annual precipitation, mean temperature of driest quarter, etc.). Using a cumulative frequency curve (see figure 3), one can identify possible outliers. The Diva-GIS implementation has enhanced the original BIOCLIM output by providing an interactive link to the GIS.



1-Annual mean temperature (°C)

Percentile 0.975 ☑ Show

□ Show 1.5 IQR

X -82.9667 Y 8.8333

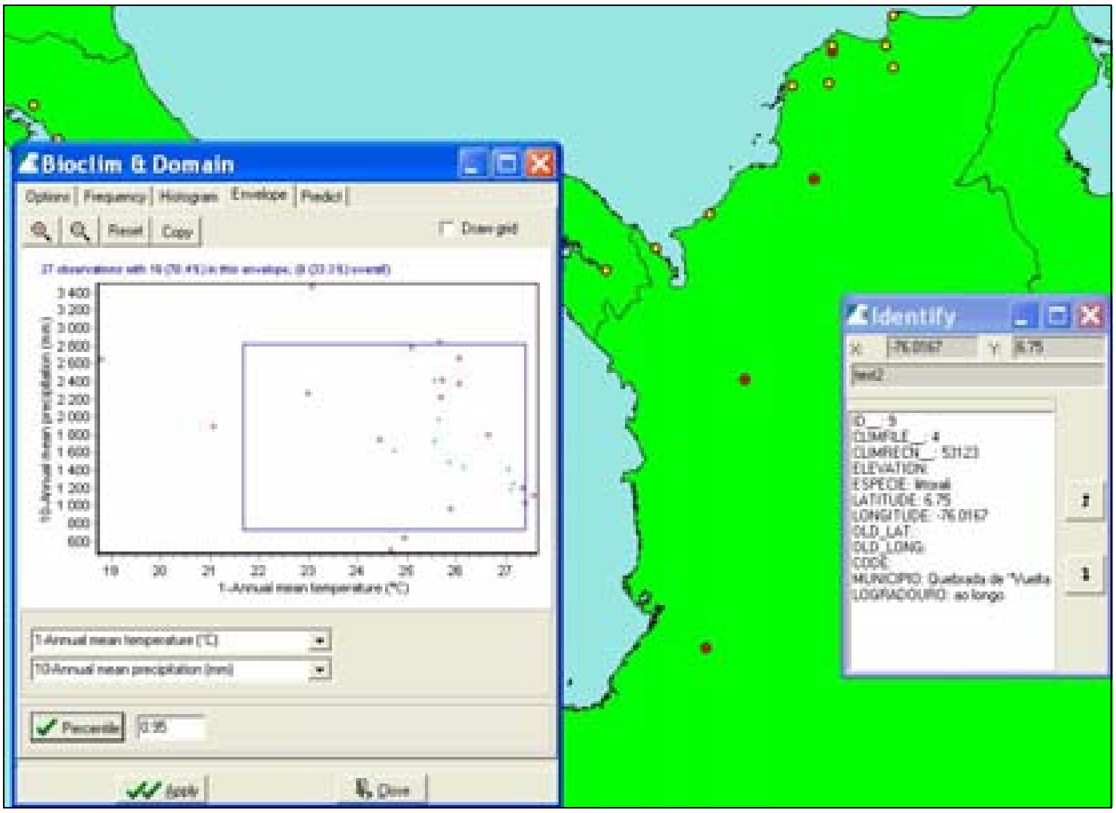
ID_ 15
CLIMFILE_ 4
CLIMRECN_ 54700
ELEVATION
ESPECIE [illegible]
LATITUDE 8.8333
LONGITUDE -82.9667
OLD_LAT
OLD_LONG
CODE
MUNICIPIO
LOGRADOURO Campo dos

Fig 23. *Results from Diva-GIS showing the use of the Cumulative Frequency curve from BIOCLIM to identify possible geocoding errors in* Rauvolfia littoralis. ***A1** and **A2** show possible outliers in climate space, **B1** and **B2** the corresponding mapped records. The Blue lines represent the 97.5 percentile.*

Using *Rauvolfia littoralis* again as an example (see under 8.2.1 FloraMap above), but this time using the DivaGIS program, we are able to identify the same errors as was identified in FloraMap, but using a completely different methodology. The Cumulative Frequency curve for Annual Mean Temperature (fig.23) shows one (**A1**) and possibly a second (**A2**) outlier in the climate space. Click on these points and the corresponding points (**B1** and **B2**) are highlighted on the map. As pointed out above, both these points are likely errors – B1 having within its record a statement that it is from the "littoral" and B2 probably had the wrong "Cali" selected when the geocode was determined, there being another place with the same name close to the coast.

8.2.2.5 BIOCLIM Envelope

The BIOCLIM modelling program is one of a suite of Bioclimatic Envelop methods for modelling species (Nix 1986, Chapman and Milne 1998). The envelope identifies records that fall inside a bounding box at a set percentile range for all the climate parameters used. Diva-GIS uses this method to identify, not only those records, but any records that fall outside the envelop for any one climate parameter (fig. 24). This information can again be used to look for possible outliers. In reality this is an extension of the Cumulative Frequency method above, but shows all climate parameters at once rather than one at a time. It does not allow, however, for the identification of individual records.



27 observations with 19 (70.4%) in the envelope; 8 (33.3%) overall

1-Annual mean temperature (°C)
10-Annual mean precipitation (mm)
☑ Percentile 0.95

X: -76.0167 Y: 6.75
test2

ID_: 9
CLIMFILE_: 4
CLIMRECN_: 53123
ELEVATION:
ESPECIE: Mirzal
LATITUDE: 6.75
LONGITUDE: -76.0167
OLD_LAT:
OLD_LONG:
CODE:
MUNICIPIO: Quebrada de "Vuelta
LOGRADOURO: ao longo

Fig 24. *Results from Diva-GIS showing the use of the Bioclimatic Envelope from BIOCLIM to identify outliers in climate space. In this case the percentile cut off is set at 95. Red points on the envelope correspond with red points on the map, green points in the envelope correspond with yellow points on the map.*

8.2.3 ANUCLIM

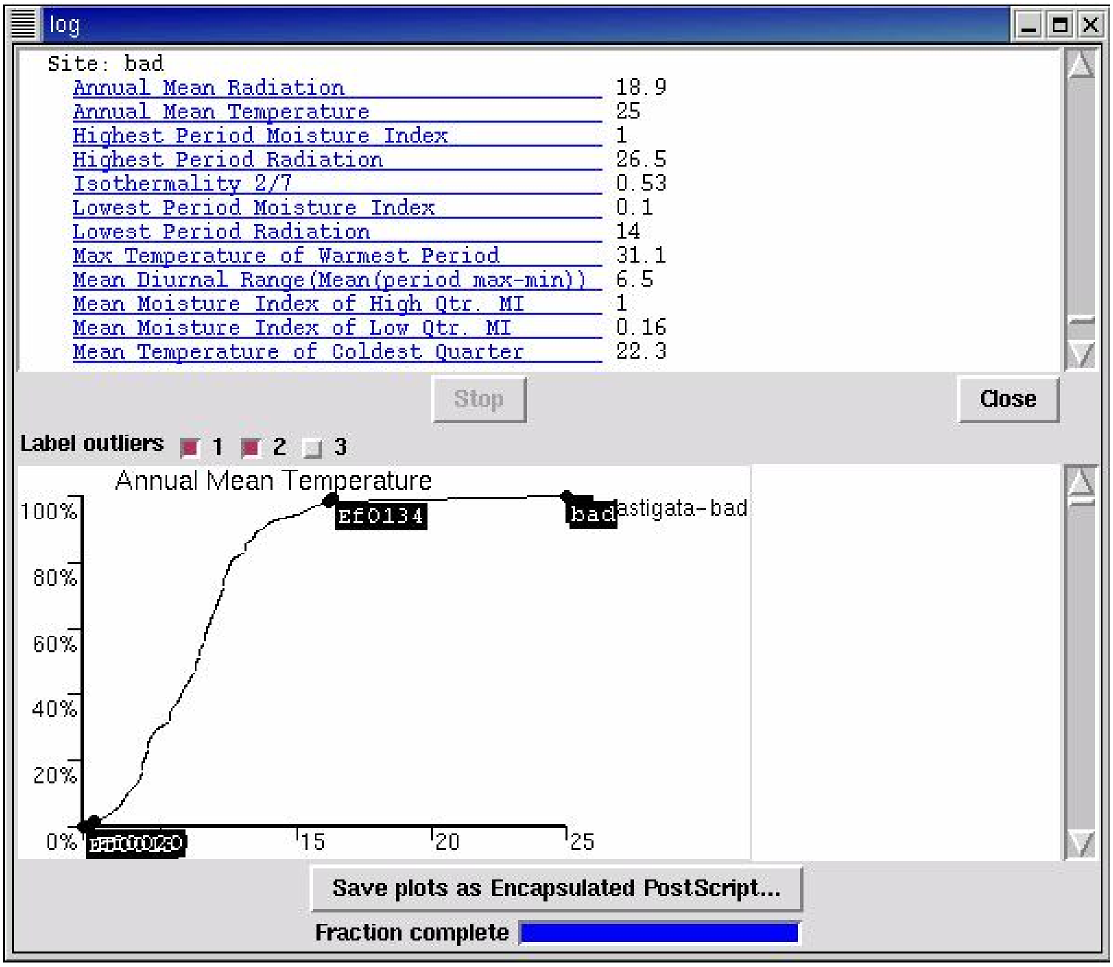
ANUCLIM (Houlder *et al*. 2000) is a software package containing a suite of programs that enable the user to obtain estimates of mean monthly climate variables, bioclimatic parameters, and indices relating to crop growth. The suite includes in its package the most recent version of BIOCLIM (Nix 1986, Busby 1991). Similar to earlier versions (see discussion under Item 8.2.2 Diva-GIS above), the program includes a number of methods for identifying errors in the input specimen data. ANUCLIM is a commercial product available from the Centre for Resource and Environmental Studies (CRES) at the Australian National University for around 1,000 Australian Dollars. Ordering information can be obtained from http://cres.anu.edu.au/outputs/software.html.

BIOCLIM includes two features that enable the identification of suspicious data points in specimen input files. This first is 'parameter extremes' which lists each site that appears as either a maximum or minimum value on one or more parameters. The second is the labelling of outlier points on cumulative frequency plots.

8.2.3.1 Cumulative Frequency Plot

BIOCLIM can be used to generate a species profile that includes two files that can be used to help identify outliers (**.pro** and **.bio** files) in the profile. A log window is produced (fig.24) which has a number of features to help the user to check for possible errors in the input sites file.

The example in figure 24 shows the cumulative frequency plot of the annual mean temperature for *Eucalyptus fastigata*. In this case the site file used to generate the plot contains one suspect data point (labelled "bad"). The log window also shows the climatic information associated with that point to help the user determine if the record is a true error or not. In this example, the log window has been scrolled to show that the site named "bad" is listed as being a maximum or minimum for many parameters. The cumulative frequency plot of annual mean temperature shows a large gap between the bad point and the 2nd and 3rd most highly ranked sites (which are almost coincident on the plot). The labelling check-buttons (see below) have been set to identify the most extreme and 2nd most extreme data points.



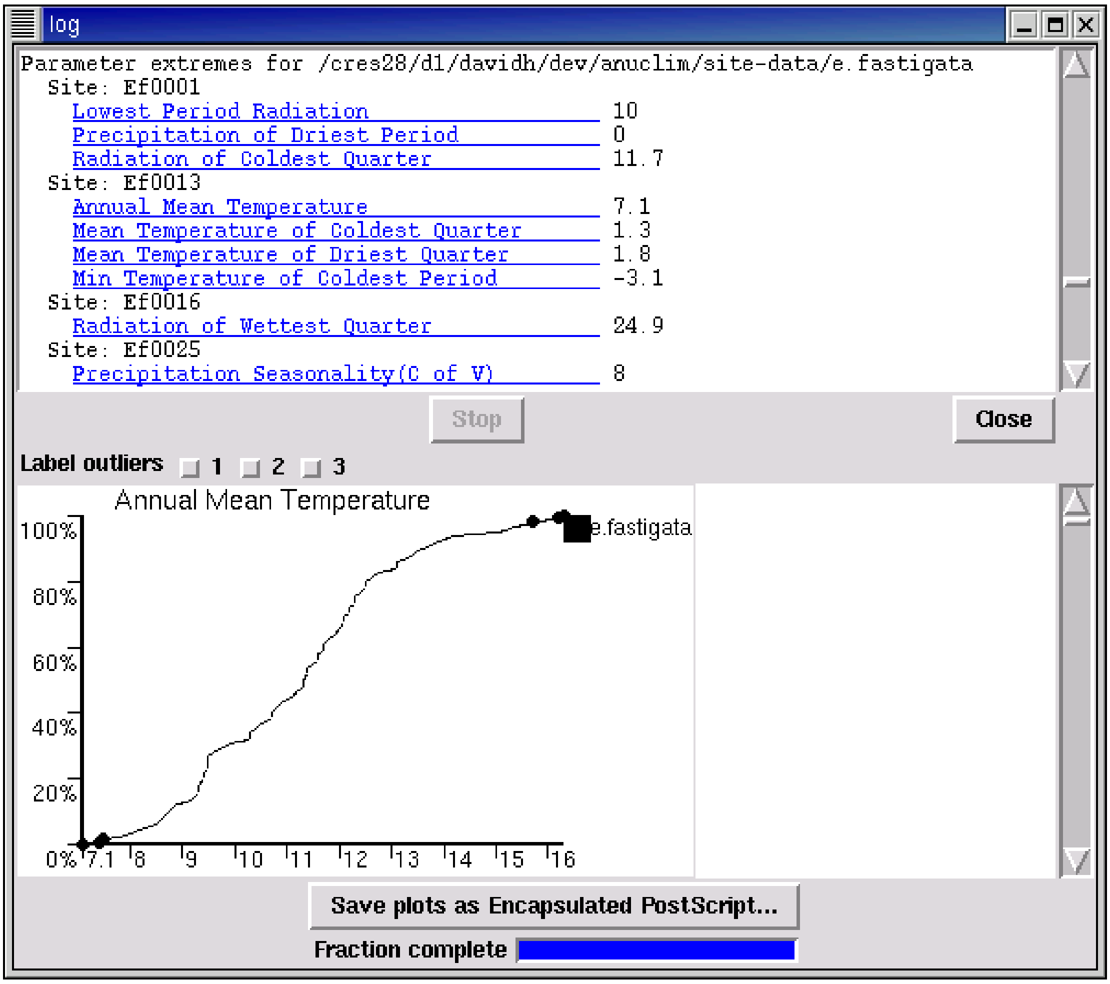
```
Site: bad
   Annual Mean Radiation                          18.9
   Annual Mean Temperature                        25
   Highest Period Moisture Index                  1
   Highest Period Radiation                       26.5
   Isothermality 2/7                              0.53
   Lowest Period Moisture Index                   0.1
   Lowest Period Radiation                        14
   Max Temperature of Warmest Period              31.1
   Mean Diurnal Range(Mean(period max-min))       6.5
   Mean Moisture Index of High Qtr. MI            1
   Mean Moisture Index of Low Qtr. MI             0.16
   Mean Temperature of Coldest Quarter            22.3
```

Label outliers ☒ 1 ☒ 2 □ 3

Fig 25. *Log file of* Eucalyptus fastigata *from ANUCLIM Version 5.1 showing the species accumulation curve with an identified outlier (labelled "bad"). Information from the "bad" record is displayed at the top of the log file (from Houlder* et al. *2000).*

These cumulative frequency curves should be smooth 'S' curves, but if errors in the data are present they can have long tails at either end and can be split so that the two parts of the curve are disjointed. If any of these conditions occur then the data needs to be checked for errors. A single record containing an error in the geocoding will produce a long tail to the curve (see example in fig. 25). Splits in the curve can be caused by the species location data being for two different populations, be due to incomplete sampling of the species, or may be caused by geocoding errors in the dataset. Whatever the cause, these inconsistencies should be checked by checking the geocoding information on the source, or the original data.

The log window includes “Outlier-labelling” check-buttons (fig. 25). These buttons allow the user to display the site labels for outlier points on the cumulative frequency plots. The first check-button will label the minimum and maximum site on each graph. The second will label the next most extreme sites and so on. The number of outliers that can be labelled defaults to 3, but can be changed on the options panel in BIOCLIM or on the “Show parameter profiles” main window.



```
Parameter extremes for /cres28/d1/davidh/dev/anuclim/site-data/e.fastigata
  Site: Ef0001
    Lowest Period Radiation                    10
    Precipitation of Driest Period             0
    Radiation of Coldest Quarter               11.7
  Site: Ef0013
    Annual Mean Temperature                    7.1
    Mean Temperature of Coldest Quarter        1.3
    Mean Temperature of Driest Quarter         1.8
    Min Temperature of Coldest Period          -3.1
  Site: Ef0016
    Radiation of Wettest Quarter               24.9
  Site: Ef0025
    Precipitation Seasonality(C of V)          8
```

Fig 26. *Log file of* Eucalyptus fastigata *from ANUCLIM Version 5.1 showing the parameter extremes (top) and associated species accumulation curve (bottom).*

8.2.3.2 Parameter Extremes

The second method of error detection used in BIOCLIM is the identification of the extreme records for each climate layer in the profile. The parameter extremes list each site that appears as a maximum or minimum value in one or more parameters. Sites that are listed as being a maximum or minimum value for many parameters can be regarded as being of particular concern. Clicking on the parameter name in this display will scroll the window containing the cumulative frequency plots to show the plot in question (fig. 26).

8.2.4 eGaz

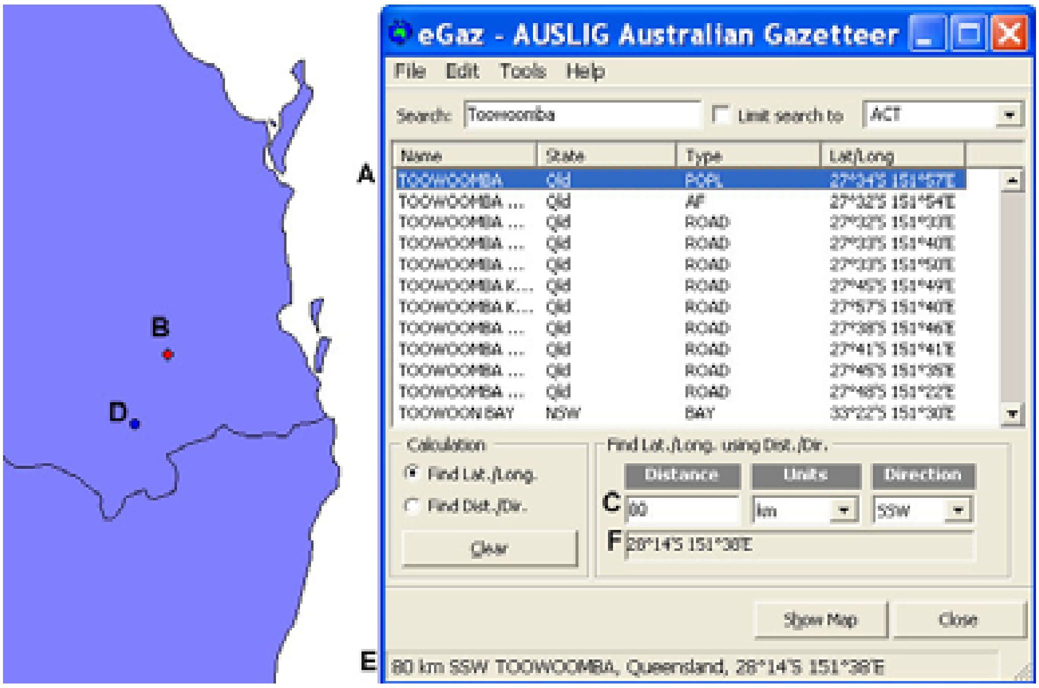
eGaz (Shattuck 1997) is a program developed at the CSIRO's Australian National Insect Collection to assist museums and herbaria to identify and add geocodes to their specimen records. With the development of the data entry and specimen management software, BioLink (Shattuck and Fitzsimmons 2000), it was incorporated into that software package. eGaz is available as part of the Biolink package (see below), but may also be obtained by downloading the stand-alone software from the CSIRO site at http://www.biolink.csiro.au/egaz.html.

BioLink is a software package designed to manage taxon-based information such as nomenclature, distribution, classification, ecology, morphology, illustrations, multimedia and literature and is available for free from CSIRO at http://www.biolink.csiro.au/.

eGaz eliminates the need for paper based maps and rulers to determine the latitude and longitude for cities, towns, mountains, lakes and other named places. eGaz can also calculate latitude and longitude for sites a known distance and direction from a named place. The program allows for the easy inclusion of Gazetteers from any region, and Gazeteers for much of the world are available for download from the CSIRO site (http://www.biolink.csiro.au/gazfiles.html).

EGaz is a Microsoft Windows based product that provides two windows, a Gazeteer window and a Map window (fig.27). It allows the user with a location in the form of a "Named Place", "Distance" and "Direction" to obtain a geocode and transfer that to a file.

The example shown in fig.26 is of obtaining the latitude and longitude of a position "80 km SW of Toowoomba", Queensland, Australia. The first step is to load the appropriate Gazetteer and select "Toowoomba" from it (**A**). There are a number of options, but I have selected the Toowoomba City (labelled POPL for Populated Place). The location of Toowoomba appears on the map in red (**B**). The distance "80" is typed into the Distance field and the pull down menus used to select "km" and "SSW" (**C**). The selected location appears on the map as a blue dot (**D**). The location, along with the latitude and longitude also appears on the bottom of the Gazetteer window (**E**). By right clicking on this area and selecting "Copy" that information can be copied and pasted into any Microsoft Windows compatible file (Word, Excel, Access, etc.). The Latitude and Longitude (to 1 arc-minute resolution) also appears (**F**), and this can similarly be copied to a file. Alternatively, by going to the Edit menu and select "Copy Lat/Long" the geocode can be copied to an accuracy of one arc-second.



Search: Toowoomba □ Limit search to ACT

Name	State	Type	Lat/Long
TOOWOOMBA	Qld	POPL	27°34'S 151°57'E
TOOWOOMBA ...	Qld	AF	27°32'S 151°54'E
TOOWOOMBA ...	Qld	ROAD	27°32'S 151°33'E
TOOWOOMBA ...	Qld	ROAD	27°33'S 151°40'E
TOOWOOMBA ...	Qld	ROAD	27°33'S 151°50'E
TOOWOOMBA K...	Qld	ROAD	27°45'S 151°49'E
TOOWOOMBA K...	Qld	ROAD	27°57'S 151°40'E
TOOWOOMBA ...	Qld	ROAD	27°38'S 151°46'E
TOOWOOMBA ...	Qld	ROAD	27°41'S 151°41'E
TOOWOOMBA ...	Qld	ROAD	27°45'S 151°35'E
TOOWOOMBA ...	Qld	ROAD	27°48'S 151°22'E
TOOWOON BAY	NSW	BAY	33°22'S 151°30'E

Calculation
● Find Lat./Long.
○ Find Dist./Dir.
Clear

Find Lat./Long. using Dist./Dir.

Distance	Units	Direction
80	km	SSW

28°14'S 151°38'E

Show Map Close

80 km SSW TOOWOOMBA, Queensland, 28°14'S 151°38'E

Fig 27. *Sample output from eGaz, showing the determination of latitude and longitude for a position 80 km SSW of Toowoomba, Queensland, Australia.* **A.** *Information on Toowoomba from Gazetteer.* **B.** *Mapped location of Toowoomba.* **C.** *Input showing 80 km SSW of highlighted location.* **D.** *Mapped location 80 km SSW of Toowoomba.* **E.** *Details on location.* **F.** *Latitude and Longitude of new location.*

One can also go to the map itself and zoom in to the point. Other coverages such as a road network (in ESRI Shape file format) can be loaded to allow more accurate positioning of the point – i.e. perhaps move it to the nearest road if collecting was done from a vehicle, etc. The selection tool can then be used to click on the point to obtain the geocode to one arc-second resolution. Again by right clicking with the mouse, or using Edit/Copy Lat/Long, that information can be copied to an appropriate file.

At present, the use of the eGaz program for areas outside Australia has a problem and the use of Distance and Direction from a point does not operate as it should. The rest of the program does work in these areas. The developers have been notified and it is hoped that this bug will be fixed shortly.

A similar internet-based program for use in Brazil is being developed within CRIA (see 8.1.3 above).

8.3 Scripts

Various scripts have been developed around the world to check for geographic outliers or to help in geocoding specimen data. I have mentioned just a couple of scripts here that I am aware of, but there is likely to many dozens, if not hundreds of similar useful scripts is use around the world’s museums and herbaria. The difficulty is in finding them.

8.3.1 CPBR Database script

A database (SQL) script written by the Centre for Plant Biodiversity Research in Canberra, allows for data entry personnel to check the database for similar location records that have already been databased. It works by looking through the database at records already entered for a record from a similar area to that being added. For example, if the data entry operator is adding a record for a species with the location information:

> "26 km NW of Bourke"

The database can be asked to look for records already databased with "Bourke" in the location (or Named-Place) field. It will then sort them, and return a list, for example:

Bourke	30°05'S	145°56'E
Bourke, 10 km N	30°00'S	145°56'E
Bourke, 18 km SW	30°12'S	145°48'E
Bourke 26 km NW	**30°15'S**	**145°45'E**
Bourke 27.2 km NW	30°15'S	145°44'E

Because another collection has already been databased from the same location (may even have been collected by the same collector on the same day), then one just has to accept that, and not spend time recalculating the geocode.

8.3.2 ERIN DEM-Altitude Script

A script was written at the Environmental Resources Information Network, in Canberra, about 12 years ago, to assist in entering altitude or elevation records to specimen data. The script uses a Digital Elevation Model (DEM) at 9-second (250-meter square) grid resolution (but will work with any DEM resolution) to extract the altitude using the geocode of the specimen, and putting the figure into the database. It can be set to either override existing elevation figures in the database, or skip those record that already have an elevation. The Script also uses the DEM to add a code to an extra field in the database to record if the record was offshore ('O'), on the mainland ('M') or on an island ('I'). By far the majority of specimen collections in a museum or herbarium lack elevation information, however, this attribute can be important for many biogeographic, taxonomic, ecological or predictive modelling studies.

The script was written in AML (ArcInfo GIS script) with imbedded SQL scripts, and there is no reason it could not be modified to work with any appropriate database and DEM. The script extract the 'x', 'y', 'z' and 'specimen_id' from the database, uses the Digital Elevation Model to check for onshore, mainland or island and puts that code in the database, finds the elevation for the coordinate and adds that to the database along with the source of the altitude into a source field, and an accuracy figure into an altitude-accuracy field.

8.4 Guidelines

There are a number or written documents available on the Internet that provide valuable Guidelines to the databasing of museum and herbarium specimen data, specimen data management techniques, the geocoding of specimen records, and handling data quality with respect to herbarium and museum specimen data.

8.4.1 HISPID

The Herbarium Information Standards and Protocols for Interchange of Data (HISPID) (Conn 1996, 2000) were first written in 1989 (Croft 1989) to assist in the interchange of data from one herbarium to another. They were originally developed from an earlier document called ABIS, the Australian Biotaxonomic/Biogeographic Information System (Busby 1973). Although not actually used for the purpose of information exchange for many years, HISPID became the standard throughout Australia for the design of herbarium specimen databases. The Australian Herbarium Information Systems Committee (HISCOM) has coordinated the development of the Standard since 1995. The standard is now in its third edition (adopted as a TDWG Standard in about 1997) (TDWG 2003) as a paper based and Internet edition (Conn 2000), and fourth, Internet only, edition (Conn 2002).

The standard's data dictionary is concerned primarily with data interchange but has considerable relevance to database structure. The fields discussed in the data dictionary cover most of the herbarium and botanic gardens sphere of activity and are arranged in groups of similar types of information. In many cases these groups may coincide with separate defined database tables of structurally similar records.

The HISPID standard provides a good basis for developing and managing not only herbarium databases, but also specimen databases generally. The latest standards (Versions 3 and 4) are available from the Royal Botanic Gardens, Sydney site at: http://www.rbgsyd.gov.au/HISCOM/.

8.4.2 MANUS Georeferencing Guidelines

The Georeferencing Guidelines written by John Wieczorek (Wieczorek 2001a) are a valuable resource for anyone wanting information on anything to do with the georeferencing of specimens. It discusses issues associated with adding latitude and longitude to specimen data and the importance of geographic datums. Importantly, the document goes into considerable detail on uncertainty and the determining of accuracy and error from location records. The guidelines were developed at the Museum Networked Information System (MANIS) at the University of California, Berkeley, and are available on the Internet from that site: http://dlp.cs.Berkeley.edu/manis/GeorefGuide.html.

The MANIS site offers some valuable tips to their geocoders that are more generally applicable, viz (Wieczorek 2002):

1. Do not necessarily georeference every locality.
2. Try to recognize particularly difficult localities before spending too much time on them. It is legitimate to add a comment in the "NoGeorefBecause" column that simply says, "Too time consuming to do now." As a rule of thumb, if a locality looks like it is going to take significant amount of time (they suggest 9 per hour for US, 6 per hour for non-US North American, and 3 per hour for non-North American localities), to leave it for later.
3. Group localities from a given region before beginning to georeference them. Work by county or similar administrative subdivision when possible.
4. Wherever possible, filter the records so that you can see at once all of the localities that refer to a given named place.

8.4.3 MaPSTeDI Geocoding Guidelines

The MaPSTeDI (Mountain and Plains Spatio-Temporal Database-Informatics Initiative) Guide to Geocoding (University of Colorado Regents 2003) is a collaborative effort between the University of Colorado Museum, Denver Museum of Nature and Science, and Denver Botanic Gardens. It was developed as part of a project to convert the separate collections into one distributed biodiversity database and research toolkit covering the southern and central Rockies and adjacent plains http://mapstedi.colorado.edu/geocoding-howto.html#toc.

The guide provides a good common sense approach to adding geocodes to a database. It also provides advice on quality checking, an oft neglected feature. One aspect that these guidelines recommend is the importance of using experienced personnel as checkers, and for those adding geocodes to get into a routine that speeds up the process, while at the same time improving on accuracy. An example of a routine from the guidelines is:

- Locate and Plot Your Locality
- Determine a Margin of Error
- Record the Information
- Document the Methods Used
- Mark for Further Review, if Necessary

9. Documentation

Documentation is an often-neglected aspect of data quality checking and validation. It is important that all stages of a data quality checking and validation system be documented. If, due to a validation process, a record is altered, for example a geocode is altered or an altitude altered, then it is important that the change be documented.

Another aspect, often not incorporated in databases, is a flag to indicate that the record has been checked and is correct. An automated validation or checking procedure such as one of those discussed above, may identify an outlier in geographic or environmental space, and thus identify a record as needing to be checked. Once the record has been checked, it may prove to indeed be a good record and a valid outlier. The next time the validation check is run, the record is likely to again be identified as suspect and, if a flag has not been added to show that the record has been checked and is a good record, then valuable time may again be wasted rechecking.

Data storage is not a major issue with today's computers. It is not a major space problem to add a few extra fields for validation and documentation into a specimen database. The fields may be simple "who", "when", "how" and "what" – i.e. who carried out the validation, when they did it, the methodology used and what was the result. These should be attached to each specimen record.

Each institution should also provide clear data entry and checking guidelines for their operators. It is very easy for new or badly trained operators to make a mistake in data entry that is difficult to identify and hard to correct. Good, simple guidelines can help reduce the likelihood of such errors.

10. Risk Assessment

"Risk Assessment" has seldom been considered in detail with respect to the data held in museums and herbaria. As the data are databased, however, it becomes of value to a wide range of people and organizations involved in making major environmental decisions. Many of those decisions involve risk of one sort or another. The quality of the data, upon which the decisions are made, may be a contributor to that risk, and thus needs to be assessed and documented.

Data from museums and herbaria are being used in a number or new ways. For example, the data may be used to assess the likelihood of an agricultural or ornamental introduction spreading and where its likely range may be if it does escape. The success or otherwise of a development application may depend upon its likely impacts upon sensitive environments (Australian Government 1999) – its impact on a threatened species, for example. Modelled distributions based on museum data may be the basis upon which such decisions are made. If the data are incorrect, then there may be both a financial and environmental risk associated with the decision. Museums and herbaria thus need to consider risk assessment as an integral part of their work. The Assessment of risk is generally not an onerous task and the documentation of data quality checking and validation procedures mentioned above should supply most of the information necessary to properly assess and document the risk inherent in the data. Once the data, etc. are documented, it is then up to the user of the information to assess the risk inherent in the use they are putting the data to.

> *Risk assessment is a tool to facilitate informed decision making*
> (Beer and Ziolkowski 1995).

The majority of papers on environment-related risk assessment relate either to pollution or health, or to the interrelationship of pollution and health. There is virtually nothing written on the "green" environment – on the uncertainty in predictions of locations of species or communities, for example, or on the risks associated with the decision making processes based on those uncertainties (Chapman 2002). The application of risk assessment techniques to flora and fauna has been termed ecological risk assessment (Suter 1992) or analysis (Beer and Ziolkowski 1995). These terms, however, are still largely used for the effects of contaminants on plants and animals, on bioaccumulation of toxins and on the use of plants and animals as stressors in screening for pollutants or pollutants (eg. Suter *et al.* 1995), rather than the types of risks inherent in point specimen data in museums and herbaria. While there has been some studies on point-sourced biological data (Austin and Meyers 1995, DNR *et al.* 1997) it has all been done using high quality survey data over small geographic regions.

In an area as complex as the environment, it can be argued that it is impossible to provide a fully objective measure of risk because there is often subjective judgement involved in choosing appropriate data sets, the data sets that are available are often inadequate and not representative, and the massive uncertainties that are usually inherent in biological data can not be quantified. However, there are examples of this risk being quantified as it becomes more important to do so (Beer and Ziowlkowski 1995, Chapman 1999).

I do not intend to delve further into this topic here other than to stress the need to document the accuracy of the data and of the procedures conducted to validate the data and to test its accuracy.

11. Useful Links

There is a wealth of information available on the Internet. For a number of reasons, I do not intend providing an extensive list of links in this document. Apart from anything else, such a list is difficult to maintain, particularly as many URL's are altered and the links become useless, and new links are created. There are, however, a few very valuable sites that are reasonably stable, that are worth citing. In most cases, the sources cited provide links to other resources that readers may find useful. These links are additional to those cited throughout the text and in the References, below. Further links may be found in the Guidelines for Nomenclature (Chapman 2003a). [All links accessed 26 January 2004].

- **BIOSIS Systematics, Taxonomy & Nomenclature Software -** http://www.biosis.org.uk/zrdocs/zoolinfo/stn_soft.htm
- **BIOSIS General Software -** http://www.biosis.org/zrdocs/zoolinfo/software.htm
- **BIOSIS Curation and Collections Management -** http://www.biosis.org/zrdocs/zoolinfo/curation.htm
- **TDWG Subgroup on Biological Collection Data -** Software for Biological Collection Management http://www.bgbm.org/TDWG/acc/Software.htm
- **Internet Directory for Botany –** http://www.botany.net/IDB/
- **Museum Resources on the World Wide Web –** http://www.museumsalberta.ab.ca/network/resource.html
- **Search engines such as Google®** http://www.google.com

12. Ideas for future research/tools etc.

A number of ideas for future research into data validation and cleaning methods spring to mind. They are cited here without any priority or order intended.

- Most GIS packages include a Euclidean distance/direction algorithm that would allow the writing of a simple script to carry a function to move records to the nearest land. Quite often, the geocode given to a specimen that is close to the coast, has an accuracy such that the actual point may appear offshore, although this was not the intention of the geocoder. For example, a record to the nearest minute of latitude or longitude (approx. 2 km) may see the record 500m out to sea. A function such as this could use supervised automatic procedures to move the record to the nearest land. [The FloraMap program (see above) includes an algorithm that allows for the automatic movement of the mismatched points to the centre of the nearest 10-minute climate grid. This is of value if the user wishes to use the records in modelling the potential distribution of the species, but has little value for museums and/or herbaria in correcting possible data errors].
- Many collections include an offset location along a path such as "10 km by road north of …" (Wieczorek *et al.* in press). Some GISs (eg. ArcIMS) include an

algorithm for the calculation of vector distances – distance along a road, a river, etc.)

- It is often known that a collection is restricted to an area associated in some way with another feature. For example, aquatic plants are often associated with streams or rivers. A tool (available in most GISs) could be used to buffer along streams. Also, it is known that many collections (especially in the past) have been collected within a certain distance of a road. A study I did some years ago (Chapman, unpublished), showed that plant collections were highly correlated with road networks and that most botanists apparently walked no more than 0.5 km from the nearest road.
- This buffering of features, for example, may provide a more accurate estimation of error than, for example, the point-radius method (Wieczorek *et al.* in press). This would especially be so in cases mentioned in the point above.
- Once a record is automatically geocoded using a program such as BioGeoMancer (see above), it would be advantageous for the user to be able to then drag that point – to the nearest road, for example, to a place that they believe more accurately reflected its true location. Such a method can be used with the EGaz program, however it is not as intuitive as it may be.
- Gradual build up of collector's itineraries (Peterson *et al.* in press), within collective databases, could lead to the development of a simple tool that could indicate possible error if, for example, the date of collection didn't fit the particular pattern of that collector. This could be particularly useful for collectors from the 18th and 19th centuries. In the example in figure 28, collections between 11 and 25 Apr should be in the Pentland-Lolworth area, if outside that, it is likely to be an error in the date of the collection, or in the geocode. Again, buffering, using a GIS could prove useful.

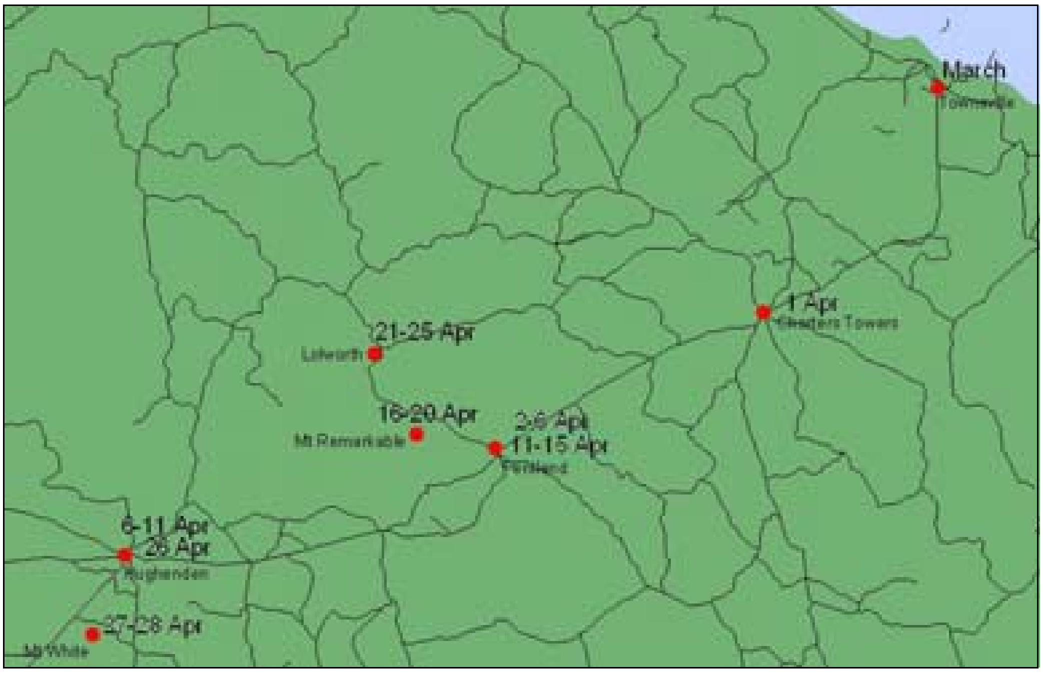
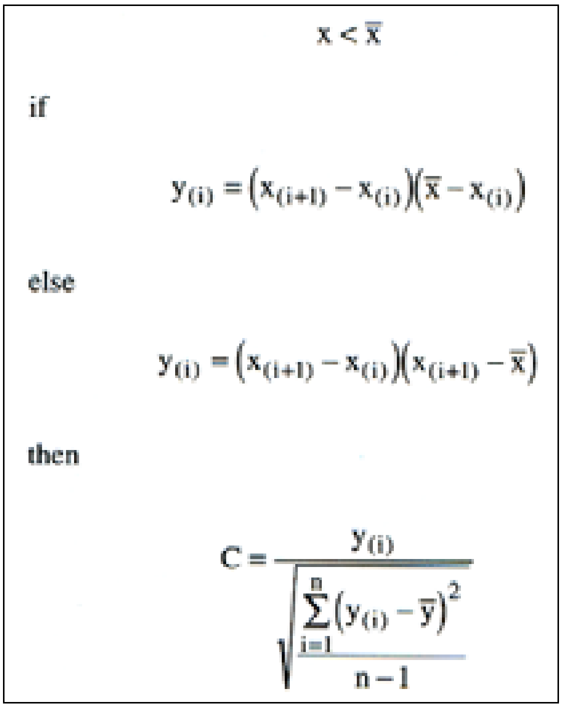


Fig. 28. *Collecting localities of Karl Domin in Queensland, Australia in 1910 (Chapman 1988). He travelled by train from Townsville to Hughenden, stopping at Charters Towers and Pentland. He then returned and spent 15 days in the Pentland, Mount Remarkable, Lolworth area on horseback, before returning to Hughenden by train.*

- The development of a simple tool, see under 8.3.1 above, that could check on entry if the same location had already been used within the database, and thus have already been geocoded.
- Development of simple tools that check for outliers in altitude, climate parameters (see Diva-GIS), latitude, etc. Simple routines could be used to look for statistical outliers within any of these using similar jack-knifing thresholds to those used in Chapman 1999.

  In this formula, the distance between each record and its neighbour is calculated. This figure is multiplied by the distance between the mean and the outer record (i.e. for records less than the mean, the lower of the two records and for records larger than the mean, the higher of the two records is used). The result is divided by the standard deviation to give the Critical value C (fig. 28). If C is greater than the Threshold value (fig. 31) for that number of records, then the record is regarded as an outlier and thus a “suspect” record.

  This method is used to identify an unknown number of outliers at both the top and bottom of an array, unlike many other methods that select only a known number of outliers, or only outliers at one end of the array (Chapman 1999).



if $x < \bar{x}$

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

Fig. 30. *Formula for determining the Critical Value (C) in an outlier detection algorithm where C = Critical Value (from Chapman 1999).*

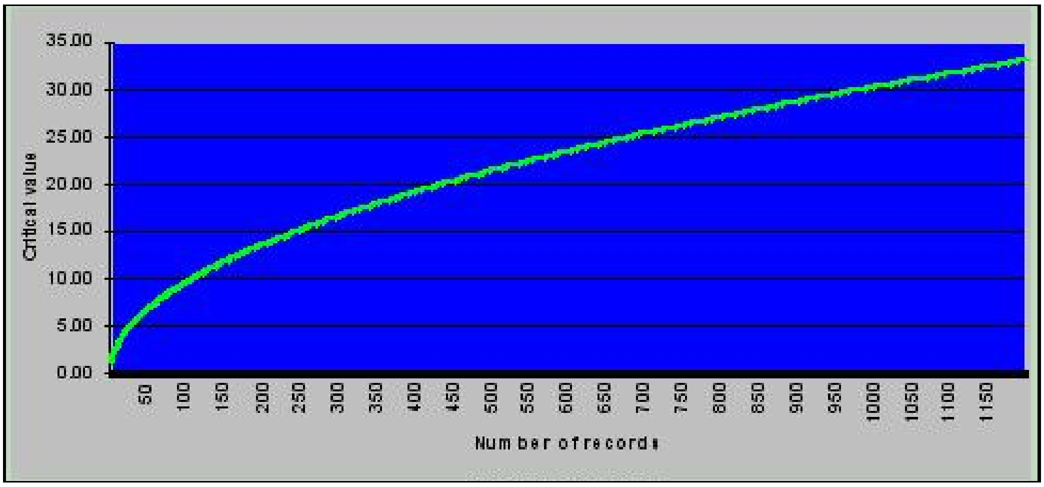


Fig. 31. *Threshold Value Curve (T=0.95(√n)+0.2 where 'n' is the number of records). Values above the curve are regarded as "suspect", values below the curve as "valid" (from Chapman 1999).*

NB. Since writing this in June 2003, CRIA have developed a prototype on-line outlier detection tool as described in 8.1.3 above.

13. Options

There are a number of optional ways of proceeding from here. I recommend that a data-cleaning and validation toolkit be prepared and made available to institutions via a CD-Rom. As well as containing stand-alone software, the CD-Rom would contain guidelines and documentation and links to Web-based tools. I would suggest that the CD contain:

- *Stand-alone software*
- *Guidelines*
- *Links to Web-based tools*
- *Links to documentation and software*
- *Documentation*

Consideration would need to be given to what was included. Some of the software mentioned above is public-domain software, however, I believe that in some cases, some of the software could be made more valuable for such a toolkit with some modification. Other software mentioned is not free, and there may be value in extracting some of the methodologies and developing them into a separate software package.

13.1. Stand-alone software

Several examples of stand-alone software have been discussed above. The possibility and advisability of inclusion of some of the methods is discussed below.

13.1.1 FloraMap

FloraMap contains several methodologies for data validation and cleaning. The first of these looks for records that fall outside of a climate grid. Diva-GIS, on the other hand, has similar routines that can be used to identify records that fall outside a polygon. The polygon method is a preferable method and I would recommend is the better option in this case.

The second method uses Principal Components Analysis to identify possible outliers. In comparing this method with the BIOCLIM methods used in Diva-GIS, in the examples tested, both methods seem to identify similar records as possible errors. Ideally, one would like to include both options, however, with FloraMap not being public-domain software, I believe that the better option in this case is to go the way of the Diva-GIS software. The alternative of using FloraMap should be mentioned, however, for those organizations that may wish to outlay the cost of purchasing FloraMap.

The third FloraMap method uses Cluster analysis of the climate profile to identify possible disjunctions, caused by either outliers or the possibility of their being two distinct populations or species. Testing of this method has shown it to be a quite powerful methodology. It also showed great promise for analysis of species patterns, however that is not what I am looking at here. Although this methodology seems exclusive to FloraMap, it would not seem a difficult programming problem to include within other software (including Diva-GIS). This may need to be discussed with the developers of both FloraMap and Diva-GIS.

In summary, FloraMap is not public-domain software, and is unlikely to be available free for inclusion on the CD-Rom. It may be worth discussing the possibility of inclusion, however, with the developer. Two of the three routines can be replaced with routines in other software. The third would ideally be included, but requires work to do so.

13.1.2 Diva-GIS

The Diva-GIS is public-domain software. It has some very good data cleaning and validation algorithms. The program, as is, however, has some aspects that I believe are unlikely to be extensively used by herbarium or museum workers, and may be confusing to them. It would not be difficult, through collaboration with the developer of the program, to modify some of these aspects. Discussions with the developer have indicated that this may be possible – see attached email below.

Both the Check Coordinate and Assign Coordinate algorithms would prove valuable tools to any museum or herbarium. They are simple to use. The first of these, as mentioned above, is a more powerful and useful tool (for the purpose we are examining) than the corresponding routine in FloraMap. The Assign Coordinates algorithm has some overlap with BioGeoMancer, GeoLoc-CRIA and EGaz, and could be enhanced by the inclusion of accuracy values in the determinations.

The BIOCLIM Cumulative Frequency method is a method that has been around for a long time, and has proven its usefulness over that period. It is still included in the latest releases of ANUCLIM as mentioned above. This method has proven to identify many of the similar outliers to the PCA method in FloraMap, and in many ways is the more intuitive of the methods.

Discussions with Robert Hijmans, the developer of Diva-GIS on the possible inclusion of the software on a Data-cleaning Toolkit has elicited a positive response. His response included:

> "*It would be great if you would put DIVA on such a CD. I could make a special install (with some tutorial stuff and climate data included).*"
> (R.J.Hijmans – personal email, 14 Jan 2004).

13.1.3 ANUCLIM

Unless Institutions wish to purchase ANUCLIM for other purposes (for which it is a very valuable tool), due to its cost, I would not recommend consideration of including it, or any part on the CD-Rom. The main data-cleaning algorithms are identical (but use more climate layers) than the earlier version of the software included in the Diva-GIS.

13.1.4 EGaz

EGaz is a valuable tool for finding latitude and longitude values for inclusion in a specimen database. At present, the output links directly to the BioLink specimen management software (see discussion under Chapman 2003c), but intermediate routines, I believe, could easily be programmed to link to other specimen management tools.

EGaz is a user-friendly, intuitive tool, and one I believe should be included in the Toolkit. It does have a bug at the moment that causes it not to work outside Australia, however I am confident that this will be fixed before long.

An alternative is the on-line tool ‘Geo-Loc-CRIA’ mentioned above (8.1.3) being developed at CRIA. This may be a better alternative for use, at least in Brazil. There may be good reason to include both, if the problems with EGaz are fixed.

13.2 Additional Algorithms

If it is decided to modify a program such as Diva-GIS (or alternatively write a new program using many of the same methodologies), there are a number of additional algorithms that I believe should be included.

13.2.1 Altitude

A simple SQL script could be added to the CD to assist with the adding of elevation records into the database. This would use an SQL script to extract the latitude and longitude from the specimen database, and check that against a databased Digital Elevation Model (DEM) in the form of x, y, and z values. Such a DEM exists for South America at a grid scale of 1km (NGDC 2000). Once an elevation is extracted from the DEM, it can be inserted back into the specimen database. The script would need to be modified, and perhaps could be prepared for each of the main specimen management programs such as Biota (Colwell 2002), BRAHMS (University of Oxford 2003), Specify (University of Kansas 2003a), BioLink (Shattuck and Fitzsimmons 2000), etc. Alternatively, a simple on-line tool could be developed to do this through the submission and return of an updated file. Perhaps it could be done in conjunction with CRIA’s Outlier Detection tool spOutlier-CRIA (CRIA 2004b).

13.2.2 Other Algorithms

A number of additional algorithms may be included, for example:

- Algorithm from Centre for Plant Biodiversity Research for tracking already databased geocodes
- Collector-tracking algorithm
- others

13.2 Guidelines

The CD should include a number of simple Guidelines. These would include:

- Introduction to Data Quality – probably a Powerpoint presentation,
- HISPID (Conn 1996) – permission already obtained,
- Geocoding Guidelines (Wieczorek 2001a)
- Data Validation Guidelines (to be written)
- Guidelines to Nomenclature (Chapman 2003a)
- Searchable Help document (to be written – similar to a Microsoft Help)

13.3. Pick Lists

I would envision the CD including a number of Standard Pick Lists that could be incorporated into users databases. In the case of the Species2000 Checklist, this may just be as an additional companion CD.

- Plant Collectors of Brazil (Koch 2003).
- Checklist of Neotropical Plant Species
- Species2000 Catalogue of Life
- Other relevant checklists
- Links to ECAT (GBIF 2003) once available.

13.4. Gazetteers

A number of Gazetteers could be included, including all the South American Countries, and perhaps even Central America. These are freely available.

13.5. Links to Web Tools

- BioGeoMancer
- Lifemapper
- Desktop GARP
- Additional Gazetteers
- Software sources
    - BioLink
    - BIOTA
    - Brahms
    - Platypus
    - Specify
- Catalogue of Life
- ECAT

13.6. Other links

Links to key sites – some of which are listed in the Guidelines to Nomenclature (Chapman 2003a).

14. Conclusions
I believe is required by biologists and collection managers in Brazil are::

1. guidelines and tools for users managing databases of plant and animal collections
2. more efficient and accurate methods of geocoding specimen records
3. methodologies and tools for users to check and validate records already databased

I believe the best way to achieve this is through the provision of a simple data quality and validation toolkit on CD-Rom to include both stand-alone software and links to web-based solutions.

The CD should be available free, or at a very reasonable cost (cost of production only). Initially I would envisage an English version (because much of the documentation and many of the tools are already in English), and Portuguese version, with possibly a Spanish version to follow.

The CD would include Stand-alone software tools, guidelines, pick-lists of names (collectors and species), computer algorithms, and links to on-line resources.

Some of the software may need negotiated licences and/or modifying. I suggest that this be done collaboratively with the developers of the software wherever possible.

I recommend that funding be sought for the production of such a CD, and that CRIA consider beginning the project as soon as possible. I would continue to be available to assist with the writing of documentation and methodologies.

15. Acknowledgements
Many people have contributed to the writing of this document – through providing software and computer scripts for testing, data – both good and bad, a ready audience for lively discussion, or just answering my many questions. The format of the paper and the elaboration of ideas therein are mine and any errors or misinterpretations are solely due to my ignorance or misunderstanding. Many of the ideas, however, could not have been elaborated upon without the many discussions on data quality and validation by many, many people around the world over the past 15-30 years.

In particular, I would like to thank the staff at my previous places of employment, especially, Karl Bossard, Simon Bennett, John Busby, Jim Croft, Gaston Rozenbilds, Kate Sanford-Readhead, Tony Rosling, Jeff Tranter and Judy West for helpful discussions and ideas over a period of nearly 30 years. Jim Beach, Reed Beaman, Stan Blum, Jim Croft, Towsend Peterson, Ricardo Scachetti-Pereira, David Stockwell and John Wieczorek have provided lively debate on all issues associated with data quality and validation over a period of 15 years along with many ideas and solutions to what is a difficult problem. Robert Hijmans, Peter Jones and Steve Shattuck have made available to me their valuable software for testing. Robert Hijmans has also supplied valuable data and assistance. The staff of CRIA have been particularly helpful in not

only providing a forum for me to express my ideas, a place to work in peace and tranquillity, but many ideas and much support. The museums and herbaria of the world have been an endless source of data (both good and bad) upon which to experiment, and in particular those of Australia where much of my early work was done. Ingrid Koch of CRIA was also a source of data with unique problems with which I was able to test and compare the various software tools. Last, but not least, I would like to thank Vanderlei Canhos, Dora Lange-Canhos and Carlos Joly for providing me with the opportunities of coming to Brazil on a number of occasions and to be able to spend some time working here. This would not have been possible without their support and that of the Biota Program of FAPESP which has provided the funding for this and other project through grant no. 02/10039-7.

16. References:

Armstrong, J.A. (1992). The funding base for Australian biological collections. *Australian Biologist* **5(1):** 80-88.

Austin, M.P. and Meyers, J.A., 1995, *Real data case study, Sub-project 4, Modelling of landscape patterns and processes using biological data*. Canberra: Division of Wildlife and Ecology, CSIRO.

Australian Government (1999). *Environment Protection and Biodiversity Conservation Act 1999. No. 91, An Act relating to the protection of the environment and the conservation of biodiversity, and for related purposes.* Canberra: AGPS.

Beaman, R.S. (2002). Automated georeferencing web services for natural history collections **in** Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002 http://www.cria.org.br/eventos/tdbi/flora/reed [Accessed 26 Jan 2004].

Beaman, R.S. and others (2003). BioGeoMancer Lawrence, Kansas: University of Kansas – prototype http://www.biogeomancer.org/ [Accessed 26 Jan 2004].

Beer, T. & Ziolkowski, F. (1995). *Environmental risk assessment: an Australian perspective.* Supervising Scientist Report 102. Canberra: Commonwealth of Australia. http://www.deh.gov.au/ssd/publications/ssr/102.html [Accessed 27 Jan 2004].

Burrough, P.A and McDonnell, R.A (1998). *Principals of Geographical Information Systems*. Oxford, UK: Oxford University Press

Busby, J.R. (ed.) (1973). *Australian Biotaxonomic/Biogeographic Information System (ABIS).* Canberra: Australian Biological Resources Study.

Busby, J.R. (1991). BIOCLIM – a bioclimatic analysis and prediction system. Pp. 64-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO

Carpenter, G., Gillison, A.N. and Winter, J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2:** 667-680.

Chapman, A.D. (1988). Karl Domin in Australia **in** *Botanical History Symposium. Development of Systematic Botany in Australasia. Ormond College, University of Melbourne. May 25-27, 1988*. Melbourne: Australian Systematic Botany Society, Inc.

Chapman, A.D. (1991). Land cover project. *Erinyes* **9:** 4-6.

Chapman, A.D. (1992). Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth land Information Forum.

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

Chapman, A.D. (2002). Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia. pp. 31-40 **in** Hunter, G. and Lowell, K. (eds). *Accuracy 2002. 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Melbourne: RMIT.

Chapman, A.D. (2003a). Guidelines on Biological Nomenclature – Brazil edition. Report to Centro de Referência em Informação Ambiental. Campinas: CRIA.

Chapman, A.D. (2003b). *Lifemapper – comments and ideas.* Internal report No.7 to CRIA and University of Kansas, Draft - July 2003.

Chapman, A.D. (2003c). *BioLink –an evaluation.* Internal report No.8 to CRIA, Draft in Preparation - July 2003.

Chapman, A.D. (2004). *Environmental Data Quality – a. Discussion Paper*. Report No. 5 to CRIA. Campinas: CRIA.

Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. (2001). Environment Protection and Biodiversity Conservation Act, 1999 – Information System. Proceedings of the 17th Annual Meeting of the Taxonomic Databases Working Group, Sydney, Australia 9-11 November 2001. Powerpoint: http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 26 Jan 2004].

Chapman, A.D. *et al*. (2002). *Guidelines on Biological Nomenclature*. Canberra: Environment Australia. http://www.deh.gov.au/erin/documentation/pubs/nomenclature.doc [Accessed 26 Jan 2004]

Chapman, A.D. *et al.* (in prep). *The use of expert validation of modelled species distributions in Australia's EPBC Act, decision support system.*

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

Chapman, A.D. and Milne, D.J. (1998). *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia

CHAH (2002). AVH - Australian's Virtual Herbarium. Australia: Council of Heads of Australian Herbaria. http://www.chah.gov.au/avh/avh.html [Accessed 26 Jan 2004].

Christidis, L. & Boles, W.E. (1994*). Taxonomy and Species of Birds of Australia and its Territories*. Royal Australasian Ornithologists Union, Melbourne. 112 pp.

Colwell, R.K. (2002). *Biota: The Biodiversity Database Manager*. Connecticut, USA: University of Connecticut http://viceroy.eeb.uconn.edu/Biota [Accessed 26 Jan 2004]

Conn, B.J. (ed.) (1996). *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. http://www.bgbm.org/TDWG/acc/hispid30draft.doc [Accessed 26 Jan 2004].

Conn, B.J. (ed.) (2000). *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data.* Version 4 – Internet only version. Sydney: Royal Botanic Gardens. http://plantnet.rbgsyd.nsw.gov.au/Hispid4/ [Accessed 26 Jan. 2004].

Croft, J.R. (ed.) (1989). HISPID – *Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.

CRIA (2002). SpeciesLink. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/ [Accessed 26 Jan. 2004].

CRIA (2004a). *GeoLoc-CRIA*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [Accessed 26 Jan 2004].

CRIA (2004b). *spOutlier.* Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [Accessed 26 Jan 2004]

Department of Natural Resources (DNR), Department of Environment, and Environment Australia (1997) *Systematic Vertebrate Fauna Survey Project. Stage II – Assessment of Habitat Quality for Priority Species in Southeast Queensland Bioregion*. Brisbane: Queensland Government. http://www.rfa.gov.au/rfa/qld/se/raa/fauna/eh1.1.2b/eh2b.pdf [Accessed 26 Jan. 2004].

ESRI (2003). *ArcSDE: The GIS Gateway to Relational Databases*. http://www.esri.com/software/arcgis/arcinfo/arcsde/overview.html [Accessed 26 Jan. 2004].

FAO (2000). *Ecocrop*. Rome: Food and Agricultural Organization of the United Nations. http://ecocrop.fao.org/ [Accessed 26 Jan. 2004].

Francki, R.I.B., Fauquet, C,M., Knudson, D.L., Brown. F. (1990). Classification and Nomenclature of Viruses. *Archives of Virology Supplement* **2:** 1-445. [see http://www.biosis.org/zrdocs/codes/icvcn.htm Accessed 26 Jan. 2004].

GBIF (2003). GBIF Work Program 2004. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/GBIF_org/wp/wp2004/GB7_20WP2004-v1.0-approved.pdf [Accessed 23 Jan 2004].

Hijmans, R.J., Schreuder, J. De la Cruz, J. and Guarino, L. (1999). Using GIS to check coordinates of genebank accessions. *Genetic Resources and Crop Evolution* **46:** 291-296.

Hijmans, R.J., Guarino, L., Bussink, C., Barrentes, I. and Rojas, E. (2004) *DIVA-GIS Version 4. A geographic information system for the analysis of biodiversity data*. http://www.diva-gis.org [Accessed 23 Jan 2004].

Houlder, D. Hutchinson, M.J., Nix, H.A. and McMahaon, J. (2000). ANUCLIM 5.1 Users Guide. Canberra: Cres, ANU. http://cres.anu.edu.au/outputs/anuclim.html [Accessed 26 Jan. 2004].

International Code of Botanical Nomenclature (2000). International Code of Botanical Nomenclature (St Louis Code). *Regnum Vegetabile* **138**. Königstein: Koeltz Scientific http://www.bgbm.fu-berlin.de/iapt/nomenclature/code/SaintLouis/0001ICSLContents.htm [Accessed 26 Jan. 2004].

International Code of Zoological Nomenclature (2000). *International code of zoological nomenclature adopted by the International Union of Biological Resources International Commission on Zoological Nomenclature*. 4th edn. London : International Trust for Zoological Nomenclature. [see http://www.iczn.org/code.htm Accessed 26 Jan 2004].

Jennings, M. & Scott, J.M. (1997). Gap Analysis Program. Official Description of the National Gap Analysis Program. Moscow, ID: USGS. http://www.gap.uidaho.edu/About/Overview/GapDescription/default.htm [Accessed 26 Jan. 2004].

Jones P.G. and Gladkov, A. (2001). *Floramap Version 1.01*. Cali, Colombia: CIAT. http://www.floramap-ciat.org/ing/floramap101.htm [Accessed 26 Jan. 2004].

Koch, I. (2003). Coletores de plantas brasileiras. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 26 Jan. 2004].

Lindemeyer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. and Tanton, M.T. (1991). The Conservation of Leadbeater’s Possum, *Gymnobelidus leadbeateri* (McCoy): A Case Study of the Use of Bioclimatic Modelling. *J. Biogeog.* **18:** 371-383.

Marino, A., Paverin, F. de Souza, S. and Chapman, A.D. (in prep). *Simple on-line tools for geocoding and validating biological data*. To be submitted to CODATA Journal.

Nix, H.A. (1986). A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). Atlas of Australian elapid snakes. *Australian Flora and Fauna Series* No. **7:** 4-15. Canberra: Australian Government Publishing Service.

NGDC (2000). *Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Data* Version 1.0. http://www.ngdc.noaa.gov/seg/fliers/globedem.shtml [Accessed 26 Jan. 2004].

Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. (1998). The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140:** 288-294.

Peterson, A.T., Navarro-Siguenza, A.G. and Scachetti-Pereira, R. (in press). Detecting Errors in Biodiversity: Collector’s Itineraries Flag Mislabeled Specimens. Submitted to *Bulletin of the British Ornithologists' Club*.

Scachetti-Pereira, R. (2002). *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research.

Shattuck, S.O. (1997). eGaz, The Electronic Gazetteer. *ANIC News* **11:** 9 http://www.ento.csiro.au/research/natres/anicnews/anicnews11_09.html [Accessed 26 Jqan. 2004].

Shattuck, S.O. and Fitzsimmons, N. (2000). *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. http://www.biolink.csiro.au/ [Accessed 26 Jan. 2004].

Sneath, P.H.A. (ed.) (1992). *International Code of Nomenclature of Bacteria*, 1980 Revision. Washington: International Committee on Systematic Bacteriology (ICSB). [see http://www.biosis.org/zrdocs/codes/icnb.htm Accessed 26 Jan. 2004].

Specht, D. (1997). *RapidMap. Geocoding locality descriptions associated with herbarium collections*. http://users.ca.astound.net/specht/rm/ [Accessed 26 Jan. 2004].

Species 2000 (2002). *Catalogue of Life- Indexing the world's known species. Year 2002 Annual Checklist*. http://www.sp2000.org/AnnualChecklist.html [Accessed 26 Jan. 2004].

Suter, G.W., II (1992). *Ecological Risk Assessment*. Chelsea, Michigan: Lewis Publishers Inc.

Suter, G.W., II, B.E. Sample, D.S. Jones, T.L. Ashwood, and J.M. Loar. 1995. *Approach and strategy for performing ecological risk assessments for the U.S. Department of Energy's Oak Ridge Reservation*: 1995 revision. Oak Ridge TN: Oak Ridge National Laboratory,

TDWG (2003). International Union for Biological Standards – Taxonomic Database Working Group (TDWG). http://www.tdwg.org/index.html [Accessed 26 Jan. 2004].

Trehane, P., Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Spongberg, S.A. & Vrugtman, F. (1995). *International Code of Nomenclature for Cultivated Plants.* Winbourne, UK: Quarterjack Publishing. [see http://www.ishs.org/sci/icracpco.htm Accessed 26 Jan. 2004].

University of Colorado Regents (2003). *MaPSTeD. Geocoding*. Denver: University of Colorado MaPSTeDI project. http://mapstedi.colorado.edu/geocoding.html [Accessed 27 Jan. 2004]

University of Kansas (2003a). *Specify.* Biological Collections Management. Lawrence, Kansas: University of Kansas http://usobi.org/specify/ [Accessed 26 Jan. 2004].

University of Kansas (2003b). *LifeMapper.* Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/ [Accessed 26 Jan. 2004].

University of Kansas (2003c). *BioGeoMancer*. Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://biogeomancer.org/ [Accessed 26 Jan. 2004].

University of Oxford (2003). *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford http://storage.plants.ox.ac.uk/brahms/ [Accessed 26 Jan. 2004].

Wieczorek, J. (2001a). *MaNIS: Georeferencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://dlp.cs.Berkeley.edu/manis/GeorefGuide.html [Accessed 26 Jan. 2004].

Wieczorek, J. (2001a). *MaNIS: Georeferencing Calculator*. Berkeley: University of California, Berkeley - MaNIS http://elib.cs.berkeley.edu/manis/gc.html [Accessed 26 Jan. 2004].

Wieczorek, J. (2002). *Summary of the maNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charels, LA, June 16, 2002*. Berkeley: University of California, Berkeley - MaNIS. http://elib.cs.berkeley.edu/manis/ASM2002.html [Accessed 26 Jan. 2004].

Wieczorek, J. and Beaman, R.S. (2002). Georeferencing: Collaboration and Automation in Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002 http://www.cria.org.br/eventos/tdbi/bis/georeferencing [Accessed 26 Jan. 2004].

Wieczorek, J., Guo, Q. and Hijmans, R.J. (in press). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty.

Williams, P.H., Marguiles, C.R. and Hilbert, D.W. (2002). Data requirements and data. sources for biodiversity priority area selection. *J. Biosc.* **27(4):** 327-338.