Report No. 5 DATA

Environmental Data Quality
- A Discussion paper

Arthur D. Chapman
20 Jan. 2004

Aim:
To discuss issues of data quality and error, particularly as it applies to CRIA, and the FAPESP/Biota Program.

Background:

This paper attempts to discuss general data quality and error issues with respect to the data being used in the FAPESP/Biota Program. A second paper, Report No. 6 – Environmental Data Quality – Data Cleaning Tools – deals with more specific data cleaning methodologies and tools.

Environmental data quality and error in data are often neglected issues with databases, modelling systems, GIS, decision support systems, etc. Too often, data is used uncritically without consideration of the error contained within, and this can lead to erroneous results, misleading information, and even unwise environmental decisions.

This paper particularly concentrates on the issue of data quality with respect to species and environmental data from a spatial viewpoint. It does not attempt to examine other, more general issues of data quality such as species names, etc. other than as they apply to spatial issues, however these are discussed in the following report on environmental data cleaning tools.

The custodians and owners of data (individual collection agencies such as museums and herbaria) are largely responsible for the quality of their data. None-the-less, those supplying the data and those using the data, also have responsibilities. The data suppliers – in the case of museum and herbarium data, the collectors - need to ensure that the collections they make are adequately documented, that the label information is clear and unambiguous, that the location information is accurately recorded and documented and that their collection methodologies are fully documented. Users of the data need to feed back to the custodians information on any errors they may come across and additional information they may need recorded in the future, etc. Users and collectors have important roles to play in assisting custodians in maintaining the quality of the data in the collections, and both have a vested interest in the data being of the highest possible quality.

CRIA, aside from being a user of data through projects such as *species*Link, various species modelling projects, etc., may also have a role to play in assisting in the development of methodologies and tools for the identification of errors in the data and in improving the overall quality of the data infrastructure through development of "data cleaning" tools. As improving techniques for managing data and data quality, and distributing tools and guidelines, an improved overall data consistency and quality will be achieved. This issue is more fully discussed in Report No. 6.

Data Error

Two basic types of data exist – primary data, such as individual point-referenced plant and animal collection data and meteorological data; and secondary or derived data such as climate surfaces and species models. Primary data, which is collected and referenced to individual points, largely eliminates problems of scale and those posed by categorisation (Chapman and Busby 1994). Categorised information commonly used to produce natural resource maps including soil types, vegetation categories, tree height classes and species (i.e. a collection of individual specimens) can present problems in a number of ways, but are essential for information presentation and analysis. One of the problems with pre-classified data occurs when the concepts underpinning the classification change, and thus the underlying data may become useless. Data stored as primary attributes – individual specimens, actual tree heights, etc. – can be used to produce classified entities for display and communication while remaining available for use in alternative classifications, or as individual records in their own right.

Data stored in its primary format allows a far greater scope for validation, detection of error and correction. Once the data have been converted to secondary information, errors may well be detected (and often the use of secondary information provides valuable error detection methods), but unless the primary data is retained, those errors cannot be corrected. It is important, therefore, that feedback mechanisms be developed whereby errors detected in the data – at any stage in the process – can be reported back to the primary data custodians.

a. Species Data

Plant and animal specimen data held in museums and herbaria provide a vast information resource, providing not only present day information on the locations of these entities, but also historic information going back several hundred years (Chapman and Busby 1994). It is estimated that there are approximately 2.5-3 billion collections worldwide in museums, herbaria and other collection institutions (Duckworth *et al.* 1993, OECD 1999). The number housed in Brazilian collections is unknown, as is the number of Brazilian collections housed elsewhere, but would not be insignificant. Projects to digitise this information are underway in many institutions, and most others are at either the discussion or planning stage. The process of digitising collections is a tedious and time-consuming exercise. It can take between 5 and 30 minutes to database one herbarium plant collection with about half this time dedicated to the addition of geocoding (latitude and longitude) information (Greg Whitbread *pers. com.* 2003). Some museum collections can take considerably longer (Armstrong 1992).

There are many uses for the data in museums and herbaria. Traditionally, these collections were only made with one main purpose in mind – that of taxonomic study. The introduction of computer processing and computer databases, however, opened up this vast data store to many other uses (Chapman 1999). These uses include biogeographic studies (Longmore 1986, Peterson *et al.* 1998), conservation planning

(Faith *et al*. 2001), reserve selection (Margules and Pressy 2000), development of environmental regionalizations (Thackway and Cresswell 1995), climate change studies (Chapman and Milne 1998, Pouliquen and Newman 1999, Peterson *et al.* 2002a), agriculture and forestry production (Booth 1996, Nicholls 1997, Cunningham *et al.* 2001), species translocation studies (Mackey 1996, Soberón *et al.* 2000, Peterson and Veiglas 2001), etc., etc. Many of these studies have used environmental modelling using software such as BIOCLIM (Nix 1986, Busby 1991), GARP (Stockwell and Peters 1999, Pereira 2002) or GLM (Austin 2002). Most of these species distribution models rely on specimen or observation records, generally of a presence-only nature (usually including records from herbaria or museums as well as observation data) or occasionally presence-absence data from systematic surveys. It is this use in biogeographic studies, including for species modelling, that has caused a focus on data quality issues that had previously been largely ignored.

Because of the historical nature of many museum and herbarium collections, many records carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998). Where geocoding is given it is often not very accurate (Chapman 1999) and has generally been added at a later date by those other than the collector (Chapman 1992). Many of these data thus have drawbacks when it comes to use for species' distributional studies. New tools are now being developed around the world to assist institutions in the process of adding geocode information to databased collections. Such tools include Egaz (Shattuck 1997), BioGeomancer (Beaman *et al*. 2003) and Localidade (CRIA 2004a). At least two of these (BioGeoMancer and Localidade) are looking at incorporating validation algorithms. The need for further and more varied validation tools cannot, however, be denied.

Additionally, much of the data has been collected opportunistically rather than systematically (Chapman 1999, Williams *et al.* 2002) and this can result in large spatial biases – for example, collections that are highly correlated with road or river networks (Margules and Redhead 1995, Chapman 1999, Peterson *et al.* 2002b). Museum and herbarium data generally only supply information on the presence of the entity at a particular time and say nothing about absences in any other place or time (Margules and Austin 1994, Peterson *et al.* 1998).

These data cannot be neglected, however, as they constitute the largest database of biological information we are ever likely to have. The cost of replacing these data with new surveys would be prohibitive. It is not unusual for a single survey to exceed $1 million to conduct (Burbidge 1991). They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanization, climate change, or been modified in some other way (Chapman 1999).

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 major causes of error in modelling and biogeographic studies, for example. Assessment of the accuracy of input data is essential otherwise the results in any of these studies 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 it cannot 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. This last issue – spatial bias - is very evident in herbarium and museum data (e.g. collections along roads, etc.) but is more an issue for future collections, and future survey design rather than being related to the quality of individual plant or animal specimen records (Margules and Redhead 1995, Williams *et al*. 2002). Indeed – collection bias is more related to the totality of collections of an individual species (i.e. the secondary or classified data), than it is to any one individual collection. In order to improve the overall spatial and taxonomic coverage of biological collections within an area, and hence reduce spatial bias, existing historical collection data can be used in determining the most ecologically valuable locations for future surveys (e.g. see Cofinas *et al.* 1995, Neldner *et al.* 1995).

Methods have been developed to identify possible 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 both 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 (see Report 6 on Data Cleaning Tools – Appendix G). 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, and EGaz (Shattuck 1997), BioGeoMancer (Beaman *et al.* 2003) and Localidade (CRIA 2004a), that assist in the georeferencing of collections. 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 2001), the MaPSTeDI Guide to Georeferencing (University of Colorado 2003), and HISPID (Conn 1996, 2000).

As stated in the MaPSTeDI Guidelines "While geocoding is not an exact science and no collection can be geocoded 100% correctly, quality checking can drastically improve the percentage of the collection that is correctly geocoded. Every project should take it into account when planning their geocoding operation" (University of Colorado 2003).

Ecological data

Ecological data is another form of data being used in biodiversity studies. In virtually all cases, this data is obtained from outside sources, and users have to rely on the metadata associated with those data layers in order to determine their accuracy. These data include polygon-based vegetation and remotely sensed raster data, and may include both primary and secondary data. Again the quality of this data is likely to be variable, especially with vegetation data, and the level of documentation is also likely to be variable. There are a number of issues associated with accuracy and error with this type of data, and the type of error may vary depending on whether the data is raster data or polygon data.

With raster data, for example remotely sensed satellite data, one key aspect of error is the accuracy with which a pixel is able to be located against a position on the ground. It is generally accepted that the margin of error, depending upon the method of geometric registration, is about +/- one pixel. Thus AVHRR (Advanced Very High Radiation Resolution) which is commonly used in vegetation studies and which has a pixel size of approximately one kilometre, has an accuracy of approximately +/- one kilometre (Mao and Huang 1999). In many cases, however, especially in areas of the earth with few identifiable registration points (the deserts of Australia, the Amazon Basin, many marine areas), pixel accuracy cannot be relied on at better than +/- two pixels (S.Cridland *pers. com.*).

Polygon data can have quite varying levels of accuracy. The TOPO-250K data for Australia (1:250 000 topographic data) is a well-researched data set and its accuracy is described as “not more that 10% of well-defined points are in error by more than 160 metres; and in the worst case, a well defined point is out of position by 300 metres” (Geoscience Australia 2003a). This accuracy can be quite important, for example if one is trying to determine if a species (with accuracy of say +/- 1 km) occurs within a national park (with an accuracy of +/- 160 meters). With paper topographic maps, drawing constraints may also restrict the accuracy of where certain lines are placed. A one (1) mm wide line depicting a road on a 1:250 000 map represents 250 meters on the ground. If there is a railway running beside the road, there usually needs to be a separation of 1-2 mm (250-500 meters) and then the line for the railway – another 1 mm or 250 meters – makes a total of between 750 meters and 1 km as a minimum representation. If you are using these features to determine the location of your specimen, for example, then you cannot be certain at better than about 1 km accuracy. Obtaining accurate coastlines, for example, can also be a nightmare. Does the map use the high water mark or mean sea-level. Has neap, spring or king tides been taken into account (Bannerman 1999). When it crosses the opening of a river – does it take a direct line across, or does it follow the river upstream for a way, and if so, how far. Are rivers depicted by two lines – one for each bank, or just by a centre line? How narrow is the river, before it changes from one to the other? How much of the river curve has been removed? Is a town shown as just a point, and if so what part of the town does that point represent? If the town is represented by a boundary is it the municipal boundary or the boundary of outer development, etc.? (Wieczorek 2001). How is terrain represented – is it just by contours and thus excludes the highest and lowest points, or does it show spot heights, and how accurate are they? Many features such as coastlines and rivers change over time. In many tropical areas of the world, seasonal billabongs are formed - how are they represented? (Bannerman 1999). Again, the scale of the map can make a significant difference.

FIGURE of Coastlines at different scales

Fig 1.

The depiction of phenomena that don't have discreet boundaries (vegetation, soils, geology, etc.) can be another problem (Burrough and McDonnell 1998). Vegetation types seldom have discreet boundaries and gradually meld into the neighbouring vegetation type. Where does one draw the line, and how has this been done on the map one may be using? One of the troubles in nature is that lines shown on a map are seldom obvious on the ground. In some cases, a vegetation description may be a mosaic of vegetation types (Sattler and Williams 1999). The classification of data instead of the use the primary data as mentioned above, may be necessary, but can cause problems in interpretation and thus lead to error (Chapman and Busby 1994). Trees and shrubs may occur in two different datasets. In one, a tree may be defined as anything above 8 meters and a shrub below 8 meters, while in the other, the definition may have had the cut off at 6 meters. If you are attempting to use this information, what do you do with those areas where the plants are between 6 and 8 meters? Here you have an error of around 25% in the figure.

Another issue with polygon data is that of Datums. More often than not, the Datum is not given, especially on older paper maps. The use of a different Datum can cause quite a difference in the position on the ground. The difference between the Australian Geodetic Datum (AGD66) and the World Geodetic System (WGS84) in Australia, for example, means a shift of around 170 meters, depending on where are you are (Geoscience Australia 2003b). This can get even worse off-shore, where often the old AGD66 datum was wrongly used on islands hundreds or thousands of kilometres off the coast, meaning that the island was misplaced by hundreds of kilometres on many maps. In the USA, the recent change in Datum caused a variation from between 0 and 470 meters (Wieczorek 2001).

Environmental data

As mentioned above, a lot of biogeographically related studies are using environmental modelling tools. So just as important as the species data, are the environmental layers used to model the species against. The theory of most environmental models is that species have certain habitat preferences that have an environmental basis (Nix 1986). Environmental data can thus be used as surrogates for biodiversity in modelling of distribution patterns of species or populations from the point records obtained from collection data (Faith and Walker 1996, Ferrier and Watson 1997, Williams *et al.* 2002). Many models use climatological information such as temperature, rainfall, radiation, evaporation, soil moisture etc. as the basis on which to broadly define the habitat or ecological niche. Other models use vegetation characteristics such as vegetation classes, detailed habitat information, correlated species. Environmental data are generally more widely available, usually more accessible, and generally exist in more consistent form than most biological data (Williams *et al*. 2002).

Terrestrial environmental data fall into three basic categories: terrain, climate and substrate.

Terrain refers to surface morphology and includes parameters such as elevation, slope, relief and aspect. Digital Elevation Models (DEMs) are representations of surface morphology and can be developed at varying scales. The development of DEMs allows for consistent and repeatable interpolation across whole regions and constitutes the necessary first step in generating environmental surfaces (Hutchinson 1991). The construction of a DEM is time-consuming and technically demanding (Hutchinson 1991). Errors in this type of data can arise in any number of ways, and the method used to create the DEM can be quite important in determining both the type and dimension of the likely error.

Climate data are generally available from national meteorological agencies, but may have to be digitised and spatially interpreted for use in species' modelling programs. The spatial interpolation of climate data can be carried out with the aid of DEMs by fitting surfaces as smooth tri-variate functions of latitude, longitude and elevation (Hutchinson 1995). These are usually done at the scale of the DEM, and when done properly, involve a lot of data cleaning and quality control. The development of appropriately scaled climate surfaces is essential for modelling of species' distributions if the models are to have any environmental meaning at the scales usually required for management or decision-making.

Substrate data, both physical and chemical, can be one of the most difficult to obtain, and the quality from one layer to another can be extremely variable. Some mapping has been done in most regions of the world, but this is at varying scales and completeness. Substrate layers include soils, lithology, surficial geology, hydrology and landform.

One of the important considerations in choosing appropriate environmental layers is that of scale. Too fine a scale will lead to errors due to mismatching with the biological data being modelled against it. Too coarse a scale may not adequately delineate the appropriate environmental niches. Too often modellers give little consideration to scale in their selection of environmental layers. This may, of course be due to availability as they may only be able to use those layers that are available at the time. If this is the case, then it is important that this be noted, and the assumptions arising be elaborated. The use of poor environmental layers often leads to models that prove of very little practical value in understanding the environmental factors that drive the species' niche preferences.

The preparation of environmental layers is often one of the most time-consuming, and computer-intensive areas of modelling. Fortunately, it only has to be done occasionally and once the surfaces are prepared can be used for any number of models. Climate surfaces, for example, have been prepared for much of the world's land surface and are available for use by researchers free, or at nominal charge. Until recently, these have been at varying scale with surfaces at a suitable scale for modelling not available for South or Central America. In early 2004, this situation has changed (see separate report (Chapman 2004), as globally consistent surfaces have now been prepared at 30-second (about 1 km) resolution for the whole world. A separate case for the development of such surfaces (at 3-arc minutes) was made early in 2003 (Chapman 2003a). These new layers have been prepared using ANUSPLIN (Hutchinson 2001), which includes extensive error checking and validation routines.

Errors in environmental data must not be ignored. The mis-use of environmental layers in environmental modelling, the mixing of scales (see fig. 2), the use of inappropriate scales (again see fig. 2), and the mis-registration of layers (fig. 2) can also cause major errors in models relying on that data to create the rules or profiles.

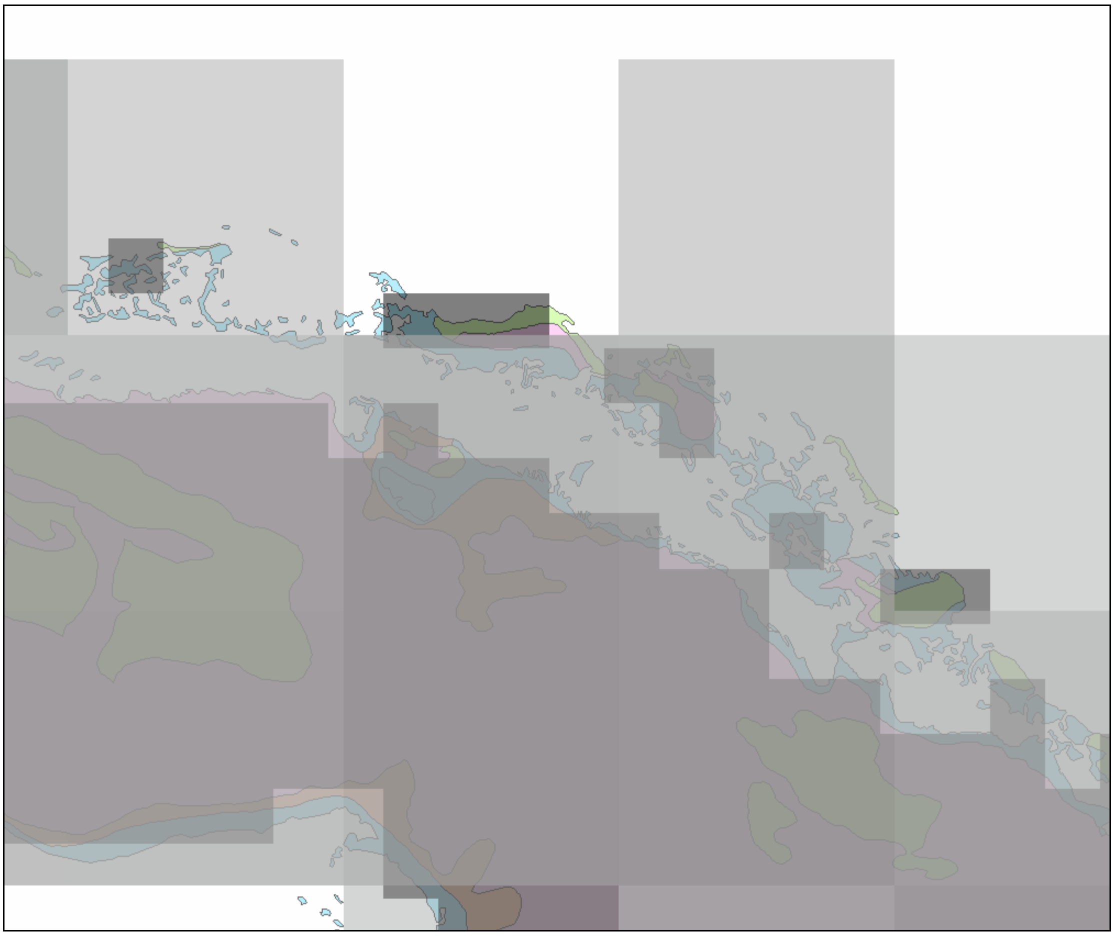

Fig. 2. Example of environmental layers being used at different scales. This figure also illustrates the problem with using mis-registered layers. Layers are from Central America and include slope, precipitation and radiation.

All too often, environmental layers are selected without critical analysis. This is partly due to the difficulty of such analysis, and thus modelling often tends to be not only testing the species against the environment layers, but the environment layers against themselves. Often great care goes into selecting species records and checking their accuracy, but the same cannot be said of the environmental layers, although the latter are likely to be used over and over again in model after model.

In a similar manner as to species data, the mis-location of weather stations, or errors in readings at those stations, can cause major errors in resultant climate surfaces. Geocoding of weather stations can be just as tedious, and error-prone as geocoding of species data – but at least there are fewer weather stations than species collections! When prepared critically, climate layers will have an extremely low level of error. Errors will generally be of two major types – positional error and attribute error. Positional accuracy will depend on the accuracy of the underlying DEM. The DEM prepared for South Africa, for example, at 10 arc-minutes has a standard error of between 20 and 150 meters (Hutchinson 2003a). The methods used in ANUDEM, developed by Hutchinson and others has led to the development of DEMs with a much lower error than previous methods (Hutchinson 1996). The attribute error for the climate data, on the other hand, is different for the temperature and rainfall. Because the ANUSPLIN method (Hutchinson 2001) employs dependence on elevation, it is significantly more accurate than methods that use bi-variate functions of longitude and latitude (Margules and Redhead 1995). The standard errors of the temperature are about 0.5 degrees centigrade and of the rainfall grids range between about 5 and 15 per cent, depending on data density and the spatial variability of the actual monthly mean rainfall (Margules and Redhead 1995, Hutchinson *et al.* 1996, Hutchinson 2003a)

Environmental data layers for South and Central America, as used in species modelling software such as GARP (Stockwell and Peters 1999, Scachetti-Pereira 2002), are at a mix of scales, and have been prepared using different methodologies. A brief examination of the situation would indicate that not as much effort has gone into the selection and cleaning of the environmental layers as has gone into preparing the species data. Many of the layers have used the USGS 1km DEM (NGDC 2000) as the basis for the development of the climate layers. This DEM is variable in accuracy across different areas, and in some areas of South America has considerable error (NGDC 2002). McKenney and others (*pers. com.*) recommend the preparation of a new DEM using different methodologies such as those of Hutchinson (2003b), especially if new and finer climate layers are to be based on them.

The climate layers (precipitation, temperature, radiation, etc.) being used for modelling in South America, until recently at least, have generally been at 0.5º resolution, and this is far too coarse a scale for use in effective species modelling. Climate grids do exist for South America at finer resolution (10 arc minutes, for example), and these were introduced into GARP modelling at CRIA during 2003 (Chapman 2004). Until recently (early 2004), the finest scale climate layers covering all of South America were those created by CIAT using the methods of Jones (Jones *et al.* 1990, Jones 1995, Hijmans 1999). Jones uses these in his own species modelling tool – FloraMap (Jones and Gladkov 2001) and after downloading these from the CIAT site in June 2003, they began to be used in other modelling programs such as GARP. The recent (first quarter 2004) creation of Global 30-second climate surfaces (Hijmans *et al.* in prep.) will now allow, for the first time, layers at a suitable scale to be used for species' modelling in South and Central America (Chapman 2004). [NB This data is still in draft form and is not scheduled for public release until March or April 2004].

I believe that the ideal grid size for modelling on a continental basis is about 1/20th of a degree or 3-arc minutes (about 5 km). It may be necessary to reclassify the 30-second surfaces up to a scale approaching this and I understand the developers of the 30-second surfaces are planning to create 2.5-minute surfaces from those in the near future.

One problem with layers at too coarse a scale can be seen in figure 3 where the grids do not necessarily cover all land surfaces, especially offshore islands. The collection shown (green circle) is outside the climate grid (shown in light grey), although validly placed on the island.

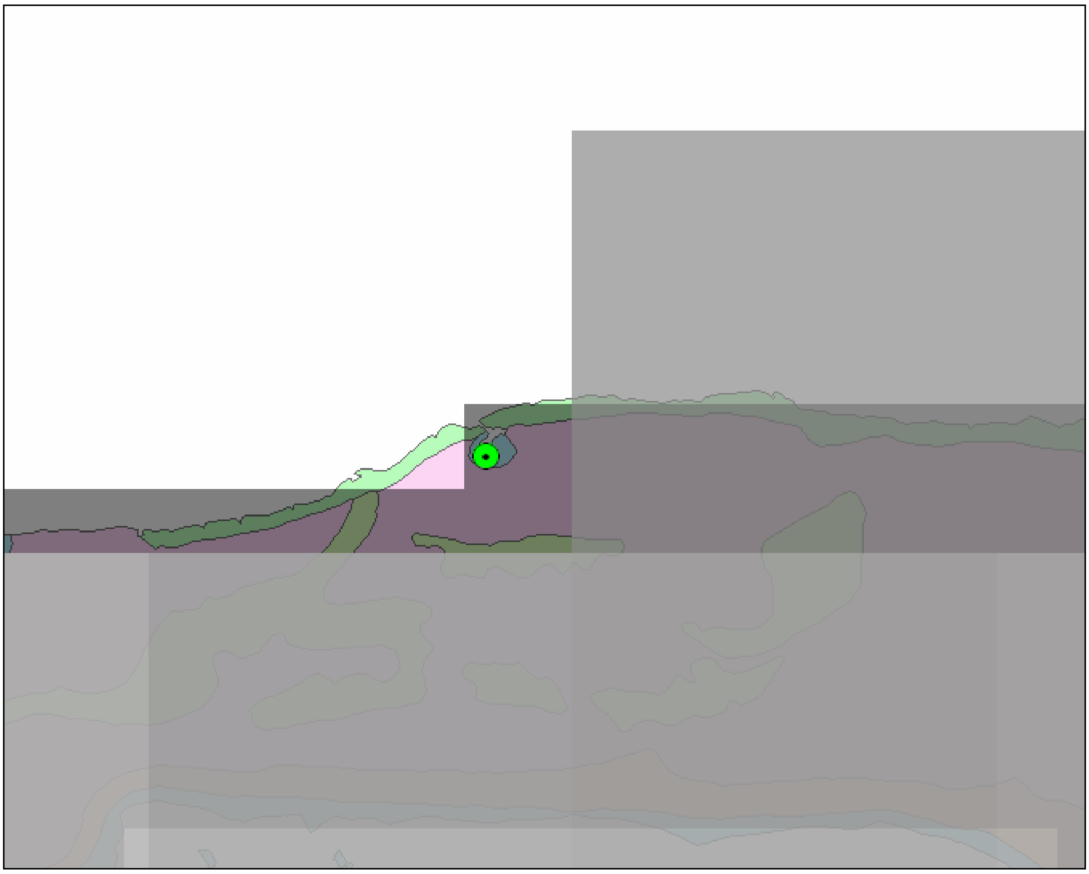

Fig 3. Example of environment layers of varying scale, showing how, in some cases, the grids do not cover the complete land surfaces.**[Get better example]**

Marine environmental data includes data such as depth, seabed surface, tidal ranges and flows, currents, prevailing winds, sea surface temperature and bathymetry (Margules and Redhead 1995). These have varying degrees of accuracy depending on where they have been obtained. As these layers are little used in species modelling at this stage, I don't intend to elaborate on them further. A similar situation exists with aquatic data which includes water temperature, pH, turbidity, oxygen content, chemical composition, water flow, etc. Although much of this data is available, its use in environmental modelling is till in its infancy, and thus is not covered further in this paper.

Fitness for Use

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, is not necessarily of low quality. Quality only comes into being once the data is being used and is not a character of the data *per se*. For example, a collection record that only has 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 a relative term, and is a factor of fitness for use.

All data will include error – there is no escaping it. It is knowing what the error is that is important, and knowing if the error is within acceptable limits for the purpose to which the data is being put. 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. This is where metadata comes to the fore for datasets as a whole, and indeed it is in the area of metadata development that the term “fitness for use” has come to prominence. The concept did not became fully recognised as an important one with spatial information until the early nineties, and it wasn’t until the mid 90s that it started to appear in the literature in this context (e.g. Agumya and Hunter 1996). Recording information only at the dataset level, however, will not always supply the information that the user requires. Recording error at the record level, especially with species data, can be extremely important for determining the fitness of that record for use. Thus the level of accuracy of each given geocode should be recorded within the database. I prefer this to be in non-categorical form, and be 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 that data that is better than a certain metric value – e.g. better than 5,000 meters. It is also important that automated georeferencing tools include calculated accuracy as a field in the output. The CRIA-developed Localidade (CRIA 2004a) already includes the feature, and I understand that the BioGeoMancer program (Beaman *et al.* 2003), which is still under development, intends to include this feature (Beaman *pers. com.* 2002).

It is also important that users of the data understand the concept of fitness for use. All too often species data are extracted from a database in a “record no., x, y” format regardless of any accuracy information that may be present. The geocode itself is always a point, but it seldom, if ever, represents a true point. Some records may have been entered into a database with an arbitrary point (for example a collection that just has “South America” on the label), and given an accuracy of 5 000 000 meters in the accuracy field. There are some databases that do this! To extract the record and use it’s arbitrary point will be extremely misleading. Users need to be made aware that there is an accuracy field, and be advised on how to use it. Otherwise, data reports could include a field for accuracy that must be filled in before the data can be supplied.

Identifying Error

Identifying errors in data in which the error is not fully documented is not an easy task. The majority of specimen based museum and herbarium data falls into this category. As mentioned above, some methods have been developed to identify possible errors in species’ data, however more can, and needs, to be done.

The testing of errors in already assigned geocodes can involve:

- checking against other information internal to the record itself or between records within the database - 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?;
- checking against an external reference using a GIS – does the record fall on land rather than at sea?;
- checking for outliers in geographic space; or
- checking outliers in environmental space.

Details on the application of these tests is given in Report 6 on Data Cleaning Tools (Chapman 2003b). It is the principle only that is elaborated here.

a. Internal checks

Internal checks check for logical consistency in relationships between fields within a record or between records. This generally involves checking information in one field against information in one or more other fields within the database. Most of these checks are simple, and generally can only be used to check textual information.

Most species and species-related databases include a certain amount of redundant information. For example, the State in which the collection was made as well as a field for textual location information. Some databases also include a "nearest named place" and this may also duplicate information within the locality field. Checks can therefore be made to check that the cited town or nearest named place in one field, is located within the correct State or district, or even country as cited in another field.

.

Not all databases are set up correctly initially, and this can allow errors that should never occur. For example, latitudes greater than 90º or less than –90º and longitudes greater than 180º or less than -180º. If these are permitted by the database, then the database needs to be modified, otherwise checks need to be run on a regular basis to ensure that errors like these do not occur and are corrected.

b. external reference using a database

More sophisticated use of databases can be used to check the accuracy of certain fields by comparing one database against another. For example if a database includes a figure for altitude, as well as a geographic reference, the altitude can be checked against a Digital Elevation Model (DEM).

Gazetteers exist for most of the world in one form or another, and generally these are available as databases. This means that they can be used to check appropriate fields within the specimen database for accuracy. Care needs to be exercised with the use of many of these databases as often they, themselves, contain errors (see for example figure 4). Also, many named places may be ambiguous (e.g. there are hundreds of "Sandy Creek"s in Australia) (Chapman and Busby 1994), or involve historic place names that do not occur in the modern gazetteer. There is also the problem of what a place name may actually mean (Wieczorek 2001).

One method that is seldom used, but that has great potential, is cross checking against databases of collectors' localities. Very few databases exist at present on collectors and their itineraries, but they are gradually being developed. Peterson *et al.* (in press) recently suggested a novel statistical method 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, they identified possible errors in specimens that fall outside the calculated range. Similar methods to this could be carried out in the database itself. Such a method will only work, however, if the databased collections from the collector are large enough to create such an itinerary.

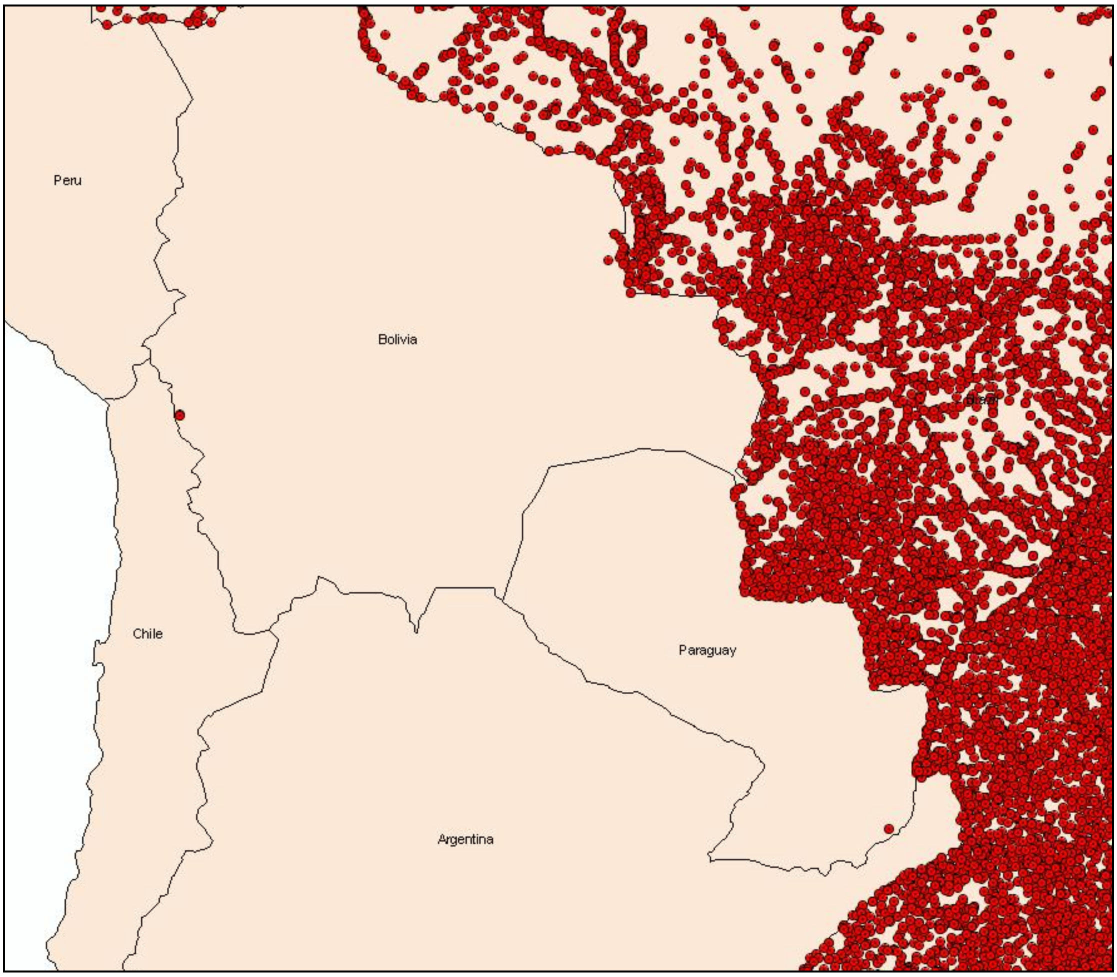

Fig. 4. Records from a Gazetteer of Brazilian place names showing a number of errors, with one obvious error sitting on the Chile-Bolivian border and another in southern Paraguay, and a number of others just over the Brazilian border in Bolivia.

c. external reference using a GIS

There are any number of ways in which a GIS (Geographic Information System) can be used to check and identify geocoding errors in a specimen database. Many of the methods mentioned above, that use checks against other databases, can also be carried out using a GIS, and often more efficiently. For example, checking against gazetteer locations, checking against collector's itineraries, etc.

The simple plotting of records in a GIS will often quickly identify obviously misplaced records. The most obvious of these is the misplacement of terrestrial records out to sea, and vice versa. A common problem with herbarium records is artificially cultivated records, for example those cultivated in a Botanic or private garden. Although the geocode may be accurate, one would not want to include these records in most biogeographic or species modelling studies. Ideally, the database would include a field to flag these records, but this is not always the case, and when the field is present data entry personnel often forget to properly fill it out. A GIS is a handy way to identify these records and thus allow for subsequent flagging.

A GIS can also be useful in identifying records that occur within particular geographic regions – within a country, a division, on an island, national park, etc. If the collections are all from a particular country, for example, it is quickly obvious if there are records that fall outside those boundaries. Figure 4 shows an example using gazetted place names. Misplaced specimen records can be identified in a similar manner as misplaced locations.

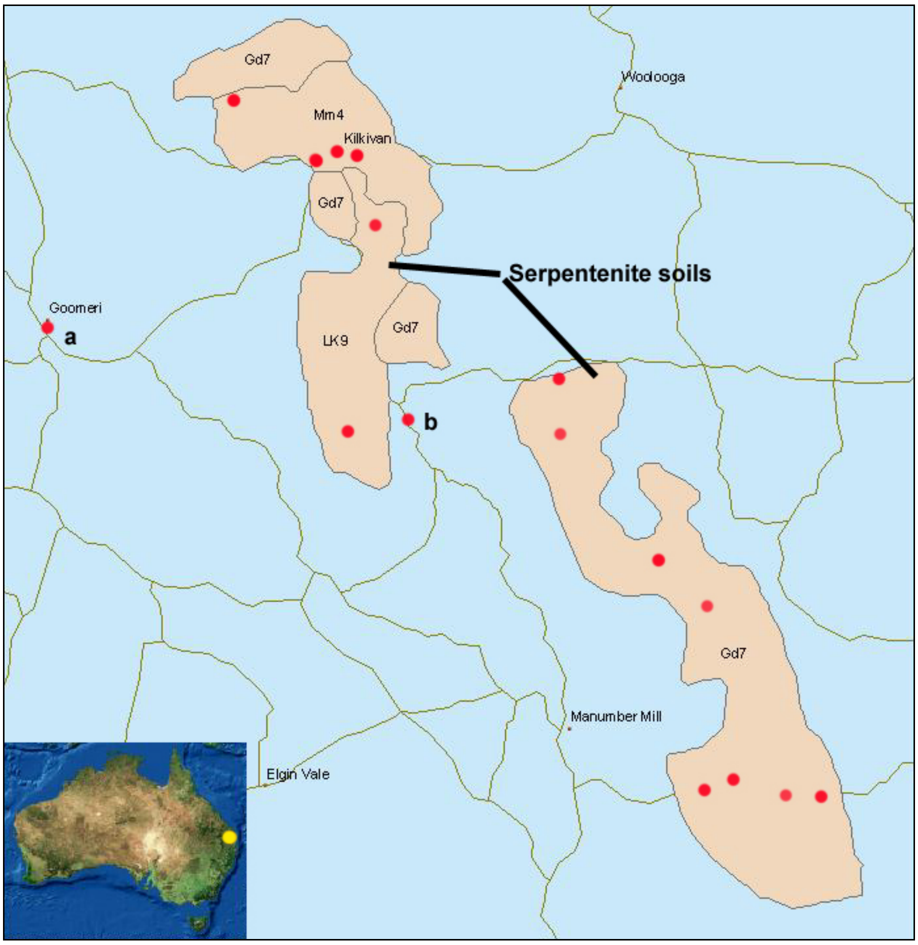

Fig. 5. Records of a species (red) that is only found on highly mineralised Serpentenite soils. Records marked ‘a’ and ‘b’ have likely errors in geocoding.

The GIS can also be used to check that records fall outside a particular vegetation type, soil type or geology, etc. Some species are highly specific to certain geological types - limestone, sandstone, serpentenite (figure 5). If you have the boundaries of these, any record that falls outside may be regarded as a possible outlier and flagged for further checking (Chapman *et al.* 2001 and in prep.). In figure 5, a species that occurs only on highly mineralised Serpentenite soils is mapped and two records (marked ‘a’ and ‘b’) show up as likely errors. On checking, record ‘a’ only has the locality ‘Goomeri’ – the nearest town to the Serpentenite outcrop, and has been geocoded with the latitude and longitude of the town. Record ‘b’ is quite near the outcrop and is likely misplaced due to the precision of the cited geocode. However, as outliers, it may be important to flag them for further checking against the original specimen locations.

The identification of collector's itineraries (Chapman 1988, Peterson *et al.* in press), allows for checking for possible error if, for example, the date of collection doesn't fit the particular pattern of that collector. This could be particularly useful for collectors from the 18th and 19th centuries prior to collectors being able to cover vast distances within the one day using helicopters, planes or motor vehicles. In the example given in figure 6, collections between 11 and 25 April 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 (Chapman 1988). Again, using a GIS to map both the itinerary and the collections can be very valuable.

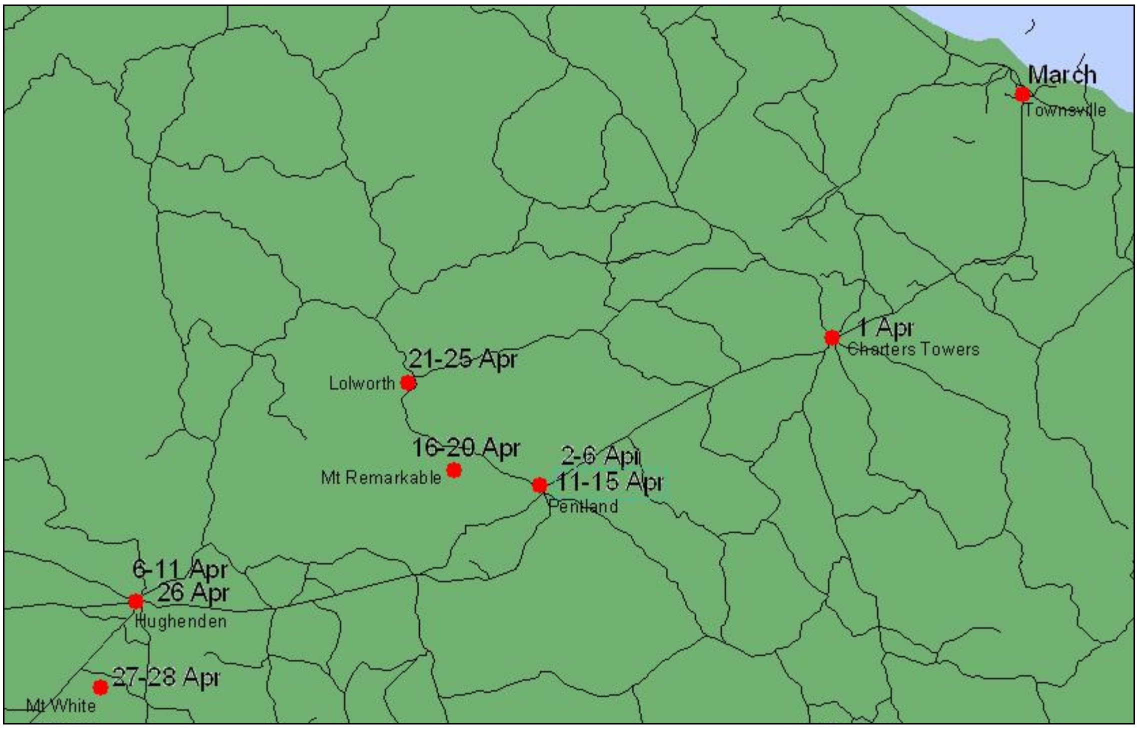

Fig. 7. Collecting localities of Karl Domin in Queensland, Australia in 1910 (Chapman 1988). He travelled by train from Townsville to Hughenden, stopping at Charters Towers and Pentland. He then returned and spent 15 days in the Pentland, Mount Remarkable, Lolworth area on horseback, before returning to Hughenden by train. [NEEDS TO BE CHECKED BEFORE PUBLICATION – TAKEN FROM MEMORY]

Another use of a GIS is in buffering likely locations – e.g. streams for fish and aquatic plants, the coast for littoral species, altitudinal ranges for alpine species or others known to have a distinct altitudinal range. In this way, anything outside the buffer may need to be checked. Care needs to be exercised, as with the fish, for example, it may mean that those records outside the buffer zone are not errors at all, but the species may be occurring in small streams too small for mapping. These tests can generally only flag suspect records, and then it is up to individual checks to determine what may be real errors in the record, and what may be true outliers.

d. outliers in geographic space

The use of statistics to identify outliers in geographic space is an option that has seldom been used. This could be especially valuable with latitude, as many species tend to have a significant latitudinal correlation. Similarly, many species tend to have an altitudinal correlation. Any number of statistical formulae could be used to detect outliers, for example, a similar (reverse-jackknifing) method previously used by me (Chapman 1999) to identify outliers in climate space, has recently been used by CRIA to develop an on-line outlier detection tool (CRIA 2004b).

e. outliers in environmental space.

Environmental modelling, although around for more than 20 years, is becoming more and more important as environmental layers are improved in quality and resolution. Using environmental modelling software to detect possible errors in data has been around for many years (Nix 1986, Busby 1988, 1991, Lindemeyer *et al.* 1991). Many of these methods have used the identification of outliers using cumulative frequency curves (Nix 1986, Busby 1991, Lindemeyer *et al*. 1991, Houlder *et al*. 2000, Hijmans *et al.* 2003) and this has been quite successful. It has limitations where there may be no, or many outliers, but these can generally be worked around. Others methods have used Principal Components Analysis (Jones and Gladkov 2001), Cluster Analysis (Jones and Gladkov 2001) and reverse-Jackknifing (Chapman 1992, 1999) all with considerable success. Beaman and others (2003) have suggested using previously modelled distributions, for example those produced using GARP (Stockwell and Peters 1999, Scachetti-Pereira 2002) or Lifemapper (University of Kansas 2003b) to look for records that fall outside the previously modelled distribution. This method has a major flaw, where the previously computed model does not include records covering the complete range of the species – for example if it was restricted to records from one State (Chapman 2003c).

See Report 6 on Data Cleaning Tool for details of each of the methods mentioned above, along with software tools that incorporate some of the methods (Chapman 2003b).

Documenting Error

One of the keys to knowing what error exists in data is documentation. It is of very little use to anyone if checks of data quality are carried out, and corrections made if it is not fully documented. This is especially important where these checks are being carried out by other than the originator of the data. There is always the possibility that perceived errors are not errors at all, and that changes that are made, add new error. It is also important that checking not be done over and over again. We cannot afford to waste resources in this way. For example, data quality checks carried out on data using one of the above methods may identify a number of suspect records. These records may then be checked and found to be perfectly good records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect. This person may then exclude the records from their analysis, or spend more valuable time rechecking the information. This is basic risk management, and should be carried out routinely by all data custodians and users. The value and need for good documentation cannot be stressed too heavily. It assists users in knowing what the data is, what the quality is, and what purposes the data are likely to be fit for. It also aids curators and data custodians to keep track of the data and its quality and to not waste resources rechecking supposed errors.

One of the ways of making sure that error is fully documented is to include it in the early planning stages of database design and construction. Additional data quality/accuracy fields can then be incorporated. Fields such as positional or geocode accuracy, source of information such as for the geocode and elevation, fields for who added the information – was the geocode added by the collector using a GPS, or a data entry operator at a later date using a map at a particular scale, was the elevation automatically generated from a DEM, if so, what was the source of the DEM, its date and scale, etc. All this information will be valuable in later determining whether the information is of value for a particular use or not, and the user of the data can then decide.

Two good examples of documenting error are those used by MaPSTeDI – a collaborative effort between the University of Colorado Museum, Denver Museum of Nature and Science and the Denver Botanic Gardens (University of Colorado 2003), and those used by the MaNIS project in California (Wieczorek 2001). These are projects initiated to integrate collections between different institutions, and in order to do this efficiently, developed guidelines that are of value not only to those institutions, but more generally.

Information presentation

A neglected area of information presentation with biological data is the presentation of error or uncertainty (Agumya and Hunter 1996). Very few collections institutions provide information on the accuracy of individual records, or on the methods used to detect error or improve the quality and accuracy of the data. Species based environmental models usually present just one hypothesis out of many, and show the possible, or likely distribution of a species. The outputs of most models are presented as a map of two or more colours that attempt to inform the user that the species is likely to or may occur in a particular area. Very seldom are certainty intervals displayed on the map, or information given as to the likely error inherent in the map. There are a number of ways that this may be done, and one that I advocate is the inclusion of a probability surface on model outputs rather than using a straight "yes/no" approach. This approach provides the users – decision makers, environmental managers, etc. – with increased information on which they can base their decision. As part of good Risk Assessment policy, information should be supplied with the outputs of all models, of the quality of the various input data that went to make up the model.

Conclusion

Data quality is an important issue with all data, be they museum or herbarium collection data, meteorological data, or derived information such as climate surfaces, species' models, vegetation maps, etc. The associated report: Report 6 on Data Cleaning Tools (Chapman 2003b), sets out a range of methods that can be used to improve the quality, especially of museum and herbarium specimen data, and recommends the development of a Data Cleaning Toolkit to include, among other things, various software tools, guidelines, and links to on-line services.

The importance of data quality and error checking can not be stressed too strongly. It is essential if the data are to be of real value in developing outputs that will lead to improved environmental decisions and management.

References:

Agumya, A. and Hunter, G.J. (1996) Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. *Proceedings of the GIS/LIS '96 Conference*, Denver, Colorado, pp. 359-70

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

Austin, M.P. (2002). Case Studies of the Use of Environmental Gradients in Vegetation and Fauna Modeling: Theory and Practice in Australia and New Zealand pp. 73-82 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press.

Bannerman, B.S. (1999). *Positional Accuracy, Error and Uncertainty in Spatial Information*. Howard Springs, NT, Australia: Geoinovations. http://www.geoinnovations.com.au/posacc/patoc.htm

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.

Beaman, R.S. and others (2003). BioGeoMancer Lawrence, Kansas: University of Kansas – prototype http://www.biogeomancer.org/.

Booth, T.H. (1996). Matching Trees and Sites. Proceedings of an international workshop held in Bangkok, Thailand, 27-30 March 1995, *ACIAR Proceedings* No. 63.

Burbidge, A.A. (1991). Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO

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

Busby, J.R. (1988). Potential impacts of climate change on Australia's flora and fauna pp. 387-398 **in** Pearman, G.I. (ed.). *Greenhouse: Planning for Climate Change*. Melbourne:

Busby, J.R. (1991). BIOCLIM – a bioclimatic analysis and prediction system. Pp. 4-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO

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. (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. (2003a). *The Case for a 3 minute Climate Surface for South America*. Internal report No. 3 to CRIA- May 2003.

Chapman, A.D. (2003b). *Environmental Data Quality – b. Data Cleaning Tools*. Internal report No. 6 to CRIA.– June 2003.

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

Chapman, A.D. (2004). 1 km Climate Surface for South America. Internal report No. 3b to CRIA. January 2004.

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

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.

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

Cofinas, M., M.P. Bolton, A.J. Bryett, D.C. Crossley and A.L. Bull (1995). *Flora Data and Modelling for Cape York Peninsula*. Cape York Peninsula. Land Use Strategy Natural Resources Analysis Program, Brisbane/Canberra.

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

Conn, B.J. (ed.) (1996). *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3. Sydney: Royal Botanic Gardens.

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://www.rbgsyd.gov.au/HISCOM/.

CRIA (2004a). *Lista de Localidades Brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://www.cria.org.br/localidade/ [Accessed 20 Jan 2004].

CRIA (2004b). *Outliers in Geographic Space*. Campinas: Centro de Referência em Informação Ambiental. http://www.cria.org.br/outlier/ [Accessed 20 Jan 2004]

Cunningham, D., Walsh, K. and Anderson, E. (2001*). Potential for Seed Gum Production from Cassia brewsteri.* RIRDC Project No. UCQ-12A. Kingston, ACT: Rural Industries Research and Development Corporation. http://www.rirdc.gov.au/reports/NPP/UCQ-12A.pdf.

Duckworth, W.D., Genoways, H.H. and Rose, C.L. (1993). *Preserving Natural Science Collections: Chronicle of our Environment Heritage*. Washington, DC: National Institute for the Conservation of Cultural Property 140pp.

Faith, D.P. and Walker, P.A. (1996). Environmental diversity: on the best possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation* **5:** 399-415

Faith, D.P., Walker, P.A., Margules, C.R., Stein, J. and Natera, G. (2001). Practical application of biodiversity surrogates and percentage targets for conservation in Papua New Guinea. *Pacific Conservation Biology* **6:** 289-303 http://wwwscience.murdoch.edu.au/centres/others/pcb/toc/pcb_contents_v6.html

Ferrier, S. and Watson, G. (1997). *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*. Canberra: Department of Environmnet, Sport and Territories.

Geoscience Australia (2003a). *Geodata TOPO-250K (Series 1) Topographic Data* . Canberra Geoscience Australia. http://www.auslig.gov.au/meta/meta.htm

Geoscience Australia (2003b). *Differences between coordinate systems*. Canberra Geoscience Australia. http://www.ga.gov.au/nmd/geodesy/datums/aboutdatums.jsp

Hijmans, R.J. (1999). Global Climate Surfaces at a 10' Resolution. *Production Systems and Natural Resource Management* Working Paper No. 3. Lima, Peru: International Potato Centre.

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

Hijmans, R.J., Cameron, S., Para, J., Jones, P., Jarvis, A. and Richardson, K. (in prep.). *Worldclim Version 1.1*. Berkeley, CA: Museum of Vertebrate Zoology. [http://bnhm.berkeley.museum/gisdata/worldclim/worldclim.htm].

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

Hutchinson, M.F. (1991). The application of thin plate smoothing splines to continent-wide data assimilation. **in** Jasper, J.D. (ed), *Data Assimilation Systems*, Bureau of Meteorology Research Report No. 27, Bureau of Meteorology, Melbourne, pp. 104-113.

Hutchinson, M.F. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of GIS* **9:** 305-403.

Hutchinson, M.F. (2001). ANUSPLIN Version 4.2. Canberra: Centre for Resource and Environmental Studies, Australian National University. http://cres.anu.edu.au/outputs/anusplin.html.

Hutchinson, M.F. (1996). A locally adaptive approach to the interpolation of digital elevation models. In: NCGIA (ed.), *Proceedings of the Third International Conference Integrating GIS and Environmental Modeling*, Santa Fe, New Mexico, 21-25 January, 1996. University of California, Santa Barbara, National Center for Geographic Information and Analysis: CD-ROM and http://www.ncgia.ucsb.edu/conf/SANTA_FE_CD-ROM/main.html

Hutchinson, M.F. (2003a). Topographic and Climate Database for Africa. Version 1.0. Canberra: Centre for Resource and Environmental Studies, Australian National University. http://cres.anu.edu.au/outputs/africa.html.

Hutchinson, M.F. (2003b). ANUDEM Version 4.6.3. Canberra: Centre for Resource and Environmental Studies, Australian National University. http://cres.anu.edu.au/outputs/anudem.html

Jones, P.G. (1995). *Centro Internacional de Agricultura (CIAT) climate database version 3.41, Digital data tape*, Cali, Columbia: CIAT.

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

Jones, P.G., Robison, D.M. and Carter, S.E. (1990). *A geographical information approach for stratifying tropical Latin America to identify research problems and opportunities in natural resource management for sustainable agriculture in Centro Internacional de Agricultura Tropical (CIAT)*. Cali, Colombia: Agroecological Studies Unit, CIAT.

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.

Longmore, R. (ed.) (1996). Atlas of Elapid Snakes of Australia. *Australian Flora and Fauna Series* No. **7**. Canberra: Australian Government Publishing Service.

Mackey, A.P. (ed.) (1996). Prickly Acacia (*Acacia nilotica*) in Queensland. Brisbane: Queensland Department of Natural Resources and Mines. http://www.nrm.qld.gov.au/pests/psas/pdfs/Pricklyacacia.pdf

Mao, Z., P.D. and Huang, H. (1999). Automatic Registration of Sea WIFS and AVHRR Imagery **in** Xu, G. and Chen, Y. (eds). *Towards Digital Earth. Proceedings of the International Symposium on Digital Earth*. Science Press.

Margules, C.R. and Austin, M.P. (1994). Biological models for monitoring species decline: the construction and use of data bases. *Philos. Trans. R. Soc., London* **B344:** 69-74.

Margules, C.R. and Pressey, R.L. (2000). Systematic Conservation Planning. *Nature* **405:** 243-253.

Margules, C.R. and Redhead, T.D. (1995). *BioRap. Guidelines for using the BioRap Methodology and Tools*. Canberra: CSIRO. 70pp.

Neldner, V.J., Crossley, D.C., and Cofinas, M. (1995). Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biol. Conserv*. **73**: 1-17

Nicholls, N. (1997). Increased Australian wheat yield due to recent climate trends. *Nature* **387:** 484-485.

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

NGDC (2000). *Global Land One-kilometer Base Elevation (GLOBE) Digital Elevation Data* Version 1.0. Available on-line form: http://www.ngdc.noaa.gov/seg/fliers/globedem.shtml.

NGDC (2002). *Global Land One-kilometer Base Elevation (GLOBE) Documentation Version 1.0*. http://www.ngdc.noaa.gov/seg/topo/report/

OECD (1999). *Final Report of the Megascience Forum Working Group on Biological Informatics*. Paris: OECD

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., and Vieglais, D.A. (2001). Predicting species invasions using ecological niche modeling. *BioScience* **51:** 363-371.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R.H. and Stockwell, D.R.B. (2002a). Future projections for Mexican faunas under global climate change scenarios. *Nature* **416:** 626-629.

Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. (2002b). Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale.* Washington: Island Press.

Peterson, A.T., Navarro-Siguenza, A.G., Scachetti Pereira, R. (in press). Detection of errors in biodiversity data: Collectors' itineraries flag mislabeled specimens. Submetido à Bulletin of the British Ornithologists' Club

Pouliquen-Young, O. and Newman, P. (1999). *The Implications of Climate Change for Land-Based Nature Conservation Strategies*. Final Report 96/1306, Australian Greenhouse Office, Environment Australia, Canberra, and Institute for

Sustainability and Technology Policy, Murdoch University, Perth, Australia, 91 pp.

Sattler, P. and Williams, R. (eds) (1999). *The Conservation of Queensland's Bioregional Ecosystems*. Brisbane,Qld: Environment Protection Authority.

Pereira, R.Scacheiti (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.

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

Soberón, J., Golubov, J. and Sarakhán, J. (2000). Predicting the Effects of Cactoblastis cactorum Berg on the Platyopuntia of Mexico: A Model on the Route of Invasion pp. 95-97 **in** *Assessment and Management of Alien Species that Threaten Ecosystems, Habitats and Species*. CBD Technical Series No. 1. Montreal, Canada: Convention on Biological Diversity. http://www.biodiv.org/doc/publications/cbd-ts-01.pdf

Stockwell, D. and D. Peters (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**(2): 143-158.

Thackway, R. and Cresswell, I. (eds) (1995). *An Interim Biogeographic Regionalisation for Australia: A Framework for Setting Priorities in the National Reserves System Cooperative Program*. (Version 4.0) Canberra: Australian Nature Conservation Agency. http://www.ea.gov.au/parks/nrs/ibra/version4-0/index.html.

University of Colorado (2003). MaPSTeDI. A Guide to Georeferencing. Denver, CO: University of Colorado. http://mapstedi.colorado.edu/geocoding-howto.html [Accessed 20 Jan 2004].

University of Kansas (2003a). *Specify*. Biological Collections Management. Lawrence, Kansas: University of Kansas http://usobi.org/specify/.

University of Kansas (2003b). *LifeMapper*. Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/

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

Wieczorek, J. (2001). MaNIS: Georeferencing Guidelines http://dlp.cs.Berkeley.edu/manis/GeorefGuide.html

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.

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