Data Quality
Data Cleaning
&
Data Validation Techniques

Arthur D. Chapman

Organized by the
Belgian Biodiversity Platform, the Belgian GBIF node

GBIF

.be

ULB

2010 International Year of Biodiversity

Belgian Science Policy Office
belspo

List of participants

Name	Organization	Email
Brosens Dimitri	BBPF	Dimitri.brosens@inbo.be
Cael Garin	RMCA	Garin.cael@africamuseum.be
Davy James	RMCA	Davy.james@africamuseum.be
De Wever Aaike	RBINS	Aaike.dewever@naturalsciences.be
Debusscher Bos	KBIN	bosdebusscher@hotmail.com
Engledow Henry	NBGB	Henry.engledow@br.fgov.be
Gonzalez Barbara	BBPF	bgonzale@ulb.ac.be
Groom Quentin	NBGB	Quentin.groom@br.fgov.be
Hanquart Nicole	NBGB	hanquart@br.fgov.be
Heughebaert André	BBPF	aheugheb@ulb.ac.be
Kapel Michel	KBIN	mkapel@naturalsciences.be
Nicolas Noé	BBPF	niconoe@ulb.ac.be
Schmidt-Kloiber Astrid	BOKU	Astrid.schmidt-kloiber@boku.ac.at
Strobbe Francis	RBINS	Francis.strobbe@natuurwetenschappen.be
Theeten Franck	RMCA	Frank.theeten@africamuseum.be
Vanreusel Wouter	Natuurpunt	Wouter.vanreusel@natuurpunt.be

Data Quality
Data Cleaning
&
Data Validation Techniques

Training Manual for Courses

for

French, Belgian and Dutch Data Providers

Arthur D. Chapman

1-2 July 2010
(France)

5-6 July 2010
(Belgium)

8-9 July
(The Netherlands)

Organised by
GBIF France, BeBIF & NLBIF

Introduction

This document is for training courses on Data Quality, Data Cleaning and Data Validation Techniques for French, Dutch and Belgian data providers run by the GBIF Nodes in The Netherlands (NLBIF), France (GBIF France) and Belgium (BeBIF).

The first training event is being conducted in Paris (1-2 July 2010), then Brussels (5-6 July 2010), and finally in Amsterdam (8-9 July 2010).

Workshop content specifically aims at GBIF data providers and would-be data providers, who would like to improve the quality of their databases. The workshops are managed by Arthur Chapman (Australian Biodiversity Information Services) an authority in the field of data quality and data validation. Training sessions will be conducted in English.

Chapman is author of GBIF documents on topics to be covered in the workshop:

- *Uses of Primary Species-Occurrence Data*
- *Principles of Data Quality*
- *Principles and Methods of Data Cleaning*
- *Guide to Best Practices for Generalising Sensitive Species Occurrence Data*

Another important document will be the *BioGeomancer Guide to Best Practices for Georeferencing*.

All five of these documents are available as PDF files, under Print and Online Resources: Booklets on the GBIF website and have also been incorporated into the GBIF Training Manual (GBIF 2008).

Key References:

Chapman, A.D. (2005a). *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2005. 61pp. Copenhagen: GBIF. Available in English, Chinese and Korean. http://www2.gbif.org/DataQuality.pdf

Chapman, A.D. (2005b). *Principles and Methods of Data Cleaning.* Report for the Global Biodiversity Information Facility 2005. 75pp. Copenhagen: GBIF. Available in English, Chinese and Korean. http://www2.gbif.org/DataCleaning.pdf

Chapman, A.D. (2005c). *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2005. 111pp. Copenhagen: GBIF. Available in English, Chinese and Korean. http://www2.gbif.org/BioGeomancerGuide.pdf

Chapman, A.D. and Wieczorek, J. (eds). (2006). *Guide to Best Practices for Georeferencing*. BioGeomancer Consortium. 90pp. Copenhagen: GBIF. Available in English and shortly in French. http://www2.gbif.org/BioGeomancerGuide.pdf

Chapman, A.D. and Grafton, O. (2008). *Guide to Best Practices for Generalizing Sensitive Species-Occurrence Data*, version 1.0. 27 pp. Copenhagen: Global Biodiversity Information Facility. Available in English. http://www2.gbif.org/BPsensitivedata.pdf

GBIF (2008). *GBIF Training Manual 1: Digitisation of Natural History Collections Data*. 518pp. Copenhagen: GBIF. http://www2.gbif.org/TM1.pdf

Agenda

Hands-on data validation
1-2, 5-6, 8-9 July 2010

Day 1 : France — Thursday, July 1
Belgium — Monday, July 5
Netherlands — Thursday, July 8

09:00	Welcome
	Introduction to Principles of Data Quality
	Introduction to Data Cleaning : • Taxonomic and Nomenclatural Data
10:00	Coffee Break
	Exercise 1 – Taxonomic and Nomenclatural Data
	• Introduction to Georeferencing Best Practices
12:45	Lunch Break
14:00	Exercise 2 – Determining Maximum Uncertainty
	Exercise 3 – Assigning Coordinates
15:00	Discussion
15:30	Finish

Day 2 : France — Friday, July 2
Belgium — Tuesday, July 6
Netherlands — Friday, July 9

09:00	Welcome
09:00	Introduction to Data Cleaning : • Spatial Data
	Diva GIS Software Installation (see http://www.diva-gis.org)
	Exercise 3 – Coordinates checking with Terrestrial Data
10:45	Coffee Break
	Exercise 4 – Coordinates checking with Marine Data
	Exercise 5 – On-line Data Cleaning Tools
	Data Generalization
	Discussion
12:45	Lunch Break
14:00	More Exercises on your own dataset
15:00	Discussion
15:30	Finish

CONTENTS

EXERCISE ON TAXONOMIC AND NOMENCLATURAL DATA QUALITY

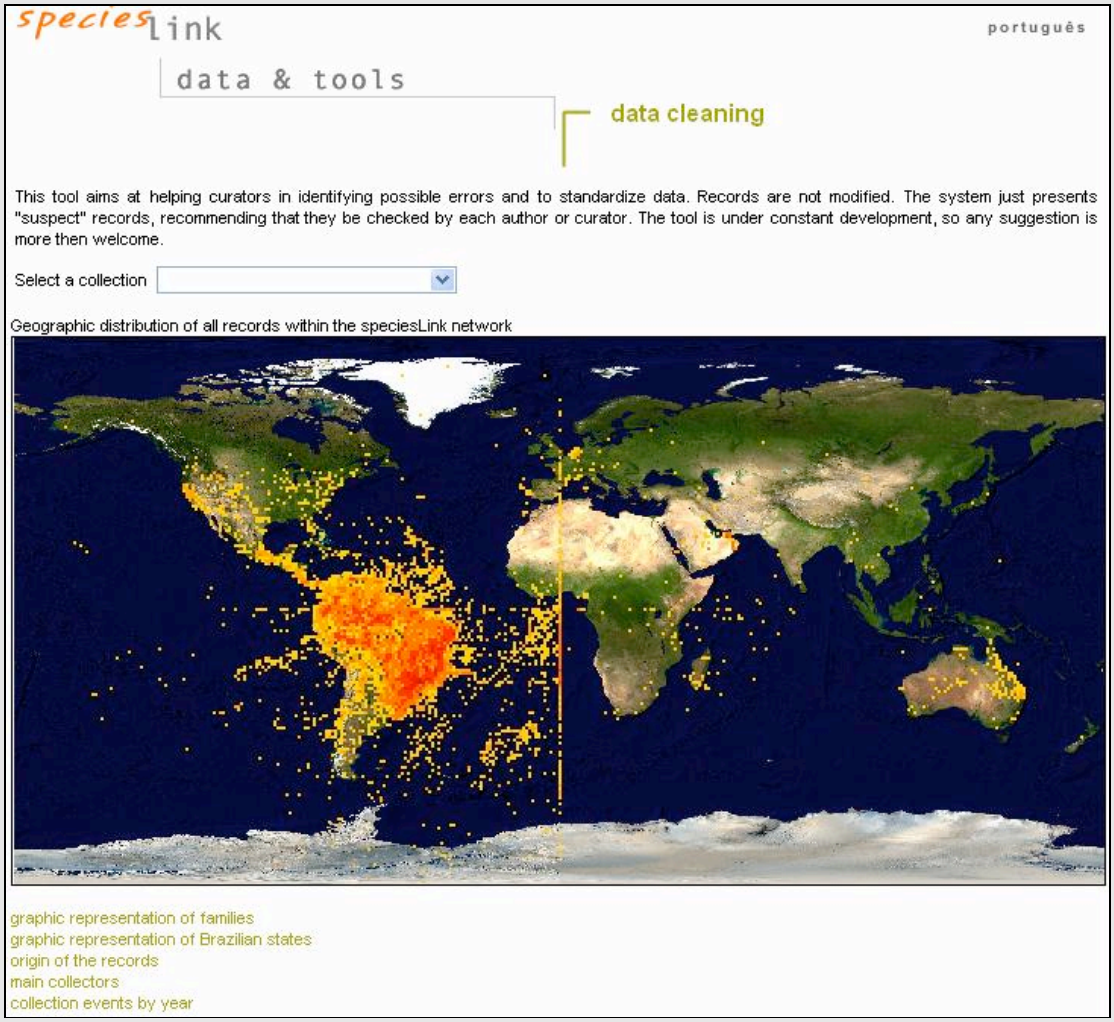
1. CRIA – Reference Center for Environmental Information, Brazil.

Using your Web Browser:

Go to **http://splink.cria.org.br/dc**

[**NB.** You must allow your Web browser to pop up additional windows]

Click on english unless you are more comfortable in Portuguese.



NB. The databases sited here are constantly updating and cleaning their data – so errors documented here may have already been corrected by the time you read this so I suggest if one database doesn't work, try another.

Under : **Select a collection:** select a collection from the drop down menu – for example, '**HSJRP**' (Herbário de São José do Rio Preto)

Look at the area '**Taxonomic data**'

taxonomic data	
inventory	scientific name - collector - types
family	380 suspect records
genus	295 suspect records
species	211 suspect records
subspecies	not found
author	112 suspect records
duplicate	820 suspect records

Let us look first at the Family names. One would suspect that these should be entered using a pick list, but with this database that is not the case, and it is a good example of the problems that arise.

Click on: '**family**'

Note the errors (identified in **Red**).

Close that window and now look at the Genera by clicking on '**genus**'

When finished there we can look at the species by clicking on '**species**'

Click on one of the light coloured numbers under 

and it will bring up information on the collection.

When you are ready – we will click on one of the 

buttons.

This will create a new window and do a search through a number of in-house and external databases as well as showing what the situation is with that name in the ***Catalogue of Life***.

You can now explore some of the other categories such as **author**, **subspecies**, etc.

Duplicates

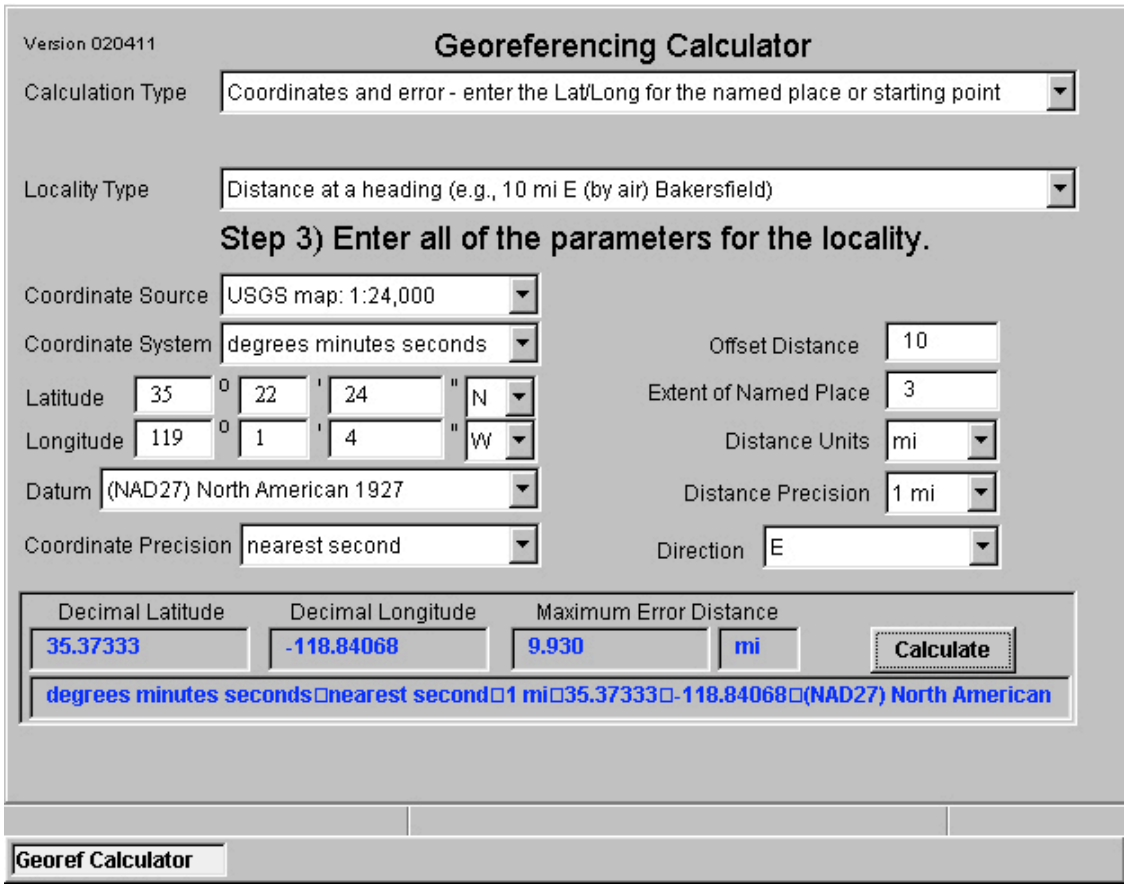
We will now look at Duplicates of collections that may be cited in other collections.

Click on: '**duplicate**'

This will show you differences that may occur between this collection and collections in other databases, but which are based on the same "Collector" and "Collector number".

DETERMINING MAXIMUM UNCERTAINTY IN GEOREFERENCED DATA

Use the Georeferencing Calculator (Wieczorek 2002[1])



We are going to work through several examples in order to determine a value for the Maximum Uncertainty Distance.

[1] Wieczorek, J. 2002. Manual for Georeferencing Calculator. MaNIS/HerpNet/ORNIS. University of California, Berkeley: Museum of Vertebrate Zoology. http://manisnet.org/CoordCalcManual.html [Accessed 28 Jan. 2006].

1. Named Place

Example 1
Locality: "Liège", Belgium.

Using your Web Browser

Go to **http://www.manisnet.org/gc.html**

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Named place only".

Coordinate Source: gazetteer
Coordinate System: degrees minutes seconds
Latitude: 50º 38' N
Longitude: 5 º 40' E
Datum: not recorded; 79 m uncertainty
Coordinate Precision: nearest minute; 2400 m uncertainty
Extent of Named Place: 3 km
Distance Units: km
Decimal Latitude: 50.63333
Decimal Longitude: 5.66667
Maximum Uncertainty Distance: 6.197 km

2. Offset Distance

Example 2 (Distance Only)
Locality: "5 km from Feurs", France.
Suppose the coordinates for the locality were interpolated to .001 of a degree obtained from a Gazetteer and the distance from the centre of Feurs to the furthest city limit is 3.5 km.

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Distance only (e.g., 5 mi from Bakersfield)".

Coordinate Source: gazetteer
Coordinate System: decimal degrees
Latitude: 45.744
Longitude: 4.223
Datum: European 1979; no uncertainty
Coordinate Precision: 0.001 degrees; 135 m uncertainty
Offset Distance: 5 km
Extent of Named Place: 3.5 km
Distance Units: km
Decimal Latitude: 45.744
Decimal Longitude: 4.223
Maximum Uncertainty Distance: 9.636 km

3. Offset at a Heading

Example 4

Locality: "15 km ENE (by air) from Oss, The Netherlands"

Suppose the coordinates for the locality were interpolated to the nearest second from the relevant 1:25,000 map and the distance from the centre of Oss to the furthest city limit is 5 km.

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Distance at a Heading".

Coordinate Source: non-USGS map: 1:25,000; 12 m uncertainty
Coordinate System: degrees, minutes, seconds
Latitude: 50º 40' 23" N
Longitude: 5º 32' 15"E
Datum: WGS 84; no uncertainty
Coordinate Precision: nearest second; 0.024 mi uncertainty
Offset Distance: 15 km
Extent of Named Place: 5 km
Distance Units: km
Distance Precision: 1 km
Direction: ENE - Precision: 11.25 degrees either side of ENE
Decimal Latitude: 50.67306
Decimal Longitude: 5.5375
Maximum Uncertainty Distance: 6.995 km

4. Offset along a Path

Example 5

Locality: "13 mi E (by road) Bakersfield, USA"

Suppose the coordinates for this locality were interpolated to the nearest 1/10th minute from the USGS Taft 1:100,000 Quad map and the distance from the center of Bakersfield to the furthest city limit is 2 mi.

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Distance along a Path".

Coordinate Source: USGS map: 1:100,000; 0.032 mi uncertainty
Coordinate System: degrees, decimal minutes
Latitude: 35º 26.1' N
Longitude: 118º 48.1'W
Datum: NAD27; no uncertainty
Coordinate Precision: 0.1 minutes; 0.148 mi uncertainty
Extent of Named Place: 2 mi.
Distance Units: mi
Distance Precision: 1 mi.
Decimal Latitude: 35.435
Decimal Longitude: -118.80167
Maximum Uncertainty Distance: 3.211 mi

5. Latitude and Longitude Coordinates

Example 6
Locality: "35 ° 22' 24" N, 119º1' 4" W**"**

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Coordinates Only".

Coordinate Source: locality description
Coordinate System: degrees, minutes, seconds
Latitude: 35º 22' 24" N
Longitude: 119º 1' 4" W
Datum: not recorded; 79 m uncertainty
Coordinate Precision: nearest second; 40 m uncertainty
Distance Units: km, m, mi, yds, or ft
Decimal Latitude: 35.37333
Decimal Longitude: -119.01778
Maximum Uncertainty Distance: 0.119 km, 118.8 m, 0.074 mi, 129.9 yds, or 390 ft

Example 7
Locality: "35.37,-119.02, NAD27, USGS Gosford Quad 1:24 000**"**

For **Calculation Type**: use:
"Error – enter Lat/Long for the actual locality"
For **Locality Type:** use:
"Coordinates Only".

Coordinate Source: USGS map: 1:24,000; 12 m uncertainty
Coordinate System: decimal degrees
Latitude: 35.37
Longitude: -119.02
Datum: NAD27; no uncertainty
Coordinate Precision: .01 degrees; 1434 m uncertainty
Distance Units: km, m, mi, yds, or ft
Distance Precision: 1 mi.
Decimal Latitude: 35.37
Decimal Longitude: -119.02
Maximum Uncertainty Distance: 1.446 km, 1446 m, 0.899 mi, 1582 yds, or 4745 ft

6. Using the BioGeomancer Workbench

Example 1
Locality: "13 km WNW, Liège, Belgium".

Using your Web Browser

Go to **http://www.biogeomancer.org/**

Click on **APPLICATIONS**
Click on **> Biogeomancer Workbench**
In the First paragraph – Click on **HERE**

Type "**13 km WNW, Liege, Belgium"** into text box

NB – don't use the accent in this example!

You will notice that four circles have come up –
Two (the blue pointers) are derived from the GADM dataset and have large uncertainties (24 926 meters and 58 107 meters).
The other two (the green pointers) are derived from the GeoNet dataset and have finer uncertainties (8522 meters and 4880 meters).
These are for "Liège" – one perhaps for the county.

If you had used Liège instead of Liege – only three circles would have appeared.

Click on the  in the centre of the smaller circle.

You can now click on "**delete others**" as this is the one you are interested in.

Click on "**zoom in**"

Click "**edit uncertainty**"

Click on the  and drag it to the intersection of N614 and the A3

[For example you know that you collected it at this corner]

You can now click on the  and it will reset the uncertainty circle.
You can also drag it to a smaller or larger circle if you think you know the limits of the uncertainty better than shown. You may wish to turn on the Satellite, or Hybrid in the top right hand corner to help you with this.

You will notice that the uncertainty value has altered, along with the geographic coordinates

For fun, now – delete the Belgium from the text box and Georeference again!

DOWNLOAD AND INSTALL DIVA-GIS

1. Download Diva-GIS from the Internet

NB – if you have a slow Internet connection, you can skip this step and install from the CD, however, to get the latest version and to register the software we recommend you carry out this step.

Using your Web Browser

Go to **http://www.diva-gis.org**

Click on the **Download the DIVA-GIS 7.1.7**

NB This is a Beta test version

Save to Disk in a directory where you can find it again.

2. Install Diva-GIS

Double click on the Zip file and extract the files.

This is the file you have just saved, or the **diva-GIS 7.1.7.zip** file in the Software section of your CD.

Double click on the **SETUP.EXE** file and install the software.

Once you have installed the software, you will need to upgrade to the latest version

Copy this file (**diva.exe**) to the **C:\Program Files\DIVA-GIS** directory and click **Yes** when it asks if you wish to replace the file there.

3. Download Tutorial Files and Manual

NB. These files have been downloaded and are on your CD, however, you may wish to download them to get the latest versions.

For this course – please copy the contents of the folder on your CD labeled **"Copy this folder to C_program files_Diva-GIS"** to the DIVA-GIS directory on your computer. This should be in **C:\Program Files\DIVA-GIS**

Using your Web Browser

Go to **http://www.diva-gis.org/documentation**

Download the Manual (available in English and Spanish)
NB this is to Version 5.2

Click on the **data** part of the **Download the tutorial (pdf) and the accompanying data.**

You can also download the **tutorial (pdf)** if you wish, however, you do not need it for this series of exercises.

Save to Disk in a directory where you can find it again.

Double click on the Zip file and extract the files
Save them to the **C:\Program Files\DIVA-GIS\tutor** directory.

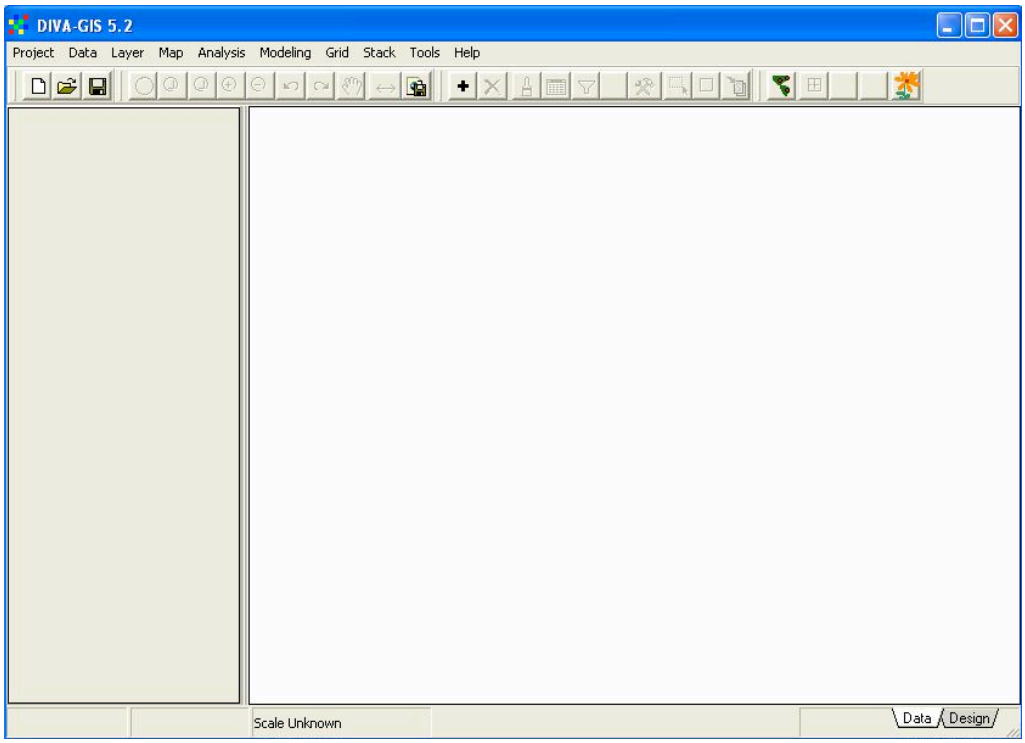
You can also download the **Manual** – (in either English or Spanish).

NB. You can also download various datasets from this site that may be of use to you in your work.

CHECKING POINT DATA AGAINST TERRESTRIAL REGIONS

NB. This can be carried out against any Boundaries and is one way to check if the georeferencing of a species actually places it in the region it is supposed to be, whether it is on the land or out to sea, etc. Routines will also be shown for detecting outliers in Climate Space.

1. Open Diva-GIS



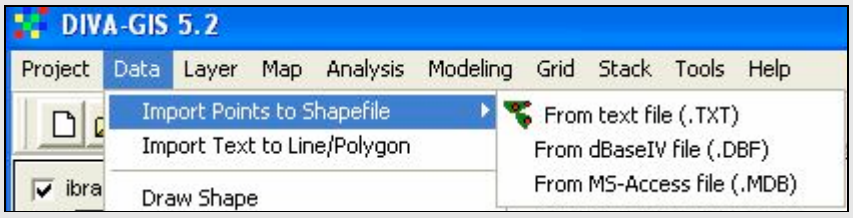
2. Import point data from a .dbf file

<u>In Diva-GIS</u>

Go to **Data**:

Import Points to Shapefile:

From dBaseIV file (.DBF)



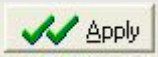
Click on Input File and navigate to C:\Program files\DIVA-GIS\tutor and highlight the dbf file <**bol_wildpot.dbf**>. Click on Open

The **X/Longitude** and **Y/Latitude** fields should be automatically filled in – but use the drop-down menu to change these to "**LONGITUDE**" and "**LATITUDE**" respectively.

This file is a point file of Wild Potato localities in Bolivia.

Give the file an output name. I suggest using Control_C (^C) to copy the information in the Input File field and paste (^V) the information into the Output Field. Change the shape file name to '**bo_wildpot.shp**'. Alternatively type:

'C:\Program Files\DIVA-GIS\tutor\bo_wildpot.shp'

Click the 

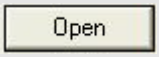
button.

3. Import Polygon data

Click on the  button and navigate to **C:\Program files\DIVA-GIS\tutor**

Highlight '**bo_provinces.shp**'.

This is a polygon file of Bolivia and its Provinces.

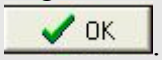
Click on 

and it will load.

Let us label the Provinces as this will be useful to us later.

Left click on the **bo_provinces** in the left hand panel. In the top Diva-GIS Menu click on '**Layer**' and then on '**Add Labels**'

In the Dialogue Box use the down arrow to add **PROVINCE** in the field.

Click on 

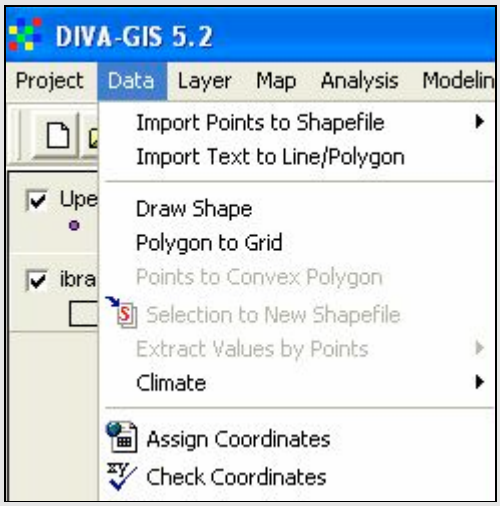
. Close the dialogue box.

4. Check Coordinates

In Diva-GIS

Go to **Data**:

Check Coordinates



Click on '**Shape of Point**' Input File and navigate to the shape file <**bo_wildpot.shp**> that you have just created. Click on Open and it will load.

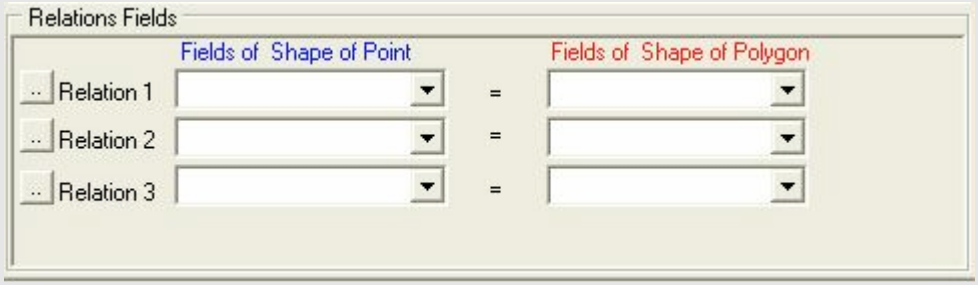
Make sure you set the fields for longitude and latitude:

In this file it is easy as they are labeled LONGITUDE and LATITUDE respectively. In other data sets you may load it may be 'dec_long', 'y', etc.

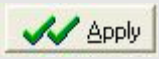
Click on '**Shape of Polygon**' Input File and again navigate to the shape file <**bo_provinces.shp**>. Click on Open and it will load.

NB If you have only one shape file loaded it will already be filled in for you.

Now set the relationships between the two files using the various down arrows.



In Relation 1:	For the '**Field of Shape of Point**' set	'**COUNTRY**'
	For the '**Field of Shape of Polygon**' set	'**COUNTRY**'
In Relation 2	For the '**Field of Shape of Point**' set	'**STATE**'
	For the '**Field of Shape of Polygon**' set	'**DEPARTMENT**'
In Relation 3	For the '**Field of Shape of Point**' set	'**PROVINCE**'
	For the '**Field of Shape of Polygon**' set	'**PROVINCE**'

Click on 

1. Points outside all Polygons
You can now check for points that fall outside all the polygons (i.e. fall outside Bolivia). In this case there are 4 records identified out of a total of 169.

Go to the top Diva-GIS menu and with **bo_wildpot** highlighted, click on the "**Full Extent**" button () you will get a full picture of all the points.

In the **'Check Coordinates'** Dialogue Box, click on the Tab:
'Points outside all Polygons'
you can now highlight a number and click on the 

button and the point will flash on the screen. You can Zoom to the point if you wish (we will do that shortly).

NB. If you have Zoomed or Panned and then wish to return to where you were click on the  (Return) button.

If you click on record "**52**" you can see that the latitude and Longitude coordinates actually places the record in Brazil.

Note: If you check out all the other collections, the Longitude values are all around -66º whereas these records have longitudes of either -46º or -56º. This would indicate that these are data entry errors – but always check and don't just automatically 'correct'.

2. Points that don't match Relations
Now we can check Points where the relations you set earlier don't match.

In the **'Check Coordinates' Dialogue Box**, click on the Tab:
'Points do not match Relations'
you can now highlight a number and click on the 

button and the point will flash on the screen.

Let us now highlight record '**6**'

Click on the Zoom To button, then on the Highlight Button.

You can now see that the highlighted point is in the province Arque just over the border from Quillacollo. You can also see in the **'Check Coordinates'** Dialogue Box that the field labeled 'Point: Province' shows 'Quillacolla' and the 'Polyg: Province' shows 'Arque'.

Obviously the collector has said he was in Quillacolla when he collected the point, but he/she may have crossed the border into Arque without knowing it, or the precision of the point was such that the margin or error could actually place it in either province (the point being only 775m from the border), etc. You can check this by using the ↔ button and measuring on the screen. In the LOCALITY field you can see that a map at 1:250,000 scale was used to determine the coordinates, and the distance given was to a precision of about 0.5 km. Earlier today we discussed how to determine the Maximum error, and for this record it is in the order of 2.265 km, and thus the 775 m falls well within the maximum error.

You need to decide how this needs to be dealt with in the database.

As noted elsewhere in this course – DO NOT just replace the original information, but add new information. For example you may wish to replace the 'Quillacolla' in the Point 'PROVINCE' field with Arque, and place in a field called 'PROVINCE_ORIG' the original province information (i.e. Quillacolla') and give reasons for the change in a REMARKS field. Alternatively you may wish to leave as 'Quillacolla' and make a comment in the Remarks field that the Maximum Error of 2.265 km could place the record in either province.

NB. You can create a file of these 'suspect' records by clicking on the Export button and saving the file into the appropriate area.

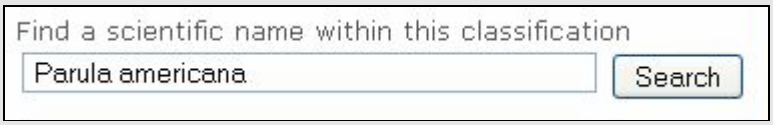
5. Import point data from GBIF into Diva-GIS

In your Web Browser

Go to: **http://www.gbif.org**

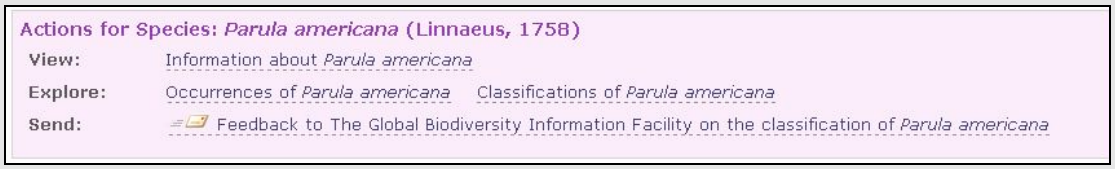
Then: **"VISIT THE GBIF DATA PORTAL"**

Then: **EXPLORE SPECIES**



Use name: '**Parula americana**'

Submit



Select: '**Occurrences of *Parula americana***'

Select: '**Spreadsheet of results**'

Choose Format: **Tab-Delimited**

Tick the radio buttons for:

Catalogue number
Scientific Name
Country
State or Province
County
Longitude
Latitude
Coordinate Precision

Click: on 

NB. I have saved an edited copy to your CD – we can use this to save time

Save File as: '<Parula-americana.txt>'
(somewhere where you can find it again)

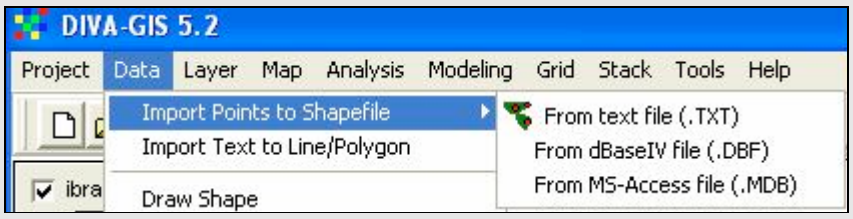
NB You will need to edit this file in EXCEL – to remove records without Lats and Longs, and some fields with long records as these may have problems loading into Diva-GIS. Data extracted from GBIF often has problems and can't just be loaded into Diva-GIS (or most other GIS programs) and needs considerable editing.

In Diva-GIS

Go to **Data**:

Import Points to Shapefile:

From Text file (.TXT)



Click on Input File and navigate to '<**Parula-americana.txt**>'. Click on Open

The **X/Longitude** and **Y/Latitude** files should be automatically filled in – if not – fill them in.

The Output File will be given the name '<**……/Parula-americana.shp**>'

Click the Apply button.

Note. You may have to hit **OK** on the Warning Dialogue Box several times as the field names are too long – a quirk of Diva.

Close out of the dialogue box

6. Import GBIF Data from your own country

In this case we are using the example of data from the Netherlands, but you can do similarly for your own country. [Incidentally – when I tried this with some French data for the same species – I had extreme difficulties in getting it to load, and I am not sure of the reason].

In your Web Browser

Go to **:** http://www.gbif.org

Then: **"VISIT THE GBIF DATA PORTAL"**
Then: **EXPLORE COUNTRIES**

Scroll down to **"Netherlands"**

Click on the **"Species recorded in Netherlands"**
again Click on the **"Species recorded in Netherlands"**

Use the name ***Gryllotalpa gryllotalpa*** for the filter

Click on **"Add Filter"** and then **"Search"**
Click on "**View**" at end of the line

Select: "**Occurrences of Gryllotalpa gryllotalpa**"

Select: "**Download: Spreadsheet of Results**"

Choose Format: **Tab-Delimited**

Tick the radio buttons for:

Catalogue number
Scientific Name
Country
State or Province
County
Longitude
Latitude
Coordinate Precision

Click: on 

NB. I have saved an edited copy to your CD – we can use this to save time

Save File as: '<**Gryllotalpa-gryllotalpa.txt**>'
(somewhere where you can find it again)

NB You will need to edit this file in EXCEL – to remove records without Lats and Longs, and some fields with long records as these may have problems loading into Diva-GIS.

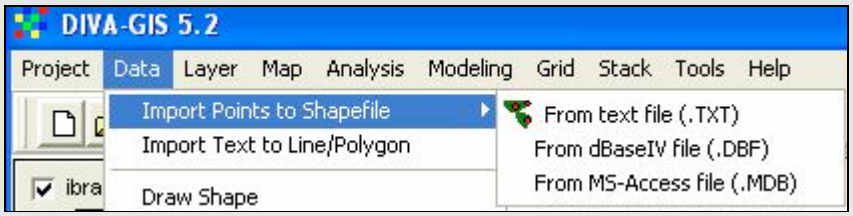
Note that the file includes all records for the species – not only those for the Netherlands – for this exercise I have deleted all but the Netherlands records.

In Diva-GIS

Go to **Data**:

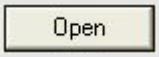
Import Points to Shapefile:

From Text file (.TXT)



Click on 

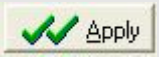
and navigate to '<**Gryllotalpa-gryllotalpa-NL.txt** >'.

Click on 

The **X/Longitude** and **Y/Latitude** files should be automatically filled in – if not – fill them in.

The 

will be given the name '<**……/Gryllotalpa-gryllotalpa-NL.shp**>'

Click the 

button.

Note. You may have to hit **OK** on the Warning Dialogue Box several times as the field names are too long – a quirk of Diva-GIS.

7. Import Polygon data for <Your Country>

In this case we are using the example of data from the Netherlands, but you can do similarly for your own country.

Point your Browser to: **http://www.diva-gis.org/gData**

Select <***Country***> from Drop-down Menu
Select <**Administrative areas**> from the second Drop-down Menu

Click on OK

Click on the Download

Save to your desktop.

Use **UnZip** to extract the files.

2. Import Polygon data

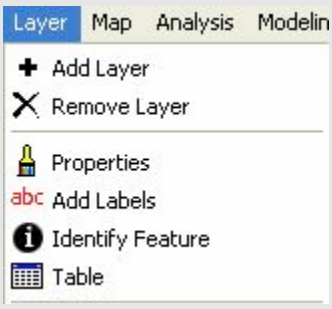
Click on the + button and navigate to wherever you saved the data (above).

Highlight one of the files you just extracted (use the one with the highest number – e.g. '**FRA_adm1.shp**' (for France)
'**BEL_adm1.shp**' (for Belgium)
'**NLD_adm1.shp**' (for The Netherlands)

Click on Open and it will load.

Let us label the Provinces as this will be useful to us later.

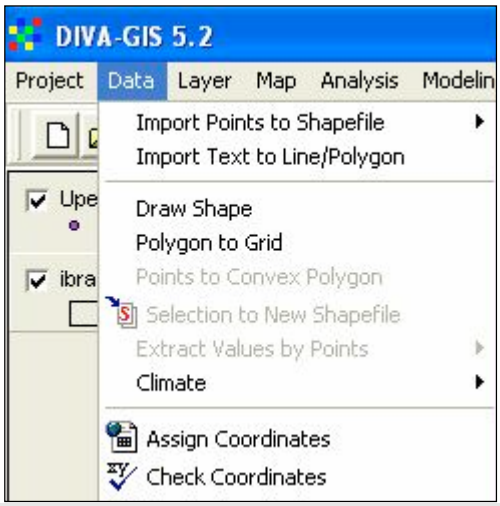
Highlight the **shape file you just loaded** in the left hand panel.
Click on the **Layer** tab and then on '**Add Labels**'



Using the '**Field**' dropdown select the highest numbered '**NAME_1**' field
And then on OK

8a. Check Coordinates

In Diva-GIS
Go to **Data**:
Check Coordinates



Click on '**Shape of Point**' Input File and navigate to the shape file <**Parula-americana.shp**> file that you have just created N[B. Due to the need for lots of editing of the files – I have prepared a file in the Tutor section of your C-Drive – I have taken just the first 1000 records. Click on Open and it will load.

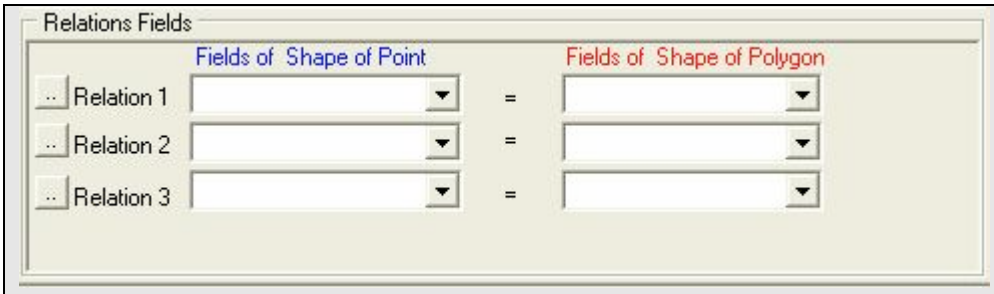
Make sure you set the fields for longitude and latitude:

In this file is easy as they are labeled LONGITUDE and LATITUDE respectively. In other data sets you may load it may be 'dec_long', 'y', etc.

Click on '**Shape of Polygon**' Input File and again navigate to the shape file <**n-america-counties**>. Click on Open and it will load.

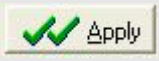
NB If you have only one shape file loaded it will already be filled in for you.

Now set the relationships between the two files using the various down arrows.



In Relation 1: For the '**Field of Shape of Point'** set '**STATEPROVI**'
For the '**Field of Shape of Polygon**' set '**STATE_NAME**'

Note! We could do Country and Country as well as State/State, however in the Point File these are spelt differently between the two layers, so without fixing these we will get a 100% mismatch.

Click on 

Import the <**n-america-counties**> file – by clicking on the  and navigating to it.

1. Points outside all Polygons
You can now check for points that fall outside all the polygons (i.e. fall outside North America). In this case there are 1586 records identified out of a total of 10,098.

Go to the top Diva-GIS menu and click on the "Full Extent" button () you will get a full picture of all the points.

In the **'Check Coordinates'** Dialogue Box, click on the Tab:
'Points outside all Polygons'
you can now highlight a number and click on the 

button and the point will flash on the screen. You can Zoom to the point if you wish (we will do that shortly).

NB. If you have Zoomed or Panned and then wish to return to where you were click on the  (Return) button.

Many of these records are just off the coast and are probably a result of mismatching resolutions. Records **1** and **8396** would appear to be gross errors.

2. Points that don't match Relations
Now we can check Points where the relations you set earlier don't match.

In the **'Check Coordinates'** Dialogue Box, click on the Tab:
'Points do not match Relations'
you can now highlight a number and click on the 

button and the point will flash on the screen. Again some of these are because we don't have States in the polygon file for Central America, however records **9034**, **9043**, **9044**, **9049** and

9050 occur on the border between Mississippi and Louisiana and again may be a result of resolution differences between the layers.

Click on the 

button, Then on the 

Button.

NB. You can create a file of these 'suspect' records by clicking on the 

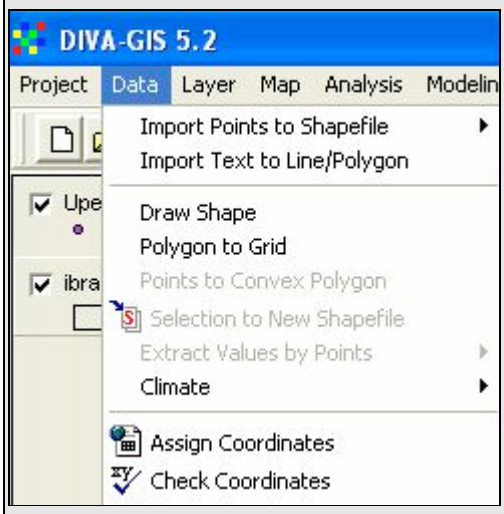
button and saving the file into the appropriate area.

8b. Check coordinates using your own data

In Diva-GIS

Go to Data:

Check Coordinates



Click on '**Shape of Point**' Input File and navigate to the shape file < **Gryllotalpa-gryllotalpa-nl.shp>** file that you have just created. Click on Open and it will load.

NB I have modified the file on your CD – use this one as I have added States into the file.

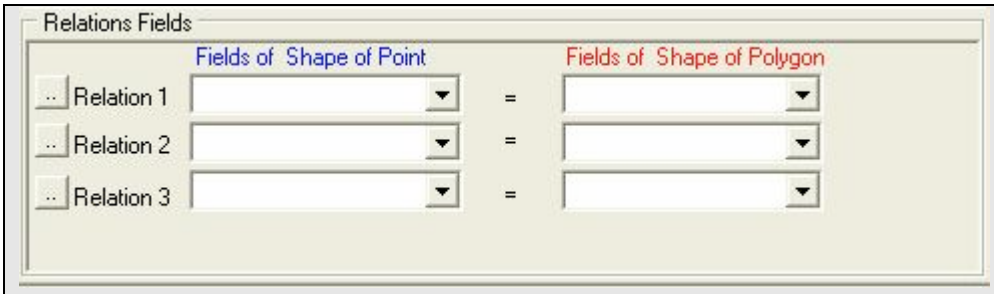
Make sure you set the fields for longitude and latitude:

In this file is is easy as they are labeled LONGITUDE and LATITUDE respectively. In other data sets you may load it may be 'dec_long', 'y', etc.

Click on 'Shape of Polygon' Input File and again navigate to the shape file <**your country**> - e.g. in this example <**NLD1.shp**>. Click on Open and it will load.

NB If you have only one shape file loaded it will already be filled in for you.

Now set the relationships between the two files using the various down arrows. (NB in the example I have here – we can't do this as the data doesn't include counties, etc.)



In Relation 1: For the 'Field of Shape of Point' set 'STATEPROVI'
For the 'Field of Shape of Polygon set 'NAME_1'

Click on Apply

1. Points outside all Polygons
You can now check for points that fall outside all the polygons (i.e. fall outside The Netherlands). In this case there are 125 records identified out of a total of 1077 – many of these are duplicates.

Go to the top Diva-GIS menu and click on the "Theme extent" button ( ) you will get a full picture of all the points.

In the **"Check Coordinates" Dialogue Box**, click on the Tab:
'Points outside all Polygons'
you can now highlight a number and click on the Highlight button and the point will flash on the screen. You can Zoom to the point if you wish (we will do that shortly).

NB. If you have Zoomed or Panned and then wish to return to where you were click on the (Return) button.

2. Points that don't match Relations
Now we can check Points where the relations you set earlier don't match. There are 60 here – most are as a result of resolution of the data.

In the **"Check Coordinates" Dialogue Box**, click on the Tab:
'Points do not match Relations'
you can now highlight a number and click on the Highlight button and the point will flash on the screen.

Click on the Zoom To button, Then on the Highlight Button.

NB. You can create a file of these 'suspect' records by clicking on the Export button and saving the file into the appropriate area.

9. Assign Coordinates

In Diva-GIS

Go to Data:

Assign Coordinates

NB – This used to be done in Diva-GIS, but is now done via BioGeomancer

Go to: **http://bg.berkeley.edu/latest/**

You will need to Register and get a login

How do I format a file for batch georeferencing?

Save your data in a tab-delimited text file.The first row of the must contain column names. BioGeomancer can understands concept names from the DarwinCore 1.4 and its Geospatial Extension.

Specifically, BioGeomancer currently interprets the following fields (the order and case are not relevant):

Locality
HigherGeography
Continent
WaterBody
IslandGroup
Island
Country
StateProvince
County
VerbatimLatitude
VerbatimLongitude
VerbatimCoordinates
VerbatimElevation

Any other fields submitted in the upload are retained, but uninterpreted. Of those in the list above, the Verbatim fields are currently unused in the spatial description, and the WaterBody, IslandGroup, and Island are likely to be heavily under-represented in the gazetteer.

Pay attention to the encoding system used. You should use UTF-8, a character encoding system for the Unicode standard to represent any of the world's scripts. Text encoded in the standard windows/English "Latin-1" will not be corrupted if it contains pure ASCII characters.

I have prepared a file on your CD under "Data"

Once you have logged on – Browse to the file on the CD called "**Accessions without coordinates.txt**"

NB this file must be in UTF-8 – to convert a file to UTF-8 you can open it in Notepad and 'Save as" UTF-8.

Once the project is completed the results will be posted with a download

Click on "**Download**" – call the file "Accessions with coordinates"

Save the file as an XML document

I had a problem opening the file – so if you have a problem

1. Open the file in WordPad
2. Delete the first character on the first line
3. Save the file as a "txt" file
4. You can open this file in EXCEL

You can see that for a number of records (e.g. 4 and 8) there are more than one option, and for some (e.g. records 5 and 6) there were no results.

Load the file into Diva-GIS (NB – you must close the files in the other programs)

You will have to put in the **X/Longitude** and **Y/Latitude** fields

Scroll down to the fields as shown

Load the **world_adm0.shp** file

Note that one record (record 5) is way out – not sure why this has happened – the file shows a blank field, but here it has been filled in with the altitude. Ignore this for now.

Zoom to extend of World_adm and then zoom in around Peru and Bolivia.

Load the files **PER_adm 3.shp** and **BOL_adm3.shp** (in the tutorial area)
Highlight the PER_adm3.shp file and go to **Layer** - **Add Labels** – use the **NAME_3** field to label the records

Highlight the point file and go to **Layer** - **Add Labels** and use the **ID** field to label the records

Highlight the Point file and open the Table by clicking on the  tab.

By clicking on the three records labelled "**4**" – the point on the map will flash. You can Zoom to them. Use the Zoom button to get in closer

You can see in the Table – in the Locality field – that for record 4 it states "**Prov. Cusco**" On the map – you can see that one of the records falls into the Province of "Cusco" – the others in "Maras" and "Santiago" – you can now select one of those three records and you have the latitude and longitude.

> Some of the problems have arisen here due to conflict in the data set between what may be labeled State, Department or Province in Peru. Also spelling differences between Cuzco and Cusco.

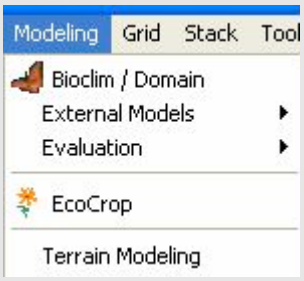
You can now work your way through the other records or work with one of your own datasets

10. Using Environmental Data

In Diva-GIS

In the left-hand panel, tick the box beside the **bo_wildpot**. If it is already ticked – click on the name.

In the Diva-GIS Main menu
Go to **Modeling**
Bioclim/Domain



Note – if Bioclim/Domain is grayed out – make sure you have highlighted the file **bo_wildpot.**

NB. If under Diva Climate data there are no options – we will need point the program to the location of the Climate layers.

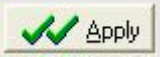
Go to **Tools**
Options

And navigate to the General Data Directory where the Climate data is

NB. We may have to load up the Climate data from the CD – open the world_wc_2_5m.zip file and extract these to the C:/Program Files/Diva-GIS/environ directory.

Check '**Remove Duplicates'** … '**From same grid cell**'

Click on the '**Outliers**' Tab. (We will leave all ticked for now and the 'Min # vars.' at 3).



You will see that a number of lines are in red. These are identified outliers in at least 3 climate parameters.

Click on one of the red lines. A point will flash, and an information box with details of the collection involved. You will note that the three records so identified are the records we previously identified as being outside all polygons.

Now click on the '**Frequency**' Tab.

Tick '**Outliers**'

You will notice that there are two records (top right) that are a darker green than the others.

Click on one of the **Dark Green** points. Again, you will notice that the point on the map will flash, an information box will appear, and a small box that gives an "outlierness' value

If the value is >1, then the record is regarded as being an outlier and thus a suspect record. If <1 then it is regarded as a good record (for this climate parameter). The greater the outlierness value, the greater the likelihood that the record is in error.

Different climate parameters can be examined in a similar way.

This method uses a Reverse Jackknifing methodology (Chapman 1998, 2005), and has shown to be quite accurate with data using both climate and geographic criteria.

This Cumulative Frequency curve also shows records that fall outside a user-defined Percentile area. This is a method commonly adopted by statisticians, however it always identifies records at the top and bottom end whether outliers or not, and commonly identifies 'good' records and misidentifies 'bad' records.

The 1.5 Interquartile range can also be shown and this is another method commonly used for identifying outliers. Various tests using this and the Reverse Jackknifing procedures have shown the Reverse Jackknifing to be more reliable for this type of data.

If you click on the **Envelope** tab

You will see a blue square with green and red dots. The green are records that fall within the 0.025 Percentile (see bottom left) for the two climate layers identified. The red dots are ones that fall outside the percentile area for any combination of climate parameters, the ones outside the blue box are those that fall outside the percentile for this combination of climate parameters.

The map shows yellow points (equivalent to the green dots), and green points (equivalent to the red points outside the blue square)

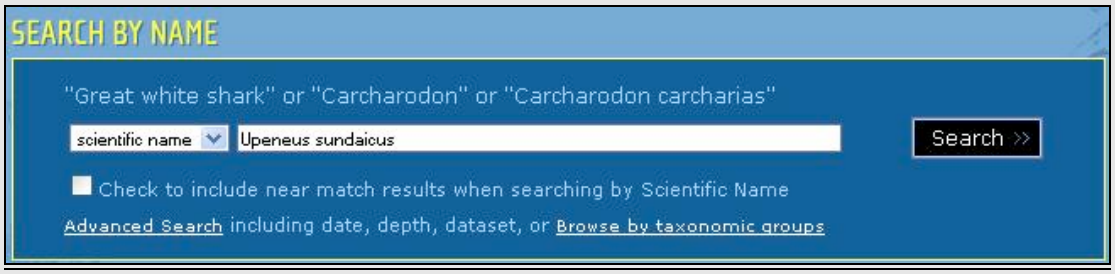
CHECKING POINT DATA AGAINST MARINE REGIONS

NB. This can be carried out against any Boundaries and is one way to check if the georeferencing of a marine species actually places it on the land.

1. Obtain point data from OBIS Portal

Using your Web Browser

Go to **http://www.iobis.org/Welcome.htm**



Search for ***Upeneus sundaicus*** (Ochre-banded goatfish)

Photo Copyright: john_e_randall

Chose: "**View Results as .TXT**"

Save the resultant text file in your working directory
I suggest giving it the name: **Upeneus-sundaicus.txt**

2. Open Diva-GIS

3. Import Shape File

Go to the "**Marine Data**" area on your CD and load the file:

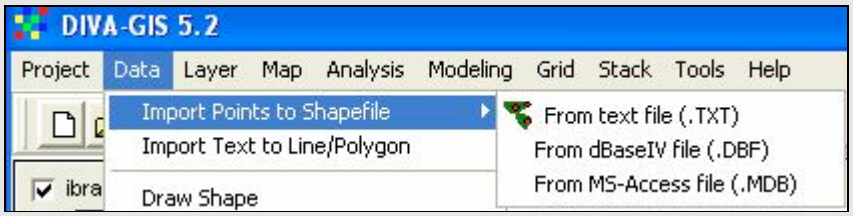
<**World_Seas.shp**>

[Do this by clicking on the ]

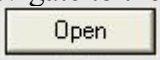
Navigate to "**Marine Data**" area on your CD and click on the **World_Seas.shp** file.

4. Import Point data

In Diva-GIS
Go to Data:
Import Points to Shapefile:
From text file (.TXT)



Click on 

and navigate to the text file <**Upeneus-sundaicus.txt**> you previously saved. Click on 

and it will load.

[**Note** – the 

will now be filled in with the name of the shape file – you can modify this if you wish]

Set the Field Delimiter to "Other |"
NB Make sure you add the "|" character in the required field

When you hit 

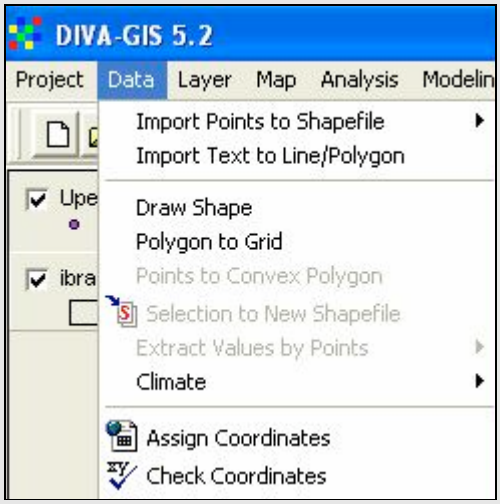
- a series of Warning Dialogue boxes will appear

Just keep clicking OK (about 30 or so times)
[This is a quirk of Diva-GIS where field names are limited to 10 characters]

Close the loading dialogue box by clicking on Close.

5. Check Coordinates

In Diva-GIS
Go to Data:
Check Coordinates

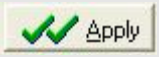


Click on '**Shape of Point**' Input File and navigate to the shape file <**Upeneus-sundaicus.shp>** that you have just created. Click on Open and it will load.

Make sure you set the fields for longitude and latitude:

In the OBIS data this is easy as they are labeled LONGITUDE and LATITUDE respectively. In other data sets you may load it may be 'dec_long', 'y', etc.

Click on '**Shape of Polygon**' Input File and again navigate to the shape file <**World_Seas.shp**>. Click on Open and it will load.



Check the Tab "**Points outside all polygons**"

You can see 12 records listed there out of a total of 119 – these are records NOT in the seas but that fall on the land.

Now click on records "**22**" or "**23**" and "**Zoom To**" the record you can see that these points are situated in the middle of the Cape Your Peninsula of Australia, and thus one would have to regard the georeference information as highly suspect.

Now, lets look at records "**19-21 and 119**". These are all located on the North West Cape in Western Australia. The points are about 7 km (you can use the  measure button to measure the distance once Zoomed in) from the Coast. Now, if you examine the 'X' and 'Y' values in the 'Check Coordinates' dialogue box, you can see that they are only recorded to a precision of 1 degree (about 130 km in this part of the world). The record is thus probably OK, being well within the resolution of the point, but it may be worth checking if the institution has better precision data for these records.

Now click on the Tab "**Points do not match X,Y**"

This uses a form of outlier detection similar to the Reverse Jackknifing using climate data we discussed earlier. Three of the 119 records are listed here.

If you click on record '**4**' you can see that it is off the coast of Mexico; record "**9**" you can see is on the Eastern Coast of Africa and '**10**' is in the middle of the Indian Ocean. As these are along way (geographically) from the other records – you may wish to check their identifications.

NB. You can create a file of these 'suspect' records by clicking on the 

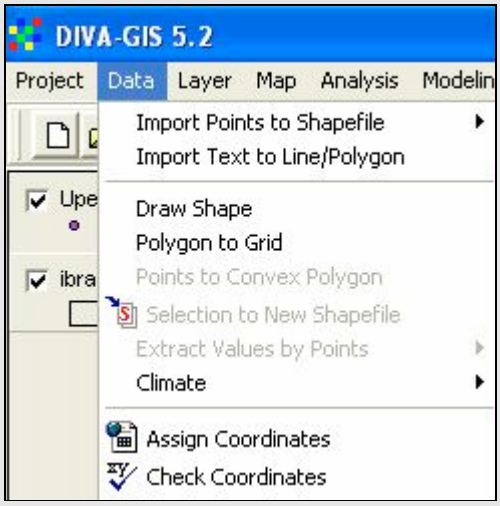
button and saving the file into the appropriate area.

6. Check against EEZ

In Diva-GIS

Go to the + and navigate to the general data area and import the file "**World-EEZ.shp**"

Go to Data:
Check Coordinates



Click on '**Shape of Point**' Input File and navigate to the shape file <**Upeneus-sundaicus.shp**> that you have just created. Click on Open and it will load.

Make sure you set the fields for longitude and latitude:

In the OBIS data this is easy as they are labeled LONGITUDE and LATITUDE respectively. In other data sets you may load it may be 'dec_long', 'y', etc.

Click on '**Shape of Polygon**' Input File and again navigate to the shape file <**world_eez.shp**>. Click on Open and it will load.

This file should also be available from the Diva-GIS Web site (if not there now it should be shortly) and is a layer of the Worlds Exclusive Economic Zones.

NB If you have only one shape file loaded it will already be filled in for you.

Click on Apply

You can now check for points as before, and identify any that fall outside the polygons (i.e. outside all the World EEZs - in this case there are 15 records identified out of a total of 119). One will identify records here that are either on-land, or a long way from land (and thus possibly in deeper water than the other records). This species generally occurs in shallower water areas, so occurrences outside the EEZ are suspect records and should be checked. Not that there is also one record listed under the Tab "**Points that do not match X, Y**"

NB. You can create a file of these 'suspect' records by clicking on the 

button and saving the file into the appropriate area.

USING ON-LINE DATA CLEANING TOOLS

There are a couple of on-line data cleaning tools that have been set up. Generally, these have been set up for use in-house use, however, we will demonstrate one of those to show the possibilities. Be aware that these are maintained outside my control, so I give no guarantee of what is written here will actually happen!

1. CRIA Outlier Detection

In your Web Browser

Go to http://splink.cria.org.br

Change to **English** (unless you are more familiar with Portuguese)

Click on "**data and tools**"

Click on "**spOutlier**"

Let us use the data file we extracted before

Start Excel

<File/Open>: and navigate to: **"Upeneus-sundaicus.txt"**

[Alternatively, you can use one of your own files]

Allow file type: ".txt"

Tick the "Delimited" tick box
<Next>

In 'other': add: **"|"**
<Next>
<Finish>

Highlight Columns "B-E"

<Edit/Delete>

Optionally you can now also Delete all columns from "D onwards" so that we are left with just a file with 'index, 'Latitude' and 'Longitude'.

NB Due to blank lines being in the file extracted from the Internet – sometimes I find it necessary to highlight the fields your require (ID, Latitude, Longitude) Copy and paste into a new file

Save file as EXCEL file

<File/Save As …>

Set File type to "**.xls**".

Give the file the name "**Upeneus-sundaicus.xls**"

Now return to your Web browser:

On the "**import excel file**" and Browse to "**Upeneus-sundaicus.xls**"

When it loads:
In "test" check box – use "**marine**"
Tick "**see map**"
Tick "**show labels for outliers**" click on "**Check**"

You should get the following results:

Result

2	104.3	10.35	marine
3	109.166666666667	12.25	Vietnam
4	122.5	5	marine
5	-103.533333333333	10.8333333333333	marine
6	107.072222222222	20.7791666666667	marine
7	107.073611111111	20.7583333333333	marine
8	107.077777777778	20.7749999999999	marine
9	107.079166666667	20.7861111111111	marine
10	51.2600000063577	11.271666653951	marine
11	61.1499999999999	-10.2666666666666	marine
12	120.989999999999	14.589	Philippines
19	150.25	-21.66	Australia
20	114	-22	Australia
21	114	-22	Australia
22	114	-22	Australia
23	143	-18	Australia
24	143	-18	Australia
117	51.26	11.271667	marine
118	61.15	-10.266667	marine
119	94.9	15.5	marine
120	114	-22	Australia

The red shows **suspect** records and shows if it is regarded as
an outlier in Latitude
an outlier in Longitude, or
on shore (i.e. not 'marine') and shows what country it lands in.

The suspect records are shown on the map with a red triangle and the index number.

2. CRIA Data Quality

Go to **http://splink.cria.org.br/dc**

This system is part of the speciesLink project that links databases across Brazil (mainly in the State of São Paulo).

[**NB.** You must allow your Web browser to pop up additional windows]

Click on english unless you are more comfortable in Portuguese.

In the **Select a Collection:** Go to the **UEC** Database

This is the database of the University of Campinas.
We will use this one because it illustrates the issues we wish to highlight

We will look at the '**locality data'**

We will go through each of these in-turn and examine what it shows.

As you do this – think about how a similar system may be developed for GBIF France, BeBIF or NLBIF.