Genetic Resources and Crop Evolution **46:** 291–296, 1999.

Using GIS to check co-ordinates of genebank accessions

Robert J. Hijmans[1,*], Marianne Schreuder[1], Jorge De la Cruz[1] & Luigi Guarino[2]
[1]*International Potato Center (CIP), Apartado 1558, Lima 12, Peru. E-mail: r.hijmans@cgiar.org;* [2]*IPGRI, Office for the Americas, c/o CIAT, AA 6713, Cali, Colombia (*[*]*Author for correspondence)*

Received 26 February 1998; accepted 5 October 1998

Key words: biological collections, database, documentation, genebank, geographic co-ordinates, GIS, maps

Abstract

The geographic co-ordinates of the locations where germplasm accessions have been collected are usually documented in genebank databases. However, the co-ordinate data are often incomplete and may contain errors. This paper describes procedures to check for errors, to determine the cause of these errors and to assign new co-ordinates, using Geographical Information Systems (GIS). These procedures can assist in improving the quality of genebank databases, and with that, increase the capability for analysis and use of crop genetic diversity.

Introduction

Information on the accessions (i.e., entries, genotypes) conserved in genebanks is usually documented in a database. The completeness and the quality of such databases are important determinants of the usefulness of the germplasm collection to which they refer. A genebank's database may include passport, characterisation and evaluation data. Passport data include a description of the location where the accessions were collected. The location is usually specified by the country and at least one administrative subdivision (e.g., state or department), and by a description of the locality where the accession was found. Often, the location is also specified with geographic co-ordinates (latitude and longitude).

When location data are in co-ordinate form, they can be used in a Geographical Information System (GIS). With a GIS, a number of analyses can be carried out that are of importance for managing and using the germplasm collection, and in planning further collecting. For example, Jones et al. (1997) used genebank databases together with climate surfaces to identify areas where the wild common bean *Phaseolus vulgaris* might appear, but had not actually been recorded. Other activities in which the analysis of georeferenced genebank data can make a considerable contribution include the investigation of the taxonomic structure of collections (Jones et al., 1997); the identification of areas of high diversity (Nabhan, 1990; Frankel et al., 1995; Hijmans, 1997); the targeting of genetic resources for breeding programs (Nabhan, 1995; Guarino et al., 1998); the development of core collections (Guarino et al., 1998); and the selection and design of sites for *in situ* conservation (Guarino et al., 1998).

Unfortunately, in many cases the co-ordinates in the databases are (wholly or partly) missing, imprecise or wrong. As curators of genebanks strive to improve their databases, they face the task of completing and correcting the co-ordinate data (Hazekamp & Frese, 1992). The present paper intends to assist in that activity by describing how the co-ordinates of genebank accessions can be checked and improved, using GIS. We do not discuss the use of GIS in the further analysis of genebank data. The procedures described in this paper are, however, the first steps that are needed to undertake such analyses.

GIS

A GIS is a computer-based tool for managing geographical referenced databases and analysing spatial relationships. There are many different GIS software packages available. In this article we include, as

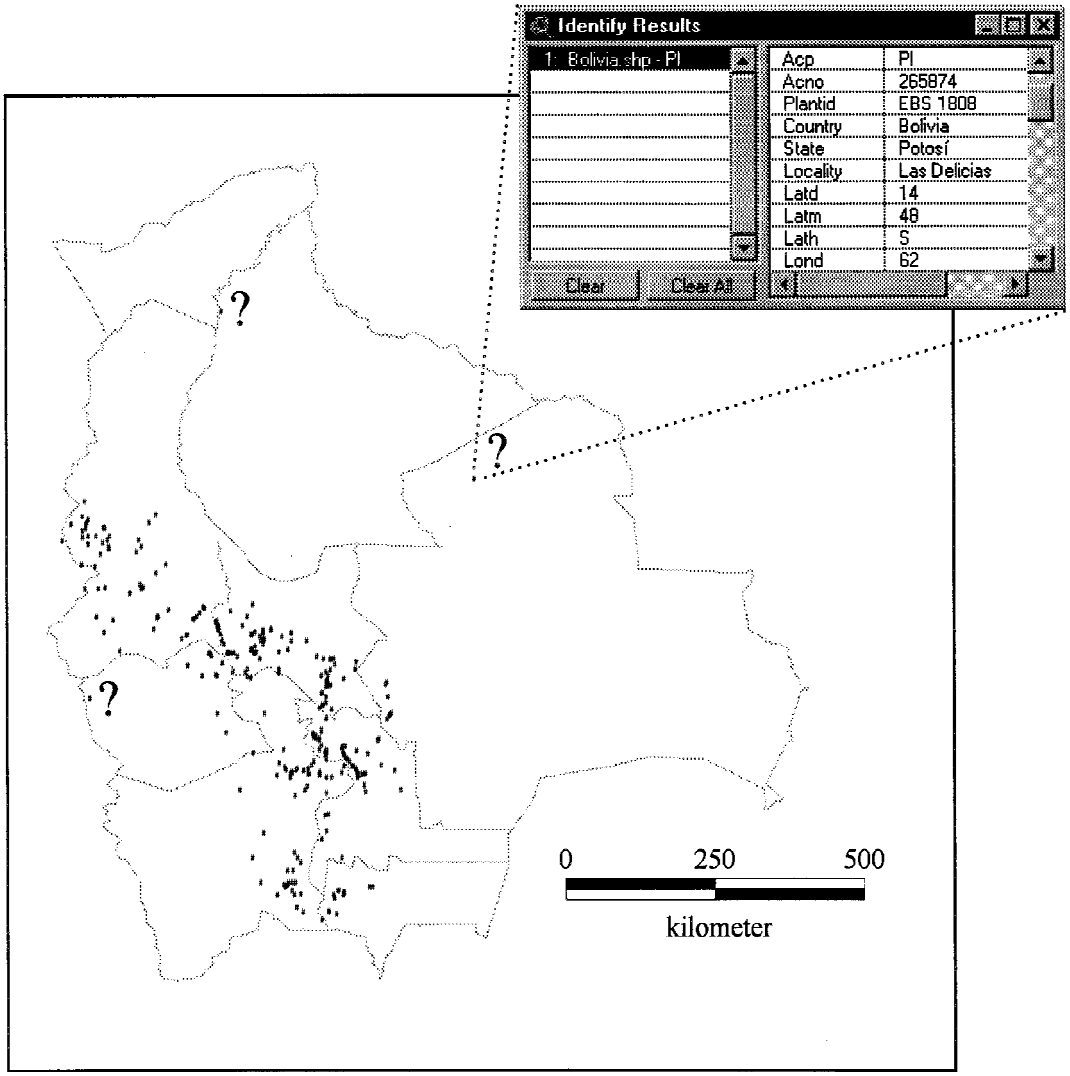

Figure 1. Bolivia, its departments, and locations where wild potatoes were collected (black dots). Probable errors are indicated with a question mark. The ArcView 'Identify Results' window is shown for one of the probable errors.

an example, some commands from the widely used Arc/Info and ArcView software (trademarks of ESRI; http://www.esri.com).

To be able to use genebank data in a GIS, the data need to be in a format that the GIS supports. This is usually done by creating a text file with three numbers on each line: a unique identifier, longitude and latitude. Longitude and latitude should be in decimal degrees. In many genebank databases longitude and latitude are given in degrees (°), minutes (′), and in some cases also in seconds (″), together with a hemisphere (North or South, and East or West). In a GIS, co-ordinates in decimal degrees are needed. Conversion to co-ordinates in decimal degrees is done with the following formula:

$$d^{\circ}m's'' = h * (d + m/60 + s/3600)$$

where h = 1 for the Northern and Eastern hemispheres and −1 for the Southern and Western hemispheres (e.g., 30° 30′ 0″ S = −30.5 and 30° 15′ 55″ N = 30.265). The unique identifier could be the collection number, but because this usually is a combination of alphanumeric and numeric characters, such as 'SVGU 6505', this can easily lead to errors. It may, therefore, be better to use a unique *numeric* identifier that is related to the collection number. The text file is then

imported into the GIS ('Generate' in Arc/Info). With a relational database operation, using the unique identifier, the points can be linked to the passport and other data from the original database.

Finding errors

Errors can be spotted by plotting the collection sites on a map with administrative boundaries. This can lead to the detection of *impossible* locations, that include, e.g., accessions located in an ocean or a lake, and *unlikely* locations, that are far away from all other accessions. In both cases, one should access the database to see if the suspicious locations are really wrong. Perhaps an accession in the ocean was actually found on a small island, that is not shown on the map, or, the accession may truly be isolated for other reasons, such as lack of collection efforts in that area.

This visual inspection method is illustrated in Figure 1, using data on wild potato species of Bolivia from the Intergenebank Potato Database (Huamán et al., 1996). Three suspiciously isolated locations are indicated with a question mark. For one of the locations, the 'Identify Results' window is shown that appears after clicking on the point, using ArcView. On the map, the point is located in the Department of Santa Cruz, according to the co-ordinates in the genebank database. However, the locality description in the same database indicates that it should be located in 'Las Delicias', in the Department of Potosí. The co-ordinates in the database, and thus the location on the map, are therefore likely to be wrong. The visual inspection method only works for the grossest of errors. These errors and other less conspicuous errors can be identified using the capabilities of a GIS to a greater extent.

By simultaneously querying the accessions database and the administrative boundaries database, a new database can be created ('overlay' analysis in GIS language; Arc/Info: 'identity'). For each accession, the new database contains the location names according to the genebank database *and* according to the administrative boundaries database. These names should be the same, and any mismatches reflect errors. This is illustrated in a simple example for an imaginary island that has three provinces, called A, B and C. Six accessions have been collected and stored in a genebank. The co-ordinates of the collection locations, according to the genebank database, have been plotted on a map of the provinces (Figure 2). By querying

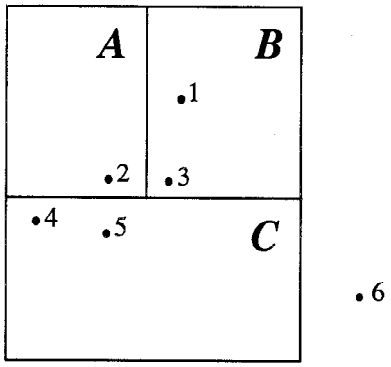

Figure 2. Imaginary island with three provinces, A, B, and C and the location of 6 germplasmaccessions.

Table 1. The location of six germplasm accessions according to the genebank database and according to the administrative boundaries database. Discrepancies between the two databases (bold entries) point at likely errors

Genebank database		Administrative boundaries
Accession	Province	Province
1	B	B
2	**B**	**A**
3	B	B
4	**E**	**C**
5	C	C
6	**C**	

the two databases, Table 1 is generated, pointing at accessions 2, 4 and 6 as possible errors.

Determining the cause of errors

It is easier to spot errors than to determine their causes. The causes of errors include incorrect reading of maps, sometimes caused by duplicate location names and confusion about the co-ordinate system. Typographical mistakes are perhaps the most common cause of errors in the database. Thus, if a name in the genebank database does not correspond with the name in the administrative boundaries database, this could be due to wrong co-ordinates or to wrong names in the genebank database.

In general, it is more likely that co-ordinates are wrong than that the name is wrong. Co-ordinates are often derived from the names in the first place. Moreover, because co-ordinates are, unlike geographical

names, rather abstract entities to most people, it is more likely that errors be made in assigning and digitising them. Some common errors of this type are the switching of latitude and longitude, typing the wrong hemisphere or typing two digits in a number in the wrong order.

If an error is due to a wrong name in one of the databases, this is often caused by differences in spelling or by typographical mistakes. These errors are easy to trace and correct by inspecting alphabetical lists of the names in the databases. Such errors are particularly common if names have been transliterated from other scripts (e.g., Arabic, Chinese or Cyrillic).

It is not always immediately obvious why names do not correspond. Such is the case, for example, when a collection is made near a border of an administrative unit. Apparent errors can be due to lack of precision in administrative boundaries maps, or because explorers did not exactly know in which administrative area they were when they made the collection. In such cases, the co-ordinates may be correct, and the discrepancies may be due to a wrong name. Another common cause of discrepancies in the records is the change in names of administrative units, or the creation of new ones.

When there is doubt about the location of an accession, it is useful to reconstruct the expedition's itinerary. As collection numbers are assigned sequentially, this may help to determine where an accession was collected. It is likely that accession z was collected in between the locations of $z-1$ and $z+1$. This is not always the case, however, as collectors may travel up-and-down a road, so this rule should not be applied blindly.

One can also plot sub-groups of the database, and then compare the doubtful location of an accession in question with the location of the other accessions of the same taxon. For example, if all accessions of a species were found in the Bolivian Amazon, and a doubtful location is in the desert coast of Peru, the doubtful location is likely to be wrong. However, one should be very cautious when applying this procedure. To allow future interpretation of the genebank data, one should avoid downgrading the database by creating artificially reinforced spatial relationships. The exceptions to the general spatial patterns should not be changed/removed just because they are exceptions, but only when they are clearly wrong.

Additional variables, like altitude or vegetation, can also be used to verify the co-ordinate data. For example, when an overlay is made of the collection locations and an altitude map, a new database is generated, analogous to what we showed for administrative boundaries. If there is a big difference between the altitude according to the map and according to the collectors, the co-ordinates may be wrong. However, one should bear in mind that the altitude in the genebank might be wrong too, because of typographical errors or because altitude is often estimated without proper use of instruments. The US Geological Survey has published a high resolution (approximately 1 km^2) altitude grid for the whole world that is available on CD-ROM and on the Internet (http://edcwww.cr.usgs.gov/landdaac/gtopo30/gtopo30.html). This grid can be used to accurately verify the altitude of a given location. Problems with using vegetation maps to check co-ordinates include the fact that many genebank databases do not have much information on vegetation; that the use of different classifications systems complicate comparison; that patchy or mosaic-like vegetation may be missed on small scale maps; and that vegetation patterns change (Hall, 1994).

Assigning co-ordinates

If the co-ordinates of an accession are wrong, or absent, new co-ordinates need to be assigned where possible. Many old accessions generally have very little location data. In these cases, one cannot do much. If a record has a description of the locality where the accession was found, co-ordinates can be determined, using maps. The precision with which this can be done depends on the scale of the maps and on the locality description. Locality descriptions are sometimes very detailed: e.g., 'Franz Tamayo, 10 m west (towards Pelechuco) of bridge crossing over Río Chullumuyo, on horse trail from Pelechuco to Mojos, ca 6 km east of Quiara, 1:250,000-scale map SD 19-10'; but others are very short, like 'near Mojo' or 'about 20 km from Cochabamba'. It is not clear what 20 km from Cochabamba means. In what direction? 20 km from the centre, or from the outskirts of town (and in what year)?

For reasons of precision, large scale (high resolution) maps, i.e., of 1:100,000 and larger, should be used where available. Especially in mountainous areas, it can be difficult to estimate distance on a map, because of the winding roads. In some cases, more precise descriptions can still be found in the field books or expedition reports of the explorers, rather

Table 2. Initial number of records, the number of errors per category (one accession may have more than one error), and the final number of accessions with acceptable co-ordinates, after applying our methods, for a database of wild potato germplasm from Bolivia

Number of records	Number	Percentage
Total	1420	100
Wrong names (province and/or department)	344	24
Wrong co-ordinates	202	14
Missing co-ordinates	483	34
Final number with co-ordinates (after corrections)	1039	73

than on the collecting forms and the databases that were derived from them.

Searching for names on maps can be time-consuming. Gazetteers, or lists of geographic names and their co-ordinates, make searching quicker and easier. There is no comprehensive world gazetteer yet available, but the Times' Atlas of the World (Times Books, 1988) contains an extensive one. The US Board of Geographical Names constructs the Official Standard Names Gazetteer that is available in country volumes. Another reference, containing more than 3 million names from all over the globe, is available on CD-ROM (the GEOName Digital Gazetteer, for more information see http://gdesystems.com/IIS/SlipSheets/GEONAME.html) and directly accessible on the World Wide Web (http://164.214.2.59/gns/html/index.html). Herbaria sometimes develop their own unpublished gazetteers, perhaps in a card catalogue or computer database. Details of the localities mentioned in standard Floras are sometimes published in separate volumes or appendices (Hall, 1994). Historical maps, atlases and gazetteers, and even travel books, can be useful sources of localities if names or boundaries have changed (Maxted et al., 1995).

If new co-ordinates are assigned, the co-ordinate checking procedures described in the previous paragraphs should be applied again. Changes made in the database should be documented so that others may understand the reason for any change that was made. This would be especially useful in the event that any new error is introduced.

A case study

The importance of checking and assigning co-ordinates with GIS is illustrated by a data set from a case study on wild potatoes from Bolivia. Because the database consists of records from 18 expeditions from a period of more than 40 years, many errors could be expected. Applying the procedures described above, we found that more than 50% of the accessions had an error of one kind or another. By carefully studying the sources of the errors, and the location descriptions, most of the errors could be corrected (Table 2). Even recent data, collected with a Global Positioning System, contained some errors, due to typographical mistakes.

Conclusions

We have described methods to verify co-ordinates of germplasm accessions and to assign new co-ordinates where they are absent or when errors are detected. These are important steps in improving the quality of a genebank database, and with that, the usefulness of the germplasm collection.

Using GIS, three kinds of errors can be detected:

i. Accessions in impossible places, like oceans
ii. Accessions in unlikely places, e.g., widely separated from all others, or, at an unlikely altitude
iii. Accessions in the wrong place according to passport data.

The first two kinds of errors are due to wrong co-ordinates and can be detected by *visual inspection* of the data. The third type of error is typically due to wrong co-ordinates and/or wrong names and can be detected using *overlay analysis* methods.

The different kind of errors and the difficulties in detecting and correcting them, highlight the importance of more precise bookkeeping by germplasm collectors and curators. A detailed and unequivocal locality description is crucial. The availability of Global Positioning System (GPS) greatly facilitates taking geographical co-ordinates. However, in our case study, even the data of the accessions that were collected with a GPS had errors, both in the geographical names and in the co-ordinates, caused by typographical mistakes.

Checking and improving the co-ordinates of a germplasm database is tedious and time consuming. And even after applying the procedures described here, the database will likely still contain errors. However, given the dramatic increase of the data quality that is feasible, as shown by the data from the case study, it seems that the effort would be justified for many genebanks.

Acknowledgements

We thank Zosimo Huamán and Meredith Bonierbale for comments on an earlier version of this paper and Steve Kearl and Lorena Corbin for editorial suggestions.

References

Frankel, O.H., A.H.D. Brown & J.J. Burdon, 1995. The Conservation of Plant Biodiversity. Cambridge University Press, Cambridge.

Guarino, L., N. Maxted & M. Sawkins, 1998. Analysis of georeferenced data and the conservation and use of plant genetic resources. Linking Genetics and Geography: Emerging Strategies for Managing Crop Biodiversity. ASA/CSSA/SSSA Annual Meetings, 26–31 October 1997. Anaheim, California.

Guarino, L., 1995. Mapping the ecogeographic distribution of biodiversity. In: L. Guarino, V. Ramanatha Rao & R. Reid (Eds.), Collecting Plant Genetic Diversity, Technical Guidelines, pp. 287–328. CAB International, Wallingford.

Hall, J.B., 1994. Mapping for monographs: Baselines for resource development. In: R.I. Miller (Ed.), Mapping the Diversity of Nature, pp. 21–35. Chapman & Hall, London.

Hazekamp, Th. & L. Frese, 1992. Application of mapping systems for the analysis of the geographical origin of collected material. In: L. Frese (Ed.), International *Beta* Genetic Resources Network. A report on the 2nd International *Beta* Genetic Resources Workshop held at the Institute for Crop Science and Plant Breeding, Braunschweig, Germany, 24–28 June 1991. International Crop Network Series No. 7, pp. 69–70. International Board for Plant Genetic Resources, Rome.

Hijmans, R.J., 1997. Diversity of wild potato species. In: N. Denisov, C. Heberlein, L. Czaran & O. Simonnet (Eds.), GIS in Agricultural Research: Awareness Package. UNEP/DEI/TR.97-9, Case-study 5.

Huamán, Z., R. Hoekstra & J.B. Bamberg, 1996. The Intergenebank Potato Database. In: Abstracts of Conference Papers, Posters and Demonstrations of the 13th Triennial Conference of the EAPR, Veldhoven, The Netherlands, 14–19 July 1996, p. 315.

Jones, P.G., S.E. Beebe, J. Tohme & N.W. Galwey, 1997. The use of geographical information systems in biodiversity exploration and conservation. Biodiversity and Conservation 6: 947–958.

Maxted, N., M.W. van Slageren & J.R. Rihan, 1995. Ecogeographic surveys. In: L. Guarino, V. Ramanatha Rao & R. Reid (Eds.), Collecting Plant Genetic Diversity, Technical Guidelines, pp. 255–285. CAB International, Wallingford.

Nabhan, G.P., 1990. Wild Phaseolus ecogeography in the Sierra Madre Occidental, Mexico: Aeorographic techniques for targeting and conserving species diversity. Systematic and Ecogeographic Studies on Crop Genepools 5. IBPGR, Rome.

Times Books, 1988. The Times' Atlas of the World – Comprehensive edn. Times Books, London.