Validation of georeferenced data

Outlier detection

The detection of outliers in spatial data is an effective way to identify possible errors in the identification and/or georeferencing of primary species occurrence data, thus improving its fitness for use.

Arthur Chapman outlines several methods for geocode checking and validation for primary species occurrence data. In order to inform the development of sound procedures for the systematic validation of spatial data, a preliminary appraisal of the suitability of these methods for validating AVH data was undertaken.

Method	Software available	Comments
Geographic outlier detection	spOutlier-CRIA	This is the easiest and most efficient method for detecting outliers in large batches of data; however, spOutlier is not quite sensitive enough for our needs. We would also need to use a better map of Australia, especially if it will be used to pick up marine localities.
Cluster analysis & Multidimensional Scaling	FloraMap PATN vers 3.01	These methods are too time consuming to use for a large numbers of species, and are too sensitive for large-scale data validation.
Cumulative frequency curves	Diva-GIS (BIOCLIM) ANUCLIM	These methods are too time consuming to use for a large numbers of species.
Climatic envelope	Diva-GIS (BIOCLIM)	
Parameter extremes	ANUCLIM	Not investigated.

It is recommended that a program like spOutlier be used for broad-scale detection of geographic outliers in AVH species distribution data. Rather than outlier detection being carried out at each herbarium, it would be more efficient and more effective to perform this sort of validation at the AVH level, in order to take advantage of the larger data set, and to avoid duplication.

It would be beneficial to incorporate an outlier detection function into the AVH restricted access page, and link it to the proposed annotation system. This would allow restricted access users to perform outlier detection on taxa of interest, and for suspect records to be flagged and herbaria notified. This would also allow the status of potential outliers to be recorded, once they have been checked at the relevant herbarium. Each herbarium could also be assigned the task of checking a subset of the AVH priority taxa, thus distributing the data validation workload, and reducing duplication.

Internal checking

Internal checking of related fields in each database is another useful method for detecting errors in georeferenced data. It would probably be more productive for these checks to be performed by individual institutions, given that there is no benefit to using combined data, and each suspect specimen will need to be physically checked. It may be beneficial to create a priority list and timeline for internal checks (marine/terrestrial, state/territory, botanical region), to align the data validating efforts between herbaria.

Other considerations

- Do georeferencing protocols need to be more consistent between and within herbaria? This is difficult (if not impossible) to achieve in retrospect, so perhaps providing a description of the georeferencing protocols at each institution will help inform data users about the fitness for use of our data for their needs.
- Do different methods of assigning precision values to geocodes create problems when they are transferred to AVH?
- Are cultivated records from any herbaria returned in AVH search results? This would impair the effectiveness of any AVH–level spatial data validation efforts.