Chapter 4. The Dataset.

4.1. Introduction.

This chapter describes the dataset used to address the research question. It describes the history and limitations of the dataset for achieving the research aims. It also assesses the dataset as a source of training data for the analysis and determines which indices from the field area are appropriate for addressing the research question.

4.2. Dataset Documentation.

The dataset has been built from three sources: drill core, elevation and remotely sensed (Landsat Thematic Mapper) data. It is stored in a GIS to enable the co-registration of layers and ensure the analysis is operating on data from correct spatial locations. All components have been projected into the Australian Map Grid (AMG) Zone 54 using the Australian National spheroid (Australian Geodetic Datum 1966).

A point data structure is used to store the drill core dataset. The remainder of the data is stored in raster format because this best represents continuous surfaces (see Section 2.2.2.1).

4.2.1. Drill Core Data.

Point data from the Weipa geological database has been provided by Comalco. It covers an area of 1,100 km² and contains drill core data from 57,642 cores. Of these cores 14,740 points did not intersect bauxite. The remaining 42,902 points contain assay percentages of $\text{Al}_2\text{O}_3$, $\text{Fe}_2\text{O}_3$, $\text{SiO}_2$, $\text{TiO}_2$, depth of overburden, and depth to the ironstone layer. The mineral data are for pisolithic samples from above the ironstone layer.

Drilling was originally done at 1000 feet (304.8 m) intervals on a grid aligned to magnetic north. Later drilling programmes filled in detail using spacings of 500, 250 and 125 feet (152.4, 76.2 and 38.1 m) to better prove the resource.
The total number of samples reduces to 47,422 when converted to a 30m raster grid to match the DEM and Landsat data. This is due to resampling where two points occur on the same grid cell. In such cases, the value of the most northeasterly point is assigned to the cell.

An additional 7,563 points covering the breakaway country have been added to provide more information about the absence of bauxite. This was done interactively using features visible in the Landsat image as breakaway country and coastal plain (see Figure 3.2). The spatial distribution of drill cores and additional points is shown in Figure 4.1.

This dataset was collected before 1980 and bauxite properties have been multiplied by a factor by Comalco before release to protect commercially sensitive information. The factor is not known, but is constant and so will not change the relative variation across the deposit. Consequently it should not impact on the results of this study.

4.2.1.1. Limitations and Known Errors.

The values in the database are composite drill core data, which is the average of two drill holes located one to two metres apart. Geochemistry was measured using Atomic Absorption Spectrometry, which probably biases results towards the mean values for the deposit (Morgan, 1995, p11). The value reported is that for the entire core, no vertical variation was measured. The stated accuracy of samples is +/- 1% abundance (Dunster, 1983).

The lack of information about the vertical variation within the bauxite restricts the analysis in that dimension. However, such extra information would require a further level of complexity in the analysis and thus would introduce more uncertainty. It is also not common for regolith sampling programmes to have detailed downhole sampling for such a large number of samples.

Prior to 1980 the <0.6 mm size fraction of the sample was removed before assaying was performed (Morgan, 1995, p12), because the fine fraction is excluded from the production process. Consequently the values in this dataset are for pisolith geochemistry. While this may introduce some bias relative to the surface indices, particularly vegetation response, there is a linear association...
between bauxite pisolith and matrix chemical composition (M. Morgan, pers. comm., 1996). The geochemical values are therefore representative of the entire bauxite profile.

From 1980, samples from Andoom were screened at 0.3 mm, and 1.7 mm for Weipa. For mine operations the pre-1980 sample values were regressed to this size fraction. This dataset has not been regressed and remains as original values (A. Crossing, pers. comm., 2000).

Drilling was done using a bucket auger. This has the effect of breaking some pisoliths along fracture zones and causes an approximately 10% breakage of pisoliths (Morgan, 1995, p4). This should not have a significant effect on the dataset, as the resulting grades are extremely close to that obtained without breakage (Morgan, 1995, p75). Bucket augers also use finite advances of 15 cm, which limits the precision of the measurements of the bauxite contacts with overburden and ironstone. Commonly the bauxite thickness is underestimated within that 15 cm precision.

Regolith characteristics may vary significantly over short distances (one to ten metres; see Section 3.2.2.4.1), which means the size of the sample plot can have a significant impact on the representation of the process investigated. In the case of the Weipa dataset there is some structural variation induced by tree roots and the morphology of the ironstone layer. However, the geochemical data is measured from pisoliths which should be more resistant than the matrix to geochemical change induced by tree roots. The data should therefore represent the local scale geochemistry without being controlled by the finer scale variation caused by biotic influences.
Figure 4.2

DEM used in the analyses. Contours used for DEM interpolation are superimposed.
4.2.2. Digital Elevation Model.

A grid DEM (Figure 4.2) was interpolated from digitised five metre interval contour data using the ANUDEM algorithm (Hutchinson, 1989), implemented as the TOPOGRID function in Arc/Info version 7. Drainage enforcement was applied and melon hole centres were included as sinks.

A 30 m cell size is used to conform to the pixel size of the Landsat data. Zhang and Montgomery (1994) found a cell size of 10 m best represented hydrological indices. Ideally this study would also use a 10 m resolution, however it is believed the spacing of the elevation contours does not realistically support it.

The elevation contours were manually digitised by the author from two 1:25,000 scale map sheets of the Comalco lease. The contours were interpolated from regularly sampled spot heights sampled by Aerometrex, an aerial survey company, sometime around 1973. The maps contain no reliability information and have elevation contours at a five metre interval with some other basic information. The elevation datum is the Weipa high tide gauge. No data was available for the gap in the north-east corner of the two maps, as it is not part of the Comalco lease. The true quality of this data is unknown but it is assumed the map information is internally consistent and so is adequate for this research. This is the best available source of elevation data covering the study area. The next best source is the Australian 1:100,000 topographic map series which has a 20 m contour interval and so lacks the detail required in this study. A detailed elevation survey was planned by Comalco for 1996 but was not made available for this study.

4.2.2.1. Limitations and Known Errors.

As with all interpolations, the reliability of fine detail away from contour lines, such as near the coast in the Andoom Peninsula, is low for derivative datasets. This is particularly the case for curvatures, as they are second derivatives of the elevation surface (see Section 4.3.3.2.1). The low relief of the study area also makes it difficult for automatic drainage delineation as there is very little information represented by the contours. This should not be a serious problem given that most drainage in the area is vertical except during saturated overland flow conditions during the wet season.
A specific error in the DEM occurs on the east side of the Botchet Swamp melon hole where a drainage line derived from the DEM passes around, instead of through, the melon hole. This is caused by the melon hole boundary having the same elevation as the nearest contour line, causing it to be interpolated as a ridge.

The shorelines on the DEM are mapped differently from those represented on other topographic maps of the Weipa area. This is due to the high tidal range of the area influencing the mapping interpretation. Errors will occur in these coastal areas, and may be accentuated when the topographic data is combined with the remotely sensed data. This should not greatly influence the analysis, as bauxitic areas are some distance from the coastline. Normally there is mangrove vegetation in these regions.

Given these known limitations, and the undocumented quality of the contour data, a cautious approach is adopted with the analyses conducted.

### 4.2.3. Remotely Sensed Data.

A Landsat 5 Thematic Mapper (TM) scene covering most of the Comalco mining lease was acquired from the Australian Geological Survey Organisation (Figure 3.2). The drill core and elevation data extend beyond the eastern edge of the TM scene, but this excludes only 231 drill cores. The scene was captured on 16 June 1988. A subset containing the field site was extracted and rectified using the ER-Mapper image analysis software, version 5 (Figure 3.2). Twenty-seven ground control points (GCPs) were defined using the DEM and ancillary vector datasets. The ancillary datasets included roads and rail tracks mapped by Comalco cartographers, as well as stream lines calculated from the DEM. To allow for positional errors in the DEM a bilinear quadratic rectification with nearest neighbour resampling was used. The average root mean squared error (RMSE) of the GCPs was 0.624 m and the total RMSE was 16.844 m.

Gamma radiometric imagery was also available for the study site, but was not used because the flight lines were spaced at 500 m. Such a spacing is too coarse for this study.

The Landsat image is cloud free and was taken in the middle of the dry season. It contains some band striping typical of Landsat 5 data. Topographic shading is
not evident in the field area. The season of acquisition should provide some
spectral discrimination resulting from vegetation stress through low water
availability during the dry season. Ideally the highly bauxitised areas will be better
drained than the less bauxitised areas and show a different vegetation response.
For example, the melon holes are clearly visible in the image as areas of higher
water availability. The bauxite plateau (light green) is easily differentiated from the
breakaway country (light brown to white; Figure 3.2).

Vegetation disturbance by mining activities are evident (locations B and C in
Figure 3.2). Some areas near major haul roads are also obscured by dust from
the mining operation. Seasonal fire events, many overlapping, also confuse the
vegetation signatures. Fire scars are typical of much of the north of the Australian
continent (eg. location D in Figure 3.2).

There is a scanner distortion in bands 5 and 7 (location E in Figure 3.2) where the
sensor has not adjusted after an abrupt spectral change. This appears to be
caused by the high reflectance of the mine floor at that point.

Roads cleared by bulldozer for drill rig access are aligned magnetic north-south
and east-west at intervals varying from 125 feet to 1000 feet (Section 4.2.1). These introduce a variable source of disturbance. Some are kept clear for mining
purposes while others are left for the vegetation to grow over. Almost the entire
study area has been drilled since mining operations began in 1955, some parts
more than once. Consequently, vegetation at all locations in the point database
is regrowth, and any geobotanical relationships identified by the analysis will
mostly be broad scale vegetation community relationships. There is some mid-
storey information visible in the image near the melon holes, and this should be
detected in the analyses if it has a relationship with the bauxite properties.

This study has only one Landsat TM scene available. Ideally another scene,
possibly from another sensor, would be used to provide seasonal information of
any geobotanical relationships in future work. Higher resolution information may
not provide more information due to the disturbance history of the site and the
generally broad scale variation visible in the image. Multi-date imagery will not be
considered further in this thesis.
4.2.4. Other Data.

Other datasets have been provided by Comalco, and these are collected as part of the ongoing mine management. This data includes haul roads, mining blocks, coastline, and melon holes. In this study these are mainly used for spatial reference, although the melon holes are used in the multivariate analyses (Section 6.3.2).

4.3. The Training Set.

The multivariate analyses used in this study to address the research question are data driven. That is, they make few assumptions about the distribution of the data, although they remain highly parameterised (see Sections 5.3.3, 5.4.3) rather than non-parametric. They are data driven in the sense that they describe relationships among the data, without necessarily imposing deterministic rules on the system. This section describes the dataset as they relate to the analysis.

4.3.1. Quantity and Distribution of Training Data.

Data driven models need large quantities of high quality training data to derive sensible relationships. Where such data is available, care must be taken to ensure it is representative of all the relevant study area conditions, and that any bias is known and can be dealt with.

The spatial density and areal extent of training data used in this study indicate it should cover the full range of environmental conditions in the field area (Figure 4.1). There are few areas of the bauxite plateau that have not been drilled at a spacing of at least 500 feet (152 m; see Section 4.2.1), and the information for non-bauxite locations is equally extensive (Figure 4.1). However, there may be issues of bias to deal with (see Section 5.3.3.4).

4.3.2. Drill Core Data Preparation.

The datasets used in this study have several known errors and biases (see Sections 4.2). Where possible these need to be corrected for, or they must be excluded from the analysis. This should improve the ability of the analysis methods to discover and apply sensible relationships with this dataset.
Areas identified in the Landsat image as mining areas (open pit, vegetation regrowth and haul roads) and as artificial (urban and industrial) are excluded from the analysis when the Landsat data is used, as they will generate false relationships. The most spatially dense drilling is normally done immediately before mining, so excluding mining areas reduces the number of drill cores intersecting bauxite from 47,422 to 14,833, and the number of cores that did not intersect bauxite from 14,740 to 7,261. The number of additional non-bauxite points remains at 7,563.

4.3.3. Independent Variables.

Dependence and independence are used here to describe the relationship between indices. In the context of this study the independent variables are the surface measurable features. They are independent because any relationships found by the analysis are in terms of the dependent variables, which are the regolith properties.

The surface measurable features used in this study must be sourced from the available datasets, in this case the DEM and the Landsat TM scene. The DEM is used to derive geomorphometric indices, while the Landsat data will be used to assess possible vegetation response to regolith variation.

This section considers and tests which indices are useful as independent variables in the context of the analysis. In this case they must be reliable representations of the landscape as it relates to regolith processes.

4.3.3.1. Remotely Sensed Information.

It is not necessary to use all Landsat bands in an analysis because adjacent bands are highly correlated (Table 4.1), which means that little additional information is introduced to the analysis by using all. Also, if all bands are used then the degrees of freedom are decreased, causing reduced confidence in the results. It is therefore assumed that much of the variation in a scene should be adequately accounted for using three of the TM bands. This point will be explored further if the Landsat data contributes a large amount of information to the results.
The thermal (6) and visible blue (1) bands will not be used, as they have little utility for this study. The thermal band has a very coarse spatial resolution (120 m) and there is low vegetation response in thermal wavelengths (Horler et al, 1980; Suresh et al, 1989). Visible blue wavelengths also provide little information about vegetation response. Which of the remaining visible and infrared bands to use in the analysis will be investigated using visual assessment.

<table>
<thead>
<tr>
<th></th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
<th>Band 4</th>
<th>Band 5</th>
<th>Band 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band 2</td>
<td></td>
<td>1.0000</td>
<td>0.4134</td>
<td>-0.5647</td>
<td>-0.3259</td>
<td>-0.1588</td>
</tr>
<tr>
<td>Band 3</td>
<td>1.0000</td>
<td>0.3778</td>
<td>0.6243</td>
<td>0.6974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Band 4</td>
<td>1.0000</td>
<td></td>
<td></td>
<td>0.8489</td>
<td>0.7206</td>
<td></td>
</tr>
<tr>
<td>Band 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
<td>0.9451</td>
</tr>
<tr>
<td>Band 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 4.1. Correlations (Pearson's r) between visible and infra-red bands in the Landsat image (see Figure 3.2).

4.3.3.2. Geomorphometry.

Three types of geomorphometric indices are used in this research. The first type is the general geomorphometric indices devised by Evans (1972; Section 2.2.1.2) which will be calculated using the method of Zevenbergen and Thorne (1987). The second type is the hydrologic indices (Section 2.2.2.3), used to describe drainage conditions of a cell. The third type, multi-scale indices using watersheds, are described in Section 5.2.3.2.

The potential for general geomorphometric and hydrologic indices in addressing the research question is discussed here.

4.3.3.2.1 General Geomorphometric Indices.

Distributions of general geomorphometric indices are shown in Figure 4.3.

Elevation is the base from which all other geomorphometric variables are calculated, but is also useful in its own right. For broad, regional scale studies it
can be used to infer temperature variation by adiabatic relationships, and any resulting changes in regolith properties. For smaller scale studies, such as this one, some correlation should be found with past sea levels and groundwater pressure gradients. The effect of sea level transgressions (Rhodes, 1980; Foster, 1996; Section 3.2.2.4.2) will vary depending on the depth and temporal extent of inundation. Groundwater pressure gradients will, in general, increase as the elevation decreases, although better indications may be obtained using hydrological indices.

The Weipa landscape is extremely flat, and vertical relief across the area of drill core sampling is less than 55 m. Consequently there is limited variation in either plan, profile or total curvature across the dataset, and most parts of the study area have curvatures close to zero (Figure 4.4). In addition, the accuracy of the DEM is not considered high enough to generate valid second derivatives of elevation, particularly when using a nine cell kernel.

Despite the low relief, slope may be useful as an indicator of landscape processes. Slope is a first derivative of the elevation surface, and so will have less associated uncertainty than indices of curvature. There is some variation of slope across the study area, particularly at the boundary of the bauxite plateau and the breakaway country (see Figure 4.4). This should be sufficient to separate some variation in regolith properties in the analysis.

Aspect affects regolith development by controlling the amount of energy from solar radiation available for weathering (see for example Rech et al, 2001), and can also be used as a surrogate for topographic shading in any spectral response measured by remote sensing devices. In general, aspect influences soil formation and vegetation only with slopes greater than 15% (Klingebiel et al, 1988, cited in Cialella et al, 1997). Most slopes at Weipa are much less than 15% so it is doubtful if aspect will provide any direct information about the regolith properties. The low relief also means there is very little spectral effect through topographic shading, and this is evident in the Landsat image (Figure 3.2). Aspect will thus provide little useful information for the analysis and will not be used.

The remainder of the landscape morphology information will be derived from hydrological indices.
Figure 4.3  General geomorphometric indices. There is almost no variation in the curvature indices.
4.3.3.2 Hydrological Indices.

There are many hydrological indices that may be calculated from a DEM. For reasons previously discussed (Section 2.2.2.3), only the non-random flow algorithms will be considered.

The D8 algorithm generates many false parallel flow lines using this DEM (Figure 4.4) and so will not be further considered. This leaves only the flow dispersal fD8 index (Freeman, 1991) and stream tubes (Costa-Cabral and Burges, 1994).

As discussed in Section 2.2.2.2, estimates of aspect will bias to the cardinal and diagonal directions where slope is less than approximately five degrees. Consequently the low slopes at Weipa will direct the stream tubes in the eight directions of the cardinals and diagonals. The results would therefore be similar to the fD8 index (Figure 4.4). The fD8 index was calculated without using a threshold for convergent flow, and appears to generate sensible results where flow is dispersed across flat drainage zones.

The extent of the DEM will have some effect on the accumulation results, as it does not cover the entire catchment. This will have most effect in East Andoom where the catchment areas continue beyond the edge of the DEM. However, the effect of this will mostly be in the drainage lines rather than on hillslopes. Consequently it should have little impact on the analysis, as drainage lines are clipped to an upper threshold before being used in the analysis (see below). In addition, Andoom Creek flows across the top of this edge from north-east to south-west. As a result there are only short hillslope sections affected by this limited extent.
Figure 4.4  Flow accumulation in cells, D8 (top) and fD8 (bottom) algorithms.
Other hydrological information will be gained from using distance from streamlines and distance from melon holes. These are assumed to represent other drainage effects occurring in the Weipa landscape. Streamlines have been calculated using a threshold of 200,000 m², determined using visual assessment. Melon hole boundaries have been provided as part of the extra data provided by Comalco (Section 4.2.4).

4.4. Chapter Summary.

There are some limitations and problems with the dataset used in this study, but it is suitable for analysis. This is because many problems are consistent and thus can be dealt with computationally.

The selection of analysis tools is discussed in the next chapter.