5. SPATIAL DATA INTEGRATION AND ANALYSIS USING A RASTER GIS

5.1 Concepts

Hazard zonation mapping using remote sensing and a geographical information system (GIS) involves the assembly of various spatially co-registered data sets and their comparison with landslide distribution. The composition of the database will depend on what data is available and what is considered likely to be useful. Primary data sets might include: a distribution map of old and recent landslides; geology; structure; site investigation and material test results; soils; topography; drainage; roads; centres of population; etc.

Combining data sets is a common activity in many areas of geoscience. The purpose is generally to represent meaningful relationships between the inputs. The resulting map is merely one possible interpretation of the data, which may need to be modified as more information becomes available. Printed maps are usually for general purpose use; consequently, they may include too much information for the non-specialist but still not provide what is actually important to a user. This can be a problem for those who do not have a geological background, such as planners and regional authorities, whose needs are for simple thematic products. In the case of landslide hazard, such users require a map that clearly indicates zones of potential risk. Recent advances in computer technology and GIS now allow more flexibility in map outputs.

The probability that a landslide will occur at a particular location depends on a number of conditions which may be regarded as (1) controlling factors and (2) triggering events. Controlling factors may be broadly divided into material properties (rock/soil type; *in situ* and bulk strength; etc) and terrain conditions (slope angle; fracturing; cultivation etc), whereas triggering events might include earthquakes, intense rainstorms and possibly new construction/development. If the controlling factors were completely understood and *full ground survey information available* it should be possible to 'predict' where landslides are likely to occur *given a particular triggering event.*

In the case of Fiji the problem is rather different; although much is known about the general causes of landsliding (Chapter 2), little mapping of landslides on a regional basis has been done. In terms of the present study, we are trying to determine whether correlations between past landsliding and other data sets exist that can provide a crude measure of regional hazard probability. It is evident in Fiji that intense rain storms are the main triggering event responsible for major mass movement events such as the Serua Hills landslides of 1980. What is uncertain is the extent to which regional variations in ground and geological conditions influence the location and/or severity of landsliding (given that such rainfall events could occur anywhere).

The approach taken here is empirical and the results therefore only indicative; the method is at best semi-quantitative and at this stage provisional. The potential benefits are that the method is relatively cheap and rapid as compared to a full ground survey (which may, in fact, be unachievable over large areas). Using a GIS, the geologist can test for spatial relationships between landslides and other potential influencing factors, can quantify their importance, and can mathematically combine them to produce a hazard 'probability map'. The objective of the pilot study is to develop an approach to modelling landslide occurrences

based on a minimum of existing data. If encouraging results are obtained, additional, relevant data can be added to the database to improve and further develop the technique.

The fundamental indicator, obtainable from remote sensing, is a map showing the distribution of both old and recent landslides. This is then used as a basis for deciding where further landslides are likely to occur. Even on its own, the landslide inventory map may be regarded as a crude hazard map, but in order to try to understand what factors influence landsliding, and thus to be able to rank the degree of hazard, the relationship of different variables to landsliding is tested in turn. Significant variables can be thought of as 'controlling factors' for the purposes of the analysis *even though the **manner** of their relationship to landsliding may not be known*. Once these empirical relationships are established, the geologist must decide, and weight, the importance of each variable, and combine the weights to provide an overall probability estimate across the entire area. However, it should be remembered that the available database layers may not be sufficient to completely model the true situation. It is also possible that the apparent 'controls' modelled using this approach may be relatively insignificant compared to the overriding importance of triggering events (i.e. intense rainfall).

The resulting hazard map produced in this way will be similar, but by no means identical, to the landslide distribution map used to develop the model. In the hazard map, the zones will not be restricted to the precise areas where landslides are recorded but will extend beyond these limits based on the correlations found in the data. The use of a raster GIS enables these operations to be carried out quickly, and provides flexibility that allows inputs to be easily varied.

5.2 Analysis

The variables comprising the Fiji database were described in Section 4.4. Two layers form the basis of the analysis. The first is a Boolean mask formed by combining all the landslide polygon information irrespective of age and interpretation criteria. The second is the distribution of landslide 'initiation points' for recent landslides. It is against these control layers that the other variables are examined. The variables are summarized in Table 5.1.

Table 5.1. Variables used in the landslide analysis for Fiji

Variable	Number of classes
Landslide polygons (total)	1
Landslide start points (total)	1
Elevation (250 m classes)	17
Slope (5° classes)	16
Aspect (45° sectors)	8
Geology	22
Forestry	8
Soils	43

The GIS used for the analysis was IDRISI. This was chosen in preference to ILWIS since it seems likely that this system, in its soon-to-be-released Windows version, will become more widely used in the south west Pacific region.

GIS analysis relies on examining the spatial relationships between interpreted landslides and variables, individually and in combination. To begin with, each class within each variable (e.g. each lithology of the geology layer) was cross tabulated with the map of landslide polygons to determine the number of landslide pixels and non-landslide pixels comprising that class: e.g.

$$\frac{\text{Number of landslide pixels in lithology Class 12*}}{\text{Total number of pixels comprising Class 12}} = \frac{171}{602} = 0.284$$

(* where Class 12 is equivalent to the Basal Conglomerate of Navua Mudstone)

This provides information about the variables and classes which have an association with the presence of landslides. For example, different rock types may have a higher or lower tendency to slip due to cohesive strength related to composition, grain size, degree of fracturing etc. In the above example, 28.4% of the ground area mapped as Basal Conglomerate corresponds to landslides. The same analysis was carried out for all classes of all variables for the south east Viti Levu test area. The results of the cross tabulations are presented in Appendix 1 (Tables A1.1 to A1.6). They allow important deductions to be made regarding the role and possible significance of the variables, as described below:

1. **Geology**: As might be expected, different rock types provided different responses (Appendix 1: Table A1.1). Formations showing the highest incidence of landslides are: Lokalevu Keratophyre, Basal Conglomerate (Navua Mudstone), Nubuonaboto Volcanic Conglomerate, Serua Conglomerate, Navutulevu Polymict Conglomerate, and the Tawavatu Tuff.

2. **Slope angle**: The relationship between landsliding and slope is complicated (Appendix 1: Table A1.2). Because most slopes in this region area are gentle, the majority of landslides occur on slopes of 25° or less. In these areas, the highest risk of landslides (21.1%) would appear to be on slopes of 10-20°. Although there are relatively few very steep slopes, there is an increased incidence associated with angles greater than 45°.

3. **Slope aspect**: There appears to be only a slight relationship with slope aspect, with a tendency for landsliding to occur on northerly or northeast facing slopes (Appendix 1: Table A1.3).

4. **Soils**: There is an apparent relationship with a number of soil categories (Appendix 1: Table A1.4). However, it is difficult to interpret the significance of these results in the context of the MPI classification. Further examination of these results is warranted to see if any correspondence exists between the observed weightings and the fivefold division suggested in section 2.3.1.

5. **Elevation**: The majority of landslides occur at elevations of between 50 m and 350 m (Appendix 1: Table A1.5). The highest risk of landsliding (23.5%) would appear to correspond to the elevation range 250-300 m.

6. **Forestry**: The strongest correlation is between landsliding and hardwood, followed by coconut plantation and non-forest (Appendix 1: Table A1.6).

7. **Lineaments**: Although there is a suggestion in some areas that regional fractures exert a controlling influence on the location of landslides, lack of time prevented this data set from being compiled and digitised. It is recommended that in any follow-up study such information be added, and included in the model. Intuitively, one would expect there to be an association between lineaments and landslides since lineaments mainly represent faults or lines of weakness along which movement could occur, or simply zones of more broken ground. One way of testing the significance of lineaments is to compare the relationship of lineaments to landslide initiation points. To do this requires a cross tabulation to be carried out of the lineament buffer map against landslide start points. This is done by comparing the number of landslide initiation points that fall within each distance zone against the 'expected' number, taken as the average incidence of start points over the land area as a whole.

8. **Catchments & landslide density**: The density of landslides within catchments can be included as a variable in the model to improve the correspondence between observed distributions and the model predictions. In the present study, time limitations prevented the catchment data being digitised. However, landslide density is in any case not strictly a 'variable' in the sense that cannot be used to extend the model beyond the area of landslide photointerpretation. In the present case, landslide density was calculated as a moving average using image processing software external to the GIS. Rather than incorporate it in the actual analysis, the map was used as a means of visual comparison for the goodness of fit of the final model.

Having considered the relative significance of each class of each variable, the analysis was continued by calculating whether a class contained more or fewer landslides than was typical for the area as a whole. To do this, the earlier calculated class percentage values were normalised by dividing by the *regional average* incidence of landslides, calculated as:

$$\frac{\Sigma \text{ landslide pixels}}{\Sigma \text{ pixels comprising study area}} = \frac{42498}{285362} = 14.89\%$$

Taking the earlier example of lithology Class 12 and dividing by the regional average, the result is 28.40/14.89 = 1.91. This means that the incidence of landsliding in the Navua Mudstone Basal Conglomerate is 1.91 times greater than for south east of Viti Levu as a whole. Considered as a measure of prediction, one could say that landsliding is almost twice as likely to occur within this rock type than on average over the area.

The same calculation was repeated for every class of every variable, and weightings derived (Appendix 1: Tables A1.1 to A1.6). To avoid the use of decimals, weights were multiplied by 10 and rounded. A value of 9 or less indicated that the class had a lower that average incidence of landslides, a value of 10 an average incidence, and values of 11 and above a higher than average incidence. The *overall* (or average) importance of a variable was judged by how far the class weights diverged from 10. If all classes were close to 10, then the effect of the variable was neutral, whereas if some classes had a much higher value than 10, then the variable was likely to be significant. The larger the weight, the greater was the chance of landsliding within the class.

Quantifying the significance of a variable as a predictor of landsliding is not straightforward since this can be considered in different ways. Two possible measures of performance are provided in Table 5.2. The first of these - here termed '**accountability**' - calculates the percentage of the total landslide population accounted for by each variable. It is computed for each variable as:

$$\frac{\Sigma \text{ landslide pixels in classes having a weighting } \geq 11}{\Sigma \text{ landslide pixels over the entire study area.}}$$

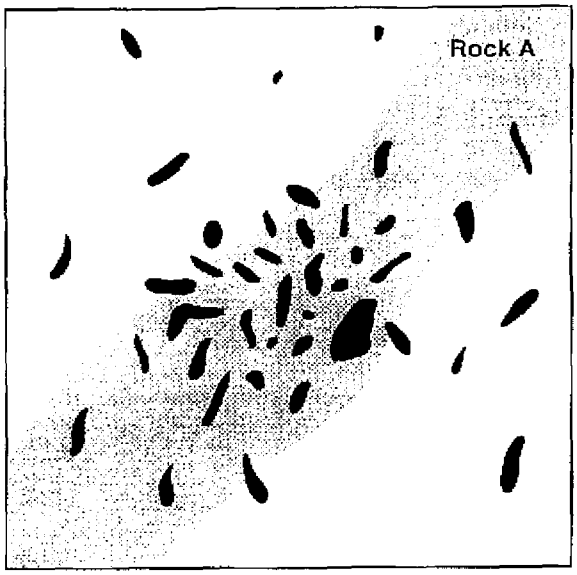
Another way to regard performance is in terms of '**reliability**' (i.e the chance, or probability, that a pixel in a class will be a landslide), calculated as the percentage area of a variable corresponding to landslides. It is computed for each variable as:

$$\frac{\Sigma \text{ landslide pixels in classes having a weighting } \geq 11}{\Sigma \text{ landslide \& non-landslide pixels in the same classes}}$$

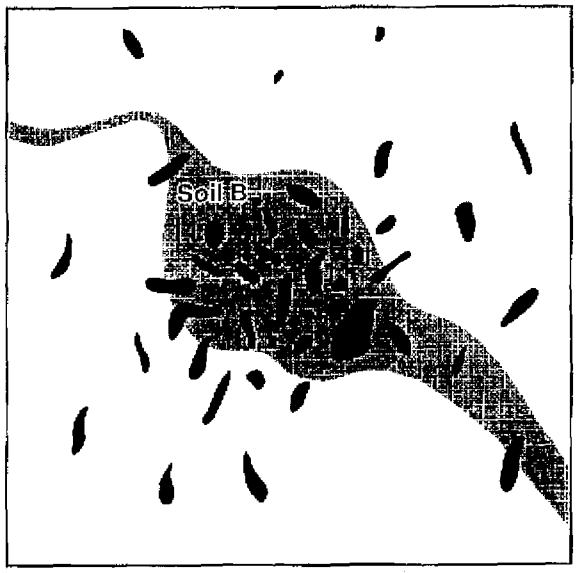
(Note: in both performance measures, only classes ≥ 11 (i.e. showing above average landslide incidence) are considered).

It can be seen from Table 5.2 that the two indicators do not provide the same information, nor is it obvious which is the better measure. For example, whereas the soils variable is the most reliable indicator in that 23.3% of the variable corresponds to landslides (1st ranked), it accounts for only 64.5% of all landslides in the area (4th ranked). On the other hand, slope angle accounts for more of the total landslide population (77.3%) than any other variable, although less of this variable corresponds to landslides (19.0% - 4th ranked).

The reason why the two performance indicators give different results can be explained by a simple illustration involving 2 variables each comprising one class. In Figure 5.1A, Rock A accounts for 80% of all landslides but only 20% of the unit corresponds to landslides. In Figure 5.1B, Soil B accounts for 60% of total landslides, but here 25% of the variable is landslide. So, whereas Rock A accounts for more of the total landslides population, Soil B more reliably indicates the likelihood of a landslide. The reason is that within each category the landslides are not evenly distributed but form clusters. Nevertheless, in their own way,



5.1A Rock A accounts for 80% of the landslides in the region but only 20% of Rock A corresponds to landslides



5.1B Soil B accounts for 60% of the landslides in the region but only 25% of Soil B corresponds to landslides

5.1C The combined category of 'Rock A + Soil B' now accounts for 55% of the region's landslides and 45% of this category corresponds to landslides

Figure 5.1 Hypothetical situation where landsliding is controlled by 2 independent variables, Rock A and Soil B.

each of these measures provides a good predictor of landslides, although the reliability factor may be overall the more important.

Table 5.2. Measures of performance of variables in predicting occurrence of landslides in SE Viti Levu (for explanation see text)

Variable	**'Reliability'** **% of variable corresponding to landslides**	**'Accountability'** **% of landslides accounted for by variable**	**Overall ranking**
	Calculated as:- Σ landslide pixels in all classes having weightings ≥11 divided by Σ pixels in same classes	**Calculated as:- Σ landslide pixels in all classes having weightings ≥11 divided by Σ all landslide pixels**	
slope angle	19.0% (4)	77.3% (1)	2
geology	21.8% (2)	76.7% (2)	1
elevation	19.6% (3)	71.0% (3)	4
soils	23.3% (1)	64.5% (4)	3
aspect	16.7% (6)	46.2% (5)	6
forest cover	17.6% (5)	41.0% (6)	5

Having assessed the variables separately, it is necessary to consider their combined relationships to landsliding. Logically, if two variables *individually* relate (in some undefined way) to landsliding, then the two *taken together* should provide a still better indicator. Such combinations may, for example, help explain, and model, particular spatial patterns evident in landslide distributions resulting from multivariate interactions. This is further illustrated in Figure 5.1C. Here, the combination of Rock A-plus-Soil B (occupying the shaded ground) accounts for 55% of the landslides, but the reliability factor of the combined variable is now 45%. Thus, the use of two variables improves some aspects of prediction at the expense of others: fewer landslides are accounted for than by Rock A on its own, but the higher risk areas are predicted much more reliably than for either Rock A or Soil B alone. It might be concluded in this example that, whereas areas where Rock A and Soil B occur together are particularly prone to landsliding, such high risk conditions do not apply over most of the region. Thus, it appears that a more *reliable* model can also be one that *accounts* for less of the total landslide population.

Based on the above logic, the next step was to combine variables to see if the reliability and/or accountability could be improved. Before this could be done, however, it was necessary to test whether the variables were truly independent. Correlated variables contain redundant information and should not be included together in the analysis without risking duplication and biased weighting. This was achieved by pairwise cross tabulation between each of the variables. In most cases, only limited correlation was found, and the variables were accepted as being independent.

Multivariate analysis involved calculating successive combinations of the variables and comparing their performance, both visually and statistically. In theory, the two variables individually showing the strongest relationship to landsliding should be combined first and lower ranked variables sequentially added. However, as noted above, there is no single measure of performance on which to base this sequence. Consequently, the rankings were decided somewhat subjectively by taking account of both the reliability and accountability values. The final column in Table 5.2 above shows the ranking assigned to the 6 variables.

The combinations of variables ('models') were produced on the GIS by first recasting each class of each variable in terms of the weightings previously calculated, and then adding these weights pixel by pixel across the area. (N.B. all classes were included in the analysis regardless of their individual weights). This resulted in a new combined (or *logical*) weight for every pixel. In order that different models could be more easily compared, these new weights were divided by the number of variables used in the combination (i.e by 2 in the case of the 2-variable combination).

To test whether the first combination (geology + slope angle) provided a better or worse model of landslide distribution, the logical weights were considered as 'classes' of the *new variable* 'geology-plus-slope' and tested in the same way as the single variables; that is, the number of landslide pixels in each new 'class' was divided by the total pixels making up that class, and divided by the regional average. For most classes it was observed that the recalculated weights were higher than the logical weights, suggesting that more of the variability was being accounted for in that class than might be expected by a simple linear addition (Table 5.3). It was concluded that higher-than-expected weightings were the result of variables combining in a multiplicative manner due to interaction (synergy) between them (i.e. certain conditions in each variable supported each other). Lower-than-expected values in some classes (particularly classes ≤ 10 - not included in Table 5.3) suggested that interactions in certain cases had the effect of increasing ground stability. This concept of logical and recalculated weights is illustrated in Figure 5.2.

The overall significance of the first model as a predictor of landsliding was assessed in terms of each of the two measures of performance discussed earlier. Compared with the individual variables, the combination of geology plus slope angle showed a higher value for reliability (23.7%) at the expense of reduced accountability (66.7%) (Model 2 in Table 5.4: compare with the separate values for geology and slope angle in Table 5.2). Performance measures for the combinations are discussed in more detail in section 5.3 below.

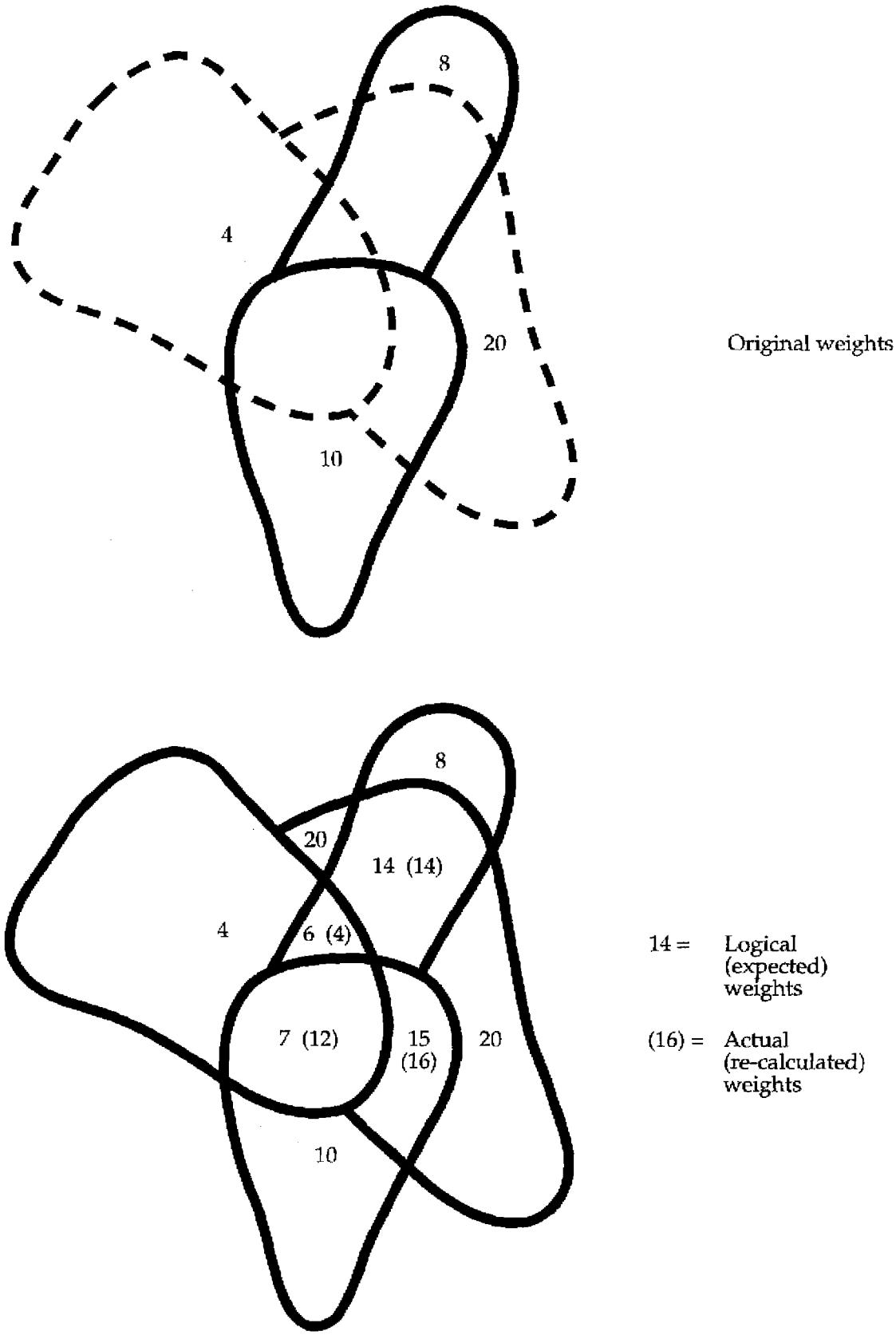


Figure 5.2 The upper diagrams show 2 variables each consisting of 2 weighted classes. In the lower diagram, summation of the 2 variables produces new 'logical weights' for the new polygon areas. However, the 're-calculated weights' (in brackets) do not necessarily match the logical weights.

Table 5.3 Comparison of logical and recalculated weightings for the Fiji hazard models

Logical weightings of class	Recalculated weightings within combinations (models)				
	2	3	4	5	6
11	10	11	10	11	12
12	13	15	14	14	16
13	15	14	15	19	22
14	17	19	21	22	21
15	24	20	23	22	33
16	20	22	22	33	39
17		25	32	34	
18	24	35	36		
19		26			
20	18	36			

The analysis was continued by adding a third variable (soils) to the first two, summing the weights pixel by pixel and dividing by 3. Again, new weightings were calculated, these weights compared with the expected (logical) weights, and reliability and accountability computed. This procedure was repeated for each further combination of variables with elevation, forest cover and aspect respectively being added in turn. The results are given in Appendix 1 (Tables A1.7 to A1.11).

In order to present the results in simple map form, each model was sub-divided into a few levels (representing high, medium and low hazard) and smoothed. The sub-division was based on the final *recalculated* weightings: high was taken as all classes ≥ 21, medium 11-20, and low ≤ 10. In order to improve the appearance of the final map, all unnecessary, unrealistic and complicating detail was removed by 'smoothing' the map. This was achieved by replacing the central value of a moving 3 x 3 window by the mode of the window (the most commonly occurring pixel value within the window), the procedure being applied repeatedly until broad, consistent classes resulted.

5.3 Discussion and evaluation

Table 5.4 compares reliability and accountability values for the successive combinations of variables.

Table 5.4. Reliability (R) and accountability (A) measures for the Fiji hazard models (values in per cent).

Model	2		3		4		5		6	
Re-calculated weights	R	A	R	A	R	A	R	A	R	A
$\Sigma \leq 10$	*8.5*	33.2	*6.7*	23.5	*7.4*	31.6	*6.0*	22.5	*6.3*	24.4
Σ 11-20	*22.9*	60.6	*22.4*	61.9	*20.8*	33.0	*21.6*	56.6	*21.4*	46.5
$\Sigma \geq 21$	*36.4*	6.1	*35.3*	14.7	*32.8*	35.4	*33.5*	20.8	*32.4*	29.0
Totals for ≥ 11	***23.7***	66.7	***24.1***	76.6	***25.7***	65.4	***23.9***	77.4	***24.6***	75.5

[Where: $\Sigma \leq 10$ is the sum of pixels in classes with a recalculated weighting less than or equal to 10. 2 = geology+slope; 3 = geology+slope+soil; 4 = geology+slope+soil+elevation; 5 = geology+slope+soil+elevation+aspect; 6 = geology+slope+soil+elevation+aspect+forest cover]

The *totals* for reliability in Table 5.4 are comparable to the single (average) reliability values used previously in discussing the individual variables (Table 5.2). These values show little variation and seem to imply that Models 2 to 6 are very similar. Based on this information alone, one might conclude that successive models show no improvement. That this is not the case can be seen by comparing 5.3A to 5.3E (Models 2 to 6). As described earlier, these plots are divided into 3 simple levels of hazard (low, medium and high). The plots show discrete zones of higher and lower hazard over the study area. From Model 2 to Model 6, the shape of these zones exhibits a decreasing dependency on the boundaries of the individual input layers (compare with Figures 4.6 to 4.11). For example, the influence of geology (Figure 4.6) in defining the zones is much less evident in the higher combination models than in the lower ones. Although this improvement is apparent from visual inspection, the reliability **totals** by themselves fail to make this distinction.

For a more quantitative assessment, it is necessary to consider how the reliability and accountability values vary from model to model. The statistical results in Table 5.4 relate to the same threefold sub-division of levels used to plot the models. By thus grouping classes, any undue importance attaching to individual classes is avoided (for example, classes with high weights but composed of very few landslide pixels). In the following paragraph, the changes evident in the plots are described on the basis of the statistics in Table 5.4.

Model 2 (Figure 5.3A) shows only a small region (A = 6.1%) of high hazard probability (R = 36.4%), and the total of landslides accounted for in the medium-plus-high zones is 66.7%. In Model 3 (Figure 5.3B), the size of the high reliability (R = 35.3%) area has increased (A = 14.7%), as has the total accountability for the medium-plus-high zones (A = 76.6%). In Model 4 (Figure 5.3C), there is a large increase in the size of the high reliability zone (A = 35.4%). Model 5 (Figure 5.3D) shows a reduction in the size of the high hazard zone (A = 20.8%), but a significant increase in the medium hazard zone

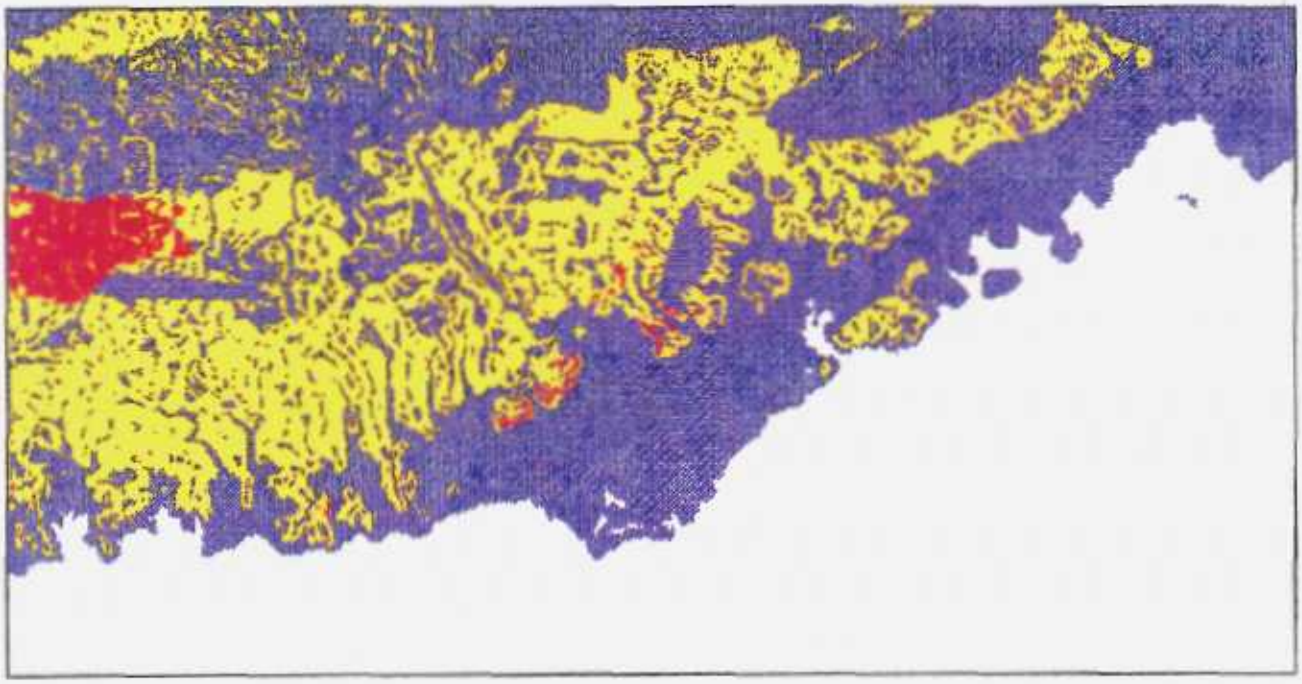


Figure 5.3A Landslide hazard map based on 'geology + slope' (Model 2).

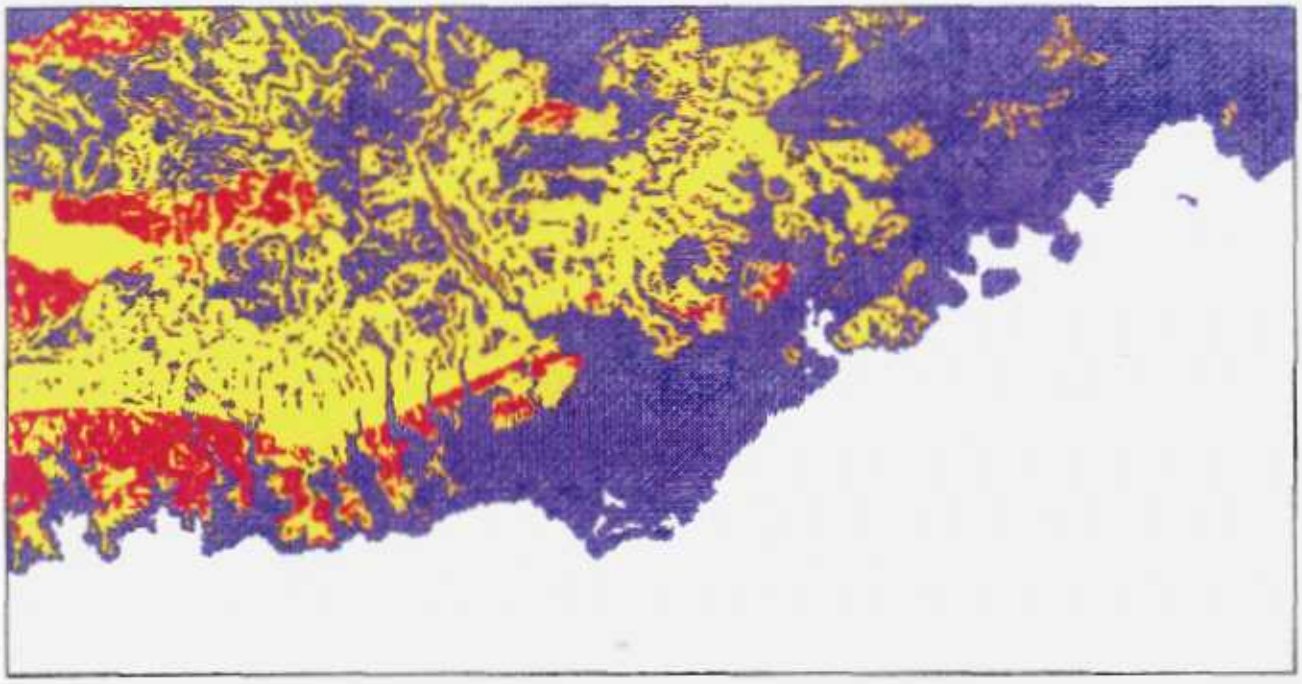


Figure 5.3B Landslide hazard map based on 'geology + slope + soils' (Model 3).

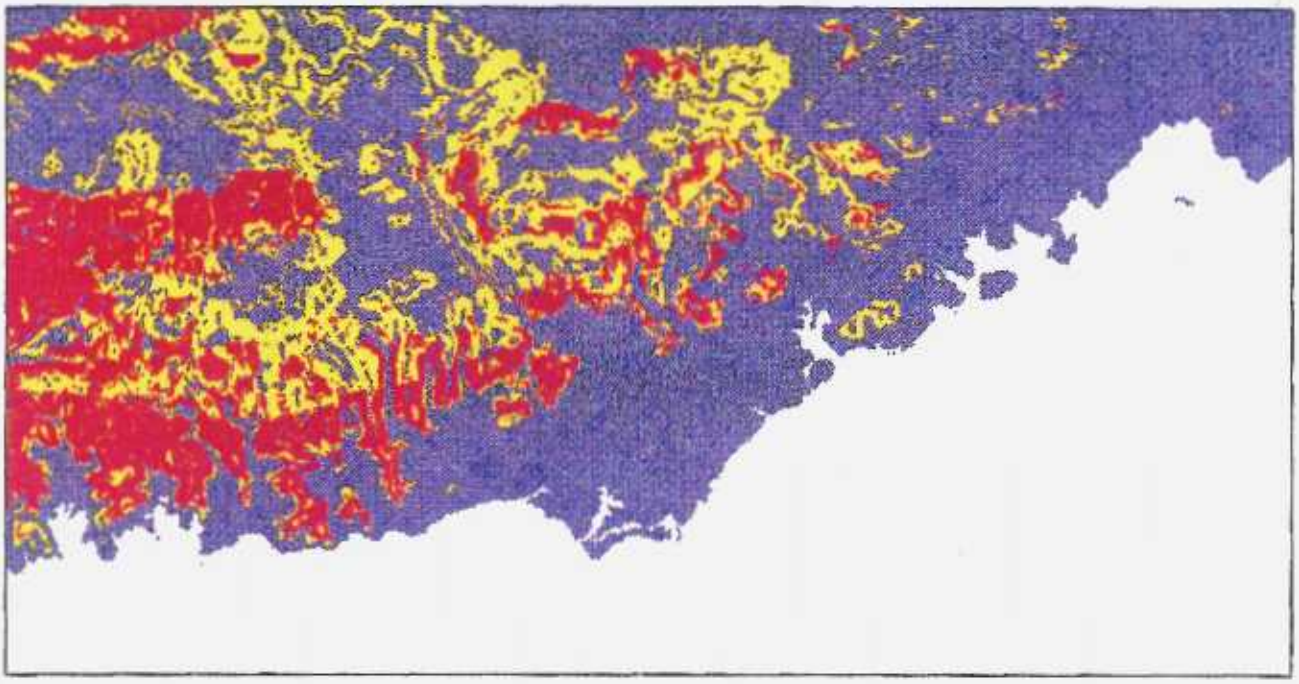


Figure 5.3C Landslide hazard map based on 'geology + slope + soils + elevation' (Model 4).

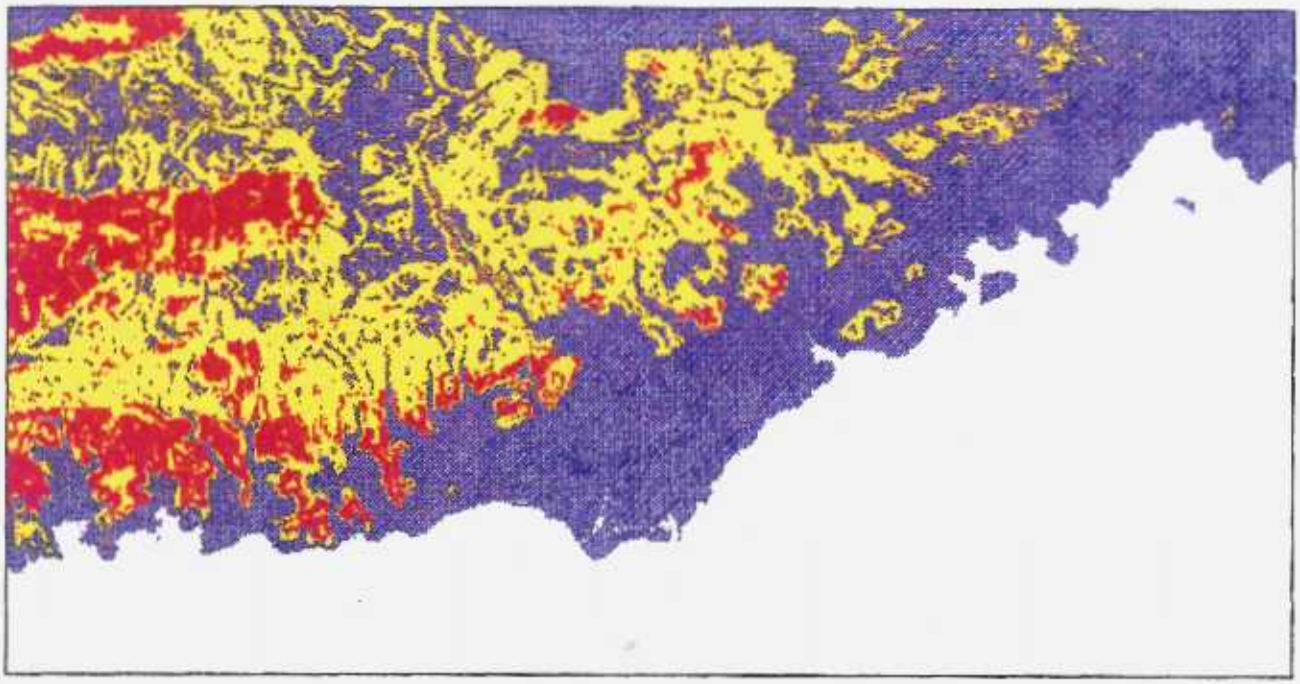


Figure 5.3D Landslide hazard map based on 'geology + slope + soils + elevation + forest' (Model 5).

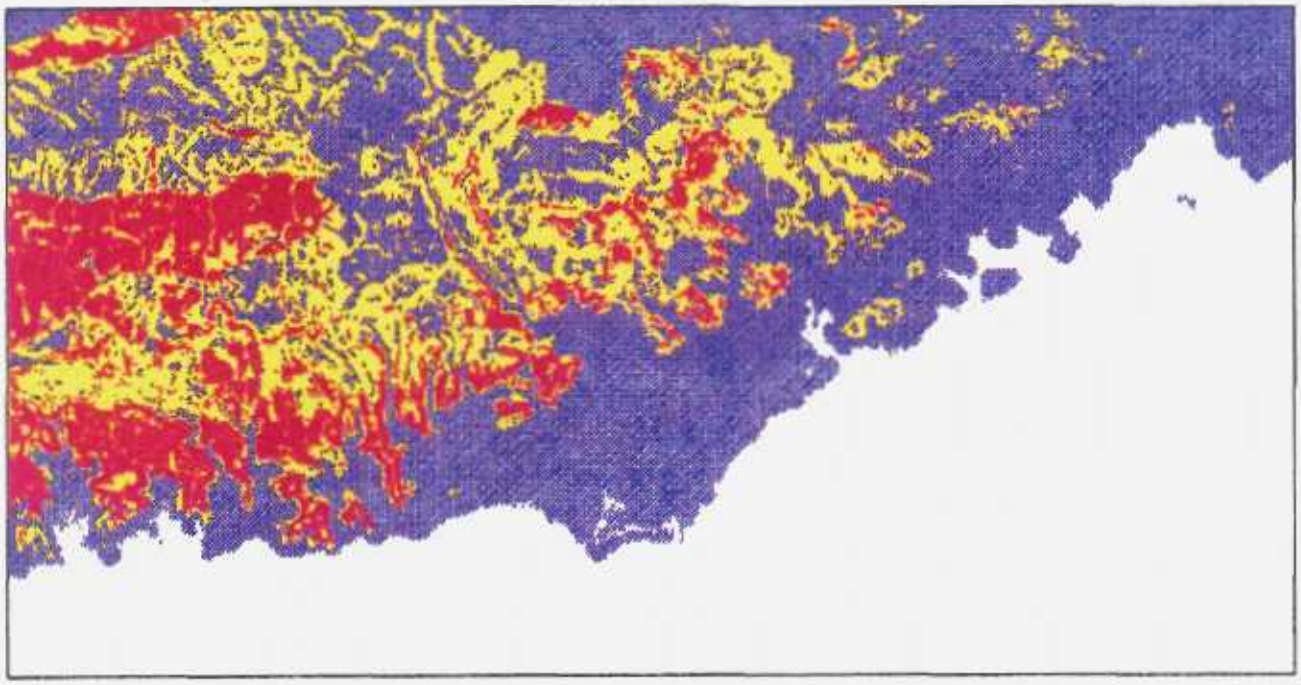


Figure 5.3E Landslide hazard map based on 'geology + slope + soils + elevation + forest + slope aspect' (Model 6).

(A = 56.6%). In Model 6 (Figure 5.3E), the size of the high hazard zone has increased again (A = 29.0%) and the reliability factor is nearly unchanged (R = 32.4%).

Whereas the statistics can be used to describe and characterise the plots, they do not provide a simple quantitative basis on which to measure the improvement, or otherwise, of the successive models. This aspect still requires further work so that models can be more objectively assessed. For the present, it is necessary to consider both the statistics of the models and to visually compare the hazard plots with the total landslide distribution map or, still better, a landslide density plot derived from it (Figure 5.4). Not surprisingly, the models show each varying degrees of similarity with the density plot but the comparison is strongest for Models 4 and 6. Interestingly however, neither of these models show a particularly high correlation with the main high density area seen in the western part of Figure 5.4. This difference is apparently real and demonstrates that the hazard maps are not the simple equivalents of the density plot.

Figure 5.5 is Model 6 in its final form, smoothed to remove unnecessary detail. If compared with Figure 5.3E it can be seen that smoothing has the effect of eliminating spurious small groups of pixels and providing a thematic plot comprising more uniform broad groups.

Based on the statistics in Table 5.4, the hazard levels displayed on this map may be interpreted as follows:

Blue	*Low*:	below average incidence of landslides
Yellow	*Medium*:	21% average probability of landsliding
Red	*High*:	32% average probability of landsliding

Figure 5.5 represents an attempt to model landslide hazards in south east Viti Levu. It is both preliminary, in the sense that further GIS analysis of the existing data is warranted, and provisional in that modified versions could be produced as more information becomes available. For example, lineaments should be added to the GIS and more use made of the separate categories of old, transitional and new landslides, and landslide initiation points. Time prevented this from being done.

The plots show interesting patterns that need explanation and/or validation. For example, the high hazard zone in the southern coastal area does not appear to be wholly a reflection of the Serua Hills landslides. The interactions between the variables are clearly complex, possibly because the variables used do not relate in a direct and simple manner to landsliding. For example, it is possible that the mapped lithostratigraphic sub-divisions do not equate in a simple manner to geotechnical rock properties (e.g. weathering characteristics). It is clear from comparing the recalculated weights with the logical weights in each combination that the effects are non linear but the true nature of what is happening is still far from clear.

Perhaps the greatest limiting factor of the present data set is the variability in the landslide data caused by several workers contributing to the aerial photograph interpretation. This is fundamental since all correlations are based on this information. Future work must address