Reasoning With Uncertainty in Remote Sensing

M. R. Ahmadzadeh

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

February 2001

© M. R. Ahmadzadeh 2001
Summary

This thesis is on information fusion in remote sensing. Several fusion approaches are investigated and some of them are successfully implemented. Assessing the risk of desertification of a forest after a fire, which is the main motivation of this work, depends on many factors. Aggregation of these factors, which are derived from different sources, gives a basis for the evaluation of the risk of desertification.

Different ways of considering the uncertainty and imprecision due to data, concepts, measuring instruments etc in decision making systems, lead scientists to develop different information fusion approaches. In this work I concentrate on uncertainty in the data due to errors in interpolation. The slope and aspect of the terrain are among those factors which influence the risk of desertification. The slope and aspect of the terrain are derived by a Geographic Information System (GIS) from the Digital Elevation Model (DEM). The problem is that although the sources of these data are usually of diverse resolution, all of them are resampled to refer to the same resolution. Re-sampling, which is done by interpolation, introduces errors in the data. Most commercial GISs, in spite of these errors, deal with the data during decision making as if they were precise. Modelling the errors of slope and aspect when computed from interpolated data is the first objective of this thesis. The proposed error models may be used subsequently in the decision making process.

Studying different fusion of information approaches to combat the problem at hand is the second contribution of this thesis. Especially I focus on the Dempster-Shafer evidence theory and its application in combining multi source data.

First I use Dempster's rule of combination as a tool for combining two classifiers: a Bayesian network and a fuzzy logic classifier. These two classifiers have been proposed in the past to assess the risk of desertification of burnt forests. The problem is that one of the classifiers has 3 classes (Bayesian network classifier) and the other one 5 classes (fuzzy logic classifier). To combine these two classifiers a superset of classes is defined, with the help of which the classes of each classifier can be defined by the union of few superset classes. The novelty of the proposed methodology is that not only the two classifiers are of different types (a probabilistic classifier and a Fuzzy logic-based classifier), but also the number of output classes are different.

Finally I examine three other combining approaches: neural network approach, fuzzy neural network approach and application of Dempster-Shafer evidence theory in propagating the belief functions through a network in an expert system. In each experiment the results are compared with the expert results which are derived by inspecting the field data.

Key words: Information Fusion, Remote Sensing, Risk of Desertification, Error Modelling, Dempster-Shafer Theory, Belief Propagation, Fuzzy Neural Network.

Email: M.Ahmadzadeh@eim.surrey.ac.uk
WWW: http://www.eim.surrey.ac.uk/
Acknowledgements

First of all I would like to express my deepest respect and gratitude to my supervisor Professor Maria Petrou. Her advice was always very constructive and helpful.

I would like to thank my friends at CVSSP for their help. Particularly Dr. Paul Duree for providing me with the fractal generation software, Dr K. R. Sasikala for helping me to extract GIS data and Dr. Maria Faraklioti, Reza Ghaderi and Fuad Alkoot for their valuable discussions and suggestions.

I am grateful to the Ministry of Science, Research and Technology of Iran for its financial support during my research studies.

Finally many thanks to my family especially my parents and my wife who encouraged me to continue my study. Their tireless support and patience are highly appreciated.
3.4.1 Statistical Approach ........................................ 55
3.4.2 Dempster-Shafer Evidence Theory ............................ 56
3.4.3 Neural Network Approach .................................... 59
3.4.4 Knowledge-based Systems ................................... 60
3.4.5 Bayesian Network Approach .................................. 60
3.4.6 Fuzzy Logic Approach ..................................... 61

3.5 Previous Work .................................................. 62
3.5.1 Bayesian Network Approach .................................. 62
3.5.2 Fuzzy Logic Approach ..................................... 64

3.6 Conclusions ..................................................... 67

4 Dempster-Shafer Theory ......................................... 69

4.1 Fundamentals of Dempster-Shafer Theory ....................... 69
4.1.1 Bayesian Statistics ........................................ 70
4.1.2 Mass Functions and Belief Functions ...................... 71
4.1.3 Dempster's Rule of Combination ............................ 73
4.1.4 Properties of the Orthogonal Sum ......................... 75

4.2 Comparison of Bayesian Theory and Dempster-Shafer Theory .... 76

4.3 Decision Making in Dempster-Shafer Theory .................... 77

4.4 A Simple Example ............................................. 78

4.5 Dempster-Shafer Theory in Expert Systems .................... 79

4.6 Belief Combination in Rule-based Systems ..................... 82

4.7 Belief Conjunction ............................................. 86
4.7.1 A Simple Example ........................................ 89

4.8 Real Application ................................................ 92

5 Classifier Combination Using the Dempster-Shafer Theory ....... 95

5.1 Introduction ...................................................... 95
5.2 Data Specifications ............................................ 98

5.3 Problem Definition ............................................ 101
5.4 The Proposed System of Classifier Combination ................ 102

5.5 Experimental Results .......................................... 105
5.6 Discussion and Conclusions .................................... 106
6 Application of the Belief Propagation in a Real Problem 109
6.1 Introduction .......................................... 109
6.2 Dempster-Shafer Theory in a Rule-based System .......... 110
6.3 Experimental Results ................................... 111
6.3.1 Lack of Information About Uncertainty in Data and Expert Rules (Experiment I) ......................... 111
6.3.2 Uncertainty in the Data: Error Models (Experiment II) .......... 112
6.3.3 Uncertainty in the Data: Redistribution of Mass Functions (Experiment III) .......................... 114
6.3.4 Reliability of Expert Rules (Experiment IV) ............... 115
6.4 Critical Discussion of Combination of Information for Risk of Desertification . 116
6.5 Conclusions ........................................... 117

7 Artificial Neural Network Classifiers 121
7.1 Multilayer Perceptrons .................................. 122
7.1.1 Algorithm ........................................... 123
7.1.2 Generalisation and Training Data ..................... 124
7.2 Experimental Results of the Neural Network Approach ......... 125
7.2.1 Method I ........................................... 125
7.2.2 Method II .......................................... 126
7.3 Neural Network with Fuzzy Training .......................... 126
7.3.1 Fundamentals of Fuzzy Neural Networks ................. 127
7.3.1.1 Learning Algorithm ................................ 131
7.4 Experimental Results of the Fuzzy Neural Network Approach .... 133
7.4.1 Method I ........................................... 134
7.4.2 Method II .......................................... 136
7.5 Discussion and Concluding Remarks .......................... 137

8 Summary and Conclusions 141
8.1 Possible Applications and Suggestions ........................ 144

A Fractional Brownian Motion 147

B Kriging 149
B.1 Definitions ............................................ 149
B.2 Spatial Continuity Analysis ............................... 151
B.3 Estimation ............................................ 152
List of Figures

2.1 Delaunay triangulation and Voronoi graphs. The circles are data points and each asterisk is the centre of the circumcircle of a triangle and vertex of a Voronoi tile. .................................................. 18

2.2 Definition of slope and aspect. $\vec{N}$ is the normal vector. ......................... 22

2.3 The compass direction of the aspect. ......................................................... 22

2.4 The averaged normal vector for eight adjacent points. ............................. 22

2.5 (a) Mean of absolute error in slope (in percent) versus sampling rate and fractal dimension by Delaunay triangulation method (b)-(e) some statistics of error in slope (in percent) versus fractal dimension for 5% sampling rate (f) Histogram of the error in slope for 5% sampling rate and $D=2.2$; the bin width is 0.229. ........................................ 25

2.6 (a) Standard deviation of error in aspect (in degrees) versus fractal dimension and sampling rate by Kriging method (b)-(e) some statistics of error in aspect versus fractal dimension for 5% sampling rate (f) Histogram of the error in aspect for 5% sampling rate and $D=2.2$; the bin width is 5.54°. ............... 26

2.7 (a) Standard deviation of error in slope and fitting curve (first method) versus fractal dimension for 5% sampling rate (b) Standard deviation of error in slope and fitting curve (first method) versus sampling rate for $D=2.1$ (c) Standard deviation of error in aspect and fitting curve (first method) versus fractal dimension for 5% sampling rate (d) Standard deviation of error in aspect and fitting curve (first method) versus sampling rate for $D=2.1$. .................... 27

2.8 (a) Standard deviation of error in slope and fitting curve (second method) versus fractal dimension for 5% sampling rate (b) Standard deviation of error in slope and fitting curve (second method) versus sampling rate for $D=2.1$ (c) Standard deviation of error in aspect and fitting curve (second method) versus fractal dimension for 5% sampling rate (d) Standard deviation of error in aspect and fitting curve (second method) versus sampling rate for $D=2.1$. ............. 28

2.9 (a) A subsampled fractal; 20% of the original data retained (b) The fractal features of the surface of figure (a) can be calculated from this line as: $H=0.82, 2 \log \sigma = -1.91$. .................................................. 31

2.10 PDF of error in slope and its fitting function (Gaussian) for an artificial terrain. 34

2.11 PDF of error in aspect and its fitting function (Laplacian) for an artificial terrain. ................................................................. 34

2.12 PDF of error in slope and its fitting function (Laplacian) for a real terrain. 34
2.13 PDF of error in slope and two fitting functions (Gaussian and Laplacian) for a real terrain. ................................................................. 34

2.14 Histogram of increments (solid line) and the Gaussian fitting function (dashed line) for $|\Delta \bar{x}| = 8$ and 10 bins. Vertical axis is the number of pairs which fall into bins. ................................................................. 35

2.15 Histogram of increments (solid line) and the Laplacian fitting function (dashed line) for $|\Delta \bar{x}| = 8$ and 10 bins. Vertical axis is the number of pairs which fall into bins. ................................................................. 35

2.16 Standard deviation of error in slope versus fractal dimension for different sampling rates of the terrain. The open circles are the experimental results of 5% sampling rate of validation fractals. ................................................................. 36

2.17 Standard deviation of error in aspect versus fractal dimension for different sampling rates of the terrain. The open circles are the experimental results of 5% sampling rate of validation fractals. ................................................................. 36

2.18 The mean of the absolute error in (a) slope (b) aspect, and the standard deviation of the absolute error in (c) slope (d) aspect for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Delaunay triangulation. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line. ......................... 38

2.19 The standard deviation of the absolute error in (a) slope (b) aspect, and the mean of the absolute error in (c) slope for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Delaunay triangulation. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line. ......................... 39

2.20 The mean of the absolute error in (a) slope (b) aspect, and the standard deviation of the absolute error in (c) slope (d) aspect for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Kriging. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line. ......................... 40

2.21 The standard deviation of the absolute error in (a) slope (b) aspect, and the mean of the absolute error in (c) slope for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Kriging. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line. ......................... 41

2.22 A real terrain. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. The same scale is used for the vertical axis too. ................................................................. 43

2.23 The fractal features of the real terrain of figure 2.22 can be calculated from this line as: $H=0.75$, $2 \log \sigma = -1.63$. ................................................................. 43

2.24 The terrain of figure 2.22 when subsampled with 5% sampling rate. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 43

2.25 The fractal features of the subsampled terrain in figure 2.24 can be calculated from this line as: $H=0.78$, $2 \log \sigma = -1.71$. ................................................................. 43
2.26 Interpolated terrain of figure 2.24 by the Delaunay triangulation method. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 44

2.27 Error in the elevation of the interpolated terrain shown in figure 2.26. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 44

2.28 Error in slope (in percent) of the interpolated terrain in figure 2.24. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent slope in percent. ................................................................. 44

2.29 Error in aspect (in degrees) of the interpolated terrain in figure 2.24. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent aspect in degrees. ................................................................. 44

2.30 The PDF of errors in slope and its fitting function (Gaussian) for the subsampled terrain of figure 2.24. ................................................................. 45

2.31 The PDF of errors in aspect and its fitting function (Laplacian) for the subsampled terrain of figure 2.24. ................................................................. 45

2.32 A real terrain. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. The same scale is used for the vertical axis too. ................................................................. 46

2.33 The fractal features of real terrain of figure 2.32 can be calculated from this line as: $H=0.78$, $2\log\sigma = -1.71$. ................................................................. 46

2.34 The terrain of figure 2.32 when subsampled with 3% sampling rate. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 46

2.35 The fractal features of the subsampled terrain of figure 2.34 can be calculated from this line as: $H=0.80$, $2\log\sigma = -1.80$. ................................................................. 46

2.36 Interpolated terrain of figure 2.34 by the Kriging method. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 47

2.37 Error in the elevation of the interpolated terrain shown in figure 2.36. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. ................................................................. 47

2.38 Error in slope (in percent) of the interpolated terrain in figure 2.34. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent slope in percent. ................................................................. 47

2.39 Error in aspect (in degrees) of interpolated terrain in figure 2.34. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent aspect in degrees. ................................................................. 47

2.40 The PDF of errors in slope and its fitting function (Gaussian) for the subsampled terrain of figure 2.34. ................................................................. 48
2.41 The PDF of errors in aspect and its fitting function (Laplacian) for the sub-sampled terrain of figure 2.34. ........................................ 48

3.1 Relation between variables and natural regeneration potential, soil erosion and risk of desertification. ........................................ 63

4.1 Combination of mass functions. ........................................ 74
4.2 Combination of mass functions. ........................................ 79

5.1 Definition of the superset of classes. ............................. 103
5.2 Combination of mass functions. ........................................ 104
5.3 Proposed combination of classifiers system to assess the risk of soil erosion. 105

7.1 A multilayer perceptron with two hidden layers. .................. 123
7.2 Arithmetic addition of two fuzzy numbers. .......................... 128
7.3 S and π functions are used to define the membership functions. .......................... 135

B.1 Exponential covariance function (a) and variogram function (b) for $C_0 = 0$, $C_1 = 10$ and $a = 10$. ............................... 157
B.2 Spherical covariance function (a) and variogram function (b) for $C_0 = 0$, $C_1 = 10$ and $a = 10$. ............................... 157
B.3 Gaussian variogram function model $\tilde{\gamma}(h) = C_0 + C_1(1 - \exp(-\frac{3|h|}{a})^2)$. ............................... 158
B.4 Linear variogram function model $\hat{\gamma}(h) = C_0 + C_1(h/a)$. ............................... 158
B.5 Power variogram function model $\gamma(h) = C_1 h^{2H} = Ch^b$ for b=0.6. ............................... 158
B.6 Power variogram function model $\gamma(h) = C_1 h^{2H} = Ch^b$ for b=1.6. ............................... 158
List of Tables

2.1 Empirical formulae for error (ε) in slope and aspect. In these formulae $0 \leq P \leq 0.4$ and $2 \leq D \leq 2.5$ and $K = \log^{-1}(\log(\sigma) - \log(\sigma_{ref}))$. .............................. 30

2.2 Fractal features of fractals which are used in this research. .............................. 31

2.3 Coefficients of polynomials which are used to fit the error curves of the Delaunay and Kriging methods respectively. Columns 1 to 4 are statistics of error in slope (mean and SD of absolute error and mean and SD of error respectively) and columns 5 to 7 are statistics of error in aspect (mean and SD of absolute error and SD of error respectively). .............................. 33

2.4 Statistics of the absolute relative difference between the errors predicted by the empirical formula and those computed directly from real data, computed over 100 different experiments of terrain subsampling, interpolation, slope and aspect computation, and error estimation. .............................. 42

2.5 Experimental results and predicted statistics of error (ε) in slope (in percent) and aspect (in degrees) by using the empirical formulae for the terrain of figure 2.22. .............................. 45

2.6 Experimental results and predicted statistics of error (ε) in slope (in percent) and aspect (in degrees) by using the empirical formulae for the terrain of fig. 2.32. .............................. 48

3.1 Definition of some operators: original and generalised forms. .............................. 66

4.1 Logical interpretation of rules. .............................. 84


5.2 Rules for Natural Regeneration Potential. NL: No Limitation, SL: Slight Limitation, ML: Moderate Limitation, SG: Strong Limitation, SE: Severe Limitation. .............................. 100

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Experimental results. Under CCS the number of the sites which were classified in the same class as that assigned by the expert. Under CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>112</td>
</tr>
<tr>
<td>6.2</td>
<td>Experimental results. Under CCS the number of the sites which were classified in the same class as that assigned by the expert. Under CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>118</td>
</tr>
<tr>
<td>7.1</td>
<td>Experimental results for the risk of desertification. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>126</td>
</tr>
<tr>
<td>7.2</td>
<td>Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD). CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>127</td>
</tr>
<tr>
<td>7.3</td>
<td>Experimental results for the risk of desertification. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>135</td>
</tr>
<tr>
<td>7.4</td>
<td>Experimental results for the risk of desertification when expert rules are used. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>136</td>
</tr>
<tr>
<td>7.5</td>
<td>Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD). CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>137</td>
</tr>
<tr>
<td>7.6</td>
<td>Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD) when expert rules are used. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.</td>
<td>138</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Problem Motivation and Definition

Desertification is land degradation in arid, semiarid and dry sub-humid areas. This can happen due to overgrazing, climate changes, land mismanagement, human activities etc. Fire in the forests, which is due to lightning or is human made, is one of the most important factors contributing to the desertification of forests in the Mediterranean region.

Two major factors which influence the risk of desertification of a forest after a fire are the natural regeneration potential and soil erosion. Some kinds of plants like maquis can regenerate naturally by re-sprouting. Some other kinds of plants and trees can regenerate naturally after a fire if there are enough good seeds and minerals in the soil. However in some situations some sites can not regenerate naturally. If these regions are not afforested they are highly at risk of desertification. Regions which are left without cover, gradually loose the soil, degrade, and change to bare area which can not be afforested easily due to lack of high quality top soil. Because of the fire, some chemical and physical properties of the soil are affected, and this may lead to erosion by heavy rains. Afforestation, therefore, of the affected sites, which are at high risk of desertification, is of paramount importance.

The problem is that usually resources are not enough to manually afforest all the affected areas. However, some regions are not at high risk and can naturally regenerate 2-5 years after the fire. So evaluation of the risk of desertification of burnt forests can be used to prioritise the resources for afforestation. Usually experts examine the affected areas, measure the variables which influences the natural regeneration potential and soil erosion, and then they
make decisions by using some rules which relate the variables of interest. However, training an expert and collecting ground data are very expensive processes. Geographic Information Systems (GIS) can analyse the various layers of data and make decisions using the expert rules. For example, slope, aspect, soil depth and rock permeability are factors which have been singled out in this study to assess the natural regeneration potential and risk of soil erosion. Other factors, common to all areas of study have been ignored. The chosen factors are those which influence the relative ranking of desertification of the small areas under study.

Multispectral remote sensing data and spatial data on rock permeability and soil depth are fed to a GIS. A Digital Elevation Model (DEM) which is derived from topographic maps and satellite data is used to derive secondary data, slope and aspect. The problem is that although the sources of these data are usually of diverse resolution, all of them are resampled to refer to the same resolution. Re-sampling is done by interpolation. This process introduces errors in the data. Most commercial GISs, in spite of these errors, deal with the data during decision making as if they were precise. Modelling the errors of slope and aspect when computed from interpolated data is the first objective of this thesis.

Given the remote sensing and spatial data, the GIS developed [57] calculates the slope, aspect, soil depth and rock permeability of each pixel of the region after re-sampling. To decide upon the risk of desertification, the GIS uses the expert rules which have been keyed in. These rules express the relation between the data and the final class which should be assigned. For example, one of the rules says “IF the soil depth is shallow and the rock is permeable and the slope is gentle THEN the risk of soil erosion is moderate”. The GIS uses these rules directly, without paying attention to the uncertainty in the data caused by errors in the measurements and/or interpolation and uncertainty in the rules. The issue of combining information and taking into consideration the uncertainty is a major topic of research. There are various approaches one may use: Bayesian, Fuzzy, Neural Networks and Dempster-Shafer theory. In the past, two major research projects have been carried out on this problem by Stassopoulou et al. [70, 71, 72] and Sasikala et al. [60, 61, 62]. Stassopoulou et al. developed a Bayesian network to solve the problem of predicting the risk of burnt forest desertification. They used Pearl’s algorithm for propagation of the items of evidence through the network. Sasikala et al. [60, 61, 62] used fuzzy logic to solve the same problem. They used disjunctive and conjunctive operators to combine membership functions of variables based on expert rules. The fuzzy conjunctive operators were used for
aggregating the conditions which variables should meet to trigger a rule, and the disjunctive operators were used to combine different expert rules which lead to the same conclusion. In this thesis we deal with all the above fusion of information approaches, with emphasis on the application of Dempster-Shafer evidence theory in the problem under study.

In some applications, results of individual classifiers are imprecise such that decision making based on individual "experts" is unreliable or perhaps impossible. Classifier combination has increasingly been drawing the attention of scientists in many fields of research. The aim of classifier combination is to use the results of different classifiers in order to increase accuracy and improve performance. This need is also prominent in remote sensing where we need much more reliable and accurate decision making systems, based on the combination of multi source / multi sensor data. In this thesis we propose the Dempster-Shafer theory as a mechanism for classifier combination.

1.2 Scope of the Thesis: An Overview

The first contribution of this thesis is studying the distribution of errors in slope and aspect when these variables are computed from interpolated data. Roughness of surface of terrain and the number of data points which are used for interpolation are two important factors which influence the magnitude of the introduced error. The rougher the terrain, the more interpolation error for fixed rate of sampling, and the less data used for interpolation (small sampling rate), the more interpolation error for a fixed roughness of the terrain. It is the first time that this type of error is being modelled. These models express error statistics (mean of error, standard deviation of error, mean of absolute error and standard deviation of absolute error) versus the roughness of the terrain and the sampling rate (percentage of data points which are originally used for the interpolation over the final number of points).

The roughness of the terrain is expressed by the fractal dimension which theoretically can vary between 2 and 3, but for realistic looking landscapes one should not exceed the value 2.5. Fractal modelling has been used to create many terrains of various degrees of roughness. Created fractals which are subsequently subsampled by various rates, are interpolated using either Delaunay triangulation, or Kriging. The slope and aspect of the reconstructed terrains are computed and their values are compared with the values that would have been computed from the original data. Error statistics are calculated and fitting functions as functions of the fractal dimension and the rate of sampling are created. The roughness of the surface
can be calculated from original data with the help of the variogram. The error distribution has also been studied and modelled by fitting some parametric functions.

The second contribution of this thesis is studying different fusion of information approaches to combat the problem at hand. Especially we focus on the Dempster-Shafer evidence theory and its application in combining multi source data.

As it is clear from the above brief introduction to our problem, the main objective of our study is aggregation of data which are extracted from satellite images and thematic maps to assess the risk of soil erosion, the natural regeneration potential and finally the risk of desertification. In the Stassopoulou et al. approach uncertainty of the data was considered by selecting the probabilities of labels of root nodes, which are nodes with no incoming link from another node, from the interval \([0, 1]\). This means that the uncertainty of data was modelled by the prior probabilities. However they assumed that the distribution of error was Gaussian with arbitrary parameters which where selected intuitively. Since they used a training phase to derive the conditional probability matrices, the available expert rules were not used directly. A problem with their method was that the size of probability matrices depended on the number of nodes and the number of states which each node could take. For example, to evaluate the risk of soil erosion three attributes were combined. If we assume that each variable involved, takes \(N\) possible values, the matrix must be \(N \times N \times N \times N\). So, for \(N = 5\), there should be 625 elements of the matrix, each expressing the probability of the site to belong to a certain class of soil erosion, given that the site attributes have certain combination of classes. The calculation of such a large number of probabilities, however, required the availability of a large amount of data. That is why they had to reduce the number of classes of the attributes from 5 to 3. This caused less accuracy in assessing the risk of desertification as their system could categorise the output into only 3 classes like low, medium and high risk. Their method is explained with more detail in chapter 3.

Sasikala and Petrou [61] used a Fuzzy approach to solve the same problem. They solved the problem of the relative reliability of the different sources of information by a novel approach. They generalised the conventional fuzzy operators by allowing the membership functions to get values which reflect their relative importance. These values were not restricted to the interval \([0, 1]\) as in conventional fuzzy sets. Their approach has a shortcoming: their classifier is blunt as some times several outputs get the highest possible value. The correct class is usually among the classes selected, but such a behaviour of a classifier is not desirable.
The second contribution of this thesis is the combination of the above two mentioned classifiers i.e. the Bayesian network and the fuzzy logic classifier using the Dempster-Shafer evidence theory. While most classifier combination approaches are used to combine results of classifiers which have the same number of outputs, in the approach proposed in this thesis we combine two classifiers which have different classes. One of the classifiers has 3 classes (Bayesian network classifier) and the other one 5 classes (fuzzy logic classifier). To combine these two classifiers a superset of classes is defined, with the help of which the classes of each classifier can be defined by the union of few superset classes. Furthermore, uncertainty in the expert rules which has been ignored in the fuzzy logic approach is considered in the proposed system. The relative reliability of the classifiers is also taken into consideration in the combination process. The results show that this system is superior to both individual classifiers in having the advantages of both systems and lacks their shortcomings. The proposed combination approach is applicable to multi neural network classifiers in modular mode.

However, Dempster-Shafer theory can be used directly to solve the problem at hand, and not just as a combiner of classifiers. So, we used this theory for propagating belief functions in an expert system. The mass functions are derived from information regarding the distribution of errors as derived in chapter 2 or the percentage of pixels of each site which fall into the same data class. The mass functions are used to define the belief function of the IF part of the expert rules which contain a combination of few conditions. The defined belief functions are to be propagated through the expert rules. Uncertainty of the expert rules can be considered easily in this step. Using the procedure which has been proposed by Hau and Kashyap [19] we show two different approaches by which to incorporate expert rules within the framework of Dempster-Shafer theory. In this part of the work the error models derived in chapter 2 are used. A training-based approach is used for the variables for which error models are not available.

The last fusion of information approaches which is examined are neural network and fuzzy neural network approaches. A neural network with backpropagation algorithm is implemented to classify the sites into defined classes for the natural regeneration potential, risk of soil erosion and the risk of desertification. In this experiment expert rules are not used and according to the neural network philosophy, the input-output relations are derived in a training phase. However, expert rules can be used to contribute in the training of a specific class of a neuro-fuzzy network. In the implemented fuzzy neural network, which has
been proposed by Ishibuchi et al. [24], fuzzy expert rules can be used to train the network. Interval valued feature vectors extracted from examples or fuzzy expert rules are presented to the network. Results show that if the training samples are in agreement with the expert rules, the performance and accuracy of the network is improved.

1.3 Layout of the Thesis

In this section the overall organisation of the thesis is presented.

In chapter 2 terrain modelling in general is discussed, followed by an introduction to simulated terrains and fractal modelling. Two different interpolation approaches, namely Delaunay triangulation and Kriging interpolation, which usually are used by GIS for interpolation, are discussed subsequently. Error models for slope and aspect, when they are computed from interpolated surfaces, are derived empirically. These models can be used within a GIS in parallel with the interpolation packages. It is shown that the error distribution of slope follows a Gaussian form while that of aspect follows a Laplacian form. The proposed error models are validated with real terrain data.

In chapter 3 various data fusion approaches which have been used in remote sensing is surveyed. Especially we review statistical, neural network, Dempster-Shafer-based, fuzzy logic, rule based and Bayesian network approaches for multi source / multi sensor problems. We also review two major projects which have been carried out in the past on the same problem that we study.

The fundamentals of Dempster-Shafer evidence theory are reviewed in chapter 4. Dempster-Shafer theory is compared with Bayesian theory and the relation between Bayesian belief functions and probability is given. The incorporation of expert rules in the evidence theory is discussed briefly.

Fusion of information at decision level by Dempster-Shafer evidence theory is studied in chapter 5. The results of two classifiers, a Bayesian network and a fuzzy logic system, are used to improve the accuracy of the overall system. The novelty of this work is that the two classifiers have different information classes which span the same classification space.

In chapter 6 the problem of risk of desertification is studied based on belief propagation in Dempster-Shafer theory. In other words, chapter 6 is an application based on the theory
of chapter 4. Error distribution models which have been proposed in chapter 2 are used to re-distribute the mass functions.

Since in some experiments we used training to identify weights, reliability factors etc. We considered it appropriate to implement for comparison some of the classical training based approaches, namely neural networks and fuzzy neural networks. Results obtained by a 3 layer feedforward neural network, which is trained by the backpropagation algorithm, are presented in chapter 7. Also a neuro-fuzzy network which is proposed by Ishibuchi [24] for training a network by IF-THEN rules is implemented for the problem. Again the error distributions derived in chapter 2 are used to calculate the mean of each variable of a site. Some popular membership functions are used to derive the interval-valued feature vectors which are used as input signals. The implemented neuro-fuzzy network is able to be trained by examples as well as expert rules which are expressed in IF-THEN format.

Eventually discussion and conclusions are presented in chapter 8.
Chapter 1. Introduction
Chapter 2

Error Modelling of Slope and Aspect

Error models of slope and aspect of a terrain are presented in this chapter. Such data are often extracted from a GIS which may contain information from digital maps and remote sensing images. Although the sources of these data are usually of diverse resolution, all of them are usually re-sampled to refer to the same resolution. In this chapter we examine the error which is associated with such data because of subsampling. The error distributions are modelled empirically.

2.1 Introduction

Very often the data stored in a Geographical Information System (GIS) have originated from maps or images of different resolutions. GIS contains packages which allow the interpolation and re-sampling of the data so that all corresponding layers are of the same resolution. Inevitably, such processing of the data introduces errors. The errors of interpolation techniques have been the subject of many studies. What we are concerned here is the way these errors propagate to subsequent levels of processing, i.e. to the derivation of secondary data from the original ones. In particular, we have in mind problems like those of predicting the risk of erosion, the risk of desertification, etc. Among others, some variables which influence these conditions are the ground slope and the aspect of a region. Both these factors are secondary data, derived from the digital elevation models of the terrain
by digital differentiation. Differentiation is a process known to greatly amplify noise. Any error therefore in the original data is bound to become more significant upon differentiation.

Given that data often stored in a GIS are of varied degree of reliability, spatial combination of attributes by simple spatial superposition ignores this effect totally. Spatial superposition of causes and the application of some production rules is the standard way of reasoning with such systems in order to predict effects. This way, however, one forces out of the data more information than they actually contain. It produces crisp results of absolute confidence which cannot be justified given the different levels of certainty in the data. There are various ways by which one can try to incorporate the uncertainty of the input data to the outcome. Among others, one can use probability theory [72], fuzzy logic [61] or Dempster-Shafer theory [2, 63]. All these methods assume the knowledge of the statistical distributions of errors in the input data. They provide then mechanisms of propagating these distributions in the output layers. A common assumption is that the input variables carry additive errors with Gaussian distributions. This assumption is not entirely wrong, as according to the central limit theorem, if we have many independent sources of error, the overall result will be a Gaussianly distributed error, irrespective of the actual original distributions. However, in these cases, very little effort has been made to correlate the error with the actual nature and state of the data. If there is any degree of non-linearity in the processes involved, the error distribution becomes a function of the data themselves and it cannot be modelled in a completely independent way. Making the Gaussian assumption leads to membership functions expressed in terms of the error function \( \text{erf}(x) \) [62]. The other extreme is to assume total ignorance about the exact value of an input variable, and simply say that it is uniformly distributed within a range of extreme values. This assumption leads to the classical triangularly shaped membership functions of Fuzzy logic. Dempster-Shafer theory on the other hand, follows a more cautious approach: it attaches a minimum and a maximum probability to each possible outcome, called belief and plausibility respectively. These two probabilities are derived on the premise that positive evidence leads to a certain probability of an event to happen, but the complement of this probability is not the probability of the opposite conclusion. It is simply the plausibility of the opposite conclusion. The probability of the opposite conclusion is defined on the basis of positive supporting evidence for the opposite conclusion. This approach leads to the issue of propagating to the output layers intervals of probabilities rather than single probabilities that express the confidence in the data.
2.1. Introduction

If all these elaborate mechanisms of propagating errors in a system are to be of any value, then the error distributions that are propagated have to be realistic and model faithfully the real error distributions. In most cases, by far the most significant source of error is that of under-sampling the data, rather than instrumentation error which is often represented well by Gaussian distributions. The effect of data under-sampling is an aliasing effect: the more information the data contain in high frequencies, the worse the error will be for a given sampling rate. For example, in the case of terrain interpolation for predicting the degree of erosion, the rougher the terrain, the more inaccurate the interpolation error for fixed rate of sampling, and the more significant the error in the calculation of the secondary variables, like slope.

It becomes clear from the above, therefore, that it would be very useful if we had models of the error distributions for the data, either for the primary or the secondary variables, and if these models were appropriate for different terrain types and different sampling rates.

This chapter is exactly on this problem: We derive error distributions for calculating the slope and aspect for terrains of various degrees of roughness and for various sampling rates. Ideally one should have real terrains in high resolution and of various degrees of roughness, subsample them by various degrees, and perform statistics on the interpolated outputs. Alternatively, one can create artificial terrains and follow the same procedure. The creation of artificial terrains has the advantage of allowing us to control the data and create all types of surfaces, all with the same accuracy, and of predetermined roughness. One can perform a large numbers of experiments with such data. In addition, the process of terrain creation allows us to express in a quantitative way the degree of roughness of the terrain. Of course, it is very important that the model terrains we create are for all practical purposes realistic. In fact, they do not need to be realistic in all aspects, but they have to be realistic with respect to those aspects that enter into the calculation of the error distributions. For example, when calculating the slope and aspect of a terrain, we use differences in height values between neighbouring samples. If we have models, therefore, that are realistic with respect to the second order statistics, we expect that the conclusions we shall draw concerning the errors will be reliable and valid for real terrains. The second order statistics of a real terrain can be expressed with the help of the variogram of the terrain (see appendix B for the definition of variogram). Computer Graphics people have been using the linear approximation of the variogram in the logarithmic \((\ln - \ln)\) space, for many years now, in order to create very realistic looking terrains: the method they use is based on fractal geometry [52, 78].
The roughness of the terrain in such models is expressed by the fractal dimension which theoretically can vary between 2 and 3, but for realistic looking landscapes one should not exceed the value 2.5.

In this chapter we use fractal modelling to create many terrains of various degrees of roughness. We subsequently subsample them by various rates and interpolate them using either Delaunay triangulation, or Kriging. The slope and aspect of the reconstructed terrains are computed and their values are compared with the values that would have been computed from the original data. Error distributions are calculated and fitting functions of their means and variances as functions of the fractal dimension and the rate of sampling are created.

In section 2.2 we discuss representation of a terrain. Then in section 2.3 we present a brief overview of the method with which the model terrains were constructed. In section 2.4 each of the two interpolation methods used is reviewed. In sections 2.5 and 2.6 we present the way we calculate the slope and aspect from these model terrains, and the error distributions extracted for some of the cases. We also present the empirical formulae derived for the parameters of these error distributions as functions of the roughness of the terrain and the rate of sampling. Modelling of the distribution of error in slope and aspect is given in section 2.7. Validity of the proposed empirical formulae is discussed in section 2.8. We present our conclusions in section 2.9.

### 2.2 Terrain Modelling

Digital Terrain Modelling (DTM) is the most important part of any Geographical Information System (GIS). The DTM is used to model, analyse and display the topography and other related surfaces. Any interpolation method is bound to introduce some local error. The issue is even more relevant to digital terrain models than to other interpolated surfaces as terrain is known to exhibit variation in all scales and it is best modelled by a fractal rather than a smooth surface [52].

Digital Terrain Modelling as an important part of any GIS was used in this area by Miller and Laflamme for the first time [45]. After that, the DTM as a digital representation of the surface of the Earth has been the basis for many applications in the Earth and engineering science. A DTM can be often represented by fields in which every point \((x, y)\) in a Cartesian coordinate system has a unique \(z\)-value because overlapping cliffs and faults are very rare.
in nature [80]. Since we are discussing the surface of the Earth specifically, sometimes we use the term Digital Elevation Model (DEM) instead of the term DTM.

### 2.2.1 Data Sources of Digital Terrain Models

The quality of a DTM depends on the source of data. In addition to elevation, some other information such as drainage channels, ridges and discontinuities are effective in the consistency of the model with the real terrain.

The three most important sources of data for a DTM are: Ground surveys, photogrammetric data capture and digitised cartographic data. Ground survey data are accurate but are time consuming to collect. So this method is used for small areas and specific projects. Photogrammetric data are based on stereoscopic interpretation of satellite imagery or aerial photographs. There are several photogrammetric sampling methods such as regular pattern sampling, selective sampling patterns, progressive sampling, composite sampling and digital stereo image correlation. The aim of these methods is to minimise the number of samples needed and to maximise the accuracy of the model [53, 80]. Digitised cartographic data are derived from cartographic documents such as contour maps and profiles. This is done by the digitisation of maps and by processing the digitised data to drive the DTM.

### 2.2.2 Data Structures for Digital Terrain Models

After the extraction of the data from the terrain, the next step is the representation of data to enable us to handle them. Two major data structures which are used frequently are the Rectangular Grid and the Triangulated Irregular Network (TIN). Grids are represented by matrices because the spacing between the data points is uniform. Redundancy in the data is one of the features of this method. So, from the point of view of storage, this is a far too expensive method for satisfactory accuracy. Further structural features, such as topographic features, cannot be described by this method. However, handling of the data is easy.

TIN structures are based on continuous and planar triangles which are constructed by data points. It is not necessary for the data points to be regularly spaced and actually they can be scattered. In this method structural features of the terrain can be described easily. Either method is used in different applications and sometime it is necessary to convert from one method to the other.
2.3 Simulated Terrains

Fractal geometry was proposed by Mandelbrot for the first time. Fractal geometry is a very useful tool for describing natural things such as clouds, trees, coastlines, rocks, mountains etc [1, 49, 78]. The main idea of fractal geometry is self-similarity of a shape by scaling. The self similarity says that an object retains its shape under scaling. This is expressed by $1 = Nr^D$ or $D = \frac{\log(N)}{\log(1/r)}$ where $N$ is the number of parts of the object under scaling down by $r$, and $D$ is the fractal dimension. For example, in two dimensional space and for $r = \frac{1}{2}$ we have $N = 4$. It means that by scaling down a square by ratio $\frac{1}{2}$ we will have four parts (squares) which are similar to the original one. The self-similarity can be extended to statistical properties of a set $S$ which is called statistical self-similarity. A set $S$ is statistically self-similar if it is composed of $N$ distinct subsets which are scaled down by ratio $r$, and all statistical properties of the subsets and set $S$ are similar. Sometimes it is necessary to scale different geometrical directions of points by different ratios to preserve statistical moments. For example, if the $x$ direction is scaled by factor $a$, the $y$ direction maybe scaled by factor $a^H$. A set with this property is called statistically self-affine.

Fractional Brownian motion (fBm), which is a mathematical generalisation of the Brownian motion, is one of the models of such fractals [84]. In fBm, $V_H(t)$ is a single valued function of one variable, $t$ (usually time). Its increments $V_H(t_2) - V_H(t_1)$ have a Gaussian distribution with variance

$$\text{Var} = \langle |V_H(t_2) - V_H(t_1)|^2 \rangle > \alpha |t_2 - t_1|^{2H} \quad (2.1)$$

where the brackets $\langle \rangle$ denote average over all samples of $V_H(t)$. Parameter $H$ takes values in the range $0 < H < 1$. This function is stationary, isotropic and continuous but it is not differentiable [78]. For $H = 1/2$ we have $\text{Var} = \langle \Delta V^2 \rangle > \alpha \Delta t$ which is the familiar Brownian motion. The derivative of normal Brownian motion, corresponding to $H = 1/2$, is uncorrelated white Gaussian noise. For $H > 1/2$ there is a positive correlation and for $H < 1/2$ there is a negative correlation for both increments of $V_H(t)$ and its derivative fractional Gaussian noise [78].

Let $Z_H(\vec{x})$ be a random function which exhibits pure fractional Brownian motion and $\vec{x}$ a point in $E$ dimensional Euclidean space. So we have:

$$Pr\left[ \frac{Z_H(\vec{x} + \Delta \vec{x}) - Z_H(\vec{x})}{\|\Delta \vec{x}\|^H} < y \right] = F(y) \quad (2.2)$$
2.3. Simulated Terrains

$F(y)$ is a cumulative distribution function of a random variable $y$. The self-affine parameter $H$ is a constant which lies in the range $0 < H < 1$ and is called Hurst dimension [1, 10, 55, 84]. If $F(y)$ is assumed to be a zero-mean Gaussian distribution $N(0, \sigma^2)$ with a variance $\sigma^2$, the following equation can be derived from equation 2.2 [14, 84]:

$$< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})| > = C\|\Delta \bar{x}\|^H$$

(2.3)

Constant $C$ is equal to the mean of the random variable $|y|$ and is given by $C = \sqrt{2/\pi}$. It can be shown that the following relations are valid as well (see appendix A):

$$< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 > \propto \|\Delta \bar{x}\|^{2H}$$

$$< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 > = C_1\|\Delta \bar{x}\|^{2H}$$

(2.4)

where $C_1 = \sigma^2$. By taking the logarithm of equation 2.4 we have:

$$\log(< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 >) = \log(\sigma^2) + 2H \log(\|\Delta \bar{x}\|)$$

(2.5)

If we plot $\log(< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 >)$ against $\log(\|\Delta \bar{x}\|)$ we obtain a line with slope $2H$. The vertical axis intercept of the line is equal to $\log(\sigma^2)$. The relation between $H$, the Hurst dimension, and $D$, the fractal dimension, is given by $D = (E + 1) - H$ in an $E$ dimensional Euclidean space [15]. For example, in the two dimensional Euclidean space $D = 3 - H$.

Fractals have been used by many people in different fields. In geology to generate realistic looking terrains, in computer graphics to simulate flight and animation, in pattern recognition for texture analysis [30]. In the literature there are several applications of fBm in terrain modelling. Although it is not claimed that a fractal is the perfect model for real terrain, nevertheless it is one of the best options. Yokoya et al. [84] showed that real terrain preserves fractal features, $H$ and $\sigma$, in a wide range of distances. Although an fBm function has the same parameter values in all ranges, Yokoya et al. showed that the fractal parameters of a terrain vary smoothly from position to position.

There are several methods for constructing fractal-based artificial terrains, such as random midpoint displacement, fast Fourier transform filtering, successive random additions etc [78]. The artificial terrains which we have used in this research were generated by Fourier transform filtering but in the validation step fractals which were generated by the successive random additions method were also used.
These results are reported in row Exp. IV.-2 in table 6.1. In all other experiments we used steps equal to 0.1 only for the exhaustive search.

In another experiment we followed the above procedure but we also used the method proposed in experiment III to account for the unknown uncertainty in the soil depth by using the matrix of equation 6.3. The result of this experiment is given in the row marked Exp. IV.-3. The results presented were obtained with $W_{RP} = \{0.2, 0.1, 0.3, 0.1, 0.3\}$ and $W_{SE} = \{0.3, 0.1, 0.3, 0.1, 0.2\}$. In the third version of this experiment we again tried to compensate for the uncertainty in soil depth but matrix $W$ was re-deduced by simultaneous training for its elements and those of $W_{RP}$ and $W_{SE}$. The following values were used:

$$W = \begin{bmatrix} 0.5 & 0.15 & 0.1 \\ 0.3 & 0.7 & 0.3 \\ 0.2 & 0.15 & 0.6 \end{bmatrix}$$

$W_{RP} = \{0.1, 0.1, 0.2, 0.6, 0.0\}$

$W_{SE} = \{0.1, 0.1, 0.4, 0.3, 0.1\}$

The results of this experiment are shown in table 6.1 in the row marked with Exp. IV.-4.

### 6.4 Critical Discussion of Combination of Information for Risk of Desertification

As we mentioned before there are some methods for combining different rules the outputs of which are defined in the same frame or space. In the last section we incorporated the belief of the individual rules which lead to the same conclusions and considered it as the overall belief to that data class. For example we have 3 rules with output data class, “No to slight risk” in the group of rules regarding the risk of soil erosion. Summation of the belief of 3 rules is considered as the total belief to the class “No to slight risk”. The final classification was based on finding the class with the maximum belief. Although this interpretation of rules is applicable in many cases, another interpretation is also possible. For example, if we consider each rule as a piece of evidence which gives us some degree of belief to one of the classes (depending on its conditions), and also if all rules satisfy independence criteria, i.e. reliability of one rule does not affect the reliability of the other rules, then we can combine the beliefs of rules by the Dempster’s combination rule.
Based on the Dempster-Shafer theory framework, each item of evidence assigns mass functions to focal elements according to the amount of information it has. If any source can not assign a total mass of 1 to the various propositions, the remaining mass is assigned to the frame of discernment which is interpreted as the lack of information.

In this application each rule assigns a mass function to a proposition (data class) based on the evidence which is available in the fact part of the rule. If we consider each rule as an independent source then in this application each source has information only for one single data class or singleton. In this sense we may assign the remaining mass function to the frame of discernment as expressing the lack of information. However, if we had any rule which could not distinguish between a few classes it would be very easy to deal with by the Dempster-Shafer theory, because we can assign a mass function to a union of propositions (data classes here) by the Dempster-Shafer theory.

In the case when each rule assigns belief to singletons, we do not expect so much difference in comparison with the approach in which just we add the beliefs of the same data classes. However we run all the experiments reported earlier with the new strategy of combining different rules, the Dempster's combination rule.

Table 6.2 shows the result of this group of experiments using Dempster's combination rule. As it can be seen the number of correctly classified sites is not very different from the result of the previous group of experiments. All conditions of experiments are the same except in Exp IV.1 where we used $W_{RP} = \{0.1, 0.1, 0.4, 0.3, 0.1\}$ and $W_{SE} = \{0.1, 0.4, 0.2, 0.2, 0.1\}$ and also in Exp IV.3 where we used $W_{RP} = \{0.1, 0.1, 0.5, 0.3, 0.0\}$ and $W_{SE} = \{0.1, 0.1, 0.4, 0.4, 0.0\}$.

6.5 Conclusions

The problem of assessing the risk of desertification using four variables, slope, aspect, soil depth and rock permeability was discussed. By using some rules which were extracted from verbal rules given by the experts it was possible to find the natural regeneration potential and the risk of soil erosion. By some other rules, again extracted from verbal rules, assessments for the natural regeneration potential and the risk of soil erosion were used to assess the risk of desertification of the site under study. The problem is that the results of classification by the experts mostly disagree with those which are obtained by the expert rules.
2.4.2 Kriging Interpolation

Kriging is one of the linear estimation methods which is used especially in geostatistics. Kriging is synonymous with optimal prediction and is a Best Linear Unbiased Estimator (BLUE). In this approach we use linear combination of known samples to estimate the unknown at a point. So the estimated value \( \hat{v} \) at the point in question is given by

\[
\hat{v} = \sum_{i=1}^{n} w_i v_i
\]

where \( v_i \) are the values at the known points and \( w_i \) are the weights we wish to specify. The set of weights may change at different locations. The difference between the estimated value and the true value is defined as error or residue.

\[
Error = Residue = r_i = \hat{v}_i - v_i
\]

If the random function is stationary it can be shown that the expectation value of the errors will be 0 if \( \sum_{j=1}^{n} w_j = 1 \). This condition guarantees unbiasedness of the estimation. The Kriging method tries to minimise the variance of the error of estimation. If \( \sigma^2_R \) is the error variance of \( K \) estimates, we have [25]:

\[
\sigma^2_R = \hat{\sigma}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \hat{C}_{ij} - 2 \sum_{i=1}^{n} w_i \hat{C}_{io}
\] (2.6)
where $\sigma^2$ is the variance of each random variable $v_i$, $\tilde{C}_{ij}$ is the covariance of random variables $v_i$ and $v_j$, and $\tilde{C}_{i0}$ is the covariance of random variables at point $i$, $v_i$, and at the point in question, $\hat{v}$. In the above equation the variance of the modelled error is expressed as a function of $n$ unknown weights, $w_1, \ldots, w_n$, assuming the model is completely known. To minimise this error it is necessary to take the partial derivatives of the expression on the right hand side with respect to all variables, $w_1, \ldots, w_n$, and set them equal to 0. Solving this system of $n$ equations for the $n$ variables we obtain the solution of minimising the variance of the error. However, we have another constraint which says that the estimation must be unbiased so the summation of all weights must be equal to 1. Therefore the above mentioned system of equations is augmented and really we have $n$ variables and $n+1$ equations. Solving this problem is possible by using the Lagrange parameter technique [25].

In the Lagrange parameter technique the system is converted to an unconstrained system by adding another parameter, $\mu$, which is called Lagrange parameter.

$$\hat{\sigma}_e^2 = \sigma^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} - 2 \sum_{i=1}^{n} w_i \tilde{C}_{i0} + 2\mu \left( \sum_{i=1}^{n} w_i - 1 \right) \quad (2.7)$$

If we set equal to 0 the partial derivative of the above expression with respect to variable $\mu$, we obtain $\sum_{i=1}^{n} w_i = 1$, which is the unbiasedness condition. So we have $n+1$ equations and $n+1$ unknowns. This system can be solved easily.

By taking $n+1$ partial derivatives of the above equation and set them equal to 0 we have the following system of equations:

$$\begin{cases} 
\sum_{j=1}^{n} w_j \tilde{C}_{ij} + \mu = \tilde{C}_{i0} & \forall i = 1, \ldots, n \\
\sum_{i=1}^{n} w_i = 1 
\end{cases} \quad (2.8)$$

This system of equations is called the ordinary Kriging system and can be written in matrix form as:

$$\mathbf{C} \mathbf{W} = \mathbf{D}$$

$$\begin{bmatrix}
\tilde{C}_{11} & \cdots & \tilde{C}_{1n} & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\tilde{C}_{n1} & \cdots & \tilde{C}_{nn} & 1 \\
1 & \cdots & 1 & 0
\end{bmatrix} \begin{bmatrix}
w_1 \\
\vdots \\
w_n \\
\mu
\end{bmatrix} = \begin{bmatrix}
\tilde{C}_{10} \\
\vdots \\
\tilde{C}_{n0} \\
1
\end{bmatrix} \quad (2.9)$$

The solution of the above system is the set of weights which satisfy the unbiasedness condition and also minimise the error variance. It can be shown that this minimum is
\( \hat{\sigma}_H^2 = \sigma^2 - W^T D \). This minimum value of variance is called the ordinary Kriging variance, and sometimes is denoted by \( \sigma_{OK}^2 \).

The first step in Kriging interpolation is the selection of a model for representing the spatial continuity of data. The correlation coefficient (correlogram), \( \rho(h) \), covariance function, \( C(h) \) and the semivariogram, \( \gamma(h) \), where \( h \) is the vector distance, are used to represent the spatial continuity. The semivariogram is given by:

\[
\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j) | \|h_{ij}\| = h} (v_i - v_j)^2
\]  

(2.10)

where \( N(h) \) is the number of pairs of samples whose locations are separated by \( h \). All these statistics can be calculated from each other, for example \( \gamma(h) = C(0) - C(h) \). Given the covariance function or semivariogram it is easy to solve the ordinary kriging system of equations.

If the semivariogram is only a function of \( h \) the semivariogram is called isotropic. The semivariogram should have a property called conditional negative definiteness which allows the ordinary kriging system 2.9 to have a solution. In the literature some functions which have this property are suggested which are relevant for some applications [28, 41]. Some commonly used models are: the nugget effect, the exponential, the spherical, the circular, the Gaussian, the linear, the power, the logarithmic and the quadratic. From all these, we are most interested in the power model which is given by \( \gamma(h) = c_0 + c_1 h^A \). Comparing this function and equation 2.10 with equation 2.4 reveals that the power model is equivalent with the fractal model with parameter \( H = \lambda/2 \). In other words the semivariogram of fractal terrains can be represented by the power model.

More detailed discussion of the Kriging interpolation method is given in appendix B.

### 2.4.2.1 Fractional Brownian Motion and Semivariogram

As it was mentioned before, spatial continuity can be represented by the correlogram, the covariance and the variogram (see Appendix B). The variogram is equal to half of the moment of inertia of the scatterplot. It is given by the equation

\[
\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j) | \|h_{ij}\| = h} (v_i - v_j)^2
\]  

(2.11)
2.5. Slope and Aspect of the Terrain

where $h$ is the lag vector of certain size and certain direction. This equation can be represented also in another notation:

$$\gamma(\|\Delta \vec{x}\|) = \frac{1}{2N} \sum_{i}^N [Z_H(\vec{x}) - Z_H(\vec{x} + \|\Delta \vec{x}\|)]^2$$

(2.12)

Here $N$ is the number of pairs of samples with the given lag distance, $\|\Delta \vec{x}\|$ is the lag distance, $\vec{x}$ is a point in 2D Euclidean space and $Z_H(\vec{x})$ is the value of variable $Z_H$ at point $\vec{x}$. If we compare equation 2.4 with equation 2.12 we conclude that:

$$\gamma(\|\Delta \vec{x}\|) = C_1 \|\Delta \vec{x}\|^{2H}$$

(2.13)

In other words, modelling the terrain by fractional Brownian motion is equivalent to selecting a power model for the variogram of spatial continuity. There are some techniques for extracting parameter $H$ from the variogram [55, 84]. Although parameter range does not appear explicitly in this equation, this relationship between the variogram and the Hurst dimension is valid only for a limited range of lags.

2.5 Slope and Aspect of the Terrain

The derivation of the thematic derivates of elevation (slope, aspect, curvature etc) from the Digital Terrain Modelling (DTM) is studied under general geomorphometry. Slope and aspect can be defined by the plane tangent to the surface at the given point. Slope is the maximum rate of change in the altitude, and aspect is the compass direction of this maximum rate of change in the altitude. Since the terrain is presented digitally, we do not have the gradient of the surface at all points. If the DTM is represented by a Triangulated Irregular Network (TIN), the calculation of slope and aspect is very simple. First we need to find the triangle that covers a given point. Slope and aspect of the plane of the triangle are assigned to all its points. To find the slope and aspect of a plane it is necessary to find the normal vector of the plane. The normal vector can be easily calculated by the cross product of two vectors in the plane. Two edges of the triangle can be used for this purpose. If $\vec{N} = (A\vec{e}, B\vec{y}, C\vec{z})$ is the normal vector, the slope and the aspect both in degrees are (see figure 2.2 for the definition of slope and aspect):

$$\text{Slope} = \theta = \cos^{-1} \frac{|C|}{\sqrt{A^2 + B^2 + C^2}}$$

(2.14)

$$\left\{ \begin{array}{l}
\text{Aspect} = - \arctan(B, -A) \times \frac{180}{\pi} \\
\text{If Aspect} < 0 \rightarrow \text{Aspect} = \text{Aspect} + 360.
\end{array} \right.\quad (2.15)$$
It must be noted here that function \(\text{arctan}\) must compute the phase of a point in radians in a Cartesian coordinate system in the range \([-\pi, \pi]\), i.e. it must take into account the signs of the sine and cosine of the angle separately. It is assumed that the direction of the \(X\) axis is South and the direction of the \(Y\) axis is East. The value of aspect is chosen as shown in figure 2.3. If the DTM is represented by a grid, slope and aspect at a given point \((i, j)\), are calculated by using its neighbours [80]. The partial derivatives of the elevation with respect to \(X\) and \(Y\) and also with respect to the diagonal directions are taken by using the following formulae (see figure 2.4):

\[
\begin{align*}
\bar{S}_x &= \frac{Z(i-1, j) - Z(i+1, j)}{2\Delta h} \times (1, 0) \\
\bar{S}_y &= \frac{Z(i, j-1) - Z(i, j+1)}{2\Delta h} \times (0, 1) \\
\bar{S}_{d1} &= \frac{Z(i+1, j-1) - Z(i-1, j+1)}{2\sqrt{2}\Delta h} \times (1, 1) \\
\bar{S}_{d2} &= \frac{Z(i-1, j-1) - Z(i+1, j+1)}{2\sqrt{2}\Delta h} \times (1, -1) \\
\bar{S} &= (\bar{S}_x + \bar{S}_y + \bar{S}_{d1} + \bar{S}_{d2})/2 \\
slope &= |\bar{S}|
\end{align*}
\]

Aspect is defined as before, using the components of the vector of slope in formula 2.15. Here \(\bar{S}_{d1}\) and \(\bar{S}_{d2}\) are the slope vectors in the diagonal directions and \(\Delta h\) is the grid spacing. In this technique eight immediate neighbours are used. Some people have used four immediate neighbours in the \(X-Y\) or in the diagonal directions. If we have no data for all neighbours,
we copy the values of given points to points where data are not available. This happens at
the boundary points of the convex hull of a region of interest.

2.6 Error Assessment

One of the important issues in interpolation is accuracy. Accuracy can be defined as the
closeness of the results or estimates to the true value or the values that are considered to be
true. The accuracy of a DEM may be expressed by two factors: the average difference of the
value of DEM from the real surface, or the distribution of the errors. So the accuracy of a
DEM depends on the density and distribution of the data points as well as on the roughness
of the terrain surface.

Since we do not have enough real data with different degrees of roughness to compare with
the result of interpolation, we shall construct a series of fractal surfaces and treat them
as if they were true terrains. It has been shown that such simulations can reproduce very
realistic looking terrains [52, 78].

The artificially constructed terrains will be subsampled and interpolated either by Delaunay
triangulation or Kriging and the errors in calculating the slope and aspect will be assessed.
The fractals used for modelling are created by the Fourier transform method but fractals
used for validation are generated by midpoint displacement with successive random additions
method as well [78]. Parameters which produce different kinds of fractals are the fractal
dimension $D$ and the seed of the random number generator used. Parameter $\sigma$ of the fractal
only controls the amplitude of the roughness. Changing this parameter has no effect on
the aspect but it changes the magnitude of the slope. If slope is measured in percent,
changing parameter $\sigma$ by a factor $k$ will change the magnitude of slope by $k$ too. So this
parameter will be used just to decide whether a fractal is scaled relatively to the training
fractals which are used for modelling. The “terrain” produced is subsampled randomly
with uniform distribution. The percentage of retained points and the seed of the random
number generator used for subsampling are two parameters that may create different data
configurations. For all our experiments fractals of size 128 $\times$ 128 pixels were generated.
Our goal was to find the effect of the roughness of the terrain, expressed by the fractal
dimension $D$, and the density of samples on the accuracy of the calculation of the slope
and aspect of the terrain when it is interpolated by Delaunay triangulation and Kriging.
Fractals with different degrees of roughness but with the same variance of random number
generator were created. A fractal generation algorithm almost never produces a fractal with exactly the desired fractal dimension. So, a large number of fractal terrains was produced and the fractal dimension of each one was carefully measured. We disregarded almost all generated terrains except 60 of them with fractal dimensions 2.07, 2.11, 2.20, 2.29, 2.41 and 2.48. Ten terrains from each category were used. A "terrain" with fractal dimension 2.5 is a limiting case of a totally uncorrelated surface. These terrains were subsampled uniformly with different rates, namely 3%, 5%, 10%, 20%, 30% and 40%. Terrains of the same fractal dimension were randomly sampled with different seeds for the random number generator in order to avoid the repeated selection of the same points. After interpolation, the slope and aspect of the triangulated surface were derived by using the 8 neighbours method and compared with the slope and aspect of the original surface. The maximum, the mean and the standard deviation of the error and absolute error were computed. The mean error distribution over all terrains with the same fractal dimension is adopted as the error model appropriate for the particular surface roughness and sampling rate. Some results are plotted in figures 2.5 and 2.6 for both Delaunay and Kriging methods. In figure 2.5a the mean of the absolute error in slope by the Delaunay triangulation method is plotted versus sampling rate and fractal dimension. Figures 2.5b to 2.5e show other statistics of the error in slope as a function of the fractal dimension. In figures 2.5f and 2.6f the histograms of errors in slope and aspect are plotted respectively. In figure 2.6a the standard deviation of aspect by the Kriging interpolation method is plotted versus sampling rate and fractal dimension. Some statistics of the error in aspect are shown in figures 2.6b to 2.6e. The mean error in aspect is nearly zero and the maximum error in aspect is nearly 180 degrees so these are not shown.

By using the results of this experiment one may predict the statistics of errors in slope and aspect, given the surface roughness (quantified by the fractal dimension and the variance of the random process - the fractal modelling parameters) and the percentage of the retained data. As it can be predicted our results show that the errors in slope and aspect depend on both these quantities. In other words error statistics are bivariate functions. Figures 2.5a and 2.6a show this pictorially. It would be more useful if we could express this dependency by some empirical formulae to avoid having to use look up tables or diagrams for different combinations of sampling rates and degrees of roughness. We chose two methods to identify such empirical formulae.

From plots of each statistic versus one of the two variables while the other variable is kept
2.6. Error Assessment

Figure 2.5: (a) Mean of absolute error in slope (in percent) versus sampling rate and fractal dimension by Delaunay triangulation method (b)-(e) some statistics of error in slope (in percent) versus fractal dimension for 5% sampling rate (f) Histogram of the error in slope for 5% sampling rate and $D=2.2$; the bin width is 0.229.
Figure 2.6: (a) Standard deviation of error in aspect (in degrees) versus fractal dimension and sampling rate by Kriging method (b)-(e) some statistics of error in aspect versus fractal dimension for 5% sampling rate (f) Histogram of the error in aspect for 5% sampling rate and $D = 2.2$; the bin width is 5.54°.
2.6. Error Assessment

Figure 2.7: (a) Standard deviation of error in slope and fitting curve (first method) versus fractal dimension for 5% sampling rate (b) Standard deviation of error in slope and fitting curve (first method) versus sampling rate for $D=2.1$ (c) Standard deviation of error in aspect and fitting curve (first method) versus fractal dimension for 5% sampling rate (d) Standard deviation of error in aspect and fitting curve (first method) versus sampling rate for $D=2.1$. 
Figure 2.8: (a) Standard deviation of error in slope and fitting curve (second method) versus fractal dimension for 5% sampling rate (b) Standard deviation of error in slope and fitting curve (second method) versus sampling rate for $D=2.1$ (c) Standard deviation of error in aspect and fitting curve (second method) versus fractal dimension for 5% sampling rate (d) Standard deviation of error in aspect and fitting curve (second method) versus sampling rate for $D=2.1$. 
fixed, one can guess that all statistics have exponential dependence on each variable. The general form of an exponential function can be written as \( y = c(1 - e^{ax}) \). We assumed that the dependence of each statistic on the two functions is separable, i.e. each statistic is assumed to be the product of two such exponential functions, one of which depends on the fractal dimension \( D \) and the other one on the sampling rate \( P \). The function of each statistic with respect to each variable was fitted separately by the least square error method and the two fitting functions were multiplied to produce the function that would fit the whole 2D surface as a function of the fractal dimension and rate of sampling. Some of the results of the empirical formulae derived are plotted in figures 2.7a to 2.7d. The fitting process was only applied for fractal dimensions in the range \( 2.0 \leq D \leq 2.5 \) and sampling rates from 3\% to 40\% because this range is more relevant to applications. Extracted empirical formulae are given in table 2.1. In the given formulae \( D \) is the fractal dimension and \( P \) is the fraction of retained pixels after subsampling. \( K \) is the scaling factor for the slope. This has to be determined for each set of data separately as follows: Given the data from an unknown terrain for which we wish to identify the error distributions in slope and aspect, we should plot first the logarithm of semivariogram, \( \log(\gamma(h)) \), versus logarithm of distance, \( \log(h) \), in linear axes. The slope of the regression line is equal to \( 2H \) and the intersection with the vertical axis is equal to \( 2 \log(\sigma) \). The fractal dimension \( D \) is \( D = 3-H \). Using this value for \( D \) we read the value of \( \sigma_{ref} \) from the look up table given in table 2.2. If there is no entry in the table with exactly the same \( D \), we use linear interpolation to estimate \( \log(\sigma_{ref}) \). Then we can calculate the scaling factor as \( K = \log^{-1}(\log(\sigma) - \log(\sigma_{ref})) \). Figure 2.9a shows a sampled fractal and figure 2.9b shows the plot of its semivariogram and extracted fractal features.

In the second method for deriving empirical formulae, we choose polynomial functions of degree 3 instead of the exponential function. A polynomial function \( y = ax^3 + bx^2 + cx + d \) which has four coefficients is more flexible than the exponential function which has two coefficients to fit curves. In this case we do not assume that the formula is separable in its dependence on the two variables \( D \) and \( P \).

\[
F(D, P) = a_1 D^3 P^3 + a_2 D^3 P^2 + a_3 D^3 P + a_4 D^3 + a_5 D^2 P^3 + a_6 D^2 P^2 + a_7 D^2 P + a_8 D^2 + a_9 DP^3 + a_{10} DP^2 + a_{11} DP + a_{12} D + a_{13} P^3 + a_{14} P^2 + a_{15} P + a_{16}
\]  

(2.21)

where \( F(D, P) \) is the statistic of error distribution we are concerned with, \( D \) is the fractal dimension of the terrain, and \( P \) is the sampling rate. For 6 different sampling rates and
### Chapter 2. Error Modelling of Slope and Aspect

#### Table 2.1: Empirical formulae for error ($\varepsilon$) in slope and aspect. In these formulae $0 \leq P \leq 0.4$ and $2 \leq D \leq 2.5$ and $K = \log^{-1}(\log(\sigma) - \log(\sigma_{ref}))$. 

<table>
<thead>
<tr>
<th>Variable</th>
<th>Method</th>
<th>Empirical formulae</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Slope</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt; \varepsilon &gt;$</td>
<td>Delaunay tri.</td>
<td>$2.13K(e^{4.35(D-2)} - 0.75)(e^{-10.64P} + 0.29)$</td>
</tr>
<tr>
<td>$&lt; \varepsilon &gt;$</td>
<td>Kriging</td>
<td>$1.84K(e^{5.26(D-2)} - 0.9)(e^{-8.9P} + 0.13)$</td>
</tr>
<tr>
<td>$\sigma = &lt; (\varepsilon - \varepsilon)^2 &gt;$</td>
<td>Delaunay tri.</td>
<td>$10.7K(e^{2.1(D-2)} - 0.88)(e^{-9.9P} + 0.45)$</td>
</tr>
<tr>
<td>$\sigma = &lt; (\varepsilon - \varepsilon)^2 &gt;$</td>
<td>Kriging</td>
<td>$9.3K(e^{2.1(D-2)} - 0.95)(e^{-6.67P} + 0.39)$</td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
</tr>
<tr>
<td>$\sigma = &lt; (</td>
<td>\varepsilon</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma = &lt; (</td>
<td>\varepsilon</td>
<td>-</td>
</tr>
<tr>
<td><strong>Aspect</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
</tr>
<tr>
<td>$\sigma = &lt; (</td>
<td>\varepsilon</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma = &lt; (</td>
<td>\varepsilon</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma = &lt; (\varepsilon - \varepsilon)^2 &gt;$</td>
<td>Delaunay tri.</td>
<td>$63.7(1 - e^{-6.4(D-2)})(e^{-5.9P} + 0.4)$</td>
</tr>
<tr>
<td>$\sigma = &lt; (\varepsilon - \varepsilon)^2 &gt;$</td>
<td>Kriging</td>
<td>$63.8(1 - e^{-4.65(D-2)})(e^{-5.93P} + 0.4)$</td>
</tr>
</tbody>
</table>
2.6. Error Assessment

<table>
<thead>
<tr>
<th>$D_{ref}$</th>
<th>2.07</th>
<th>2.11</th>
<th>2.20</th>
<th>2.29</th>
<th>2.41</th>
<th>2.48</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2\log_{10}(\sigma_{ref})$</td>
<td>-4.44</td>
<td>-4.16</td>
<td>-1.84</td>
<td>-1.55</td>
<td>-1.10</td>
<td>-0.84</td>
</tr>
</tbody>
</table>

Table 2.2: Fractal features of fractals which are used in this research.

Figure 2.9: (a) A subsampled fractal; 20% of the original data retained (b) The fractal features of the surface of figure (a) can be calculated from this line as: $H=0.82, 2\log \sigma = -1.91$.

6 degrees of roughness we have data for 36 combinations of $D$ and $P$ which should satisfy equation 2.21. In other words we have 36 equations and 16 unknown variables (coefficients). In matrix form we can write these equations as:

$$ A \times X = B \quad (2.22) $$

where $X^T = (a_1, a_2, \ldots, a_{16})$ is the vector of unknown coefficients, $B$ is the 36 x 1 vector of the known values of the function for different values of $D$ and $P$, and $A$ is a 36 x 16 matrix of the known values of $p^3D^3, p^3D^2, \ldots, 1$ for different values of $D$ and $P$. Since $A$ is not square we first multiply both sides of equation 2.22 by $A^T$. So, the unknown coefficients are given by:

$$ X = (A^T A)^{-1} \times A^T B \quad (2.23) $$

The least square error solution is computed and the results are listed in table 2.3. Columns 1 to 4 refer to the statistics of errors in slope, i.e. mean and standard deviation of absolute error and mean and standard deviation of error respectively, and columns 5 to 7 refer to the statistics of error in aspect, i.e. mean and standard deviation of absolute error and standard deviation of error respectively. Mean of error in aspect is not included in this
table because in all cases this statistic is near zero irrespectively of the sampling rate and fractal dimension. In figures 2.8a to 2.8d real curves and fitted curves by this method are plotted for some functions. These plots can be compared with plots which were extracted by the first method. Polynomial surface (curve) fitting is much more accurate than by the first method, however the dependency of each statistic on each of the two variables is not as clear as in the case of exponential fitting.

2.7 Distribution of Errors in Slope and Aspect

The error models we derived in the previous section refer only to the first and second order statistics of the error distributions. The question then arises about the most appropriate function that can model these error distributions. Here we investigate this issue with the help of an artificial fractal and a real terrain. The distribution of errors in slope and aspect of a fractal can be inferred from the histograms of figures 2.5f and 2.6f. We fitted some well known error distribution functions to the distributions of all training fractals by using mean least square error method. We found that the distribution of error in slope can be best represented by the Gaussian function, \( G(\sigma, \mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \). However the distribution of error in aspect is best modelled by a Laplace function, \( L(\sigma, \mu) = \frac{1}{\sqrt{2\sigma}} e^{-\frac{|x-\mu|}{\sigma}} \). In figures 2.10 and 2.11 the PDF of error in slope and aspect and the corresponding fitting functions are shown. The distribution of errors in aspect for real data is the same as for the artificial data, i.e. the best fitting function is the Laplacian. However, the distribution of errors in slope for real data depends on how well the data can be modelled by a fractal. If the given terrain exhibits to some extent fractal properties, the distribution of errors in slope will be the same as for the artificial data, i.e. a Gaussian function. To check how fractal a terrain is, we plot the distribution of increments, \( |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})| \), for a few values of distance \( |\Delta \bar{x}| \). In a fractal this is a Gaussian function. The approximation of the distribution of errors in slope by a Gaussian function is as good as the approximation of the distribution of increments by a Gaussian function. In other words in the process of extracting the fractal features of a terrain, we must plot the histogram of increments for some distances and compare it with the Gaussian and Laplace function, in order to identify which of these two functions is to be used for the distribution of errors in slope. In figures 2.12 and 2.13 the PDFs of errors in slope in two cases of real terrains are shown: one when the distribution is best approximated by the Gaussian function and one when it is best approximated by the
### 2.7. Distribution of Errors in Slope and Aspect

Table 2.3: Coefficients of polynomials which are used to fit the error curves of the Delaunay and Kriging methods respectively. Columns 1 to 4 are statistics of error in slope (mean and SD of absolute error and mean and SD of error respectively) and columns 5 to 7 are statistics of error in aspect (mean and SD of absolute error and SD of error respectively).

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Slope</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$a_2$</td>
</tr>
<tr>
<td></td>
<td>$a_8$</td>
<td>$a_9$</td>
</tr>
<tr>
<td></td>
<td>$a_{17}$</td>
<td>$a_{18}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Slope</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_{26}$</td>
<td>$a_{27}$</td>
</tr>
</tbody>
</table>

Table 2.3: Coefficients of polynomials which are used to fit the error curves of the Delaunay and Kriging methods respectively. Columns 1 to 4 are statistics of error in slope (mean and SD of absolute error and mean and SD of error respectively) and columns 5 to 7 are statistics of error in aspect (mean and SD of absolute error and SD of error respectively).
Figure 2.10: PDF of error in slope and its fitting function (Gaussian) for an artificial terrain.

Figure 2.11: PDF of error in aspect and its fitting function (Laplacian) for an artificial terrain.

Figure 2.12: PDF of error in slope and its fitting function (Laplacian) for a real terrain.

Figure 2.13: PDF of error in slope and two fitting functions (Gaussian and Laplacian) for a real terrain.

Laplacian. In figures 2.14 and 2.15 we show the corresponding histograms of increments for $|\Delta z| = 8$ for the two terrains.

2.8 Evaluation of the Error Models

In this section we investigate the validity of the error models which are proposed in this thesis. We do this in two ways: using artificial terrains and real terrains.

First we generated fractal terrains using the Fourier transform and the midpoint displacement with random additions method. These terrains are subsampled uniformly with different sampling rates. Fractal features are extracted from the subsampled data. During this
Figure 2.14: Histogram of increments (solid line) and the Gaussian fitting function (dashed line) for $|\Delta z| = 8$ and 10 bins. Vertical axis is the number of pairs which fall into bins.

Figure 2.15: Histogram of increments (solid line) and the Laplacian fitting function (dashed line) for $|\Delta z| = 8$ and 10 bins. Vertical axis is the number of pairs which fall into bins.

process the approximation of the distribution of increments by the Laplacian and Gaussian functions is measured by using the mean least square error criteria. Using the extracted fractal features and with the help of look up table 2.2 we calculate the scaling factor for errors in slope. Results of over 100 fractals which were subsampled with different sampling rates show that our models can precisely predict the statistics of the validation fractals. As we extracted the models from fractals which were generated by the Fourier transform method and we tested with the fractals which were generated by the midpoint displacement method, we conclude that the models derived do not depend on the method of fractal terrain generation. Figures 2.16 and 2.17 show the standard deviation of errors in slope and aspect respectively over 30 of the 100 artificial terrains used for validation, when the terrains were subsampled to keep only 5% of the original number of points. The continuous curves in these graphs represent the values predicted by the error models derived, for different combinations of $D$ and $P$. As it can be seen, the predictions of the error models agree well with the results of the experiments. Similarly good agreement could be obtained with other sampling rates as well.

In addition, we selected 80 DEMs of size $128 \times 128$ from data which are available from the USGA homepage [22]. The space between pixels is $1/400$ arc-second which is equal to approximately 70 meters on the ground. We followed the same process as above, i.e. we subsampled the real terrains with different sampling rates and interpolated them by the Delaunay triangulation and Kriging methods. The errors in slope and aspect and their
Figure 2.16: Standard deviation of error in slope versus fractal dimension for different sampling rates of the terrain. The open circles are the experimental results of 5% sampling rate of validation fractals.

Figure 2.17: Standard deviation of error in aspect versus fractal dimension for different sampling rates of the terrain. The open circles are the experimental results of 5% sampling rate of validation fractals.

distributions were computed. Figures 2.18 and 2.19 show all statistics of error in slope and aspect by the Delaunay triangulation method, while the model curves are also plotted. Figures 2.20 and 2.21 show similar results when the Kriging method is used. Real terrains were sampled by 5% sampling rate and after interpolation the statistics were derived. As it can be seen statistics of error in slope are in agreement with the model, except the mean of error in slope which is less than what the model predicts. Also, all statistics of error in aspect are less than what the model predicts. Actually in these cases the error for 5% sampling rate coincides with the error predicted by the model for 10% sampling rate. This overestimation can be seen in other sampling rates as well, i.e. the error for 10% sampling rate coincides with 20% sampling rate model. So, there seems to be a bias in the prediction. We can account for that when using the model formulae. For example, we should substitute $P = 0.1$ instead of $P = 0.05$ and $P = 0.2$ instead of $P = 0.1$ etc. in the formulae in order to predict the errors due to 5% or 10% subsampling respectively. In table 2.4 we summarise some statistics of the results over all evaluation experiments we performed for 5% sampling rate and for both interpolation methods. The statistics concern the relative difference between the experimental result and that predicted by the empirical formula, expressed as a percentage. Mean and standard deviation of difference, which indicate precision of the proposed formulae, are given in the table for both methods. As it can be seen the worst case is mean error in slope which is as high as 92% for the kriging method. Actually in this
case the error for 5% sampling rate coincides with the model for 20% sampling rate. It is interesting to observe that in general our empirical formulae for the errors produced by the Kriging method are worse than our empirical formulae that concern Delaunay triangulation. It is worth noting that proposed formulae are perfect for the fractal terrains or real terrains which behave like fractals. However in the real data that we have used there are many small patches which are very smooth or even absolutely flat. Fractal surfaces do not contain absolutely flat patches.

Indeed, these are limitations in the use of fractal terrain models, due to the inadequacy of the measuring devices to capture the fractality of real terrains to the full: every measuring instrument, and every representation has some finite resolution. Although real terrains show variation down to the molecular level, measuring methods may be able to differentiate heights only if they are more than a minimum threshold (which could be of the order of meters), either because the user is not interested in higher accuracy, or because the instrument used may not be able to have higher resolution, or even because the digitisation process used may quantise the values to a limited number of levels. This will result in large areas represented by absolutely flat patches, having smoothed out all variations at scales smaller than the scales of the limiting factors mentioned. On the other hand, fractal models do not exhibit this property as they are supposed to show variation at all scales.

Figure 2.22 shows one of these terrains and figure 2.23 shows the estimation of its fractal features. Figure 2.24 shows the same fractal when subsampled with 5% sampling rate and figure 2.25 shows its fractal features extracted from the subsampled terrain. From the extracted fractal features and using the look up table we can calculate $K = 1.12$ which should be used for the prediction of error in slope. In figure 2.26 we show the interpolated terrain by the Delaunay triangulation method and in figure 2.27 the error surface in the estimation of elevation are shown. In figures 2.28 and 2.29 the error in slope and aspect of every individual pixel is shown. The PDF of error in slope and aspect and the best fitting functions are plotted in figures 2.30 and 2.31 respectively. All statistics of the errors for this terrain using only 5% of the data ($P = 0.05$) are shown in table 2.5. In the second column the results from the proposed formulae are given and in the third column the experimental results by using the Delaunay triangulation method are given. It can be seen how the experimental results agree with the model. It is worth noting that in the calculation of errors in aspect we took into consideration the bias of our models as explained earlier. In figures 2.32 to 2.41 similar results are shown for another real terrain and using only 3%
Figure 2.18: The mean of the absolute error in (a) slope (b) aspect, and the standard deviation of the absolute error in (c) slope (d) aspect for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Delaunay triangulation. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line.
Figure 2.19: The standard deviation of the absolute error in (a) slope (b) aspect, and the mean of the absolute error in (c) slope for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Delaunay triangulation. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line.
Figure 2.20: The mean of the absolute error in (a) slope (b) aspect, and the standard deviation of
the absolute error in (c) slope (d) aspect for 80 real terrains, as a function of the fractal dimension of the
terrain, for 5% sampling rates, and Kriging. The lines have been computed from the empirical formulae
proposed. The 5% sampling line is the solid darkest line.
2.8. Evaluation of the Error Models

![Graphs](image)

**Figure 2.21:** The standard deviation of the absolute error in (a) slope (b) aspect, and the mean of the absolute error in (c) slope for 80 real terrains, as a function of the fractal dimension of the terrain, for 5% sampling rates, and Kriging. The lines have been computed from the empirical formulae proposed. The 5% sampling line is the solid darkest line.
Chapter 2. Error Modelling of Slope and Aspect

|       | Method       | \langle |\epsilon| \rangle | \sigma = \langle (|\epsilon| - |\bar{\epsilon}|)^2 \rangle | \langle \epsilon \rangle | \sigma = \langle (\epsilon - \bar{\epsilon})^2 \rangle |
|-------|--------------|----------------|---------------------------------|----------------|---------------------------------|
| Mean of difference | Delaunay | -23.56% | -2.48% | -38.37% | -5.63% |
| SD of difference    | Delaunay   | 9.57%   | 8.33%  | 34.00%  | 7.71%  |
| Mean of difference   | Kriging    | -21.50% | -4.28% | -92.93% | -4.05% |
| SD of difference     | Kriging    | 9.36%   | 8.86%  | 86.20%  | 7.74%  |

|       | Method       | \langle |\epsilon| \rangle | \sigma = \langle (|\epsilon| - |\bar{\epsilon}|)^2 \rangle | \sigma = \langle (\epsilon - \bar{\epsilon})^2 \rangle |
|-------|--------------|----------------|---------------------------------|---------------------------------|
| Mean of difference | Delaunay | 0.94% | 1.59% | -0.57% |
| SD of difference    | Delaunay   | 8.68%   | 6.28%  | 7.27%  |
| Mean of difference   | Kriging    | -10.94% | -6.79% | -7.38% |
| SD of difference     | Kriging    | 9.40%   | 10.26% | 9.6%   |

Table 2.4: Statistics of the absolute relative difference between the errors predicted by the empirical formula and those computed directly from real data, computed over 100 different experiments of terrain subsampling, interpolation, slope and aspect computation, and error estimation.

of the data when the Kriging method is used for interpolation. All statistics of the errors of this terrain using only 3% of the data \((P = 0.03)\) are shown in table 2.6. From the fractal features extracted from the subsampled data and by using the look up table we can calculate \(K = 1.04\) which is used for the prediction of the errors in slope by using the proposed empirical formulae.

2.9 Conclusions and Discussion

In this chapter we derived empirical formulae for the mean and variance of model error distributions of slope and aspect of a terrain as functions of the roughness of the terrain and the rate of subsampling.

The roughness of the terrain is expressed by its fractal dimension. The fractal dimension of a real terrain can be easily computed from its variogram by following, for example, the method of Arakawa and Krotkov [1]: We consider all pairs of points in the terrain. We find the difference in height between them and their distance from each other. We find the average difference in height for a given distance and plot the logarithm of them against each
2.9. Conclusions and Discussion

Figure 2.22: A real terrain. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. The same scale is used for the vertical axis too.

Figure 2.23: The fractal features of the real terrain of figure 2.22 can be calculated from this line as: $H=0.75$, $2\log \sigma = -1.63$.

Figure 2.24: The terrain of figure 2.22 when subsampled with 5% sampling rate. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

Figure 2.25: The fractal features of the subsampled terrain in figure 2.24 can be calculated from this line as: $H=0.78$, $2\log \sigma = -1.71$. 
**Figure 2.26:** Interpolated terrain of figure 2.24 by the Delaunay triangulation method. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

**Figure 2.27:** Error in the elevation of the interpolated terrain shown in figure 2.26. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

**Figure 2.28:** Error in slope (in percent) of the interpolated terrain in figure 2.24. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent slope in percent.

**Figure 2.29:** Error in aspect (in degrees) of the interpolated terrain in figure 2.24. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent aspect in degrees.
2.9. Conclusions and Discussion

Figure 2.30: The PDF of errors in slope and its fitting function (Gaussian) for the subsampled terrain of figure 2.24.

Figure 2.31: The PDF of errors in aspect and its fitting function (Laplacian) for the subsampled terrain of figure 2.24.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Empirical formulae</th>
<th>Experimental results</th>
<th>Error in prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt;\varepsilon&gt;$</td>
<td>3.85</td>
<td>2.72</td>
<td>41.5%</td>
</tr>
<tr>
<td>$\sigma = &lt;(\varepsilon - \bar{\varepsilon})^2&gt;$</td>
<td>8.92</td>
<td>9.06</td>
<td>1.6%</td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
<td>7.75</td>
</tr>
<tr>
<td>$\sigma = &lt;</td>
<td>\varepsilon</td>
<td>-</td>
<td>\bar{\varepsilon}</td>
</tr>
<tr>
<td>Aspect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
<td>46.56</td>
</tr>
<tr>
<td>$\sigma = &lt;</td>
<td>\varepsilon</td>
<td>-</td>
<td>\bar{\varepsilon}</td>
</tr>
<tr>
<td>$\sigma = (\varepsilon - \bar{\varepsilon})^2$</td>
<td>34.7</td>
<td>32.69</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

Table 2.5: Experimental results and predicted statistics of error ($\varepsilon$) in slope (in percent) and aspect (in degrees) by using the empirical formulae for the terrain of figure 2.22.
Figure 2.32: A real terrain. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters. The same scale is used for the vertical axis too.

Figure 2.33: The fractal features of real terrain of figure 2.32 can be calculated from this line as: $H=0.78, 2\log \sigma = -1.71$.

Figure 2.34: The terrain of figure 2.32 when subsampled with 3% sampling rate. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

Figure 2.35: The fractal features of the subsampled terrain of figure 2.34 can be calculated from this line as: $H=0.80, 2\log \sigma = -1.80$. 
2.9. Conclusions and Discussion

Figure 2.36: Interpolated terrain of figure 2.34 by the Kriging method. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

Figure 2.37: Error in the elevation of the interpolated terrain shown in figure 2.36. All numbers are in pixels. An inter-pixel distance is approximately equal to 70 meters.

Figure 2.38: Error in slope (in percent) of the interpolated terrain in figure 2.34. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent slope in percent.

Figure 2.39: Error in aspect (in degrees) of interpolated terrain in figure 2.34. The numbers along the horizontal axes are in pixels. An inter-pixel distance is approximately equal to 70 meters. The numbers along the vertical axis represent aspect in degrees.
Figure 2.40: The PDF of errors in slope and its fitting function (Gaussian) for the subsampled terrain of figure 2.34.

Figure 2.41: The PDF of errors in aspect and its fitting function (Laplacian) for the subsampled terrain of figure 2.34.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Empirical formulae</th>
<th>Experimental results</th>
<th>Error in prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slope</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt;\varepsilon&gt;$</td>
<td>1.22</td>
<td>3.36</td>
<td>63.7%</td>
</tr>
<tr>
<td>$\sigma = &lt;(\varepsilon - \bar{\varepsilon})^2&gt;$</td>
<td>8.92</td>
<td>9.06</td>
<td>1.6%</td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
<td>7.22</td>
</tr>
<tr>
<td>$\sigma = &lt;</td>
<td>\varepsilon</td>
<td>-</td>
<td>\bar{\varepsilon}</td>
</tr>
<tr>
<td>Aspect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$&lt;</td>
<td>\varepsilon</td>
<td>&gt;$</td>
<td>25.15</td>
</tr>
<tr>
<td>$\sigma = &lt;</td>
<td>\varepsilon</td>
<td>-</td>
<td>\bar{\varepsilon}</td>
</tr>
<tr>
<td>$\sigma = &lt;(\varepsilon - \bar{\varepsilon})^2&gt;$</td>
<td>39.39</td>
<td>37.61</td>
<td>4.7%</td>
</tr>
</tbody>
</table>

Table 2.6: Experimental results and predicted statistics of error ($\varepsilon$) in slope (in percent) and aspect (in degrees) by using the empirical formulae for the terrain of fig. 2.32.
other in a linear coordinate system. The fractal dimension can be derived from the slope of the linear part of the curve. It is possible that one knows a priori the fractal dimension that characterises a region: for example, valleys and plains have fractal dimensions lower than weathered mountains and weathered mountains have fractal dimensions lower than new mountains. In that case an estimate of the expected errors can be obtained without plotting the variogram.

The rate of sampling is related to the resolution with which the original data were collected and the resolution with which the data are presented in the GIS. For example, if geological data were collected with resolution of 100 x 100 m\(^2\), but for the purpose of combining them with TM Landsat images they were interpolated to a resolution of 30 x 30 m\(^2\), the rate of subsampling is 0.09, i.e. we have only 9% of the data we ought to have had given that the conclusions we wish to draw concern resolution of 30 x 30 m\(^2\).

One can use then this number in conjunction with the fractal dimension of the terrain to predict the error in slope and aspect as computed after the interpolation process has taken place. It is these error distributions that have to be propagated in a decision support system combining information from the input causal variables to draw accurate conclusions on possible effects.

Limitations in the resolution with which elevation data are collected and represented in digital form, result in large flat regions which are better modelled by flat triangular patches than by the more realistic continuous variation models. These flat patches are probably responsible for some systematic difference observed between the errors predicted by our formulae and those calculated. This bias can be taken into consideration in using the formulae by increasing the sampling rate by 5% over its true value.
Chapter 3

Remote Sensing Data Fusion: A Literature Survey

In this chapter we survey different approaches which have been used for integrating and combining multi source remote sensing data. Generally we discuss different techniques of information fusion especially applicable to remote sensing. Among different approaches we survey Dempster-Shafer evidence theory approach, probabilistic approach, fuzzy logic approach, neural network approach, Bayesian network approach and rule based inference systems. First we discuss different levels of fusion and general categorisation of multi source integration in section 3.1. In sections 3.2 and 3.3 we review two data fusion approaches at decision and feature level. In section 3.4 we review different techniques which have been used in this field. In section 3.5 we discuss two major projects which have been carried out in the past on the problem that we are studying. Conclusions are given in section 3.6.

3.1 Introduction

In recent years with improvement of technology, scientists have access to high quality and various remote sensing data. The need to combine various sources which some times are incomplete or inaccurate is increasing continuously. Apart from different approaches which are employed to combine multi source / multi sensor remote sensing data, it is possible to categorise them into 3 groups based on the definition of Benediktsson and Landgrebe [3]: data level fusion, feature level fusion and decision level fusion. In the data level fusion the raw data acquired from sources are combined directly. In the applications which
are related to remote sensing image fusion, this type of fusion sometimes is called fusion at pixel or measurement level. In the feature level fusion, feature vectors extracted from different sources are the subject of fusion. Sometimes this type of fusion in the applications related to remote sensing image fusion is called fusion at attribute level. In the decision level each source acts as a classifier and can assign a label from different possible labels to the outputs. Then the outputs of sources, which can be treated as classifiers, are combined. In remote sensing image data fusion, sometimes this level is called symbol level fusion in which higher level of information are combined [75]. Different fusion methods which we survey in this chapter cover all three levels.

From another point of view, information fusion systems can be categorised into two groups: distributed systems and centralised systems. In distributed approaches data acquired from different sources are processed independently and the results of their decisions are combined. Based on Benediktsson and Landgrebe’s definition these approaches are equivalent to fusion at decision level. Classifier combination problems are distributed systems when we look at a classifier as a source of information. In the centralised systems data acquired from sources are used with or without further processing to produce the final result. Unlike the distributed systems, different sources are not able to make any decisions. These approaches are fusion at data and feature level according to the previous definitions. Most traditional approaches in multi source data fusion belong to this category.

Although information extraction can be done by multi spectral measurements from a single source, it is often preferred to use multi source measurements. In some applications where information is extracted from a single sensor (although in different wave bands) the data may be incomplete and this may lead to misclassification. Further, measurements that are made by different sensors for the same site are partially redundant and they may include some imprecision because of the acquisition device. Therefore, the main task of data fusion is combining multi source and/or multi sensor information to reduce uncertainty and incompleteness [8, 39, 42]. In all cases precision of classification should be improved in combination, over the use of one data source only. Some data that are often combined with remotely sensed data are geological and geophysical data. There are many methodologies for integration of remote sensing, geological and geophysical data. Each method models the degree of belief (usually a number in the interval [0, 1]) according to its own mathematical framework. In methods which are based on probability and Bayesian theory, the degree of belief is a conditional probability, in fuzzy set methods the degree of belief is a membership
function, in the MYCIN system it is the certainty factor and in Dempster-Shafer theory it is a belief function [8]. Also, other methods such as neural networks, Bayesian networks and fuzzy neural networks have been used in this area [60]. Depending on the problem at hand, one of the systems may be preferred.

Petrou and Stassopoulou [54] surveyed several data fusion approaches in remote sensing. Especially they reviewed the rule based inference approach, fuzzy logic-based systems, Dempster-Shafer evidence theory approach, Bayesian Network approach, probabilistic approach and neural network approach. They reported results of many researches in remote sensing data fusion. They separately discussed various approaches which fall into the distributed and centralised systems. They concluded that all methods that combine information from different approaches perform better compared to individual source classification.

3.2 Decision Level Fusion

Mason et al. [43, 13, 81] developed a novel system which they called multi sensor image processor (MuSIP). This system was designed to fuse and analyse multi source/multi sensor data with other ancillary data like maps, ground data etc. This system was aimed to be used in medical image processing and remote sensing applications like monitoring forestry by using datasets from radar, maps and ground data. MuSIP is a knowledge based system which includes four types of knowledge: general image processing knowledge or procedural knowledge, heuristic knowledge, application knowledge and algorithm knowledge. Having the segmented image and using the information from the knowledge base, MuSIP uses data fusion at symbol level to analyse the image. This system fuses data at region level. Regions of images should be segmented and after pre-processing, labels from knowledge should be assigned. The level of confidence for the attached label is also determined during the pre-fusion. Final fusion takes place over all data which are available from different sources/sensors.

Jeon and Landgrebe [26] proposed two fusion methods at decision level for multi-temporal classification: multi-temporal classification with the joint likelihood decision fusion and multi-temporal classification with the weighted majority decision fusion. In the first approach they assumed that $p$ local decisions, $\{u_1, \ldots, u_p\}$, by a classifier using separate temporal data are available. The problem was making an optimal decision, $u_0$, among
When the local decisions are given. They defined the expected cost of determining $u_0$ given $\{u_1, \ldots, u_p\}$ when the true class is $\omega_j$ as:

$$E\{J(u_0; u_1, \ldots, u_p, \omega_j)\} = \sum_{\omega_j \in \Omega_0} J(u_0; u_1, \ldots, u_p, \omega_j) P\{u_1, \ldots, u_p, \omega_j\}$$

where $J(\cdot)$ is the cost function, $u_0$ is optimum global decision, $\Omega_0 = \{\omega_1, \ldots, \omega_M\}$ are $M$ user-defined information classes. They used the cost function which is defined by Tang et al. [73] as: $J(u_0; u_1, \ldots, u_p, \omega_j) = J(u_0; \omega_j) = [1 - \delta(u_0, \omega_j)]$ where $\delta(u_0, \omega_j) = 1$ if $u_0 = \omega_j$, and 0 otherwise. It means that the cost function is equal to 0 when the optimum global decision is exactly the true class and equal to 1 otherwise. They assumed conditional independence of $u_k$'s given $u_0$ i.e. $P\{u_k|u_{k-1}, \ldots, u_1, u_0\} = P\{u_k|u_0\}$. So they defined the maximum of the $P\{u_0\} \prod_{k=1}^p P\{u_k|u_0\}$ as the global optimum decision based on Bayesian minimum cost. They called this classifier the jointly likelihood decision fusion multi-temporal classifier.

In the second approach they modified the cost function of the first method to consider the reliability of local decisions. The cost function was defined as $J(u_0; u_1, \ldots, u_p, \omega_j) = \sum_{k=1}^p J(u_0; u_k, \omega_j)$ where $J(u_0; u_k, \omega_j)$ is a local cost function regarding the $k$th data set and is equal to: $J(u_0; u_k) = 1 - A_k(u_k)\delta(u_0, u_k)$ where $0 \leq A_k(u_k) \leq 1$. $\delta(u_0, u_k) = 1$ if the spectral class $u_k$ belongs to the information class that $u_0$ expresses, otherwise it is 0. After defining the expected cost function and minimising it they came up with the following classifier which should be maximised.

$$H_{TP-WHTM}(u_0) = \sum_{k=1}^p A_k(u_k)\delta(u_0, u_k).$$

They called this classifier the weighted majority decision classifier. If all local decisions have the same reliability i.e. all $A_k(u_k)$'s are 1, the classifier reduces to the majority rule. So the votes of local decisions are honoured by the incorporation of $A_k(u_k)$'s.

They implemented the above mentioned methods for classification of multi temporal Landsat Thematic Mapper (TM) data into four information classes: corn, soybean, wheat and alfalfa/oat. They used the ML classifier for local decisions. Using the first approach they had better overall classification accuracy. The second method also had comparable results with those of first method.
3.3  Feature Level Fusion

Jimenez et al. [27] used two different types of data fusion to deal with hyper-dimensional data. They used feature level data fusion to project from one feature vector space to another feature vector space. Their aim was reduction of dimensionality of multispectral data while preservation as much information as possible. Also they employed data fusion at the decision level to combine local decisions of each sensor. They used different fusion methods such as the majority voting schemes, the maximum, minimum and average rule and a neural network in this step. They defined 5 information classes, namely soybean, corn, grass, woods and soil, to be classified from 200 spectral channels of images on a pixel-by-pixel basis. In the paper they compare four methods which they used to classify the hyperspectral data. Using supervised projection pursuit for reduction of dimensionality from 200 to 10 and ML for classification had the best result.

3.4  Data Fusion Approaches

3.4.1  Statistical Approach

There are several statistical based techniques for integrating multi source data [31, 39]. The first and simplest technique is the stacked vector. In this method all data from different sources are aligned into one vector. This then is considered as the data vector as if it were from only one source. Although this method is straightforward, it can be used only when all sources are similar and can be described by a common model. In many cases this condition is very difficult. The second technique is based on stratifying the data. Some data are used to stratify and others are used to analyse each stratum [39]. A third approach is similar to the second one. It was introduced by Hutchinson [23] and is called ambiguity reduction. In this approach some sources are used to stratify the data, then the others are used to reduce ambiguities. The main disadvantage of this method is that the result depends on the ordering of the sources. The last approach is used by Lee et al. [39] and is called global membership function. In this technique, the membership functions are driven by Bayes formulae. Since these functions are commutative, there is no problem arising from the ordering of the sources. In this technique the relative quality or uncertainty of the sources can also be involved.
3.4.2 Dempster-Shafer Evidence Theory

Kim and Swain used evidential reasoning approach for integrating multi source data [31, 32]. At first they used six data sources: Airborne multispectral scanner (A/B MSS), synthetic aperture radar (SAR) shallow and steep mode and topographic data (elevation, aspect, slope). The area under study was the Anderson river and six information classes were selected. They used interval-valued probability for representation of degree of belief and maximum plausibility was used for decision making. This is the opposite of point-value probability which is used in the Bayesian inference method. They compared the maximum likelihood (ML) classification based on the stacked vector approach with the multi source data method. For the multi source data method they used consonant belief function (CBF) and partially consonant belief function (PCBF). Their results show that classification with the multi source data (CBF and PCBF) is better than with the maximum likelihood method. They used various degrees of reliability for different sources in the multi source data classification and showed that the result of incorporating different relative reliability is better than that of the equal reliability. In a second test they used High Resolution Imaging Spectrometer (HIRIS) data which are data of very high dimensionality. Reasoning with data of high dimensionality is not only computationally expensive, but it also requires many more training data. There are some techniques for reducing dimensionality of the feature vector, but they require the calculation of statistical parameters. Finally, Kim and Swain separated HIRIS data into three parts based on the correlation between them. Then they dealt with these data as if they were from three sources. They used Dempster-Shafer theory for combination of masses which were assigned by these three dummy sources. The result showed that this method has better classification relative to maximum likelihood method.

Le-Hegarat et al. [42] used Dempster-Shafer theory for unsupervised classification in multi source remote sensing. They used multi frequency polarimetric radar images and multi spectral optical images for the classification of an agricultural site. In unsupervised classification they used information from multi source as a complementary tool. They assumed that there is no ambiguity between clusters so they assigned null empty mass functions to compound hypotheses. It means that they assigned mass functions only to single hypothesis (singletons). In the specific case when there were two classes that were not distinguishable by one sensor, they assigned the same mass to their union. For definition of the mass function they
used conditional probability. In the decision making step they chose the proposition with

\[ \text{max} [\text{bel}(A)] \quad \text{if} \quad \text{bel}(A) \geq \text{bel}(\neg A). \]

This rule was triggered only for singleton classes. If there was not any case to satisfy this condition, the pixel was labelled as unclassified. Their results show that data fusion based on Dempster-Shafer theory is much better than stacked vector and class subdivision approaches. For example, they could get 20% improvement in the classification for corn.

Lee et al. [39] have used two different approaches for the integration of multi source data, probabilistic and Dempster-Shafer evidence theory. They used a subscene of a Landsat MSS image of an agricultural region. By a supervised analysis they divided the region into 11 information classes. They used information from visible and infrared sources separately. They assumed that the distribution is normal and the sources are independent.

When the prior probabilities are available, there exist two methods for incorporating them. The first technique considers prior probabilities as a separate source and uses Dempster-Shafer theory for integrating them. The second treats the prior probabilities as normalising factors. Their experimental results showed that the first technique had better accuracy but when they used prior probability as a normalising factor in Dempster-Shafer theory, both approaches had the same results. Their results also showed that they may be improved by using an uncertainty factor.

Peddle and Franklin improved a software (MERCURY⊕) for multi source classification of surface cover and frozen ground based on evidential reasoning [51]. Their system is able to process many different levels of measurement (nominal, ordinal, interval and ratio) without any restriction about the statistical model of the data and incorporates uncertainty by using Dempster-Shafer theory. They used a new frequency-based technique to build a Knowledge Look Up Table (KLUT) from which the belief functions are derived automatically. In the first test they used 3 parameters from field observations (nominal land cover class, terrain aspect and potential insulation). The results showed 85% agreement with a soil probe. In a second test they used the same parameters which were extracted from remotely sensed data. The accuracy of classification in this test was 82%. They concluded that this high accuracy shows close relation between remotely sensed data and field observations. In the decision making step they used the greatest sum of support and plausibility.

Moon used evidential reasoning for integration of four data sets, Airborne Elevation Model (EM), Airborne total field magnetic, ground EM and bedrock geology maps of Faley lake area
in Canada to explore the presence of iron ore deposit and base metal deposit [48]. Airborne EM was available only for two-thirds of the test area and only a small portion of the test area was covered by ground EM survey. Geological maps represented only approximate location of the basement rocks, however the aeromagnetic field had full coverage on the test area. The probability assignment to a proposition, which is an important task in Dempster-Shafer approach, was made by the author's intuitive and qualitative knowledge of the theory of mineral deposits because no systematic statistical approach for mineral exploration was available. Using Dempster's rule of combination, he plotted belief, disbelief, plausibility and ignorance maps. These maps were in agreement with the expected results and could convey all information or lack of information. For example, the highest support for iron formation was located in the area where pyrrhotite had been found by drilling.

Wilkinson and Megier [82] used the Gordon-Shortliffe algorithm, which is an implementation of Dempster-Shafer theory for hierarchical hypotheses to integrate image classifiers with expert systems and geographic information systems (GIS). They demonstrated how Dempster's rule of combination is used to combine the information from a maximum likelihood image classifier with five contextual expert rules. Expert rules used geographic information such as terrain altitude, slope, aspect etc which were available from a GIS. While an image classifier alone could not produce an accurate ground cover, using new evidence from expert rules could improve the accuracy of classification.

Tupin et al. [76] proposed a method for automatic interpretation of SAR images. Low level information and the speckle statistics which were extracted by detectors were presented to the Dempster's combination rule. The information classes that they defined were urban, industrial, homogeneous (forest or sea), relief, road, river and bright field. Different operators where employed to detect their dedicated objects. Measures of confidences of operators were used to define the mass functions. After fusion of responses from detectors by Dempster-Shafer evidence theory, final decisions were made by using contextual information within the Markovian framework. By their proposed method after setting many parameters in a training phase, the system could automatically interpret any new SAR images which came from the same sensor.
3.4.3 Neural Network Approach

Dai and Khorram [12] proposed a framework based on artificial neural networks for land cover change detection. Their aim was detection of multitemporal changes in land cover from two Landsat TM images which were taken at two different dates. They used four information classes: forest, agriculture/bare/urban, cypress/wet deciduous scrub/marsh and water. They used six non-thermal TM spectral bands. So they employed a neural network with 12 input nodes, 6 for each image, and 16 output nodes, one for each change from one class to the other classes. By experiments they found that two hidden layers with 36 and 48 nodes for the second and third layer respectively have the best results. Other parameters of the neural network were found experimentally too. They used the multilayer backpropagation algorithm. They defined two types of accuracy; overall and categorical accuracy. Overall accuracy is calculated over both categories, changes and no-change, while categorical accuracy is the minimum of accuracy of two categories. They could have 97% overall change mask accuracy with a categorical change accuracy of 95.6%. They compared their result with the Maximum Likelihood approach which could have 89.9% and 86.5% overall and categorical accuracy respectively.

Benediktsson et al. [4] proposed several methods to combine multiple classifiers which were used to classify land cover into 6 information classes. Six data sources were used: airborne multispectral scanner data source, two synthetic aperture radar data sources and three topographic data sources; slope, aspect and elevation. They combined classifiers based on consensus theory. In consensus theory single probability distributions are combined to summarise estimates from multiple experts, which are based on Bayesian decision theory. They used linear and non-linear optimisation methods to optimise the outputs of individual classifiers. They compared various combing schemes including minimum Euclidean distance, maximum likelihood method, linear opinion pool and logarithmic opinion pool (consensus rules), single-stage conjugate gradient backpropagation (CGBP) neural network and parallel consensual neural networks. Different weighing schemes and optimisation also were tested. Their results showed that the logarithmic opinion pool optimised with CGBP had the best overall and average accuracies.
3.4.4 Knowledge-based Systems

Srinivasan and Richards used two knowledge-based techniques for integration of multi source remote sensing data, numerical and qualitative reasoning [69]. In the numerical reasoning method, they used a knowledge base which was formed from many subsidiary knowledge bases, which were related to sources. Each source had a knowledge base. All possible cover type labels (vegetation, water, soil) were organised in a hierarchical form. The appropriate rules fired when data were driven to the system, then masses were assigned to various nodes. By assuming independence between sources, they used the Dempster’s rule of combination for final belief assignment to specific data. They used implementation of Dempster’s rule for a hierarchy of propositions which has been proposed by Shafer and Logan [64]. They used the following rules for final labelling (decision making).

1. If there is a single label with maximum belief, that label is chosen.

2. If two or more labels have the same belief, the label with the most plausibility is chosen.

3. If two or more labels have the same belief and plausibility, the system suspends judgement.

The authors suggest two ways in assigning mass to a node. The first one is using functions which assign a fix mass to a node. The other one is using heuristic functions for calculating the mass function. In this method when a rule is fired it is possible that a fixed mass is assigned to the node as a fuzzy value which must be calculated by the system by using some functions. They compared the result of rule based classification using theory of evidence with maximum likelihood classification. The results show the result of rules based classification using theory of evidence are much cleaner than the result of maximum likelihood classification. From computational expenses consideration the evidence theory approach (using Shafer and Logan approach) was better, to be $O(N)$ order in comparison with the maximum likelihood method which was of $O(N^2)$ order.

3.4.5 Bayesian Network Approach

Hellwich and Gunzl [21] proposed a Bayesian network for scene interpretation by fusion of optical image data and SAR data. They defined three levels between the real world objects
and image data: the real world level which contains topographic data, the sensor level which contains image data and the geometry and material data which links between the two end levels, real world level and sensor level. This is because often cause-effect relation between measurements and the objects is through geometric properties of the objects. They constructed a Bayesian network in which the real world level is used for landuse classification. They defined forest, agricultural vegetation and built up areas as different states of the landuse node. Green vegetation, wood and soil nodes were used for connecting the landuse node to multispectral data and small scale roughness and large scale roughness for SAR data. They proposed another Bayesian network to include contextual information and also they extended the proposed Bayesian network for multitemporal multi sensor landuse classification.

3.4.6 Fuzzy Logic Approach

Binaghi et al. used a fuzzy set-based approach for classification of multi source remote sensing images [5, 6]. They proposed a fuzzy logic system that integrates contextual information with the multi source data. They applied the proposed approach to identification of the glacier equilibrium line in two different zones in the Italian Alps. The fuzzy classifier was able to classify the images into 3 classes: snow, ice and other. The numerical results showed that the proposed fuzzy classifier could correctly classify the two images by 88% and 87% accuracy.

Solaiman et al. [68] developed a fuzzy-based multi sensor data fusion classifier to integrate multi sensor, contextual and a priori information. Their aim was land cover classification using ERS-1/JERS-1 SAR composites. Fuzzy membership functions were computed by using multi sensor data. They updated the membership functions by using contextual information in an iterative procedure.

Chanussot et al. [11] used fuzzy fusion operators to combine SAR images acquired at different dates (multitemporal) to automatically detect lines in connection with road network extraction. They used two approaches to combine the images. In the first approach they used an operator with compromise behaviour as they could not get good results from extension of union and intersection operators. In the second approach they combined the result of two extreme operators, one a severe and another one indulgent to obtain global intermediate results. They tested different fuzzy operators for both approaches. They could get
satisfactory improved automatic detection results although none of them were comparable with manual extraction.

### 3.5 Previous Work

In this section we survey two major research projects which have been carried out on the problem that we are working. The first project was carried out by Stassopoulou et al. [70, 71, 72] by using a Bayesian network. The second project was presented by Sasikala et al. [60, 61, 62] by using fuzzy set theory. Although they used the same data and tried to solve the same problem there was a small difference between them in the number of output classes. While Stassopoulou et al. used 3 output classes, Sasikala et al. used 5 output classes. The reason behind this was a technical problem in implementing the Bayesian network with 5 output classes. We shall explain this in chapter 5.

#### 3.5.1 Bayesian Network Approach

Stassopoulou and co-authors [70, 71, 72] developed a Bayesian network to assess the risk of desertification of burnt forests. They used Pearl's algorithm to propagate the evidences through the network. Based on Geosciences experts the risk of desertification depends on two factors; the forest regeneration potential and the risk of soil erosion. The former itself is affected by two spatial variables, aspect and soil depth. The later, however depends on slope, rock type and soil depth. They showed dependence of the risk of desertification by a graph which is shown is figure 3.1. This graph shows the cause-effect relation between variables and the risk of desertification. Therefore they established a Bayesian network similar to the graph in figure 3.1. The circles are the nodes and the arcs between the nodes express the cause-effect relation between nodes. Labels, which are assigned to nodes, are shown in the legend of the figure.

In the Bayesian network prior probabilities should be assigned to the labels of root nodes. The relation between root nodes and their child nodes which are represented by links should be quantified by conditional probability matrices. Because there were no statistics available for them to derive prior probabilities, they used equal prior probabilities, for the root nodes. To compute the conditional probabilities they used two methods. For the soil erosion, where an analytic equation was available to relate contributing factors slope, aspect and soil depth
Figure 3.1: Relation between variables and natural regeneration potential, soil erosion and risk of desertification.

to the soil erosion, they used the equation to derive the conditional probability matrix. For the natural regeneration potential and relation between the natural regeneration potential and soil erosion with the risk of desertification, where no analytic equation was available, they used expert rules which relate these variables together to derive the conditional probability matrices. After the definition of primary conditional probabilities they used a training set to tune the values to have the best results.

Stassopoulou et al. incorporated two kinds of uncertainties in the process. They considered the uncertainty in data by the probability with which a label was assigned to input nodes. For example, instead of assigning a label to the node $S$ as $(1,0,0)$ which implies that the forest belongs to data class gentle, they incorporated the uncertainty of data by assigning probabilities to each label. They assumed that the error in the measurements
is distributed Gaussianly with a specific standard deviation. They derived the standard deviations intuitively from the expert's opinion.

The second uncertainty that they considered was uncertainty in the expert rules. Since the expert rules are realised as conditional probabilities, tuning them by using a training set can be interpreted as introducing uncertainty of the expert rules.

Stassopoulou et al. implemented the above mentioned Bayesian network to assess the risk of desertification of a forest after a fire. They used Pearl's algorithm to propagate the items of evidence through the network. Final classification was based on the maximum probability. They used 39 sites for training the network and deriving the conditional probability matrices. Their results show that the system could classify 37 sites out of 39 training sites and 13 sites out of 14 test sites. They compared these results with a rule-based system results when no uncertainty was considered during inference. Because in rule-based inference system no training was needed, they just implemented the expert rules. The results showed that only 18 sites out of 53 sites were correctly classified. The great improvement in results shows the effectiveness of the approach especially incorporating the uncertainty in data and expert rules.

3.5.2 Fuzzy Logic Approach

Sasikala and co-authors [60, 61, 62] tried to solve the same problem that we explained above by a fuzzy logic approach. The only difference is that they defined 5 classes for variables and also for the output. They used a finer set of classes in comparison to Stassopoulou et al.

Because of vague concepts and also imprecision included in the data, they preferred to solve this problem by a fuzzy logic approach. They had 3 sets of expert rules that govern the relations between influencing factors (variables) and the natural regeneration potential and the risk of soil erosion. And also relation between these two assessments and the risk of desertification. These three sets of expert rules were used by Stassopoulou to derive the conditional probability matrix elements. These expert rules are given in chapter 5.

Using fuzzy logic for decision-making tasks involves two steps: membership function evaluation and combination of fuzzy set by fuzzy operations. Sasikala et al. for the first step assumed that the variables which were given by the measurements (slope and aspect) have
3.5. Previous Work

Gaussian distribution error. They assumed that these variables have zero mean error with some standard deviations which were intuitively selected. Then they derived a formula, discussed in detail in chapter 6, which given a measurement with a specified mean and standard deviation, yields the probability of the object belonging to a predefined class. They used this probability as membership function of the variables which were given by their values not the classes.

For the second step of fuzzy reasoning, Sasikala et al. presented a novel approach. Instead of using common fuzzy operators to combine the fuzzy sets, they considered the relative importance of the variables (or the fuzzy set that are being combined) by a new approach. They considered the relative importance of variables by allowing their membership functions get values which reflect their relative importance. By their method the membership function is not restricted to the interval $[0,1]$ but can take any value in $[0, w_i]$ which $w_i$ is the largest value that the membership function of a variable can get. They believe that this approach is different from scaling the membership function based on their importance, because some of the aggregation operators are not linear and have to be modified.

Generalisation of the fuzzy aggregation operators to allow the fuzzy variables to have any membership function grades produced few changes in the definitions and properties of the operators. Table 3.1 shows some of the operators which they defined for the generalised operators. They examined all properties of the original fuzzy operators for the generalised form.

Sasikala et al. applied their proposed fuzzy operators to solve the problem. Expressing the expert rules, which some of them are given in tables 5.1-5.3, in plain language it will reveal that there are two kinds of combination of information: disjunctive and conjunctive. The fuzzy conjunctive operations were used for aggregating the conditions which variables should meet to trigger a rule and when they wanted to combine few expert rules which lead to the same conclusion they used disjunctive operations. The same way as Stassopoulou et al., they used 39 sites to find the best weights or factors of relative importance. Also, they examined different disjunctive and conjunctive operators with different importance factors of fuzzy sets. They used symmetrical sums and mean operators as both disjunctive and conjunctive operators. Their results show that they could correctly classified 26 sites out of 39 training sites and 7 sites out of 14 test sites. This result came out when they used $\min_2$ and $\text{sum}$ as disjunctive and conjunctive operators respectively for the natural regeneration.
potential and $max_3$ and $sum$ as disjunctive and conjunctive operators respectively for the soil erosion and $min_2$ and $\sigma_+$ as disjunctive and conjunctive operators respectively for the risk of desertification. The importance weights that led to the best result were 1 for soil depth and 0.1-0.8 for the aspect in finding the limitation to the natural regeneration potential, 1 for soil depth and 9 for rock permeability and slope for the risk of soil erosion. They also gave importance factors in the final combination to the natural regeneration potential and risk of soil erosion as 1 and 0.2 respectively. In other words they used generalised operators in all combination levels.

Sasikala et al. used also another method to assess the risk of desertification. They used the same aggregation operators in all steps. Using $min_2$ and $am$ could produce the best results. They could correctly classify 37 out of 39 training sites and 9 out of 14 test sites. In all methods they compared their results with a power weighting method in which the importance of the fuzzy operands is considered by raising the membership functions to some

<table>
<thead>
<tr>
<th>Operator</th>
<th>Definition</th>
<th>Generalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection ($min_1$)</td>
<td>$\min(x, y)$</td>
<td>$\min(x, y)$</td>
</tr>
<tr>
<td>Union ($max_1$)</td>
<td>$\max(x, y)$</td>
<td>$\max(x, y)$</td>
</tr>
<tr>
<td>Probabilistic sum ($sum$)</td>
<td>$x + y - xy$</td>
<td>$\frac{w_1 x + w_2 y - xy}{\max(w_1, w_2)}$</td>
</tr>
<tr>
<td>Algebraic product ($prod$)</td>
<td>$xy$</td>
<td>$xy$</td>
</tr>
<tr>
<td>Bounded sum ($min_2$)</td>
<td>$\min(1, x + y)$</td>
<td>$\min[\min(w_1, w_2), x + y]$</td>
</tr>
<tr>
<td>Bounded difference ($max_2$)</td>
<td>$\max(0, x + y - 1)$</td>
<td>$\max[0, x + y - \max(w_1, w_2)]$</td>
</tr>
<tr>
<td>Sigmoid ($sig$)</td>
<td>$\frac{1}{1 + e^{-(x+y)}}$</td>
<td>$\frac{1}{1 + e^{-(x+y)}}$</td>
</tr>
<tr>
<td>Arithmetic mean ($am$)</td>
<td>$\frac{x + y}{2}$</td>
<td>$\frac{x + y}{2}$</td>
</tr>
<tr>
<td>Geometric mean ($gm$)</td>
<td>$\sqrt{xy}$</td>
<td>$\sqrt{xy}$</td>
</tr>
<tr>
<td>Harmonic mean ($hm$)</td>
<td>$\frac{2xy}{x+y}$</td>
<td>$\frac{2xy}{x+y}$</td>
</tr>
<tr>
<td>$max_3$</td>
<td>$\frac{\max(x, y)}{1 +</td>
<td>x-y</td>
</tr>
<tr>
<td>$min_3$</td>
<td>$\frac{\min(x, y)}{1 +</td>
<td>x-y</td>
</tr>
</tbody>
</table>

Table 3.1: *Definition of some operators: original and generalised forms.*
power. They could have slightly better results with their proposed method.

3.6 Conclusions

We surveyed many data fusion approaches in this chapter. We paid special attention to those which have been used in remote sensing. Based on definition of Benediktsson and Landgrebe that categorise data fusion into three levels, data level fusion, feature level fusion and decision level fusion, we discussed different approaches which are used in the remote sensing area. Applications based on Dempster-Shafer evidence theory, probabilistic approaches like probabilistic relaxation, fuzzy logic, neural networks, Bayesian networks and rule based systems were discussed. Based on studies which have been carried out by researchers it is clear that usually combination of classifiers (data fusion at decision level) can offer better result in comparison with individual decision makers.

Statistical methods are well known as having sound theoretical foundation [75]. Those which usually are based on Bayesian theory have been used extensively in the information fusion. Where enough samples are available for the data, these methods are much more reliable than other algorithms.

Artificial neural network approaches however, are known as distribution free methods [12]. They do not need any information about modelling the data. These methods usually rely on a training phase of the neural network implementation. If the network is trained properly (having enough examples to show to network and avoiding over-fitting) and selecting proper structure, good results are expected. Further attempts to derive models of data are not required. However they are more at the risk of misclassification for those samples that have not been seen during the training phase.

One of the drawbacks of the rule-based systems is their rigidity in defining the boundaries between different classes [54]. However, rules which are expressed by humans can not properly convey human cognitive process. Fuzzy logic was developed to solve this problem. The vagueness in the human recognition is included in the fuzzy logic framework. So, incorporating expert rules and human knowledge in the recognition process of especially complex systems is very smooth in the fuzzy logic system. The soft gradual result of the fuzzy logic system in comparison with the conventional approaches, which offer rigid results, is much more reliable and easy for the experts for further interpretation and investigation [6].
Among the reviewed approaches of fusion of information, Bayesian networks have not been used widely in remote sensing applications, because their implementation is a rather complex task. However, they have been used successfully in applications in which the relations between data classes and information classes can be expressed in cause-effect form (directly or indirectly).

I also presented an overview over two major studies which were carried out in the past on our problem. I shall report the results of combination of these methods by Dempster-Shafer theory in chapter 5.

From various studies it is understood that no optimal fusion method has been proposed yet and for each application and each type of data, one of the fusion methods may be preferred over the others.
Chapter 4

Dempster-Shafer Theory

The theory of evidence was introduced by Glean Shafer in 1976 as a mathematical framework for the representation of uncertainty. This theory has been used in knowledge-based/expert systems, because many other techniques in such systems cannot deal with uncertainty. There are two approaches for surveying the mathematical foundation of this theory, the probabilistic approach and the non-probabilistic approach [35]. In the former, probability theory is the basis of evidence theory. In the latter, the theory is expressed in an axiomatic way. In this research the second approach is selected because it is straightforward and does not need much probabilistic background.

In section 4.1 after reviewing some aspects of Bayesian theory, the belief functions are introduced. In section 4.2 the Bayesian theory and Dempster-Shafer theory are compared. The problem of decision making in the Dempster-Shafer theory is discussed in section 4.3. Application of the Dempster-Shafer theory in expert systems will be discussed in sections 4.5 and 4.6. Some operations over belief interval is given in section 4.7. In section 4.8 critical discussion of a real problem is given.

4.1 Fundamentals of Dempster-Shafer Theory

As it will be seen the Dempster-Shafer theory is a generalised form of Bayesian statistics. So it is useful to review some principles of Bayesian theory first.
Chapter 4. Demster-Shafer Theory

4.1.1 Bayesian Statistics

In this section Bayesian density functions and Bayesian functions are introduced [18]. Let \( \Theta \) be a finite and non empty set that is called **frame of discernment**. This set contains all possible true values that a quantity can take and is exclusive and exhaustive in propositions. Also let \( 2^\Theta \) be a power set of \( \Theta \) which is a set of all possible subsets of \( \Theta \). For example if \( \Theta = \{A, B, C\} \) so \( 2^\Theta = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\} \).

A function \( d : \Theta \to [0, 1] \)

is called a **Bayesian (probabilistic) density function** if:

\[
\sum_{x \in \Theta} d(x) = 1. \tag{4.1}
\]

A function \( \text{bay} : 2^\Theta \to [0, 1] \)

is called **Bayesian (probabilistic) function** or **Bayesian belief function** if:

1. \( \text{bay}(\emptyset) = 0 \)
2. \( \text{bay}(\Theta) = 1 \)
3. \( \text{bay}(A \cup B) = \text{bay}(A) + \text{bay}(B) \) when \( A \cap B = \emptyset \)

The third one is called **Bayes’ rule of additivity**, and it can be shown that it is equivalent to:

4. \( \text{bay}(A \cup B) = \text{bay}(A) + \text{bay}(B) - \text{bay}(A \cap B) \).

Generally if \( A_1, A_2, \ldots, A_n \) are subsets of \( \Theta \) then

\[
\text{bay}(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{I \subseteq \{1, 2, \ldots, n\}, I \neq \emptyset} (-1)^{|I|+1} \text{bay}(\cap_{i \in I} A_i) \tag{4.2}
\]

where \( |I| \) is the number of elements of \( I \) and is called **cardinality** of set \( I \).

**Theorem 1** (bay – d inversion) [18, 63] If \( d \) is a Bayesian density function, \( \text{bay} \) will be a Bayesian function such that

\[
\text{bay}(A) = \sum_{x \in A} d(x) \quad \text{for all } A \subseteq \Theta \tag{4.3}
\]

and we have

\[
d(x) = \text{bay}(\{x\}) \quad \text{for all } x \in \Theta \tag{4.4}
\]

and inversely, if \( \text{bay} \) is a Bayesian function then \( d \) defined above is a Bayesian density function and equation 4.3 holds.
4.1.2 Mass Functions and Belief Functions

A function
\[ m : 2^\Theta \rightarrow [0, 1] \]
is called mass function if the following conditions are satisfied.

1. \( m(\emptyset) = 0 \)

2. \( \sum_{A \subseteq \Theta} m(A) = 1 \) for all \( A \subseteq \Theta \)

The mass function is also called basic probability assignment (bpa) and it really defines a probability distribution on \( \Theta \). \( m(A) \) is the exact amount of belief committed to the subset \( A \) of \( \Theta \). \( m(\Theta) \) represents the amount of belief that is not assigned to any other subsets of \( \Theta \) and in fact it is a measure of lack of information. \( m(\emptyset) = 0 \) means that no bpa can be assigned to the empty set. If \( m(A) > 0 \), the subset \( A \) is called focal element of mass function \( m \). The union of all focal elements of a mass function is called core of \( m \):

\[ C_m = \bigcup_{A, m(A) > 0} A. \]

**Theorem 2 (d – m relation) [18]** If \( d \) is a Bayesian density function then \( m(\ldots) \) defined as

\[ m(\{x\}) = d(x) \text{ for all } x \in \Theta \]

and

\[ m(X) = 0 \text{ for all non singleton subsets } X \text{ of } \Theta \]
is a mass function. Conversely, if a mass function is non zero for all non singletons, the function \( d(x) = m(\{x\}) \) is a Bayesian density function for all \( x \in \Theta \).

Summarising, a mass function is a generalisation of the Bayesian density function. A mass function not only assigns values to singletons, but also to any other subsets of \( \Theta \). Now the belief function as a generalisation of the Bayesian function is introduced.

A function
\[ \text{bel} : 2^\Theta \rightarrow [0, 1] \]
is called a belief function if it satisfies following conditions:
1. \( \text{bel}(\emptyset) = 0 \)

2. \( \text{bel}(\Theta) = 1 \)

3. for any collection \( A_1, A_2, \cdots, A_n \) of subsets of \( \Theta \)

\[
\text{bel}(A_1 \cup A_2 \cup \cdots \cup A_n) \geq \sum_{I \subseteq \{1,2,\cdots,n\}, I \neq \emptyset} (-1)^{|I|+1} \text{bel}(\cap_{i \in I} A_i)
\]

A belief function \( \text{bel}(A) \) is the total belief or support that is committed to the set \( A \). If we compare the third condition with equation 4.2 it is obvious that every Bayesian density function is a belief function. Later, the conditions under which a belief function is a Bayesian function will be stated. It can be shown that the following relations exist between the basic functions \( \text{bel} \) and \( \text{m} \).

\[
\text{bel}(A) = \sum_{X \subseteq A} \text{m}(X) \quad \text{for all } A \subseteq \Theta
\]  

(4.5)

\[
\text{m}(A) = \sum_{X \subseteq A} (-1)^{|A-X|} \text{bel}(X) \quad \text{for all } A \subseteq \Theta
\]  

(4.6)

So the function \( \text{bel} \) brings together all the masses that imply belief in \( A \) with some certainty. We see that \( \text{bel} \) and \( \text{m} \) both have the same information and one can be obtained from the other.

**Theorem 3 (bel – bay Relation)** [18] If \( \text{bel} \) is a belief function and its mass function is given by

\[
\text{m}([x]) = \text{bel}([x]) \quad \text{for all } x \in \Theta
\]

and

\[
\text{m}(X) = 0 \quad \text{for all non singleton subsets of } \Theta
\]

then \( \text{bel} \) is a Bayesian function and its Bayesian density function \( d \) is

\[
d(x) = \text{m}([x]) \quad \text{for all } x \in \Theta.
\]

Another function that is related to mass and belief functions is the commonality function which is defined as follows:

A function \( \text{com}(A) : 2^\Theta \rightarrow [0, 1] \) is a commonality function if there is a mass function such that

\[
\text{com}(A) = \sum_{A \subseteq B} \text{m}(B) \quad \text{for all } A \subseteq \Theta.
\]

On the other hand if \( \text{com}(A) \) is a commonality function, then

\[
\text{bel}(A) = \sum_{B \subseteq A} (-1)^{|B|} \text{com}(B) \quad \text{for all } A \subseteq \Theta
\]
is a belief function. Then mass, belief and commonality functions are into one to one correspondence and thus all convey the same information [2].

Given a belief function \( \text{bel}(A) \), the function \( \text{dou}(A) \equiv \text{bel}(\neg A) \) is called doubt function, where \( \neg A \) is the complement of \( A \) in the \( \Theta \). The function \( \text{pls}(A) \equiv 1 - \text{dou}(A) = 1 - \text{bel}(\neg A) \) is called the plausibility function. While \( \text{dou}(A) \) represents the total belief against \( A \), \( \text{pls}(A) \) expresses how much we should believe in \( A \) if all unknown information were to support \( A \). In other words \( \text{pls}(A) \) is the maximum belief that can be assigned to \( A \) if all unknown information supports \( A \). Since \( \text{bel}(A) \) is the total belief in \( A \) based on the existing information in \( A \), the true belief in \( A \) will be somewhere between these two limits.

Generally \( \text{bel}(A) + \text{bel}(\neg A) \leq 1 \) so \( \text{bel}(A) \leq \text{pls}(A) \). This means that the total support to proposition \( A \) and the total support to \( \neg A \) (support against \( A \)) do not add up to 1. This property is called the non additivity and indicates that some facts or information are unknown. The interval of \([\text{bel}(A), \text{pls}(A)]\) is called belief interval. The basic functions \( \text{bel}(A) \) and \( \text{pls}(A) \) are also called lower probability and higher probability respectively. The width of this interval i.e. \( \text{ign}(A) = \text{pls}(A) - \text{bel}(A) \) is a measure of our ignorance (or lack of knowledge) about proposition \( A \).

4.1.3 Dempster’s Rule of Combination

In this section the Dempster’s rule of combination of the mass functions will be introduced as the fundamental operation in evidential reasoning. Let \( m_1 \) and \( m_2 \) be two mass functions on the same frame of discernment, \( \Theta \). The mass function \( m = m_1 \oplus m_2 \) which is called orthogonal summation of \( m_1 \) and \( m_2 \) is defined as follows [2, 63].

\[
m(A) = \frac{\sum_{X \cap Y = A} m_1(X)m_2(Y)}{1 - K} \quad (4.7)
\]

for all non empty sets \( A \subseteq \Theta \) where

\[
K = \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y).
\]

\( K \) is considered as a normalisation factor and is needed to make sure that no mass is assigned to the empty set. In addition it is a measure of conflict between the two masses. If \( K \neq 1 \) it means that there is partial conflict. If \( K = 1 \) the combination of \( m_1 \) and \( m_2 \) does not exist and they are totally or flatly contradictory. Sometimes \( \log(1 - K) \) is used as a
measure of conflict weight and it is denoted by \( wett(m_1, m_2) \). If \( wett(m_1, m_2) = 0 \) it means that there is no conflict and if \( wett(m_1, m_2) = 1 \) it means that \( m_1 \) and \( m_2 \) are totally contradictory. For illustrating purposes it is convenient to represent the Dempster’s rule graphically.

\[
\begin{array}{c|c|c|c|c|c}
 & m_1(A_1) & m_1(A_2) & \cdots & m_1(A_i) & \cdots \\
\hline
m_2(B_1) & & & & & \\
\hline
m_2(B_2) & & & & & \\
\vdots & & & & & \\
\hline
m_2(B_j) & & & & & \\
\vdots & & & & & \\
\end{array}
\]

Figure 4.1: Combination of mass functions.

Let \( A_1, A_2, \ldots, A_i, \ldots \) and \( B_1, B_2, \ldots, B_j, \ldots \) be focal elements of mass functions \( m_1 \) and \( m_2 \) respectively. Along the horizontal side of a unit square we show the mass functions of all elements of \( m_1 \) (figure 4.1). The width of each strip is proportional to the value of its corresponding mass function. Note that the sum of all mass functions is 1, so that all parts of the square side will be covered. Also the mass functions of \( m_2 \) are shown along the vertical side of the same square. The area of intersection of strips \( m_1(A_i) \) and \( m_2(B_j) \) (dashed area) represents the amount of mass that is assigned to \( A_i \cap B_j \). According to the Dempster’s rule, \( m(C) = m_1 \oplus m_2(C) \) is proportional to the sum of the areas of all rectangles that \( C = A_i \cap B_j \). It is possible that for some \( i \) and \( j \) we have \( A_i \cap B_j = \emptyset \), so to satisfy the second condition for mass functions it is necessary to scale them. This is done with the help of \( K \) in the Dempster’s rule. If \( A_i \cap B_j = \emptyset \) for all \( i \) and \( j \), all mass of the combination goes to the empty set and we say that \( m_1 \) and \( m_2 \) are not combinable or totally contradictory.
Theorem 4 [18] Let \( m_1 \) and \( m_2 \) be mass functions. Then

1. \( m = m_1 \oplus m_2 \) exists if and only if the cores of mass functions \( m_1 \) and \( m_2 \), respectively \( C_{m_1} \) and \( C_{m_2} \) are not disjoint i.e.
   \[
   C_{m_1} \cap C_{m_2} \neq \emptyset
   \]

2. If \( m = m_1 \oplus m_2 \) exists, then the core of \( m \) is equal to intersection of \( C_{m_1} \) and \( C_{m_2} \) i.e.
   \[
   C_{m_1 \oplus m_2} = C_m = C_{m_1} \cap C_{m_2}
   \]

4.1.4 Properties of the Orthogonal Sum

It is easy to show that the orthogonal summation obeys the commutative and associative laws. According to the commutative law:

\[
m_1 \oplus m_2 = m_2 \oplus m_1.
\]

This means that combination of two items of evidence does not depend on the order with which they are combined. According to the associative law

\[
m_1 \oplus (m_2 \oplus m_3) = (m_1 \oplus m_2) \oplus m_3.
\]

This means that the Dempster's rule can be extended to \( n \) mass functions.

Let \( m_1, m_2, \cdots m_n \) be different mass functions on the same frame \( \Theta \). We have

\[
m(\emptyset) = 0
\]

\[
m(A) = m_1 \oplus m_2 \oplus \cdots \oplus m_n (A) = \sum_{X_1 \cap X_2 \cap \cdots \cap X_n = A} \frac{m_1(X_1)m_2(X_2) \cdots m_n(X_n)}{1-K} = \sum_{X_1 \cap X_2 \cap \cdots \cap X_n = A} \prod_{i=1}^{n} m_i(X_i)
\]

(4.8)

for all non-empty \( A \subseteq \Theta \) where

\[
K = \sum_{X_1 \cap X_2 \cap \cdots \cap X_n = \emptyset} m_1(X_1)m_2(X_2) \cdots m_n(X_n) = \sum_{X_1 \cap X_2 \cap \cdots \cap X_n = \emptyset} \prod_{i=1}^{n} m_i(X_i).
\]

By using the Dempster's rule of combination of two masses successively it is possible to combine several masses. The orthogonal summation of the other evidence functions,
bel, pls, com and dou can be drawn from their definitions [2, 18]. For example, let bel be the orthogonal sum of bel₁, bel₂, ..., belₙ. So we have

\[ \text{bel}(A) = (\text{bel₁} \oplus \text{bel₂} \oplus \cdots \oplus \text{belₙ})(A) \]

\[ \text{bel}(\emptyset) = 0 \]

\[ \text{bel}(A) = \sum_{B \subseteq A} m(B) = \sum_{B \subseteq A} \sum_{\emptyset \neq B \subseteq A} \frac{\prod_{i=1}^{n} m_i(X_i)}{1 - K} \]

\[ = \sum_{\emptyset \subseteq X_1 \cap X_2 \cap \cdots \cap X_n \subseteq A} \frac{\prod_{i=1}^{n} m_i(X_i)}{1 - K} \]

\[ = \sum_{\emptyset \neq \cap_{i=1}^{n} X_i \subseteq A} \frac{\prod_{i=1}^{n} m_i(X_i)}{1 - K} \]

\[ (4.9) \]

### 4.2 Comparison of Bayesian Theory and Dempster-Shafer Theory

In this section the Dempster's rule of combination is compared with the Bayesian rule of conditional probability. In section 4.1.1 some properties of the Bayesian functions were introduced. Another rule says that if we have bay : 2^θ → [0, 1], and we learn that A ⊆ Θ is true, we should replace bay with bayₐ : 2^θ → [0, 1] which is given by

\[ \text{bayₐ}(B) = \text{bay}(B \mid A) = \frac{\text{bay}(B \cap A)}{\text{bay}(A)} \quad \text{for all } B \subseteq Θ \]

(4.10)

under the condition bay(A) > 0. This is called Bayes’ rule of conditioning.

**Theorem 5** Suppose that bel₁ and bel₂ are defined over Θ. Let

\[ \text{bel₂} = \begin{cases} 
1 & \text{if } B \subseteq A \\
0 & \text{if } B \not\subseteq A 
\end{cases} \]

and subset B be the only focal element of bel₂ and m₂(B) = 1. bel₁ and bel₂ are combinable if and only if bel₁(⁻B) < 1. If bel₁ and bel₂ are combinable, then:

\[ \text{bel}(A \mid B) = \text{bel₁}(A) \oplus \text{bel₂}(B) = \frac{\text{bel₁}(A \cup ⁻B) - \text{bel₁}(⁻B)}{1 - \text{bel₁}(⁻B)} \]

(4.11)

and

\[ \text{pls}(A \mid B) = \text{pls₁}(A) \oplus \text{pls₂}(B) = \frac{\text{pls₁}(A \cap B)}{\text{pls₁}(B)} \quad \text{for all } A \subseteq Θ. \]

(4.12)
Although in practical cases often new evidence occurs with some uncertainty, there is similarity between equations 4.10 and 4.12.

The Dempster-Shafer rule is a tool for combining new evidence with prior opinions. So we can change our belief in the light of new evidence. From this point of view, we deal with items of evidences as new and old evidence, although the resultant combination does not depend on which of them is old or new. This task i.e. changing the belief when new evidence arrives can be done by Bayes’ rule of conditioning. New evidence is represented as a proposition and our prior belief function as condition. Here no symmetry between old and new evidence is seen. The result of new evidence must be a single proposition with some certainty [63]. While the Dempster-Shafer theory can deal with ignorance and uncertainty, Bayesian theory cannot accommodate any ignorance. In other words Bayesian theory cannot distinguish between disbelief and lack of belief. For example if $A \subset \Theta$, Bayesian theory says that $\text{bay}(A \cup \neg A) = 1$, and since $A \cap \neg A = \emptyset$ by using Bayes’ rule of additivity $\text{bay}(A) + \text{bay}(\neg A) = 1$. It means by assuming some belief with $A$, we have to assign the remaining belief to its complement, in contrast with the Dempster-Shafer theory where $\text{bel}(A) + \text{bel}(\neg A) \leq 1$. Thus belief in $A$ does not imply any disbelief to $\neg A$. In other words we can have some evidence which brings some belief to $A$, but nothing about $\neg A$. When $\text{bel}(A) = \text{pls}(A)$, $\text{bay}(A) + \text{bay}(\neg A) = 1$. This happens when our information is complete and the Dempster-Shafer theory reduces to the Bayesian theory. So the Dempster-Shafer evidence theory is a generalisation of the Bayesian theory.

Another advantage of the Dempster-Shafer theory is that a mass function can be assigned to a union of propositions, when the information at hand can not distinguish between these propositions. So, without any further judgement about them the decision is suspended until more evidence or information becomes available.

4.3 Decision Making in Dempster-Shafer Theory

The final step of every evidential reasoning system after gathering all the evidence is decision making. In probability theory, for example, decision making is based on maximum probability. Unfortunately in Dempster-Shafer theory decision making is a problem. This is because uncertainty during integration remains, although reduces. In other words after the combination of items of evidence, we have three basic functions, mass, belief and plau-
sibility. The question is how to take a decision with these functions. Here are some of the most popular rules for decision making:

1. **Maximum belief function:** According to this rule, the proposition with the maximum belief function is selected because it is the most probable.

2. **Maximum plausibility function:** According to this rule, the proposition with the maximum plausibility function is selected because it has the most potential to become the most probable.

3. **Maximum belief and plausibility function:** This is a sound decision because it has the most probability and the most possibility. In many cases this condition is not satisfied.

### 4.4 A Simple Example

Now we demonstrate Dempster’s rule of combination with an example. Suppose we are going to aggregate two pieces of evidence which are available regarding risk of desertification of a burnt forest. Two experts have been asked to express their views about the risk of desertification based on information that are given from slope, aspect, soil depth and rock permeability of site. Suppose the risk of desertification is categorised into 5 levels, 1 for the lowest risk and 5 for the highest risk. In response to our question, “How much is the risk of desertification?” they expressed their views as follows:

\[
\begin{align*}
m_1(\{w_1\}) &= 0.4 & m_1(\{w_2, w_3\}) &= 0.3 \\
m_2(\{w_1\}) &= 0.3 & m_2(\{w_3, w_4\}) &= 0.5
\end{align*}
\]

It means that the first expert believes that the risk of desertification is level 1, or class \(\{w_1\}\) with confidence 0.4. But he could not discriminate between \(w_2\) and \(w_3\), so he has assigned his belief to the union of them, i.e. \(\{w_2, w_3\}\) with confidence 0.3. Also he was not able to assign the remaining belief i.e. 1-(0.4+0.3)=0.3 to any of the subsets. So we should assign this mass which represents lack of information to the whole frame of discernment i.e. to \(\Theta = \{w_1, w_2, w_3, w_4, w_5\}\). In a similar way, for the second expert we have: \(m(\Theta) = m(\{w_1, w_2, w_3, w_4, w_5\}) = 0.2\). By using the Dempster’s rule of combination we can combine the evidence from these two experts. Figure 4.2 demonstrates combination...
4.5 Dempster-Shafer Theory in Expert Systems

In this section we are going to study the Dempster-Shafer theory in conjunction with rule-based expert systems. Before explaining the methods some definitions are given.

<table>
<thead>
<tr>
<th>$m_1({w_1}) = 0.4$</th>
<th>$m_1({w_2, w_3}) = 0.3$</th>
<th>$m(\emptyset) = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_2({w_1}) = 0.3$</td>
<td>$m({w_1}) = 0.12$</td>
<td>$m(\emptyset) = 0.09$</td>
</tr>
<tr>
<td>$m_2({w_3, w_4}) = 0.5$</td>
<td>$m(\emptyset) = 0.2$</td>
<td>$m({w_3}) = 0.15$</td>
</tr>
<tr>
<td>$m(\emptyset) = 0.2$</td>
<td>$m({w_1}) = 0.08$</td>
<td>$m({w_2, w_3}) = 0.06$</td>
</tr>
<tr>
<td>$m(\emptyset) = 0.06$</td>
<td>$m(\emptyset) = 0.06$</td>
<td>$m(\emptyset) = 0.06$</td>
</tr>
</tbody>
</table>

**Figure 4.2:** Combination of mass functions.

of these mass functions graphically. From graphical representation of Dempster’s rule we see that the the amount of mass that is assigned to $\emptyset$ is in total 0.29. As mass can not be assigned to the empty set, the mass functions of the focal elements should be scaled. So we have:

$m(\{w_1\}) = 0.29/0.71 = 0.41; \ m(\{w_3\}) = 0.15/0.71 = 0.21; \ m(\{w_3, w_4\}) = 0.15/0.71 = 0.21$

$m(\{w_2, w_3\}) = 0.06/0.71 = 0.085; \ m(\emptyset) = 0.06/0.71 = 0.085$

From equation 4.5 we can calculate the belief functions.

$bel(\{w_1\}) = 0.409; \ bel(\{w_3\}) = 0.211; \ bel(\{w_3, w_4\}) = 0.422$

$bel(\{w_2, w_3\}) = 0.296; \ bel(\emptyset) = 1$

If we use maximum belief function rule for decision making, class $w_1$ will be selected. Note that the belief function of the union of the classes $w_3$ and $w_4$ is greater than that for class $w_1$, but usually union of classes does not specify which of them has more confidence and this is not so informative.
Simple Belief Function: A mass function is called simple if there is a subset $A \subseteq \Theta$, $A \neq \emptyset$ such that:

$$m(A) = s \text{ where } 0 < s \leq 1$$

and

$$m(\Theta) = 1 - s$$

and

$$m(\text{elsewhere}) = 0.$$

The subset $A$ is called the focus of $m$ and $s$ is called support degree of $m$. The belief function of a simple mass function $m$ is called simple support function and it is equal to the sum of all the values the mass function takes over all subsets of $A$.

$$S_m(A) = \text{bel}(A) = \sum_{X \subseteq A} m(X)$$

For example, in our problem suppose that we are given the class to which a slope belongs, i.e., $\Theta_S = \{G, M, S\}$ where $G$, $M$, $S$ stand for Gentle, Medium and Steep respectively. If a piece of evidence assigns 0.6 belief to the subset $\{G\}$, without any extra information, $m(\{G\}) = s = 0.6$, $m(\Theta_S) = 1 - s = 0.4$ and $m(\text{elsewhere}) = 0$ (i.e. $m(\{M\}) = m(\{S\}) = m(\{G, M\}) = m(\{G, S\}) = m(\{M, S\}) = 0$). Combining simple mass functions can be done according to some simple formulae which are discussed next.

Theorem 1 [18, 63]

Let $m_1, m_2, \ldots, m_n$ be simple mass functions with the common focus $A$ and respective support degrees $s_1, s_2, \ldots, s_n$ which are given by $n$ independent sources of information. Let $\oplus$ mean orthogonal sum, and symbolise the combination of these mass functions. Then $m_1 \oplus m_2 \oplus \ldots \oplus m_n$ is still a simple mass function with focus $A$ and we have:

$$S_{m_1 \oplus m_2 \oplus \ldots \oplus m_n}(A) = 1 - \Pi_{i=1,2,\ldots,n}(1 - s_i)$$

where $S_{m_1 \oplus m_2 \oplus \ldots \oplus m_n}(A)$ is the support function of $A$. The total mass function of $A$ is the same as this support function, and satisfies the following equations:

$$m_1 \oplus m_2 \oplus \ldots \oplus m_n(A) = 1 - \Pi_{i=1,2,\ldots,n}(1 - s_i)$$

$$m_1 \oplus m_2 \oplus \ldots \oplus m_n(\Theta) = \Pi_{i=1,2,\ldots,n}(1 - s_i)$$

and

$$m_1 \oplus m_2 \oplus \ldots \oplus m_n(\text{elsewhere}) = 0.$$
4.5. Dempster-Shafer Theory in Expert Systems

Also

\[ S_{m_1} \oplus S_{m_2} \oplus \ldots \oplus S_{m_n}(X) = \begin{cases} 
0 & \text{if } A \not\subseteq X, \\
1 - \prod_{i=1,2,\ldots,n}(1 - s_i) & \text{if } A \subseteq X \subseteq \Theta, \\
1 & \text{if } X = \Theta 
\end{cases} \tag{4.13} \]

This theorem is the simple version of the Dempster's combination rule used in conjunction with simple support functions.

A separable support function is either a simple mass function or the orthogonal sum (if it exists) of two or more simple support functions. There are some formulae for combining simple mass functions with different but intersecting focuses and also with different and non-intersecting focuses [18, 63].

Dichotomous mass function: If in addition to the evidence for a certain set of propositions \( A \) we have also separate evidence for the opposite propositions, we are dealing with a dichotomous function. If we have several different pieces of evidence in support of a certain set \( A \) of propositions and several different pieces of evidence against these propositions, we may use the combination rule of theorem 1 to produce a single mass function for the combined evidence in support of \( A \) and a single mass function with the combined evidence against \( A \). Suppose that \( A = \{x\} \) where \( x \) is a certain proposition and \( m_{x1}, m_{x2}, \ldots, m_{xI} \) are mass functions that represent evidence in favour of the proposition \( x \) and \( m'_{x1}, m'_{x2}, \ldots, m'_{xJ} \) are mass functions that represent evidence against proposition \( x \). By using orthogonal summation we may collect all pieces of evidence in favour of proposition \( x \) as follows:

\[ m_x = m_{x1} \oplus \ldots \oplus m_{xI} \]

\[ m_x(\{x\}) = 1 - \prod_i(1 - m_{xi}(\{x\})) \]

\[ m_x(\Theta) = 1 - m_x(\{x\}) \]

and

\[ m_x(\text{elsewhere}) = 0. \]

In a similar way we have:

\[ m'_x = m'_{x1} \oplus \ldots \oplus m'_{xJ} \]

\[ m'_x(\{x\}) = 1 - \prod_j(1 - m'_{xj}(\{x\})) \]

\[ m'_x(\Theta) = 1 - m'_x(\{x\}) \]
and

\[ m'_{x}(\text{elsewhere}) = 0. \]

Now we can combine all the items of evidence in favour of or against a proposition i.e. 
\[ m = m_{x} \oplus m'_{x}. \] This mass function is called dichotomous mass function with dichotomy \{x\} and \( \Theta \setminus \{x\} \). The focal elements of this mass function are \{x\}, \( \Theta \setminus \{x\} \) and \( \Theta \). For example, in our problem suppose that a piece of evidence assigns mass equal 0.6 to a given slope belonging to \( \{G\} \) (Gentle) and mass equal to 0.2 belonging to its complement, \( \overline{G} \} = \{M, S\} \). Without any extra information we have \( m(\{G\}) = 0.6, m(\overline{G}) = m(\{M, S\}) = 0.2, m(\Theta \setminus S) = 0.2 \) and \( m(\text{elsewhere}) = 0 \) (i.e. \( m(\{M\}) = m(\{S\}) = m(\{G, M\}) = \ldots = 0 \)).

### 4.6 Belief Combination in Rule-based Systems

In this section we will discussed some techniques to propagate belief functions through rule-based systems and networks.

**Representation of Facts** [19]

In rule based systems a fact can be represented by a subset \( F \) and the belief interval \([a, b]\). The width of this interval, \( b - a \), represents the amount of ignorance in the fact. Modelling of uncertain facts is possible by a dichotomous belief function.

\[
m(F) = a
\]

\[
m(F) = 1 - b
\]

\[
m(\Theta) = b - a
\]

**Representation of Rules** [19]

Usually rules in a rule-based system take the format

\[
\text{if } E \text{ then } H \text{ with uncertainty } [c, d].
\]

Two interpretations are possible for such a rule.

**Interpretation 1:** The lower bound \( c \) defines the degree to which \( E \) supports \( H \) and \( 1 - d \) represents the degree to which \( E \) supports \( \overline{H} \). It means that the rule is equivalent to a pair of implication relations in logic as:

\[
p : E \rightarrow H \text{ with support } c
\]
4.6. Belief Combination in Rule-based Systems

\[ q : E \rightarrow \overline{H} \text{ with support } 1 - d \]

By this interpretation the mass function is defined over \( \Theta = \Theta_E \times \Theta_H \) where \( \Theta_E \) and \( \Theta_H \) are frames of discernment for \( E \) and \( H \) respectively. We are going to distinguish from now on mass functions which express evidence for a proposition and mass functions which express evidence for an inference rule. We shall denote the latter by subscript \( R \), i.e. \( m_R \). The only time inference \( p \) is false is when \( E \) is true but \( H \) is not true, symbolically, when \( E \land \overline{H} \) is true. Therefore, \( p \) is correct when the opposite of \( E \land \overline{H} \) is true, i.e. when \( \overline{E} \land \overline{H} = \overline{E} \lor H \) is true. So, we may write:

\[
m_R(p) = m_R(\overline{E} \lor H) = c
\]

In a similar way, we can infer the mass function of rule \( q \) to be:

\[
m_R(q) = m_R(\overline{E} \lor \overline{H}) = 1 - d
\]

And the mass function that expresses ignorance as:

\[
m_R(\Theta) = d - c
\]

The focal elements \( \overline{E} \lor H \) and \( \overline{E} \lor \overline{H} \) are defined as:

\[
\overline{E} \lor H = \{(x, y)|x \in \overline{E} \text{ or } y \in H\}
\]

\[
\overline{E} \lor \overline{H} = \{(x, y)|x \in \overline{E} \text{ or } y \in \overline{H}\}
\]

It is worth mentioning here that in probabilistic terms this interpretation rule is expressed as:

\[
P(H|E) = c
\]

\[
P(\overline{H}|E) = 1 - d
\]

Note that although lower bound, \( c \), can be interpreted as our total belief (in some papers called credibility) to \( \overline{E} \lor H \), the upper bound, \( d \), is not its plausibility. However, the amount of mass, which is not assigned to the focal elements, i.e. \( 1 - (c + (1 - d)) = d - c \) and is assigned to \( \Theta \), is considered as ignorance. The case when the upper bound, \( d \), is the plausibility of \( \overline{E} \lor H \) constitutes the second possible interpretation which will be discussed below.

**Interpretation 2:** The lower bound \( c \) shows the belief of the implication relation being true and \( d \) shows the plausibility of the implication relation i.e.

\[
m_R(\overline{E} \lor H) = c
\]
\[ m_R(\overline{E} \lor H) = 1 - d \]
\[ m_R(\Theta) = d - c \]

For the sake of clarification and in order to show the difference between the two interpretations the truth table of inference \( p \) is given in table 4.1. According to the first interpretation we have:

\[ m(E \rightarrow H) = m(\overline{E} \lor H) = c \]
\[ m(E \rightarrow \overline{H}) = m(\overline{E} \lor H) = 1 - d \]
\[ m(\Theta) = m(\Theta_E \times \Theta_H) = d - c \]

According to the second interpretation we have:

\[ m(E \rightarrow H) = m(\overline{E} \lor H) = c \]
\[ m(\overline{E} \rightarrow H) = m(E \lor \overline{H}) = 1 - d \]
\[ m(\Theta) = m(\Theta_E \times \Theta_H) = d - c \]

The lower bound of belief interval, \( c \), has the same meaning in both interpretations and it shows the degree to which \( E \) supports \( H \) or the total belief which is assigned to the implication \( E \rightarrow H \). In other words it is the mass function which is assigned to the 5th column of table 4.1. The upper bounds in the two interpretations have different meanings. According to the first interpretation, \( 1 - d \) is the degree by which \( E \) supports \( \overline{H} \), or the mass which is assigned to the 6th column of table 4.1. However in the second interpretation, the upper

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>( \overline{E} )</th>
<th>( \overline{H} )</th>
<th>( E \rightarrow H )</th>
<th>( E \rightarrow \overline{H} )</th>
<th>( \overline{E} \rightarrow \overline{H} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 4.1:** Logical interpretation of rules.
bound, \( d \), shows the plausibility of the implication, i.e. \( 1 - d \) is the mass which is assigned to the last column of the table 4.1.

**Belief Conjunction:** Belief conjunction is when we have a belief function which is associated with propositions \( A \) and \( B \) and we want to deduce the belief in \( A \land B \). Let both propositions be defined in the same frame of discernment, \( \Theta \). If we have two pieces of evidence which assign functions

\[
\begin{align*}
& m_1(A) = a \\
& m_1(\overline{A}) = b \\
& m_1(\Theta) = b - a \\
& m_2(B) = c \\
& m_2(\overline{B}) = d \\
& m_2(\Theta) = d - c
\end{align*}
\]

and if these propositions are independent, it can be shown that [19]

\[
\begin{align*}
& bel(A \land B) = a \times c \quad \text{lower bound} \\
& pls(A \land B) = b \times d \quad \text{upper bound}
\end{align*}
\]

**Belief Interval of** \( (A \land B) = [a \times c, b \times d] \).

For the case of dependent propositions Hau and Kashyap have proposed some other formulae [19].

**Belief Propagation:** Aggregation of uncertainty in the antecedent of a rule and uncertainty in the rule itself is called belief propagation. Let \( P \) be a given proposition with belief interval \([a, b]\) which shows our confidence in the condition, and \( P \rightarrow Q \) be the given rule with belief interval \([c, d]\) which shows our confidence in the rule. The aim of belief propagation is to find the belief interval associated with proposition \( Q \). Hau and Kashyap in [19] have claimed that in terms of multi-valued logic, the belief propagation is identical to *modus ponens* \(^1\) so they used their belief conjunction and the modified Dempster's rule procedure to propagate the belief in the rule. The procedure they proposed is as follows:

By using the first interpretation, rule \( P \rightarrow Q \) with belief interval \([c, d]\) has focal elements \( P \lor Q, P \lor \overline{Q} \) and \( \Theta \) with mass function \( c, 1 - d \) and \( d - c \) respectively. By applying their proposed method the interval of proposition \( Q \) can be calculated.

\[
\begin{align*}
& m(Q) = m(P) \overline{\Theta} m(P \rightarrow Q) \\
& bel(Q) = \sum_{B \rightarrow Q} m(B) = a \times c
\end{align*}
\]

where \( \overline{\Theta} \) represents the belief propagation operators and the summation is over all sets \( B \) which imply \( Q \).

\(^1\)The deduction rule known as *modus ponens* states that from \( P \rightarrow Q \) and \( P \) one deduces \( Q \).
Chapter 4. Dempster-Shafer Theory

\[ \text{pl}s(Q) = 1 - \text{bel}(\overline{Q}) = 1 - \sum_{B \rightarrow \overline{Q}} m(B) = 1 - a \times (1 - d) \]

\text{belief interval of } Q = [a \times c, 1 - a \times (1 - d)].

We see that the lower bound, which is our total belief to \( Q \), is equal to the multiplication of the lower bound of the condition and the rule. However the upper bound, the plausibility of \( Q \), is one minus the multiplication of our belief in condition, \( a \), and the degree to which the condition supports the complement of \( Q \), \( 1 - d \).

By using the second interpretation, the rule \( P \rightarrow Q \) with belief interval \([c, d]\) has focal elements \( \overline{P \lor Q}, P \land \overline{Q} \) and \( \Theta \) with mass function \( c, 1 - d \) and \( d - c \) respectively. By applying Hau and Kashyap's method the interval of proposition \( Q \) can be calculated:

\[ m(Q) = m(P) \bowtie m(P \rightarrow Q) \]

\[ \text{bel}(Q) = \sum_{B \rightarrow Q} m(B) = a \times c \]

\[ \text{pl}s(Q) = 1 - \text{bel}(\overline{Q}) = 1 - \sum_{B \cap \overline{Q} \neq \emptyset} m(B) = 1 - b \times (1 - d) \]

\text{belief interval of } Q = [a \times c, 1 - b \times (1 - d)].

This means that the lower bound is the same as according to the first interpretation of the rule but the upper bound, the plausibility of \( Q \), is one minus the multiplication of plausibility of the condition, \( b \), and the degree to which the condition supports the complement of \( Q \), \( 1 - d \).

4.7 Belief Conjunction

So far we have discussed belief conjunction when the propositions are from the same frame of discernment. In this section we are going to discuss the same problem when the propositions are from different frames of discernment. Let \( x_1 \) and \( y_1 \) be propositions from frames of discernment \( \Theta_X \) and \( \Theta_Y \) respectively which are independent. If we have two pieces of evidence which assign belief functions

\[ m_1(x_1) = a \quad m_1(\overline{x_1}) = b \quad m_1(\Theta_X) = b - a \]

\[ m_2(y_1) = c \quad m_2(\overline{y_1}) = d \quad m_2(\Theta_Y) = d - c \]

how much is the belief interval of \( x_1 \land y_1 \)? It is better to clarify that belief functions of \( x_1 \land y_1 \) are defined over the frame of discernment \( \Theta_X \times \Theta_Y \). In the simplest case let \( \Theta_X = \{x_1, x_2\} \)
and \( \Theta_Y = \{y_1, y_2\} \). Let us also use the notation \( 2^{\Theta_X} \) to indicate the power set \( X \), i.e. the set of all its subsets. So we have:

\[
2^{\Theta_X} = \{\emptyset, \{x_1\}, \{x_2\}, \{x_1, x_2\}\}
\]

\[
2^{\Theta_Y} = \{\emptyset, \{y_1\}, \{y_2\}, \{y_1, y_2\}\}
\]

\[
\Theta_X \times \Theta_Y = \{(x, y) \mid x \in \Theta_X \text{ and } y \in \Theta_Y\} = \{(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)\}
\]

\[
2^{\Theta_X \times \Theta_Y} = \{\emptyset, \{(x_1, y_1)\}, \{(x_1, y_2)\}, \ldots, \{(x_1, y_1), (x_1, y_2)\}, \ldots, \{(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)\}\}
\]

Here I start with very simple examples to be sure that the principles of this problem are understood properly.

**Example 1:** \( A = x_1 \wedge y_2 \) means that:

\[
A = \{(x, y) \mid x \in \{x_1\} \text{ and } y \in \{y_2\}\}
\]

\[
A = \{(x_1, y_2)\}
\]

**Example 2:** \( A = \{x_1, x_2\} \wedge y_2 \) means that:

\[
A = \{(x, y) \mid x \in \{x_1, x_2\} \text{ and } y \in \{y_2\}\}
\]

\[
A = \{(x_1, y_2), (x_2, y_2)\}
\]

**Example 3:** \( A = x_1 \vee y_2 \) means that:

\[
A = \{(x, y) \mid x \in \{x_1\} \text{ or } y \in \{y_2\}\}
\]

\[
A = \{x_1\} \times \Theta_Y \cup \Theta_X \times \{y_2\}
\]

\[
A = \{(x_1, y_1), (x_1, y_2), (x_2, y_2)\}
\]

**Example 4:** If we have two pieces of evidence which assign the following masses:

\[
m_1(x_1) = 0.6 \quad m_1(x_1) = m_1(x_2) = 0.2 \quad m_1(\Theta_X) = 0.2
\]

\[
m_2(y_2) = 0.3 \quad m_2(y_2) = m_2(y_1) = 0.5 \quad m_2(\Theta_Y) = 0.2
\]

What is the belief interval of \( x_1 \wedge y_2 \)?

Actually we have a dichotomous belief functions with the following belief intervals:

\[
x_1 : [0.6, 0.8]
\]

\[
y_2 : [0.3, 0.5]
\]

We may define all combinations of the focal elements of these two frames. We also may assign a mass function equal to the multiplication of masses of focal elements which are
combined to form their combination (conjunction). In this example we have 9 combinations \( x_1 \land y_2, x_1 \land y_1, x_1 \land \Theta_Y, x_2 \land y_2, x_2 \land y_1, x_2 \land \Theta_Y, \Theta_X \land y_2, \Theta_X \land y_1 \) and \( \Theta_X \land \Theta_Y \) which can be assigned respectively with 0.18, 0.30, 0.12, 0.06, 0.10, 0.04, 0.06, 0.10 and 0.04. All these combinations are subsets of the power set of \( \Theta_X \times \Theta_Y \) and all masses add up to one as they should. Other subsets of \( \Theta_X \times \Theta_Y \) which do not appear in these combinations are assigned zero mass because by this method of assigning mass no more mass is left to be assigned and on the other hand no evidence is available to support them. After assigning a mass function to all focal elements it is an easy task to calculate the other belief functions. In this example we have:

\[
\begin{align*}
bel\{(x_1, y_2)\} &= m\{(x_1, y_2)\} = m_1(x_1) \times m_2(y_2) = 0.18 \\
pls\{(x_1, y_2)\} &= 1 - bel\{(x_1, y_2)\}
\end{align*}
\]

As the total mass (which is equal to 1) has to be distributed between the three possibilities \( A, \overline{A} \) and \( \Theta \) we have that in general \( m(A) + m(\overline{A}) + m(\Theta) = 1 \). Applying this identity for \( A = \{x_1, y_2\} \) and \( \Theta = \Theta_X \times \Theta_Y \) we obtain:

\[
\begin{align*}
pls\{(x_1, y_2)\} &= bel\{(x_1, y_2)\} + m(\Theta_X \times \Theta_Y) = 0.18 + 0.04 = 0.22 \\
So the belief interval of \( x_1 \land y_2 \) is: \([0.18, 0.22]\).
\end{align*}
\]

Generalising, suppose that we have two independent frames \( \Theta_X = \{x_1, x_2, \ldots, x_n\} \) and \( \Theta_Y = \{y_1, y_2, \ldots, y_m\} \) and two pieces of evidence which can be used to define the following dichotomous belief functions:

\[
\begin{align*}
m_1(A) &= a & m_1(\overline{A}) &= 1 - b & m_1(\Theta_X) &= b - a \\
m_2(B) &= c & m_2(\overline{B}) &= 1 - d & m_2(\Theta_Y) &= d - c
\end{align*}
\]

where \( A \) and \( B \) are subsets of \( \Theta_X \) and \( \Theta_Y \) respectively. We can define the belief interval of \( A \land B \) as:

\[
\begin{align*}
bel(A \land B) &= m(A \land B) = m(A) \times m(B) = a \times c \\
pls(A \land B) &= 1 - bel(\overline{A} \land \overline{B}) = m(A \land B) + m(\Theta_X \times \Theta_Y) = a \times c + (b - a) \times (d - c)
\end{align*}
\]

\[
\text{belief interval of } A \land B : \ [a \times c, a \times c + (b - a) \times (d - c)]
\]

In other words by the conjunction operator over different frames of discernment, the lower bound of conjunction is equal to the multiplication of the lower bounds of operands and the belief interval width (ignorance) of conjunction is equal to the multiplication of the belief interval widths (ignorance) of operands.
4.7. Belief Conjunction

4.7.1 A Simple Example

We are given two variables (sources) $X$ and $Y$. Variable $X$ is classified into two classes $x_1$ and $x_2$ and $Y$ is classified into two classes $y_1$ and $y_2$. These variables are used to decide about the class of $Z$. Also $Z$ is classified into two classes say $z_1$ and $z_2$. We have some rules that establish a relationship between the classes of variables $X$ and $Y$ and $Z$:

Rule 1: IF $(X=x_1$ and $Y=y_1)$ THEN $z_1$
Rule 2: IF $(X=x_2$ and $Y=y_2)$ THEN $z_1$
Rule 3: IF $(X=x_2$ and $Y=y_1)$ THEN $z_2$
Rule 4: IF $(X=x_1$ and $Y=y_2)$ THEN $z_2$

In terms of Dempster-Shafer theory we have the relation $\Theta_X \times \Theta_Y \rightarrow \Theta_Z$ where $\Theta_X = \{x_1, x_2\}$, $\Theta_Y = \{y_1, y_2\}$ and $\Theta_Z = \{z_1, z_2\}$ are frames of discernment.

Suppose that we are given some pieces of evidence for each of these variables which say:

$m_1(x_1) = 0.6 \quad m_2(y_1) = 0.3$
$m_1(x_2) = 0.2 \quad m_2(y_2) = 0.5$
$m_1(\Theta_X) = 0.2 \quad m_2(\Theta_Y) = 0.2$

Question: What is the result of classification and how much is its reliability?

Answer 1: We need to derive the mass function of a fact which is the conjunction of two variables. We have the mass function of each variable which is defined in its frame. So we need to define the mass function in the frame $\Theta_X \times \Theta_Y$. We may multiply the mass functions of the facts (conditions) to deduce the mass function of the descendants:

Rule 1: $m = 0.18 \rightarrow z_1$
Rule 2: $m = 0.10 \rightarrow z_1$
Rule 3: $m = 0.06 \rightarrow z_2$
Rule 4: $m = 0.30 \rightarrow z_2$

Note that the sum of these masses is not 1, because some part of the mass has been assigned to frames as ignorance. If we assume that the rules are exact without any uncertainty, then the masses of facts can be assigned to the results (descendants of the rules). Note that $z_1$ has appeared in two rules and so has $z_2$. Aggregation of results can be done by simply taking summation of similar results and decision making can be done by taking the maximum. In this example we have:

$m(z_1) = 0.18 + 0.1 = 0.28$
Chapter 4. Dempster-Shafer Theory

\[ m(z_2) = 0.06 + 0.30 = 0.36 \]

So \( z_2 \) wins. Some obvious questions which that arise from this approach are: Why summation? What can Rule 1 say about the other subsets of \( \Theta_2 \)?

**Answer 2:** Now we shall discuss a second way of solving the problem. We may interpret these rules as Hau and Kashyap did. In this view first we need to define the belief interval of the fact and then that of the rule. Since the fact in every rule is conjunction of some variables we must first define the belief intervals of these variables. This can be done by using evidence which we have for each variable separately and also error or uncertainty models (if they are available). Lower and upper bounds of conjunction of variables can be calculated by multiplying lower bounds and upper bounds respectively as we did when propositions are from the same frame (although in this case propositions are from different frames).

Propagation of belief through rules can be done by the proposed method of Hau and Kashyap. If \([a, b]\) is the belief interval of the fact and \([c, d]\) is the belief interval of the rule according to the first interpretation, the belief interval of the descendent is:

\[ \text{bel}(.) = [ac, 1 - a(1 - d)] \]

For this example we have:

Rule 1: \([a, b] = [0.6 \times 0.3, 0.8 \times 0.5] = [0.18, 0.40]\)

Rule 2: \([a, b] = [0.2 \times 0.5, 0.4 \times 0.7] = [0.10, 0.28]\)

Rule 3: \([a, b] = [0.2 \times 0.3, 0.4 \times 0.5] = [0.06, 0.20]\)

Rule 4: \([a, b] = [0.6 \times 0.5, 0.8 \times 0.7] = [0.30, 0.56]\)

Now suppose that for all rules \([c, d] = [0.7, 0.9]\) so by using the belief propagation for every rule we have:

Rule 1: *belief interval of z₁* : \([0.126, 0.982]\)

Rule 2: *belief interval of z₁* : \([0.07, 0.99]\)

Rule 3: *belief interval of z₂* : \([0.042, 0.994]\)

Rule 4: *belief interval of z₂* : \([0.21, 0.97]\)

The next step is aggregation of these belief functions. This can be done by using the Dempster’s rule. In this technique we assume that our belief functions are dichotomous and it seems that the results of these rules can be aggregated. So for example from Rule 1 we deduce:

\[ m_1(z_1) = 0.126 \quad m_1(\overline{z_1}) = 1 - 0.982 = 0.018 \quad m_1(\Theta_2) = 0.982 - 0.126 = 0.856 \]
Similarly mass functions for the other rules can be obtained. Eventually by applying Dempster's combination rule we have:

\[ m(z_1) = (m_1 \oplus m_2 \oplus m_3 \oplus m_4)(z_1) = 0.164 \]

\[ m(z_2) = (m_1 \oplus m_2 \oplus m_3 \oplus m_4)(z_2) = 0.2146 \]

so \( z_2 \) wins because it is supported more than \( z_1 \).

We may use the second interpretation of a rule as already discussed. The belief interval is

\[ bel(.) = [ac, 1 - b(1 - d)] \]

In this interpretation the upper bound of the fact, \( b \), plays a role in the belief interval. It is worth noting that calculating the belief interval of conjunctive variables based on multiplication is just a suggestion.

Another possibility is to use multiplication for the lower bound and calculate the upper bound by the plausibility formula which was given in the previous section. For this example we have:

Rule 1: \([a,b]= [0.6 \times 0.3 , 0.6 \times 0.3 + 0.2 \times 0.2] = [0.18 , 0.22]\)

Rule 2: \([a,b]= [0.2 \times 0.5 , 0.2 \times 0.5 + 0.2 \times 0.2] = [0.10 , 0.14]\)

Rule 3: \([a,b]= [0.2 \times 0.3 , 0.2 \times 0.3 + 0.2 \times 0.2] = [0.06 , 0.10]\)

Rule 4: \([a,b]= [0.6 \times 0.5 , 0.6 \times 0.5 + 0.2 \times 0.2] = [0.30 , 0.34]\)

The next step is to use one of the belief propagation procedures to find the belief interval of the result. With the first interpretation of a rule we have:

Rule 1: \(belief\ interval\ of\ z_1: [0.126, 0.982]\)

Rule 2: \(belief\ interval\ of\ z_1: [0.07, 0.99]\)

Rule 3: \(belief\ interval\ of\ z_2: [0.042, 0.994]\)

Rule 4: \(belief\ interval\ of\ z_2: [0.21, 0.97]\)

which are the same as before because \( b \) does not play a role in the first interpretation.

However, according to the second interpretation we have:

Rule 1: \(belief\ interval\ of\ z_1: [0.126, 0.987]\)

Rule 2: \(belief\ interval\ of\ z_1: [0.07, 0.986]\)

Rule 3: \(belief\ interval\ of\ z_2: [0.042, 0.994]\)

Rule 4: \(belief\ interval\ of\ z_2: [0.21, 0.966]\)

Although the upper bounds of the results are a little different, in practice the lower bound influences our decision. After using Dempster's combination rule to aggregate the results (assuming that the rules act as independent sources) we have:

\[ m(z_1) = (m_1 \oplus m_2 \oplus m_3 \oplus m_4)(z_1) = 0.164 \]
Chapter 4. Dempster-Shafer Theory

\[ m(z_2) = (m_1 \oplus m_2 \oplus m_3 \oplus m_4)(z_2) = 0.2146 \]

So \( z_2 \) is selected by the decision maker because its belief function is maximum.

4.8 Real Application

The example of the previous section was chosen to be very similar to the real problem we have to solve, namely to deduce the risk of soil erosion or desertification on the basis of some given attributes and a set of inference rules. For example, for the problem of soil erosion, we have 3 variables: Slope, Soil depth and Rock Permeability. Each of these is quantised into the following classes.

**Slope:** 0 - 20% (Gentle), 21% - 40% (Middle or medium) and > 40% (Steep)

**Soil depth:** 0 - 5cm (Bare), 5 - 30cm (Shallow) and > 30cm (Deep)

**Rock Permeability:** (Permeable) and (Impermeable)

These variables are going to be used in assessing the risk of soil erosion. We have some expert's rules which express the dependence of soil erosion on these variables (see table 4.2).

From these rules we find that the risk of soil erosion is classified into 6 classes. So we can define the frame of discernment \( \Theta_{SE} \) as:

\[ \Theta_{SE} = \{ \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 \} \]

where \( \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 \) stand for classes NSR, SR, MR, HR, VHR and AE respectively.

We can represent the above rules by the IF - THEN format.

<table>
<thead>
<tr>
<th>PERMEABILITY &amp; SOIL DEPTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>PERMEABLE</td>
</tr>
<tr>
<td>BARE SHALLOW DEEP</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>IMPERMEABLE</td>
</tr>
<tr>
<td>BARE SHALLOW DEEP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SLOPE</th>
<th>PERMEABLE</th>
<th>IMPERMEABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENTLE</td>
<td>*</td>
<td>SR NSR</td>
</tr>
<tr>
<td>MEDIUM</td>
<td>*</td>
<td>MR SR</td>
</tr>
<tr>
<td>STEEP</td>
<td>*</td>
<td>MR SR</td>
</tr>
</tbody>
</table>

4.8. Real Application

Rule 1: IF (D and G and P) THEN $\theta_1$
Rule 2: IF (Sh and G and P) THEN $\theta_2$
Rule 3: IF (D and M and P) THEN $\theta_2$

\vdots
Rule 18: IF (B and S and I) THEN $\theta_6$

where $D$ stands for Soil depth is Deep and so on.

The example of the previous section exposed major difficulties in applying Dempster-Shafer theory to this problem.

1. We do not know the confidence we can assign to each rule, and how one may go about in order to measure it. For example, one might need a large amount of retrospective data which will allow the determination of the reliability of each rule. Usually such data are not available.

2. There is a degree of arbitrariness when we decide how to combine the mass functions of the conditions of a rule. For example, multiplication of their belief functions implies independence (or orthogonality) between them. One may use other operands which will allow for the possible partial dependence of the conditions. This leads to the conclusion that either the necessary information is missing, or the applicability assumptions are violated for the strict application of the Dempster's rule. Once this is accepted, it becomes evident that arbitrary decisions have to be taken, and possibly alternative rules of combination may be tried. However, the fact remains that the confidences we may have in each of the classes of the descendants will probably vary within certain range.

For example, if the value of the slope is 19%, it will be classified into class Gentle. However, this value is possibly corrupted by high levels of error in the measuring process, the interpolation method, the sampling rate and the roughness of the terrain. Such errors make the classification of a 19% slope almost equally likely in the class Gentle as well as in the class Medium. If we have an indication of the distribution of errors in each variable of the problem, we may come up with intervals of confidence associated with each classification. Our problem then is the propagation of these intervals through the inference process so that we associate an interval of confidence to the final result. The propagation can be done either by using the Dempster's rule of combination where the higher value of the confidence interval...
will be loosely interpreted as plausibility, or some other rules used in Fuzzy Reasoning which will have to be adopted to deal with interval value arithmetic.
Chapter 5

Classifier Combination Using the Dempster-Shafer Theory

In this chapter we present the Dempster-Shafer theory as a framework within which the results of a Bayesian and a fuzzy classifier can be combined to produce a better final classification. We deal with the case when the two original classifiers use different classes for the outcome, and we also present a method by which different reliability is associated with different expert rules used.

5.1 Introduction

It has been established recently that combining classifiers improves the classification accuracy for many problems. This has been established both theoretically, mainly within the framework of probability theory [33], and experimentally by many researchers. In addition, in the neural network field several approaches have been introduced to combine several nets to improve the accuracy and performance.

Combining multiple neural network classifiers can be categorised into two categories: Ensemble and modular [66, 67]. In ensemble-based approaches the number of output classes is the same for all classifiers. Each classifier has a complete solution for the problem, however combination of classifiers is used to improve the classification rate. On the other hand, in the modular approach, a problem is broken into several simple sub-problems. For example, a problem with 5 output classes can be changed into several sub-problems with 2
output classes. Each sub-problem can be solved using a neural network. Combination of all classifiers provides a solution for the original problem. We can see that in this approach each classifier does not provide a solution for the problem, but all classifiers together are complementarily used to find the final classification.

Combination of classifiers has also been investigated extensively when other types of classifier are used. Generally, classifiers can be combined at different levels: abstract level, ranking level and measurement level [29, 83]. In the abstract level only the top choice of each classifier is used for the combination purpose. In the second type, the ranking level, ranks of all output classes are available and are used in the combining process. In the measurement level complete information for the outputs of the classifiers e.g. score for each possible output, is available and is used in the combination process. Although the combination at the abstract level uses the least information (only the top choice of the classifiers), it has been used frequently because all kinds of classifiers, such as statistical and syntactic, can be combined easily [83]. However, in the ranking level approach one cannot combine any kind of classifiers. For example, syntactic classifiers which give only one label at the output can not be used. In other words, the classifiers used should be able to produce measurements at the output. These measurements can be easily converted into ranking information. Approaches that use measurement levels can combine any kind of classifiers that output measurements, but for the propose of combination, these measurements should be translated into the same kind of measurement. For example, a classifier which supplies information at the output based on distances can not be directly combined with a classifier which outputs post-probability.

Xu et al. in [83] used 3 methods of combining classifiers, all at the abstract level. They combined classifiers by using the Bayesian formalism, the voting principle and the Dempster-Shafer formalism. Their results on a case study show that the Dempster-Shafer theory of evidence had the best results in comparison with the other methods. They used the recognition rate and substitution rate of classifiers to define the mass functions. The mass function of the selected output (top choice) was defined from the recognition rate and the mass function of the other outputs (complement of the selected output) was defined from the substitution rate. If the summation of the recognition rate and the substitution rate is less than 100%, the remaining is called rejection rate and is assigned to the frame of discernment because the classifier was not able to decide, so it can be interpreted as lack of information in the Dempster-Shafer theory.
5.1. Introduction

Rogova in [56] used the Dempster-Shafer theory of evidence to combine neural network classifiers. All classifiers had the same number of outputs. So the frame of discernment, \( \Theta = \{ \theta_1, \theta_2, \ldots, \theta_k \} \), was the same for all classifiers. \( \theta_k \) represents the hypothesis that the output vector is of class \( k \). Let the \( n \)th classifier be denoted by \( f^n \), the input vector by \( \overline{X} \), and the output vector by \( \overline{Y}^n \in R^K \), where \( \overline{Y}^n = f^n(\overline{X}) \). Further, let the mean vector of the outputs of classifier \( f^n \) be denoted by \( \overline{E}_k^n \) when the input is an element of the training set for class \( k \). A proximity measure can be defined using \( \overline{E}_k^n \) and \( \overline{Y}^n \). Rogova used proximity measure, \( d_k^n = \phi(\overline{E}_k^n, \overline{Y}^n) \), to define the mass functions. She defined different proximity measures to find the best choice. For any classifier \( f^n \) and each class \( k \) the proximity measure \( d_k^n \) was defined which represented the pro-hypothesis \( \theta_k \). Any evidence against \( \theta_k \) or pro \( \theta_i, i \neq k \) was denoted by \( d_k^i \). Proximity measures which were defined as mass functions of the simple support functions were combined using simplified version of the Dempster's rule of combination. Having combined evidence from all classifiers, the Dempster's rule of combination again was used to find the total confidence for each class \( k \). The class with maximum confidence was singled out as the output of classification. Rogova claimed that this method of combining classifiers could reduce misclassification by 15-30% compared with the best individual classifiers [56].

There are other classifier combination approaches in the literature, some of which were compared in [77]. The average [74], the weighted average, the Borda count [7], the fuzzy integral, the fuzzy connectives [36], the fuzzy templates and neural network approaches are among those which have been investigated in the literature.

In this chapter we address two major problems of classifier combination:

1) The case when the classifiers combined are not of the same type. For example, one approach used to demonstrate classifier combination is to use different sources of data and apply a Bayesian classifier to each source and then combine the classifications of the two classifiers to obtain an improved final answer. In this work we combine the results obtained by classifiers of different nature not only from the point of view of yielded confidences in the classification results, but also from the more fundamental philosophical point of the approach they use, namely a probabilistic classifier and a Fuzzy logic-based classifier.

2) The classifiers combined are expected to use the same classes to classify the objects in question. In this chapter we address the problem of different classes, which however span the same classification space.
We address both these problems using the Dempster-Shafer theory of evidence, where the results of the two classifiers are considered as items of evidence in support of a certain proposition. The problem of different classes is solved by using a superset of finer classes which can be combined to produce classes according to either of the two classifiers.

Our method is demonstrated in conjunction with the problem of predicting the risk of soil erosion of burned forests in the Mediterranean region using data concerning relevant factors like soil depth, ground slope and aspect and rock permeability. This problem has been solved in the past using Pearl-Bayes networks [72] and Fuzzy Logic [61, 62]. The results of these classifiers are combined to produce a more reliable classification.

5.2 Data Specifications

The main idea of this research is the aggregation of spatial remote sensing data (slope, aspect, soil depth and rock permeability) to decide about the risk of desertification of a burned forest. Geophysical data are available from 53 sites of four areas in Greece. Slope, soil depth, and rock permeability are aggregated by the rules which are recommended by experts to decide about the risk of soil erosion. Aspect and soil depth are combined to obtain the natural regeneration potential. By some other rules, again made available by experts, the risk of soil erosion and the natural regeneration potential are combined to find the risk of desertification. So, actually, all four types of data are factors which influence the risk of desertification.

The data set at hand are driven from the Arc/Info software package. For each site four data sources (variables) are available. Measurements in these sources can be classified into several classes which are called data classes. Data classes for these variables are:

**Slope:** 0 - 20% (Gentle), 21% - 40% (Middle or medium) and > 40% (Steep)

**Soil depth:** 0 - 5cm (Bare), 5 - 30cm (Shallow) and > 30cm (Deep)

**Rock Permeability:** (Permeable) and (Impermeable)

**Aspect:** 0 - 45°, 315 - 360° (North), 45 - 135° (East), 135 - 225° (South) and 225 - 315° (West)

These variables are going to be used in assessing the natural regeneration potential and the risk of soil erosion. We have some expert rules which express the dependence of the risk of soil erosion and regeneration potential on these variables and are given in tables 5.1 and
5.2. Expert rules which relate the natural regeneration potential and risk of soil erosion to the risk of desertification are given in table 5.3.

<table>
<thead>
<tr>
<th>PERMEABILITY &amp; SOIL DEPTH</th>
<th>PERMEABLE</th>
<th>IMPERMEABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BARE</td>
<td>SHALLOW</td>
</tr>
<tr>
<td>GENTLE</td>
<td>*</td>
<td>SR</td>
</tr>
<tr>
<td>MEDIUM</td>
<td>*</td>
<td>MR</td>
</tr>
<tr>
<td>STEEP</td>
<td>*</td>
<td>MR</td>
</tr>
</tbody>
</table>


The values of slope and aspect are given for all pixels of sites in percent and degrees respectively. However, the values of the soil depth and rock permeability are not available. Instead, the data classes of pixels are given directly. Given the values of slope and aspect of pixels, it is very simple to find the corresponding data classes for each pixel. Assessing the risk of desertification can be done for each pixel. The site will be classified to the data class which most of the pixels belong to. In another approach, instead of decision making for each pixel we decide about each site collecting information from the whole site. An option is to define the fractions of pixels which belong to the various data classes as features of the site and then apply the decision making process once for the whole site. For example, in site number 1, the proportion of pixels which belong to class “bare” for soil depth is 11% and the proportion of pixels which belong to class “shallow” is 89%. One may use these fractions as mass functions in the Dempster-Shafer theory framework, as a prior probability in probabilistic approaches [72] or as membership functions in fuzzy logic based systems [60].

5.3 Problem Definition

As it was mentioned above, in our problem soil erosion depends on three variables: slope, soil depth and rock permeability. Other factors that may influence soil erosion are not taken into account as they were uniform in the area of study which our data refer to. Each of the problem variables, as well as the result takes values from a small set of possible classes. Stassopoulou et al. [72] implemented a Pearl-Bayes network with which they solved the problem of combining the values of the attributes, alongside the uncertainties associated with them, in order to infer the probability with which the risk of soil erosion belonged to one of the possible classes. The use of a Pearl-Bayes network involves the use of conditional probability functions. For the case when the combined attributes and inferred conclusions are discrete valued quantities, these conditional probabilities are matrices. In the particular case, as three attributes are combined to assess the risk of soil erosion, if we assume that each variable involved takes \( N \) possible values, the matrix must be \( N \times N \times N \times N \). So, for \( N = 5 \), there should be 625 elements of the matrix, each expressing the probability of the site to belong to a certain class of soil erosion, given that the site attributes have certain combination of classes. The calculation of such a large number of probabilities, however, requires the availability of a large number of data. In research problems one very seldomly has at one's disposal enough relevant data for such estimation. To reduce the severity of the problem, Stassopoulou et al. quantised all variables of the problem into three classes each, thus having to compute only 81 conditional probability values. Their results were quite satisfactory: They obtained consistent results on the training set for 28 out of the 30 training sites, and hardening their conclusions produced agreement with the expert in 7 out of the 9 test sites. However, in spite of their accuracy, these results use gross classes, as each variable is quantised only into one of 3 possible classes.

Sasikala and Petrou [61], solved the same problem, using the same data, but as no numerical restriction existed, their results classified the risk of soil erosion into one of five possible classes, the same ones used by the expert who had made the assessment in the first place. Sasikala et al., in order to solve this problem developed a new fuzzy methodology, which involves a training stage: weights were used for the membership functions to reflect the relative importance of the combined attributes, and many different combination rules were tried. The system was trained for the selection of the best set of weights and the best combination rule. Upon hardening the final classification, they could have consistency in
the training data in 18 out of the 30 sites and they could predict correctly the class of the test sites in 5 out of the 9 cases. However, the use of weights and a variety of combination rules, produced a blunt decision system: in some cases more than one possible classes had equally high membership functions.

The idea we propose is to combine the results of the accurate probabilistic classifier, which uses gross classes, with the results of the blunt fuzzy classifier, which uses finer classes, in order to obtain a final classification which will be more accurate and less blunt.

5.4 The Proposed System of Classifier Combination

In this research we intend to use the Dempster-Shafer theory to combine the items of evidence that come from the Bayesian network and fuzzy logic approaches. As it was discussed in chapter 4, one of the conditions to be able to use the Dempster-Shafer theory is that all sources should have the same frame of discernment. In our case this is not true, as for example, risk of soil erosion is classified into 3 classes, which we denote by $A_1, A_2, A_3$, in the Bayesian network method, and into 5 classes, which we denote by $B_1, \ldots, B_5$, in the fuzzy logic method.

To be able to use the Dempster-Shafer theory in this application, we look for a definition of a frame of discernment in which both methods can be defined. Since both methods span the same classification space, we quantise the classification space into 15 classes, $C_1$ to $C_{15}$. These classes can be expressed in both methods because 15 can be divided by 3 and 5. In other words, the union, for example, of the first 5 new classes i.e. $\{C_1, \ldots, C_5\}$ is the same as the first class of the Bayesian network method, i.e. $A_1$. Also the union of the first 3 new classes i.e. $\{C_1, C_2, C_3\}$ is the same as the first class, i.e. $B_1$, of the fuzzy logic method. Figure 5.1 shows schematically the idea of defining this superset of classes.

The next step is defining the mass functions from the available sources (classifiers here). We interpret the beliefs of the Bayesian network system as mass functions in the Dempster-Shafer theory. Since the output measurements of a Bayesian network are in the form of probabilities, no further conversion is needed to use them as mass functions. However, the membership grades of classes in a fuzzy logic based system can get any value in the interval $[0,1]$ and they do not sum to 1. Therefore we cannot interpret them as mass functions directly. Instead, we use them in order to distribute the unit mass, proportionally, to the
corresponding classes.

Before using the defined mass functions in Dempster's combination rule, another factor that should be noticed is the relative reliability of the classifiers. If we have the recognition rate, substitution rate (error rate) and rejection rate of a classifier, its reliability can be defined as [83]:

\[
\text{Reliability} = \frac{\text{Recognition}}{100\% - \text{Rejection}}
\]  

(5.1)

If a classifier does not include a rejection option, its reliability is the same as its recognition rate. We are going to use this way of defining classifier reliability for the Bayesian classifier.

The fuzzy logic classifier, however, is based heavily on the use of individual production rules, which themselves may be treated as a collection of individual classifiers. One does not necessarily expect that all rules used are equally reliable; indeed, some of them may even be wrong. To assign, therefore, an overall reliability factor to the fuzzy classifier would be equivalent to ignoring the peculiarities of the individual classifiers this is a collection of. We decided instead, to examine with the help of a training phase the reliability of the individual firing rules. They are these individual reliability factors that are used to moderate the mass functions of the fuzzy classifier.

In the Dempster-Shafer theory we can interpret unreliability of a source as lack of information. So, after we scale down the mass functions which we have already defined for each classifier, by taking into consideration the reliability of the classifier, we assign the remaining mass to the frame of discernment as lack of information.

In figure 5.2 the mass functions derived from the Bayesian network and the fuzzy logic
system after considering the relative reliability of classifiers are denoted by \( m_1 \) and \( m_2 \) respectively. The combination of mass functions before normalising is denoted by \( m' \). Note that the square area denoted by, for example, \( m(C_6) \) is equal to \( m_1(A_2) \times m_2(B_2) \). This value is used in Dempster’s rule of combination given by equation 5.2 below, in order to assign mass to \( C_6 \). As it can be seen, in sixteen cases, the mass functions which resulted from the combination of the two sources can be assigned to non empty sets. Dempster’s rule of combination is repeated below for ease of reference.

\[
m(Z) = \frac{\sum m_1(X)m_2(Y)}{1 - K}
\]

where

\[
K = \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y).
\]

Here the normalisation factor, \( K \), is:

\[
K = m_1(A_2)m_2(B_1) + m_1(A_3)m_2(B_1) + m_1(A_3)m_2(B_2) + m_1(A_1)m_2(B_3) + m_1(A_3)m_2(B_3) + m_1(A_1)m_2(B_4) + m_1(A_1)m_2(B_5) + m_1(A_2)m_2(B_5)
\]

For example, we have:

\[
m(C_1, C_2, C_3) = \frac{m_1(A_1)m_2(B_1) + m_1(\emptyset)m_2(B_1)}{1 - K}
\]
5.5 Experimental Results

Although we have classified the risk of soil erosion into 15 classes, we would like to have the result in 5 classes as used by the expert and by the fuzzy logic system. Thus, we calculate the belief function of the classes of interest, by using the mass functions of the focal elements. So, after scaling all mass functions which are assigned to non-empty subsets, the summation of masses of the classes in each row will be the belief function of the corresponding class. For example, summation of $m'(C_4, C_5)$, $m'(C_6)$ and $m'(C_4, C_5, C_6)$ in the second row after normalisation will be assigned as $bel(C_4, C_5, C_6)$ which is the belief of the second class out of the 5 possible classes, i.e.

$$bel(B_2) = bel(C_4, C_5, C_6) = \frac{m'(C_4, C_5) + m'(C_6) + m'(C_4, C_5, C_6)}{1 - K}$$

Figure 5.3 shows schematically the proposed combination system.

5.5 Experimental Results

If we denote the output beliefs of the Bayesian network by $BEL(A_1)$, $BEL(A_2)$, $BEL(A_3)$ and its recognition rate by $P_A$, we used:

$$Reliability \text{ of } A = \frac{P_A}{100}.$$  

$$m_1(A_i) = \frac{BEL(A_i) \times P_A}{100}, \quad i = 1, 2, 3$$
with $P_A = 77.7\%$.

To deal with the reliability of the fuzzy classifier, we multiplied with weights, $0 < W_i \leq 1$, $i = 1, 2, \ldots, 5$, the different mass functions which resulted from different expert rules used by the fuzzy system. We used 30 training sites to identify the best weights which will give the best results. It is worth mentioning that we used exhaustive space search to find the best weights. However, in every set of weights we fixed one of the weights to be 1 in order to make the space search smaller, and because this way the weights measured the relative importance of the various rules. We found that the best results can be achieved when the mass functions of classes $B_2$ and $B_4$ are scaled down by 0.5.

After the reliability of each classifier was taken into consideration, the sum of its mass functions was not 1. The difference of the sum from 1 was assigned to the frame of discernment which is interpreted as the lack of information. For example, for the fuzzy classifier

$$m_2(\emptyset) = m_2(B_1, B_2, B_3, B_4, B_5) = 1 - \sum_{i=1}^{5} W_i m_2(B_i).$$

where $W_i$ are the weights identified during the training phase. For the Bayesian classifier

$$m_1(\emptyset) = m_1(A_1, A_2, A_3) = 1 - \sum_{i=1}^{3} \frac{BEL(A_i)P_A}{100}.$$

By using the maximum belief function criterion in the decision making step, 26 out of the 30 training sites were correctly classified and the other 4 sites were classified in the next class from that assigned by the expert. Also 6 out of the 9 testing sites were correctly classified and the other 3 sites were classified in the next class from that assigned by the expert. This should be compared with the 5 sites which were correctly classified by the fuzzy classifier alone.

### 5.6 Discussion and Conclusions

One of the shortcomings of the generalised fuzzy logic approach is that the classifier is blunt for some operators i.e. more than one out of the 5 possible output classes may be picked if we consider the maxima of the membership functions.

The probabilistic approach is also blunt because practical constraints do not allow us to use fine classification of the output. For example, conditional probability matrices become too big if we want to have 5 output classes.
The Dempster-Shafer theory as a classifier combination method allowed us to deal with the different number of classes used by the two different classifiers. In addition we were able to take into consideration the reliability of the classifiers in the process of mass definition. Experimental results showed that not only the accuracy of classifiers is improved but also we have a fine set of output classes.

We used the recognition rate of the Bayesian classifier to define its reliability. Original mass functions were scaled down by the reliability factor before applying Dempster's rule. The remaining mass functions were assigned to the frame of discernment. This is justified in the Dempster-Shafer theory.

We treated the fuzzy classifier as a collection of individual partial classifiers consisting of the expert rules used. However we could not define the reliability of each expert rule, because we had too few examples for each rule to be able to define recognition rates. For example, only 2 sites had been classified into class 4 (high risk) by the experts. These are not enough to derive a recognition rate. So we chose a training based approach to find the relative reliability of expert rules.

We found that rules that lead to classes $B_2$ and $B_4$ were half as reliable as the other rules. This is a very interesting outcome as it shows that less emphasis should be placed on rules that lead to classes other than the two extremes and the middle. People find easy to classify things in classes like "low", "medium" and "high", or "good", "average" and "bad" etc. It is more difficult to ask people to classify things in classes like "very low", "low", "medium", "high" and "very high", or "very good", "good", "average", "bad" and "very bad". It seems that heuristics devised by people to yield classification in the new classes inserted between the 3 gross ones, may not be as reliable as rules that classify into the clear cut and well defined two extreme and a medium class.
Chapter 6

Application of the Belief Propagation in a Real Problem

6.1 Introduction

In this chapter we discuss the application of the mathematical theories which were introduced in chapter 4, in the problem of assessing the risk of desertification of burnt forests.

In summary, from the expert’s point of view, the problem of risk of desertification depends on the natural regeneration potential and risk of soil erosion of the site under study. Four variables, slope, aspect, soil depth and rock permeability were singled out as those which influence the natural regeneration potential and risk of soil erosion, and eventually the risk of desertification. Aspect and soil depth of a site determine the level of the natural regeneration potential of the site. The natural regeneration potential may be classified into one of 5 possible classes. Class 1 represents the lowest and class 5 the highest regeneration potential. According to the expert’s opinion, the risk of soil erosion depends on the slope, soil depth and rock permeability of the site. The natural regeneration potential and risk of soil erosion are used to determine the risk of desertification.

In this chapter we discuss how the experts’ rules can be used to decide about the risk of desertification.
6.2 Dempster-Shafer Theory in a Rule-based System

In this section we report and discuss the results of experiments in which different strategies are used to classify the natural regeneration potential, risk of soil erosion and risk of desertification.

We express each expert rule in the IF p THEN q format or \( p \rightarrow q \). We have four variables and for each variable few data classes are defined as given in section 5.2. Based on the Dempster-Shafer theory, we define the set of data classes as the frame of discernment of the corresponding variable. So we have:

\[
\begin{align*}
\theta_{sl} &= \{G, M, St\} \\
\theta_{as} &= \{N, E, S, W\} \\
\theta_{sd} &= \{B, Sh, D\} \\
\theta_{rp} &= \{P, I\}
\end{align*}
\]

where \( \theta_{sl}, \theta_{as}, \theta_{sd} \) and \( \theta_{rp} \) denote the frames of discernment of the slope, aspect, soil depth and rock permeability respectively. Also we have 5 propositions for the natural regeneration potential, the risk of soil erosion and the risk of desertification. So we have:

\[
\begin{align*}
\theta_{RP} &= \{\omega_{RP1}, \ldots, \omega_{RP5}\} \\
\theta_{SE} &= \{\omega_{SE1}, \ldots, \omega_{SE5}\} \\
\theta_{RD} &= \{\omega_{RD1}, \ldots, \omega_{RD5}\}
\end{align*}
\]

where \( \omega_{RPi}, \omega_{SEi} \) and \( \omega_{RDi} \) denote the ith class of the natural regeneration potential, the risk of soil erosion and the risk of desertification respectively. From table 5.2 we have two facts which are conjuncted in the IF part of the rules. The THEN parts of the rules are the classes to which they belong. For example, we have:

\[
\begin{align*}
IF \text{ (the aspect is North and the soil depth is Shallow)} \\
THEN \text{ the natural regeneration potential is class 2 (}\omega_{RP2})
\end{align*}
\]

Using mathematical symbols we have:

\[
(N \land Sh) \rightarrow \omega_{RP2}
\]

where the first part is defined over the Cartesian product of the frames, \( \theta_{as} \times \theta_{sd} \), and the second part is defined over \( \theta_{RP} \).

According to the discussed mathematical fundamentals, we should define belief intervals of the facts and the rules, if this information is available. After the definition of belief intervals, or in the simplest cases belief values, we should use a procedure to propagate them through rules and then combine the results of all rules. In the next section we discuss a group of
experiments that have been planned for this problem.

6.3 Experimental Results

6.3.1 Lack of Information About Uncertainty in Data and Expert Rules (Experiment I)

In the first experiment we use the proportion of pixels which belong to a data class as being the mass function of the class. In this experiment we do not use any information regarding errors which are probably associated with the data. The mass functions of conjuncted facts are simply calculated by multiplying the mass functions of individual facts. For example, in calculating the natural regeneration potential, the product of the mass functions of aspect and soil depth data classes is considered as the belief of the antecedent (fact) part of the rule. Since in some of the experiments that follow we use training approaches, the 53 available sites are divided into two sets: 39 sites constitute the training set and 14 sites the test set. As mentioned before, we have 12 rules for regeneration potential and 18 rules for the risk of soil erosion. Calculating the mass function of the data classes and using products of those which are used in the rules, we shall have belief functions of the fact part of the rules. Suppose that all rules are absolutely reliable and we do not have any evidence that some rules are uncertain. In this case the belief interval of rules is decreased to a point since the lower and upper part of the interval can be considered as belief and plausibility of the validity of the rule, both being considered equal to 1. In chapter 4 we showed that if the belief of the fact is given by \([a, b]\) and the belief of the rule is given by \([c, d]\), the belief of the descendent of the rule can be calculated by one of the following formulae depending on two different interpretations:

\[ [a \times c, 1 - b \times (1 - d)] \]

\[ [a \times c, 1 - a \times (1 - d)] \]

Since in this problem we have not defined the belief interval, the belief of the descendent part will also appear as a point instead of an interval.

As it can be seen some rules have the same descendents, i.e. different conditions or facts may lead to the same results. In other words a data class may appear in more than one rule. If we interpret each rule as a piece of evidence then we should accumulate all items
of evidence regarding a specific data class. Depending on the interpretation of rules, there are different approaches to combine beliefs which come up with different rules. In this experiment we simply add the belief of the same data classes to combine beliefs from different rules. Therefore after combining the items of evidence from different rules we have 5 belief functions for 5 data classes. As it was mentioned before, data classes of the outputs of rules are the classes for the natural regeneration potential and the risk of soil erosion. The belief functions of the natural regeneration potential and the risk of soil erosion will be used in conjunction with relevant rules to assess the risk of desertification i.e. $\theta_{RP} \times \theta_{SE} \rightarrow \theta_{RD}$. These rules again are supposed to be absolutely true. So the belief of the first part of a rule propagates exactly to the result of the rule. In Table 6.1 the results of this experiment are given in the row marked with Exp. I. Note that there is no difference between columns under “39 site” and “14 sites”. This distinction is only relevant for the other experiments.

<table>
<thead>
<tr>
<th>RP</th>
<th>SE</th>
<th>RD</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>39 sites</td>
<td>14 sites</td>
<td>39 sites</td>
<td>14 sites</td>
</tr>
<tr>
<td>CCS</td>
<td>CCSE</td>
<td>CCS</td>
<td>CCSE</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>32</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>32</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>26</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 6.1: Experimental results. Under CCS the number of the sites which were classified in the same class as that assigned by the expert. Under CCSE is the number of sites that were correctly classified or misclassified by one class.

6.3.2 Uncertainty in the Data: Error Models (Experiment II)

In the second experiment we make use of the error models which were derived in chapter 2 in the process of mass function assignment. Given the percentage of pixels which have been used in Arc/Info software to interpolate the slope and aspect at all pixels and also the roughness of the terrain, it is possible to estimate the error distributions for the values of slope and aspect and some statistics of these distributions.

We experimentally showed that the distribution of error in slope is Gaussian, $G(\sigma, \mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Sasikala in [60] and Sasikala and Petrou in [61] showed that if we suppose that
the maximum and minimum values that a variable \( t \) can take are \( t_{\text{max}} \) and \( t_{\text{min}} \) respectively and any value in the interval \([t_1, t_2]\) is defined as class \( c \), then the probability of the variable \( t \) belonging to class \( c \) is

\[
Pr(t \in c|\mu; t_1, t_2) = \frac{\text{erf}\left(\frac{t_{\text{min}} - \mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{t_1 - \mu}{\sqrt{2}\sigma}\right)}{\text{erf}\left(\frac{t_{\text{max}} - \mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{t_{\text{min}} - \mu}{\sqrt{2}\sigma}\right)}
\]  

(6.1)

if the measured value of the variable is \( \mu \) with standard deviation \( \sigma \). For example, consider the definition of data classes for slope and assume that the slope of a given pixel is 30\% and the mean of error in slope is 3\% and the standard deviation of error is 10\%, from equation 6.1 the probability of this pixel belonging to class \( G \), \( M \) and \( St \) is 0.24, 0.66 and 0.10 respectively.

The distribution of errors in aspect as worked out in chapter 2 is Laplacian or double exponential which can be expressed by 

\[
L(\mu, \sigma) = \frac{1}{\sigma\sqrt{2}} e^{-\frac{\sqrt{2}(x-\mu)}{2\sigma}}.
\]

As before if we suppose that the maximum and minimum values that a variable \( t \) can take are \( t_{\text{max}} \) and \( t_{\text{min}} \) respectively and any value in the interval \([t_1, t_2]\) is defined as class \( c \), then the probability of the variable \( t \) belonging to class \( c \) is:

\[
Pr(t \in c|\mu; t_1, t_2) = \frac{1}{\sigma\sqrt{2}} \int_{t_1}^{t_2} e^{-\frac{\sqrt{2}(x-\mu)}{2\sigma}} dx = \begin{cases} 
\frac{1}{2}(e^{\frac{\sqrt{2}(t_2-\mu)}{2\sigma}} - e^{\frac{\sqrt{2}(t_1-\mu)}{2\sigma}}) & \text{if } t_1, t_2 < \mu \\
\frac{1}{2}(e^{\frac{\sqrt{2}(t_2+\mu)}{2\sigma}} - e^{\frac{\sqrt{2}(t_1+\mu)}{2\sigma}}) & \text{if } t_1, t_2 > \mu \\
1 - e^{\frac{-\sqrt{2}(t_2-\mu)}{2\sigma}} - e^{\frac{-\sqrt{2}(t_1-\mu)}{2\sigma}} & \text{if } t_1 < \mu < t_2
\end{cases}
\]

(6.2)

For example, if the mean of error in aspect is zero and the standard deviation of error in aspect is 60\% and if the measured value of aspect at a given pixel is 35\% then the probability of this pixel belonging to class \( N \), \( E \), \( S \) and \( W \) is 0.53, 0.35, 0.042 and 0.005 respectively.

Given the degree of roughness of the terrain and the percentage of the data which is used for interpolation, we can calculate the mean and standard deviation of the error in slope and aspect. For the data we have, we used \( D=2.25 \), which is reasonable for the rolling hills type of landscape of southern Greece, and \( P=0.03 \) which means that the data points had been interpolated from DEM of 150 meters resolution. Then we have \( \sigma_{\text{slope}} = 12.30\% \), \( \mu_{\text{slope}} = 3.03\% \) and \( \sigma_{\text{aspect}} = 63.65\% \).

Given the values of aspect and slope of all pixels, the probability of each pixel belonging to each data class can be calculated using equations 6.1 and 6.2. The mean probability of all pixels for belonging to a given data class is assigned as the mass function of the whole site. Note that by this methodology we do not define a belief interval because of the uncertainty
Chapter 6. Application of the Belief Propagation in a Real Problem

in the data but a single value of the mass function. The other steps of this experiment are the same as those of the first experiment. The results of this experiment are shown in the second row of table 6.1. Again, the distinction between the 39 and the 14 sites is not meaningful.

6.3.3 Uncertainty in the Data: Redistribution of Mass Functions (Experiment III)

In the previous experiment we included the error models of aspect and slope in calculating the mass functions. However no error model is available for the rock permeability and soil depth. If we compare the mass functions of the first and second experiments we see that all that we did by using the measurement error models was to redistribute some part of the mass functions to other data classes. We decided to apply such a redistribution to the third variable of the problem, namely to soil depth as well. To achieve this, we used an exhaustive search for different fractions of the mass functions. In other words, if we represent the mass functions of soil depth data classes by a vector \( SD = [m_{SD}(G), m_{SD}(M), m_{SD}(St)] \), by “redistribution of mass functions” we mean multiplying the vector of mass functions by a matrix \( W \) the summation of the elements of each column of which is 1. For example, the first column of this matrix shows how the mass function of the first data class should be redistributed to the second and third data classes. One of the drawbacks of this method is that in the process of minimising the error of classification we may deduce some form of matrix \( W \) which is not logically acceptable because the training data we inevitably have to use will contain other errors, in addition to those induced by the errors in the measurement of soil depth. To overcome this problem we may use some heuristic restrictions. In table 6.1 the row marked \( Exp. \ III \) shows one of the best results obtained when matrix \( W \) was:

\[
W = \begin{bmatrix}
0.2 & 0.2 & 0.5 \\
0.7 & 0.7 & 0.3 \\
0.1 & 0.1 & 0.2
\end{bmatrix}
\] (6.3)

As it can be seen the result of the first experiment shows that there are large discrepancies between the expert results and the result that we extract from the data which are derived from the GIS by using expert results. One of the sources of these misclassifications is the error of the interpolation step in GIS. We included this information for slope and aspect
6.3. Experimental Results

in the second experiment. However in the third experiment by extending the idea of redistributing the mass function to the variable of soil depth as well, and searching for the best weights not only we account for the error due to errors in slope and aspect, but also in soil depth. The errors in soil depth are compensated by the weights we use to redistribute its mass function, and which we choose by training.

6.3.4 Reliability of Expert Rules (Experiment IV)

In all previous experiments we used for the calculation of the risk of desertification the mass functions of the natural regeneration potential and the risk of soil erosion directly as they were derived by the supplied rules. In other words we considered all rules to be absolutely correct and of equal importance. However, some rules may be less reliable than others. Although we have only as few as 39 sites, we tried to use these sites for training and searching all possible relative validity of belief functions of the natural regeneration potential and the risk of soil erosion derived by using these rules. Since we have 5 data classes for the natural regeneration potential and 5 classes for the risk of soil erosion, we have 5 different combinations of weights for each of these variables. This makes a large number of possibilities to be examined when calculating the risk of desertification. We chose weights that vary from 0.1 to 1.0 with step of 0.1. Although one may use some optimisation algorithm such as a Genetic algorithm, the possibilities we investigate are sufficiently small in number to be able to use an exhaustive search. In table 6.1 the row marked Exp.IV.1 shows the results obtained when the error models for slope and aspect were used but no weighting for soil depth. The weights used for the reliability with which the different classes for the natural regeneration potential and the risk of soil erosion can be worked out by the rules, are $W_{RP} = \{0.4, 0.1, 0.2, 0.2, 0.1\}$ and $W_{SE} = \{0.1, 0.4, 0.2, 0.2, 0.1\}$ respectively. As it can be seen the number of correctly classified sites is increased, but not so much. Using finer steps in the exhaustive search, for example, 0.05 gives better results but with the cost of a big exhaustive search space which can not handled easily.

We did this experiment with exhaustive search with steps equal to 0.05 to see if there is improvement in the results. We could get 17 correctly classified sites and 38 correctly classified within one class from that of the expert. It means we have 3 more correctly classified sites in comparison with the case when we use steps equal to 0.1. The weights which have been used are $W_{RP} = \{0.25, 0.60, 0.05, 0.05, 0.05\}$ and $W_{SE} = \{0.05, 0.15, 0.10, 0.35, 0.35\}$. 
These results are reported in row Exp. IV.2 in table 6.1. In all other experiments we used steps equal to 0.1 only for the exhaustive search.

In another experiment we followed the above procedure but we also used the method proposed in experiment III to account for the unknown uncertainty in the soil depth by using the matrix of equation 6.3. The result of this experiment is given in the row marked Exp. IV.3. The results presented were obtained with $W_{RP} = \{0.2, 0.1, 0.3, 0.1, 0.3\}$ and $W_{SE} = \{0.3, 0.1, 0.3, 0.1, 0.2\}$. In the third version of this experiment we again tried to compensate for the uncertainty in soil depth but matrix $W$ was re-deduced by simultaneous training for its elements and those of $W_{RP}$ and $W_{SE}$. The following values were used:

$$W = \begin{bmatrix}
0.5 & 0.15 & 0.1 \\
0.3 & 0.7 & 0.3 \\
0.2 & 0.15 & 0.6
\end{bmatrix}$$

$$W_{RP} = \{0.1, 0.1, 0.2, 0.6, 0.0\}$$

$$W_{SE} = \{0.1, 0.1, 0.4, 0.3, 0.1\}$$

The results of this experiment are shown in table 6.1 in the row marked with Exp. IV.4.

### 6.4 Critical Discussion of Combination of Information for Risk of Desertification

As we mentioned before there are some methods for combining different rules the outputs of which are defined in the same frame or space. In the last section we incorporated the belief of the individual rules which lead to the same conclusions and considered it as the overall belief to that data class. For example we have 3 rules with output data class, “No to slight risk” in the group of rules regarding the risk of soil erosion. Summation of the belief of 3 rules is considered as the total belief to the class “No to slight risk”. The final classification was based on finding the class with the maximum belief. Although this interpretation of rules is applicable in many cases, another interpretation is also possible. For example, if we consider each rule as a piece of evidence which gives us some degree of belief to one of the classes (depending on its conditions), and also if all rules satisfy independence criteria, i.e. reliability of one rule does not affect the reliability of the other rules, then we can combine the beliefs of rules by the Dempster's combination rule.
Based on the Dempster-Shafer theory framework, each item of evidence assigns mass functions to focal elements according to the amount of information it has. If any source can not assign a total mass of 1 to the various propositions, the remaining mass is assigned to the frame of discernment which is interpreted as the lack of information.

In this application each rule assigns a mass function to a proposition (data class) based on the evidence which is available in the fact part of the rule. If we consider each rule as an independent source then in this application each source has information only for one single data class or singleton. In this sense we may assign the remaining mass function to the frame of discernment as expressing the lack of information. However, if we had any rule which could not distinguish between a few classes it would be very easy to deal with by the Dempster-Shafer theory, because we can assign a mass function to a union of propositions (data classes here) by the Dempster-Shafer theory.

In the case when each rule assigns belief to singletons, we do not expect so much difference in comparison with the approach in which just we add the beliefs of the same data classes. However we run all the experiments reported earlier with the new strategy of combining different rules, the Dempster's combination rule.

Table 6.2 shows the result of this group of experiments using Dempster's combination rule. As it can be seen the number of correctly classified sites is not very different from the result of the previous group of experiments. All conditions of experiments are the same except in Exp IV.1 where we used $W_{RP} = \{0.1, 0.1, 0.4, 0.3, 0.1\}$ and $W_{SE} = \{0.1, 0.4, 0.2, 0.2, 0.1\}$ and also in Exp IV.3 where we used $W_{RP} = \{0.1, 0.1, 0.5, 0.3, 0.0\}$ and $W_{SE} = \{0.1, 0.1, 0.4, 0.4, 0.0\}$.

6.5 Conclusions

The problem of assessing the risk of desertification using four variables, slope, aspect, soil depth and rock permeability was discussed. By using some rules which were extracted from verbal rules given by the experts it was possible to find the natural regeneration potential and the risk of soil erosion. By some other rules, again extracted from verbal rules, assessments for the natural regeneration potential and the risk of soil erosion were used to assess the risk of desertification of the site under study. The problem is that the results of classification by the experts mostly disagree with those which are obtained by the expert rules.
Table 6.2: Experimental results. Under CCS the number of the sites which were classified in the same class as that assigned by the expert. Under CCSE is the number of sites that were correctly classified or misclassified by one class.

If we assume that the classification of experts is reliable, because it is the only reference which is available to us, we should accept that the expert rules are reliable too, since the experts must have used these rules in their assessment directly or indirectly. However, it is plausible that the experts are not very good in expressing the rules, especially when the original rules are verbal and expressed with the natural language which includes vague concepts such as high, low etc. Another important factor is that the experts used the ground data in their assessment for the natural regeneration potential, the risk of soil erosion and eventually the risk of desertification. This seems to be the main source of discrepancies between our results and the experts' results. While we use secondary data which are derived from a GIS package, experts used ground data. There are different sources of error in the processing of data such as digitisation error, measurement error, interpolation error etc. Among these factors the interpolation error is the most important.

Usually all data are not available in the same resolution and it is impossible or very difficult to collect all ground data with the desired resolution. Usually the data are collected (acquired) at the points which are possible (practically or economically). Other points of interest are interpolated by some usual interpolation methods which are available in most GIS packages such as the Delaunay triangulation method, the Kriging method etc. During estimation of an unknown variable from the known values at its neighbourhood, error is committed. The amount of error depends on different factors which were studied for the slope and aspect in chapter 2.

The uncertainty of data because of the interpolation error cause error in the decision making
6.5. Conclusions

step. In chapter 2 we proposed some empirical formulae for the modelling of the distribution of error in the slope and aspect when computed from interpolated data. In this chapter we used these formulae to calculate the probability of a given site belonging to each data class. Introducing these models we obtained more reliable mass functions which translated into getting marginally better results in comparison with experiment I in which we assumed that the given data were absolutely reliable.

Such models were not available for soil depth and on the other hand we did not have access to the exact values of soil depth at pixel level. So it was impossible to apply the same procedure for soil depth. Imitating the process of using error models to account for the uncertainty in the data, we redistributed the mass function for soil depth. Experiments showed that this enhances the results of classification. Since the best results were obtained with some unreasonable weights, we concluded that the weights were compensating not only for errors in soil depth, but also for other sources of errors. Using some heuristic restrictions, we could get better results in comparison with those obtained when we used equal (unit) weights for all the soil depth data classes. However we could not get the best possible results.

To improve the classification rate of the risk of desertification other factors can be considered such as importance of the different agents (or data) in the process of decision making. For example, the importance of different classes of the natural regeneration potential and the risk of soil erosion may be different in the process of decision making for the risk of desertification. Experiment IV showed that this assumption can be true since using different importance factors could improve the accuracy of classification.

We conclude that when the data are associated with errors, using expert rules can not be enough to get accurate results. Instead, using information about the uncertainty of data and also using different importance factors in the decision making process can improve the classification result.
Chapter 7

Artificial Neural Network

Classifiers

In this chapter one class of neural networks and one class of fuzzy neural networks, which are used to classify the risk of desertification, is discussed. As it has been clear so far, the data at hand are noisy and associated with errors. That is why using expert rules can not lead us to the same classification as that of the experts for most sites. Therefore we decided to investigate whether anything can be gained if we do not use expert rules explicitly but extract the input-output relations by training a neural network. Also we implemented a class of fuzzy neural networks that has capability to learn from examples and expert rules.

In section 7.1 we discuss briefly the fundamentals of multilayer feedforward networks. In section 7.2, first we present the neural network that gives directly the result of risk of desertification. This is the first method of handling the problem. Then in the second method we follow the way that experts advised us to tackle this problem: we construct two independent neural networks to classify the natural regeneration potential and risk of soil erosion. The outputs of these two networks are applied to another neural network to classify the risk of desertification. The results of both methods are presented for 39 training sites and 14 test sites. In section 7.3 we discuss the use of a fuzzy neural network in this application. All network parameters are real numbers as in conventional neural networks, however the input signals can be fuzzy concepts such as gentle, steep etc. We fuzzify all variables which are not given in fuzzy form, because the experts used vague concepts to draw their conclusions. Then we use a special feedforward back propagation algorithm which has been proposed by Ishibuchi at al. [24] for fuzzy variables. Experimental results
of implementing this fuzzy neural network using both methods of classifying the risk of desertification will be presented in section 7.4. Fuzzy IF – THEN rules may also be used to train this fuzzy neural network. The results of applying expert rules in the training phase will be explained later. Finally, discussion and conclusions are presented in section 7.5.

7.1 Multilayer Perceptrons

Multilayer perceptrons have been used in many classification applications, pattern recognition and expert systems. Usually the network contains one sensory layer which serves as input layer, an output layer of computation nodes and one or more hidden layers of computation nodes which are not part of the input or output layer and enable the network to learn complex applications. The input signal propagates through the network layer by layer. This kind of neural network is usually called multilayer perceptron (MLP) [20]. The algorithm which is usually used to train the network is called error back-propagation algorithm. This algorithm consists of two passes, forward and backward. In the forward pass the input signal passes through the network layer-by-layer. The synaptic weights are fixed in this pass. The difference between the network output and the desired output is called the error signal. In the backward pass the error signal is used to adjust the synaptic weights to make the error less and less. In this phase the error signal propagates in the opposite direction to the input signal. The algorithm which is used to adjust the synaptic weights is called back-propagation algorithm [20, 40]. The basis of the algorithm to update the synaptic weights is the gradient-descent method.

Each neuron in the network passes its input through a nonlinear function which is differentiable. Sigmoidal nonlinearity is commonly used for this propose and is defined by the logistic function:

\[ y_j = \frac{1}{1 + \exp(-net_j)} \]  

(7.1)

where \( net_j \) is the net input to the neuron and \( y_j \) is the output. Figure 7.1 shows a fully connected multilayer perceptron with two hidden layers.

The input signal propagates layer-by-layer from left to right. At each neuron the input values are multiplied with weights and summed up to form \( net \) which is passed through the sigmoid function to form the output of the neuron. The error signal on the other hand propagates from the output layer to the left layer-by-layer.
7.1. Multilayer Perceptrons

7.1.1 Algorithm

Let $t_j(n)$ and $y_j(n)$ be the desired and actual output at neuron $j$ respectively, for the $n$th training pattern. So the error signal is defined as:

$$e_j(n) = t_j(n) - y_j(n)$$

The sum of the squared errors of the network are:

$$E(n) = \frac{1}{2} \sum_{j \in c} e_j^2(n)$$

where set $c$ includes the whole output layer of the network. The average sum of the error signal over $N$ examples, $E_{av}(n) = \frac{1}{N} \sum_{n=1}^{N} E(n)$ can be defined as a cost function which is a measure of the training set learning performance. Minimising $E_{av}$ is the objective of the learning process. This task is done by adjusting the weights with which we multiply the input values to the individual neurons, on a pattern-by-pattern basis or in a batch mode.

It can be shown that the amount of adjustment for the weight $w_{ji}$ which is the weight
connecting the output of neuron $i$ to the input of neuron $j$ is given by [20]:

$$\Delta w_{ji}(n) = \eta \delta_j(n)y_i(n)$$  \hspace{1cm} (7.4)

where $\eta$ is the learning rate and $\delta_j(n)$ is the local gradient and is equal to:

$$\delta_j(n) = e_j(n)\varphi'_{j}(v_j(n))$$  \hspace{1cm} (7.5)

where $\varphi'_{j}$ is the derivative of the activation function of neuron $j$.

When $j$ is the output neuron it is very simple to compute formula 7.5. However, when neuron $j$ is in a hidden layer it gets more complicated because there is no desired output at the hidden layers. It can be shown that in this case we have:

$$\delta_j(n) = \varphi'_{j}(v_j(n)) \sum_k \delta_k(n)w_{kj}(n)$$  \hspace{1cm} (7.6)

where the sum is over all nodes in the layer to the right of node $j$ (see figure 7.1).

Sometimes it is better to use an improved version of equation 7.4 as follows

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n - 1) + \eta \delta_j(n)y_i(n)$$  \hspace{1cm} (7.7)

where $\alpha$ is called the momentum constant and usually is a positive number. This improvement makes the network more stable when the learning rate is selected too large to speed up the rate of learning [59].

### 7.1.2 Generalisation and Training Data

When a network has been trained, for example by the back propagation algorithm, we expect it to be able to interpret a new input which is fed. In other words the network should be able to establish a correct input-output relationship. This property is said to be the property of good generalisation.

We may think of the learning process as a curve fitting process. If the curve is properly fitted to the training data, it can interpolate new data as well. However, if the fitting curve is over-fitting, it can not handle new data. Usually if there are too many parameters to be trained for, and a limited number of patterns, over-fitting occurs. On the other hand, if these are too few parameters to be adjusted, there is no room for the network to learn and it acts very poorly on the test set. So, one of the key issues is the number of parameters and the number of patterns in the training set. There are some rules of thumb for the number of training examples [20, 40].
7.2 Experimental Results of the Neural Network Approach

In this section we demonstrate the neural network which we have used to classify the risk of desertification. We used a multilayer perceptron network with one hidden layer.

7.2.1 Method I

In the first experiment we set up one network to classify the risk of desertification directly. Variables, which in this study serve as the input signal, are slope, aspect, soil depth and rock permeability. These variables have already been defined and classified into data classes. For example, slope is classified into classes gentle, medium and steep. In a preprocessing stage, the value of the membership function of a given slope of a pixel, for example, is calculated by using a Gaussian membership function. After calculating the values of the membership functions of all pixels of a site, the mean value of the membership functions of each data class is calculated and assigned to the whole site. At this stage the error distributions of the variables slope and aspect are considered in the calculation of the membership functions. However, the values of the rock permeability and soil depth are not available but only the class to which each pixel belongs. For these two variables we use the proportion of pixels which belong to a data class. For example, for the site number 1 we know that 11% of the pixels belong to class shallow and 89% belong to class deep of the soil depth classes. These proportions are used as membership functions of this site to the classes shallow and deep.

From the above discussion and also from previous discussions on data specifications, it is obvious that we have 12 different data classes. Therefore the size of the input signal or feature vector is 12 x 1. Each feature is the membership function of the site to the relevant data class which is a real number. Thus the neural network contains 12 input nodes. The number of output nodes is the same as the number of classes, which is 5.

As in previous chapters 39 sites are used to train the network and 14 sites are used to test it. Since several factors affect the convergence of the back propagation algorithm, such as initial weights, the learning constant, the number of hidden layers and the number of nodes in the hidden layers, we chose different combinations of these parameters. In table 7.1 the network parameters are given. To initialise the weights we use a uniform random number generator. Since the final results depend on the initial weights, to stabilise the results we use the mean result of 20 different runs with different initial weights. Since the output of
Table 7.1: Experimental results for the risk of desertification. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.

<table>
<thead>
<tr>
<th>Number of epochs</th>
<th>Number of hidden nodes</th>
<th>( \eta )</th>
<th>( \alpha )</th>
<th>CCS (39 sites)</th>
<th>CCSE (39 sites)</th>
<th>CCS (14 sites)</th>
<th>CCSE (14 sites)</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>2</td>
<td>0.1</td>
<td>0.35</td>
<td>27</td>
<td>35</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>1000</td>
<td>2</td>
<td>0.3</td>
<td>0.35</td>
<td>32</td>
<td>37</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>1000</td>
<td>5</td>
<td>0.6</td>
<td>0.35</td>
<td>36</td>
<td>37</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

a neural network can be interpreted as a post-probability, the output with the maximum value is selected. In table 7.1 results of this experiment for the training set are given under the “39 sites” column and for the test set under the “14 sites” column.

### 7.2.2 Method II

In the second experiment we classify the risk of desertification in two steps. First we find the natural regeneration potential and the risk of soil erosion separately by using two independent neural networks. The first one has 7 input nodes, 4 for aspect and 3 for soil depth and 5 output nodes, and the second one has 8 input nodes, 3 for slope, 3 for soil depth and 2 for rock permeability and 5 output nodes. The outputs of these two networks are presented to another neural network which has 10 input nodes and 5 output nodes. The outputs of this network are the classes of the risk of desertification. The training of each network is done separately and independently. Table 7.2 shows the results of this experiment for two different sets of parameters.

### 7.3 Neural Network with Fuzzy Training

In this section we discuss one kind of fuzzy neural network in which fuzzy logic is incorporated into the neural network model. From all different kinds of fuzzy neural networks we discuss only the one proposed by Ishibuchi et al. [24] which can be trained by fuzzy inputs. This network not only can be trained by fuzzy input numbers, but also by fuzzy IF–THEN rules. First we present an overview of this network. Then we present its results when applied to our problem.
### Table 7.2: Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD). CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>RP</th>
<th>SE</th>
<th>RD</th>
<th>RP</th>
<th>SE</th>
<th>RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of hidden nodes</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>No of epochs</td>
<td>230</td>
<td>230</td>
<td>230</td>
<td>400</td>
<td>200</td>
<td>400</td>
</tr>
<tr>
<td>Momentum const. ((\alpha))</td>
<td>0.5</td>
<td>0.5</td>
<td>0.35</td>
<td>0.5</td>
<td>0.65</td>
<td>0.8</td>
</tr>
<tr>
<td>Learning rate ((\eta))</td>
<td>0.9</td>
<td>0.6</td>
<td>0.06</td>
<td>0.1</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>CCS (39 sites)</td>
<td>24</td>
<td>21</td>
<td>26</td>
<td>24</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>CCSE (39 sites)</td>
<td>36</td>
<td>28</td>
<td>36</td>
<td>35</td>
<td>21</td>
<td>35</td>
</tr>
<tr>
<td>CCS (14 sites)</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>CCSE (14 sites)</td>
<td>13</td>
<td>5</td>
<td>9</td>
<td>13</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

7.3.1 Fundamentals of Fuzzy Neural Networks

The problem that we are trying to solve is training a neural network when a fuzzy vector is presented to the input. Shortly we shall show that arithmetic operations on fuzzy numbers reduce to arithmetic operations on closed intervals. So we first review some equations from interval arithmetic [34]. Let \(X = [x^L, x^U]\), \(Y = [y^L, y^U]\) be intervals and \(k\) be a real number. We have:

\[
X + Y = [x^L + y^L, x^U + y^U]
\]  
(7.8)

\[
kX = \begin{cases} 
[kx^L, kx^U] & \text{if } k \geq 0 \\
[kx^U, kx^L] & \text{if } k < 0
\end{cases}
\]  
(7.9)

The activation function when the input is an interval can be defined as:

\[
f(\text{net}) = f([\text{net}^L, \text{net}^U]) = [f(\text{net}^L), f(\text{net}^U)]
\]  
(7.10)

where \(\text{net}\) is an interval input and \(f(.)\) is the logistic (sigmoid) function.

Relying on the extension principle it is possible to define calculations on fuzzy numbers [47, 85, 86]. Let us assume that we wish to add two fuzzy numbers \(A\) and \(B\) so that \(C = A + B\). For each value \(x\) of \(A\) we have a value of the membership function \(\mu_A(x)\) and for each value \(y\) of \(B\) we have a value of the membership function \(\mu_B(y)\). We wish to assign
a value to the membership function for each value $z$ of $C$: $\mu_C(z)$. As there will be many different pairs $(x, y)$ which lead to the same value $z$, we apply the min–max rule: For each pair $(x, y)$ which satisfies $x + y = z$, we choose the minimum between $\mu_A(x)$ and $\mu_B(y)$. Then among all these minima corresponding to different pairs $(x, y)$ that lead to the same value $z$, we choose the maximum and assign it to $\mu_C(z)$.

Pictorially this is shown in figure 7.2. Suppose that we wish to find $\mu_C(12)$ for $A$ and $B$ with membership functions as shown in the figure. As $x + y = 12$, and $x, y \geq 0$, $x$ and $y$ must be in the range $0 \leq x, y \leq 12$. For fixed $y$, $x = 12 - y$ and we have to choose $\mu = \min(\mu_B(y), \mu_A(12 - y))$. For $y \leq 6$, $\mu_B(y) = 0$ and so $\mu = 0$. For $y \geq 8$, $\mu_A(12 - y) = 0$ and again, $\mu = 0$. For $6 \leq y \leq 7$, $x$ must be $5 \leq x \leq 6$ and $\mu = \mu_A(12 - y)$ while for $7 \leq y \leq 8$, $x$ must be $4 \leq x \leq 5$ and $\mu = \mu_B(y)$. The maximum of all these values of $\mu$.
is 0.5, and this is the membership value $\mu_C(12)$. These process has to be repeated for all
values of $z$ in order to form the complete function $\mu_C(Z)$. Formally this is expressed as:

$$C(z) = (A \oplus B)(z) = \sup_{x,y; z=x+y} [A(x) \land B(y)]. \quad (7.11)$$

or

$$\mu_{A+B}(z) = \max \{ \mu_A(x) \land \mu_B(y) : z = x + y \} \quad (7.12)$$

where the symbols $\oplus$ and $\land$ denote the arithmetic addition and the intersection fuzzy operator (minimum) respectively and $sup$ is supremum. Generally four basic arithmetic operations (addition, subtraction, multiplication and division) on fuzzy numbers can be defined in a similar way. If $*$ denotes any of the basic arithmetic operations we have:

$$C(z) = (A * B)(z) = \sup_{x,y; z=x*y} [A(x) \land B(y)]. \quad (7.13)$$

We can also define the membership function for a variable that is the result of the multiplication of a real number, $k$, with a fuzzy number, as follows:

$$\mu_{kA}(z) = \{ \mu_A(x) : z = kx \} \quad (7.14)$$

This means that when a fuzzy number is multiplied by a real number, the membership function of the product at point $z$ is equal to the membership function of the fuzzy number at point $x$ when variable $x$ is scaled inversely by the same factor.

When a fuzzy number is used as input to an activation function $f$, we have:

$$\mu_{f(net)}(z) = \{ \mu_{\text{net}}(x) : z = f(x) \} \quad (7.15)$$

where $\text{net}$ and $f(\text{net})$ are a fuzzy input and a fuzzy output. This means that the membership function of the output of the activation function at a point $z$ is equal to the membership function of the input at point $x$ with constraint $z = f(x)$.

Any fuzzy number can be considered as a family of nested $h$-levels or $h$-cuts. An $h$-level of a fuzzy number $A$ is a crisp set all the elements of which have membership grade in $A$ greater or equal to $h$. That is,

$$A_h = \{ x : \mu(x) \geq h \} \quad \text{for} \quad 0 < h \leq 1$$

As all numbers in set $A_h$ have membership function with value at least $h$, if we add two such sets $A_h$ and $B_h$, we shall find a set of numbers with membership value also at least $h$
according to the rules of addition mentioned earlier. Hence from equations 7.11 to 7.15 we can derive the following relations:

\[
[A + B]_h = A_h + B_h \tag{7.16}
\]

\[
[kA]_h = kA_h \tag{7.17}
\]

\[
[f(A)]_h = f(A_h). \tag{7.18}
\]

where \([\cdot]_h\) is \(h\)-level or \(h\)-cut set of a fuzzy number. For example, formula 7.16 denotes that the \(h\)-level set of the arithmetic addition of two fuzzy numbers is equal to the arithmetic addition of their \(h\)-levels. Note that the \(h\)-level set of a fuzzy number is an interval valued number.

This way a fuzzy membership function which is a continuous valued function, is converted into a series of intervals, one for each \(h\)-level into which we quantise it. For each \(h\)-level then we perform the interval arithmetic we described earlier in order to process it through the network.

Suppose that we have an \(n\)-dimensional fuzzy vector which should be classified into one of \(c\) classes. We employ a multi-layer feedforward neural network which has \(n\) input nodes, one hidden layer with \(m\) nodes and \(c\) output nodes. Operations which are carried out by a neural network unit are multiplication by synaptic weights, summation over all connected links and eventually calculation of the output of the activation function. Using equations 7.11 to 7.18 we can carry out these operations by using the \(h\)-levels of the fuzzy vector. The \(h\)-level of a fuzzy number coincides with an interval. So actually we should operate on intervals instead of fuzzy numbers. In other words, given a fuzzy input vector, we use input-output relations for operation on its \(h\)-levels which are simply intervals. We can do this for as many \(h\)-levels as is needed. By this method the problem of training by fuzzy numbers is reduced to training by intervals.

Now suppose that the interval vector \(X_p = (x_{p1}, \ldots, x_{pn})\) is presented to the input of the neural network. Here \(p = 1, 2, \ldots, s\) and it represents the \(p\)th pattern. The input-output relations of each unit can be expressed as follows:

**Input nodes:** \(Y_{pi} = [y_{pi}^L, y_{pi}^U] = X_{pi} = [x_{pi}^L, x_{pi}^U] \quad \text{for } i = 1, \ldots, n \tag{7.19}\)

**Hidden nodes:** \(Y_{pj} = [y_{pj}^L, y_{pj}^U] = [f(\text{net}_{pj})^L, f(\text{net}_{pj})^U] \quad \text{for } j = 1, \ldots, m \tag{7.20}\)
### 7.3. Neural Network with Fuzzy Training

#### 7.3.1. Learning Algorithm

If we denote by \( t_{pk} \) the desired output for the \( p \)-th pattern at the \( k \)-th output node, we can define the cost function as:

\[
E_{ph} = \sum_{k=1}^{c} \max \{ (t_{pk} - [Y_{pk}^L]^2)/2, (t_{pk} - [Y_{pk}^U]^2)/2 \}
\]  

(7.26)

where the maximum is over the two values of squared error calculated for the lower and upper part of the output interval. If the input is a real vector, the output vector, \( Y_{pk} \) will be a real vector too, and the cost function reduces to the squared error as in the case of the back propagation algorithm.

The learning algorithm involves minimising the cost function. As in a conventional neural network the updating of the weights between the output and the hidden layer is done using

\[
\Delta w_{kj}(n + 1) = \alpha \Delta w_{kj}(n) + \eta (-\partial E_{ph}/\partial w_{kj})
\]

(7.27)

and between the hidden and the input layer layer is done using

\[
\Delta w_{ji}(n + 1) = \alpha \Delta w_{ji}(n) + \eta (-\partial E_{ph}/\partial w_{ji})
\]

(7.28)

where \( \alpha \) is the momentum constant parameter and \( \eta \) is the learning rate parameter [59, 58].

Similar formulae are used to update the biases \( \theta_j \) and \( \theta_k \) which appear in equations 7.21-7.25. The two terms \( \partial E_{ph}/\partial w_{kj} \) and \( \partial E_{ph}/\partial w_{ji} \) can be calculated as follows [24]:

1) \( \partial E_{ph}/\partial w_{kj} \):
1. If $t_{pk} = 1$ and $w_{kj} \geq 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}} \{(t_{pk} - y_{pk}^L)^2/2\} = \frac{\partial}{\partial y_{pk}^L} \{(t_{pk} - y_{pk}^L)^2/2\} \frac{\partial y_{pk}^L}{\partial \text{net}_{pk}^L} \frac{\partial \text{net}_{pk}^L}{\partial w_{kj}} = (t_{pk} - y_{pk}^L) y_{pk}^L (1 - y_{pk}^L) y_{pj}^L
\nonumber
$$

$$
= -\delta_{pk}^L y_{pj}^L
$$

(7.29)

2. If $t_{pk} = 1$ and $w_{kj} < 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{kj}} = -\delta_{pk}^U y_{pj}^U
$$

(7.30)

3. If $t_{pk} = 0$ and $w_{kj} \geq 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}} \{(t_{pk} - y_{pk}^U)^2/2\} = \frac{\partial}{\partial y_{pk}^U} \{(t_{pk} - y_{pk}^U)^2/2\} \frac{\partial y_{pk}^U}{\partial \text{net}_{pk}^U} \frac{\partial \text{net}_{pk}^U}{\partial w_{kj}}
\nonumber
$$

$$
= (t_{pk} - y_{pk}^U) y_{pk}^U (1 - y_{pk}^U) y_{pj}^U
\nonumber
$$

$$
= -\delta_{pk}^U y_{pj}^U
$$

(7.31)

4. If $t_{pk} = 0$ and $w_{kj} < 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{kj}} = -\delta_{pk}^U y_{pj}^L
$$

(7.32)

II) $\partial \varepsilon_{ph}/\partial w_{ji}$:

1. If $t_{pk} = 1$, $w_{kj} \geq 0$ and $w_{ji} \geq 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}} \{(t_{pk} - y_{pk}^L)^2/2\} = \frac{\partial}{\partial y_{pk}^L} \{(t_{pk} - y_{pk}^L)^2/2\} \frac{\partial y_{pk}^L}{\partial \text{net}_{pk}^L} \frac{\partial \text{net}_{pk}^L}{\partial w_{ji}}
\nonumber
$$

$$
= -\delta_{pk}^L w_{kj} y_{pj}^L (1 - y_{pj}^L) y_{pi}^L
$$

(7.33)

2. If $t_{pk} = 1$, $w_{kj} \geq 0$ and $w_{ji} < 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{ji}} = -\delta_{pk}^L w_{kj} y_{pj}^L (1 - y_{pj}^L) y_{pi}^U
$$

(7.34)

3. If $t_{pk} = 1$, $w_{kj} < 0$ and $w_{ji} \geq 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{ji}} = -\delta_{pk}^L w_{kj} y_{pj}^U (1 - y_{pj}^U) y_{pi}^U
$$

(7.35)

4. If $t_{pk} = 1$, $w_{kj} < 0$ and $w_{ji} < 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{ji}} = -\delta_{pk}^L w_{kj} y_{pj}^U (1 - y_{pj}^U) y_{pi}^L
$$

(7.36)

5. If $t_{pk} = 0$, $w_{kj} \geq 0$ and $w_{ji} \geq 0$:

$$
\frac{\partial \varepsilon_{ph}}{\partial w_{ji}} = \frac{\partial}{\partial w_{ji}} \{(t_{pk} - y_{pk}^U)^2/2\} = \frac{\partial}{\partial y_{pk}^U} \{(t_{pk} - y_{pk}^U)^2/2\} \frac{\partial y_{pk}^U}{\partial \text{net}_{pk}^U} \frac{\partial \text{net}_{pk}^U}{\partial w_{ji}}
\nonumber
$$

$$
= -\delta_{pk}^U w_{kj} y_{pj}^U (1 - y_{pj}^U) y_{pi}^U
$$

(7.37)
7.4 Experimental Results of the Fuzzy Neural Network Approach

In this section we apply the above fuzzy neural network to the problem of assessing the risk of desertification. Although the expert rules which are available are actually fuzzy rules, some of the variables are not given in fuzzy form. For example, we have the value of aspect and slope at all pixels of the site, however, we have classes for soil depth and rock permeability. Although in the above fuzzy neural network there is no obligation to have all inputs in fuzzy form, since in the expert rules all variables are used in fuzzy form we changed aspect and slope to fuzzy numbers too. This task is done by the following few steps. First, the mean of the variables for each site is calculated. Calculation of mean of slope is very easy. However mean of the aspect can not be calculated just by simple averaging, because mean of, for example, 0° and 360° is 180°, but we know that this is not the correct answer. Averaging of the aspect can be calculated by using algebraic sum of the vectors which are defined in the same direction as the aspect values.

When the mean values of the variables are available we can classify them into the predefined classes (see chapter 5 for the definition of data classes). In this approach we classify the
site to one of the predefined classes to which the mean value of the variable belongs to. As usual we have two methods for assessing the risk of desertification. In the first method all variables are used to assess the risk of desertification. In the second method the natural regeneration potential and the risk of soil erosion are classified independently and in parallel, then the results of them are used to assess the risk of desertification.

7.4.1 Method I

Given the mean of all variables we construct 4 feature vectors, one for each of the variables slope, aspect, soil depth and rock permeability. We use just the class of the site under study as the fuzzy input vector. We are going to use these fuzzy classes to train a fuzzy neural network which obviously contains 4 input nodes, one for each variable, and 5 output nodes, one for each output class. It is worth mentioning that this method of defining features is totally different from all methods which we have considered so far. This is because we have one input for each variable. For example, for slope we have a single input node and not a separate one for each of its classes gentle, medium and steep. In the past we had one input node for each class of a variable.

When we are going to process fuzzy numbers, such as gentle, we should define its membership function. We used the π function and the S function for defining the fuzzy numbers. These functions are defined respectively by

\[
S(x; a, b) = \begin{cases} 
0 & \text{for } x < a \\
2\left(\frac{x-a}{b-a}\right)^2 & \text{for } a \leq x < \frac{a+b}{2} \\
1 - 2\left(\frac{x-b}{b-a}\right)^2 & \text{for } \frac{a+b}{2} \leq x < b \\
1 & \text{for } x \geq b 
\end{cases}
\]

\[
\pi(x; a, b) = \begin{cases} 
S(x; b-a, b) & \text{for } x \leq b \\
1 - S(x; b, b+a) & \text{for } x > b 
\end{cases}
\]

Figure 7.3 shows the defined membership functions for slope. The membership functions are used to calculate the h-levels. As it was mentioned before, the h-level of a fuzzy number is an interval. For a given h we first calculate the h-level of all variables then we present them to the input of the fuzzy neural network. In the presentation phase we use the input-output relations 7.20 to 7.25 to calculate the outputs of the neurons. The outputs, which are
7.4. Experimental Results of the Fuzzy Neural Network Approach

Figure 7.3: \( S \) and \( \pi \) functions are used to define the membership functions.

![Membership functions](image)

Table 7.3: Experimental results for the risk of desertification. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.

<table>
<thead>
<tr>
<th>Number of epochs</th>
<th>Number of hidden nodes</th>
<th>( \eta )</th>
<th>( \alpha )</th>
<th>CCS (39 sites)</th>
<th>CCSE (39 sites)</th>
<th>CCS (14 sites)</th>
<th>CCSE (14 sites)</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>10</td>
<td>0.5</td>
<td>0.7</td>
<td>26</td>
<td>35</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>300</td>
<td>10</td>
<td>0.1</td>
<td>0.5</td>
<td>26</td>
<td>35</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

In a second experiment we incorporate the expert rules in the training phase. Not only 39 sites are used in the training phase, but also all available expert rules are included. We have 20 \( IF - THEN \) expert rules that relate conditions of four variables to the class of risk of desertification (see table 7.7). We deal with each rule as if it were data from a site. Therefore, we have 59 patterns in the training set where 39 of them are real data from 39 sites and 20 of them are expert rules which may or may not be justified by the real data. If the 39 training sites are in agreement with the expert rules this might increase
the performance of the system, because the network is trained better with more consistent patterns.

Table 7.4 shows the experimental results. Results on the training set are given only for 39 sites and the expert rules are not accounted for, so the results can be compared with those of the other experiments. As it can be seen there is no such high improvement in comparison with the first experiment when we did not use the expert rules. This means that our data are not in full agreement with the expert rules. As it has already been discussed this is because of large errors in the data.

### 7.4.2 Method II

In the third experiment we solve the problem of risk of desertification in three steps. Three fuzzy neural networks are constructed for the natural regeneration potential, the risk of soil erosion and the risk of desertification. The Fuzzy neural network which is constructed for classifying the natural regeneration potential has only 2 inputs and the network for assessing the risk of soil erosion has 3 inputs. Outputs of these two networks are presented to the input of the third network which is used for assessing the risk of desertification. So the third network has 10 inputs and 5 outputs. Similar to the first experiment 5 h-levels are used for training the networks.

Table 7.5 shows the results of this experiment. As it can be seen, there is no big difference between these results and the results in table 7.2 where we used MLPs without incorporating the fuzzy concept.

In the fourth experiment we added the expert rules for training the network of the last experiment. In the training phase 12 expert rules are used for the natural regeneration
7.5 Discussion and Concluding Remarks

Table 7.5: Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD). CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>RP</th>
<th>SE</th>
<th>RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of hidden nodes</td>
<td>10</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>No of epochs</td>
<td>200</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>Momentum const. (α)</td>
<td>0.9</td>
<td>0.75</td>
<td>0.5</td>
</tr>
<tr>
<td>Learning rate (η)</td>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>CCS (39 sites)</td>
<td>21</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>CCSE (39 sites)</td>
<td>33</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td>CCS (14 sites)</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>CCSE (14 sites)</td>
<td>13</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

potential, 18 expert rules are used for the risk of soil erosion and 25 rules are used for the risk of desertification. The result of this experiment is shown in table 7.6. Results on training are given only for the 39 sites and the expert rules are not accounted for, so the results can be compared with those of the other experiments. As it can be seen, not only there is no improvement in comparison with the last experiment when expert rules were not used, but also there is a small degradation in performance. This is because of large errors in the data that make the expert rules and the data to be inconsistent.

7.5 Discussion and Concluding Remarks

In this chapter we discussed neural networks and fuzzy neural networks applied to the problem of assessing the risk of desertification. In previous chapters we discussed results of the fuzzy logic approach, the Bayesian network method and combination of these two by the Dempster-Shafer theory. As it has been clarified, the data at hand have many errors. Some sites while their features are the same or very close to each other, were classified by the experts into different classes. In these cases it is impossible to train a classifier, not only an expert system one, but even a human one. As it has been mentioned before, the expert's classification in this study was not based on the GIS data but on ground data. So
<table>
<thead>
<tr>
<th>Parameters</th>
<th>RP</th>
<th>SE</th>
<th>RD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of hidden nodes</td>
<td>10</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>No of epochs</td>
<td>200</td>
<td>200</td>
<td>500</td>
</tr>
<tr>
<td>Momentum const. (α)</td>
<td>0.75</td>
<td>0.65</td>
<td>0.75</td>
</tr>
<tr>
<td>Learning rate (η)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>CCS (39 sites)</td>
<td>21</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>CCSE (39 sites)</td>
<td>33</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td>CCS (14 sites)</td>
<td>5</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>CCSE (14 sites)</td>
<td>13</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 7.6: Experimental results for the natural regeneration potential (RP), the risk of soil erosion (SE) and the risk of desertification (RD) when expert rules are used. CCS is the number of sites which were classified in the same class as that assigned by the expert; CCSE is the number of sites that were correctly classified or misclassified by one class.

the only reason which can be claimed for incompatibility of the data and the expert results, and thus for the low classification rate, is the error in the data. The sources of errors have been discussed in chapter 2.

The results of the experiments using a neural network and also a fuzzy neural network show that using all variables in one step to decide about the risk of desertification is better than using intermediate results on the natural regeneration potential and the risk of soil erosion, although the results on the test set are more or less the same for all approaches.

The results from the fuzzy neural network show that we can have comparable results on the test and the training set with those obtained by the neural network approach, in spite of the fact that we lose much information when fuzzifying the variables slope and aspect. However, in the fuzzy neural network approach we can introduce fuzzy IF – THEN rules as well. Although with these data this did not help, it seems that this method can lead to better results because expert rules can be considered in the training of the network.

In the fuzzy neural network approach the input-output relations are extracted by presenting training patterns which can be real examples or IF – THEN rules. In applications in which the examples are in agreement with the fuzzy rules this leads to better results and much more realistic decision boundaries between different classes. Unfortunately, in our case, the data were not in agreement with the expert’s rules in many cases. We think that the fuzzy
neural network model proposed by Ishibuchi et al. [24] is a good approach to implement fuzzy rules by a neural network. However, such an implementation is much more complicated than implementing a conventional neural network.

<table>
<thead>
<tr>
<th>Soil depth</th>
<th>Aspect</th>
<th>Rock Permeability</th>
<th>Slope</th>
<th>ROD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep</td>
<td>North</td>
<td>Permeable</td>
<td>Gentle</td>
<td>NR</td>
</tr>
<tr>
<td>Deep</td>
<td>East</td>
<td>Permeable</td>
<td>Gentle</td>
<td>NR</td>
</tr>
<tr>
<td>Deep</td>
<td>South</td>
<td>Permeable</td>
<td>Medium</td>
<td>LR</td>
</tr>
<tr>
<td>Deep</td>
<td>West</td>
<td>Permeable</td>
<td>Medium</td>
<td>LR</td>
</tr>
<tr>
<td>Deep</td>
<td>South</td>
<td>Permeable</td>
<td>Steep</td>
<td>LR</td>
</tr>
<tr>
<td>Deep</td>
<td>West</td>
<td>Permeable</td>
<td>Steep</td>
<td>LR</td>
</tr>
<tr>
<td>Deep</td>
<td>South</td>
<td>Impermeable</td>
<td>Gentle</td>
<td>LR</td>
</tr>
<tr>
<td>Deep</td>
<td>West</td>
<td>Impermeable</td>
<td>Gentle</td>
<td>LR</td>
</tr>
<tr>
<td>Shallow</td>
<td>South</td>
<td>Permeable</td>
<td>Medium</td>
<td>MR</td>
</tr>
<tr>
<td>Shallow</td>
<td>West</td>
<td>Permeable</td>
<td>Medium</td>
<td>MR</td>
</tr>
<tr>
<td>Shallow</td>
<td>South</td>
<td>Permeable</td>
<td>Steep</td>
<td>MR</td>
</tr>
<tr>
<td>Shallow</td>
<td>West</td>
<td>Permeable</td>
<td>Steep</td>
<td>MR</td>
</tr>
<tr>
<td>Deep</td>
<td>South</td>
<td>Impermeable</td>
<td>Steep</td>
<td>MR</td>
</tr>
<tr>
<td>Deep</td>
<td>West</td>
<td>Impermeable</td>
<td>Steep</td>
<td>MR</td>
</tr>
<tr>
<td>Shallow</td>
<td>South</td>
<td>Impermeable</td>
<td>Gentle</td>
<td>HR</td>
</tr>
<tr>
<td>Shallow</td>
<td>West</td>
<td>Impermeable</td>
<td>Gentle</td>
<td>HR</td>
</tr>
<tr>
<td>Shallow</td>
<td>South</td>
<td>Impermeable</td>
<td>Medium</td>
<td>VHR</td>
</tr>
<tr>
<td>Shallow</td>
<td>West</td>
<td>Impermeable</td>
<td>Medium</td>
<td>VHR</td>
</tr>
<tr>
<td>Shallow</td>
<td>South</td>
<td>Impermeable</td>
<td>Steep</td>
<td>VHR</td>
</tr>
<tr>
<td>Shallow</td>
<td>West</td>
<td>Impermeable</td>
<td>Steep</td>
<td>VHR</td>
</tr>
</tbody>
</table>

Table 7.7: Combined rules for Risk of Desertification. ROD: Risk of desertification, NR: No Risk, LR: Low Risk, MR: Moderate Risk, HR: High Risk, VHR: Very High Risk. Taken from [60]
Chapter 8

Summary and Conclusions

The main motivation of this work was to evaluate the risk of desertification of burnt forests. Different sources of data were combined to evaluate the natural regeneration potential and the risk of soil erosion. The result of these two decisions were used to assess the risk of desertification. However, in another method original data from sources were used directly to assess the risk of desertification. Information fusion in remote sensing was discussed in this study. First we had an overview of different approaches which had been used in the literature. We surveyed different information fusion methods such as statistical, neural network, Dempster-Shafer theory, fuzzy logic, rule based and Bayesian network approaches in multi source / multi sensor problems in chapter 3. From this study it can be concluded that each method has its own advantages and disadvantages. In other words there is no individual method which is superior to all other methods in information fusion. Depending on the problem and the data one of the combination methods may have better results. For example, in applications in which models of data distributions are available, perhaps statistical approaches are more desirable. In applications in which there is imprecision in the concepts, especially when they are expressed by human language, the fuzzy logic approach is preferred. Neural networks have good performance particularly in applications in which models are not available but there are enough training samples. In applications in which the relationship between variables can be expressed as causes and effects, it is better to employ Bayesian networks. Often different methods applied to the same problem produce different results. In such cases classifier combination may be used to increase the overall accuracy and reliability of the individual classifiers.

Uncertainty in the data because of errors is a great factor for inaccuracy of classification
Chapter 8. Summary and Conclusions

There are different kinds of error in the data. Errors due to the measuring instruments or sensors is one of them. The environment in which data are processed may also add noise to the data. Another kind of error is interpolation error. In some applications like GIS applications different sources of data are combined in a system, however, sometimes they have different resolution. The first step to draw a decision by combining the different layers of data is re-sampling all layers into the same resolution by interpolation. Interpolation is a process which is known to add error to the data. In this thesis we extensively studied this type of error which is common in GIS. Especially we were interested in the error in slope and aspect when they are derived from interpolated surfaces. Since slope and aspect are defined by the first derivative of the surface, error in the interpolation of the surface of the terrain may affect seriously the secondary data. We modelled this error versus the parameters which influence it. Roughness of the terrain and resolution of the original data and the method which is used to interpolate are factors which influence the error in slope and aspect of surfaces.

To study the effect of each factor we used different degrees of roughness of the terrain and sampling rates and also two popular interpolation methods. Collecting real terrains with the desired roughness is very difficult, so we used fractal models of the surface which computer graphics people use to create realistic looking surfaces. The generated fractals with different fractal dimensions which denote the degree of roughness of the terrains, were subsequently subsampled with different sampling rates. Then the surfaces of the terrains were reconstructed using either Delaunay triangulation method or Kriging interpolation method. By analysing the slope and aspect of the original fractal and the reconstructed one, error statistics were calculated. The error statistics of the slope and aspect were plotted in a three dimensional Cartesian system against the fractal dimension and the sampling rate variables. Then by a semiautomatic algorithm suitable functions were fitted to the curves of error statistics which were plotted in a two dimensional plane versus one of the variables while the other one was used as a parameter. Eventually error statistics were stated as functions of the two parameters, the fractal dimension and the sampling rate. These novel error models are applicable in any problem which uses slope and aspect of surface as features or attributes in the process. Models and distributions of the errors were also derived and expressed by parametric functions (Gaussian for the slope and Laplacian for the aspect). These distributions were used for the definition of the mass function, the belief function, the membership function and the prior probability. Also these distribution functions were
used in chapter 6 for redistribution of mass functions of an expert system.

The second important part of this thesis was studying different fusion of information approaches to tackle the problem of evaluating the risk of desertification. We studied many approaches and successfully implemented the Dempster-Shafer theory (as a classifier combination in chapter 5 and an expert system in chapter 6), the fuzzy logic approach, the neural network approach and the fuzzy neural network approach. However, I especially focused on the Dempster-Shafer evidence theory and its application in combining multi source data.

First I showed how to combine the two classifiers which in the past had been proposed to solve this problem. Outputs of the Bayesian network were available, but I had to implement the fuzzy logic approach to obtain the outputs since the membership grades of the outputs were not available. As discussed before, the first problem was the number of output classes which was 3 for the Bayesian network and 5 for the fuzzy logic approach. The reason why Stassopoulou et al. [70, 71, 72] could not cope with 5 classes was that this made the conditional matrix too big and untrainable with the limited number of data, so they reduced the number of classes. However, in the fuzzy logic approach there was not such limitation. When the fuzzy logic classifier was implemented, I came up with the same results that Sasikala et al. [60, 61, 62] had reported. However, I found that there is another problem with this method which was ignored by them. In many cases the classifier is blunt. It means that more than one outputs have equally maximum membership grades. In those cases if the expert result were found in the selected outputs, Sasikala et al. [60, 61, 62] considered the site as correctly classified. I thought that although this was not completely wrong, the classifier was not clear and sharp enough. I decided to combine these two classifiers to reduce the uncertainty of the results in having a coarse number of outputs in the Bayesian network approach and having fine but blunt outputs in the fuzzy logic approach. The Dempster-Shafer evidence theory is a good tool to solve this problem. I defined a superset of classes within which the outputs of both classifiers could be defined by the union of few supersets elements. Dealing with union of propositions is one the of advantages of the Dempster-Shafer theory. In addition, taking into consideration the uncertainty of the expert rules was straightforward in the Dempster-Shafer theory. The reliability of the Bayesian classifier which is reflected by its recognition rate was used in the combination classifier. The results showed that this proposal is effective in solving this kind of problem.

The application of Dempster-Shafer theory to an expert system was also studied in chapter
6. Propagation of belief functions through a hierarchical network based on expert rules was considered. Uncertainty of the expert rules and the data was taken into consideration. The effect of uncertainty in the data because of error was considered through the use of the error distribution models to update the mass functions. This idea was extended to redistribute the mass functions of variables for which no error models were available.

Finally, I implemented a neural network and a fuzzy neural network to solve the problem at hand in chapter 7. The fuzzy neural network used can be trained by \textit{IF} – \textit{THEN} rules. I tested this network in two cases: when the network is trained by a training set only and when the expert rules and training samples together are used for training. There was not significant change in the recognition rate. This shows that the data are not fully in agreement with the expert rules. This fact can be recognised easily by inspecting the data and comparing the results which are expected by using the expert rules and the results which were given by the experts using the ground data. The inconsistency of the data is partially due to errors in the measurements and the interpolation which were studied. Other sources of error and uncertainty in the data, if there are any, were not studied here.

8.1 Possible Applications and Suggestions

The error models which were proposed in this thesis are useful in any application in which slope and aspect are used as features or attributes and are computed from interpolated surfaces. Especially in GIS these models can be used effectively as an aid to enhance the system and present much more reliable results.

Our efforts to model the interpolation error in soil depth was not successful. The idea was to model the error in soil depth by two fractals placed one on the top of the other. The bottom fractal would simulate the rock surface while the top fractal would simulate the top soil surface with appropriate fractal dimension chosen for them. However we tried to theoretically develop a formula for the soil depth which was not successful. On the other hand simulations showed that the error statistics were not stable when different fractals with the same parameters were used. Real data also were not available to see if there is any stable relationship between the error in soil depth and the sampling rate and the fractal dimensions of surfaces of rock and top soil. The results of these experiments were not reported in the thesis.
The classifier combination scheme which was developed can be used in many fusion of information problems. Especially when final information classes are defined by some linguistic concepts which can be varied by different experts' opinions. This combination method also is easily applicable in problems like the multi neural network problems (modular approach), which operate by grouping the outputs and reducing the number of outputs of the problem by breaking it into a few smaller problems. The number of classes are reduced to ease the decision by simple neural networks. The results of the neural networks are then combined. In these problems the outputs of the sub-problems are usually the union of propositions of the final classification. Thus, using the Dempster-Shafer theory is very effective, allowing the reliability of the individual classifiers to be incorporated in a sophisticated way.
Appendix A

Fractional Brownian Motion

To prove equations 2.3 and 2.4 we start from equation 2.2 which is given here again:

\[ Pr\left[ \frac{Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})}{\|\Delta \bar{x}\|^H} < y \right] = F(y) \]  
(A.1)

If \( F(y) \) is assumed to be a zero-mean Gaussian distribution \( N(0, \sigma^2) \) with a variance \( \sigma^2 \), we have:

\[ Pr[\lambda < y] = \int_{-\infty}^{y} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} dt \]  
(A.2)

where \( \lambda = \frac{Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})}{\|\Delta \bar{x}\|^H} \) is a dummy variable. Probability density function can be derived as:

\[ p(\lambda) = \frac{dPr(\lambda)}{dy} = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{\lambda^2}{2\sigma^2}} \]  
(A.3)

The mean of \( |\lambda| \) can be calculated from its definition:

\[ < |\lambda| > = 2 \int_{0}^{\infty} \lambda p(\lambda) d\lambda = \frac{2}{\sigma \sqrt{2\pi}} \int_{0}^{\infty} \lambda e^{-\frac{\lambda^2}{2\sigma^2}} d\lambda = \frac{2\sigma}{\sqrt{2\pi}} \]  
(A.4)

So we have:

\[ < \frac{Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})}{\|\Delta \bar{x}\|^H} > = \frac{2\sigma}{\sqrt{2\pi}} \]  
\[ < |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})| > = \frac{2\sigma}{\sqrt{2\pi}} \|\Delta \bar{x}\|^H \]  
(A.5)

The mean of \( |\lambda|^2 \) also can be calculated from its definition:

\[ < |\lambda|^2 > = 2 \int_{0}^{\infty} \lambda^2 p(\lambda) d\lambda = \frac{2}{\sigma \sqrt{2\pi}} \int_{0}^{\infty} \lambda^2 e^{-\frac{\lambda^2}{2\sigma^2}} d\lambda = \sigma^2 \]  
(A.6)

The following formulae have been used in the computation of the above equation:

\[ J_r = \int_{0}^{\infty} x^r e^{-\alpha x^2} dx \]
Appendix A. Fractional Brownian Motion

\[ J_0 = \frac{1}{2} \sqrt{\frac{\pi}{a}}, \quad J_1 = \frac{1}{2a}, \quad J_2 = -\frac{dJ_0}{da}, \ldots \]

From A.6 we have:

\[ \frac{< |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 >}{\|\Delta \bar{x}\|^H} = \sigma^2 \]
\[ < |Z_H(\bar{x} + \Delta \bar{x}) - Z_H(\bar{x})|^2 > = \sigma^2 \|\Delta \bar{x}\|^{2H} \quad (A.7) \]
Appendix B

Kriging

In conventional statistics all variables are assumed to be independent so there is no continuity between them. On the other hand in some applications concerned with spatially distributed data, adjoining points have some correlation with each other. In geostatistics this continuity is measured and then is used in the estimation of unknown points. Continuity is measured with a function which is called Semivariogram or Variogram. Estimated values of this function are used in various estimation approaches such as Kriging. First some necessary definitions are given in the next section.

B.1 Definitions

**Scatterplot:** The most common representation of bivariate data is the scatterplot, which is displayed as an $x - y$ graph. The $x$-coordinate corresponds to one variable and the $y$-coordinate corresponds to the other one. Any correlation between two variables can be easily visualised by such a scatterplot. If both variables are the same, all the points in the scatterplot will be along the 45-degree line, but if they are not closely correlated, the points will form a very disperse cluster which will not be close to the 45-degree line. We would like to express any existing correlation by a few numbers, so it becomes easy to compare correlations expressed by different scatterplots. There are some parameters which are used to quantify correlation, with the **Correlation coefficient** being the most commonly used one.
This statistic can be calculated from:

\[
\rho = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y}
\]  

(B.1)

Where \( m_x \) and \( m_y \) are the mean values of variables \( x \) and \( y \) respectively and \( \sigma_x \) and \( \sigma_y \) are their standard deviations. The numerator of equation B.1 is called covariance:

\[
C_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)
\]  

(B.2)

**h-Scatterplot:** An \( h \)-scatterplot displays all possible pairs of samples whose locations are separated by a certain distance in a certain direction. The relative location of each pair of samples can be shown by a vector such as \( h \). A scatterplot in a Cartesian coordinate system, in which the \( x \)-coordinate represents the value of one sample and the \( y \)-coordinate the value of the corresponding sample at distance \( h \), is called \( h \)-scatterplot. Therefore this graph consists of a cluster of points like the scatterplot, but now each point does not represent the values of two variables of a single physical point, but the values of the same variable at two different locations. For example, if we plot values of a variable \( V(t) \) against its values at distance \( h \), \( V(t + h) \), \( V(t) \) will be plotted along the \( x \)-coordinate axis and \( V(t + h) \) along the \( y \)-coordinate axis. Every \( h \)-scatterplot is plotted for a certain distance and direction and expresses how much the samples are correlated to other adjoining samples over that distance and direction. Like in the case of scatterplot, features of the \( h \)-scatterplot can be represented by some statistics.

**Correlation function or correlogram:** This statistic represents the relationship between the correlation coefficient of an \( h \)-scatterplot and \( h \) and is denoted by \( \rho(h) \). Although \( h \) is a vector and \( \rho(h) \) can be displayed in a contour map, it is usual that \( \rho(h) \) is sketched against just the magnitude of \( h \) and its direction is mentioned in the plot. So we need to have as many plots as directions which we are interested in.

**Covariance function:** The relationship between the covariance of an \( h \)-scatterplot and \( h \) is called Covariance function and is denoted by \( C(h) \).

**Moment of inertia:** Half of the average squared difference between the \( x \) and \( y \) coordinate of each point in the \( h \)-scatterplot is called moment of inertia:

\[
\text{moment of inertia} = \frac{1}{2n} \sum_{i=1}^{n} (x_i - y_i)^2
\]  

(B.3)

This statistic is a measure of the spread of the cluster of points in an \( h \)-scatterplot.

**Semivariogram:** The relationship between the moment of inertia and \( h \) of an \( h \)-scatterplot
is called semivariogram or simply variogram and is denoted by $\gamma(h)$. In other words this function expresses how the perpendicular distance of the points from the 45-degree line ($x = y$) changes with $h$.

Although the above mentioned statistics are defined on the h-scatterplot, they can be calculated directly from the data values as follows [25]:

$$C(h) = \frac{1}{N(h)} \sum_{(i,j)|h_{ij}=h} v_i v_j - m_{-h} m_h$$

(B.4)

where $N(h)$ is the number of pairs of samples whose locations are separated by $h$ and $m_{-h}$ (or $m_h$) is the mean of all the sample points whose locations are $-h$ (or $h$) away from some other samples. Generally $m_{-h}$ and $m_h$ are not equal and they are defined as follows:

$$m_{-h} = \frac{1}{N(h)} \sum_{i:j, h_{ij}=h} v_i$$

$$m_h = \frac{1}{N(h)} \sum_{j:i, h_{ij}=h} v_j$$

(B.5)

The correlation function can be calculated from:

$$\rho(h) = \frac{C(h)}{\sigma_{-h} \sigma_h}$$

(B.6)

where

$$\sigma_{-h}^2 = \frac{1}{N(h)} \sum_{i:j, h_{ij}=h} v_i^2 - m_{-h}^2$$

$$\sigma_h^2 = \frac{1}{N(h)} \sum_{j:i, h_{ij}=h} v_j^2 - m_h^2$$

(B.7)

$\sigma_{-h}$ (or $\sigma_h$) is the standard deviation of all samples whose locations are $-h$ (or $h$) away from some other samples.

The variogram is calculated from the following formula:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h_{ij}=h} (v_i - v_j)^2$$

(B.8)

### B.2 Spatial Continuity Analysis

In the previous section some statistics were introduced for quantifying spatial continuity. Among them the variogram is the most commonly used. There are some terms used to define features of the variogram, and we shall define them here.
**Range:** When the distance between samples is increased, the variogram is increased as well. Increasing the distance beyond a specific value has no effect on the variogram. This specific distance is called Range.

**Sill:** The plateau that the variogram reaches at the Range is called Sill.

**Nugget effect:** Although it is expected that for very small values of distance the variogram approaches 0, in some cases because of some factors such as sampling error, the variogram approaches a nonzero value which is called nugget effect. The ratio of jump of the variogram (nugget effect) and sill is called relative nugget effect.

### B.3 Estimation

Many phenomena in the geostatistics are random, not in the sense that we do not know anything about them, but in the sense that they are too complicated and we only know just some features of them. Thus, a random variable in geostatistics is different from a random variable in conventional statistic. To predict the value of a variable at a point by using the values of the variable at existing points, we need to know how the phenomenon described by the variable behaves at points which are not sampled. Understanding the behaviour of a phenomenon is referred to as modelling.

There are two kinds of models: deterministic and probabilistic. Deterministic models are used if the context of data is well understood. In the probabilistic approach, the value of a sample is viewed as a result of a random process. So in this method we deal with samples as outcomes of random processes and we would like to estimate this outcome at unknown points.

Kriging uses information from the variogram to find the optimal set of weights with which the values of known samples should be linearly combined to estimate the value of the unknown sample. As criterion of optimality, the minimisation of the variation of the error is used. Kriging is named after D. G. Krige, a mining engineer, who used statistical techniques in his investigations. Kriging is synonymous with optimal prediction and is a Best Linear Unbiased Estimator (BLUE). In this approach we use linear combination of known samples to estimate the unknown at a point. So the estimated value \( \hat{v} \) at the point in question is given by:

\[
\hat{v} = \sum_{i=1}^{n} w_i v_i
\]
where \( v_i \) are the values at the known points and \( w_i \) are the weights we wish to specify. The set of weights may change at different locations. The difference between the estimated value and the true value is defined as error or residue.

\[
Error = Residue = r_i = \hat{v}_i - v_i
\]

If the random function is stationary it can be shown that the expectation value of the errors will be 0 if:

\[
\sum_{j=1}^{n} w_j = 1
\]

This condition guarantees unbiasedness of the estimation. The kriging method tries to minimise the variance of the error of estimation. If \( \sigma_R^2 \) is the error variance of \( K \) estimates, we have:

\[
\sigma_R^2 = \frac{1}{K} \sum_{i=1}^{K} (r_i - m_R)^2 = \frac{1}{K} \sum_{i=1}^{K} [(\hat{v}_i - v_i)^2 - \frac{1}{K} \sum_{j=1}^{K} (\hat{v}_j - v_j)^2]
\]  

If we assume unbiasedness, \( m_R = 0 \) and we have:

\[
\sigma_R^2 = \frac{1}{K} \sum_{i=1}^{K} (r_i)^2 = \frac{1}{K} \sum_{i=1}^{K} (\hat{v}_i - v_i)^2
\]  

The problem is that we do not know the true value \( v_i \) at point \( x_i \) to be able to minimise \( \sigma_R^2 \). Here we assume that these samples are the outcomes of a random process. So we start with \( n + 1 \) random variables; \( n \) of them model the behaviour of the phenomenon and the other one is the value of the variable at the point at which we are interested in estimating its value. If we combine \( n \) random variables for the estimation we have:

\[
\hat{V}(x_0) = \sum_{i=1}^{n} w_i V(x_i)
\]

Where \( V(x_i) \) are the values of random variables at points \( x_i, i = 1, \ldots, n \) and \( \hat{V}(x_0) \) is the estimated value at point \( x_0 \). We use \( V(x_i) \) to indicate the random variable defined at the point with position vector \( x_i \) and use \( v_i \) to indicate a specific value of this random variable. The difference between this estimated value and the true value is the error or residual.

\[
R(x_0) = \hat{V}(x_0) - V(x_0)
\]

It is better to clarify again that \( V(x_0), V(x_1), \ldots, V(x_n) \) are random variables which we have assumed to take the true values at all \( n + 1 \) samples and \( \hat{V}(x_0) \) is the estimated value at point \( x_0 \). Now we will try to minimise the error variance of the modelled error \( R(x_0) \).
Here we use some formulae from the theory of random variables. It can be shown that [25, 50]:

\[
\text{Var}\left\{ \sum_{i=1}^{n} w_i V_i \right\} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \text{Cov}(V_i V_j) \quad (B.13)
\]

So by using this formula we have:

\[
\text{Var}\{R(x_0)\} = \text{Var}\{\hat{V}(x_0) - V(x_0)\} \\
= \text{Cov}\{\hat{V}(x_0)\hat{V}(x_0)\} - 2\text{Cov}\{\hat{V}(x_0)V(x_0)\} + \text{Cov}\{V(x_0)V(x_0)\} \\
\]

The first term \(\text{Cov}\{\hat{V}(x_0)\hat{V}(x_0)\}\) is equal to the variance of \(\hat{V}(x_0)\) which can be written as:

\[
\text{Var}\{\hat{V}(x_0)\} = \text{Var}\{\sum_{i=1}^{n} w_i V(x_i)\} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} \\
\]

The second term is equal to:

\[
2\text{Cov}\{\hat{V}(x_0)V(x_0)\} = 2\text{Cov}\{\left(\sum_{i=1}^{n} w_i V_i\right)V_0\} \\
= 2E\{\sum_{i=1}^{n} w_i V_i V_0\} - 2E\{\sum_{i=1}^{n} w_i V_i\}E\{V_0\} \\
= 2\sum_{i=1}^{n} w_i \text{Cov}\{V_i V_0\} = 2\sum_{i=1}^{n} w_i \tilde{C}_{i0} \\
\]

Here we have used the usual notation of \(E\{\ldots\}\) to mean the expectation value and the assumption of unbiasedness to set \(E\{V_0\} = 0\). Finally if we assume that all the random variables have the same variance, \(\tilde{\sigma}^2\), the third term in equation B.14 is equal to \(\tilde{\sigma}^2\). If we substitute these values into Equation B.14 we have:

\[
\tilde{\sigma}^2_R = \tilde{\sigma}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} - 2\sum_{i=1}^{n} w_i \tilde{C}_{i0} \\
\]

In the above equation the variance of the modelled error is expressed as a function of \(n\) unknown weights, \(w_1, \ldots, w_n\), assuming the model is completely known. To minimise this error it is necessary to take the partial derivatives of the expression on the right hand side with respect to all variables, \(w_1, \ldots, w_n\), and set them equal to 0. Solving this system of \(n\) equations for the \(n\) variables we obtain the solution of minimising the variance of the error. However we have another constraint that says that the estimation must be unbiased so the summation of all weights must be equal to 1. Therefore the above mentioned system of equations is augmented and really we have \(n\) variables and \(n + 1\) equations. Solving this problem is possible by using the Lagrange parameter technique [25].

In the Lagrange parameter technique the system is converted to an unconstrained system by adding another parameter, \(\mu\), which is called Lagrange parameter.

\[
\tilde{\sigma}^2_R = \tilde{\sigma}^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \tilde{C}_{ij} - 2\sum_{i=1}^{n} w_i \tilde{C}_{i0} + 2\mu(\sum_{i=1}^{n} w_i - 1) \\
\]

(B.17)
If we set equal to 0 the partial derivative of the above expression with respect to variable \( \mu \), we obtain \( \sum_{i=1}^{n} w_i = 1 \), which is the unbiasedness condition. So we have \( n + 1 \) equations and \( n + 1 \) unknowns. This system can be solved easily.

By taking \( n + 1 \) partial derivatives of the above equation and set them equal to 0 we will have the following system:

\[
\begin{align*}
\sum_{j=1}^{n} w_j \tilde{C}_{ij} + \mu = \tilde{C}_{i0} & \quad \forall i = 1, \ldots, n \\
\sum_{i=1}^{n} w_i = 1
\end{align*}
\]

This system of equations is called the ordinary kriging system and can be written in matrix form as:

\[
\begin{bmatrix}
\tilde{C}_{11} & \cdots & \tilde{C}_{1n} & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\tilde{C}_{n1} & \cdots & \tilde{C}_{nn} & 1 \\
1 & \cdots & 1 & 0
\end{bmatrix}
\begin{bmatrix}
w_1 \\
\vdots \\
w_n \\
\mu
\end{bmatrix} =
\begin{bmatrix}
\tilde{C}_{10} \\
\vdots \\
\tilde{C}_{n0} \\
1
\end{bmatrix}
\]

The solution of the above system is the set of weights which satisfy the unbiasedness condition and also minimise the error variance. It can be shown that this minimum is equal to:

\[
\hat{\sigma}_R^2 = \hat{\sigma}^2 - W . D
\]

This minimum value of variance is called the ordinary kriging variance, and sometimes is denoted by \( \sigma_{OK}^2 \).

As we assumed that the mean and variance of all variables are the same, the ordinary kriging system can be represented in terms of the variogram and the correlogram as follows:

\[
\begin{align*}
\sum_{j=1}^{n} w_j \tilde{\gamma}_{ij} + \mu = \tilde{\gamma}_{i0} & \quad \forall \ i = 1, \ldots, n \\
\sum_{i=1}^{n} w_i = 1
\end{align*}
\]

\[
\begin{align*}
\sum_{j=1}^{n} w_j \tilde{\rho}_{ij} + \mu = \tilde{\rho}_{i0} & \quad \forall \ i = 1, \ldots, n \\
\sum_{i=1}^{n} w_i = 1
\end{align*}
\]

The value of the ordinary kriging variance can be expressed in terms of the variogram and the correlogram:

\[
\hat{\sigma}_R^2 = \sum_{i=1}^{n} w_i \tilde{\gamma}_{i0} + \mu = \hat{\sigma}^2 \left( \sum_{i=1}^{n} w_i \tilde{\rho}_{i0} + \mu \right)
\]

To solve the ordinary kriging system, it is necessary to find the covariance (or the correlogram or the variogram). So we must first decide about modelling the spatial continuity. There
are several spatial models for a random function appropriate for describing the height of a terrain: some commonly used ones are the nugget effect, exponential, spherical, circular, Gaussian, linear, power, logarithmic and the quadratic models. The exponential model is defined as follows:

\[
\hat{\gamma}(h) = \begin{cases} 
C_0 + C_1\left(1 - \exp\left(-\frac{3|h|}{a}\right)\right) & \text{if } |h| > 0 \\
0 & \text{if } |h| = 0
\end{cases} \tag{B.24}
\]

or

\[
C(h) = \begin{cases} 
C_1 \exp\left(-\frac{3|h|}{a}\right) & \text{if } |h| > 0 \\
C_0 + C_1 & \text{if } |h| = 0
\end{cases} \tag{B.25}
\]

where:

- \(C_0\) : is called the nugget effect and describes the discontinuity at the origin.
- \(C_0 + C_1\) : is called the sill and is the value of variogram for high \(h\), \(\gamma(\infty)\), and also is the value of the covariance at \(|h| = 0\), and the variance of the random variable, \(\sigma^2\).
- \(a\) : is the range and describes a distance over which the variogram or covariance remain steady. These functions are plotted in figure B.1. The spherical model is defined as follows:

\[
\hat{\gamma}(h) = \begin{cases} 
C_0 + C_1(1.5\frac{h}{a} - 0.5(\frac{h}{a})^3) & \text{if } |h| < a \\
C_0 + C_1 & \text{if } |h| \geq a
\end{cases} \tag{B.26}
\]

or

\[
C(h) = \begin{cases} 
C_0 + C_1(1.5\frac{h}{a} - 0.5(\frac{h}{a})^3) & \text{if } |h| < a \\
0 & \text{if } |h| \geq a
\end{cases} \tag{B.27}
\]

These functions are plotted in figure B.2. The Gaussian, the Linear and the power function models are defined and plotted in figures B.3 to B.6.
Figure B.1: *Exponential covariance function (a) and variogram function (b) for \( C_0 = 0, C_1 = 10 \) and \( a = 10 \).*

Figure B.2: *Spherical covariance function (a) and variogram function (b) for \( C_0 = 0, C_1 = 10 \) and \( a = 10 \).*
Figure B.3: Gaussian variogram function model \( \gamma(h) = C_0 + C_1(1 - \exp\left(-\frac{h}{\alpha}\right)^a) \).

Figure B.4: Linear variogram function model \( \gamma(h) = C_0 + C_1(h^a) \).

Figure B.5: Power variogram function model \( \gamma(h) = C_1h^{2H} = Ch^b \) for \( b=0.6 \).

Figure B.6: Power variogram function model \( \gamma(h) = C_1h^{2H} = Ch^b \) for \( b=1.6 \).
Bibliography


List of Publications


