UNIVERSITY OF SOUTHAMPTON

On the extraction and representation of land cover information derived from remotely sensed imagery

by

John Manslow

A thesis submitted for the degree of
Doctor of Philosophy

in the

Faculty of Engineering and Applied Science
Department of Electronics and Computer Science

23rd April 2001

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

ON THE EXTRACTION AND REPRESENTATION OF LAND COVER INFORMATION DERIVED FROM REMOTELY SENSED IMAGERY

by John Manslow

This thesis considers the problem of area estimation from remotely sensed images and, in particular, the problem of estimating the proportions of subpixel area occupied by a predefined number of target classes based on a pixel's spectrum alone. Estimating cover proportions within pixels rather than producing crisp classifications of pixels is often seen as a way of increasing the accuracy of land cover maps derived from remotely sensed data. Although such improvements have been observed, many practical applications demand even greater accuracy since, over a large area, an error as low as 10 percent in the estimated proportions may represent the misclassification of many thousands of square kilometres. Unfortunately, much uncertainty remains as to how techniques for subpixel area proportion estimation should be applied and, more importantly, how much information pixel spectra can provide about subpixel land cover proportions.

The main contributions of this thesis consist of a novel probabilistic interpretation of subpixel area proportions that has a number of important implications: It is used to motivate a new probabilistic notation for area proportion information that, due to the probabilistic interpretation, is simple and intuitive, to show that certain types of fuzzy classifier have an equivalent interpretation as crisp classifiers, a relation that can be used to prove that they are capable of producing optimal proportion estimates and which suggests a number of enhancements that are shown empirically to improve the fuzzy classifiers performance. Finally, the probabilistic interpretation is used to provide insights into the application of the cross entropy error function in fuzzy classification that are shown to be supported by empirical evidence.

The thesis also presents a novel analysis of the impact of the sensor point spread function on fuzzy classifier performance that shows that the problem of extracting subpixel proportion information from pixels' spectral signatures is ill-posed. This is used to motivate the use of a new representation for subpixel proportion information – the spectrum conditional proportion distribution – that overcomes many of the limitations of the standard representation. Specifically, the distribution can fully represent the proportion information in a pixel's spectral signature, it permits this information to be propagated without loss, and it allows different sources of proportion information to be optimally combined. A number of techniques for extracting proportion distributions are described and empirical results are presented that underline the utility of the new representation.

Contents

1. Acknowledgements

I would like to thank Neusciences for supporting this research through an EPSRC CASE award, Dr. Tony Dodd for printing my minithesis, Dr. Jasmin Wason for binding this thesis, and Dr. Mark Nixon for providing encouragement and supervision during the latter half of this PhD.

2. List of Symbols

m	basis function means in spectral or area proportion space
μ	subpixel proportion/fuzzy classification
C_n	label of the n^{th} class
$\mu(C_n)$	area of the n^{th} class
$p(C_n)$	prior probability of observing class C_n
P	a pixel
$\mu(P)$	the area of pixel P
$\mu(C_n\|P)$	the proportion of the subpixel area of P covered by class C_n
$\mu(C_n,P)$	the area of the intersection of class C_n and pixel P
s	pixel spectral signature
$p(s)$	prior probability of observing spectral signature s
$p(C_n\|s)$	posterior probability that a pixel with spectral signature s is in class C_n
$p(s\|C_n)$	class conditional probability that a pixel in class C_n has spectral signature s
J	number of basis functions
j	basis function index
$p(j)$	prior probability that the j^{th} basis function generates a spectral signature
$p(s\|j)$	probability that the j^{th} basis function generates a spectral signture s
$p(j\|s)$	posterior probability that spectral signature s was generated by the j^{th} basis function
D	number of data points
d	data point index
N	number of classes
n	class index
$p(\mu\|s)$	spectrum conditional area proportion distribution – the probability that a pixel has subpixel proportions μ given that it has spectral signature s
$p(C\|\mu)$	probability that a pixel is in class C given that it has subpixel proportions μ
$f(s)$	fuzzy basis function activation as a function of a pixel's spectral signature
$p(C\|x,y)$	posterior probabiltiy that the subpixel point (x,y) belongs to class C
$C(x,y)$	equal to one if subpixel point (x,y) is in C, zero otherwise
$\Psi(.)$	the sensor point spread function (PSF)

r	distance of a subpixel point from the point of maximum PSF sensitivity

3. Overview

This thesis presents a detailed examination of the problem of estimating the proportion of the subpixel area of remotely sensed (RS) image pixels occupied by different land cover types – a process often referred to as the fuzzy classification of pixels. It is shown that there is a close relationship between conventional crisp classification and fuzzy classification and uses this relationship to derive several new and important results. In addition, a novel analysis of the effect of the sensor point spread function is given that provides insight into its effect on fuzzy classification accuracy. This analysis is used to motivate a new representation for information derived from remotely sensed images based on conditional probability distributions and several techniques are presented that are capable of deriving such representations.

The structure of this thesis traces the evolution of the fuzzy classification of RS image pixels from more conventional crisp classification – a process outlined in chapter 6. Chapter 5 describes a new way of viewing area proportions and hence fuzzy classifications as conditional probabilities and uses this interpretation as the basis of a probabilistic notation for area proportion information that is used to list the axioms governing its behaviour. This information is placed at the beginning of the thesis due to the elementary nature of the material it contains. Chapter 6 describes the standard approaches to the classification of RS image pixels that are germane to the main subject of the thesis. That is, classification techniques from which fuzzy classifiers have been derived, or provide useful insight into the problem of fuzzy classification.

Chapter 7 deals specifically with fuzzy classification algorithms and carefully examines their relationship to more conventional crisp classifiers. In particular, a new equivalence between fuzzy classifiers and crisp classifiers is established through the probabilistic interpretation and a novel analysis of the use of the cross entropy error function in fuzzy classifiers is presented. Chapter 8 discusses a number of factors that limit fuzzy classifier performance, and introduces terminology that makes it possible to list the conditions necessary to get perfect fuzzy classifications. Section 8.4 presents a new analytical description of the effect of the sensor point spread function on fuzzy classifier performance, which suggests simple ways of improving fuzzy classifier performance, but also motivates the use of a new representation for information derived from RS spectral data – the spectrum conditional probability distribution.

Chapter 9 begins by stating the three main advantages of the new representation, namely its ability to completely express all proportion information contained in a pixel's spectral signature, to facilitate the optimal combination of information from different sources, and the propagation of that information without loss. Subsequent subsections derive algorithms of increasing complexity for extracting spectrum conditional distribution models from sets of exemplars and present results obtained on a real world data set. This thesis thus traces the development of fuzzy classification for area proportion estimation from its origins in crisp classification to the current state of the art, and by examining the limitations of these algorithms, arrives at a new representation for proportion information that overcomes many of those limitations.

3.1. Contributions

The main focus of this thesis is to examine advanced non-linear techniques for performing fuzzy classification of pixels in remotely sensed images. By examining the current state-of-the-art and its limitations, this thesis concludes that a more flexible representation is required for fuzzy classifications, and that this can be provided by existing neural network algorithms. The following list contains the wholly novel contributions of this thesis.

Probabilistic interpretation:

It is argued that area proportions can be regarded as conditional probabilities. This clarifies the relationship between crisp and fuzzy classification since it is shown that fuzzy classification is equivalent to a crisp classification of subpixel points.

Probabilistic notation:

A form of notation for representing area proportions is proposed, which makes their conditional probabilistic nature explicit.

Listing of axioms:

The new notation is used to list the axioms governing the behaviour of area proportions by direct analogy with those of probability theory. In particular, traditionally probabilistic constructs, such as Bayes' theorem, are shown to be directly applicable to area proportions.

Equivalence of crisp and fuzzy classification:

The probabilistic interpretation is used to show that a particular type of fuzzy classifier is equivalent to an EM density estimator based crisp classifier. This highlights some restrictions of the fuzzy classifier and is suggestive of improvements that are shown empirically to produce dramatic improvements in performance. The EM density estimator based fuzzy classifier is shown to be capable of producing optimal fuzzy classifications under ideal circumstances.

Examination of the relationship between soft and fuzzy classifications:

A new perspective on the relationship between softened and fuzzy classifications is presented that suggests that softened classifications should not be used in place of fuzzy classifications to characterise subpixel cover unless fuzzy classifiers cannot be constructed due to a lack of a set of suitable exemplars.

Interpretation of the use of the cross entropy function in training fuzzy classifiers:

The probabilistic interpretation is used to show that the cross entropy function is appropriate for training fuzzy classifiers, and has a specific interpretation in terms of classifying subpixel points.

Examination of the factors limiting fuzzy classifier performance:

A discussion of the factors limiting fuzzy classification performance is presented, which focuses on limits imposed by the characteristics of the sensor and the target cover types rather than on the difficulties than can arise during the modelling process. This includes a set of axioms that describe the conditions classes must satisfy to permit perfect fuzzy classification.

Detailed examination of the ambiguity induced by the sensor PSF:

A Gaussian model of the sensor PSF is used to show that the PSF introduces ambiguity into the fuzzy classification process. In particular, the ambiguity is shown to be greatest when pixels are heavily mixed.

Introduction of the spectrum conditional density representation of proportion information:

A new way of representing fuzzy classifications is proposed that provides a complete representation of the ambiguity present in fuzzy classifications. The benefits of the new technique are demonstrated on a real world remotely sensed data set.

3.2. FLIERS – Fuzzy Land Information from Environmental Remote Sensing

The data set used to demonstrate the techniques described in this thesis was generated as part of the EU funded research project FLIERS. The aim of the research was to advance the state of the art in fuzzy classification through the use of sophisticated non-linear statistical modelling techniques such as neural networks. In order to use such techniques, it is generally necessary to have a large set of exemplar pixels of known fuzzy membership and for the purposes of the FLIERS project, several such data sets were prepared by a team at the University of Leicester. The particular data set used throughout this thesis covers a region of large scale agriculture to the east of Leicester called the Stoughton area and is shown in figure 1. In all, 21,081 pixels from a Landsat TM survey were available, and fuzzy memberships were derived using a combination of aerial photography and ground surveys. Unfortunately, due to the time required for the ground survey, the fuzzy memberships represent land cover at a slightly different date to that of the satellite imagery, allowing for the possibility of minor changes in land cover in the interim.

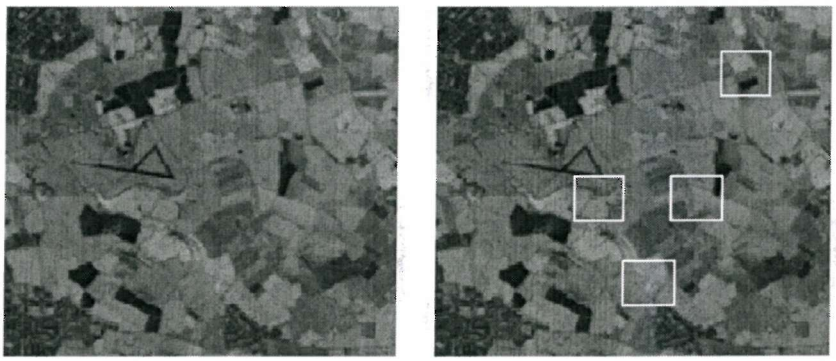

Figure 1: The Stoughton area in band 4 (left) showing the validation areas (right).

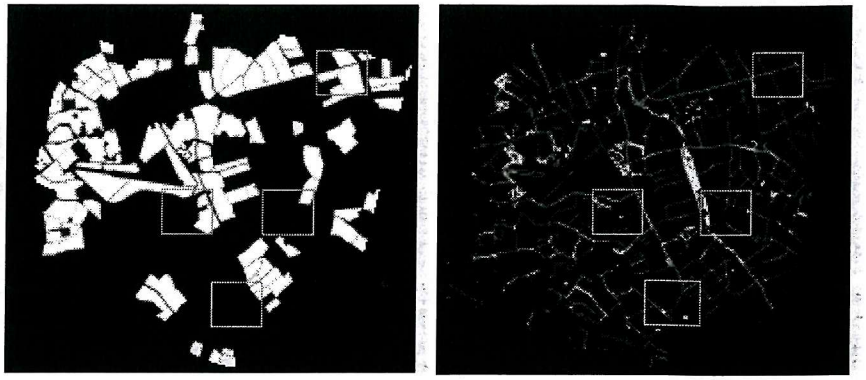

Figure 2: Cereal (left) and tall herb (right) ground truth with the validation areas (inside the small squares) shown in context of the entire data set.

Although there were as many as 26 classes of interest to the FLIERS project generally, only two were selected to permit the depth of analysis with the wide range of techniques suitable for performing fuzzy classification described in this thesis. The cereal crop and tall herb classes are considered in this thesis, and their statistics are given in table 1, along with their unconditional proportion distributions in figures 4 and 6. The unconditional proportion distributions are a useful way of visualising the distribution of proportions that actually occur in the data, and were generated by applying a standard Gaussian mixture model density estimator to the area proportion data. The data set was divided into three subsets, a training set, a test set and a validation set. The training set was used directly by training algorithms to search for the optimal parameters of the particular model being trained. The test set was used to prevent over-fitting – producing a model that was tailored to specific features in the training set that are not characteristic of the process being modelled. The validation set was not used in any way to find the optimal model and hence could be used to evaluate the performance of the models when applied to new, previously unseen areas.

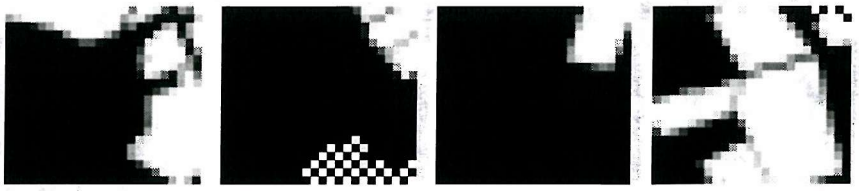

Figure 3: Cereal ground truth for the validation areas.

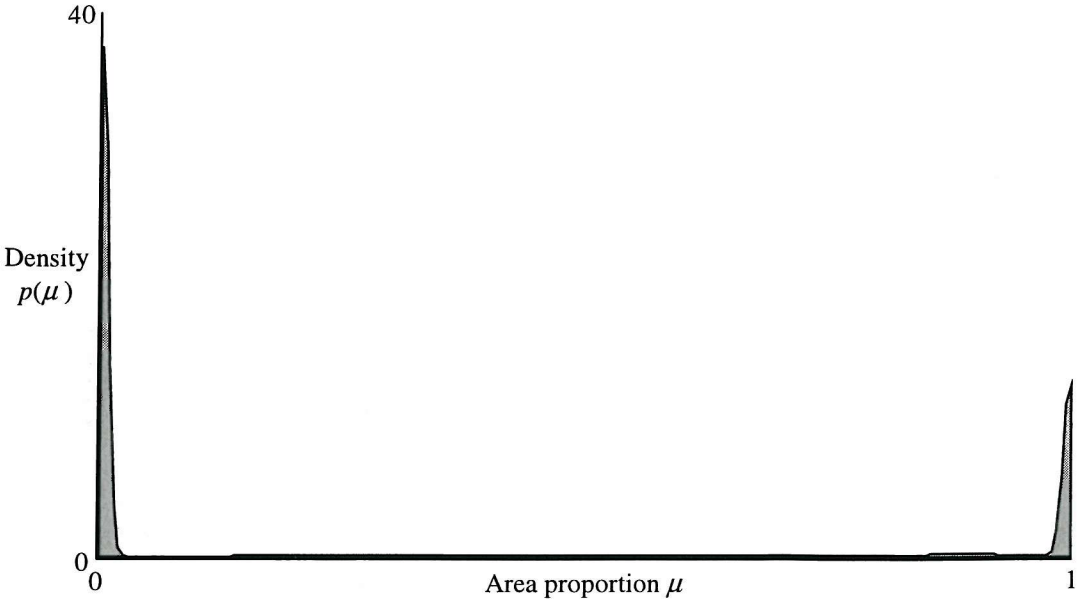

Figure 4: Cereal area proportion distribution for pixels in the training set.

To generate the three subsets of the data, there was a trade-off to be made between generating sets that were statistically representative of each other and sets that could be visualised as small sub-images. For example, the most easily interpretable form for each of the sets is that they consist of large blocks of contiguous pixels. Each set, and hence the proportion estimates made by each technique can then easily be assembled into large images to provide a clear representation of the estimates. However, spatial non-stationarity across the survey area causes marked differences in the statistics of the training, test and validation sets if they are chosen as contiguous blocks since, on average, a pixel in each data set will in the image plane be far from the closest pixel in any other data set. Such non-stationarity can have a devastating effect on the performance of statistical models, since the statistics they learn from the training set may be substantially different from those of the test and validation sets. This is essentially the same process that is described in the case of classification in [Friedl:00].

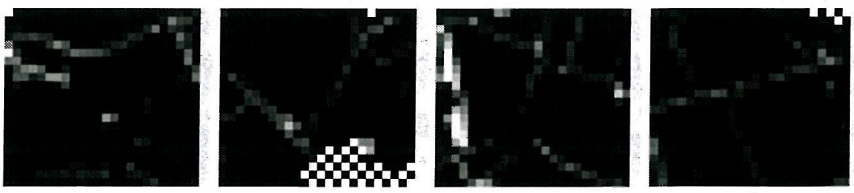

Figure 5: Tall herb ground truth for the validation areas.

The effects of non-stationarity are minimised when pixels are assigned randomly to one of the three data sets such that they form three non-overlapping sets, each distributed roughly uniformly over the survey area. In this case however, visualisation of the

proportion information in each set becomes difficult, because no image can be reconstructed from any of the three data sets without containing lots of points for which there is no data. To strike a balance between these two concerns, the image of the survey area was divided into rectangular blocks of 24×22 pixels. These were sufficiently large that they could be displayed as images to allow the estimated proportions to be visualised, but also small enough that they "covered" the image and hence limited the effects of spatial non-stationarity. The four validation regions are shown in context in figure 2 as those regions within the white squares, and in detail in figure 3.

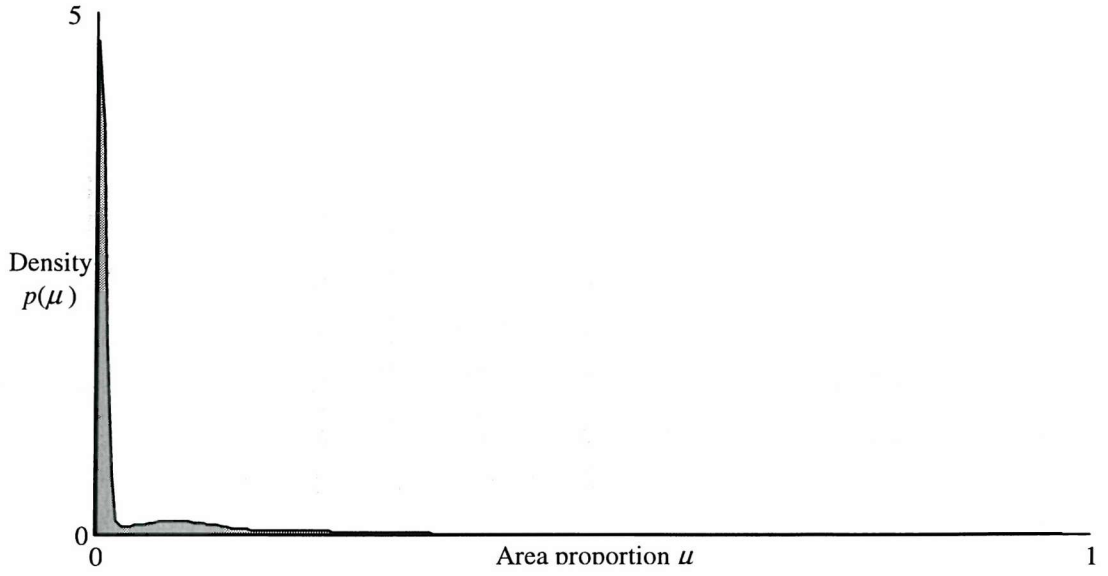

Figure 6: Tall herb area proportion distribution for pixels in the training set.

The first of the two classes used in this thesis, the cereal crop class, was a compound class composed of the main types of cereal grown in the survey area. Slightly less than half of the survey area was covered by cereal crops although most pixels were almost pure. This can be seen from the cereal proportion distribution shown in figure 4, which has pronounced peaks for fuzzy memberships close to 0 and 1 and a general lack of probability mass for most other proportions. This is due to the fact that the typical field size in the survey area is larger than the pixel size, resulting in only a small proportion of pixels straddling a field boundary. The primary statistics of the training set and the validation set are very similar, suggesting that the data subsets are representative of each other. The ground truth information – the proportions of the subpixel areas actually occupied by cereals is shown in figure 2, where white represents a pixel consisting of 100 % cereal and black 0 %. The grey squares highlight the areas from which the validation data was collected, which are also shown enlarged in figure 3. The chequered area indicates a region that was not used due to the absence of ground truth data.

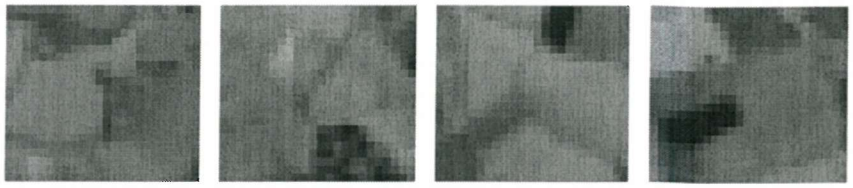

Figure 7: Validation areas in band 4.

The second class, tall herbs, consisted of a large variety of plant types that are commonly found at the sides of roads and along river banks. The survey area contained only small amounts of the tall herb class, the mean subpixel proportion being only around 0.2 % of a pixel. The tall herb area proportion distribution, shown in figure 6, also shows that most pixels contained either no tall herb or only very small quantities. The statistics of the training and validation areas for the tall herb class are quite different, suggesting that they may not be representative of each other and hence that statistical models may have problems with this partition of the survey area. The ground truth for the tall herb class is shown in figure 2 where, once again, the validation areas are highlighted by the grey squares. These areas are shown enlarged and in isolation in figure 5 where it is possible to see greater detail.

Class	Data Set	Number of Patterns	Mean	Variance
Cereals	Training	16169	0.3007	0.1882
	Test	2795	Not computed	Not computed
	Validation	2117	0.4441	0.2160
Tall herb	Training	16169	0.03642	0.01201
	Test	2795	Not computed	Not computed
	Validation	2117	0.05939	0.02437

Table 1: Summary statistics for the FLIERS data set.

4. An Introduction to Fuzzy Classification

When pixels are crisply classified, they are conventionally assigned the label of one of a number of candidate (or target) classes and are thereafter considered to belong to the set of pixels in that class. Pixels, in the applications that are of concern here, are classified according to their subpixel land cover and the class label assigned to the pixel is considered in some sense to represent the subpixel cover. There has long been concern over the inadequacy of a single class label as a representation of the often diverse range of subpixel cover [Woodcock:00] [Cracknell:98][Fisher:97][Foody:97]. Fuzzy classification offers a means of increasing the richness of these representations by assigning pixels partial degrees of membership for each of the candidate classes but, despite its numerous successes, still often receives little attention (see, for example [Cihlar:00] and [Smits:00]).

The exact meaning of the term fuzzy classification is discussed in detail later, and depends on the property of the subpixel cover that the classifications are intended to represent. The use of the term fuzzy classification does not imply any rigorous relation to the field of fuzzy logic as expounded in texts such as [Klir:95] (and, less formally in [Wang:93]), but merely highlights the fact that pixels are assigned partial degrees of membership in more than one class. The main situations in which crisp classifications of remotely sensed image pixels poorly represent true subpixel cover result from:

- pixels straddling the boundary of two or more distinct classes [Fisher:90][Fisher:97], and

- the presence of classes with boundaries that cannot be clearly delimited [Foody:92] [Wood:89].

In the first case, the true subpixel cover consists of a number of discrete classes. The conventional approach of assigning a single class label to such a pixel (which, in some instances may contain very similar proportions of the cover types) seems an inadequate representation of the subpixel process. This problem can become particularly severe when land cover transitions occur close to, or below the pixel size, since this will lead to a high proportion of image pixels containing multiple cover types. The severity of this problem thus depends on the interactions between the spatial frequency of transitions in

the target cover types, the resolution of the sensor and the sensitivity of the target application to the subpixel partition information lost during crisp classification.

Consider, for example, an image of an area that is mainly agricultural, but which also contains small settlements and farm buildings. When a crisp classifier is applied to such an area, the crop types will generally be well represented with a moderate resolution satellite, since the fields will tend to be large compared with the pixel size. The built areas, however, may never contribute significantly to the subpixel area, making it possible that no pixels in the image are classified as "built" even though "built" may constitute a significant area of the land covered by the classified image. The second difficulty with crisp classification arises when a region contains cover types that have a tendency to continuously intergrade. This means that although there may be separate regions of land cover which may satisfactorily be crisply classified as one of the target cover types, between these regions, the cover types may merge continuously, leading to land cover with characteristics resembling several different classes. Once again, it seems inadequate to represent such land cover by assigning to it the single label of any of the individual classes to which it is similar.

An apparently simple cover type such as forest can be used to illustrate this difficulty, which is relatively common when classifying many organic cover types. For example, consider a dense region of trees surrounded by open ground. If the region of trees is sufficiently large, it seems natural that the forest classification is applicable. If, on the other hand, the region of trees is actually rather small, then the forest classification seems inapplicable. Between these two extremes, however, it may be difficult to decide whether the region represents forest or not, without making a rather arbitrary distinction. Even if the region is large enough for a forest classification, tree density may decrease towards the forest edge causing difficulty in assigning a precise boundary to the forest.

In practice, fuzzy classification requires a precise definition of the fuzzy memberships that are to be used since, if such a definition is lacking, membership estimates will be difficult to interpret and it will not be possible to evaluate the relative accuracies of different membership estimates. At first, it may appear as though it is necessary to define two incompatible fuzzy memberships, each to address one of the causes of mixing outlined earlier. An appropriate definition of fuzzy membership for representing the presence of a number of otherwise crisp subpixel classes is to record the proportions of the pixel area covered by each of the target classes. Unfortunately, this definition

may not be used to tackle the problem of intergrading classes, since such classes do not have well defined boundaries and hence do not have well defined areas.

Another definition of fuzzy membership may be considered in the presence of intergrading classes, which represents the similarity of the subpixel cover to each of the target cover types. The precise nature of the definition of such a similarity measure will not be considered further here, although it will be assumed that the memberships are closed world. That is, the sum of the similarity measures over the target classes is unity for all subpixel regions. If this condition is satisfied, the mean value of a similarity measure of a crisp class is equal to the proportion of the subpixel area occupied by the crisp class. The problems of representing subpixel mixing of crisp classes and the similarity of subpixel cover to classes which intergrade are thus both the same as representing the mean subpixel similarity measure for each of the target classes.

Since a definition of a similarity measure that is clearly defined and interpretable at ground level is generally lacking, the problem of estimating similarity measures is ill-defined and is hence unlikely to be accessible to solution. For this reason, only the problem of estimating the proportions of subpixel area occupied by crisp target classes will be considered in this thesis. The following section traces the development of fuzzy classification and other concepts relevant to the content of this thesis in the literature.

4.1. The Evolution of Fuzzy Classification

In [Horwitz:71] a simple algorithm was derived for obtaining estimates of the proportions of cover types within a pixel. Their paper, the first to explicitly address the problem of area proportion estimation, made a number of simplifying assumptions such as a uniform point spread function within a pixel and Gaussian spectrum conditional densities for each class, with the help of which it was possible to show that maximum likelihood area proportion estimates could be obtained as a linear function of a pixel's spectral signature and in so doing introduced the basic concepts of linear spectral mixture modelling used in research that continues until the present day.

The use of terminology from the theory of fuzzy sets – sets that have poorly defined boundaries (see [Zadeh:65]) – in discussions of the mixed pixel problem seems to have appeared shortly after the introduction and subsequent popularisation of the fuzzy clustering algorithm described in [Bezdek:84], and was used to address the mixed pixel problem in [Robinson:85]. Fuzzy clustering is an algorithm that can be used to perform

either a supervised or an unsupervised clustering of pixel spectra into a number of fuzzy sets. By measuring the degrees of membership of new pixels in the available fuzzy sets, information about subpixel composition can be derived. Although the use of fuzzy terminology continues to be popular, the use of fuzzy set theory itself is relatively rare. This is because fuzzy set theory is only necessary if the subpixel area proportion estimation problem is regarded as one of classification. If the idea of imposing classifications on pixels is abandoned, the nature of the subpixel area estimation problem and its solution can easily be described using probability theory.

Contemporaneous with the emergence of fuzzy clustering, a new algorithm was proposed for training multilayer neural networks. Called back-propagation, it was proposed in [Rumelhart:86] (although similar ideas had earlier appeared in [Werbos:74] and [Parker:85]), and provided, for the first time, a reliable and efficient way of training multilayer artificial neural networks. This new found efficiency combined with the weak distributional assumptions made by such networks resulted in their application in conventional pixel classification in several papers that were published in the late eighties and early nineties. Gradually, it was realised that the raw outputs of neural network classifiers – which were shown to be estimates of posterior probabilities of class membership in [Baum:87] – and similarly, the outputs of other non-neural classifiers, were useful for more than choosing a class label to associate with a pixel [Gorte:98][Foody:96b][Foody:96c] [Maselli:96].

Although it was clear that there was a relationship between these outputs and subpixel composition, it was not obvious as to how the classifiers could be modified to provide more information about subpixel composition. In [Foody:95] a modification to the way in which classifiers were trained and tested was proposed that would produce fully fuzzy classifiers – neural network fuzzy classifiers with outputs that, under ideal circumstances (such as an infinite availability of data, and an infinitely flexible neural network), would produce optimal estimates of subpixel cover proportions. This synergy of the richness of the subpixel area proportion representation and the power and flexibility of neural networks has resulted in unprecedented accuracy in the land cover information that is currently derived from remotely sensed data.

Despite these advances many questions still remain about the correct application of neural networks in general and whether there are any characteristics of the subpixel area proportion estimation problem in particular that have implications for their use [Wilkinson:97]. It is these questions that have produced doubts about the current focus of

research on improving fuzzy classifier performance and led to proposals for a shift of focus towards quantifying the performance limits that are intrinsic to deriving fuzzy classifications from remotely sensed data [Wilkinson:96]. This thesis presents detailed analyses of a number of established techniques which clarify the conditions under which the techniques should be applied, and result in a number of recommendations for improving the performance of the proportion estimates they produce. In addition, it is argued that pixel spectral signatures alone contain too little information about subpixel cover to derive accurate proportion estimates and a new representation for spectrally derived proportion information is proposed that is capable of fully representing the uncertainty in the proportion estimates caused by the information-poverty of the spectral signatures. A number of models for deriving the new representation are presented along with results of their application to a real world data set.

5. Properties of Area Proportions

This section presents a novel argument that area proportions can be interpreted as conditional probabilities [Manslow:00] and uses this as motivation for an area proportion notation that is analagous to the standard notation of probability theory. The new area proportion notation is particularly convenient due to its immediate familiarity that results from the fact that the standard axioms governing the behaviour of conditional probabilities also apply to area proportions.

5.1. The Probabilistic Interpretation of Subpixel Area Proportions

In order to estimate the proportion of a pixel's area occupied by a class it must be possible, in principle, to measure the area of the class given perfect information. Consider a single pixel consisting of two cover types: grass and water. When the area is remotely observed, a mixed pixel is generated which has spectral contributions from both of the subpixel classes. If perfect information was available in the form of the true distribution of the two cover types within the pixel area, each point within the pixel could be uniquely classified as belonging to one of the cover types, and hence a subpixel map of true class membership could conceptually be constructed.

If a point is chosen at random from a uniform distribution over the conceptual subpixel cover map, it will fall within a region occupied by one of the subpixel classes. In the limit of an infinite number of such points being chosen, the proportion of points falling within each class region will equal the proportion of the subpixel area the region occupies, and also equal the probability of an individual point falling within each region. This suggests that there is a direct equivalence between these probabilities and the subpixel area proportions. It is important to emphasise that this probabilistic model does not equate the proportion of the subpixel area occupied by a specific class with the posterior probability of class membership of the entire pixel in that class, as would be estimated by most classical classification algorithms. Although estimates of these probabilities have been used to model subpixel area proportions (see, for example [Chittineni:81][Foody:96c][Gorte:98][Maselli:96]) it is shown later that they cannot, in general, be optimal estimates.

5.2. Area Proportions: Notation and Axioms

In order to describe the properties of area proportions, it is convenient to introduce a compact notation [Manslow:00]: if the area of a pixel P is represented by $\mu(P)$ and the area of the intersection of pixel P and class C_n by $\mu(C_n,P)$ then the proportion of P occupied by C_n will be denoted by $\mu(C_n|P)$. Here, the equivalence of area proportions and (conditional) probabilities is made explicit in the choice of notation. The proportion of P occupied by class C_n is found using

$$\mu(C_n \mid P) = \frac{\mu(C_n, P)}{\mu(P)} \qquad 1$$

From equation 1 the area proportion equivalent of Bayes' theorem may be derived. This can be used to convert quantities of the form 'the proportion of class C_n occupied by pixel P' to 'the proportion of pixel P occupied by class C_n' as follows:

$$\mu(C_n \mid P) = \frac{\mu(P \mid C_n)}{\mu(P)} \mu(C_n) \qquad 2$$

Clearly, the total area occupied by any object or class is found by summing the areas of its intersections with other classes. Thus, the total areas of a pixel P (where there are N classes that form a closed world partition) and of a class C_n (where '*all P*' is the set of all pixels) are given by

$$\mu(P) = \sum_{n=1}^{N} \mu(C_n, P) \qquad 3$$

$$\mu(C_n) = \sum_{allP} \mu(C_n, P) \qquad 4$$

when no two classes or pixels intersect. For two classes, C_n and C_m with $m,n \in [1,N]$ the area of their union may be computed from the sum of their individual areas minus the area of their intersection. More concisely,

$$\mu(C_n \cup C_m \mid P) = \mu(C_n|P) + \mu(C_m|P) - \mu(C_n, C_m|P) \qquad 5$$

A set of classes $C_n : 1 \leq n \leq N$ is considered to be closed world upon the target domain D if

$$\mu\left(\bigcup_{n=1}^{N} C_n \mid P\right) = 1 \quad \forall P \in D \tag{6}$$

Such a set of classes may trivially be constructed by the addition of a class that contains any subpixel region that is not assigned to any other class. Finally, area proportions lie in the closed interval [0,1] as stated in equation 7.

$$\mu(C_n \mid P) \in [0,1] \quad \forall n : 1 \leq n \leq N \tag{7}$$

All of these axioms are directly equivalent to those for manipulating probabilities (as can be found in [Cox:46][DeGroot:89]).

The following section describes the way in which land cover information can be derived from remotely sensed images by the crisp classification of pixels within such images. Although it is now widely recognised that more accurate land cover information can be obtained by other means, an examination of crisp classification is provided for the following reasons:

- three important approaches to fuzzy classification (namely parametric fuzzy classification, e.g. [Wang:90][Foody:96c], softened classification, e.g. [Foody:96], and neural network fuzzy classification, e.g. [Foody:95]) have their origins in more conventional crisp classification techniques, and can be seen as extensions and modifications of those algorithms, and

- the probabilistic interpretation indicates that there exists a close relationship between fuzzy classification and crisp classification and that many of the concepts important in understanding crisp classification are germane to the problem of fuzzy classification.

Chapter 3 described the real world data set used to illustrate the techniques and concepts developed in this thesis and presented an introduction to the ideas behind the use of fuzzy classification in extracting information about land cover from remotely sensed images. Chapter 4 introduced the basic concept of fuzzy classification and reviewed the development of techniques for extracting fuzzy proportion information from remotely sensed data by highlighting a number of seminal publications and has indicated how the work in this thesis follows from suggestions that continued experimentation with existing techniques is likely to prove fruitless unless there is a more detailed examination of the factors limiting their performance. Chapter 5 showed that fuzzy classification can be thought of as crisp classification of non-location specific subpixel points and hence that area proportions can be considered to be a specific type of posterior probability – an equivalence that was used to motivate a probabilistic notation for area proportion information with which the axioms governing its behaviour were listed.

6. Crisp Classification

As described in the introduction, crisp classification – in this case considered to be associating a class label with a pixel – was one of the earliest approaches to deriving land cover information to make use of flexible modelling techniques such as neural networks. Although the class label representation is now widely acknowledged as being an inadequate description of subpixel cover, it was some time before the more versatile fully fuzzy classification techniques that are currently used emerged. In the interim, the posterior probabilities estimated by many standard non-fuzzy classifiers were used to provide information on subpixel cover. In this sense, a transition from crisp classification to soft classification to fuzzy classification can be traced in the efforts to derive information about subpixel cover. Thus, the motivation for describing crisp classification here is that it provides a rudimentary mechanism for extracting land cover information, and one from which the current state of the art can be considered to have evolved. This chapter describes crisp classification and its immediate derivative, soft classification, and presents results of their application to the FLIERS data set. Since these techniques can no longer be considered state of the art, the results are discussed only briefly and are intended to act as a benchmark against which more recent and theoretically well founded techniques are compared in later chapters.

The problem of crisply classifying a pixel is normally considered to be one of assigning to it one or more class labels. In most practical applications, the spectral signature of a pixel will contain too little information to assign the correct class label to all pixels. The class label representation of a classification decision is too poor to represent this ambiguity, and is hence usually avoided other than as an aid to interpretation. Instead, classification decisions are usually represented by estimates of the posterior probabilities that the observed pixel lies in each of the target classes. Thus, if there are N classes of interest, the classifier output would be a vector of probabilities of length N, which, if the classes are mutually exclusive and closed world, would sum to unity. The posterior probabilities contain all the information relevant to classification, since if the probability that a pixel with spectral signature s was in class n was $p(C_n|s)$ the class label C_n would, on average be correct $100 \times p(C_n|s)$ percent of the time. The class label that minimises the misclassification rate is thus the one that maximises the posterior probability. The results of the classification experiments described in this section are presented in terms of posterior probabilities rather than class labels. This approach to

characterising land cover should not be confused with the softened classifications that will be considered later in this section.

Although the results of both crisp and softened classifications can be interpreted as posterior probabilities, there are important conceptual differences: crisp classification assumes that a particular pixel can be correctly and completely characterised by a class label, and the posterior probabilities represent the uncertainty in the classification decision due to the lack of information about class membership in a pixel's spectral signature. If all such information were available, crisp classifiers would always be able to assign the correct class label and all posterior probabilities would be either zero or one. Softened classifiers use the posterior probabilities produced by more conventional classifiers to provide information about subpixel structure even though there is no implicit assumption that it would be meaningful to assign any of the class labels to a pixel even if perfect information was available. Consider for example, the cereal class: a softened classifier would be trained on a set of pixels consisting either purely of cereal or containing no cereal at all. The outputs of such a classifier can be interpreted as estimates of the probability that a pixel is composed entirely of cereal, or conversely that it contains no cereal at all. Such a classifier will then be applied to pixels that are known to belong to neither of the classes with which its was trained. The resulting posterior probability estimates – called softened classifications – have been shown to contain information about subpixel area proportions [Foody:96].

There are essentially two ways of performing crisp classification, each of which has its own fuzzy equivalent. The first method constructs models of the way in which a set of exemplar pixels of known class membership are distributed in spectral space and uses these models to derive estimates of the probability that a new pixel of known spectral signature lies in each of the target classes. Since these probabilities are derived indirectly from a set of models that are unseen to the user, this technique is referred to as the indirect method of classification. The second method of crisp classification uses the exemplars to search for a function that can directly map the spectral signature of a pixel onto a vector of posterior probabilities of class membership. This method is referred to as the direct method of crisp classification since the function derives class membership information directly from a pixel's spectral signature. Although this section presents detailed discussion of both the direct and indirect means of crisp classification, results are presented for the direct approach only since it generally offers superior performance. The indirect method is described because of its close relationship with parametric fuzzy classification which will be described in section 7.1.1.

6.1. Direct Crisp Classification

The results reported in this section were obtained by training models on a "hardened" version of the fuzzy data set. That is, the training targets in the fuzzy data set were converted to class labels by classifying a pixel as cereal if at least 50 percent of the subpixel area was cereal, and otherwise classifying it as non-cereal. This process created a new data set with binary targets; a target of one indicating that a pixel should be classified as cereal and a zero indicating it should be classified as non-cereal. The unseen validation areas are shown for this hardened data set in figure 8.

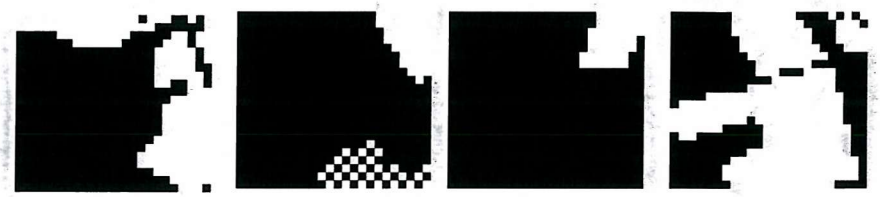

Figure 8: Hardened cereal crop data in the validation areas.

The spectral data was normalised to zero mean and unit variance before being used to train, test or query all the models applied in this thesis. The scaling information was calculated from the training pixels and was as follows:

Band	Mean	Variance
1	63.25	17.78
2	27.73	20.80
3	24.89	55.58
4	105.96	578.61
5	62.34	233.12
7	22.21	111.79

Table 2 : Summary statistics for the six spectral bands of the FLIERS data set.

Thus, for the n^{th} spectral measurement s, in a data set of N patterns in total, the new scaled value s^n_{new} is computed from the old unscaled value s^n_{old} using the mean and variance of the spectral values for that band, s_{mean} and $s_{variance}$:

$$s^n_{new} = \frac{s^n_{old} - s_{mean}}{\sqrt{s_{variance}}} \qquad 8$$

where:

$$s_{\text{mean}} = \frac{1}{N}\sum_{n=1}^{N} s_{\text{old}}^{n} \quad , \tag{9}$$

and

$$s_{\text{variance}} = \frac{1}{N(N-1)}\sum_{n=1}^{N} (s_{\text{old}}^{n} - s_{\text{mean}})^2 \quad . \tag{10}$$

Scaling inputs in this way has a number of benefits, such as improving the conditioning of the optimisation (learning) problem, usually resulting in more stable and efficient training [Haykin:94].

Three types of models are considered in this section, the linear network, the logistic network and the multilayer perceptron, typical examples of which are shown in figures 10 and 44 (the former showing both the linear and logistic networks since they have essentially the same structure). The linear network consists of six input nodes, one output node and a bias node. The bias node is held constant at a value of one and is used to provide an additive component to the model output that is independent of any of the inputs. The value of the weight connecting the bias to the output node is equal to the mean of the training targets, and the output of the linear node is simply a weighted sum of the spectral inputs and the bias node. For a pixel of spectral signature s with spectral components s_1 to s_6, the output of the linear network μ_{est} can be expressed as:

$$\mu_{est} = w_b + \sum_{m=1}^{6} w_m s_m \tag{11}$$

where w_b is the bias weight, and w_m is the weight from the m^{th} spectral input to the output. The parameters w may be found by matrix inversion, provided that the number of training patterns is not too large, or by using iterative optimisation algorithms, such as conjugate gradients [Shewchuk:94][Bishop:95][Gill:93][Axelsson:96] (ordered most to least accessible), which was used to produce the results described here.

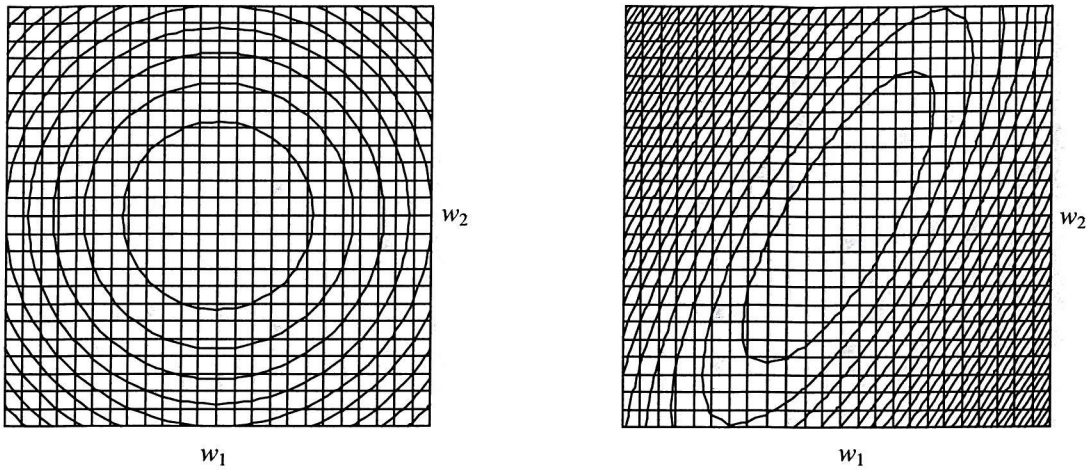

Figure 9: First order gradient descent performs well if an error function's contours are circular (left) but poorly if they are elliptical (right).

Conjugate gradients is an iterative algorithm originally designed for finding the solution to large systems of linear simultaneous equations. While normal gradient descent performs well when minimising functions that have roughly circular contours around their minimum, as shown in the left hand side of figure 9, it performs poorly when the contours are elliptical, like those in the right hand side of figure 9, as often occurs with linear networks when the inputs are correlated and with non-linear networks generally. Under such circumstances, steepest gradient descent is not guaranteed to find the solution in a finite number of steps and may in fact approach it only very slowly. Conceptually, conjugate gradients eliminates the elliptical contours by stretching the space in which the optimisation is to occur into one in which the contours around the minimum are circular. Locations in this new stretched parameter space are expressed in terms of a set of virtual parameters that can easily be mapped to, or mapped from, the original parameters using linear operators. The conjugate gradient algorithm performs steepest descent in the new space, and transforms the changes made to the virtual parameters back into the original parameter space, to derive the optimal changes that should be made to the actual weights in the network.

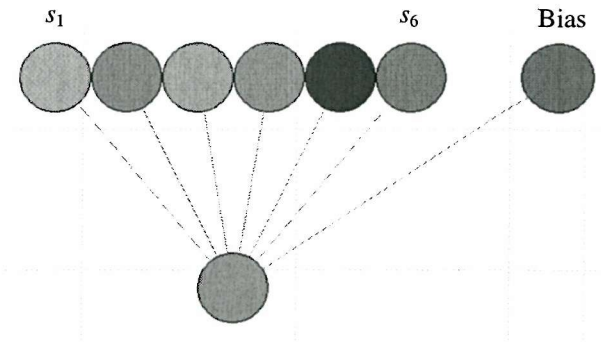

Figure 10: A linear or logistic network.

In practice, the transformation is implicit in the operation of the algorithm and is never done explicitly. As steepest descent is guaranteed to find the optimal weights for a linear network with M weights in M iterations (provided that the inputs are uncorrelated and have equal variances), conjugate gradients is guaranteed to find the optimal weights in M iterations under much more general conditions. In practice, conjugate gradients relies on a line search to find the minimum of the error function at each of the M steps of the algorithm, and hence the amount of computation required to find a solution is more than is required for M gradient computations. Conjugate gradients was chosen for the results reported here because it was supported by the software used to produce all other results in this thesis and, unlike steepest descent, it is likely to find an exact solution in a finite amount of time.

```
New_Error = Get_Training_Set_Error()
If ( New_Error > Old_Error )
{
      MLP.Weights = MLP.OldWeights
      MLP.LearningRate = 0.5*MLP.LearningRate
}
else
{
      MLP.OldWeights = MLP.Weights
      MLP.LearningRate =
      Old_Error = New_Error
}
MLP.Do_Descent_Step()
```

Figure 11 : Pseudocode for the accelerated backpropagation algorithm.

The logistic network is essentially identical to the linear network except that the output passes through the logistic function.

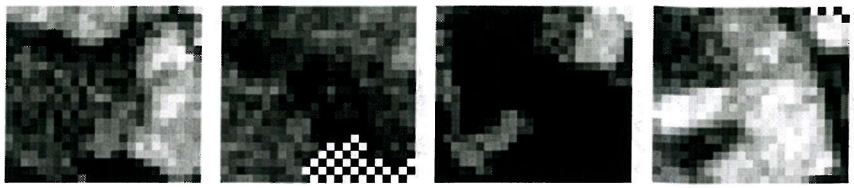

Figure 12: Cereal proportions estimated by a linear network trained using the sum of squares error on hardened data.

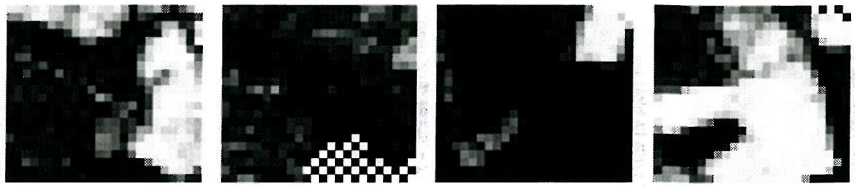

Figure 13: Cereal proportions estimated by a logistic network trained using the sum of squares error on hardened data.

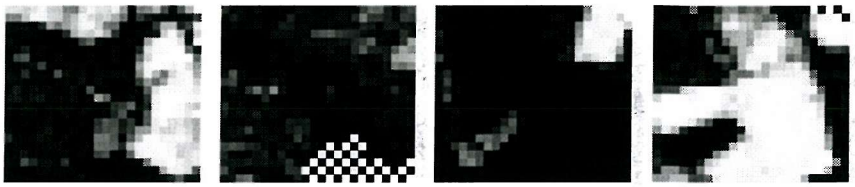

Figure 14: Cereal proportions estimated by a softmax discriminant trained using the cross entropy error on hardened data.

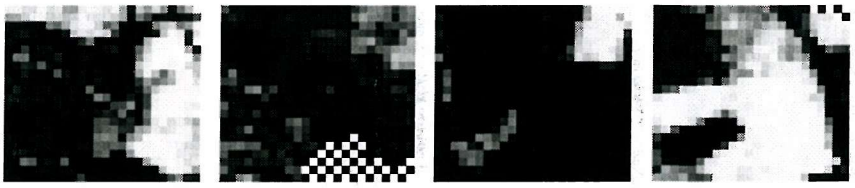

Figure 15: Cereal proportions estimated by a 6-5-1 MLP trained using the sum of squares error on hardened data.

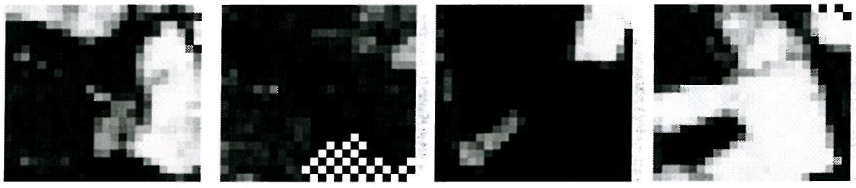

Figure 16: Cereal proportions estimated by a 6-5-2 MLP trained using the cross entropy error on hardened data.

Note that in this case the weights lose the simple interpretability that they had in the case of the linear network, and much of the advantage of using conjugate gradients to find the weights is also lost. Despite these disadvantages, the simple addition of the logistic function to the output of the linear network produces a significant improvement in the area proportion estimation performance. In particular, the squashing properties of the function make it possible for the network to predict much more homogeneous regions than the linear network, resulting in a mean squared fuzzy classification errors of 0.0556 compared with 0.0824 over the unseen areas. The fuzzy classification performance of the crisp classifiers on the validation set was used as a performance

measure rather than crisp classification performance since the main concern of this thesis is the adequacy of crisp classifiers for extracting land cover information.

The effect of the squashing function in the logistic network can clearly be seen by comparing figures 12 and 13 and the validation set performances in table 3. Figure 14 shows the proportions estimated by a single layer network trained by minimising the cross entropy error function rather than sum of squares. The cross entropy error for a proportion estimate μ_{est} and a true proportion μ is given by

$$E = \mu \ln \mu_{est} \qquad 12$$

and has much stronger theoretical justification than the sum of squares error when performing classification [Bishop:95]. The cross entropy error has also been investigated in the context of fuzzy classification in [Foody:95b], which provided good empirical evidence that the use of the cross entropy error function may be advantageous in certain applications. A new theoretical examination of the justification for the use of the cross entropy function in fuzzy classification is presented in section 7.2.1.

To train the networks using the cross entropy error function it is necessary to use two outputs that were constrained using the softmax function such that they summed to unity. This was necessary because without this constraint the cross entropy function can trivially be minimised by estimating large proportions of crops in all pixels regardless of their true composition. The training, testing, and validation sets were processed so that they contained two outputs also, one representing "proportion of crops" and the other "1-proportion of crops" to produce the required normalisation. From the figures, the behaviour of the single layer softmax network that was trained using the cross entropy function can scarcely be distinguished from that of the sum of squares trained single layer logistic network.

Next, two neural network classifiers were produced. The first was an MLP with five logistic hidden neurons and one output used to indicate membership in the crop class and was trained using the adaptive step size backpropagation gradient descent algorithm outlined in figure 11 to minimise the sum of squares error function. This algorithm, which dynamically adjusts the learning rate parameter according to changes in the training error was used to guarantee stability in learning during hours of unmonitored training. The second network that was trained was also an MLP with five logistic hidden neurons, but had two softmax output neurons indicating membership in the crop class

and was trained by minimising the cross entropy function. The results of applying these two networks to the validation data is shown in figures 15 and 16. Here it can be seen that the choice of error function has a subtle but definite effect on the behaviour of the trained networks. For example, compared to the sum of squares network, the cross entropy network appears to have improved performance in modelling the top region of the upper field in the fourth validation area at the cost of performing slightly worse in the lower left of the third region. In general, such differences are difficult to explain since they result from a complex interaction of the distribution of the data, the network parameterisation and the error function. However, it can be shown that if a sum of squares network predicts μ_{est} when the true proportion is μ, the derivative of the error function is:

$$\frac{\partial E}{\partial \mu_{est}} = \mu_{est} - \mu \qquad 13$$

which is dependent only on the size of the difference between the predicted and true proportion. For the cross entropy error however,

$$\frac{\partial E}{\partial \mu_{est}} = \frac{\mu_{est} - \mu}{\mu_{est}(1-\mu_{est})} \qquad 14$$

which means that gradient based learning algorithms that use the cross entropy function will be most sensitive to errors when the predicted proportion is close to one or zero. This could help to explain some of the differences observed in the results: if at some point during training the network tended to produce high estimates for the proportions, μ_4, in the upper field of the fourth area and low estimates, μ_3, for the area in the lower left of region three and if $1-\mu_4<\mu_3$ the cross entropy network would sacrifice accuracy in the lower left of the third region to improve performance on the field in the upper part of the fourth region, thus producing the distribution of errors that is actually observed. Finally, its interesting to note that although the increase in flexibility in moving from the logistic discriminant to the sum of squares MLP improved performance, the same increase in flexibility in moving from the softmax discriminant to the cross entropy MLP resulted in poorer performance. This difference is difficult to explain due to the interaction between the different model parameterisations, error functions and the distributions of the crisp training and test data and the distribution of the fuzzy validation data.

Algorithm	Error Function	Number of Basis Functions	Validation Set Error
Linear discriminant	Sum of squares	Not applicable	0.08237
Logistic discriminant	Sum of squares	Not applicable	0.05556
Softmax discriminant	Cross entropy	Not applicable	0.1473
MLP	Sum of squares	5	0.05201
"	Cross entropy	5	0.1487

Table 3: Model performances on the hardened data.

It is clear from these experiments that the outputs of crisp classifiers do contain information about subpixel proportions. The value of this information depends on the target application and its sensitivity to the errors in the proportion estimates that can be obtained from crisp classifiers. The following two sections describe the indirect method of obtaining crisp classifications, and the use of the softened classifier – a technique similar to that already described but designed explicitly for using a crisp classifier to derive proportion estimates.

6.2. Indirect Crisp Classification

The alternative to the direct approach to classification that was just described is to use pixels of known class membership to construct models of the class conditional probability densities for each of the target classes in spectral space. To classify a pixel of spectral signature s, one density estimator would be used for each class and would produce an estimate of the likelihood that if the pixel were of that class, it would have generated the spectral signature that was actually observed. This likelihood, written as $p(s|C_n)$ for the n^{th} class, can then be used in Bayes' theorem:

$$p(C_n \mid s) = \frac{p(s \mid C_n)p(C_n)}{p(s)}, \qquad 15$$

to obtain an estimate of the posterior probability $p(C_n|s)$ that the pixel belongs to the n^{th} class given that it has the observed spectral signature. In principle, this estimate is all that is required to perform optimal classification, since the class label that maximises the posterior minimises the misclassification rate.

The prior $p(C_n)$ in equation 15 is the probability of a pixel belonging to class n regardless of its spectral signature, and is usually estimated from the set of exemplars by

the proportion of exemplars in class n, while the unconditional density $p(s)$ is usually ignored in practice, since it is independent of the ordering of the posterior probabilities, and hence has no effect on the optimal classification decision. Classifiers based on this indirect method often perform poorly in practice because highly parametric models, such as single Gaussians, are used to estimate the class conditional densities. It should be stressed, however, that this is not an intrinsic limitation of the indirect approach to classification since more flexible models of the class conditional densities, such as a superposition of Gaussians, can be used, and will, in many cases, lead to improved performance. The following section looks at how class conditional probability densities can be accurately and efficiently modelled by a superposition of Gaussians.

6.2.1. Modelling Class Conditional Densities

To perform indirect classification, the class conditional densities of each class can be modelled independently using a separate density estimator for each class. Representing the class conditional density for the n^{th} class as a superposition of J Gaussian basis functions,

$$p(s \mid C_n) = \sum_{j=1}^{J} p(s \mid j) p(j) \qquad 16$$

where $p(j)$ are basis function priors (the probabilities that the j^{th} basis function generates an unspecified spectral signature) and are parameters of the mixture model determined during training, and the $p(s|j)$ are the probabilities that the observed spectral signature could be generated by the j^{th} basis function, which for Gaussian basis functions, are

$$p(s \mid j) = \frac{1}{(2\pi\sigma_j^2)^{1/2}} \exp\left(-\frac{(s - m_j)^2}{2\sigma_j^2} \right), \qquad 17$$

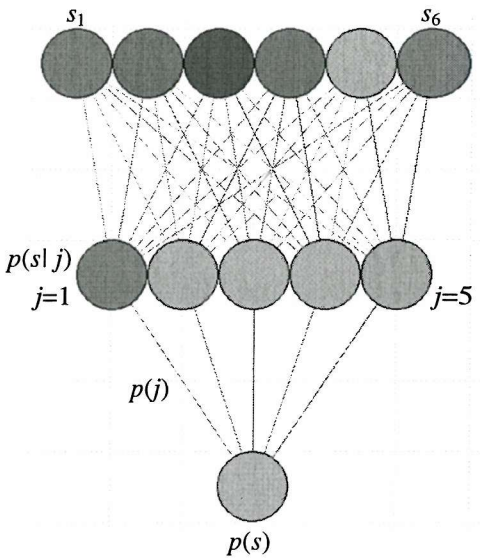

Figure 17: A mixture model density estimator with five mixture components.

where σ_j^2 is the variance (width) of the basis function, and m_j is its mean (centre). Such a density estimator is referred to as a Gaussian mixture model, since the density is modelled as a mixture of independent Gaussian components. Density estimators of this form can be visualised as a network structure (as shown in figure 17 for a model with five mixture components) where the inputs are the pixel spectral signatures, each hidden node corresponds to one of the components in the mixture model, and the hidden node activations are equal to the $p(s|j)$ terms. The priors for each of the mixture components are given by the hidden to output layer weights and the network output is the probability density at the specified point in spectral space. With a suitable choice of error function, the density estimator can be "trained" using error backpropagation gradient descent in the same way as any other neural network, provided that the priors are constrained to sum to unity.

The basis function priors, means and variances together form the complete set of parameters for the mixture model, and must in some way be inferred from a set of exemplars. This is done by dividing the available data set of D patterns into N separate data sets, the n^{th} of which contains only pixels belonging to the n^{th} class. N density estimators are then constructed by finding the mixture model parameters that minimise the negative log-likelihood of the data set given the density estimates

$$E = -\sum_{d=1}^{D} \ln p(s_d \mid C_n), \qquad 18$$

where $p(s_d|C_n)$ is the output of the n^{th} density estimator on the d^{th} exemplar. There are a variety of standard ways of minimising equation 18, one of the most efficient of which is the expectation maximisation (EM) algorithm. The EM algorithm for finding the parameters of a Gaussian mixture model is described in [Bishop:95], and so only an outline of its conceptual basis will be given here. Iterative algorithms for finding the parameters of any non-linear model proceed by using the current set of model parameters, in this case, $p_{old}(j)$, m_j^{old} and σ_j^{old}, along with some performance metric, to derive a new set of parameters, $p_{new}(j)$, m_j^{new} and σ_j^{new}. The main aim of EM is to find the new parameters such that the expected increase in the performance metric achieved by changing from the old to the new parameters is maximised. For the Gaussian mixture model, it can be shown (see [Bishop:95]) that the decrease in the cross entropy metric when changing from an old set of parameters to a new set is always less than:

$$-\sum_{d=1}^{D}\sum_{j=1}^{J} p_{old}(j \mid s_d) \ln\left|p_{new}(s_d \mid j) p_{new}(j)\right|, \qquad 19$$

where $p_{old}(j| s_d)$ is the probability that the spectral signature of the d^{th} pixel in the set of exemplars was generated by the j^{th} component in the mixture given the old parameters. Minimising the above bound with respect to the model parameters makes it possible to derive equations for the new parameters in terms of the old parameters in such a way that the minimum expected decrease in the error function is maximised. Thus, for the basis function means (the details of the derivation can be found in [Bishop:95]):

$$m_j^{new} = \frac{\sum_{d=1}^{D} p_{old}(j \mid s_d) s_d}{\sum_{d=1}^{D} p_{old}(j \mid s_d)}, \qquad 20$$

for the basis function variances,

$$\sigma_j^{new\,2} = \frac{\sum_{d=1}^{D} p_{old}(j \mid s_d)(s_d - m_j^{old})^2}{\sum_{d=1}^{D} p_{old}(j \mid s_d)}, \qquad 21$$

and for their priors:

$$p_{new}(j) = \frac{1}{D}\sum_{d=1}^{D} p_{old}(j \mid s_d), \qquad 22$$

where,

$$p(j \mid s_d) = \frac{p(s_d \mid j)p(j)}{\sum_{j=1}^{J} p(s_d \mid j)p(j)}. \qquad 23$$

Once the parameters of the density estimators have been set, classification proceeds by presenting the spectral signature s of the pixel under consideration to the set of N density estimators, to obtain the class conditional densities for each class $p(s|C_n)$. These densities can then be used with Bayes' theorem to derive the posterior probabilities upon which optimal classifications can be based:

$$p(C_n \mid s) = \frac{p(s \mid C_n)p(C_n)}{\sum_{m=1}^{N} p(s \mid C_m)p(C_m)} \qquad 24$$

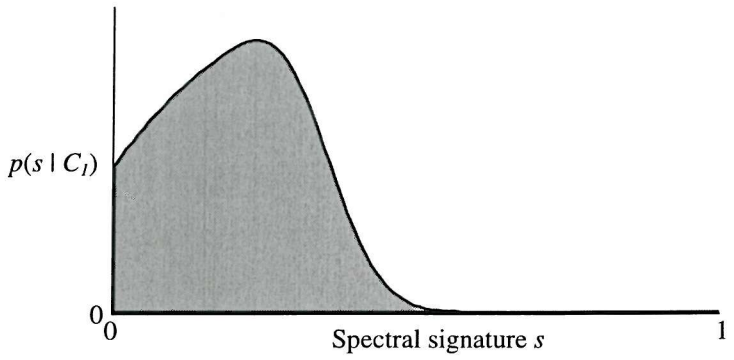

Figure 18: Class conditional probability of s given class 1.

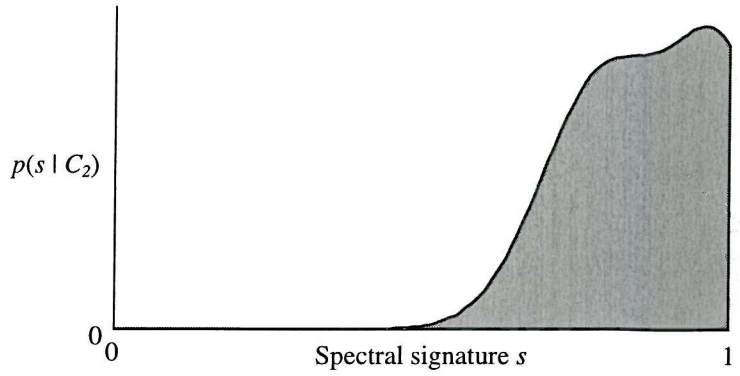

Figure 19: Class conditional probability of s given class 2.

Although in practical applications the direct and indirect approaches to classification may be used interchangeably, the direct approach will often give better performance. The reason for this can be seen from figures 18, 19, and 20. Figures 18 and 19 show two class-conditional probability densities $p(s|C_1)$ and $p(s|C_2)$ and figure 20 shows the posterior probability of class C_1 given a pixel of spectral signature s, $p(C_1|s)$, obtained by applying Bayes' theorem and assuming equal priors. The figures show that although the class conditionals are quite complex, much of the complexity lies away from the boundary between the two classes where the posterior probability makes its transition. This means that the posterior probability distribution itself is of a much lower complexity than either of the class conditionals, and hence could be described using a simpler model. This in turn implies that if the posterior distribution were modelled directly, then for a set of exemplars fixed size, the direct approach to classification would, on average, produce more accurate classifications [Ripley:96].

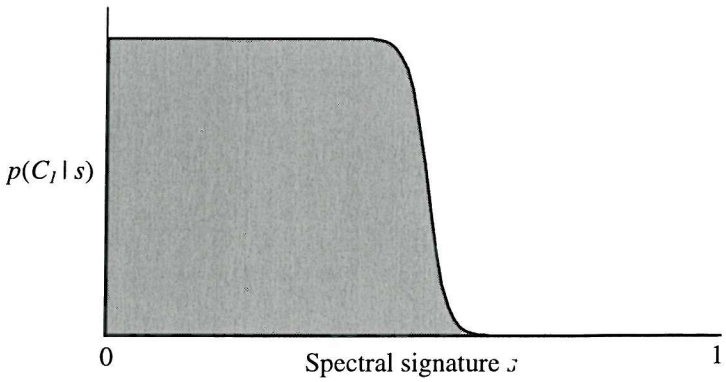

Figure 20: Posterior probability of class 1 given s.

6.3. Softened Classifications

Soft classifications were first used to derive information about subpixel cover in [Foody:96]. The technique is based on using the posterior probabilities estimated by some standard classification algorithms as fuzzy classifications, in a similar way to that of section 6.1 but with the exception that only pure pixels are used to train the classifiers. This omission means that the softened classifications are not identical to the posterior probabilities estimated at the outputs of a standard classifier, since softened classifier outputs strictly represent probabilities that hypotheses of the form "this pixel consists purely of crops" rather than "this pixel contains at least 50 % crops" as would more likely be the case with a crisp classifier.

Algorithm	Error Function	Number Of Basis Functions	Validation Set Error
Linear discriminant	Sum of squares	Not applicable	0.08469
Logistic discriminant	Sum of squares	Not applicable	0.05219
Softmax discriminant	Cross entropy	Not applicable	0.1603
MLP	Sum of squares	5	0.05260
"	Cross entropy	5	0.1594

Table 4: Model performances in producing softened classifications.

This section presents the results of experiments aimed at deriving softened classifications from neural network classifiers, and examines the relationship between soft and fuzzy classifications in more detail. To produce the soft classifiers used in this section, the training and test data sets were pre-processed so that they only contained pure pixels – pixels containing either 0 of 100 percent of the target class. Since, of the tall herb pixels, only 0.2 percent of the training pixels consisted purely of tall herb, soft classification experiments were performed only on the cereal data, for which 13,726 pixels in the training set were pure. To evaluate the potential of soft classification for fuzzy classification, five different models were trained and tested on pure data only. The MLPs were trained as though they were to be used as normal classifiers – that is, the test set was used to perform early stopping to prevent overfitting by training for a fixed period of 16 hours and selecting the parameters from that period that offered the best test set performance. These parameters were restored to the model which was then applied to the fuzzy validation data in order to evaluate its fuzzy classification performance. The results of this process are given in figures 21 to 25 and in table 4.

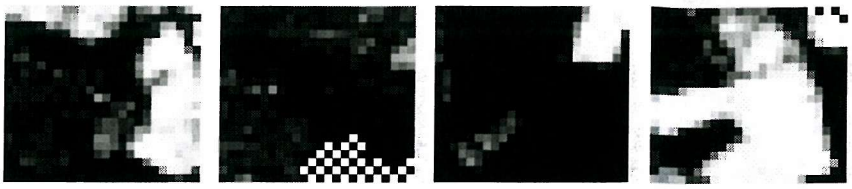

Figure 21: Cereal proportions estimated using a softmax discriminant trained using the cross entropy error on pure pixels only.

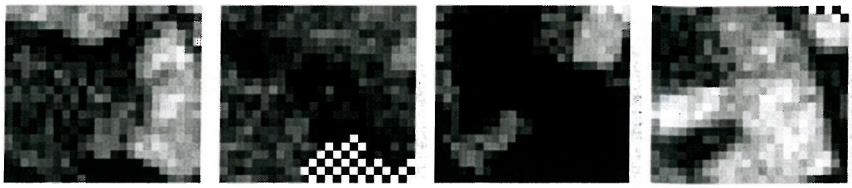

Figure 22: Cereal proportions estimated using a linear network trained using the sum of squares error on pure pixels only.

In general, the fuzzy classifications obtained by training a network on pure pixels only are slightly worse than fuzzy classifications obtained from a network trained on a hardened version of the original data set as was presented in section 6.1. The only exception to this rule is the logistic discriminant which performs better when trained on pure pixels only. As with the peculiarities of the fuzzy classification results generated by the crisp classifier, the details of the behaviour of the soft classifiers are difficult to explain. As with the crisp classifier, the addition of the logistic function to the output of the linear network produces a drastic improvement in performance. For the softened classifier however, the roles of the sum of squares and cross entropy function are reversed: sum of squares fuzzy classification performance deteriorates when the simple logistic discriminant is replaced by the more complex MLP, but cross entropy performance improves when the softmax discriminant is replaced.

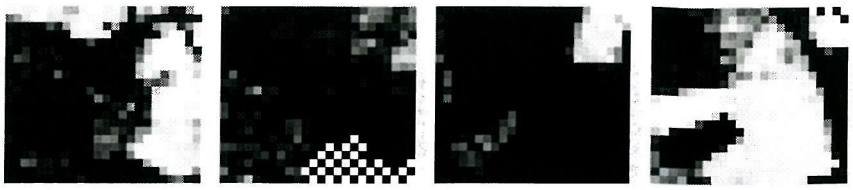

Figure 23: Cereal proportions estimated by a 6-5-2 MLP trained using the cross entropy error on pure pixels only.

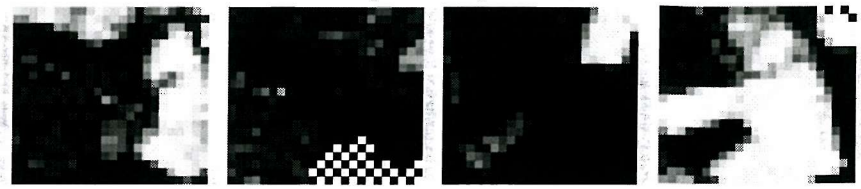

Figure 24: Cereal proportions estimated by a logistic network trained using the sum of squares error on pure pixels only.

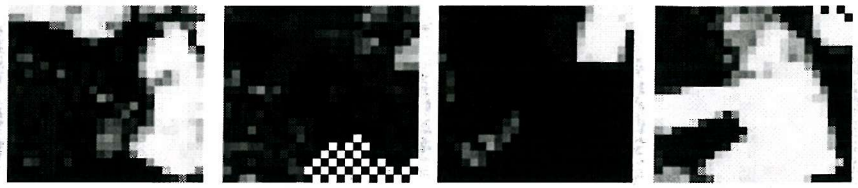

Figure 25: Cereal proportions estimated by a 6-5-1 MLP trained using the sum of squares error on pure pixels only.

As before, the differences between the techniques used here for performing fuzzy classification by soft classification are difficult to explain due to the interaction of model parameterisation, error function, and the distributions of the pure only pixels and the fuzzy validation data. Rather than discussing the specific features of the techniques' behaviour on the specific partition of the data set used in this thesis, the following section presents a short theoretical analysis of the relationship between posterior probabilities of class membership – the quantities estimated by crisp classifiers [Schurmann:96] which represent the most accurate results reported in this thesis thus far and the optimal fuzzy classifications, defined in this case as those that minimise either the sum of squares or cross entropy error functions over all possible data.

6.3.1. On the Relationship between Posterior Probabilities and Fuzzy Classifications

Since the softened outputs of conventional classifiers can often be interpreted as estimates of the posterior probabilities of class membership [Baum:87][Bishop:95][Cid-Sueiro:00], it is interesting to consider the relationship between such probabilities and the subpixel proportions that they are used to approximate when soft classification is employed in land cover mapping. This subsection considers this relationship in detail and concludes that although posterior probabilities of class membership of pixels are

likely to be positively correlated with subpixel proportions, the two quantities cannot, in general, be equal. This suggests that softened classifications should be avoided as a means of obtaining information concerning subpixel cover unless more direct means of fuzzy classification are not possible due, for example, to a lack of fuzzy membership information in the set of exemplar pixels.

The fuzzy classification that minimises the sum-of-squares and cross-entropy functions is equal to the mean of the distribution of subpixel memberships at each point in spectral space, as given below:

$$\mu_{opt} = \int \mu p(\mu \mid s) d\mu \qquad 25$$

For the purposes of this discussion, these subpixel memberships will be considered to be optimal – a reasonable assumption since they minimise the error functions (and hence maximise the equivalent likelihoods) over the true and unknown distribution of subpixel memberships. There are thus no alternative estimates that will produce, on average, smaller errors. Similarly, the posterior probability of class membership can be written in terms of the spectrum-conditional probability $p(\mu \mid s)$ as shown below:

$$p(C_n \mid s) = \int p(C_n \mid \mu) p(\mu \mid s) d\mu \qquad 26$$

In order for posterior probabilities of class membership and fuzzy classifications to be equal, it is necessary for equations 25 and 26 to be equal. This is only guaranteed for arbitrary $p(\mu \mid s)$ if the posterior probability of class membership given a certain subpixel membership is equal to the subpixel membership, i.e. that

$$\mu = p(C \mid \mu) \qquad \forall \mu \in [0,1]^N \qquad 27$$

If pixel classification is unambiguous given subpixel memberships, the vector of posterior probabilities, $p(C|\mu)$ will always have a one in the n^{th} position where $1 \leq n \leq N$ and zeros in all others, and hence cannot satisfy the above condition. This shows that the posterior probability of the membership of pixels in classes where class membership can be determined unambiguously from subpixel area proportions cannot be guaranteed to equal the optimal subpixel area proportion estimates for all forms of $p(\mu|s)$.

An alternative approach to understanding this issue is shown for a simple two class case in figure 26. This figure shows the distribution of possible subpixel proportions at some point s in spectral space. While the optimal fuzzy classification is given by the mean of the proportion distribution, the posterior probability of class membership is given by the area of the shaded region. Clearly, by modifying the probability distribution for $\mu < 0.5$, it is possible to change the distribution's mean, and hence the optimal fuzzy classification, without changing the posterior probability of class membership. The relationship between these posterior probabilities and the optimal fuzzy classification is thus a relatively weak one.

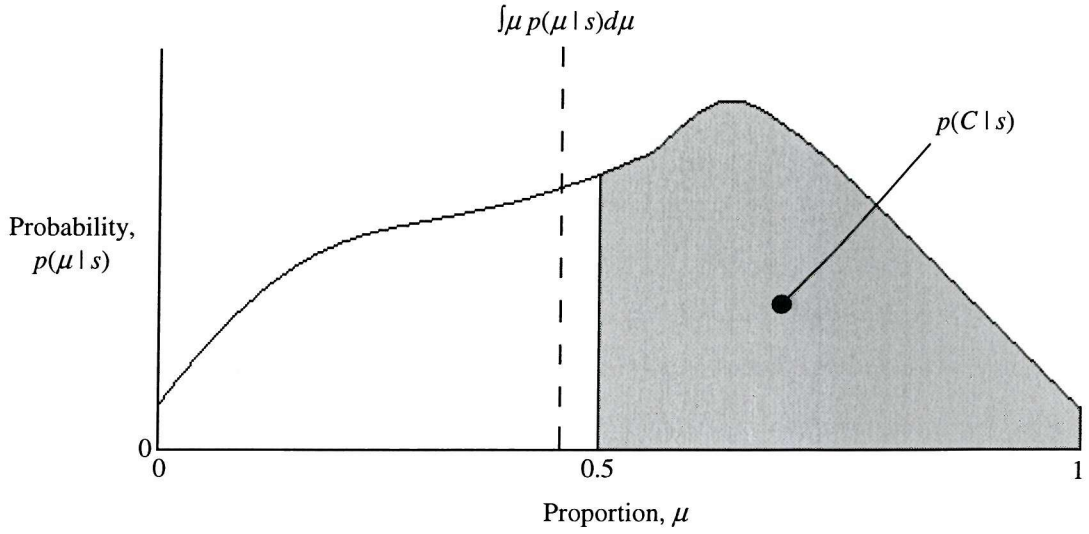

Figure 26: An illustration of the relationship between posterior probabilities and optimal fuzzy classifications.

It is interesting to note that the equivalence of posterior probabilities and fuzzy memberships does hold for special forms of $p(\mu|s)$. One such form occurs when all pixels with spectral signature s are pure (consist of a single subpixel cover class). Under these circumstances, $p(\mu|s)$ is zero except when μ has a one in the n^{th} position where $1 \leq n \leq N$, and zeros in all others. In addition to this, positive correlation between posterior probabilities defined in terms of subpixel area and fuzzy classifications, as was observed in [Foody:96c], are to be expected. This is because there is always a positive correlation between $p(C \mid \mu)$ and μ for pure pixels by virtue of the way in which class membership is defined.

7. Fuzzy Classification

There are two ways of performing fuzzy classification, each of which is analogous to one of the two ways of performing crisp classification: The first approach to crisp pixel classification is to use pixels of known class membership to construct density estimators for the class conditional distributions in spectral space. When a new pixel is observed, the class conditional densities estimated can be used with Bayes' theorem to derive estimates of the posterior probabilities of class membership of the pixel in each of the target classes upon which crisp classifications can be based. The fuzzy classification analogue of this process uses a data set of pixels of known fuzzy membership to place fuzzy basis functions in spectral space from which fuzzy memberships are derived by a process of normalisation.

Most such implementations of both crisp and fuzzy classifiers use highly parametric forms for the class conditional distributions or fuzzy basis functions and hence usually achieve only very limited performance. However, this limitation is not implicit in the algorithms but is specific to particular implementations, and significant performance benefits can be demonstrated for both crisp and fuzzy classifiers through the use of more flexible models. This is well known in the case of crisp classification, but less so for fuzzy classification where the ad-hoc choice of highly parametric basis functions dominates. The analogy between crisp and fuzzy classification presented here is further extended in section 7.1.1, where it is shown that the use of semi-parametric representations of fuzzy basis functions is a natural extension of the standard fuzzy classifier and that their application can produce drastic improvements in performance.

The second and most direct way of producing a crisp pixel classifier is to use a set of pixels of known class membership to derive a model of the relationship between a pixel's spectral signature and its class membership such that when a new pixel is observed, the model can be used to derive an estimate of the class membership of the pixel. Certain types of these models can be shown to produce approximations to the posterior probabilities that a pixel belongs to each of the target classes – a property that can be exploited in producing fuzzy classifiers and will be discussed in greater detail later. The fuzzy classification equivalent of this direct approach to classification is to use a set of pixels of known fuzzy membership to derive a model of the relationship between pixel spectral signatures and their fuzzy memberships. When a new pixel with unknown fuzzy membership is observed, the model can be used to obtain an estimate of its membership. This approach currently dominates the area proportion estimation

literature and will be discussed at length in a later section. The following sections return to the indirect methods of performing crisp and fuzzy classification by examining two algorithms, one a crisp classifier, the other a fuzzy classifier, which are shown to be closely related. Table 5 presents a summary of the relationships between direct and indirect crisp and fuzzy classifiers.

	Crisp Classification	Fuzzy Classification
Data	Requires data of known class membership	Requires data of known fuzzy membership
Indirect	Model class conditional densities	Model fuzzy basis functions
	Derive posterior probabilities of class membership using Bayes' theorem	Derive fuzzy memberships by normalisation
Direct	Model relationship between spectral signature and class membership	Model relationship between spectral signature and fuzzy membership

Table 5: Comparison of crisp and fuzzy classification.

7.1. Indirect Fuzzy Classification

As a typical example of a supervised fuzzy classifier, this section considers the influential work described in [Wang:89]. The structure of the fuzzy classifier described therein allows it to be viewed as a neural network, as shown in figure 27. At the top of the figure are the classifier inputs, which usually consist of the spectral signature of the pixel to be classified. This information is propagated to a series of non-linear basis functions in the network's hidden layer that usually take the form $p(|s-m|,\sigma)$ where s is the pixel's spectral signature, m is the basis function's centre, σ is a width parameter, which controls the rate at which p changes with s, and $p(.)$ is a monotonically decreasing function of the difference $|s-m|$. $p(.)$ therefore contains information about the distance between the basis functions' centres and pixels' spectral signatures.

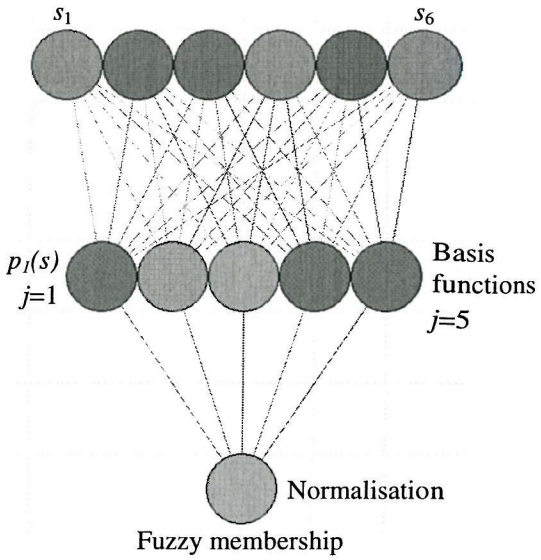

Figure 27: A fuzzy classifier with five basis functions.

One basis function is assigned to each of the target fuzzy classes, and used to represent the localisation of those classes in spectral space. The centre of the n^{th} basis function is found using:

$$m_n = \frac{\sum_{d=1}^{D} f_n(s_d) s_d}{\sum_{d=1}^{D} f_n(s_d)} \qquad 28$$

where there are D pixels in the set of exemplars, the pixels have spectral signature s_d and fuzzy membership in the n^{th} class of $f_n(s_d)$. Thus, the centre of the basis function representing class n is the mean of the spectral signatures of all pixels weighted by their membership in class n. The widths of the basis functions may be determined in a similar way using:

$$\sigma_n^{\,2} = \frac{\sum_{d=1}^{D} f_n(s_d)(s_d - m_n)^2}{\sum_{d=1}^{D} f_n(s_d)} \qquad 29$$

where, for simplicity, only one spectral band has been assumed. Both of the above formulae are independent of the estimated fuzzy memberships, and hence the parameters of the fuzzy classifier may be found from a single application of the above

equations. This is an important advantage of the fuzzy classifier over neural networks such as the MLP, which typically require many thousands of applications of the parameter update equations and hence many minutes, if not hours, of training.

Once the parameters of the fuzzy classifier have been determined, the output of the fuzzy classifier $f_n(s_d)$, which represents the fuzzy membership of the pixel under consideration in the n^{th} class, is found by normalising the output of the n^{th} hidden layer basis function by the sum of the outputs of all basis functions,

$$f_n(s_d) = \frac{p_n(s_d)}{\sum_{j=1}^{N} p_j(s_d)} \qquad 30$$

where $p_n(s_d)$ is a Gaussian basis function,

$$p_n(s_d) = \frac{1}{(2\pi\sigma_n^2)^{1/2}} \exp\left(-\frac{(s_d - m_n)^2}{2\sigma_n^2} \right) \qquad 31$$

Fuzzy memberships for all classes of interest may be derived by constructing a separate fuzzy classifier for each class. This process is more efficient than it may initially appear, since all fuzzy classifiers have the same parameters and differ only in the normalisation stage, rendering it unnecessary to duplicate either the training procedure or the input to hidden and hidden layer structures. Following the work presented in [Wang:89], the basis functions used to generate the fuzzy classifications reported here are Gaussian and for simplicity, only one spectral band has been considered. It should be noted that other applications of this type of fuzzy classifier have used different basis functions (see, for example [Foody:96c][Fisher:90][Robinson:85]), the choices of which are largely arbitrary, despite the fact that it constrains the form of the fuzzy partitions that can be realised by the classifier. The next section returns to this issue by providing a new analogy between the fuzzy classifier and EM density estimator that not only provides a meaningful interpretation of the outputs of the basis functions but also motivates the use of non-parametric substitutes to the forms normally used.

Figures 28 and 29 show the estimates made by the fuzzy classifier for the subpixel proportions on the unseen areas in the Stoughton image. Figure 28 shows the cereal proportion predicted by the fuzzy classifier giving a cross-entropy performance of 0.5161 – considerably worse than that of the softened classifiers considered earlier. The performance of the fuzzy classifier on the tall herb data is similarly poor, achieving a

cross-entropy error as high as 0.3112. There are essentially two reasons for these failures:

- The fuzzy classifier is highly parametric in the sense that the shape of the basis functions, which determine the range of functions the fuzzy classifier can realise are determined apriori. Using the fuzzy classifier is rather like using other inflexible and highly parametric models such as linear networks and would thus be expected to perform poorly on complex modelling problems.

- The fuzzy classifier contains no priors on the basis functions to adjust for the relative proportions of the different classes in an image. This problem is particularly pronounced in the case of the tall herb class, where few pixels contain more than a very small proportion. This issue is discussed again later in this thesis where it is shown that crisp classifiers and fuzzy classifiers, such as the one discussed here, are equivalent.

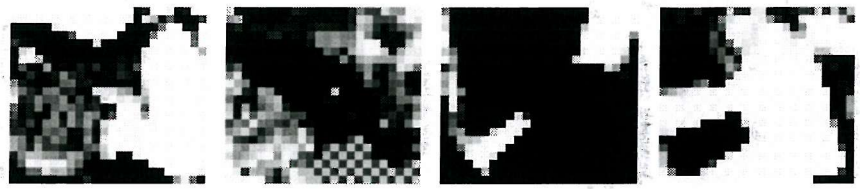

Figure 28: Cereal proportions predicted by a fuzzy classifier

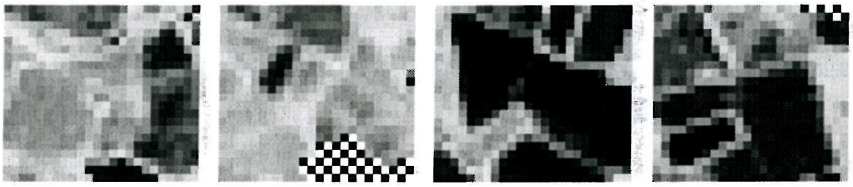

Figure 29: Tall herb proportions predicted by a fuzzy classifier.

The following section uses the novel probabilistic interpretation to highlight the close relationship that exists between the indirect approach to fuzzy classification described in this section and crisp classification. This equivalence provides new and important clues as to how the structure of the fuzzy classifier can be changed to improve its performance. The section concludes by presenting a novel analysis of the asymptotic behaviour of the fuzzy classifier, which shows that as the quantity of training data and flexibility of the classifier are increased, the fuzzy classifications it produces converge to the optimal fuzzy classifications.

7.1.1. The Equivalence of Fuzzy and Crisp Classifiers

This section shows that the fuzzy classifier used in [Wang:90] (and similar to those used in [Melgani:00] and [Chittineni:81]) is a special case of a crisp classifier under the probabilistic interpretation. In particular, if the crisp classifier uses the indirect method and the class conditional density models consist of only a single Gaussian, and the priors and posteriors for each Gaussian are unity, i.e.

$$p(j)=1 \quad \wedge \quad p(j \mid s_n)=1 \qquad \forall j : 1 \leq j \leq J \tag{32}$$

the update equations for the basis function parameters (derived in appendix A) reduce to:

$$m_j^{new} = \frac{\sum_{d=1}^{D} s_d \mu_d}{\sum_{d=1}^{D} \mu_d} \tag{33}$$

for their centres, and

$$(\sigma_j^{new})^2 = \frac{\sum_{d=1}^{D} \mu_d (s_d - m_j^{new})^2}{\sum_{d=1}^{D} \mu_d} \tag{34}$$

for their variances. Thus, the equations for finding the parameters of an EM density estimator used in a crisp classifier are the same as those for the supervised fuzzy classifier where

$$\mu_d = f_j(s_d) \tag{35}$$

and the area proportion estimate and posterior probability of class membership of a subpixel point in the m^{th} class C_m is:

$$p(C_m \mid s_d) = \frac{p(s_d|C_m)p(C_m)}{\sum_{n=1}^{N} p(s_d|C_n)p(C_n)} \tag{36}$$

where

$$p(C_n) = \frac{1}{D}\sum_{d=1}^{D}\mu_d(C_n), \qquad 37$$

and the fuzzy classifier has once again assumed all priors to be equal.

In a normal application of the EM algorithm, the re-estimation equations are repeatedly applied until the algorithm is considered to have converged. Using the probabilistic interpretation however, the posterior probabilities $p(C \mid s_d)$ are the true subpixel proportions and are therefore known. Thus, the optimal values of the basis function parameters are found from a single application of the re-estimation equations in the special case that the class conditional densities are each modelled by a single Gaussian as is the case in the equivalence described here.

The probabilistic interpretation and the comparison of the indirect approach to deriving fuzzy memberships with the indirect approach to deriving crisp classifications provides useful insight into the operation of the fuzzy classifier, and suggests ways in which it may be improved. In terms of interpretation, the similarity between the equations for the parameters suggests that the values of the intermediate fuzzy basis functions can be interpreted as representing class conditional probability densities. The normalisation process can then be seen as an application of Bayes' theorem to convert the class conditionals into the pointwise posterior probabilities that, by the probabilistic interpretation, are equivalent to area proportions.

In terms of improving performance, density estimators typically contain priors for each class – parameters that are absent from the fuzzy classifier, but which can produce a significant improvement in performance, as can be seen by comparing figures 28 and 34, figures 29 and 30 and tables 7 and 6. In the case of the tall herb class, the inclusion of the prior reduces the validation set error from 0.3112 to 0.1084. In addition to this, EM density estimators can be semi-parametric, that is, they use a superposition of a variable number of basis functions to model each class conditional density rather than the single basis function used by the fuzzy classifier. The derivation of the equations for updating the density estimator parameters using EM provided in the appendix is presented in these terms, and the results in figures 30 to 37 show that increasing the flexibility of the density estimator through the addition of basis functions improves the accuracy of the area proportion estimates made far beyond those of the standard fuzzy classifier. Note that the EM fuzzy classifier appears to overfit the tall herb data when using 10 basis

functions. This problem may be overcome by adding regularisation to the density estimators [Bishop:95] or by initialising the density estimator to have low complexity (giving the basis functions large variances), using an algorithm that updates the parameters more slowly than EM over a larger number of iterations, and using early stopping.

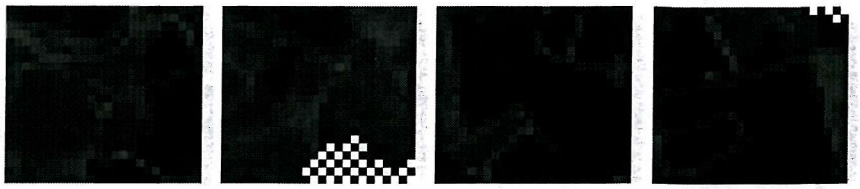

Figure 30: Tall herb proportions predicted by the EM fuzzy classifier with 1 basis function.

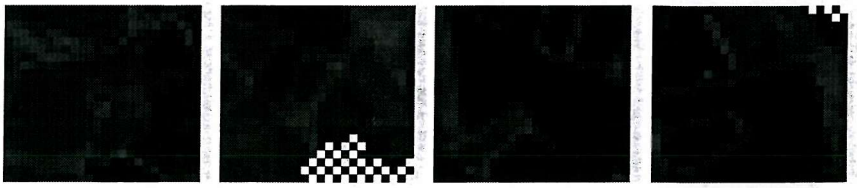

Figure 31: Tall herb proportions predicted by the EM fuzzy classifier with 2 basis functions.

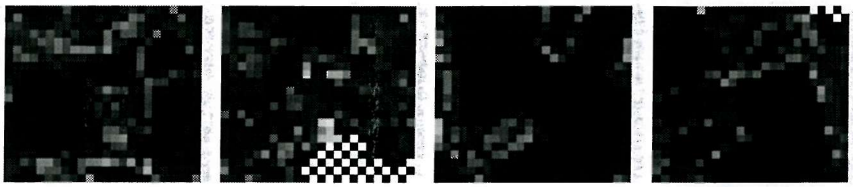

Figure 32: Tall herb proportions predicted by the EM fuzzy classifier with 5 basis functions.

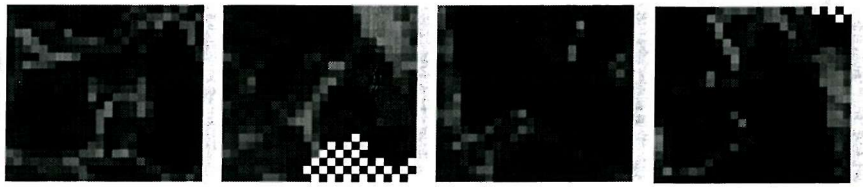

Figure 33: Tall herb proportions predicted by the EM fuzzy classifier with 10 basis functions.

Algorithm	Number of Basis Functions	Validation Set Error
Fuzzy classifier	Not applicable	0.5161
EM fuzzy classifier	1	0.4590
"	2	0.2979
"	5	0.2204
"	10	0.2210

Table 6: Comparison of the standard and EM fuzzy classifiers on fuzzy tall herb data.

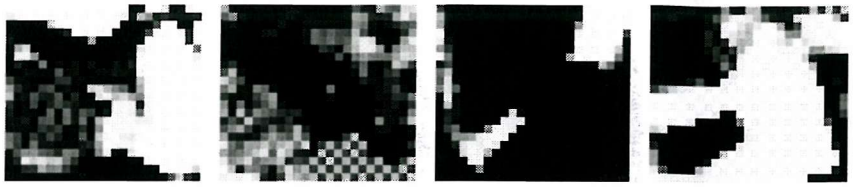

Figure 34: Cereal proportions predicted by the EM fuzzy classifier with 1 basis function.

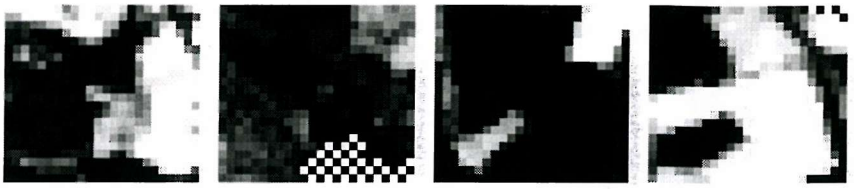

Figure 35: Cereal proportions predicted by the EM fuzzy classifier with 2 basis functions.

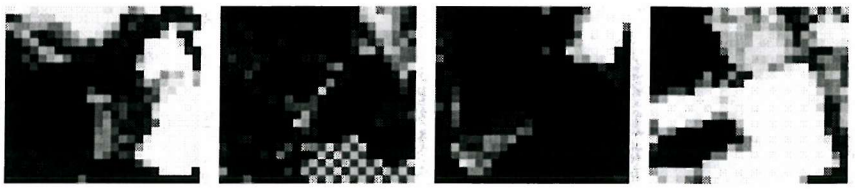

Figure 36: Cereal proportions predicted by the EM fuzzy classifier with 5 basis functions.

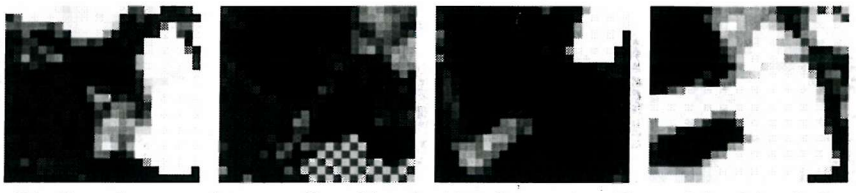

Figure 37: Cereal proportions predicted by the EM fuzzy classifier with 10 basis functions.

Algorithm	Number of Basis Functions	Best Test Set Error
Fuzzy classifier	Not applicable	0.3112
EM fuzzy classifier	1	0.1084
"	2	0.1081
"	5	0.1081
"	10	0.1055

Table 7: Comparison of the standard and EM fuzzy classifiers on fuzzy cereal data.

An important but as yet unanswered question concerning the fuzzy classifier is whether its fuzzy classification estimates approach the optimal fuzzy classifications as the quantity of training data and the flexibility of the classifier increase. A positive answer to such a question would provide some level of confidence that the fuzzy classifier is capable, in principle, of performing optimal fuzzy classification and that its outputs can reasonably be interpreted as proportion estimates. The following analysis shows that, under ideal circumstances, the indirect method of fuzzy classification can indeed produce optimal fuzzy classifications. If an unlimited quantity of data were available, the error function used to train the EM based fuzzy classifier would be:

$$E = -\int p(C \mid s)p(s) \ln p_{est}(s \mid C)ds \qquad 38$$

where $p(C|s)$ is the average posterior probability that a subpixel point is in class C (equivalent to the proportion of the subpixel area of the pixel under consideration covered by class C), given that a pixel has spectral signature s, and $p_{est}(s|C)$ is the quantity estimated by the density estimator component of the fuzzy classifier – as will be shown, this is the probability that the spectral signature s is observed given that a subpixel point is in class C. If the density estimator is allowed to become arbitrarily flexible, then when $p_{est}(s|C)$ minimises E,

$$\frac{\partial E}{\partial p_{est}(s \mid C)} = 0 \qquad 39$$

for all s. I.e. for two points s_1 and s_2 in spectral space,

$$\frac{\partial E}{\partial p_{est}(s_1 \mid C)} = \frac{\partial E}{\partial p_{est}(s_2 \mid C)} . \qquad 40$$

But, differentiating equation 38 with respect to $p_{est}(s_1|C)$ and $p_{est}(s_2|C)$ gives:

$$\frac{\partial E}{\partial p_{est}(s_1 \mid C)} = \frac{p(C \mid s_1)p(s_1)}{p_{est}(s_1 \mid C)} , \qquad 41$$

and

$$\frac{\partial E}{\partial p_{est}(s_2 \mid C)} = \frac{p(C \mid s_2)p(s_2)}{p_{est}(s_2 \mid C)} , \qquad 42$$

which, when equated give:

$$\frac{p(C \mid s_1)p(s_1)}{p_{est}(s_1 \mid C)} = \frac{p(C \mid s_2)p(s_2)}{p_{est}(s_2 \mid C)} . \qquad 43$$

Dividing each side by $p(C|s_2)p(s_2)$ and multiplying by $p_{est}(s_1|C)$ gives:

$$\frac{p(C \mid s_1)p(s_1)}{p(C \mid s_2)p(s_2)} = \frac{p_{est}(s_1 \mid C)}{p_{est}(s_2 \mid C)}, \quad 44$$

which implies that $p_{est}(s|C)$ is proportional to $p(C|s)$

$$p_{est}(s \mid C) = \alpha p(C \mid s)p(s), \quad 45$$

where α is a constant of proportionality. Fortunately, α can be determined, since the choice of Gaussian basis functions guarantees that

$$\int p_{est}(s \mid C)ds = 1, \quad 46$$

such that

$$\alpha \int p(C \mid s)p(s)ds = 1, \quad 47$$

which, since $\int p(C|s)p(s)ds=\int p(C,s)ds=p(C)$ implies that

$$\alpha = \frac{1}{p(C)}. \quad 48$$

Thus,

$$p_{est}(s \mid C) = \frac{p(C \mid s)p(s)}{p(C)}, \quad 49$$

which shows that the density estimator components of the fuzzy classifier do indeed model the class conditional probability densities. Recalling that, in practice, the EM fuzzy classifier operates by using two density estimators, their outputs representing $p_{est}(s|C)$ and $p_{est}(s|\neg C)$ ($\neg C$ meaning not class C) and being combined using Bayes' theorem:

$$p_{est}(C \mid s) = \frac{p_{est}(s \mid C)p(C)}{p_{est}(s \mid C)p(C) + p_{est}(s \mid \neg C)p(\neg C)}, \quad 50$$

where the closed world property of C and not C implies that the denominator reduces to $p_{est}(s|C)+p_{est}(s|\neg C)=p_{est}(s)$, which leads to:

$$p_{est}(C \mid s) = \frac{p_{est}(s \mid C)p(C)}{p_{est}(s)}. \quad 51$$

which, from Bayes' theorem, implies that:

$$p_{est}(C \mid s) = p(C \mid s) \quad 52$$

In other words, the EM based fuzzy classifier that is trained by minimising equation 38 produces outputs that can be interpreted as subpixel area proportion estimates, in the sense that as the classifier is given increasingly large quantities of data and allowed greater flexibility, the estimates it produces converge to the optimal proportion estimates. In practice, it is likely that even the augmented fuzzy classifiers described here would, on average, perform less well than more direct techniques such as the MLP as they suffer from exactly the same problem as the indirect crisp classifiers to which they are equivalent – they model the area proportions (equivalent to posterior probabilities under the probabilistic interpretation) via potentially complex class conditional distributions.

The following section describes the current state of the art in fuzzy classification – the use of complex flexible models such as neural networks to directly estimate proportion information from spectral data. The section describes the asymptotic behaviour of neural network learning and how it relates to the problem of area proportion estimation. It also uses the probabilistic interpretation to present a novel analysis of the role of the cross entropy function in training neural networks for fuzzy classification.

7.2. Direct Fuzzy Classification

The most successful techniques currently in use for extracting subpixel area proportion information are non-linear regression algorithms that use a set of exemplar pixels of

known fuzzy membership to learn a function that directly maps a pixel's spectral signature to a subpixel proportion estimate. Of the functions used, neural networks have been amongst the most popular due to their accessibility and robustness. This section presents the results of applying a range of non-linear regression algorithms to the FLIERS data set and then considers the question of the appropriate error function for fuzzy classification – an issue debated in [Foody:95b]. In particular, the probabilistic interpretation is used to provide new insight into the role of the cross entropy function in fuzzy classification.

Direct fuzzy classification consists of two largely independent branches, the first typified by the application of linear models to model the relationship between an observed spectral signature and subpixel proportions and the second and most important as far as this thesis is concerned is the application of flexible non-linear models such as neural networks. Although research into linear techniques continues to this day, there is good reason to believe that in practice, the relationship between a pixel's spectral signature and the optimal fuzzy classification is non-linear. This seems to suggest that robust non-linear models should be used, particularly if there are large quantities of labelled exemplar pixels available, as is the case with the FLIERS data set.

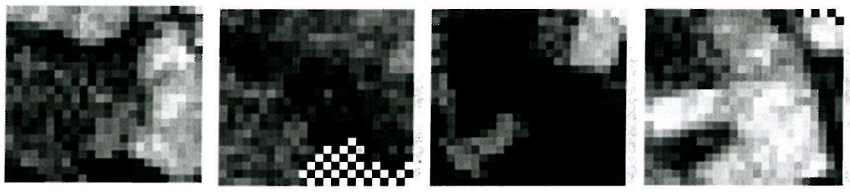

Figure 38: Cereal proportions estimated by a linear network trained using the sum of squares error on fully fuzzy data.

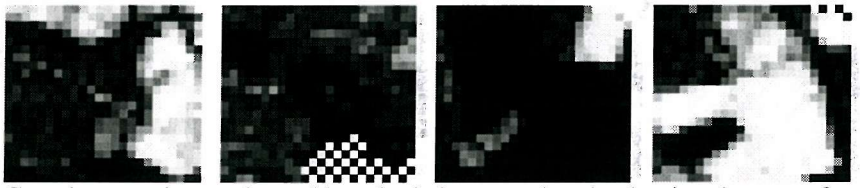

Figure 39: Cereal proportions estimated by a logistic network trained using the sum of squares error on fully fuzzy data.

The results of applying single layer linear and logistic networks to the FLIERS data set are shown in figures 38 to 43 and table 8. As usual, the simple linear network performs poorly due to its inability to suppress input variance in its output. This results in a mean squared error rate of 0.08289 as opposed to 0.05092 for the logistic network on the cereal data. Although the performance of the single layer logistic network is the best yet seen, there is very little obvious difference that can be seen between the images of

the predicted proportions in figure 38 and earlier images. This suggests that such images can only offer a fairly coarse qualitative indication of performance.

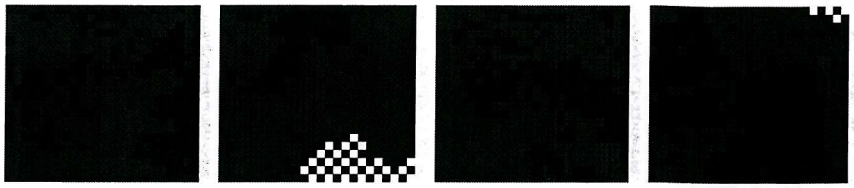

Figure 40: Tall herb proportions estimated by a linear network trained using the sum of squares error on fully fuzzy data.

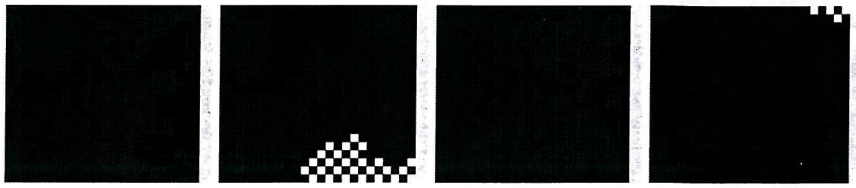

Figure 41: Tall herb proportions estimated by a logistic network trained using the sum of squares error on fully fuzzy data.

For the tall herb class, the proportions predicted are all small and little variance in predictions can be seen other than in the enhanced images of figures 42 and 43. Each of these images, representing the predictions made by the linear and logistic networks respectively have broadly similar characteristics: Although the predictions they make generally bear little resemblance to the true distribution of tall herb, they are quite similar to each other. For example, the absence of tall herb is correctly predicted for the lower field in the fourth subimage. This suggests that either the difficulty in predicting the proportion of tall herb is due to spectral confusion or neither the linear nor the logistic network can accurately predict the proportion of tall herb because they are too inflexible to learn the mapping from spectral signature to subpixel proportion.

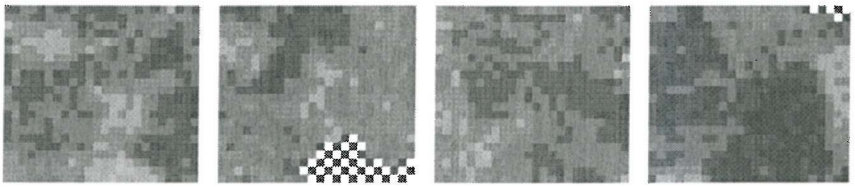

Figure 42: Enhanced image of the tall herb proportions estimated by a linear network trained by minimising the sum of squares error on fully fuzzy data.

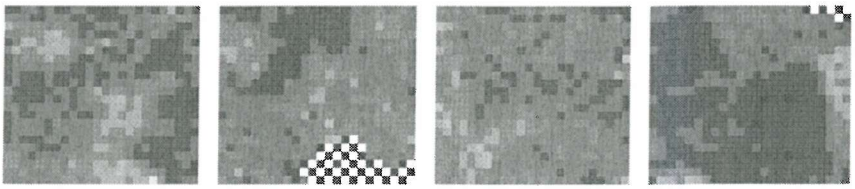

Figure 43: Enhanced image of the tall herb proportions estimated by a logistic network trained using the sum of squares error on fully fuzzy data.

The following section looks at neural networks – a class of semi-parametric models that are known to be universal approximators and hence theoretically capable of reproducing any function to arbitrary accuracy [Ripley:96]. Before applying neural networks to the fuzzy classification problem, it is interesting to consider the question of which error function should be used [Foody:95b]. This issue is discussed at the beginning of the next section, where the probabilistic interpretation is used to provide new insight into the role of the cross entropy function in area proportion estimation.

Network	Class	Error Function	Validation Error
Linear	Cereal	Sum of squares	0.08289
Logistic	Cereal	Sum of squares	0.05092
Linear	Tall herb	Sum of squares	0.02405
Logistic	Tall herb	Sum of squares	0.02397

Table 8: Comparison of fully fuzzy classifiers with linear and logistic output nodes

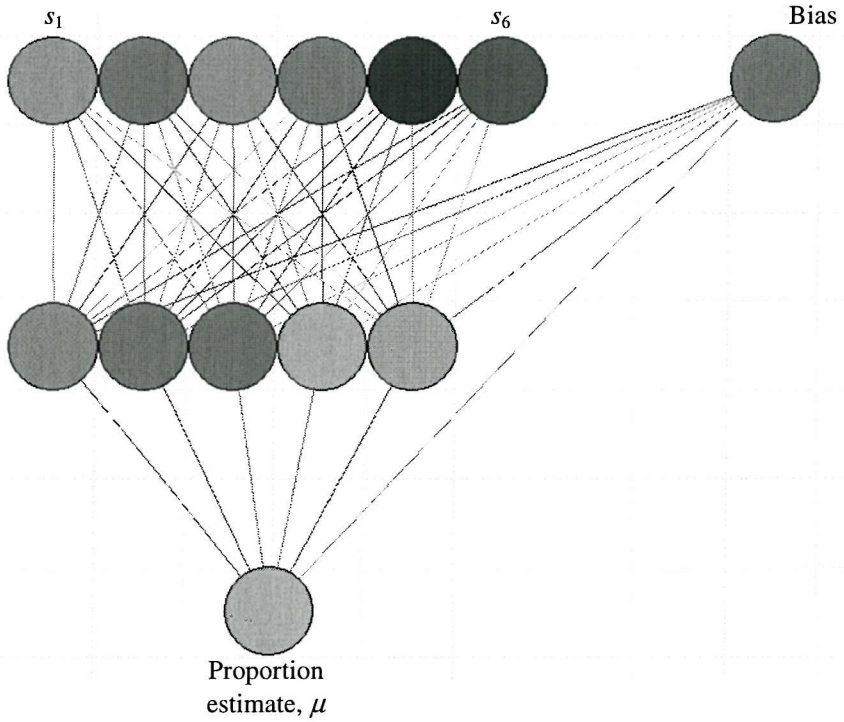

Figure 44: An MLP that uses five hidden neurons to estimate a single proportion from six spectral bands.

7.2.1. On the Cross Entropy Error and Fuzzy Classification

In most applications, neural networks are trained by minimising either the sum of squares or cross entropy error functions over the training set. While the sum of squares error function is normally used for regression problems and the cross entropy function for classification it is less clear which function should be used in the fuzzy classification of remotely sensed image pixels. This question was considered in detail in [Foody:95b] which produced several arguments and some empirical results in favour of the cross entropy function. This section uses the probabilistic interpretation to provide a novel theoretical justification for the use of the cross entropy function in the fuzzy classification of remotely sensed image pixels.

Consider a set of N classes, $C_n : 1 \leq n \leq N$ where C_n is the label of the n^{th} class. If this set is closed world and the classes are mutually exclusive, the probability distribution of class memberships obtained from random sampling of the subpixel area will be multinomial:

$$p(C \mid x, y) = \prod_{n=1}^{N} p(C_n \mid x, y)^{C_n(x,y)} \qquad 53$$

where $C(x,y)$ is a vector indicating the class membership of the subpixel point (x,y). For example, if $(x,y) \in C_n$ then $C(x,y)$ has a one in the n^{th} position and zeros in all others. $p(C_n|x,y)$ is the posterior probability of membership of the subpixel point (x,y) in class C_n. The probability of D such points having class membership C where C is now a matrix of D rows of vectors each indicating the class membership of one of the D points is given by:

$$p(C \mid D) = \prod_{d=1}^{D} \prod_{n=1}^{N} p(C_n \mid x_d, y_d)^{C_n(x_d, y_d)} \qquad 54$$

A neural network would typically be trained to classify such a set of points by using the maximum likelihood procedure. That is, the network would model the distribution parameters $p(C_n|x_d y_d)$ so as to maximise the probability that the distribution would reproduce the set of training patterns. Using the probabilistic interpretation, the distribution parameters are equal to the subpixel area proportions such that:

$$p(C \mid D) = \prod_{d=1}^{D} \prod_{n=1}^{N} \mu_{est}(C_n \mid P)^{C_n(x_d, y_d)} \qquad 55$$

where $\mu_{est}(C_n|P)$ are the neural network estimates of the distribution parameters. In other words, given a set of D subpixel points of class membership $C_n(x_d, y_d)$, maximum likelihood subpixel area proportion estimates may be obtained by finding the area proportions which maximise equation 55. It is however, possible to extend this by rearranging and considering the case when D becomes infinitely large. As this limit is approached, the proportion of the D samples belonging to class C_n approaches the proportion of the subpixel area covered by C_n. Taking the outer product inside the power thus gives:

$$p(C \mid D) = \prod_{n=1}^{N} \mu_{est}(C_n \mid P)^{D\mu(C_n|P)} \qquad 56$$

This makes it possible to simulate the effect of training a neural network on an infinitely large number of subpixel samples. To find the maximum likelihood subpixel area proportion estimates, it is convenient to minimise the negative logarithm of the likelihood given in equation 56 rather than maximise the likelihood itself. The negative log-likelihood is given by:

$$-\ln\left|p(C \mid D)\right| = -\sum_{n=1}^{N} D\mu(C_n \mid P) \ln \mu_{est}(C_n \mid P) \qquad 57$$

The multiplicative constant D is independent of the distribution parameters and hence does not change the set of parameters that maximises the likelihood. For this reason the D term may be ignored when maximising equation 57, such that the problem of finding the subpixel area proportions that maximise the likelihood is equivalent to finding the values for $\mu_{est}(C_n|P)$ which minimise:

$$E = -\sum_{n=1}^{N} \mu(C_n \mid P) \ln \mu_{est}(C_n \mid P) \qquad 58$$

which is the same as minimising the cross entropy error between the true and estimated subpixel area proportions. Multiple exemplar pixels may easily be accommodated by accumulating the expected error over the set of exemplars, so that for D pixels,

$$E = -\sum_{d=1}^{D}\sum_{n=1}^{N}\mu(C_n \mid P_d)\ln\mu_{est}(C_n \mid P_d) \qquad 59$$

The above derivation of the cross entropy function indicates that the fuzzy classifications that minimise the cross entropy function maximise the probability that a large number of samples drawn from a pixel with the subpixel proportions given by the fuzzy classification would have the same crisp class memberships as an equal number of samples drawn from a pixel with the target subpixel proportions. Alternative discussions of the use of the cross-entropy function for fuzzy classification, which focus on its information theoretic basis and the interpretability of the resulting error measures may be found in [Foody:95b] and [Foody:96c].

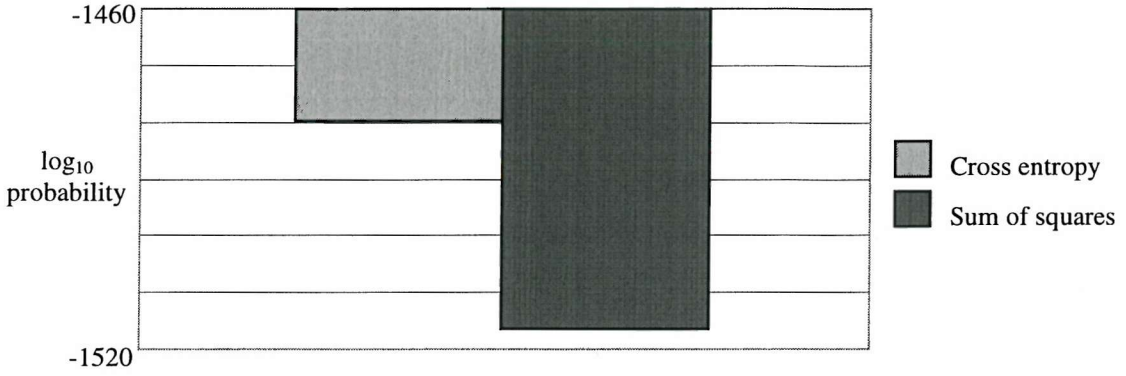

Figure 45: Log likelihoods of the validation set given cross entropy and sum of squares trained networks.

Figure 45 shows the results of an experiment conducted to evaluate the performance of a sum of squares and a cross entropy trained fuzzy classifier in terms maximising the likelihood that a set of subpixel samples from unseen pixels are generated with the correct memberships. The analysis of the cross entropy function that was presented earlier, suggests that the cross entropy trained fuzzy classifier should offer superior performance in this test. Figure 45 shows the results of this experiment in terms of the natural log joint probability of the fuzzy classifiers correctly predicting the class membership of all subpixel samples where one sample is drawn from each pixel in a validation set of 1,658 pixels. As shown, the cross entropy trained fuzzy classifier predicted the test set with log probability of approximately $-1.48\text{x}10^3$, whist the sum of squares trained fuzzy classifier only achieved a log probability of about $-1.52\text{x}10^3$. This means that the cross entropy trained fuzzy classifier is roughly 10^{40} times more likely to assign correct memberships to all the subpixel samples than the sum of squares trained fuzzy classifier.

Technique	Data Set	Error Function	Structure	Validation Set Error
MLP	Cereal	Sum of squares	6-2-1	0.05011
"	"	"	6-5-1	0.04897
"	"	"	6-10-1	0.04779
"	"	Cross entropy	6-2-1	0.1427
"	"	"	6-5-1	0.1379
"	"	"	6-10-1	0.1319

Table 9: Comparison MLPs trained with sum of squares and cross entropy functions on the cereal data.

Tables 9 and 10 and figures 46 to 57 give the results of applying sum of squares and cross entropy trained MLP fuzzy classifiers to the FLIERS data set. Note that these results are the best yet obtained for both the cereal and tall herb data sets. The minimum degree of improvement is around 10 percent for both data sets, with the exception of the logistic discriminant, which performs nearly as well as the MLP. This may be because the cereal proportions are so tightly clustered around zero and one as was shown in figure 4 that an algorithm such as the logistic network will perform well even if it has limited ability to model subpixel mixing. It is likely that the performance difference between the fully fuzzy neural networks and the other algorithms investigated thus far would increase when applied to classes that exhibit a broader distribution of mixing than the cereal class.

Technique	Data Set	Error Function	Structure	Validation Set Error
MLP	Tall herb	Sum of squares	6-2-1	0.02103
"	"	"	6-5-1	0.02046
"	"	"	6-10-1	0.02016
"	"	Cross entropy	6-2-1	0.1052
"	"	"	6-5-1	0.09615
"	"	"	6-10-1	0.09393

Table 10: Comparison of MLPs trained with sum of squares and cross entropy functions on the tall herb data.

Figures 46 to 48 and figures 49 to 51 show the results of applying MLPs with different numbers of hidden neurons to the cereal data set using first the sum of squares error and then the cross entropy error. Between the sum of squares and cross entropy MLPs with two hidden neurons, there are noticeable differences in the estimated proportions, such as the greater homogeneity of the upper field in the fourth subimage predicted by the cross entropy network at the expense of making greater errors in predicting the presence of a field of cereal in the lower left hand corner of the third subimage. These differences

are the result of interactions between the model parameterisations and the characteristics of the different error functions as was explained in section 6.1.

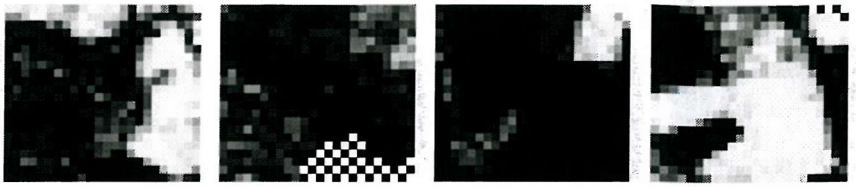

Figure 46: Cereal proportions predicted by a 6-2-1 MLP trained using the sum of squares error on fully fuzzy data.

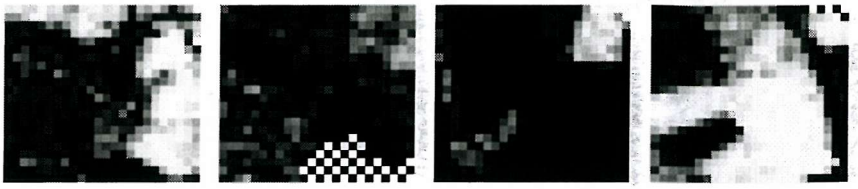

Figure 47: Cereal proportions predicted by a 6-5-1 MLP trained using the sum of squares error on fully fuzzy data.

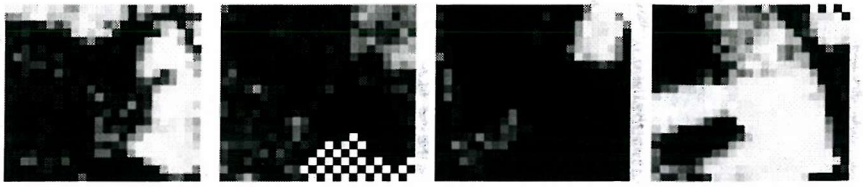

Figure 48: Cereal proportions predicted by a 6-10-1 MLP trained using the sum of squares error on fully fuzzy data.

Note that the proportions predicted by the more flexible networks – those with ten hidden neurons – are much less dependent on the choice of error function. This is because for any pixel the proportion estimates that minimise the expected sum of squares and cross entropy errors are the same (being equal to the mean proportions that would be observed for that spectral signature), suggesting that an area proportion estimator should produce the same estimate regardless of whether it is trained using the sum of squares or cross entropy functions. In practice, any particular model will be too constrained to reproduce the actual relationship between spectral signature and optimal subpixel proportion estimate and will therefore make errors in its proportion predictions that result from its parameterisation. It is the distribution of these errors between pixels that is adjusted by the choice of error function and leads to differences in behaviour. As more flexibility (in this case more hidden neurons) is added to the network, the magnitude of the error caused by the parameterisation decreases, reducing the impact of the choice of error function on the predictions made by the final model.

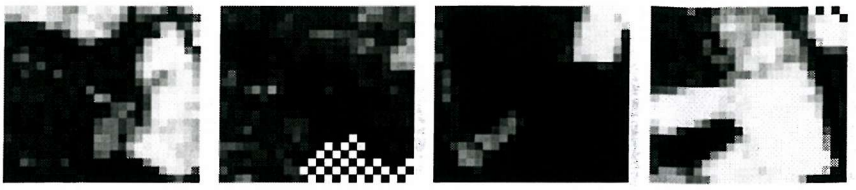

Figure 49: Cereal proportions predicted by a 6-2-2 MLP trained using the cross entropy error on fully fuzzy data.

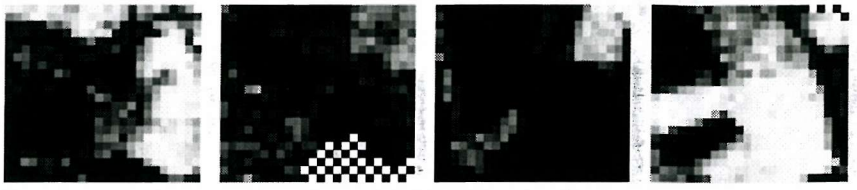

Figure 50: Cereal proportions predicted by a 6-5-2 MLP trained using the cross entropy error on fully fuzzy data.

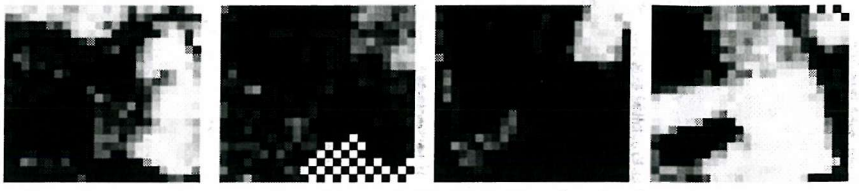

Figure 51: Cereal proportions predicted by a 6-10-2 MLP trained using the cross entropy error on fully fuzzy data.

A similar pattern of convergence of the predictions from sum of squares and cross entropy trained networks with increasing network flexibility can be seen for the tall herb data: the patterns of prediction and mis-prediction are much more similar for the networks with ten hidden neurons than for those with only two. In both cases the networks with small numbers of hidden neurons appear to be unable to pick out the systematic variance in the tall herb data and hence do little more than model the distribution mean leading to the almost uniformly dark estimate images. As more flexibility is added to the networks by increasing the number of hidden neurons, the proportion estimates display greater variation though never as much as the ground truth data, suggesting that some systematic aspects remain unlearned. The actual predictions themselves show considerable confusion, particularly between field boundaries with and without tall herb. In the fourth subimage for example, both fuzzy classifiers incorrectly predict the presence of tall herb for virtually all the field boundaries.

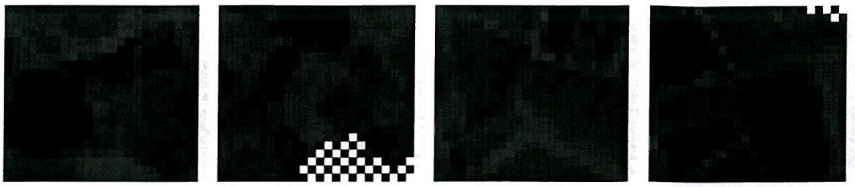

Figure 52: Tall herb proportions predicted by a 6-2-1 MLP trained using the sum of squares error on fully fuzzy data.

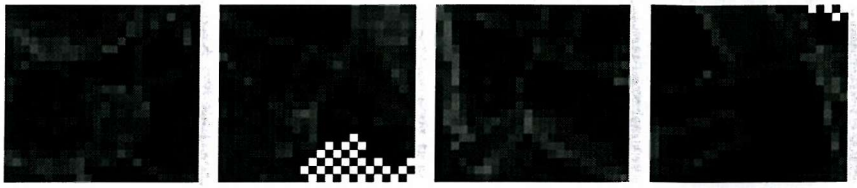

Figure 53: Tall herb proportions predicted by a 6-5-1 MLP trained using the sum of squares error on fully fuzzy data.

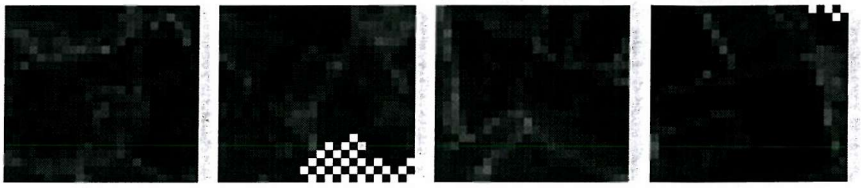

Figure 54: Tall herb proportions predicted by a 6-10-1 MLP trained using the sum or squares error on fully fuzzy data.

The following chapter presents a discussion of the factors that limit the performance of fuzzy classification algorithms. In particular, it shows that some of these are unavoidable when fuzzy classifications are based on pixel spectral signatures alone. Rather than proposing new fuzzy classification algorithms, it is suggested that a new representation for fuzzy classifications should be used – the distribution of probable fuzzy classifications of a pixel given its spectral signature.

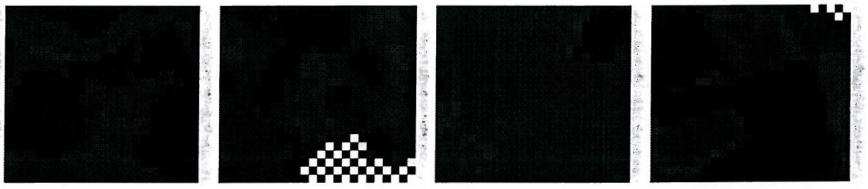

Figure 55: Tall herb proportions predicted by a 6-2-2 MLP trained using the cross entropy function on fully fuzzy data.

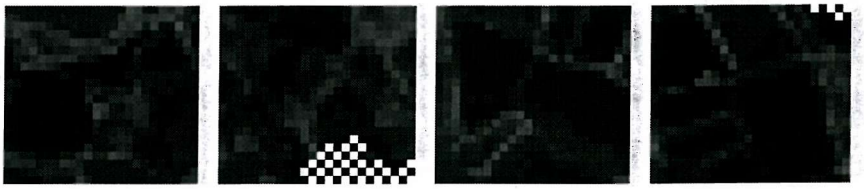

Figure 56: Tall herb proportions predicted by a 6-5-2 MLP trained using the cross entropy function on fully fuzzy data.

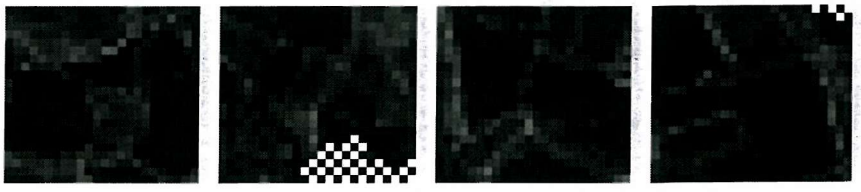

Figure 57: Tall herb proportions predicted by an 6-10-2 MLP trained using the cross entropy function on fully fuzzy data.

8. Performance Limits

The fuzzy classification literature contains many papers that report results of fuzzy classification experiments, but there is much less discussion of the limits of achievable fuzzy classification accuracy. It should be a priority to consider in detail what limits there are that are intrinsic to the fuzzy classification problem, (in the sense that they cannot be overcome by using more sophisticated fuzzy classification algorithms), since further experimentation is only justified if these limits have not already been reached [Wilkinson:96]. The following sections discuss characteristics of the fuzzy classification problem that limit fuzzy classifier performance. The discussion is not intended to be exhaustive, but rather focuses on issues that have received little attention in the fuzzy classification literature.

8.1. The Effect of the Number of Classes

It is well known in the linear mixture modelling community that difficulties arise in determining subpixel mixtures when the number of target classes exceeds the number of spectral bands [Kent:88][Sohn:97][Bosdogianni:97]. This problem is usually attributed to the difficulty of inverting the matrix of end member spectra resulting from its singularity, and may be overcome by using a regularised estimator, or mapping the original spectral measurements into some high dimensional space [Bosdogianni:97]. These solutions only eliminate one of the symptoms of a more serious problem: the multiplicity of solutions to the linear mixture model indicates that the spectral data contains insufficient information to uniquely identify the subpixel mixture responsible for the observed spectrum. If the spectral data is insufficient, an upper bound will be placed on the level of performance a fuzzy classifier can achieve, which can only be breached by providing the classifier with new information such as contextual information from surrounding pixels.

Unfortunately, it is relatively difficult to determine whether a particular set of spectra is sufficient with respect to a particular set of target classes and, if insufficient, the extent of the performance bound. The mixing which is likely to be observed in any particular application for any set of target classes forms some volume in the space of mixture proportions. If, in any region of mixture space, the volume is locally of a higher dimension than there are spectral bands, it will not always be possible to unmix the target classes with complete accuracy. In practice, it may be difficult to assess the

likelihood of this condition arising for a particular set of classes, since most classes cannot be guaranteed to interact with all others in the set. One notable example is in the case of mapping land cover on a global scale. In such a situation it is reasonable to expect that land cover types of, for example, ice and tropical forest would not be found within the same pixel. Any set of classes that includes these therefore has at least one degree of freedom less than the number of classes.

Since such detailed information may generally be lacking, it would be desirable to use an automated procedure to compute the local dimensionality of the volume of mixtures in a set of exemplars. Unfortunately, no algorithm for doing this is currently in common use and even if it were, the ultimate aim of analyses with regard to the number of target classes and their properties would be to compute the performance bound itself. In reality, this is likely to be extremely difficult, since the bound is not only dependent upon the local dimensionality of the volume of mixtures, but also on the unknown underlying distribution of the mixtures, both of which would have to be estimated from exemplars resulting in significant uncertainty in the error bound.

8.2. The Effect of Spectral Variation

Virtually all natural and most man-made cover types will exhibit some degree of spectral variation. Sources of such variation are many and varied: natural cover types may vary spectrally with age, or season. Oil seed rape provides a rather dramatic example of lifetime variation, changing from green to bright yellow over a relatively short period. More complex classes that are actually composed of a large number of simpler cover types, such as the 'urban' class will exhibit significant spectral variation due to the changing subpixel proportions of the class components [Thomas:96].

Such spectral variation dramatically increases the difficulty in collecting a set of exemplar pixels which are representative of both the types of mixtures seen in the landscape that the fuzzy classifier will be applied to in application, but also representative of the spectral properties of the target classes at the level of mixing expected. This type of problem – collecting exemplars that are representative of the application area – is ever present in statistical modelling, but is particularly severe in remote sensing image classification due to the degree of spatial non-stationarity that is present in such data. That is, the land cover statistics (as regards the relative frequency and spectral properties of cover types) vary spatially, limiting the accuracy of any

classifier that makes classifications on the basis of statistics learnt from a set of exemplars drawn from any particular location.

An additional and more critical problem with classes that exhibit spectral variation across all spectral bands is that it becomes impossible to infer subpixel cover proportions from pixel spectra without some degree of ambiguity and hence performance loss [Horwitz:71]. Including a class that exhibits spectral variation is, as far as area proportion estimation is concerned, equivalent to attempting to estimate the subpixel proportions of infinitely many classes, which, as was discussed in the previous section, cannot be done with complete accuracy. The ultimate goal of any investigation into this problem should be to quantify this component of the performance bound – a task likely to prove impossible since the bound depends not only on the distribution of the spectral variation, but also on how the spectral variation changes when classes mix, and on the mixtures likely to be present in any particular application.

8.3. Primitives and Compounds

In this section, a new phraseology is developed and used to discuss the conditions necessary to maximise the performance bound resulting from the sources of uncertainty discussed in the previous sections. The first term to be introduced, primitives, refers to the simple cover types from which more complex ones are composed. Typically, for example, the 'urban' class would be composed of much simpler classes such as 'slate' and 'tarmac' which would be classed as primitives. The 'urban' class itself would be described as a compound class, since it is composed of a number of simpler primitives. It is also useful to distinguish different types of primitives by their relation to each other and to compound classes: Shared primitives are used in the definition of two or more compound classes, whereas unshared primitives are used in the definition of only a single compound class. The relationship between two primitives is intercompound if the primitives appear only in the definition of different compound classes and intracompound if they appear in the definition of the same compound class. The conditions necessary for maximum fuzzy classification accuracy may thus be summarised as:

- Intracompound primitives should not intersect under the set of measurements to be used for area proportion estimation. i.e. for any two primitives P_a and P_b, $a \neq b$ and a set of compound classes $C_n : 1 \leq n \leq N$, if $\mu(P_a \cap C_n)>0 \wedge \mu(P_b \cap C_n)>0 \; \exists n : 1 \leq n \leq N$ then $\mu(P_a \cap P_b) = 0$.

- Intercompound primitives should not intersect under the set of measurements to be used for area proportion estimation. i.e. for any two primitives P_a and P_b, $a \neq b$ and any two compound classes C_n, C_m: $1 \leq n,m \leq N$, if $\mu(P_a \cap C_n)>0 \wedge \mu(P_b \cap C_m)>0 \; \exists n,m : 1 \leq n, m \leq N, n \neq m$ then $\mu(P_a \cap P_b) = 0$.

- Primitives should exhibit no spectral variation. I.e. the probability of observing a spectral signature s for a primitive P is given by $p(s \mid P)=\delta(s-s_P)$ where s_P is the spectral response of the primitive and $\delta(.)$ is a function which returns one when its argument is zero and returns zero at all other times.

- The number of degrees of freedom in the primitives must be equal to or less than the number of degrees of freedom in the spectral bands. This is usually the case if the total number of primitives is less than or equal to the number of spectral bands.

- Compound classes should be composed of an additive union of primitive classes. That is, for a compound C, and intracompound primitives $P_n : 1 \leq n \leq N$, $\mu(C)=\Sigma_n \mu(P_n)$.

These constraints are so strict that they are unlikely ever to be satisfied in any practical application. The value of these rules is rather that they highlight the ways in which sets of classes in particular applications deviate from the ideal and hence provide some indication as to specific sources of performance loss. The following section presents a novel examination of the impact of the sensor point spread function on the accuracy of fuzzy classifications and shows that not only does it limit performance by introducing ambiguity but also that the degree of ambiguity depends on the degree of subpixel mixing.

8.4. The Effect of the Point Spread Function

When a satellite captures a remotely sensed image, each pixel represents the spectral properties of the ground cover within and around the pixel convolved with the spatial sensitivity profile of the sensor, the point spread function (PSF). The point spread function has a 2-dimensional argument corresponding to the distance from the peak sensitivity of the sensor which occurs roughly at the pixel centre. From the pixel centre, the sensitivity of the sensor falls off monotonically with roughly circular symmetry and continues into the ground area covered by surrounding pixels. Generally, the point spread function has a number of approximations, including the Gaussian and the product of sines, depending on the degree of accuracy required (see, for example, [Justice:89]). Although the effect of neighbouring pixels is not considered here, their contribution to the proportion estimation error could be reduced by a procedure similar to that described in [Townshend:00].

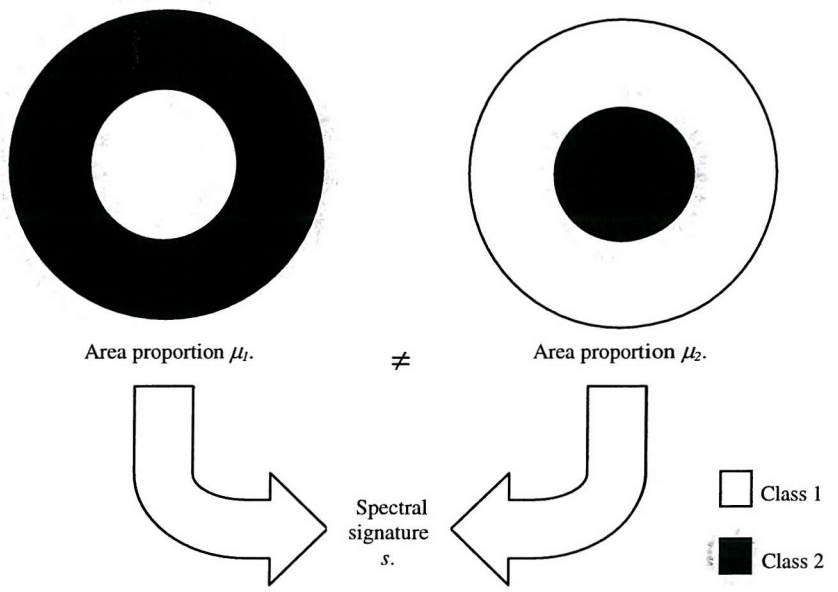

Figure 58: Two pixels with the same spectral signature but different sub-pixel composition.

Spectrally, the point spread function magnifies the contribution of land cover placed centrally in a pixel to the pixel spectrum and diminishes the contribution from land cover towards the pixel perimeter. The point spread function therefore introduces ambiguity into the spectral unmixing exercise by allowing different subpixel proportions to generate the same spectral signature [Manslow:00b]. Consider, for

example, a pixel that contains equal proportions of two land cover types, the first placed centrally, and the second arranged around the pixel perimeter. Spectrally, the pixel will appear to be most similar to the first, more centrally placed, cover type. If the locations of the cover types are now swapped such that the first cover type lies on the pixel perimeter, the spectral properties of the pixel will be most similar to those of the second class. This spectral change in the pixel may be counteracted however, by increasing the proportion of subpixel cover of the first class. Thus, as shown in figure 58, a pixel of the same spectral signature as the original one may be produced, but with different subpixel proportions.

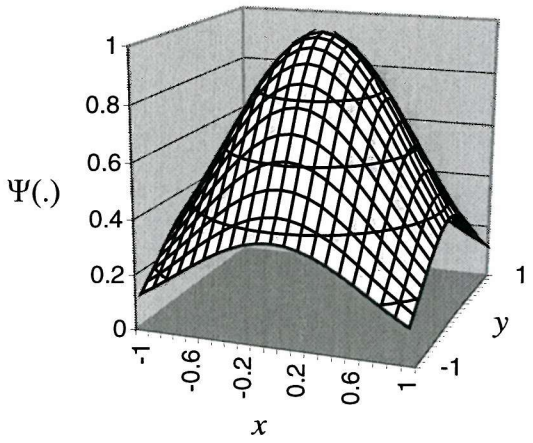

Figure 59: Gaussian PSF with α=1.

Critically, the magnitude of this effect depends on the proportions of subpixel cover. If a pixel consists of a single subpixel cover type, all such pixels would (assuming the class has no inherent spectral variation) exhibit the same spectral properties. If a pixel is mixed however, the subpixel cover can always be re-arranged to produce alternative pixels with different proportions of subpixel cover, but identical spectral signatures. To examine this phenomenon in greater detail, consider the case when a pixel consists of only two subpixel cover types covering area a_1 and a_2 respectively. We shall assume that the classes have spectral responses s_1 and s_2, which exhibit no spectral variation, and that the point spread function $\psi(.)$ is assumed to be Gaussian as given below,

$$\Psi(r) = \exp\left(-\alpha r^2\right) \qquad 60$$

where r is the distance from the pixel centre and α is a constant which controls the shape of the PSF. The pixel is assumed to have a circular footprint, that is, a point (x,y) is considered to be within the pixel area if:

$$r \leq 1 \quad | \quad r = \sqrt{x^2 + y^2} \tag{61}$$

If class 1 is concentrated in the area where the PSF is least sensitive, that is, around the pixel perimeter, all points for which

$$r(x, y) \geq \sqrt{1 - a_1 / \pi} \tag{62}$$

belong to class 1 and all others belong to class 2. The spectral response of the pixel may be written as:

$$S_A = \int_r^1 2\pi r \Psi(r) s_1 dr + \int_0^r 2\pi r \Psi(r) s_2 dr \tag{63}$$

where

$$r = \sqrt{1 - a_1 / \pi} \tag{64}$$

Now, if the subpixel cover is re-arranged so that class 1 now lies in the region of greatest sensitivity of the PSF, in the pixel centre, the spectral response of the pixel will change. It is possible, however, to restore the pixel's spectral response to its original value by adjusting the subpixel proportions of the two classes. The spectral response of a new pixel where class 1 is arranged in the central region is given by

$$S_B = \int_0^{r'} 2\pi r \Psi(r) s_1 dr + \int_{r'}^1 2\pi r \Psi(r) s_2 dr \tag{65}$$

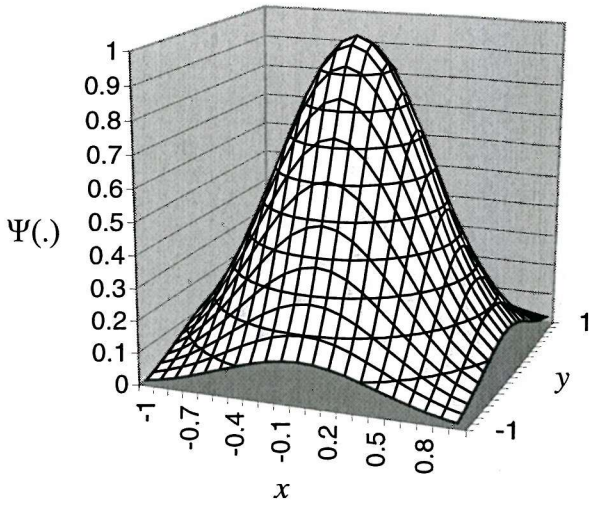

Figure 60: Gaussian PSF with α=2.

where

$$r' = \sqrt{a_1'/\pi} \qquad 66$$

and a_1' is the new area occupied by class 1. S_A and S_B are equal if each of the component integrals are equal, i.e. when

$$\int_0^{r'} 2\pi r \Psi(r) s_1 dr = \int_r^1 2\pi r \Psi(r) s_1 dr \qquad 67$$

Note that only one pair of integrals needs to be considered, since the pairs are equivalent. Substituting for the Gaussian PSF and cancelling multiplicative constants yields:

$$\int_0^{r'} r e^{-\alpha r^2} dr = \int_r^1 r e^{-\alpha r^2} dr \qquad 68$$

which gives (see appendix B):

$$-\frac{1}{2\alpha}\left(1 - e^{-\alpha r'^2}\right) = -\frac{1}{2\alpha}\left(e^{-\alpha r^2} - e^{-\alpha}\right) \qquad 69$$

Rearranging this makes it possible to compute the alternative subpixel proportions of a pixel with the same spectral response as the original under the assumption that in that pixel, class 1 was concentrated in the region where the PSF is least sensitive. Using

$$a_1' = \pi r'^2 \tag{70}$$

the alternative area can be shown to be:

$$a_1' = -\frac{\pi}{\alpha}\ln\left|e^{-\alpha} - e^{-\alpha r'^2} + 1\right| \tag{71}$$

Assuming class 1 was originally concentrated in the region where the PSF is most sensitive, gives an alternative subpixel area of

$$a_1 = \pi\left(1 + \frac{1}{\alpha}\ln\left|e^{-\alpha} - e^{-\alpha r'^2} + 1\right|\right) \tag{72}$$

obtained by using

$$a_1 = \pi r_{\max} - \pi r^2 \tag{73}$$

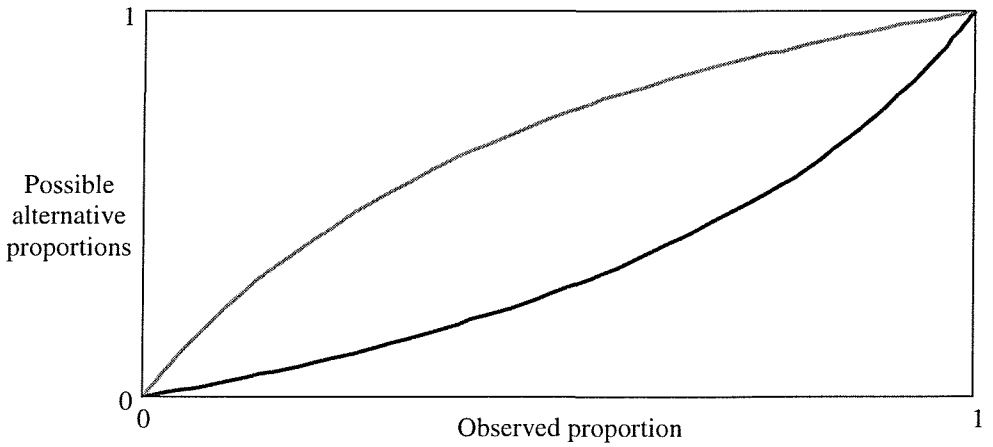

Figure 61: Ambiguity induced by a PSF with α=1.

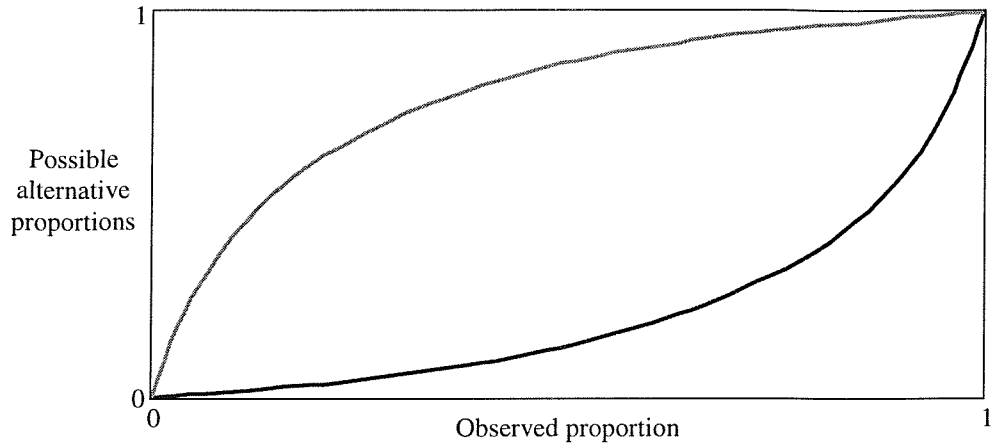

Figure 62: Ambiguity induced by a PSF with α=2.

The results derived above indicate that for the Gaussian model of the PSF, there is much higher uncertainty involved in predicting the subpixel proportions for pixels that are heavily mixed (i.e. the subpixel proportions are similar) than for ones that are more lightly mixed (the subpixel area is strongly dominated by a single class), and that the greater the range of PSF sensitivity within a pixel, the greater the induced uncertainty. The former effect suggests that estimators should model exemplars more closely when they represent pure pixels than when they represent mixed pixels since there exist no alternative pixels with different subpixel proportions *at that point* in spectral space. Learning lightly and heavily mixed exemplars to the same level of accuracy would therefore risk under fitting the nearly pure exemplars and over fitting the heavily mixed ones.

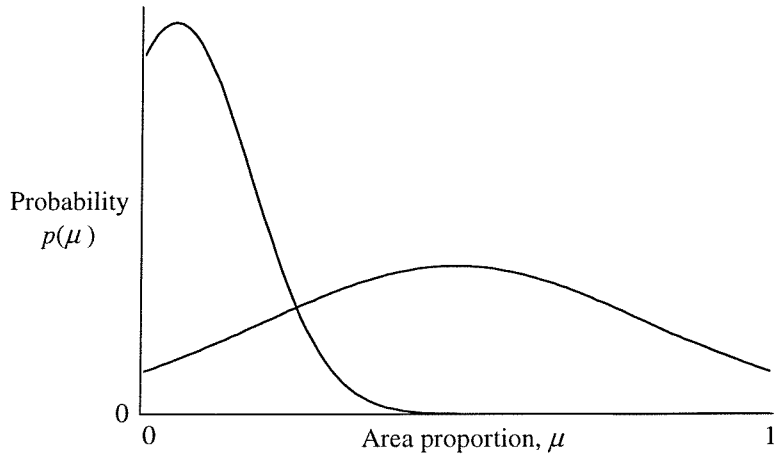

Figure 63: Small and large variance proportion distributions indicating the way in which the PSF introduces more uncertainty in mixed pixels than in almost pure pixels.

To illustrate this effect, consider a point in spectral space where a nearly pure and a heavily mixed pixel both occur. By the preceding argument, each pixel suggests some distribution of alternative pixels with different subpixel cover proportions at the same point in spectral space. The nearly pure pixel would be associated with a tight (small variance) distribution of alternative proportions, while the mixed pixel would be associated with a much broader (larger variance) distribution as shown in figure 63. Note that for illustrative purposes, Gaussian distributions have been plotted even though the actual distributions are unknown. The following section considers the implications of the effects outlined above for area proportion estimators and, in particular suggests that it should be possible to use the information derived above to obtain optimal area proportion estimators.

8.4.1. Implications for Proportion Estimation

The following analysis of the impact of the results outlined in the previous section shows, without the aid of distributional assumptions, that proportion estimators are likely to achieve greater accuracy if they lay more emphasis on nearly pure exemplar pixels than heavily mixed ones in the production of estimates. Consider the case where, at a particular point in spectral space, two pixels have been observed, one nearly pure pixel with proportion μ_1 and the other heavily mixed with proportion μ_2 and any other area proportion by μ. The proportion μ may also be written as $\mu_1+\varepsilon_1$ and $\mu_2+\varepsilon_2$ where ε_1 and ε_2 are deviations from the observed subpixel proportions. Using this notation, the expected squared error of an area proportion estimate μ_{est} given the two observed pixels can be written as:

$$E = \int (\mu_{est} - \mu)^2 \, p(\mu) d\mu \qquad 74$$

where $p(\mu)=0.5\times p(\varepsilon_1)+0.5\times p(\varepsilon_2)$, assuming that the nearly pure and heavily mixed pixels are equally likely. This expected error measure is minimised when the area proportion estimate is the mean of the means of the distributions associated with the nearly pure and heavily mixed pixels:

$$\mu_{opt} = \tfrac{1}{2}\mu_1 + \tfrac{1}{2}\mu_2 \qquad 75$$

where the above notation assumes that the observed area proportions lie at the means of the distributions. The global mean μ_{opt} will be used as the definition of the optimal area

proportion estimate in the discussion that follows, since it minimises the expected sum-of-squares and cross-entropy errors. Now consider a simple estimator μ_{wc} derived from a weighted combination of the observed area proportions in the following way:

$$\mu_{wc} = m\mu_1 + (1-m)\mu_2 \qquad 76$$

where $0 \leq m \leq 1$ is a weighting factor. It is possible to write down the expected squared error of this estimator as follows:

$$E = \iint \left(\mu_{wc} - \mu_{opt}\right)^2 p(\varepsilon_1)p(\varepsilon_2)d\varepsilon_1 d\varepsilon_2 \qquad 77$$

which, may be rearranged (see [Bishop:95] for this derivation in more conventional notation):

$$E = \left(m\mu_1 + (1-m)\mu_2 - \mu_{opt}\right)^2 + m^2\sigma_1^2 + (1-m)^2\sigma_2^2 \qquad 78$$

where σ_1 and σ_2 are the standard deviations of the distributions associated with the nearly pure and heavily mixed pixels respectively. In this form, the expected squared error of the estimator (its expected generalisation performance) can clearly be seen to be a weighted combination of the variances of the alternative area proportion distributions associated with each of the observed pixels (the second and third terms) along with an additional term dependent upon difference between the actual estimator based on the available observations, and the true optimum. Setting the derivative of the above equation with respect to the weighting factor m to zero, makes it possible to solve directly for the value of m which minimises the expected generalisation error. The optimal weighting may thus be shown to be:

$$m = \frac{\sigma_2^2 + \frac{1}{2}(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \qquad 79$$

which is plotted in figure 64 for different values of σ_1 and σ_2. It can clearly be seen that when σ_1 is small and σ_2 is large, m is close to unity and the area proportion estimator virtually ignores the data point μ_2. If, on the other hand, σ_1 is large and σ_2 is small, μ_1 is largely ignored. When the variances are similar, however, the optimal estimator combines the data points with roughly equal measure. This seems to imply that subpixel area proportion estimators that allocate computational resources more strongly towards

less mixed pixels would, on average, achieve higher performance than those that distribute the resources evenly.

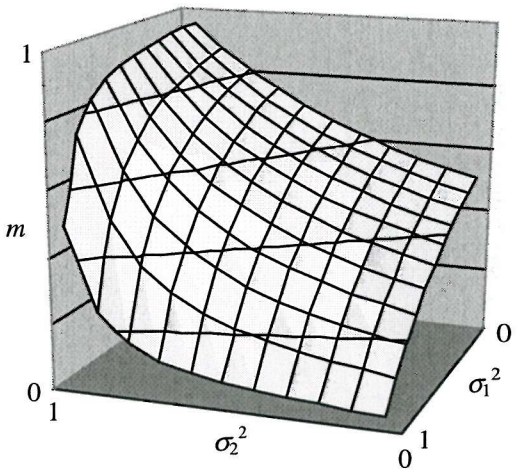

Figure 64: m as a function of σ_1 and σ_2 with $(\mu_1-\mu_2)^2=0.1$.

In practice, this effect may be difficult to demonstrate for a number of reasons. Firstly, the above discussion considered an infinitely flexible estimator operating only at a single point in spectral space. Any model that is likely to be useful in practice will generally have relatively low effective flexibility due both to the application designer's deliberate decision to limit its complexity in order to control its capacity to over fit, but also due to the relatively large size of remotely sensed data sets. This limited effective complexity will tend to prevent the variance implicit in the observed area proportions of heavily mixed pixels resulting in variance in the area proportion estimates made by the model. The equation for computing the optimal weighting factor is dependent upon the variances of the subpixel proportion distributions associated with each of the observed pixels. Unfortunately, these values are unknown, though it is certain that the variances increase with the degree of subpixel mixing. It is not therefore possible to calculate the optimal weighting factor directly, although it can be stated that the weighting factor should be lower for pixels which are highly mixed.

Finally, there are many sources of uncertainty in remotely sensed data sets that would mask any benefit of explicitly considering the effect of the PSF. One important aim of future work would be to test the hypothesis that there may be performance benefits in weighting pixels in producing models and, if no benefit is observed, to try to determine exactly why. It should be noted that the above derivation could also be used to derive an upper bound for the uncertainty introduced by the sensor PSF. It is anticipated that this bound would be of limited practical use due to the limitations of the Gaussian model of the PSF, and should be used for guidance only. The variance of the distribution of

alternative proportions induced by the PSF is maximised if all proportions lie along the bounds of the distribution as derived earlier, and if proportions are equally likely to lie on each bound. This idea could, in principle, be used to derive an upper bound on the amount of uncertainty induced by the PSF, but the analysis is not carried out here, since it is an approximation of the least useful (the upper) bound. The following section looks at the effect of the point spread function from another perspective – that is, instead of trying to improve single proportion estimates it examines the possibility of modelling the ambiguity introduced by the PSF (and all other sources) in the proportion information in a pixel's spectral signature.

8.4.2. Ill-Posedness and the Representation of Proportion Estimates

A problem is said to be ill-posed if it fails to satisfy one or more of the following conditions [Kirsch:96]:

- there exists a solution to the problem (existence),
- there is at most one solution (uniqueness), and
- the solution depends continuously on the data (stability).

From the preceding discussion, it is clear that the problem of inferring sub-pixel proportions from a pixel's spectral signature is ill-posed, since it violates the uniqueness condition. That is, for a pixel of spectral signature s with known subpixel proportion $\mu \neq 0$ or 1, it will be known that a range of alternative subpixel proportions exist that could also have generated the observed spectral signature. Existing algorithms for estimating subpixel proportions from spectral signatures implicitly choose to report one of the range of possible alternative proportions and generally provide no information as to the distribution of the alternatives. This can often be misleading, since the reported proportion may be one that could not occur in practice, or could not generate the spectral signature that was actually observed.

The following section discusses a new technique for characterising sub-pixel cover – that of modelling the probability distribution of possible proportions given the observed spectral signature. The conceptual basis of this new approach is to use a much more flexible representation for the information derived from the remotely sensed data and in

particular, one that comes as close as possible to satisfying the principles of minimum and maximum uncertainty proposed in [Klir:95]:

- the estimates should contain minimum uncertainty
- the estimates should fully represent the uncertainty that remains.

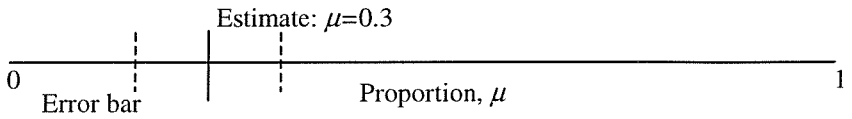

Figure 65: Representing area proportion information by a single estimate and error bars.

The conditional distribution representation cannot easily be compared to the standard single estimate in terms of the first principle, since it uses a different representation for its predictions. In general, the complexity of the conditional distribution representation makes it likely that, on average, it will contain more uncertainty than the single proportion estimate since the conditional distribution models are required to extract more information from the same number of exemplars. This point is returned to in the next section, where the conditional distributions are collapsed into single estimates to permit comparison of the performance of the standard single proportion estimation models and conditional distribution models.

It terms of the second principle, the conditional distribution representation is clearly superior to the standard single proportion estimate. As shown in figure 65, standard algorithms typically provide either no information on the uncertainty in their predictions, or summarise the information using error bars or confidence intervals. The type of uncertainty summarised by the error bars often only represents the uncertainty in the model predictions induced by uncertainty in the model parameters that remains after the model has been trained, and neglects the uncertainty due to the ill-posed nature of the area proportion estimation problem. The conditional distribution representation, shown in figure 66 on the other hand is powerful enough not only to represent uncertainty arising from the ill-posedness of the area proportion estimation problem, but also the uncertainty resulting from the modelling process itself, even though this is not supported by the model used in the next section. Note that the preceding and succeeding arguments apply to all information derived from remotely sensed data and not just the proportion information considered in detail here.

It should be emphasised that the work on the spectrum conditional area proportion distribution proposed here is fundamentally different from the distribution modelling described in [Shen:92], [Kitamoto:99] and [Erol:00], since [Shen:92] and [Erol:00] construct models of the distributions of class spectral signatures as a means to obtaining single proportion estimates (in the former case) or class labels (in the latter), rather than directly modelling the distributions of proportions themselves and [Kitamoto:99] uses fractal models of subpixel class distributions to derive prior area proportion distributions, ultimately proposing the Beta distribution also used in [Atkinson:99] and [Chittineni:81]. Indeed, the proportion distributions cannot be obtained from the class-conditional-spectral distribution without also knowing the distribution of proportions within pixels of a given classification and the relative frequencies of such classifications. Proportion distributions extracted in this way are likely to contain an excess of uncertainty since they have effectively been obtained after a conventional classification process.

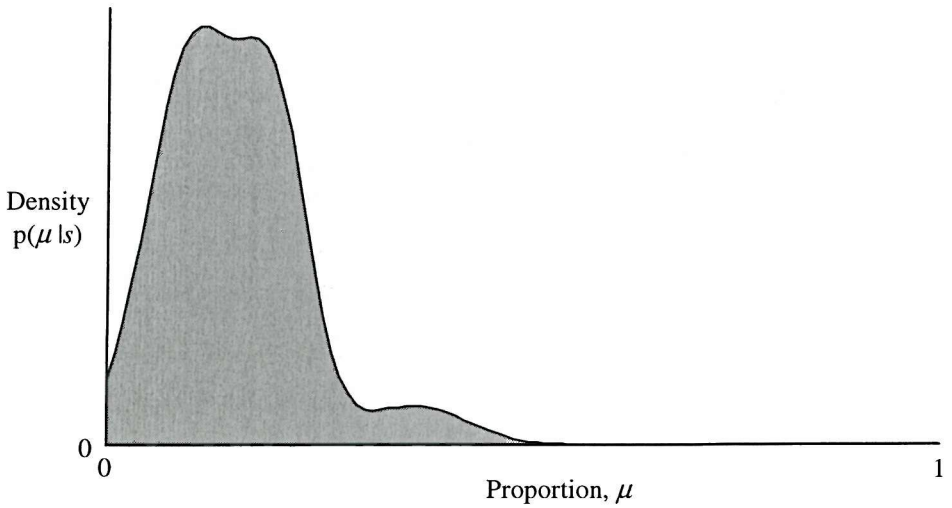

Figure 66: A hypothetical area proportion distribution showing its representative power.

9. Spectrum Conditional Probability Distributions as a Representation of Information Derived from RS Data

The preceding chapter opened with a discussion of some of the factors that limit fuzzy classifier performance. It presented a detailed analysis of the impact of the sensor PSF that provided compelling evidence that the problem of making inferences about subpixel processes on the basis of a pixel's spectrum is ill-posed. The section ended by proposing a new representation for information pertaining to subpixel processes that has been derived from pixel spectra. This new representation – the spectrum conditional proportion probability distribution – is desirable for essentially three reasons [Manslow:00b]:

Visualisation: Constructing a model of the spectrum conditional area proportion distribution permits complete representation and visualisation of all the area proportion information that can be derived from pixel spectral signatures. In contrast, traditional techniques that seek the single "optimal" area proportion estimate discard most of the information contained in a pixel's spectral signature, and provide only summary statistics of the spectrum conditional distributions, usually by estimating their means and sometimes also their variances. In many instances, these summary statistics are highly inadequate representations of the information in the spectrum conditional distributions and can often be misleading. Figure 67, for example, shows the spectrum conditional area proportion distribution for a real pixel in the FLIERS data set. The proportion being represented is that for the cereal crop class described previously and clearly suggests that pixels of the spectral signature given, are likely to consist purely of cereal crops, or to contain none at all. This sort of disjunction, that arises from multimodal spectrum conditional distributions, cannot easily be represented by a small set of summary statistics, and hence cannot be represented by conventional approaches to fuzzy classification.

One potential problem with visualising the spectrum conditional distributions is how to represent the distributions for an entire image. Fortunately, this problem may be overcome by the following considerations: In general, the standard single "optimal" proportion estimate – the distribution mean – will be a reasonable representation of the distribution for many pixels, making it unnecessary to visualise the full distributions for all pixels. Those pixels for which the single estimate is inadequate can easily be

identified, since the expected performance of a single proportion estimate is a monotonically increasing function of the variance of the distribution that it summarises. Thus, problem pixels may easily and automatically be identified by the large variance of their spectrum conditional distribution, and only for problem pixels does the spectrum conditional distribution need to be examined. This idea is shown to work well in practice for the FLIERS data set in a later section.

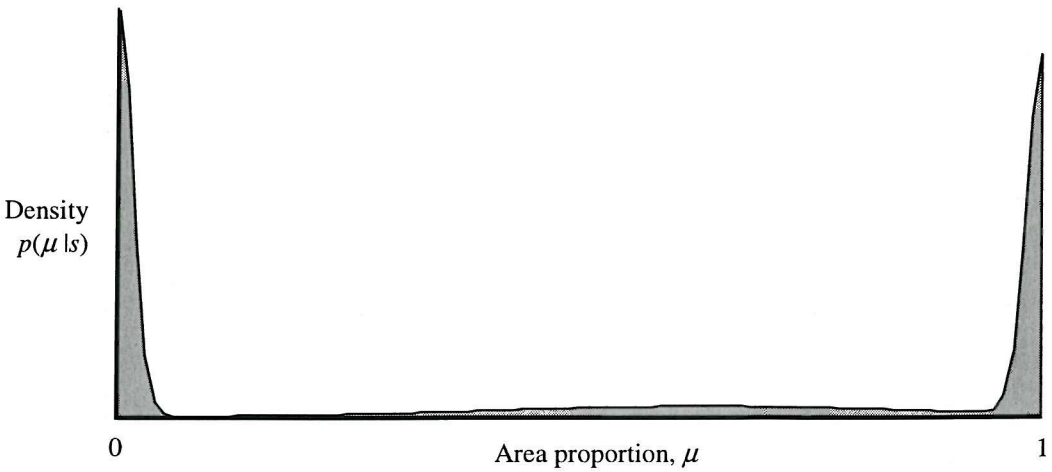

Figure 67: A cereal crop area proportion distribution for a real pixel in the FLIERS data set.

Combination: One of the ways of overcoming the ill-posedness of estimating subpixel proportions on the basis of pixel spectral signatures is to combine estimates made on this basis with information from other sources. Although many heuristics exist for doing this, there are essentially two rigorous probabilistic approaches, marginalisation and probabilistic inference using Bayes' theorem. Marginalisation is appropriate when each model is considered to be uncertain and the aim of combination is to average over the uncertainty in the choice of model. In this way, no model has a veto – if one or more of the models in the combination suggests that a particular range of proportions are possible, the combination also suggests they are possible. Given M models, H_m: $1 \leq m \leq M$, the marginalised proportion distribution is obtain from:

$$p(\mu \mid s, D) = \sum_{m=1}^{M} p(\mu \mid H_m, s)\, p(H_m \mid D) \qquad 80$$

where D is the set of exemplar pixels. $p(\mu \mid H_m, s)$ is the spectrum conditional distribution predicted by model H_m for spectrum s, and $p(H_m \mid D)$ is the probability that if the set of

exemplars D is considered to have been generated by one of the M models, it was generated by model H_m. This term can be difficult to evaluate in practice, since it is dependent on the complexity of H_m and is thus, for simplicity, often assumed to be the same for all H_m and ignored. Although marginalisation has here been presented in terms of spectrum conditional distributions, the proportion estimate given by the combination μ_{comb} is simply a weighted average of the estimates given by the M individual sources μ_m:

$$\mu_{comb} = \sum_{m=1}^{M} \mu_m \, p(H_m \mid D) \qquad 81$$

A typical application arises when several neural networks are trained on the same data set and then combined into a committee such that the output of the committee is a weighted sum of the outputs of the individual networks. In this case, a number of networks are producing different estimates of the same variable, and marginalisation is used to derive a single estimate that is only weakly dependent on any individual model.

The second probabilistic approach to combining information from different sources is to use Bayes' theorem. This method is appropriate when the distributions produced by the different models are to be considered to be accurate representations of information about the target variable conditioned on different information sources. In this case, any one model has the power of veto in that if it suggests that a particular range of proportions are impossible, the combination also suggests they are impossible. The Bayesian method of combination thus always retains a strong dependence on the behaviour of the component models regardless of the number of models present. As before, assuming there are M models, H_m, Bayes theorem can be written:

$$p(\mu \mid h_1 ... h_H) = \frac{1}{\prod_{h=1}^{H} p(y_h)} \, p(\mu) \prod_{h=1}^{H} p(y_h \mid \mu) \qquad 82$$

where y_h is the output of the h^{th} model, $p(\mu)$ is the unconditional probability of observing a proportion value μ and $p(h^m)$ is the unconditional probability of the m^{th} model outputting h_m. $p(h_m|\mu)$ is a conditional probability density that may be modelled using the technique described in the next section. For each model,

$$p(h \mid \mu) = \frac{p(\mu \mid h)p(h)}{p(\mu)} \qquad 83$$

such that:

$$p(\mu \mid h_1 ... h_M) = \frac{1}{\prod_{m=1}^{M} p(h_m)} p(\mu) \prod_{m=1}^{M} \frac{p(\mu \mid h_m)p(h_m)}{p(\mu)} \qquad 84$$

and cancelling gives:

$$p(\mu \mid h_1 ... h_M) = p(\mu)^{1-m} \prod_{m=1}^{M} p(\mu \mid h_m) \qquad 85$$

where, on the right hand side, $p(\mu \mid h_n)$ is a conditional area proportion distribution, which if derived from a pixel's spectral signature is the spectrum conditional area proportion distribution $p(\mu \mid s)$. This form of Bayes' theorem is, for the current application, more convenient than the original since it requires conditional density estimation to take place in only the one dimensional area proportion space rather than the six dimensional spectral space. This helps to minimise the effects of the curse of dimensionality [Bishop:95] and should, on average, produce more accurate results. In summary: Bayes' theorem should be used rather than marginalisation when a number of different information sources provide different information about subpixel proportions, and the use of Bayes' theorem to combine area proportion information derived from a pixel's spectral signature with information from other sources without making Gaussian approximations requires knowledge of the spectrum conditional distribution.

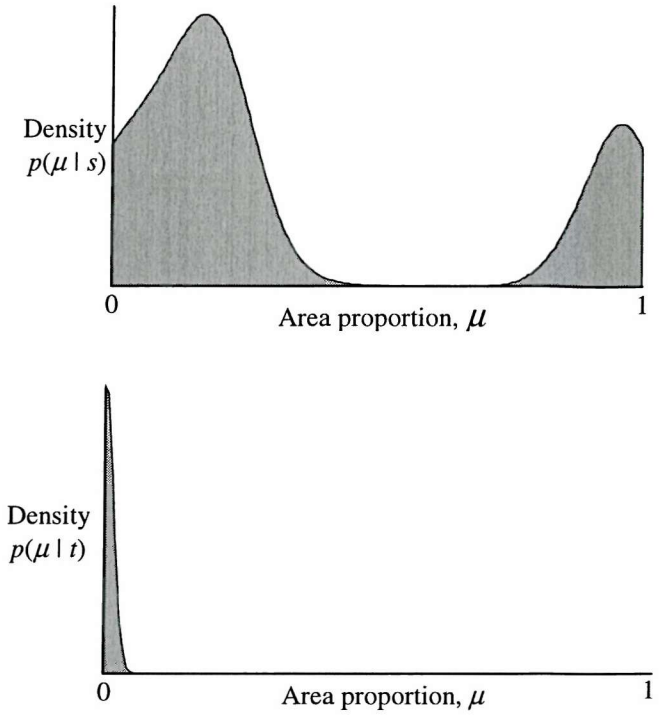

Figure 68: Proportion distributions for a pixel given the spectral signature of the pixel (top) and the texture of its neighbourhood (bottom).

As a more concrete example of how marginalisation and Bayesian combination work, consider an application which uses three models, one that estimates spectrum conditional area proportion distributions $p(\mu \mid s)$, one that extracts a metric t that characterises the texture of a region of an image and another that models the texture conditional area proportion distribution, $p(\mu \mid t)$. Assume that a new unclassified image is presented for classification and that, for a particular pixel, the spectrum conditional distribution is as shown in the top half of figure 68, and the texture conditional distribution is as shown in the bottom half. In this example, the target class could be cereal crops, and the texture metric could be distinguishing between urban and rural. This would mean that although the proportion of subpixel area covered by crops could not be decisively determined from the pixel's spectral signature (hence the high variance spectrum conditional distribution in figure 68), the pixel can be fairly confidently identified as belonging to an urban area (due to the distinctive texture of such areas that results from the high road density). The texture conditional density estimator learnt from the training set that when the texture classifier indicates that a pixel is in an urban area it is very unlikely to contain significant quantities of cereal crops and hence produces a distribution tightly centred around zero.

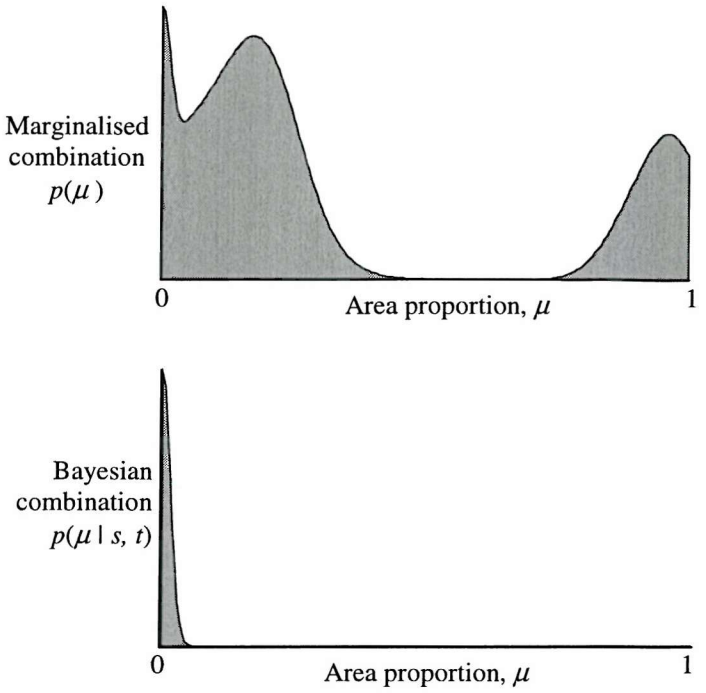

Figure 69: Proportion distributions obtained by combining texture information with direct proportion estimates using marginalisation (top) and Bayesian inference (bottom).

When these two distributions are combined using Bayes' theorem, the resulting distribution is as shown in the bottom half of figure 69 (where a uniform unconditional distribution has been assumed for the area proportions). Note that the distribution resulting from the combination of the spectrum and texture conditional distributions strongly suggests that the pixel contains either no cereals or only a very small amount. Thus, even though the pixel's spectrum contained too little information to determine the subpixel proportion with any certainty, the information provided by the texture analysis was sufficient to exclude the possibility of the pixel containing significant quantities of cereal. If these pieces of information had been combined by marginalisation, the resulting distribution would be that shown in the top half of figure 69. Clearly this method of combination is inappropriate in this instance, since even though the texture analysis effectively excludes the possibility of subpixel cereal crops, this information is lost when information from the texture analysis and the spectral analysis are combined by marginalisation.

Propagation: The problem of area proportion estimation is usually considered without reference to the final application of the proportion estimates. In practice, proportion estimates are often a means to an end rather than an end in themselves, being used as inputs to a variety of applications ranging from the monitoring land cover change, to estimating biomass. Ignoring the distribution of proportion estimates may be

problematic, especially when the application that uses them is non-linear and is expected to behave, in some sense, optimally. The reason for this is that, as already established, the proportion estimates are characterised by some distribution and so the optimal behaviour of the application that uses them is implicitly defined as optimal given their distribution. An application that uses only the single proportion estimates produced by conventional fuzzy classification algorithms will be unable to account for the distribution in the estimates, and hence will be unable to behave optimally given the uncertainty in the proportion estimates.

To illustrate the propagation of area proportion distributions through a target application, consider the apparently simple problem of estimating the percentage change in land cover from two proportion estimates for the same pixel taken at two different times. Assume that the spectral signature of the pixel remains unchanged between the two observations, such that the spectrum conditional area proportion distribution is the same for each observation, and is given in figure 70. Since the spectral signature of the pixel is unchanged between the two observations, a deterministic spectrum based fuzzy classifier that estimated a proportion of μ for the first pixel would also estimate μ for the second, which would result in an estimate of 0 percent change in subpixel cover between the observations.

Now consider percentage change computed from the spectrum conditional proportion distribution. In this case, no single percentage change estimate will result, since the proportion distribution suggests that there are a range of proportions that could have covered the pixel, there is a corresponding range of possible percentage changes, as shown in figure 71. Although the distribution agrees with the percentage change estimate that was produced by the application of the standard fuzzy classifier in the sense that it is most likely that subpixel cover did not change between the two observations, the percentage change distribution is the most complete representation of the percentage change that can be derived from the pixel spectral signatures. In particular, the distribution indicates that although it is most likely that there was no change in subpixel cover, it is also quite possible that the pixel went from 100 percent to 0 percent covered by the target class and, in fact, almost any percentage change could have taken place between the capture of the two images.

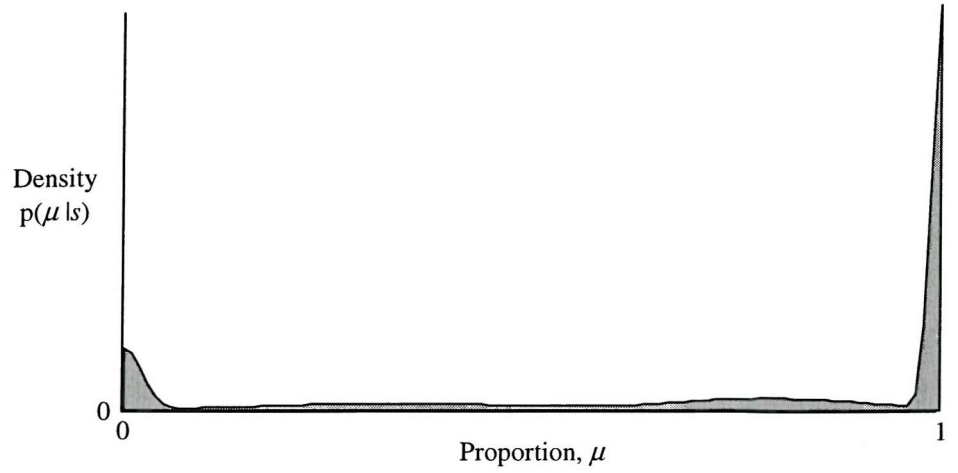

Figure 70: A typical area proportion distribution for the cereal class.

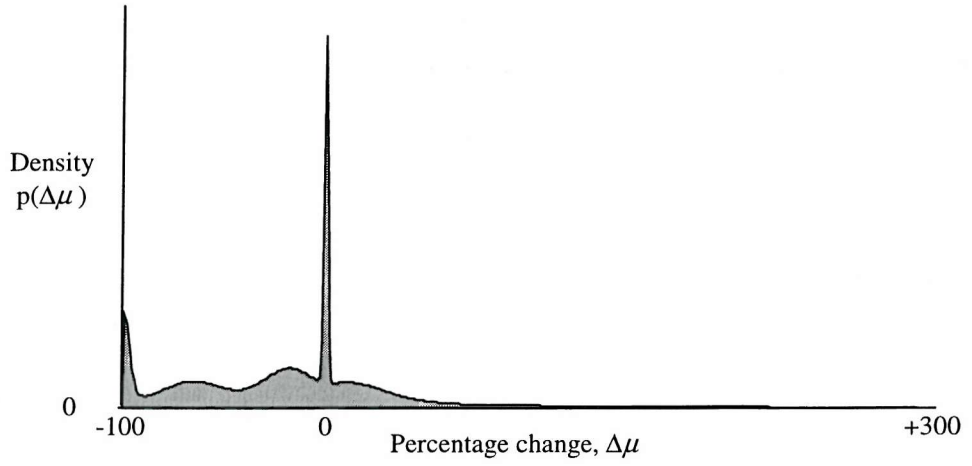

Figure 71: A percentage change distribution showing the range of possible changes in land cover.

9.1. Techniques for Modelling Spectrum Conditional Distributions

This chapter describes three algorithms of increasing complexity for extracting spectrum conditional area proportion distributions. The first is a conventional classifier that uses a stratification of area proportion space to convert the proportion distribution modelling problem into a simpler classification problem. No results are presented for the application of this technique since it is included only as a means of introducing the more complex techniques. The second technique improves on the first by increasing the flexibility of the distribution model and in so doing abandons the implicit static stratification used by the first model. Equations that allow efficient gradient based searches for the optimal model parameters are derived along with equations for summary statistics of the predicted distributions. The third and final algorithm considered is essentially the same as the second, but replaces the rectangular basis functions with Gaussians to yield the mixture density network described in [Bishop:94] and [Bishop:95].

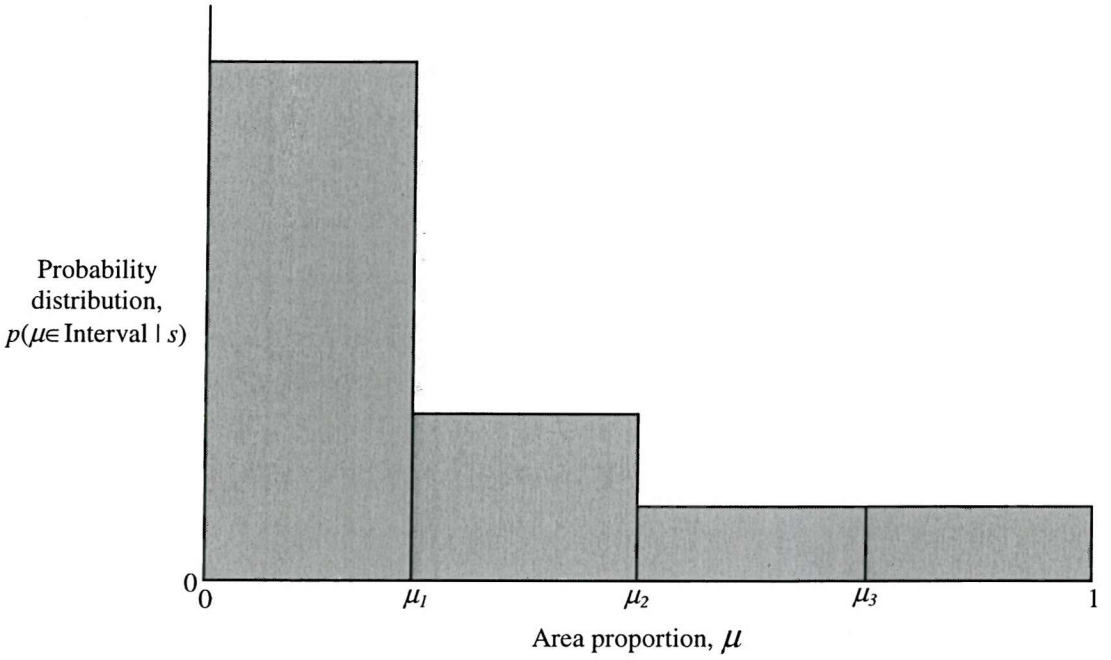

Figure 72: A representation of the spectrum conditional proportion distribution using stratification and classification.

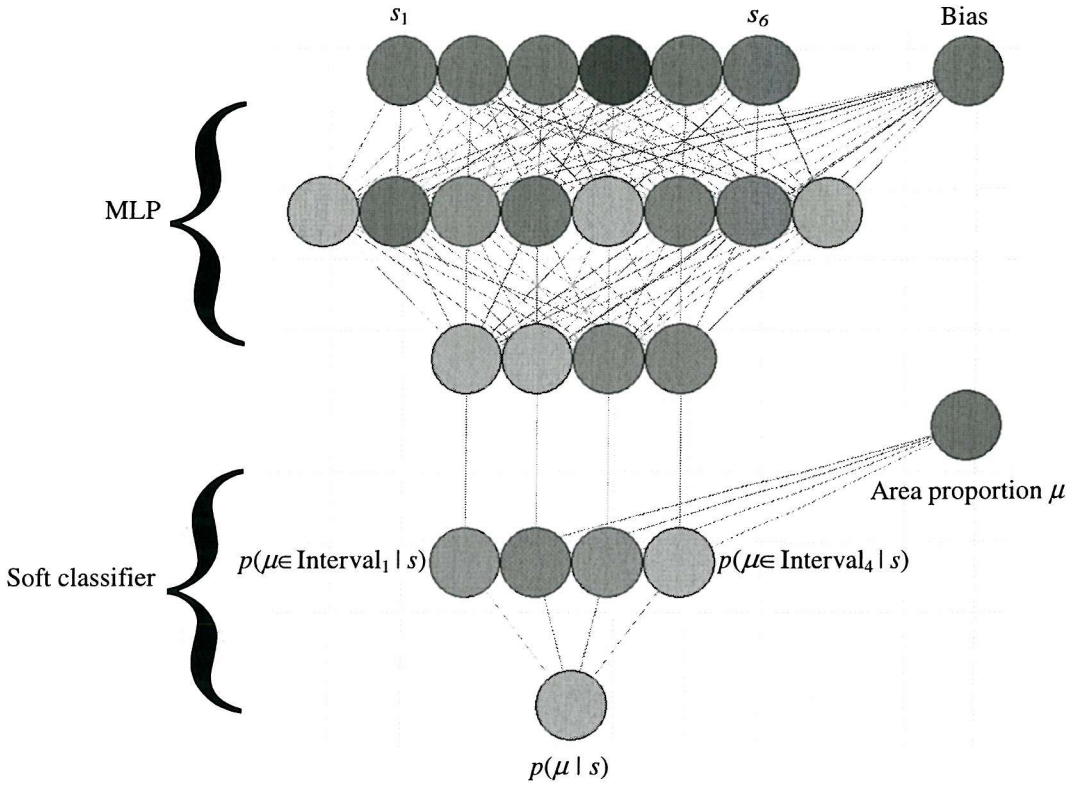

Figure 73: An MLP classifier with four outputs that can be used as a simple stratified area proportion distribution model.

9.2. Modelling Spectrum Conditional Distributions with a Stratified Classifier

The simplest approach to modelling spectrum conditional distribution information is by stratifying area proportion space and producing models that estimate the posterior probability that the true area proportion lies each stratified layer. The problem of modelling area proportion distributions is thus reduced to the more familiar one of classification. Figures 72 and 73 provide a simple example of how the technique would be applied in practice when modelling the distribution of the proportions of a single cover type. In this example, one soft classifier with four outputs is used with each output estimating the posterior probability that the true proportion μ_{true} lies within their area proportion interval. The area proportion intervals are assigned to each classifier by dividing the valid range of area proportions [0,1] into five intervals of equal size, so that the first classifier estimates the probability that the true proportion lies in the range [0.0, 0.2), the second in the range [0.2, 0.4), and the fifth in the range [0.8, 1.0].

Such classifiers can easily be produced by replacing the area proportion target information in each pattern in the set of exemplars with a five dimensional binary vector

that contains a one in the n^{th} position if the target proportion lies in the n^{th} interval and zeros in all others. The posterior probability estimating classifiers can then be produced by training a neural network with five outputs as though its was operating on a standard classification problem. One of the main limitations of this technique is that much of the structure of the model is determined apriori, but in a manner not justified by the available prior knowledge. That is, the equal spacing of the intervals in area proportion space limit the distribution of complexity in the representation used by the model to be uniform across proportion space. This is undesirable because the prior proportion distributions given at the beginning of this thesis indicate that most variation in the proportion distributions occur towards the extremes of μ=0 and μ=1, suggesting that models that can represent greater complexity in these regions will offer better performance. Unfortunately, it will, in general, be insufficient to distribute the complexity of the representation "manually" in accordance with the complexity of the prior, since the optimal distribution will be dependent on the particular spectral signature of the observed pixel. The following section describes a new technique that overcomes this difficulty by allowing the model to learn the distribution of complexity from the exemplars.

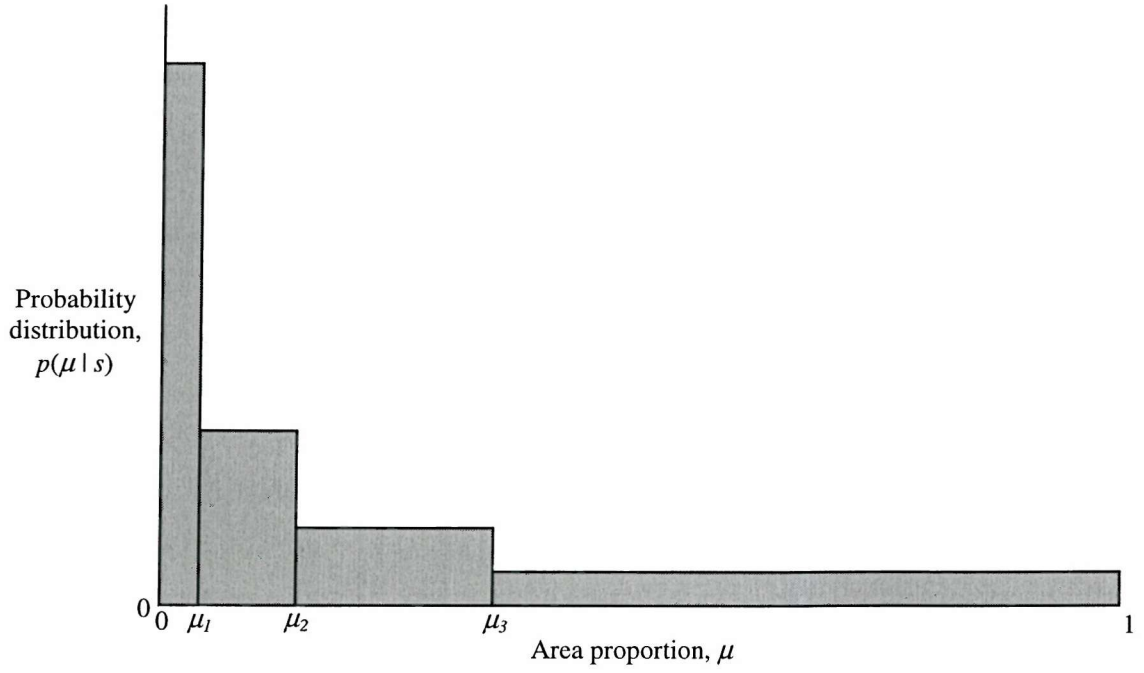

Figure 74: A histogram representation of the spectrum conditional proportion distribution.

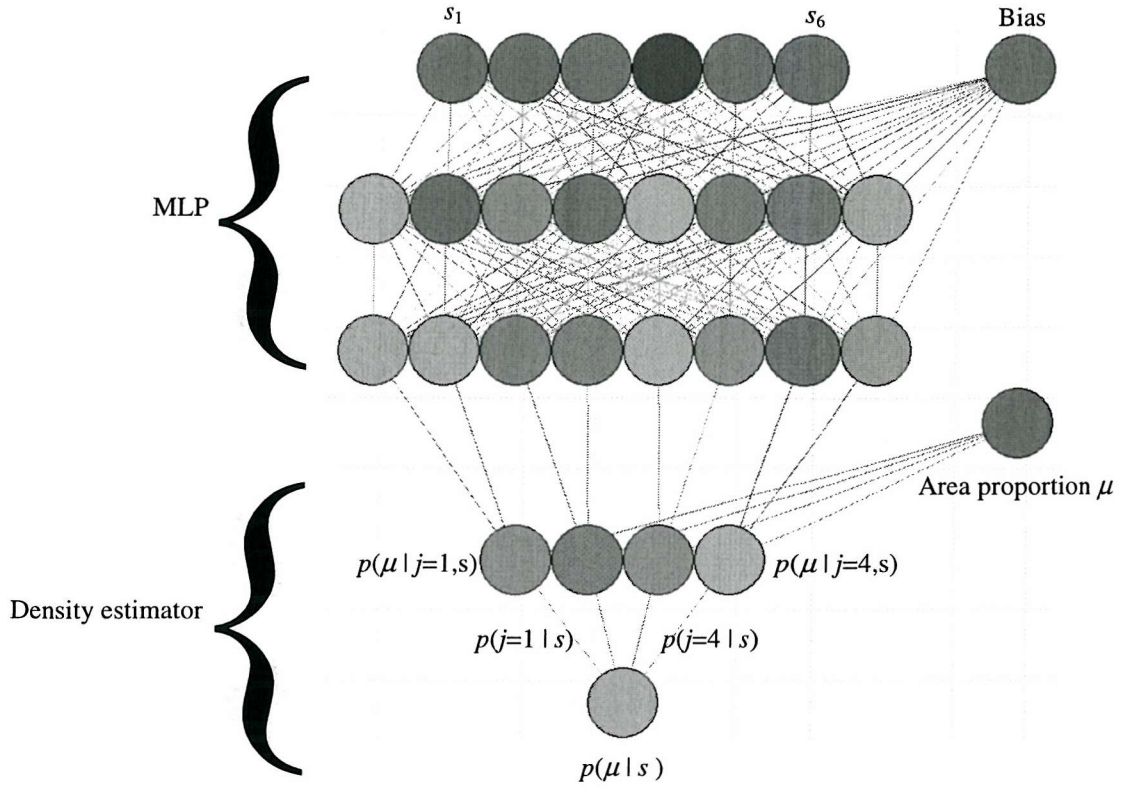

Figure 75: A histogram based spectrum conditional area proportion distribution model with eight hidden neurons and four basis functions in the density estimator.

9.3. Modelling Spectrum Conditional Distributions with a Histogram Conditional Density Estimator

This section introduces a new type of stratified classifier, shown in figure 75, that allows the complexity of the histogram representation of the spectrum conditional distribution to vary, resulting in a model that achieves a high level of flexibility with minimal additional computation. A typical distribution representation produced by the model described in this section is shown in figure 74. Notice that, unlike the previous technique, the widths of the basis functions are not fixed and equal, but are determined by the model, thus permitting it to use greater complexity in its representation of the distribution where it changes rapidly, in this case when the proportion variable is small. Rather than dividing the range of valid area proportions into a series of fixed 'bins' and estimating the posterior probability that the true area proportion lies in each bin, the histogram conditional density estimator uses a series of rectangular basis functions to approximate the spectrum conditional area proportion density $p(\mu \mid s)$. The density is modelled using a network with the structure shown in figure 75, and uses a

superposition of J non-overlapping basis functions $p_j(\mu \mid s) : 1 \leq j \leq J$, which are combined with priors $p(j) : 1 \leq j \leq J$

$$p(\mu \mid s) = \sum_{j=1}^{J} p(j) p_j(\mu \mid s) \qquad 86$$

where each basis function is controlled by a width parameter w_j such that it has unit area:

$$p_j(\mu \mid s) = \frac{1}{w_j} . \qquad 87$$

if μ lies within the support of the basis function and $p_j(\mu \mid s)=0$ otherwise. Since the model is acting as a density estimator, it is appropriate to find the model parameters using maximum likelihood, or equivalently, minimising the negative log-likelihood

$$E = -\ln p(\mu \mid s) \qquad 88$$

which gives

$$E = -\ln \sum_{j=1}^{J} \frac{p(j)}{w_j} . \qquad 89$$

Since both the priors p(j) and the basis function widths w_j must sum to unity, it is convenient to define them both in terms of a set of dummy variables y_j^p and y_j^w respectively and derive the true variables using the softmax function

$$p(j) = \frac{\exp(y_j^p)}{\sum_{k=1}^{J} \exp(y_k^p)} \qquad 90$$

and

$$w_j = \frac{\exp(y_j^w)}{\sum_{k=1}^{J} \exp(y_k^w)} \qquad 91$$

In order to minimise the negative log-likelihood, it is efficient to use a gradient based optimisation technique that requires the derivatives of the error function with respect to the dummy variables. Using the chain rule:

$$\frac{\partial E}{\partial y_k^w} = \sum_{j=1}^{J} \frac{\partial E}{\partial w_j} \frac{\partial w_j}{\partial y_k^w} \tag{92}$$

and

$$\frac{\partial E}{\partial y_k^p} = \sum_{j=1}^{J} \frac{\partial E}{\partial p(j)} \frac{\partial p(j)}{\partial y_k^p} \tag{93}$$

where the summations across all basis functions are necessary due to the coupling introduced by the softmax functions. Evaluating the second terms gives:

$$\frac{\partial p(j)}{\partial y_k^p} = p(j)\delta_{jk} - p(j)p(k) \tag{94}$$

and

$$\frac{\partial w_j}{\partial y_k^w} = w_j \delta_{jk} - w_j w_k \tag{95}$$

and similarly for the first terms

$$\frac{\partial E}{\partial p(j)} = -\frac{w_j}{p(j)} \frac{1}{w_j} \tag{96}$$

and

$$-\frac{\partial E}{\partial w_j} = \frac{p(j)}{w_j} \frac{1}{p(j)} \tag{97}$$

where the idenitity

$$-\ln x = \ln \frac{1}{x} \tag{98}$$

has been used in the second case. These simplify by cancellation to:

$$\frac{\partial E}{\partial p(j)} = -\frac{1}{p(m)} \tag{99}$$

and

$$-\frac{\partial E}{\partial w_j} = \frac{1}{w_j} \qquad 100$$

Inserting these results into the compound differentiation formulae gives:

$$\frac{\partial E}{\partial y_k^w} = \sum_{j=1}^{J} \frac{1}{w_j} w_j (\delta_{jk} - w_k) \qquad 101$$

which simplifies to

$$\frac{\partial E}{\partial y_k^w} = \sum_{j=1}^{J} (\delta_{jk} - w_k) \qquad 102$$

and

$$\frac{\partial E}{\partial y_k^p} = -\sum_{j=1}^{J} \frac{1}{p(j)} p(j)(\delta_{jk} - p(k)) \qquad 103$$

which similarly simplifies to

$$\frac{\partial E}{\partial y_k^p} = \sum_{j=1}^{J} (\delta_{jk} - p(k)) \qquad 104$$

If the dummy variables y_k^p and y_k^w are outputs of a model (such as an MLP) with spectral signature as input, equations for updating the model parameters by back propagation can be derived by dividing the model outputs into two groups, one controlling the priors and the other the basis function widths, adding softmax functions to each group and replacing the standard equations for the derivative of the error with respect to each output with equations 102 and 104. Gradient based optimisation algorithms should be used with caution when searching for the parameters of this model since the discontinuities in the basis functions result in discontinuities in the error surface.

As mentioned previously, the distributions predicted by spectrum conditional distribution models may conveniently be summarised by their mean and variance. The mean of any spectrum conditional distribution given by:

$$\mu_{mean} = \int \mu p(\mu \mid s) d\mu \qquad 105$$

which, when substituting the histogram model, becomes

$$\mu_{mean} = \int \mu \sum_{=1}^{J} p(j) p_j(\mu \mid s) d\mu \qquad 106$$

which may be re-ordered

$$\mu_{mean} = \sum_{j=1}^{J} p(j) \int \mu p_j(\mu \mid s) d\mu \qquad 107$$

The integral is the mean value of μ inside the j^{th} basis function which, due to the flat nature of the basis function, is equal to the basis function's centre, such that

$$\mu_{mean} = \sum_{j=1}^{J} p(j) \left\{ \frac{1}{2} w_j + \sum_{k=1}^{j-1} w_k \right\} \qquad 108$$

The variance of a distribution can be computed using:

$$\mu_{\text{var}} = \int \mu^2 p(\mu \mid s) d\mu - \left\{ \int \mu p(\mu \mid s) d\mu \right\}^2 \qquad 109$$

where the last term is the square of the distribution's mean:

$$\mu_{\text{var}} = \int \mu^2 p(\mu \mid s) d\mu - \mu_{mean}^2 \qquad 110$$

The second term has thus already been computed and only the first term will be considered in the following steps. Substituting in the histogram model of the distribution gives

$$\int \mu^2 p(\mu \mid s) d\mu = \int \mu^2 \sum_{j=1}^{J} p(j) p_j(\mu \mid s) d\mu \qquad 111$$

which may be reordered to give:

$$\sum_{j=1}^{J} p(j) \int \mu^2 p_j(\mu \mid s) d\mu \qquad 112$$

Recalling that the j^{th} basis function has value zero unless μ lies between $\sum_{k=1}^{j-1} w_k$ and $\sum_{k=1}^{j} w_k$, in which case it has value $1/w_j$, the integral may be rewritten as:

$$\sum_{j=1}^{J} p(j) \frac{1}{w_j} \int_0^{w_j} \left\{ \mu + \sum_{k=1}^{j-1} w_k \right\}^2 d\mu \qquad 113$$

which when evaluated symbolically gives:

$$\sum_{j=1}^{J} p(j) \left(\frac{1}{3} w_j^3 + w_j \left\{ w_j + \sum_{k=1}^{j-1} w_k \right\} \sum_{k=1}^{j-1} w_k \right) \qquad 114$$

Using these results in the equation for the distribution variance gives:

$$\mu_{\text{var}} = \sum_{j=1}^{J} p(j) \left\{ \frac{1}{3} w_j^2 + \left(w_j + \sum_{k=1}^{j-1} w_k \right) \sum_{k=1}^{j-1} w_k \right\} - \left(\sum_{j=1}^{J} p(j) \left\{ \frac{1}{2} w_j + \sum_{k=1}^{j-1} w_k \right\} \right)^2 \qquad 115$$

To evaluate the performance of the histogram based spectrum conditional density model, a model with four basis functions was trained on both the crops data and the tall herb data. Images of the means and variances of the distributions predicted for each of these problems are shown in figures 76 to 79. Since the density estimator does not produce single area proportion estimates, it is not immediately clear how to compare its performance with that of more conventional algorithms. Indeed, it is not even clear whether it is meaningful to make simple performance comparisons between the techniques, since they seek to extract different types of information, the distribution of possible proportions in the case of the density estimator as compared with the single best proportion estimate for the conventional algorithms.

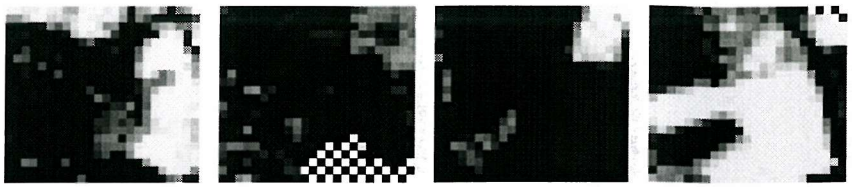

Figure 76: Means of the cereal distributions predicted by the histogram density estimator.

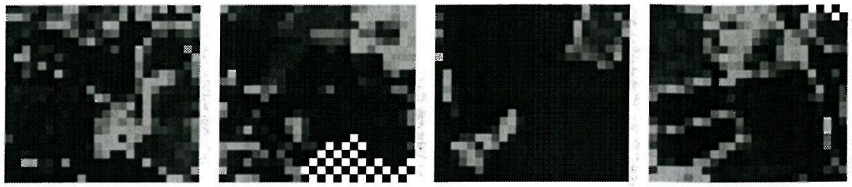

Figure 77: Variances of the cereal distributions predicted by the histogram density estimator.

The approach taken here is to propose the mean squared error between the mean of the distributions predicted by the density estimator for the validation sets and the actual subpixel proportions. The reasoning behind this is that if the distribution modelled by the density estimator were correct then the means of those distributions would be the optimal proportion estimates in the sense that they would, on average, minimise the sum of squares and cross entropy errors. Such a performance measure can be used only as a guide however, since it ignores the fact that the density estimator extracts far more information from the set of exemplars than do more conventional algorithms, and the performance is measured over a finite set of samples – the validation set – and is hence subject to variance. This latter point means that it is possible that even the best possible area proportion predictor could be outperformed by much worse algorithms on the validation set, but not on other data sets in general.

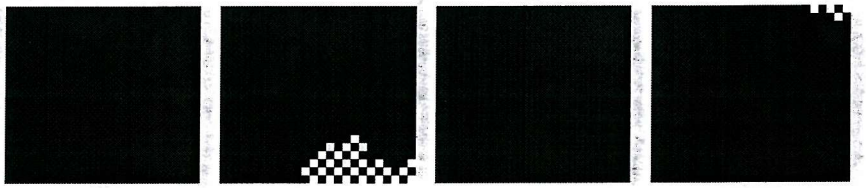

Figure 78: Means of the tall herb distributions predicted by the histogram density estimator.

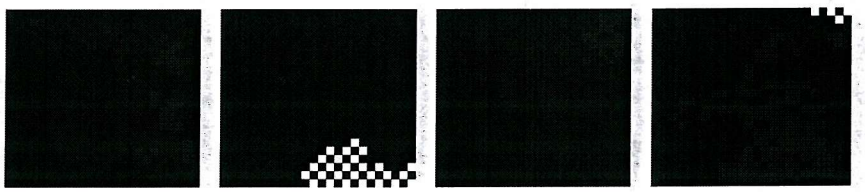

Figure 79: Variances of the tall herb distributions predicted by the histogram density estimator.

The results in figures 76 to 81 show that the means of the distributions predicted by the density estimator are very similar to the proportion estimates produced by other

algorithms. This includes the areas where the proportion estimates are poor, chiefly in lower left of the third subimage and the upper right of the fourth. The image of the distribution variances has the interesting characteristic that regions of high variance correspond quite strongly to areas where the distribution means poorly predict subpixel proportions. This was a possibility that was hinted at earlier which is now shown to occur in practice: in areas where the density estimator cannot predict the subpixel cover accurately, it is often able to provide warning of its failure. Most of the field boundaries seem to be regions of high variance, suggesting that the network has only a limited ability to model the proportions in such regions. This may be because the majority of cereal pixels are pure and the training data contains too few mixed pixels to learn the relationship between their spectral signature and their subpixel proportions, or it may be because mixed pixels with different proportions are spectrally confused and cannot be separated using the six spectral bands that are available. Whichever of these is the true cause of the observed weakness of the proportion predictions, it is important to emphasise that the density estimator, unlike the more conventional alternatives, is able to highlight the regions where its predictions are likely to be weak.

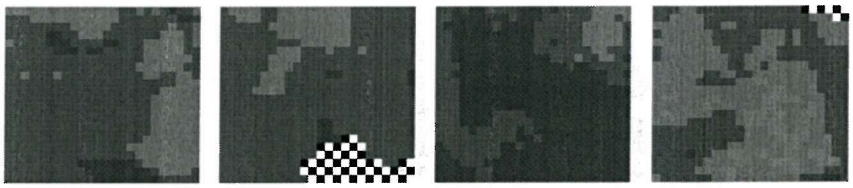

Figure 80: Enhanced images of the means of the tall herb distributions predicted by the histogram density estimator.

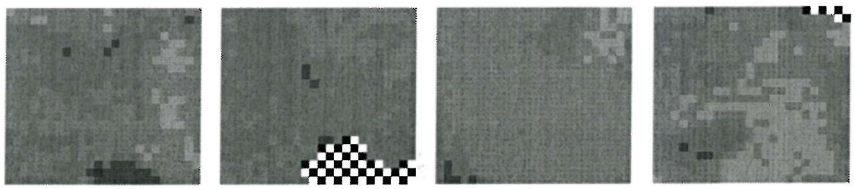

Figure 81: Enhanced images of the variances of the tall herb distributions predicted by the histogram density estimator.

The following section describes the mixture density network, a version of the histogram based density estimator that replaces the rectangular basis functions with Gaussians. This produces two main advantages, firstly the error surface is smooth and continuous meaning that efficient gradient based optimisation algorithms can be used with confidence and secondly that the density models are likely to be, on average, closer to the true spectrum conditional densities. These two advantages come at the cost of higher computational complexity both during training and query.

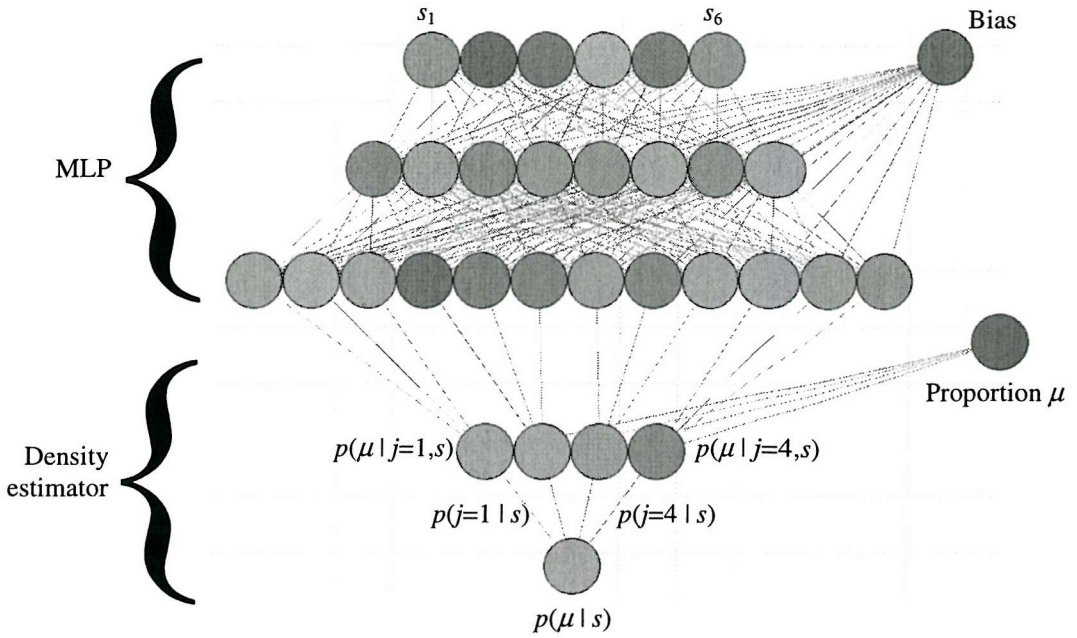

Figure 82: A mixture density network with eight hidden neurons and four components in the mixture model.

9.4. Modelling Spectrum Conditional Distributions with a Gaussian Mixture Model Conditional Density Estimator

The mixture density network, as shown in figure 82, is a hybrid structure containing both an MLP and a Gaussian mixture model density estimator. The network operates by using the MLP to modify the parameters of the Gaussian mixture model (the means, variance, and priors of the mixture components), and hence the shape of the conditional probability density it represents according to the conditioning variables. In the area proportion estimation problem, the conditioning variables are the spectral bands of the pixel under consideration, s_1 to s_6, the conditioned variable is the subpixel area proportion μ, and hence the density modelled by the network is $p(\mu \mid s)$, the spectrum conditional area proportion distribution. The density models produced by the mixture density network are similar to those of the histogram based conditional density estimation algorithm described in the previous section except that the mixture density models are smooth. This means that, on average, the mixture density network density models are likely to be closer to the true distributions and that the parameters of the mixture density network may safely be searched for using gradient based techniques. Note that the use of this mixture density terminology differs from that sometimes used in the area proportion estimation literature since it refers to the probability density in

area proportion space rather than spectral space (for examples of the latter usage, see [Tubbs76] [Peters:76], etc.).

To query the mixture density network, the spectral signature of the pixel under consideration is input to the MLP, and the information is propagated through the MLP in the usual manner. Once the outputs of the MLP have been computed, they are used as parameters in a one-dimensional density estimator, the input to which is an area proportion estimate. It may at first seem strange that one of the inputs to a model designed to estimate area proportions is an area proportion estimate, but the output of the model can be thought of as a measure of the consistency of the proportion estimate put into it with the spectral signature that was observed. By querying the mixture density network with a range of area proportions in the interval [0,1], it is thus possible to construct a graph of the spectrum conditional density (or consistency) of a range of proportions. The mixture density network therefore has the potential to provide information about subpixel cover using a representation powerful enough to describe the ambiguity implicit in inferences made about subpixel cover based on remotely sensed imagery.

The spectrum conditional density for an area proportion μ given a spectral signature s is computed by considering each of the Gaussian basis functions in the density estimator to be independent generators of pixels with Gaussian distributed area proportion distributions. Thus, the predicted spectrum conditional distribution is:

$$p(\mu \mid s) = \sum_{j=1}^{J} p(\mu \mid j, s) p(j \mid s) \qquad 116$$

where there are J components in the mixture model, each with a Gaussian distribution:

$$p(\mu \mid j, s) = \frac{1}{(2\pi\sigma_j^2(s))^{1/2}} \exp\left(-\frac{(\mu_j^{mean}(s) - \mu)^2}{2\sigma_j^2(s)} \right) \qquad 117$$

where $\mu_j^{mean}(s)$ is the mean of the jth component and $\sigma_j^2(s)$ is its variance. The terms $p(j|s)$ are the conditional priors for each of the J basis functions, that represent the probability that each basis function generates an unspecified proportion. The basis function means need to be bounded to lie between zero and one and hence should be connected to MLP outputs that use logistic activation functions, while the variances must be positive, a condition that can be guaranteed by connecting them to outputs that

have exponential activation functions. Finally, the condition that the basis function priors must sum to unity can be satisfied by connecting the priors to outputs that are constrained by the softmax activation function.

The mixture density network can be trained in essentially the same way as any other neural network; by using error backpropagation gradient descent. The details of the derivation of the equations for updating the network weights can be found in [Bishop:94] and [Bishop:95], and only an outline will be given here. As with most training algorithms, that for the mixture density network is derived by considering the network to be generating input-output pairs and then deriving a procedure for finding the network parameters that maximise the probability of generating the input-output pairs that are actually observed in the training set. Thus, the probability that the mixture density network generates the n^{th} input-output pairing of the training set (μ_n,s_n) is equal to the network output $p(\mu_n|s_n)$ and so the probability that the network with parameters w generates the entire training set D of N patterns is (assuming independence) the joint probability of correctly generating each training pattern:

$$p(D \mid w) = \prod_{n=1}^{N} p(\mu_n \mid s_n) \qquad 118$$

This is the likelihood of the weights given the data and is the quantity normally maximised during network training. Maximisation of the likelihood is equivalent to minimisation of the negative log-likelihood – a procedure generally preferred since the influence of the terms in the likelihood resulting from each individual training pattern are decoupled resulting in a simplification of the optimisation problem. The negative log-likelihood is normally thought of as the error function that is minimised during training and is hence usually labelled E:

$$E = -\sum_{n=1}^{N} \ln p(\mu_n \mid s_n) \qquad 119$$

The networks demonstrated in the following sections were trained using the standard gradient descent procedure, normally used for MLPs, with the simple addition of the step size adjustment described earlier to ensure the stability of the training procedure. More recent experiments have suggested that the "R-Prop" algorithm described in [Jervis:93] can produce a significant increase in the speed of training of the mixture

density network in particular, and it is strongly suggested that this algorithm be considered in all future work.

Example: Inverting $y=x^2$

In order to demonstrate the operation of the mixture density network, it was trained on a simple problem. The network was given a data set based on inverting the equation $y=x^2+\varepsilon$ where ε is an additive noise component uniformly distributed in the closed interval [-0.1,+0.1]. To do this a list of 1,000 numbers were chosen uniformly in the closed interval [-1,+1], their corresponding squares were computed, and a small amount of noise was added to the result. Thus, the training data set consisted of 1,000 input-output pairs such as those below:

Pattern	Training Input y	Training Target x
1	0.042	0.263
2	0.688	0.825
⋮	⋮	⋮
999	0.939	-0.943
1000	-0.041	0.120

Table 11: Some examples of the patterns in the training set for the MDN example.

The mixture density network was given the task of recovering x from y – a problem that is ill-posed in the same sense as that of extracting subpixel information from remotely sensed images. The network was trained for a fixed period of 12 hours in the manner already described using the target function shown in figure 84 producing a mixture density network with the behaviour shown in figure 85. The mixture density network was repeatedly tested with combinations of values of x and y and the probabilities $p(x|y)$ predicted by the mixture density network recorded and used to produce the contour map.

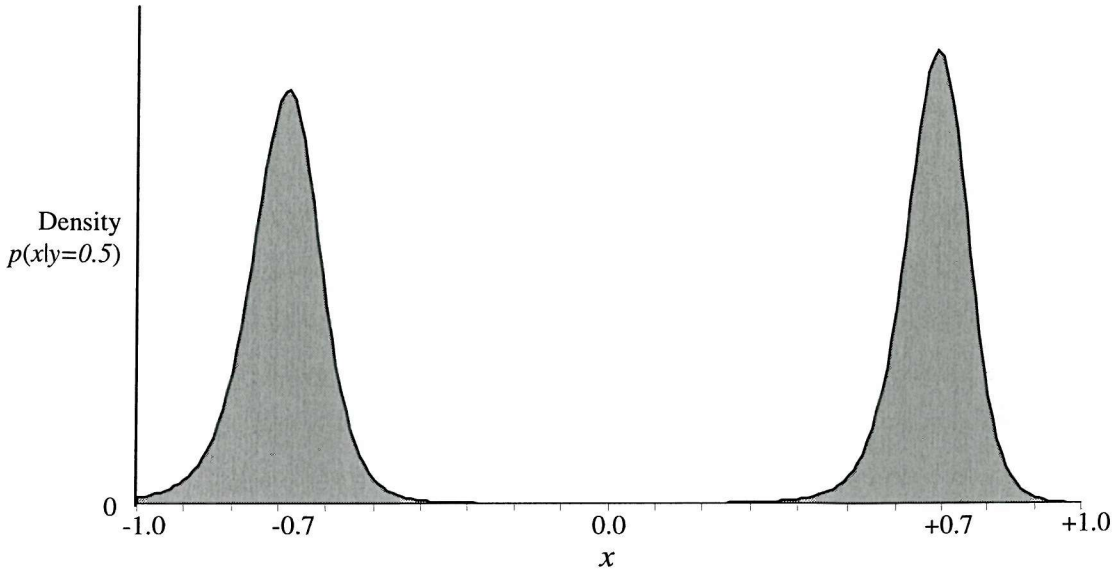

Figure 83: The distribution of x predicted by the MDN given that y=0.5. Note that the MDN correctly predicts that x must lie close to ±0.707.

By comparing figure 84 and 85, it can be seen that the mixture density network has captured the overall form of the relationship between x and y, in the sense that it predicts high probabilities for valid combinations of x and y and low probabilities for invalid combinations. The magnitudes of predictions of x for values of y close to one are positively biased due to the asymptotic behaviour of the logistic activation functions used in the hidden layer of the MLP component of the mixture denisty network. This effect could be relieved by further training. Figure 83 shows the distribution modelled by the mixture density network at y=0.5, for which x could have been ±0.707. The mixture density network correctly predicts that for this value of y, x should be around either +0.707, or –0.707, the variance in the components of the predicted distribution being due to the noise term present in the training data.

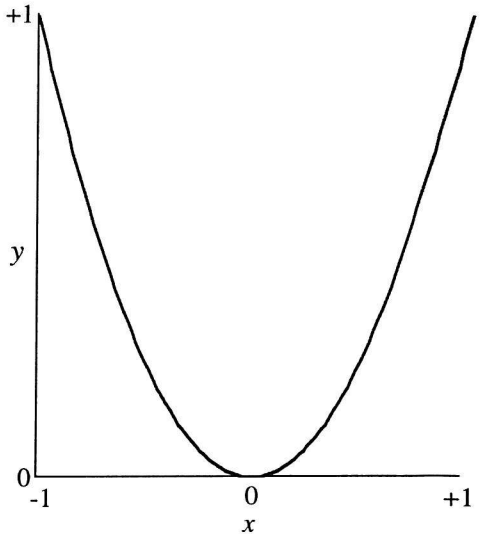

Figure 84: $y=x^2$.

This simple example provides a clear illustration of the power of the mixture density network in extracting complex information from a set of exemplars, and the flexibility of using conditional probability densities to represent ambiguity in the solution of inverse problems. Most conventional modelling algorithms would be inappropriate for solving the type of problem described here since the representations they use are incapable of describing the ambiguity implicit in the solutions of such problems. If an MLP or SVM (see [Brown:00]) had been applied to this example, they would both have predicted that x was always approximately zero for all values of y, since such predictions minimise the cost functions that the algorithms use for training. The following section presents the results of applying the mixture density network to the FLIERS data set.

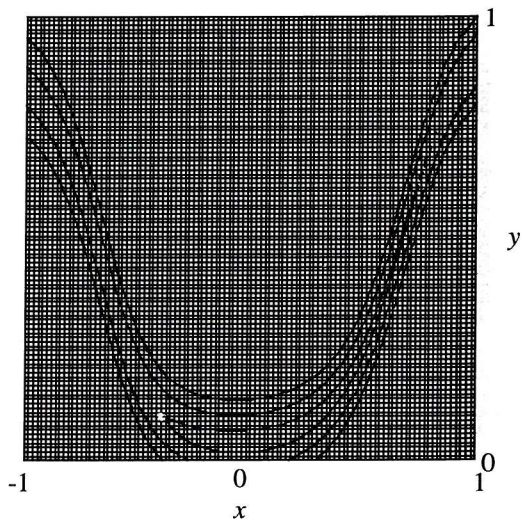

Figure 85: Contour plot of the inverse of $y=x^2$ as modelled by the MDN.

9.5. Deriving Spectrum Conditional Proportion Distributions from the FLIERS Data Set

To produce the results described in this section, a mixture density network with five hidden neurons in its MLP component and four Gaussian basis functions in its density estimator component was applied to the FLIERS data set [Manslow:00b]. The means and variances of the distributions produced by the MDN are shown in figures 86 to 88 for the cereal data and figures 92 to 95 for the tall herb data. The mean squared error performances of the means of the distributions modelled by the MDNs as single proportion estimates were 0.04854 for the unseen cereal data and 0.02093 for the unseen tall herb data. Although these are comparable to those for the fully fuzzy MLP, such performance measures should be used for guidance only in the case of the MDN, since the main strength of the conditional density estimation techniques is that they model the distribution of proportions – information that can be of great value even though it is not accounted for in any of the standard error measures.

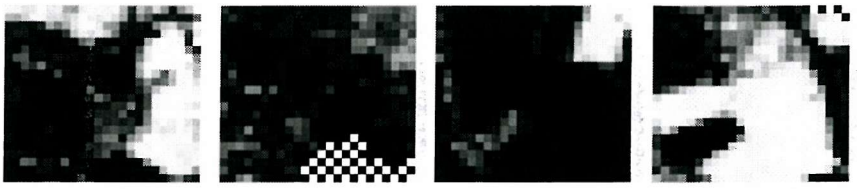

Figure 86: The means of the distributions predicted by the MDN.

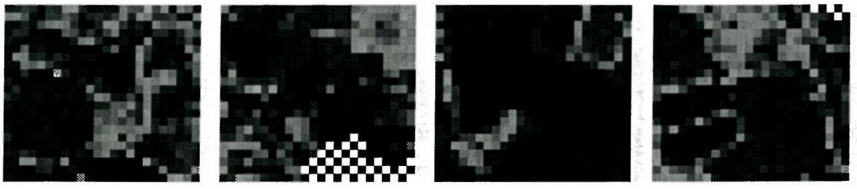

Figure 87: The variances of the distributions predicted by the MDN.

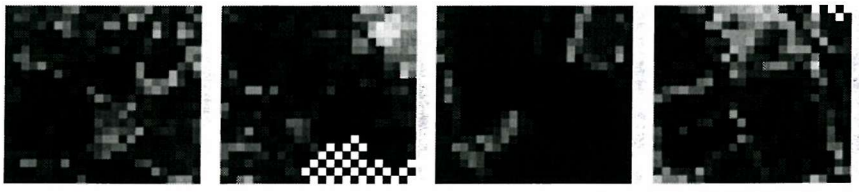

Figure 88: The magnitude of the squared errors that result from using the MDN distribution means as proportion estimates.

The mean of the distributions modelled by the MDN are shown in figure 86 and are very similar to the proportion estimates made by the fully fuzzy MLP based proportion estimators described in section 7.2.1, suggesting that, in terms of the best single proportion estimate, the MDN and the more conventional MLP fuzzy classifier are in agreement. Figure 87 shows the variances of the distributions modelled by the MDN

which, when compared to the magnitudes of the prediction errors made by the distribution means in figure 88 indicates that the distribution variances provide useful information about the accuracy of the distribution means as proportion estimates. It should be noted however that the distribution variances represent the expected sum of squares error of the distribution mean as a proportion estimate given that the distribution is correct and that for specific pixels the error that is actually observed may be greater or less than the distribution variance. As described earlier such variances can be used to direct further analysis and visualisation to those pixels where a single proportion estimate is unlikely to adequately summarise the possible scenarios of subpixel cover.

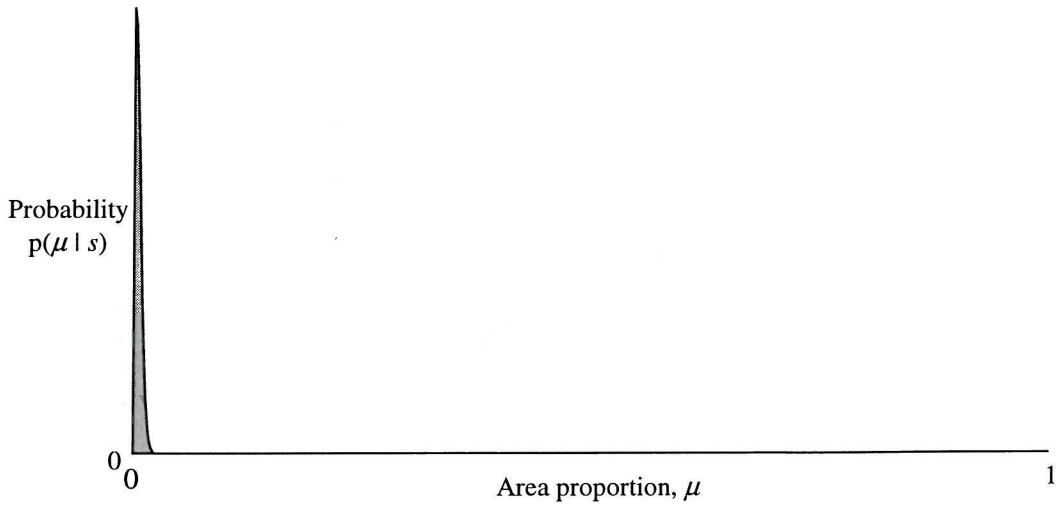

Figure 89: Distribution predicted by the MDN for a pixel that probably contains no cereal.

Figures 89, 90, and 91 show spectrum conditional distributions extracted from the mixture density network for three different pixels in the validation area. The first distribution is from a pixel in the third subimage that is known not to contain any cereal whatsoever and clearly shows that the MDN confidently predicts that the pixel contains either no cereal or only very small quantities of it. This distribution has low variance correctly suggesting that the mean proportion is likely to be accurate and effectively summarise the information contained in the distribution. Figure 90 shows the predicted distribution for a pixel taken from the fourth subimage that is known to consist purely of cereal. Once again, the distribution consists essentially of a pronounced single peak, but this time suggesting that the pixel consists purely of cereal. There are two other peaks visible in the distributions, one centred at about μ=0 and the other at about μ=0.85. The first of these is rather small in the sense that the volume under the peak is negligible

compared to the volume of either of the others. The latter peak is very common for pixels where the presence of cereal is predicted. This seems to suggest that it is relatively easy to detect the presence of cereal (or, conversely, its absence) but difficult to precisely specify the subpixel proportion.

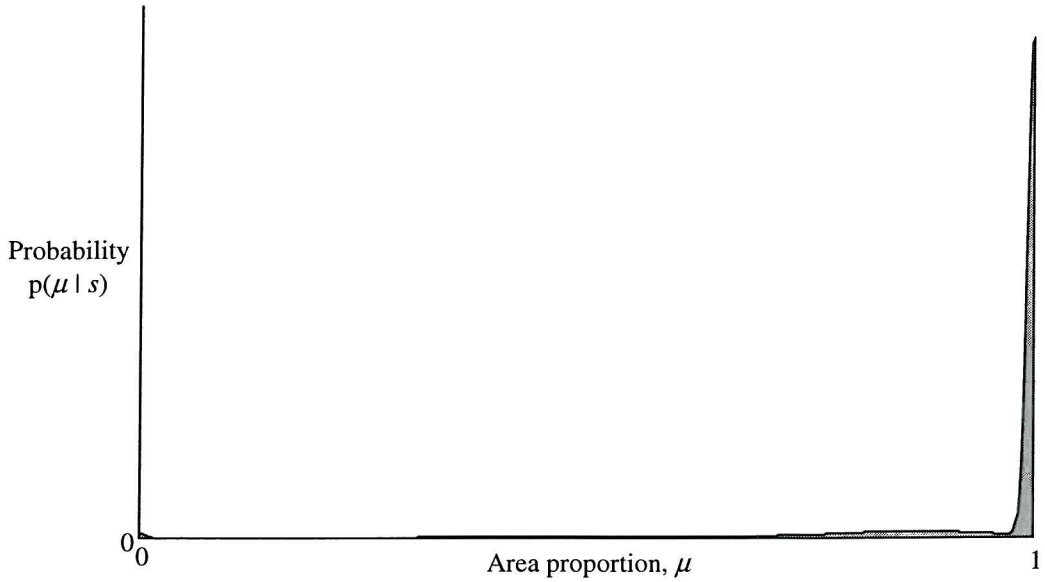

Figure 90: Distribution predicted by the MDN for a pixel that contains some cereal and is probably pure cereal.

Finally, figure 91 presents the distribution for a more problematic pixel – one taken from the lower left of the third subimage where the distribution mean incorrectly predicts the presence of cereals and where the variance is high. In this case, there is significant probability mass for all possible area proportions, suggesting that, on the basis of the set of exemplars and the pixel's spectral signature, the MDN cannot rule out any particular proportion occurring. In particular, four peaks are apparent in the distribution, one at μ=1, one at μ=0, one centred roughly at μ=0.8 and one at μ=0.4. The largest peaks are those for the pure pixels suggesting that a pixel with the observed spectral signature is likely to be pure, most likely purely cereal, but also quite likely to contain no cereal at all. Even though the distribution mean is a poor estimate of the actual subpixel proportion, the distribution itself suggests that a wide range of proportions are possible, and implies that the true proportion is one of the more probable. Note that although the standard MLP and the means of the MDN distributions predict similar proportions, the MDN variances provide a useful indicator of problematic pixels – those where the MLP and MDN means are likely to poorly predict subpixel cover, which can then be analysed in greater detail by examining the distributions of probable subpixel proportions predicted by the MDN.

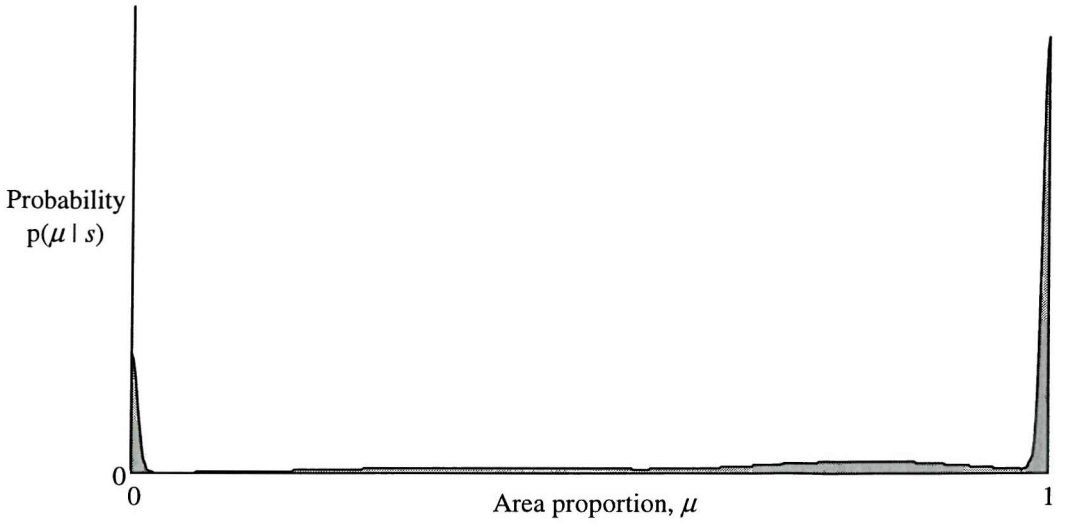

Figure 91: Distribution predicted by the MDN for a pixel that could contain almost any amount of cereal.

Figures 92 and 93 show the means and variances of the MDN predicted distribution means and variances for the validation areas in the tall herb data set. Since these values are quite small, enhanced versions of the prediction images are given in figures 94 and 95. From these images it is clear that the spectrum conditional area proportion distribution means poorly predict the actual proportions of subpixel cover. This should come as no surprise, since none of the algorithms so far applied have been able to predict the subpixel proportions of tall herb to a high degree of accuracy, suggesting that a pixel's spectral signature may contain too little information. Figures 96 to 98 show spectrum conditional distributions modelled by the MDN for three pixels from the validation areas. Figure 96 and 97 show two pixels where the MDN correctly predicts the absence and presence of tall herb. Figure 96 shows the predicted proportion distribution for a pixel from the third validation subimage known not to contain tall herb. The probability mass in the distribution produced by the MDN for this pixel is very tightly clustered around μ=0 suggesting that the pixel contains no tall herb with high probability. Figure 97 shows the predicted proportion distribution for a pixel that lies on the river also in the third validation subimage and which contains a substantial proportion of tall herb. In this case although the distribution still has a peak at μ=0, there is a large amount of probability mass for a range of proportions up to about μ=0.4 and beyond. Thus, although the pixel's spectral signature appears to contain too little information for the MDN to precisely predict the subpixel proportion of tall herb, the

MDN was able to indicate that it is possible that the pixel contains a range of proportions and is, in fact, likely to contain some tall herb.

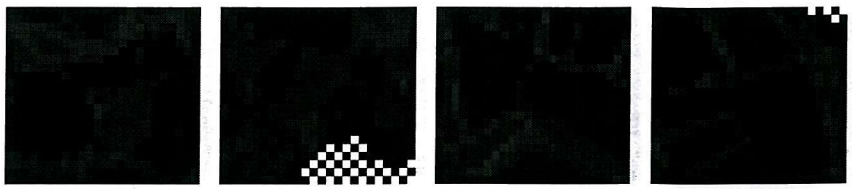

Figure 92: MDN estimated tall herb proportion distribution means.

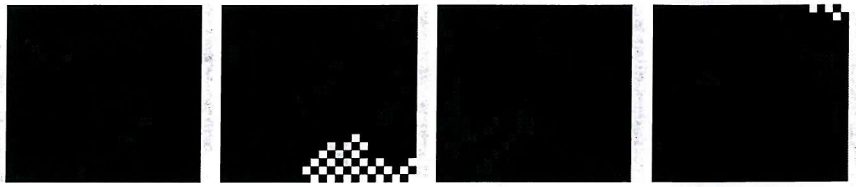

Figure 93: MDN estimated tall herb proportion distribution variances.

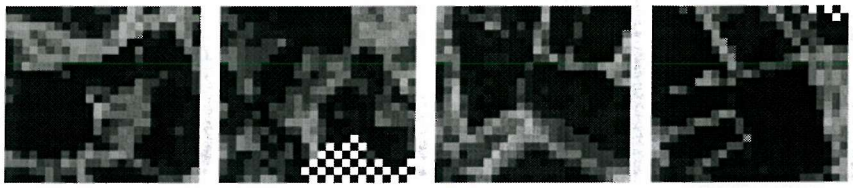

Figure 94: Enhanced MDN estimated tall herb proportion distribution means.

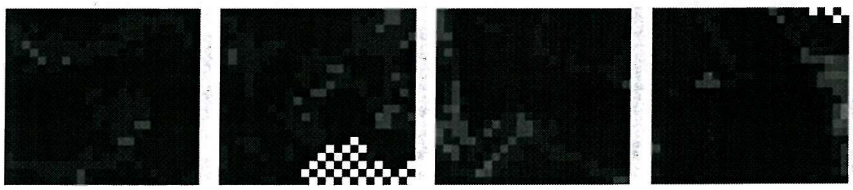

Figure 95: Enhanced MDN estimated tall herb proportion distribution variances.

Figure 98 and 99 show the spectrum conditional distributions for two more difficult pixels, one for which the mean of the distribution predicted by the MDN over estimates the quantity of tall herb, and the other for which it under estimates it. Figure 98 is typical of the distributions produced by the MDN when the distribution mean over estimates the true subpixel proportion. The distribution consists of a pronounced peak at μ=0, the true proportion, but also significant probability mass for greater proportions, suggesting that some pixels with the observed spectral signature may contain small quantities of tall herb. Figure 99 shows a pixel taken from the third subimage of the validation region which is known to contain tall herb, but for which the mean of the MDN distribution is small. In this case, the distribution is dominated by a large peak at μ=0, suggesting that most pixels of the observed spectral signature will contain no tall herb, leading to the distribution's low mean. However, the distribution does have significant probability mass for proportions right up to around μ=0.75, suggesting that although unlikely, it is possible that the pixel contains large quantities of tall herb.

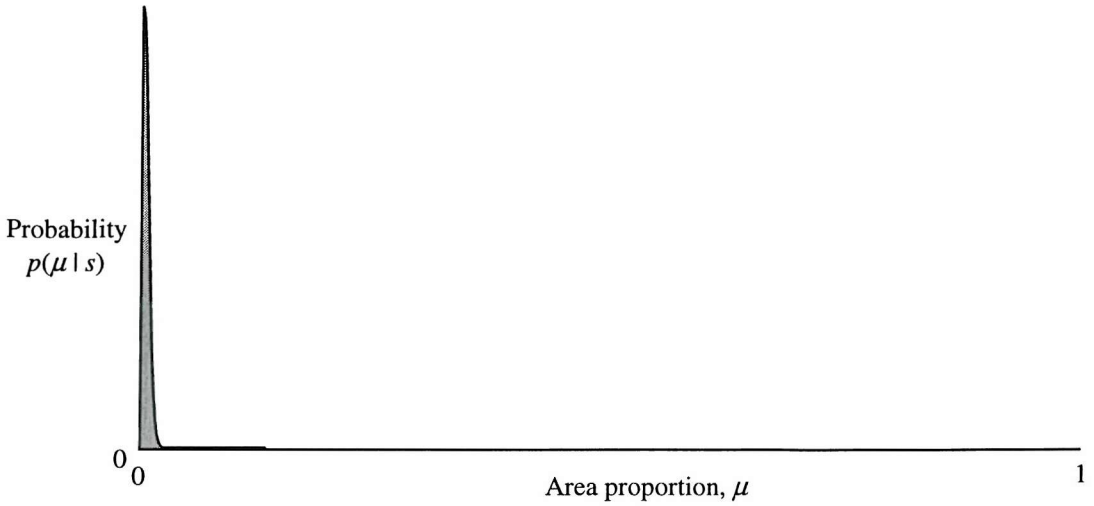

Figure 96: Tall herb distribution predicted by the MDN for a pixel containing no tall herb at all.

This section has shown how the mixture density network can be applied to the problem of fuzzy classification and subpixel area proportion estimation to provide new information into possible scenarios of subpixel cover. This can be done using extant data sets and requires little more time and effort than training other sophisticated non-linear models such as the MLP. In particular, it has been shown empirically that the MDN predicts subpixel proportions with an accuracy on a par with the MLP, that the variances of the distributions it models provide useful information as to the likely accuracy of its predictions and that when that accuracy is low, the distributions themselves can be used to provide insights into which subpixel proportions are most likely to occur. The spectrum conditional area proportion distribution representation along with the mixture density network that can be used to model it provides an efficient means of extracting new and detailed information about subpixel cover from remotely sensed data.

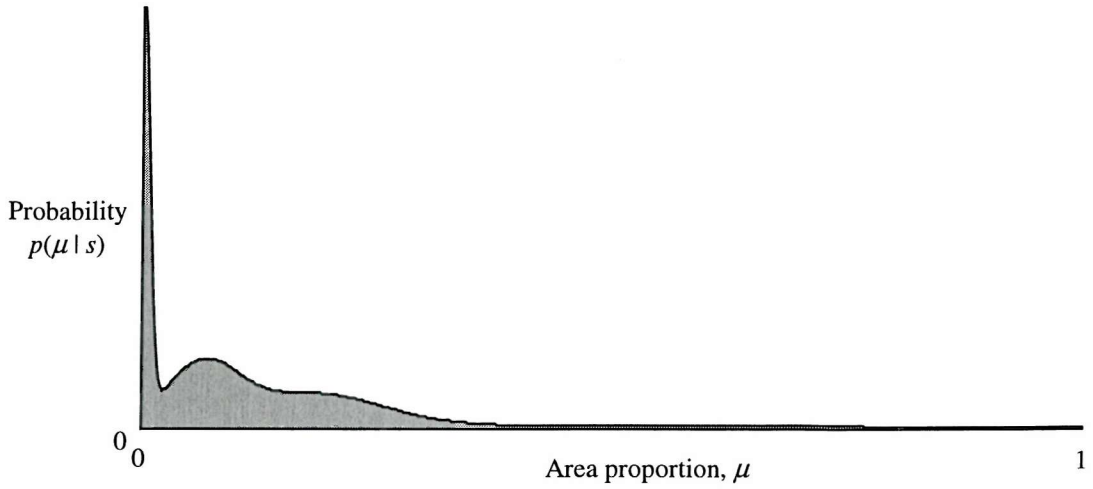

Figure 97: Tall herb distribution predicted by the MDN for a pixel containing some tall herb.

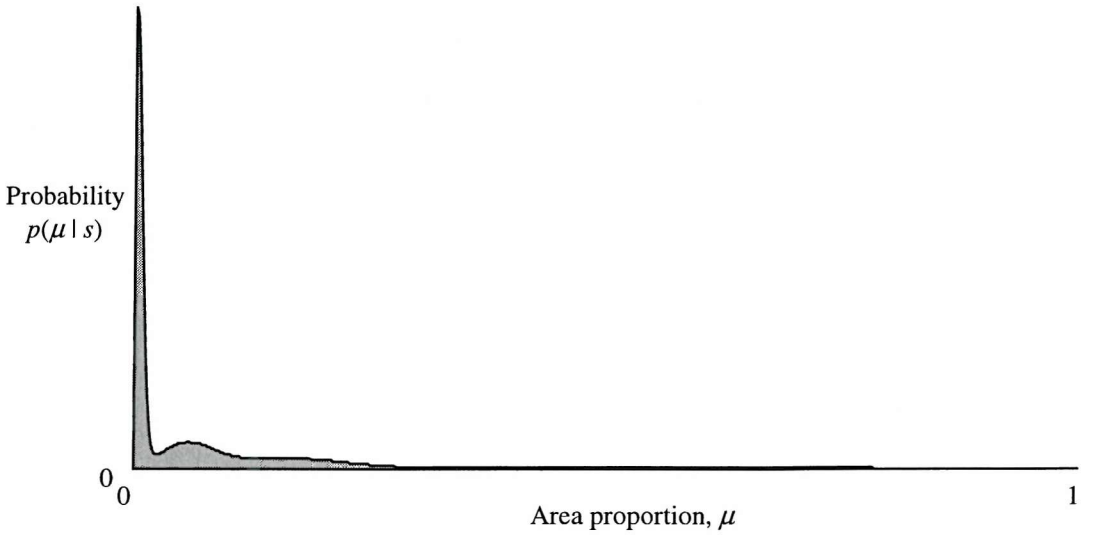

Figure 98: Tall herb distribution predicted by the MDN that has a mean less than the true proportion.

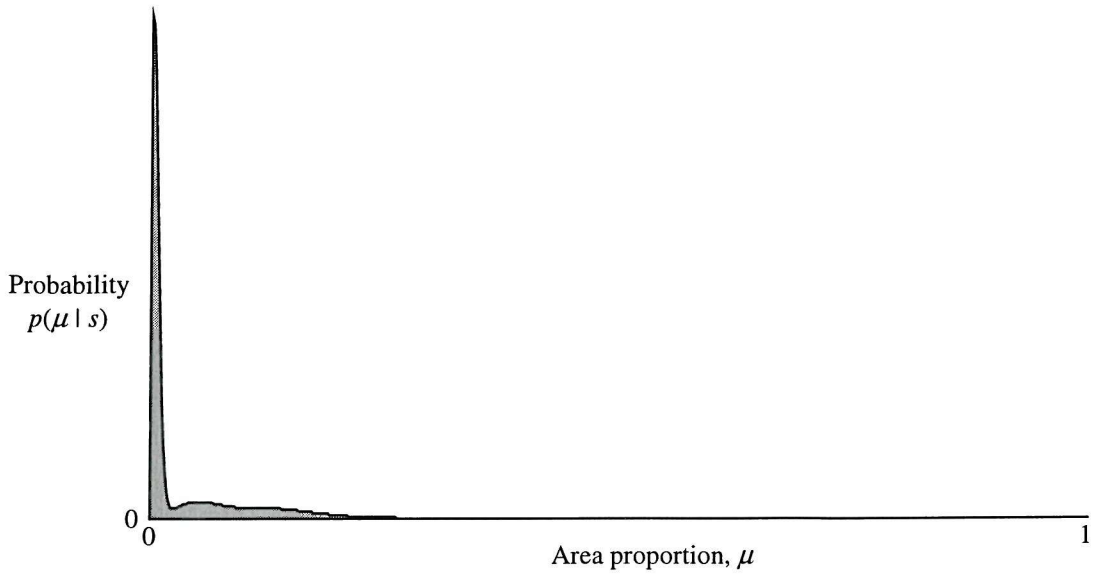

Figure 99: Tall herb distribution predicted by a MDN that has a mean greater than the true proportion.

10. Conclusions

This thesis has described a number of novel contributions in the field of subpixel land cover area proportion estimation. Foremost among these was the interpretation of subpixel area proportions as conditional probabilities of class membership of subpixel points. This interpretation provided a natural motivation for a new conditional probability-like notation for area proportions with which the primary axioms governing their behaviour were outlined. The probabilistic interpretation has made it possible to show that one of the standard approaches to fuzzy classification is equivalent to a crisp classifier that uses a Gaussian class conditional density model with fixed priors and is trained using the expectation maximisation algorithm. This equivalence is used to justify a number of extensions to the basic fuzzy classifier, including the re-introduction of priors, and the use of mixtures of Gaussians for the class conditional density models that are shown empirically to produce significant increases in performance. The fuzzy classifications produced by the new algorithm are shown theoretically to converge to the optimal subpixel proportion estimates as the flexibility of the fuzzy classifier and the quantity of training data are increased.

In addition, the probabilistic interpretation is used to show that the use of the cross entropy function in the derivation of subpixel proportion estimates has a specific meaning. That is, minimum cross entropy error estimates of subpixel area proportions maximise the probability that subpixel samples drawn from exemplar pixels and subpixel samples drawn from a set of pixels with area proportions predicted by the estimator have the same class membership. The results of an experiment that was designed to test this theory in practice showed that the cross entropy based estimator produced predictions that were, in this sense, superior to those of a sum of squares error based estimator. An analysis was presented of the relationship between posterior probabilities of class membership and optimal fuzzy classifications that suggested that although posterior probabilities and fuzzy classifications will usually be positively correlated, they will not, in general, be equal, even under ideal circumstances, such as when an infinite amount of data is available.

Chapter 8 examined some of the factors that limit the performance of area proportion estimation accuracy. The new concepts of primitive and compound classes were introduced and used to produce a concise list of the conditions necessary to minimise

the performance limit associated with the way in which classes are defined. In particular, error free proportion estimates cannot be obtained if there are more classes than spectral bands, or the classes exhibit spectral variation. Section 8.4 discussed the way in which the sensor point spread function introduces uncertainty into subpixel area proportion estimates and, for a simple Gaussian model of the point spread function, new bounds on this uncertainty were derived that showed that more ambiguity was induced in the estimation of proportions in more heavily mixed pixels. The recognition of the impossibility of deriving error free proportion estimates was used to justify the introduction of a new representation for proportion information derived from remotely sensed data: the spectrum conditional area proportion distribution.

Chapter 9 discussed the new representation in detail and outlined its three main benefits: that it is capable of fully representing the information contained in a pixel's spectral signature, that the completeness of the representation allows proportion information from a variety of disparate sources to be combined optimally, and that the proportion information may fully be propagated through down stream processes. Sections 9.2 to section 9.4 consider three increasingly complex approaches to deriving spectrum conditional proportion distributions, the first using a stratified classifier to estimate the posterior probability that the true subpixel proportion lies in one of a number of predefined bins, and the last using an MLP and Gaussian mixture model based conditional density estimator to directly estimate the spectrum conditional area proportion probability density for a given pixel. Empirical results were presented which show that in terms of discrete proportion estimation the means of the distribution models are comparable to the performance of a standard MLP. However, it is also shown that the distribution variances are useful in identifying pixels for which the discrete estimates are likely to be poor and that for such pixels, the full distribution can be obtained to provide complete information about the range of possible subpixel cover scenarios.

11. Future Work

It is the unique richness of the spectral conditional proportion distribution that makes it possible to completely represent, propagate and combine proportion information derived from different sources. These new capabilities have implications for a wide range of applications that were based on more conventional single estimate representations of derived proportion information and are likely to have produced results that are suboptimal or misleading. One simple example of the power of the proportion distribution is in the problem of computing percentage land cover change that was demonstrated in the first section of chapter 9 where it was shown that the area proportion distributions induce a distribution over the percentage of land cover change. In general, the optimal single percentage change estimate – defined as that which minimises the expected squared error over the percentage change distribution – is not equal to the percentage change calculated from the optimal area proportion estimates at each time. Thus, applications that previously used single proportion estimates to model land cover change not only produce an information poor representation – a single percentage change estimate – but also information that can be misleading. This simple example illustrates how the area proportion distribution representation can produce new insights into old problems through the completeness of the representation and the potential to propagate information without loss.

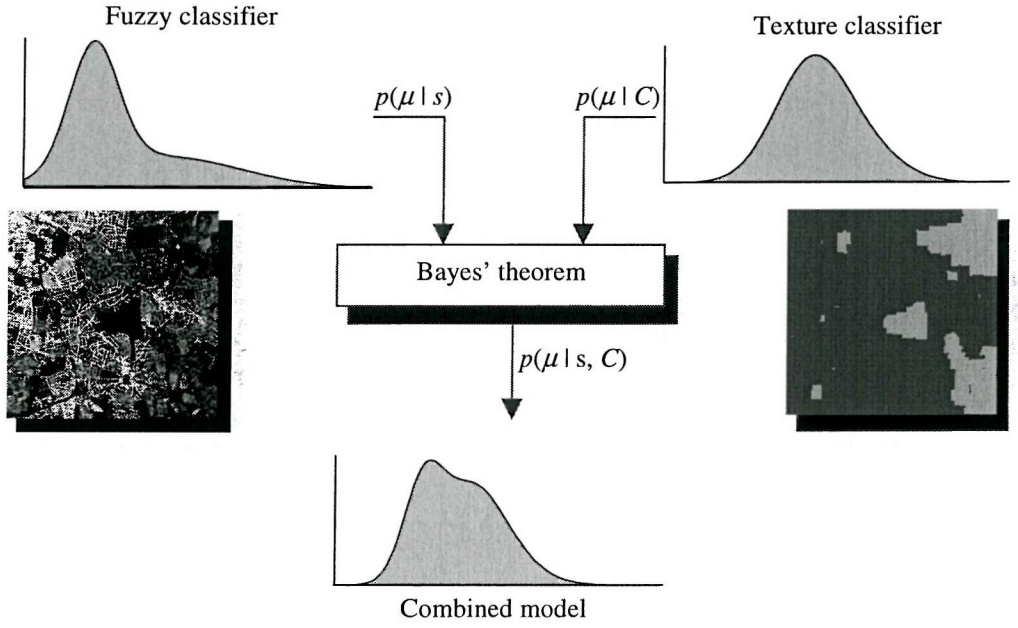

Figure 100: Combining texture information with area proportion information.

The capacity of the representation for optimal combination of proportion information from different sources was also discussed in the first section of chapter 9 and provides a new way of improving proportion estimation performance. For example, as part of the FLIERS project, a pixel classifier was produced that classified pixels according to the texture of their immediate neighbourhood. Although the classifier itself did not produce proportion information, it was possible to extract conditional proportion distributions given the classification it produced $p(\mu \mid C)$. This information could be used with Bayes' theorem as described in chapter 9 to enhance the spectrum conditional proportion information obtained from an MDN, $p(\mu \mid s)$, as illustrated conceptually in figure 100. Some preliminary results were obtained that strongly hinted at the utility of such a combination. These focussed on a problem that occurred in mapping the distribution of cereals in an image containing both urban a rural areas. In particular, it was found that small quantities of cereal would be predicted within the urban area suggesting that there were features in the urban area which were spectrally confused with cereal crops. Since the urban area has a distinctly different texture to the rural area, the texture classifier could identify it with a high level of accuracy. In addition the sparsity of cereals in the areas identified by the texture classifier as urban areas tended to suppress predictions of cereal made by the MDN when the two sources of information were combined using Bayes' theorem. Although these initial results show great promise, more work is required to fully assess the benefits of the combination and to establish a more complete representation of the information extracted by the texture classifier.

One minor problem with the MDN is that its use of Gaussian basis functions means that it always predicts illegal proportions – those less than zero and greater than one – with non-zero probabilities. The result of this is that the means of the proportion distributions tend to be biased towards μ=0.5, and the variances have a slight positive bias. In fact, the strict upper bound on the variance of the area proportion distribution (see appendix C) no longer applies. Although these effects are relatively minor, they can be overcome by mapping the area proportion space to the real line before applying the MDN. This is done by computing a set of dummy proportions μ_{dum} from the true proportions μ using the inverse logistic or logit transform:

$$\mu_{dum} = \tfrac{1}{\alpha}\ln|\mu| - \tfrac{1}{\alpha}\ln|1-\mu| \qquad 125$$

The dummy proportions are used by the density estimator and lie in the interval [-∞,+∞]. Thus, when the dummy proportions are mapped back to the original area proportion space using:

$$\mu = \frac{1}{1+\exp(-\mu_{dum})} \qquad 125$$

all the probability mass lies in the legal range [0,1]. Note that in order to maintain the correct shape and normalisation of the distribution, the spectrum conditional distribution is computed from that for the dummy variables using:

$$p(\mu \mid s) = p(\mu_{dum} \mid s)\frac{\partial \mu_{dum}}{\partial \mu} \qquad 122$$

where $p(\mu_{dum} \mid s)$ is modelled by the MDN, and

$$\frac{\partial \mu_{dum}}{\partial \mu} = \frac{1}{\alpha\mu(1-\mu)} \qquad 123$$

Although this offers a framework for bounding the distribution, it is not clear what level of performance improvement would result and whether it would be worth the extra computational cost that is necessary to compute the distribution means and variances. This can only be determined by further experimentation. This problem with the use of Gaussian distributions was also noted in [Chittineni:81] in the context of finding a suitable prior (unconditional distribution) for area proportion variables. The proposed solution consisted of artificially clipping the probability density to zero outside the range [0,1] and re-normalising within it. Although in principle, it should be possible to adapt that approach for use with the mixture density network, it is mathematically more complex than the technique outlined above which is therefore recommended for initial investigation.

In addition to using the Gaussian as a prior area proportion distribution, [Chittineni:81] also uses the Beta distribution [DeGroot:89] as does [Atkinson:99]. This distribution, given in below:

$$p(\mu) = Z\mu^{b}(1-\mu)^{c} \qquad 125$$

where Z is a normalisation constant, is very similar to the multinomial-derived proportion distribution discussed earlier, and also presented below:

$$p(\mu_{est} \mid \mu) = Z\mu_{est}{}^{\mu}(1-\mu_{est})^{(1-\mu)} \qquad 125$$

The clear similarity of these distributions suggests that a relationship must exist between them, though it is not immediately obvious what form it should take. In particular, the almost arbitrary parameters, b and c of the Beta distribution seem to have no specific relation to the "true" proportions, μ_{est}, in the conditional distribution. A detailed examination of the relationship between these two expressions may provide useful insights into area proportion distributions and yield useful results that provide guidance on the proper values for the parameters of the beta distribution.

Finally, it would be interesting to use the idea of simulating the interactions of the sensor PSF with subpixel cover of varying characteristic length scales using techniques similar to those described in [Kitamoto:99]. The aim of this work would be to derive expressions for the area proportion distribution induced by the PSF that would probably be described by a function of the ratio of the characteristic length scale of the cover type to the nominal pixel size. Although this uncertainty is implicitly modelled by the mixture density network it is mixed in with all the uncertainty from a large number of other sources and cannot easily be isolated. A model of the specific shape of the proportion distribution induced by the PSF would provide important insights into its effect, not least because its variance places an intrinsic limit on proportion estimator performance.

12. Appendix A: Derivation of Expectation-Maximisation Equations from the Kullback-Liebler Divergence

Consider the error function E measured over D pixels and M subpixel samples:

$$E = -\frac{1}{M}\sum_{d=1}^{D}\sum_{m=1}^{M} c_{dm} \ln p(s_d) \qquad 126$$

where c_{dm} is an indicator function taking on the value one if the m^{th} subpixel sample is in the class under consideration and zero otherwise and $p(s_d)$ is the probability of observing the spectral signature s_d. As the number of subpixel samples becomes large, E can be approximated by:

$$E = -\sum_{d=1}^{D} \mu_d \ln p(s_d) \qquad 127$$

where μ_d is the proportion of the d^{th} pixel covered by the class of interest. The change in this error measure when the parameters of the approximating mixture model are updated from $p_{old}(j)$, m_j^{old} and σ_j^{old} to $p_{new}(j)$, m_j^{new} and σ_j^{new} is:

$$E_{new} - E_{old} = -\sum_{d=1}^{D} \mu_d \ln \left| \frac{p_{new}(s_d)}{p_{old}(s_d)} \right| \qquad 128$$

where E_{new} and E_{old} are the new and old errors, and $p_{new}(s_d)$ and $p_{old}(s_d)$ are the estimated densities at s_d given the new and old parameters. Since

$$p_{new}(s_d) = \sum_{j=1}^{J} p_{new}(s_d \mid j) p_{new}(j) \qquad 129$$

,

the change in error can be re-written as:

$$E_{new} - E_{old} = -\sum_{d=1}^{D} \ln\left\{ \frac{p_{old}(j \mid s_d) \sum_{j=1}^{J} p_{new}(s_d \mid j) p_{new}(j)}{p_{old}(j \mid s_d) p_{old}(s_d)} \right\}. \qquad 130$$

Using Jensen's inequality, which states that for $\lambda_j \geq 0$, and $\Sigma_j \lambda_j = 1$,

$$\ln\left\{ \sum_{j=1}^{J} \lambda_j s_j \right\} \geq \sum_{j=1}^{J} \lambda_j \ln s_j \qquad 131$$

,

and substituting $p_{old}(j \mid s)$ for λ_j,

$$E_{new} - E_{old} \leq -\sum_{d=1}^{D} \mu_d \sum_{j=1}^{J} p_{old}(j \mid s_d) \ln \left| \frac{p_{new}(s_d \mid j) p_{new}(j)}{p_{old}(j \mid s_d) p_{old}(s_d)} \right|. \qquad 132$$

The right hand side of the above equation gives a lower bound on the amount by which the error decreases when the mixture model parameters are updated. It seems reasonable that the parameters should be changed in such a way as to maximise the minimum possible decrease in error suggesting that the right hand side of equation 132 should be minimised. Writing the right hand side as:

$$Q = -\sum_{d=1}^{D} \mu_d \sum_{j=1}^{J} p_{old}(j \mid s_d) \left(\ln \left| p_{new}(s_d \mid j) p_{new}(j) \right| - \ln \left| p_{old}(j \mid s_d) p_{old}(s_d) \right| \right) \qquad 133$$

where the identity

$$\ln \frac{a}{b} = \ln a - \ln b \qquad 134$$

has been used, makes it possible to drop terms that do not affect the optimal values of the new parameters, that is, terms independent of $p_{new}(s_d \mid j)$ and $p_{new}(j)$:

$$Q = -\sum_{d=1}^{D} \mu_d \sum_{j=1}^{J} p_{old}(j \mid s_d) \ln \left| p_{new}(s_d \mid j) p_{new}(j) \right| \qquad 135$$

.

Specifically for a one dimensional Gaussian mixture model,

$$p_{new}(s_d \mid j) = \frac{1}{\sqrt{2\pi\sigma_j^{new^2}}} \exp\left(-\frac{(s_d - m_j^{new})^2}{2\sigma_j^{new^2}} \right), \qquad 136$$

which gives:

$$Q = -\sum_{d=1}^{D} \mu_d \sum_{j=1}^{J} p_{old}(j \mid s_d) \left\{ \ln p_{new}(j) - \tfrac{1}{2} D \ln(\sigma_j^{new})^2 - \frac{(s_d - m_j^{new})^2}{2\sigma_j^{new^2}} \right\} + const. \qquad 137$$

where, once again, terms independent of the parameters have been ignored. The new model parameters can now be found by minimising Q – a process most easily achieved by setting the derivative of Q with respect to each of the model parameters to zero. Thus, for the basis function variances:

$$\frac{\partial Q}{\partial (\sigma_j^{new})^2} = -\sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d) \left\{ \frac{1}{2} \frac{D}{(\sigma_j^{new})^2} + \frac{(s_d - m_j^{new})^2}{2(\sigma_j^{new})^2} \right\} \qquad 138$$

which, when set to zero and multiplied through by the variance gives:

$$0 = -\sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d) \left\{ \frac{1}{2} D(\sigma_j^{new})^2 + \frac{1}{2}(s_d - m_j^{new})^2 \right\} \qquad 139$$

such that:

$$\tfrac{1}{2} D(\sigma_j^{new})^2 \sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d) = \tfrac{1}{2} \sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d)(s_d - m_j^{new})^2 \qquad 140$$

which shows the optimal estimate of the new variance parameter of the j^{th} basis function to be:

$$(\sigma_j^{new})^2 = \frac{1}{D} \frac{\sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d)(s_d - m_j^{new})^2}{\sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d)} . \qquad 141$$

Repeating the same procedure for the basis function means:

$$\frac{\partial Q}{\partial m_j^{new}} = \sum_{d=1}^{D} \mu_d \, p_{old}(j \mid s_d) \frac{s_d - m_j^{new}}{(\sigma_j^{new})^2} \qquad 142$$

which, when set to zero gives:

$$m_j^{new} \frac{\sum_{d=1}^{D} \mu_d \, p_{old}(j \mid s_d)}{(\sigma_j^{new})^2} = \frac{\sum_{d=1}^{D} s_d \mu_d \, p_{old}(j \mid s_d)}{(\sigma_j^{new})^2} \qquad 143$$

and hence the optimal estimate for the new mean of the j^{th} basis function is:

$$m_j^{new} = \frac{\sum_{d=1}^{D} s_d \mu_d \, p_{old}(j \mid s_d)}{\sum_{d=1}^{D} \mu_d \, p_{old}(j \mid s_d)} \; . \qquad 144$$

A slightly different approach must be taken when determining the basis function priors, since they must be subject to the constraint:

$$\sum_{j=1}^{J} p_{new}(j) = 1 \qquad 145$$

This can be achieved using Lagrange multipliers by minimising the Lagrangian:

$$Q = -\sum_{d=1}^{D} \mu_d \, p_{old}(j \mid s_d) \ln p_{new}(j) + \lambda \left\{ \sum_{j=1}^{J} p_{new}(j) - 1 \right\} \qquad 146$$

where terms independent of $p_{new}(j)$ do not affect the solution and have been ignored. Thus,

$$\frac{\partial Q}{\partial p_{new}(j)} = -\sum_{d=1}^{D} \mu_d \frac{p_{old}(j \mid s_d)}{p_{new}(j)} + \lambda \qquad 147$$

which gives

$$\lambda = \sum_{d=1}^{D} \frac{\mu_d p_{old}(j \mid s_d)}{p_{new}(j)} \quad . \tag{148}$$

Since $p_{new}(j)$ is independent of n, it may be taken outside of the summation:

$$\lambda p_{new}(j) = \sum_{d=1}^{D} \mu_d p_{old}(j \mid s_d) \tag{149}$$

which may further be simplified by summing over the basis functions:

$$\lambda \sum_{j=1}^{J} p_{new}(j) = \sum_{d=1}^{D} \mu_d \sum_{j=1}^{J} p_{old}(j \mid s_d) \quad . \tag{151}$$

Since the basis function priors and posteriors both sum to unity, the Langrange multiplier can be seen to be:

$$\lambda = \sum_{d=1}^{D} \mu_d \quad . \tag{151}$$

The new estimates of the priors are found by substituting this value back into equation 147 and solving for $p_{new}(j)$:

$$\frac{\partial Q}{\partial p_{new}(j)} = -\sum_{d=1}^{D} \mu_d \frac{p_{old}(j \mid s_d)}{p_{new}(j)} + \sum_{d=1}^{D} \mu_d \tag{152}$$

so that when the derivative is zero, the new estimates of the priors are:

$$p_{new}(j) = \frac{\sum_{d=1}^{D} t_d p_{old}(j \mid s_d)}{\sum_{d=1}^{D} t_d} \quad . \tag{153}$$

It is thus possible to derive a form of the EM algorithm for updating the parameters of a Gaussian mixture model density estimator that has exactly the same form as the algorithm for finding the parameters of the fuzzy classifier used in [Wang:90]. This provides a rigorous basis for the algorithm, suggests a probabilistic interpretation of its

operation, and clearly indicates ways in which its flexibility and performance may be improved.

13. Appendix B: Analytical Convolution of Ground Cover with a Gaussian Model of the Sensor PSF

This section describes in detail how the spectral signature of a pixel is calculated when land cover is either concentrated in a circular region at the pixel's centre, or as a ring on the pixel's perimeter. To achieve this, the spectral signatures of individual sub-pixel points are convolved with the sensor PSF. In the restricted case considered here, this may be achieved by solving the following integral:

$$I = \int_a^b re^{-\alpha r^2} dr \,. \tag{154}$$

Using the substitution

$$u = -\alpha r^2 \qquad \frac{du}{dr} = -2\alpha r \tag{155}$$

such that:

$$dr \longrightarrow -\frac{1}{2\alpha r} du \,. \tag{156}$$

It is also necessary to calculate new values for the limits, since the integration is now over u rather than r. Thus,

$$r = a \qquad \Rightarrow \qquad u = -\alpha a^2 \,, \tag{157}$$

and

$$r = b \qquad \Rightarrow \qquad u = -\alpha b^2 \,, \tag{158}$$

such that,

$$\int_a^b re^{-\alpha r^2} dr = -\frac{1}{2\alpha} \int_{-\alpha a^2}^{-\alpha b^2} e^{-u} du \tag{159}$$

which, when integrated gives:

$$-\frac{1}{2\alpha}\left[e^{u}\right]_{-\alpha a^2}^{-\alpha b^2} = -\frac{1}{2\alpha}\left\{e^{-\alpha b^2} - e^{-\alpha a^2}\right\}. \tag{160}$$

Thus, when the cover type of interest is at the pixel centre,

$$a = 0$$

$$b = r'$$

then

$$I = -\frac{1}{2\alpha}\left\{1 - e^{-\alpha r'^2}\right\} \qquad 161$$

and when it is on the pixel perimeter,

$$a = r$$

$$b = 1$$

then

$$I = -\frac{1}{2\alpha}\left\{e^{-\alpha r^2} - e^{-\alpha}\right\}. \qquad 162$$

14. Appendix C: Bounding the Variance of an Area Proportion Distribution by a Function of its Mean

This appendix shows that the variance of an area proportion distribution with mean μ_{mean} is always less than or equal to $\mu_{mean}*(1-\mu_{mean})$. The highest possible variance area proportion distribution has exactly half of its probability mass located at μ=0 and the other half at μ=1 as shown in figure 102.

Figure 102: The highest variance area proportion distribution.

The mean of this distribution is computed using:

$$\mu_{mean} = 0 \times \mu_0 + 1 \times \mu_1 \qquad 163$$

assuming that there is μ_0 probability mass at μ=0 and μ_1 at μ=1 and gives:

$$\mu_{mean} = \mu_1 \qquad 164$$

The variance of a distribution is computed using:

$$\mu_{var} = \int (\mu - \mu_{mean})^2 d\mu \qquad 165$$

which, for this specific distribution simplifies to:

$$\mu_{\text{var}} = \mu_0(0-\mu_{mean})^2 + \mu_1(1-\mu_{mean})^2 \qquad 166$$

but, since $\mu_1=\mu_{mean}$ and thus $\mu_0=1-\mu_{mean}$,

$$\mu_{\text{var}} = (1-\mu_{mean})\mu_{mean}^2 + (1-\mu_{mean})^2\mu_{mean} \qquad 167$$

which simplifies to:

$$\mu_{\text{var}} = \mu_{mean}(1-\mu_{mean}) \qquad 168$$

which has a maximum values of 0.25 when $\mu_{mean}=0.5$. Thus, the variances of area proportion distributions are bounded by a simple quadratic function of their means.

15. Appendix D: Finding the Optimum Weighting in a Linear Combination of Two Sources of Proportion Information

The expected squared error of a simple proportion estimator that combined proportion information from two different sources, with distributions characterised by means μ_1 and μ_2, and variances ${\sigma_1}^2$ and ${\sigma_1}^2$ was shown to be

$$E = \left(m\mu_1 + (1-m)\mu_2 - \mu_{opt}\right)^2 + m^2\sigma_1^2 + (1-m)^2\sigma_2^2 \qquad 169$$

For known means and variances, the optimal estimator can be found by minimising E with respect to the amount of weight, m, given to each source. Taking the derivative of E with respect to m gives:

$$\frac{\partial E}{\partial m} = 2(m\mu_1 + (1-m)\mu_2 - \mu_{opt})(\mu_1 - \mu_2) + 2m\sigma_1^2 - 2(1-m)\sigma_2^2 \qquad 170$$

Multiplying out and collecting terms in m gives:

$$\frac{\partial E}{\partial m} = 2m(\mu_1 - \mu_2)^2 + 2\mu_2(\mu_1 - \mu_2) - 2\mu_{opt}(\mu_1 - \mu_2) + 2m(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 \qquad 171$$

Substituting for μ_{opt}:

$$\frac{\partial E}{\partial m} = 2m(\mu_1 - \mu_2)^2 + 2\mu_2(\mu_1 - \mu_2) - 2\frac{1}{2}(\mu_1 + \mu_2)(\mu_1 - \mu_2) + 2m(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 \qquad 172$$

and multiplying out gives:

$$\frac{\partial E}{\partial m} = 2m(\mu_1 - \mu_2)^2 + 2\mu_2(\mu_1 - \mu_2) - \mu_1(\mu_1 - \mu_2) - \mu_2(\mu_1 - \mu_2) + 2m(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 \qquad 173$$

which results in a cancellation:

$$\frac{\partial E}{\partial m} = 2m(\mu_1 - \mu_2) + \mu_2(\mu_1 - \mu_2) - \mu_1(\mu_1 - \mu_2) + 2m(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 \qquad 174$$

and the following simplification:

$$\frac{\partial E}{\partial m} = 2m(\mu_1 - \mu_2)^2 - (\mu_1 - \mu_2)^2 + 2m(\sigma_1^2 + \sigma_2^2) - 2\sigma_2^2 \qquad 175$$

The minimum of E with respect to m is found by setting the derivative to zero, and collecting terms in m on the left hand side:

$$2m(\mu_1 - \mu_2)^2 + 2m(\sigma_1^2 + \sigma_2^2) = (\mu_1 - \mu_2)^2 + 2\sigma_2^2 \qquad 176$$

Isolating m and dividing throughout by 2 yields the optimum value:

$$m = \frac{\frac{1}{2}(\mu_1 - \mu_2)^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2)} \qquad 177$$

16. References

[Atkinson:99] P. M. Atkinson. Assessing accuracy in fuzzy land cover maps. *Proceedings of the 25th Annual Conference and Exhibition of the Remote Sensing Society.* pp79-86.

[Axelsson:96] O. Axelsson. *Iterative Solution Methods,* Cambridge University Press 1996.

[Barbosa:94] V. C. Barbosa, R. J. Machado, F. S. Liporace. A neural system for deforestation monitoring on Landsat images of the Amazon region. *International Journal of Approximate Reasoning* 11, pp321-359, 1994.

[Bastin:97] L. Bastin. Comparison of fuzzy c-means classification, linear mixture modelling and MLC probabilities as tools for unmixing coarse pixels. *International Journal of Remote Sensing* 18, pp3629-3648, 1997.

[Baum:87] E. B. Baum. Supervised Learning of Probability Distributions by Neural Networks. *Neural Information Processing Systems,* pp52-61, 1987.

[Bezdek:84] J. C. Bezdek, R. Ehrlich, W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences* 10, pp191-203, 1984.

[Bishop:94] C. M. Bishop. Mixture Density Networks. *Technical Report* NCRG/94/004. Neural Computing Research Group, Aston University. http://www.ncrg.aston.ac.uk/

[Bishop:95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[Bosdogianni:97] P. Bosdogianni, and M. Petrou. Mixture Models with Higher Order Moments. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 35, No. 2, pp341-353, 1997.

[Brown:00] M. Brown, H. G. Lewis, S. R. Gunn. Linear spectral mixture models and support vector machines for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 38, No. 5, pp2346-2360, 2000.

[Chittineni:81] F. Canters. Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 4, pp403-414, 1997.

[Chittineni:81] C. B. Chittineni. Estimation of proportions in mixed pixels through their region characterisation. *1981 Machine Processing of Remotely Sensed Data Symposium*, Vol. 63, No. 4, pp403-414, 1997.

[Cihlar:00] J. Cihlar. Land cover mapping of large areas from satellites: status and research priorities. *International Journal of Remote Sensing*, Vol. 21, No. 6 & 7, pp1093-1114, 2000.

[Cox:46] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, Vol. 14, No. 1, pp1-13, 1946.

[Cracknell:98] A. P. Cracknell. Synergy in remote sensing - what's in a pixel? *International Journal of Remote Sensing,* Vol. 19, No. 11, pp2025-2047, 1998.

[DeGroot:89] M. H. DeGroot. *Probability and Statistics*. Addison-Wesley Publishing Company, 1989.

[Erol:00] H. Erol. A practical method for constructing the mixture model for a spectral class. *International Journal of Remote Sensing,* Vol. 21, No. 4, pp823-830, 2000.

[Fisher:90] P. F. Fisher and S. Pathirana. The Evaluation of Fuzzy Membership of Land Cover Classes in the Suburban Zone. *Remote Sensing of Environment*. pp121-132, 1990.

[Fisher:97] P. Fisher. The pixel: a snare and a delusion. *International Journal of Remote Sensing* Vol. 18, No. 3, pp679-685, 1997.

[Foody:92] G. M. Foody. A Fuzzy Sets Approach to the Representation of Vegetation Continua from Remotely Sensed Data: An Example from Lowland Heath. *Photogrammetric Engineering and Remote Sensing*, Vol. 58, No. 2, pp221-225, 1992.

[Foody:95] G. M. Foody. Fully fuzzy supervised image classification. *Proceedings of the 21st Annual Conference of the Remote Sensing Society*, pp1187-1194, 1995.

[Foody:95b] G. M. Foody. Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 50, No. 5, pp2-12, 1995.

[Foody:96] G. M. Foody. Fuzzy modelling of vegetation from remotely sensed imagery. *Ecological Modelling* 85, pp3-12, 1996.

[Foody:96b] G. M. Foody. Relating the Land-Cover Composition of Mixed Pixels to Artificial Neural Network Classification Output. *Photogrammetric Engineering and Remote Sensing*. Vol. 62, No. 5, pp491-499, 1996.

[Foody:96c] G. M. Foody Approaches for the production and evaluation of fuzzy land cover classifications from remotely sensed data. *International Journal of Remote Sensing*, Vol. 17 No. 7, pp1317-1340, 1996.

[Foody:97] G. M. Foody. Land cover mapping from remotely sensed data with a neural network: Accommodating fuzziness. In I. Kanellopoulos, G. G. Wilkinson, F. Roli and J.Austin, editors, *Neural-computation in Remote Sensing Data Analysis*. Spinger-Verlag, 1997.

[Foody:97b] G. M. Foody. Fuzzy thematic mapping: Incorporating mixed pixels in the training, allocation and testing stages of supervised image classification. In I. Kanellopoulos, G. G. Wilkinson, F. Roli and J.Austin, editors, *Neurocomputation in Remote Sensing Data Analysis*. Spinger-Verlag, 1997.

[Friedl:00] M. A. Friedl, C. Woodcock, S. Gopal, D. Muchoney, A. H. Strahler and C. Barker-Schaaf A note on procedures used for accuracy assessment in land cover maps derived from AVHRR data. *International Journal of Remote Sensing*, Vol. 21 No. 5, pp1073-1077, 2000.

[Gill:93] P. E. Gill, W. Murray and M. H. Wright. *Practical Optimisation.* Academic Press Limited, 1993.

[Gorte:98] B. Gorte and A. Stein. Bayesian classification and class area estimation of satellite images using stratification. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 36, No. 3, pp803-812, 1998.

[Haykin:94] S. Haykin. *Neural Networks: A Comprehensive Foundation.* Macmillan Publishing Company, 1994.

[Horwitz:71] H. M. Horwitz, R. F. Nalepka, P. S. Hyde, J. P. Morgenstern. Estimating the Proportions of Objects within a Single Resolution Element of a Multispectral Scanner. pp1307-1320, 1971

[Huguenin:97] R. L. Huguenin, M. A. Karaska, D. V. Blaricom, J. R. Jensen. Subpixel classification of bald cypress and tupelo gum trees in thematic mapper imagery. *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 6, pp717-725, 1997.

[Jervis:93] T. T. Jervis and W. J. Fitzgerald. Optimisation schemes for neural networks. Technical Report CUED/F-INFENG/TR 144, Cambridge University, 1993.

[Justice:89] C. O. Justice, B. L. Markham, J. R. G. Townshend, R. L. Kennard. Spatial degradation of satellite data. *International Journal of Remote Sensing*, Vol. 10, No. 9, pp1539-1561, 1989.

[Kent:88] J. T. Kent, K. V. Mardia. Spatial Classification Using Fuzzy Membership Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 8, pp659-671, 1988.

[Kirsch:96] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems.* Springer-Verlag 1996.

[Kitamoto:99] A. Kitamoto and M. Takagi. Image classification using probabilistic models that reflect the internal structure of mixels. *Pattern Analysis and Applications*, Vol. 2, pp31-43, 1999.

[Klir:95] G. J. Klir, B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall. 1995.

[Manslow:00] J. Manslow, M. Brown and M. Nixon. On the Probabilistic Interpretation of Area Based Fuzzy Land Cover Mixing Proportions. *Artificial Neuronal Networks: Application to Ecology and Evolution*, Springer-Verlag, pp81-95, 2000.

[Manslow:00b] J. Manslow and M. Nixon. On the representation of land cover information extracted from remotely sensed imagery. *Proceedings of 26th Annual Conference of the Remote Sensing Society: Adding value to remotely sensed data.*

[Maselli:96] F. Maselli, A. Rudolfi and C. Conese. Fuzzy classification of spatially degraded thematic mapper data for the estimation of sub-pixel components. *International Journal of Remote Sensing*, Vol. 17, No. 3, pp537-551, 1996.

[Melgani:00] F. Melgani, B. A. R. Al Hashemy and S. M. R. Taha. An eplicit fuzzy supervised classification method for multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensin*, Vol. 38, No. 1, pp287-295, 2000.

[Paola:95] J. D. Paola and R. A. Schowengerdt. A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery. *International Journal of Remote Sensing*, Vol. 16, No. 16, pp3033-3058, 1995.

[Parker:85] D. B. Parker. Learning Logic. *Technical Report* TR-47, Cambridge, MA: MIT Center for Research in Computational Economics and Management Science, 1985.

[Peters:76] C. Peters and W. A. Coberly. The numerical evaluation of the maximum-likelihood estimate of mixture proportions. Commun. Statist.-Theor. Meth. Vol. 12, No. 5, pp1127-1135, 1976.

[Ripley:96] B. D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[Robinson:85] V. B. Robinson. Fuzzy set theory applied to the mixed pixel problem of multispectral land cover databases. *Geographic Information Systems in Government.* pp781-886, 1985.

[Rumelhart:86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error backpropagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* pp318-362. Cambridge, MA. MIT Press.

[Schurmann:96] J. Schurmann. *Pattern classification: A unified view of statistical and neural approaches.* John Wiley & Sons, Inc. 1996.

[Sohn:97] Y. Sohn, R. M. McCoy. Mapping desert shrub rangeland using spectral unmixing and modeling spectral mixtures with TM data. *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 6, pp707-716, 1997.

[Shen:92] S. S. Shen and B. D. Horblit.. Multispectral image classification using a mixture density model. *Neural and Statistical Methods in Image and Signal Processing*, Vol. 1766, pp270--279, 1992.

[Shewchuk:94] J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. *Technical Report*, 1994. http://www.cs.cmu.edu/~jrs/jrspapers.html.

[Smits:00] P. C. Smits, S. G. Dellepiane, R. A. Schowengerdt. Quality assessment of image classification algorithms for land-cover mapping: a review and a proposal for a cost-based approach. *International Journal of Remote Sensing*, Vol. 20, No. 8, pp1461-1486, 2000.

[Cid-Sueiro:00] Y. Sohn, R. M. McCoy. Mapping desert shrub rangeland using spectral unmixing and modeling spectral mixtures with TM data. *Photogrammetric Engineering and Remote Sensing*, Vol. 63, No. 6, pp707-716, 1997.

[Thomas:96] G. Thomas, S. E. Hobbes, M. Dufour. Woodland area estimation by spectral mixing: applying a goodness-of-fit solution method. *International Journal of Remote Sensing,* Vol. 17, No. 2, pp291-301, 1996.

[Townshend:00] J. R. G. Townshend, C. Huang, S. N. V. Kalluri, R. S. Defries and S. Liang. Beware of per-pixel characterisation of land cover. *International Journal of Remote Sensing,* Vol. 21, No. 4, pp839-843, 2000.

[Tubbs76] J. D. Tubbs and W. A. Coberly. An empirical sensitivity study of mixture proportion estimators. *Commun. Statist.-Theor. Meth.* Vol. 5, No. 12, pp1115-1125, 1976.

[Wang:90] F. Wang. Improving remote sensing image analysis through fuzzy information representation. *Photogrammetric Engineering and Remote Sensing.* Vol. 56, No. 8, pp1163-1169, 1990.

[Wang:93] P. Wang. *The interpretation of fuzziness.* Technical Report, Center for Research on Cencepts and Cognition, Indiana University.

[Werbos:74] P. J. Werbos. *Beyond Regression: new tools for prediction and analysis in the behavioural sciences.* Ph.D. thesis, Harvard University, Boston, MA. 1974

[Wilkinson:96] G. G. Wilkinson. Classification Algorithms - Where Next? *Soft Computing in Remote Sensing Data Analysis.* World Scientific. Vol. 1, pp93-99, 1996.

[Wilkinson:97] G. G. Wilkinson. Open Questions in Neurocomputing for Earth Observation. *Neuro-computation in Remote Sensing Data Analysis.* Springer-Verlag. pp3-13. 1997.

[Wood:89] T. F. Wood, G. M. Foody. Analysis and representation of vegetation continua from Landsat Thematic Mapper data for lowland heaths. *International Journal of Remote Sensing*, Vol. 10, No. 1, pp181-191, 1989.

[Woodcock:00] C. E. Woodcock, S. Gopal. Fuzzy set theory and thematic maps: accuracy assessment an darea estimation. *International Journal of Geographical Information Science*, Vol. 14, No. 2, pp153-172, 2000.

[Zadeh:65] L. A. Zadeh. Fuzzy Sets. *Information and Control*, Vol. 8, No. 3, pp338-353, 1965.