HGCA

PROJECT REPORT No. 296

# DEVELOPING A COST-EFFECTIVE PROCEDURE FOR INVESTIGATING WITHIN-FIELD VARIATION OF SOIL CONDITIONS

JANUARY 2003

Price: £11.50

## PROJECT REPORT No. 296

# DEVELOPING A COST-EFFECTIVE PROCEDURE FOR INVESTIGATING WITHIN-FIELD VARIATION OF SOIL CONDITIONS

by

R M Lark[1], H C Wheeler[1], R I Bradley[2], T R Mayr[2] and P M R Dampney[3]

[1]Silsoe Research Institute, Wrest Park, Silsoe, Bedford MK45 4HS

[2]National Soil Resources Institute, Cranfield University, Silsoe, Bedford MK45 4DT

[3]ADAS Boxworth, Battlegate Road, Boxworth, Cambridge CB3 8NN

This is the final report of a 36 month project which started in April 1999 with a grant of £141,537 from HGCA (Project No. 2116).

The Home-Grown Cereals Authority (HGCA) has provided funding for this project but has not conducted the research or written this report. While the authors have worked on the best information available to them, neither HGCA nor the authors shall in any event be liable for any loss, damage or injury howsoever suffered directly or indirectly in relation to the report or the research on which it is based.

Reference herein to trade names and proprietary products without stating that they are protected does not imply that they may be regarded as unprotected and thus free for general use. No endorsement of named products is intended nor is any criticism implied of other alternative, but unnamed product.

# CONTENTS

# Abstract

The aim of this project was to develop a cost-effective procedure for investigating the variability of soils within fields as an aid to farm-level decisions on the adoption of variable rate management of inputs. The project was based on the premise that not all fields will merit variable rate management and so cost-effective investigation, by analysis of cheap and available data, should allow us to identify the fields with the largest potential for variable rate management before substantial resources are invested in soil sampling. It was hypothesized that information to this end could be extracted from past yield maps of the field, and that generalized information on the variability based on the soil parent material would also be indicative of the scope for variable management. It was also hypothesized that division of a field into zones, *within* which the season-to-season variation of crop yield is more or less uniform should identify the spatial structure of the key variations in soil properties which influence crop performance and so should aid in the identification and mapping of these properties.

Using data obtained by intensive and extensive sampling and examination of the soil at nearly forty fields, 34 fields (10 intensive and 24 extensive) funded in the current project, we were able to validate both hypotheses. We showed that it is possible to emulate expert assessment of the potential for variable rate management of a field (based on an agronomic interpretation of soil information) by applying a classification tree procedure to statistics (spatial and non-spatial) extracted from yield data of the particular field, and generalized assessments of the opportunity for variable management of different inputs over soils developed on different parent materials. The resulting classification tree is a decision support tool which could be used by individual farmers.

We also showed that in all fields at least some measured soil variables were significantly associated with the zonation of the field based on yield data. This allows us to identify the soil factors which may be relevant to variable rate management in the field and to obtain an initial impression of their spatial variation in the field. It is possible, for example, to produce a reasonable map of the principal soil types (e.g. soil series) in a field by using a classification tree, trained with a relatively small number of point observations of the soil, with continuous class memberships (derived in the identification of the within-field zones from yield data) used as predictors. Bulk (composite) soil samples from within each zone may also be tested to look for evidence of limiting soil concentrations of key nutrients. The precision of point predictions of these properties from the zones was generally poor, although not consistently better or worse than point predictions from non-intrusive soil sensors. When detailed information on the variation of a property is needed for management there is probably no substitute for moderately detailed sampling of the soil.

## Summary

### The problem.

The project was motivated by two problems which may constrain the uptake of variable rate management of inputs to cereal crops. First, a farmer considering the possibility of responding to spatial variability on some or all of his fields needs to decide whether the variability of his fields is such that variable management is likely to be economically worthwhile and technically feasible. Ideally the farmer will want to make this decision on the basis of the smallest costs possible. It would also be useful if the farmer could identify which fields are most likely to respond to variable management, as a focus for his initial effort. Second, when a farmer has decided to proceed beyond this initial decision to more detailed examination of the variability of his fields it is not clear how this can be done most cost effectively. Eventually it may be necessary to sample certain soil properties on an intensive grid, but this expense should only be incurred when we have evidence that the spatial variation of the particular soil property is of significance for management of the particular field. We require a protocol for cost-effective exploratory sampling which will deliver maximum information (e.g. about which soil properties appear to cause yield variation) and provide, ideally, a basis for mapping the variation in these properties.

In summary, the farmer should be able to identify fields where spatial variation is likely to justify variable management, and to make this assessment at minimum cost. It is then necessary to identify, with only limited sampling, those variable factors in the field which should influence decisions on inputs to crops. Only after this stage would very detailed soil sampling be planned, and this would be focused on those properties which we have good reason to believe are important.

### A note on terminology.

In this report we use the word *site* to denote a location *within* a field nominally at a point which can be given Cartesian co-ordinates, for example Eastings and Northings in the Ordnance Survey grid. A *yield data point* is a site in the field for which a yield value is recorded by a yield monitor. A *sample site* or *sample point* is a site in a field where the soil has been sampled. A *region* of a field is an area of the field, a set of sites (not necessarily all contiguous) which we could delineate by drawing boundaries. A *zone* or *management zone* is a region of a field which we have defined according to some criteria on the assumption that all sites within a zone are expected to be subject to similar constraints on crop performance and, therefore, might be managed in the same way.

**Hypotheses.**

1) The first hypothesis is pertinent to problem 1. In a pilot study for HGCA (*The development of cost-effective methods for analysing soil information to define crop management zones*, 1998. HGCA Project Report 170) we showed that fields differ in their variability with respect to the magnitude of variation (large or small differences between key properties) and the spatial scale of this variation. If the variation in practically important soil properties of a field is small, then the rewards of managing this by variable management will be small. If most of the variation occurs at fine spatial scales (i.e. over a few metres) then it is unlikely that it can be managed. Our hypothesis is that:

> *The scale and magnitude of variation in yield, as measured by yield maps, can act as a proxy for a detailed assessment of the variability of a field made by an expert, and provide a basis for identifying fields where variable management is likely to be feasible.*

If this hypothesis is supported then it should be possible to produce a decision support procedure which will use yield maps of a field to indicate how likely it is that variable management on this field will be feasible. Since yield-mapping equipment is now standard on many combine harvesters this is a relatively cheap source of information.

2) The second hypothesis is as follows.

> *If factors, potentially limiting on crop performance, vary within a field then the most limiting factor is likely to differ from one part of the field to another. Sites within the field where potentially limiting factors have similar values are likely to show a similar pattern of season-to-season variation in yield.*

This hypothesis has been studied previously and received empirical support. The goal here is to test it across a range of conditions and to assess its practical value. Regions within a field which exhibit contrasting patterns of season-to-season variation in yield may be identified by computer analysis of sequences of yield maps. Such regions may be suitable management zones for subsequent variable application of inputs. We tested the evidence that regions identified in this way differ with respect to soil properties of interest, and assessed the scope for using this analysis to predict the values of soil properties at un-sampled sites, using a relatively small calibration data set.

**Work to test these hypotheses.**

*1) Analysis of yield data.*

Yield maps were collected from farms representative of a wide range of conditions in Great Britain. Yield data from 16 farms were collected, but only 57% of the mapped fields were usable. This was largely because of data quality problems which have been reduced by improvements in the positioning systems used in yield-mapping, but also highlight the need for careful data management on farm. The yield maps for all fields were analyzed in three stages.

First, summary statistics of the raw data were compiled, and the data were screened for questionable values using the standard protocol in the AGCO Fieldstar system.

Second, the screened data were analyzed to identify distinct patterns of season-to-season yield variation. The strength of evidence for distinct patterns in the data is measured by the normalized classification entropy statistic (NCE). This analysis allows us to:

(i) subdivide the field into regions where the recorded yields most closely resemble a particular season-to-season pattern, the so-called "class of maximum membership". These regions are treated as potential management zones.

(ii) record for each location in the field a "membership" value on a scale of 0 to 1 which measures how closely the season-to-season yield variation at that site resembles each of the typical patterns which have been identified.

Third, the spatial variation of yield in each season was described statistically by computing a variogram. This measures how strongly yield at one site resembles yield at another as a function of the distance between the two sites. Using the variogram we may calculate:

(i) the amount of variation expected within a standard area of the field (a 1 ha square). This was expressed as the standard deviation (SD) and the coefficient of variation (CV) — respectively the root mean-squared difference between a point yield and the average yield (SD) and this value expressed as a proportion of the average yield (CV).

(ii) The ratio of the mean-squared difference from the mean yield in a 1 ha square to that in a 0.01ha square (Variance ratio, VR). A small VR (close to 1) this indicates that the variation is mainly over small distances.

Average values of these statistics were computed over all seasons for which yield data were available on each field.

*2) Field studies of the soil.*

A total of 34 fields were selected for study, from among those where yield data of adequate quality were available. Ten of these fields were studied intensively and 24 extensively.

The intensive fields were sampled at 50 or more sites on a loose grid. Topsoil and subsoil samples were collected. In addition the soil series was identified at each sample point. The samples were analyzed for particle size distribution, organic carbon and bulk density. At certain fields some additional analyses were conducted as part of Reading University's HGCA funded work on the same fields. In addition, intensively sampled measurements of some chemical and physical properties of the soil on fields with cereal yield maps were available from a previous study funded by the BBRO.

The extensive fields were sampled at fewer sites. Here the objective of field-work was to arrive at an assessment of the scale and magnitude of the variation in key soil properties. The resulting assessment, made by an experienced soil surveyor, along with some specific information (e.g. on the textural classes, available water capacity and soil series) was passed to a soil scientist with specific agronomic experience. He then assigned each field to one of five categories, ranging from 1 (no significant manageable variation) to 5 (substantial variation likely to be worth managing). We call this the potential for variable rate management (*PVRM*) rating. We also generated a simplified version of this rating with three categories:

| | |
|---|---|
| **PVRM 1**: | no potential for variable rate management, |
| **PVRM 2**: | potential may justify further investigation, |
| **PVRM 3**: | potential definitely justifies further investigation. |

Yield data were not used in this assessment. This assessment was also made for the intensively sampled fields, and for fields which had been studied previously in work funded by the HGCA. Assessments where therefore made for a total of 39 fields.

As well as this field-specific assessment of variation, the scope for managing different inputs by variable management was assessed according to the parent material of the soil. This was taken from a previous study by ADAS and NSRI for MAFF. Such an assessment of any field could be made from generally available information, e.g. from the soil association mapped on the 1:250,000 scale map of the soils of England and Wales. The original study for MAFF was provisional in nature, and based on rather few field observations, so our assessment here is a test of its usefulness.

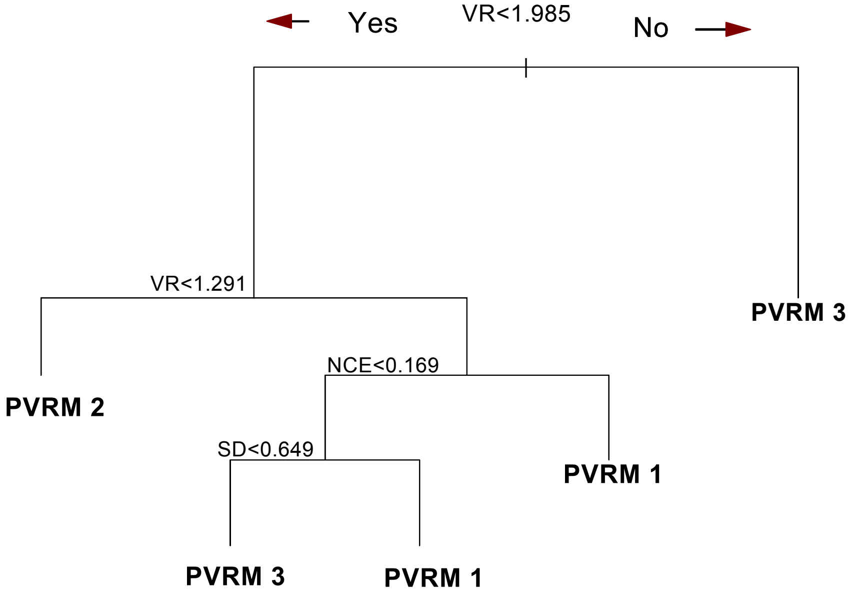
*3) Analysis of data.*

We showed that the assessments of the scope for variable management of fields based on soil parent material were significantly associated with the (reduced 3-category) field-specific PVRM assessments. This suggests that an initial assessment of the scope for variable management in a field can be made from this general information.

Analysis using classification trees was applied to test the first hypothesis. The statistics derived from the variograms of yield data (SD, CV and VR), and the NCE statistic from the analysis of the sequences of yield maps, were used as predictors (although the final tree did not use CV). We used these to derive trees to predict the 5 category PVRM rating and the reduced 3 category rating. The latter tree is shown below. At each node of the tree, working from the top down we branch left or right according to the value of one of these statistics. Each terminal node has a potential for variable rate management (PVRM) rating of maximum probability shown by the number at the end of the branch.

**Classification tree for predicting the potential for variable rate management rating from statistics extracted from yield map analyses**

**PVRM 1**: no potential for variable rate management,
**PVRM 2**: potential may justify further investigation,
**PVRM 3**: potential definitely justifies further investigation.



We also incorporated information from the soil parent material assessment of opportunity for variable rate management, along with the field specific statistics on yield variation into a classification tree. By cross-validation we were able to show that this tree correctly identifies the agronomic rating of a field from the yield data alone in 76 % of cases. This is substantially better than would be achieved by randomly allocating fields to categories, and is evidence in favour of our first hypothesis. The tree is also the key technology for making a prediction about the soil variability of a field. The full tree also indicates the uncertainty about the predicted PVRM rating at each node, and this information is contained in the full project report.

On the intensively sampled fields we conducted the following analyses.

(i) We computed the mean values of the soil properties within each potential management zone defined from the yield data, and tested statistically the evidence that the zones differ with respect to these properties.

(ii) We computed regressions of the soil properties on the membership values in the classes defined from the yield data. For those fields which were part of HGCA project 2243 we also computed regressions on the apparent electrical conductivity of the soil as measured by electromagnetic inductance (EM38 sensor).

(iii) We then used the results of the above analyses to compute the mean square error of a point prediction of each soil property using the mean value for the management zone or the regression equations.

(iv) At each sample point the soil series was known. Using classification trees we were able to show that the soil series can be predicted from the membership values in the classes derived from the yield data.

(v) At four fields, selected as test fields, we examined the effect of sampling with reduced intensity on the assessment of soil variability within a field, targeting this sampling using the management zones. It was found that a classification tree formed from the observed soil series at the reduced subset of test sites, using the class membership values as predictors, could be used to predict the soil series at un-sampled sites across the field, and that the resulting map constituted a good picture of the field's soil variability.

In all fields one or more soil properties differed significantly between the management zones and/or was significantly related to the membership values. This suggests that the zones do reflect underlying variability of the field and so may provide a basis for variable rate management of the fields. However, in most instances, prediction using the zone mean or the regressions was not very precise. This is because a good deal of soil variation remains unexplained, often the unexplained variation was very fine scales. Some of the properties best predicted with the yield data were the nutrients P and K, but the results did not consistently support the hypothesis advanced elsewhere that variation in these nutrients is driven by differential uptake (and so variable application can be based inversely on yield). In some cases the smaller concentrations of these nutrients were associated with high yielding regions (as the hypothesis suggests) but in other cases they were in low-yielding areas. However, these results do suggest that analysis of a single composite (bulked) soil sample from each management zone might identify whether these nutrients are anywhere limiting. The ECa data were not consistently better or worse than the memberships in the yield classes for prediction of soil physical properties.

## Conclusions.

The planned research to address the hypotheses outlined above has been successfully completed. Our key conclusions are as follows.

1) Our first hypothesis has been validated. Using a classification tree it is possible to allocate a PVRM rating to a field with some confidence using statistics extracted from past yield maps, and also (optionally) generalized assessments of the opportunity for variable rate management based on the soil parent material (this can be obtained given the soil association on which the field occurs in the 1:250,000 scale national soil map of England and Wales).

2) Our second hypothesis has also been validated. At least one of the measured soil variables was significantly different between the zones defined for yield data or was significantly associated with the membership values for the classes by which these zones are defined. This suggests that the zones reflect key soil variations in the field and so might be a basis for variable rate management of inputs to the field. They also provide a basis for investigation of the key soil variables within a field — e.g. by collecting a bulk (composite) soil sample for each zone and analysing it for soil nutrients (which were found to differ significantly between zones in several fields).

3) Detailed mapping of soil properties will require moderately detailed soil sampling, point predictions from the yield data alone have substantial error. It is interesting to note that electrical conductivity data on the soil measured with a sensor (EM38) are not consistently better or worse than yield data for predicting soil properties.

4) A reasonable map of the soil types within a field could be produced using a classification tree trained on data collected at half the intensity of the intensive sampling. This could be a useful tool for quick mapping and investigation of synoptic descriptions of soil variability at field scale.

To conclude, while all fields are variable, some fields are more variable than others. We have shown that a good approximation to an expert in-field assessment of the scope for managing a field with variable management can be obtained by appropriate analyses of yield maps along with readily-available information on the soil parent material.  This will enable farmers to make a rational decision whether to adopt variable management, and on which of their fields to explore this option.  Division of the field into management zones will help the next step - identification of the key factors determining variable crop performance in the field.  Some limited soil sampling, including bulk sampling of each zone to test for limiting or surplus concentrations of P and K, will then be necessary.  Using classification trees and regression equations, these limited observations can be used to generate field maps.  After this exploratory phase any properties which appear to be important may need to be sampled more intensively.

This project has successfully addressed its underlying hypotheses.  The next stage is to take the technologies which have been developed and to integrate them into a decision support system which is useable by farmers and their advisors.

# Chapter 1. Introduction

## 1.1 The problem

It is well known that some arable fields are internally variable, and that crop yields show comparable variation at these scales. It is possible that crop management could be improved by a response to this variability, perhaps by spatially variable management of inputs such as nutrients, herbicides, nematicides, fungicides, growth regulators, seeds and tillage. For example, recent research has shown that the economic optimum rate of nitrogen fertilizer can vary significantly and substantially within fields (Lark, 2002) and other research has suggested guidelines for managing this variability (Godwin *et al.*, 2002), presented as a decision tree by the HGCA (2002).

The farmer who wishes to implement these guidelines is immediately faced with the question "how much within-field variation do I have?" Subsequently, information on within-field variation of soil properties, as well as other variables, is required for many management decisions (although not all, weed management for example). The initial question is important if time and resources are to be directed to those fields where scope for precision farming is most likely. Later in the process it is likely that it will be costly to collect the requisite field-scale information (Pringle and McBratney, 1997). What is needed is a cost-effective process for assessing within-field variability in a step-wise manner, from the initial decision on whether the variability of the field is likely to merit management to the collection of detailed information on particular soil variables.

## 1.2 What has been done

In a previous study, funded by the HGCA (Project 0084/1/97 Report 171, see also Lark *et al*, 1999) we looked at the scope for using ancillary data sources, most notably yield maps, to improve the cost-effectiveness of investigation of soil variation within fields. This project provided the framework for the present project, so we summarize the key results. The general conclusions were as follows.

1) *The nature of spatial variability in soil properties differs substantially between fields*. Spatial variability has two key components - the overall **magnitude** of the variations, and the **spatial scale** at which this variability is expressed. These aspects have implications for the feasibility of precision farming within the fields.

**Magnitude**. Some fields investigated (most notably those with soils formed from chalky boulder clay) were comparatively uniform with respect to soil physical properties. Other fields showed substantial variation - often largely attributable to differences in depth to the underlying rock or to contrasting parent materials.

The potential benefits from a management response to spatial variation will be greatest where the variation is substantial.

**Spatial scale(s)**. In some of these fields variation was limited to, or dominated by, *short-range* components. By this we mean that most or all of the variation in the field was 'fine-grained' and could be seen over short distances. In other fields the variation was dominated by *long-range* components, and was seen in the difference between large and contrasting patches of soil. In the former case the variation is probably too intricate to be practically manageable, scope for precision farming will be greatest where variation is predominantly long-range.

Spatial analysis of yield maps, particularly for more than one season, identified differences between fields with respect to the magnitude and scale of spatial variation in yield which corresponded to differences identified (with greater effort) by field investigations of the soil. It was concluded that farmers should pay attention to what yield maps reveal about the magnitude and spatial scale of variation in their fields before deciding whether or not to invest in costly soil sampling within the fields.

2) *There is scope for improving the cost effectiveness of soil investigation within fields by incorporating information from analysis of time series of yield maps.*

The spatial variations of yield are not necessarily consistent from year to year. Nonetheless, it is often possible to recognise (automatically) subregions of a field which show broadly similar patterns of year to year variation (Lark and Stafford, 1997), and there is now evidence to support the hypothesis that, within these subregions, the crop may be subject to broadly the same limiting factors. The identification of such subregions should therefore be useful as a first step in identifying and mapping the factors which are critical in driving spatial variation of yield. These subregions may be used in practice as *management zones* and we use this term (or *zones*) in what follows.

Therefore, rather than immediately paying for a grid survey of a spectrum of soil variables, none or few of which may emerge to be important, the farmer and/or advisor could use the division of the field into zones using yield maps, as an initial stratification. This allows soil sampling to be targeted, both in terms of the location of sites and the selection of soil properties which are investigated. Such a procedure should allow the key sources of yield variation to be identified at minimum cost. More detailed mapping of some of these properties will then be needed, either across the field or within certain zones only.

3) *When it has been decided to map certain soil properties in more detail, ancillary data such as yield maps and information on topography are likely to be valuable.*

At one of the fields in our recent project we were able to show similarities between the spatial variation of available water capacity and yield in one season.  This implies that the yield data could be combined with limited soil information to produce a more accurate map of the soil property than could be generated from soil samples alone.  Similar techniques could be used to incorporate remote sensor data or data on the field's topography.

### 1.3  A proposed solution

The overall objective of project 2116 is **to develop a cost-effective decision-making strategy whereby the farmer considering the possibility of variable-rate application of inputs can plan field investigations of soil variation**.  The procedure which we propose is based on the findings of our previous project outlined above.  It has two key elements:

1)  Maximum use is made of data sources which are less costly than direct measurements of soil properties (e.g. yield maps) to improve the efficiency of soil investigations.

2)  Investigation of soil variation moves in a step-wise way, starting with general questions about the likely magnitude and scale of variation in the field, and only proceeding to more intensive and costly procedures (i.e. grid sampling) when and where there is good evidence that they will be justified.

Our proposed strategy starts with the assumption that the farmer has no quantitative information about spatial variation within his fields and must decide whether yield-mapping or remote sensor data is worth acquiring.  The next step is interpretation of such data (we focus on yield maps) in order to decide whether the soil variation within the field is likely to be substantial.  In the third step, a reconnaissance survey of the field is conducted to identify which soil properties should be mapped in detail, and where in the field this should be done.  This detailed mapping is the final stage.  Below we outline the four steps in more detail.

**STEP 1.**  *Question: is it worth investing in a yield-mapping system, or paying a contractor with such a system?*

This decision could be based on generalised information about local soils and physiography interpreted in the light of farmer/advisor experience.  One way of doing this would be to use the soil parent material classification derived for this purpose by the NSRI and ADAS (MAFF project CE0168/0166 *Towards identifying the soil types and associated cropping systems where pollution control and farm profitability benefit are likely from use of site specific crop inputs,* 1999.)  The local parent material class could be identified and a preliminary assessment of the likely within-field variability then used to decide on whether further investigation is worthwhile. For example, a farmer with predominantly "deep clay soils" might conclude that the scope for spatial management of soil variation is probably limited, while a farm with "mixed deep sand and clay soils" within individual fields might conclude that this variability is likely to be worth managing.

If the farmer decides to pursue the possibility of conducting site-specific management in response to soil variation, then yield data will be obtained. This takes us to the next step, in which individual fields are considered in turn.

**STEP 2.** *Question: is the magnitude and spatial scale of the variation within this field such that investigation of soil properties is justified?*

It is proposed that this decision is based on spatial analysis of yield maps for one or more seasons. Criteria will be developed for ranking fields with respect to both magnitude and spatial scale of their yield variation. By comparing specific fields to these criteria, it will be possible to judge whether further investigation of the field to quantify the variation of particular soil properties is a worthwhile step. If it is decided to investigate soil properties in a specific field then both sample sites and soil properties must be targeted. This takes us to the next step.

**STEP 3.** *Question: how can the field be partitioned into zones for investigation of the soil variation and, potentially, for management?*

In this step some carefully directed soil sampling is conducted in the field. This is planned using the zones obtained by analysis of the yield maps. Not necessarily all zones will be investigated in the same way. Some, where the characteristic season-to-season yield variation and/or farmer's knowledge is strongly indicative of one particular problem may not warrant measurement of soil properties which are not relevant to this.

The result of this step and the consequent field investigations will be a division of the field into zones with some soil information. Evidence will have been collected indicating which factors are most likely to limit crop yield within each region. Possible non-soil factors responsible for yield variation (e.g. weed patches) will also be noted during the reconnaissance. This may lead directly to management decisions in some regions (e.g. apply lime, drain or subsoil) - but it may be necessary to collect more detailed information to implement spatially variable management in other parts of the field. This leads to the next step.

**STEP 4.** *Action: mapping of specific soil properties.*

This can be done most cost effectively by exploiting relationships between the soil property and crop yield, topography or remote sensor measurements. The previous step in the procedure may indicate which of these 'ancillary' data sets is most likely to be useful. Regressions of the soil properties on the ancillary data may be suitable to map these properties with adequate precision. Alternatively the mean value of the property within each management zone may be an adequate predictor.

### 1.4 Research to test and implement this solution

This proposed strategy entails a series of hypotheses which were investigated in this project. We analyzed series of yield maps from over 70 fields across the country, representing a wide range of soil and cropping conditions (Chapter 2). The yield data were analyzed using the classification and spatial analyses used in past work and alluded to above (Chapter 3).

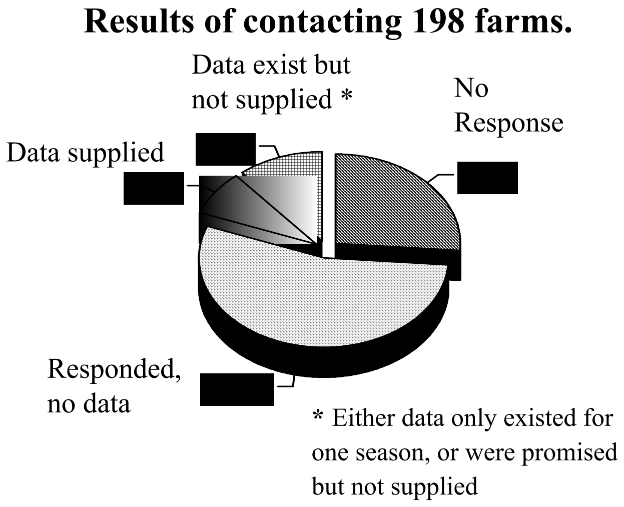
We then selected a set of 34 fields representative of a wide range of soil landscapes. Each of these fields was investigated on the ground by an experienced NSRI soil surveyor, who produced a detailed report on the observed variation (Chapter 4, section 4.3). From this information ADAS made an assessment of the likelihood that the field would benefit from variable rate application of inputs (Chapter 4, section 4). This information was then used to test two hypotheses. First that the parent material class is an indicator of the likelihood that a particular field shows manageable within-field variability (Chapter 4, section 5) and second, that the within-field variability may be predicted from analysis of yield maps of the field (Chapter 4, section 6).

Ten fields were then selected for more intensive study, again covering a range of conditions. These fields were intensively sampled collecting soil material for analysis to determine physical properties and organic carbon content (Chapter 5). In addition, comparable data from fields with cereal yield maps studied by Silsoe Research Institute, British Sugar and IACR Broom's Barn in BBRO funded research were used. These data were then analyzed to test the hypotheses that (i) zones delineated by analysis of yield data differ significantly with respect to important soil properties and (ii) regressions on yield data or variables derived from these and/or zone mean values can provide useful predictions of the variation of these soil properties across the field (Chapter 5, section 4). At two test fields a comparable analysis was undertaken. However, in addition, the zones defined from analysis of the yields maps of these fields were considered and an assessment was then made of how useful this information would be for planning efficient sampling of the field (Chapter 6).

## Chapter 2. Collection of data

We obtained customer lists from AGCO (formerly Massey Fergusson) and RDS, detailing clients who, at the time, had had yield-mapping equipment for at least 3 years. In addition to these contacts, questionnaires were sent to clients of NSRI for whom detailed farm maps had been produced in the past. In total, contact details for 251 farms were compiled in an ACCESS database. There was some overlap in names and 198 farmers were contacted by NSRI with a request for co-operation in the project. Subsequently, 16 farmers supplied data to the project. The overall response to the questionnaire is summarized in Figure 2.1 below.

**Figure 2.1 Responses to questionnaires requesting yield data.**



* Either data only existed for one season, or were promised but not supplied

It is clear from the results of this data collection exercise, and from anecdotal evidence in subsequent discussions with farmers, that this low return reflects some important factors. First, an unexpectedly small proportion of former customers of the NSRI use yield-mapping. It is possible that they have found that the soil map answers some of their questions on soil variability, or that they have decided, perhaps on the basis of the soil map, that they are not interested in considering variable rate management. Second, not all farmers have retained yield data in a useable (electronic) form. This is due in part to the problems of managing and storing large volumes of data. Third, there is some scepticism about yield-mapping with respect both to its perceived benefits, and to operational problems. Not all farmers who own the equipment are using it when harvesting. Having said that, in a separate survey of all AGCO customers conducted during the period of this project Pedersen *et al* (2001) found that 58% believed that yield-mapping was likely to have the greatest potential benefit for their farm of all 'information gathering' precision farming activities. This was the largest score of all technologies (the next largest score was 28% for grid soil sampling). This compared to 46% of producers in Nebraska and 28% in Denmark who rated yield-mapping as most important.

A particular concern is that 43% of all the data sets received were not useable. That is to say, there were not yield data of reasonably complete coverage for at least three seasons. The yield maps of many fields had substantial gaps with missing data (due to system problems, errors in positioning or where two or more combines were used in the field and not all were equipped for yield-mapping). In these cases, when the data are overlaid prior to analysis, the resulting data set is too sparse to give useful results. Some of these problems have been removed or reduced by recent improvements in GPS technology.

Data from over 70 fields were analysed, as described in Chapter 3. We describe in Chapter 4 and 5 how fields were selected for inspection (intensive or extensive) of soil variation on the ground. Despite the relatively poor response to our questionnaire it was possible to select fields which represented a wide range of soil types used for arable production across the country. Figure 2.2 below shows the places where soil data were collected, superimposed on a map of land in arable use (60-90% or more than 90% of the land area) in England in Wales. Figure 2.3 shows the distribution in England and Wales of soil types similar to those which occur at fields used in this project.

**Figure 2.2 Extent of arable land in England and Wales with fields where soil data was collected for this project.**

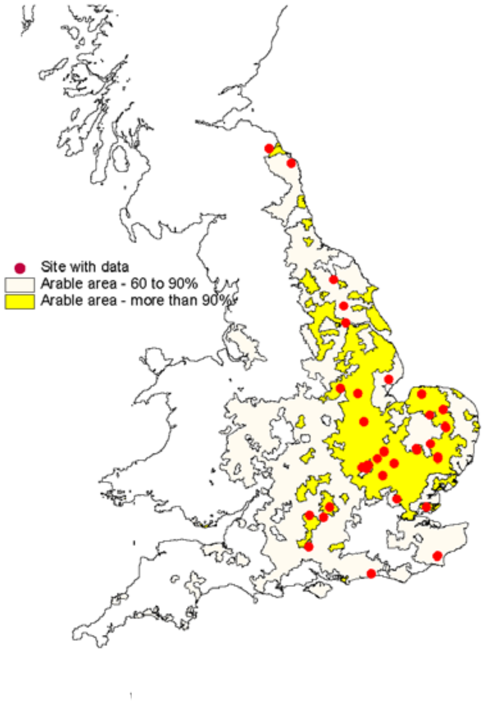
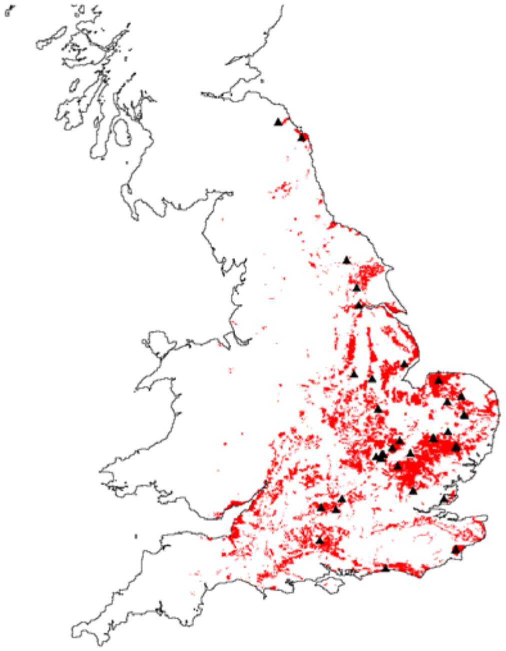


**Figure 2.3 Distribution in England and Wales of soils similar to those which occur in the fields used in this project.**



# Chapter 3. Basic Analysis of Yield Data

This chapter explains the analysis of yield data from all fields used in this project. We outline the basic procedures and supply references with full technical details. The basic principles and rationale of the analysis is summarized in the following paragraphs.

The objective of the classification method is to divide a field into zones *within* each of which there is a characteristic pattern of yield variation between the seasons. Thus, we might identify zones within which the yields are consistently high, or within which yield was high in all but one season. This is done, because zones within which broadly similar factors are limiting on yield are likely to show similar patterns of season to season variation in yield. Thus, if parts of a field are potentially droughty, then they should form a distinct zone characterised by low yields in drier years. Similarly, parts of the field with a significant nutrient deficiency may show consistently low yields. The classification thus simplifies all the complex information in a sequence of yield maps (typically 10,000 data points) to a few basic patterns.

A zone in the geographical space of a field is defined as a class of yield data points, i.e. as a set of data points at which the season-to-season yield patterns are similar. These classes are defined automatically by a method of cluster analysis. This generates as output the 'class centres' which represent the typical season-to-season fluctuation of yield for each class. Each yield data point will resemble each of these patterns to a greater or lesser degree, which is measured by a 'membership' value in each of the classes. A data point has 'maximum membership' in the class which it most closely resembles. One way of defining zones from yield data is to identify the class of maximum membership for each data point and to create a zone from all yield data points with maximum membership in the same class. This is how we have defined zones in this project, although we retain the membership values for each data point in all the classes for use in later analyses.

Usually the map showing class of maximum membership has a certain amount of short range 'speckle'. This detailed variability is unlikely to be of interest, and hinders interpretation. A smoothing method is applied to the map. This uses the 'multivariate variogram' which identifies the distance over which the important variations in yield occur, and maximises the smoothing effect at shorter distances.

## 3.1 Exploratory data analysis and editing.

The raw yield data comprised point observations of Global Positioning System (GPS) latitude/longitude, yield and system-specific information in system-specific formats. These were converted from latitude/longitude to Ordnance Survey co-ordinates. Data editing was done in line with the standard procedure used in the AGCO software for processing . This excludes all data where the nominal ground speed determined from GPS was greater than 10 metres per second since this suggests GPS error. The

AGCO software also excludes suspect yield data outside limits where the lower limit is two thirds the average yield and the upper limit is twice the average yield. This procedure was followed by default unless it led to the removal of substantial numbers of data and not just values associated with familiar sources of error (headlands and other turning points, and single passes with unusually low values suggesting a partly filled cutter bar). In these circumstances the standard statistical thresholding procedure of Tukey (1977) was used, and those data which are outside his 'outer fences' were excluded.

Sequences of yield maps for each field were overlain. In some cases, however, the correspondence of Ordnance Survey co-ordinates between successive seasons was not good. This may be due to GPS errors, because of changes in base station settings, for example. In these cases one of the yield maps was selected as a reference, after inspection of a field base map, and the offset between the reference and each of the remaining yield maps was estimated, on the assumption that it consisted of a simple shift in the eastings and northings. In all cases data coverage across the field was reasonably uniform, so the offset could be estimated by comparing the mean eastings and northings for the complete data set in each year.

The edited and spatially registered yield values were then overlain and yield estimates for each of the seasons were extracted for points across the field.

### 3.2 Continuous classification and filtering of the classes.

A detailed account of the continuous classification procedure and its scientific rationale has been given elsewhere (Lark, 2001; Lark and Stafford, 1997). We outline the key steps below.

Continuous classification was conducted on the yield data having first standardized the yield within each season to zero mean and unit variance. The classification was carried out to identify 2,3 …....8 classes. For a specified number of classes (*g*), the procedure searches the data for the *g* most distinct groups of observations in terms of standardised yield in the different seasons. Each yield data point (*I*) has a membership in each class (*j*) denoted $\mu_{I,j}$ which is subject to the following constraints:

$$\sum_{j=1}^{k} \mu_{i,j} = 1, \ \forall i \qquad (3.1)$$

$$0 \leq \ \mu_{i,j} \ \leq \ 1 \quad \forall i,j \qquad (3.2)$$

Thus observation *I* may have membership $\mu_{I,j} = 1.0$ in one class *j* (complete resemblance to class *j*) and therefore zero membership in all other classes to which it is held to have no resemblance, or it may have partial membership in two or more classes. This allows the classification to retain information on the essentially continuous nature of yield variations in space.

The centre of a class, *j*, defines its typical member and is, effectively, the set of mean (standardized) yield values for all the data weighted by their memberships in class *j*.

The normalised classification entropy, NCE, (see McBratney and Moore 1985), was then calculated. This is a measure of how distinct are the classes identified in each classification. Lark (2001) discusses this statistic, and the information which it gives in more detail. Our principal interest is to use it to decide how many classes should be used to describe the yield variation in a particular field.

Because the classification procedure can "stick" at locally optimum solutions, the classification was repeated several times for each number of classes and the one with the smallest value NCE was selected. This minimum NCE was then plotted against the number of classes, and a local minimum in the plot was identified. The corresponding classification was selected for further consideration. Where a distinct local minimum was not seen, the plots sometimes showed a distinct break at *g* classes such that NCE (*g* -1) > NCE(*g*) ≈ NCE (*g*+1) ≈ NCE (*g*+2). In this case *g* classes were selected. If the NCE plot declined exponentially with *g* there is weak evidence of distinct classes in the data and an arbitrary number of classes were selected (usually 4).

The map of class of maximum membership often shows a good deal of short range "speckle". This reflects short range variation which results in one site being placed in a different class to its neighbours. Short range variation is likely to reflect errors in the data or factors of variability which the farmer cannot hope to manage. This short-range variability was therefore removed by smoothing the membership values following Lark (1998). The membership values at each site were replaced by weighted averages of the membership values within different neighbourhoods of the site (after log-ratio transformation, needed because memberships are compositional data, constrained by Equation (3.1) to sum to a constant value, Aitchison, 1986). The smoothing weights are determined according to the multivariate variogram (Bourgault and Marcotte, 1991) of the standardised yield data such that distant points have lower weight. Neighbourhoods were defined by radii of 5, 10, 20, 30 and 40 metres and a smoothed map of maximum membership was obtained for each radius. Lark's (1998) coherence index was then calculated for each smoothed map and the map with a maximum index was selected. This index ensures that smoothing of local intricacy in the class map is maximised without creating larger scale changes in the relative frequency of the classes.

### 3.3 Spatial analysis of yield data.

The spatial variability of yield data for any season may be described by the variogram. This is based on the assumption that the yield, *y*, measured at a location, **x**, may be treated as a realisation of a random function, denoted by *Y*(**x**). The analysis is possible if the random function is intrinsic, that is if

$$\mathrm{E}\left[Y(\mathbf{x}) - Y(\mathbf{x}+\mathbf{h})\right] = 0, \tag{3.3}$$

where E[] denotes the statistical expectation, and

$$2\gamma(\mathbf{h}) = \mathrm{E}\left[\left\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\right\}^2\right] \tag{3.4}$$

depends only on the spatial separation or lag **h**. The function $\gamma(\mathbf{h})$ is the variogram. The variogram was estimated from yield data *y*(**x**) using the estimator

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(\mathbf{h})} \left\{ y(\mathbf{x}_i) - y(\mathbf{x}_i + h)\right\}^2 \quad , \tag{3.5}$$

where *N*(*h*) pairs of observations among the available data are separated by the scalar lag distance *h*. A model was then fitted to these estimates of the variogram by weighted least squares using the MVARIOGRAM procedure in Genstat (Harding and Webster, 1995).

### 3.4 Weather data.

Such weather data as are available were obtained for the nearest Meteorological Office station to each field (Table 3.1). In some cases only limited data were available from the closest station and a more distant one was used.

In each season for which yield data were available the length of the growing season was determined from the soil temperature data following Smith (1976). We then extracted the potential soil moisture deficit estimates (obtained by the Met. Office's standard procedure) from the start of the growing season to the end of July (only available from three weather stations), and mean monthly rainfall data for the same period (called the 'Spring-Summer rainfall' below). We also extracted mean monthly rainfall from the beginning of September to the end of the growing season in the autumn *preceding* each harvest for which yield data were available ('Autumn rainfall'). This was extracted because in studies on other fields we have found evidence that a wet autumn may result in low yields in parts of the field with heavy soils with poor drainage.

In the following section we present the class centres from the analysis of each of the intensively sampled fields, and the corresponding weather data:— mean (monthly) autumn and spring/summer rainfall prior to each harvest and the mean and maximum PSMD from March to July in each harvest and the May ('Mid-season') PSMD. Note that in the rainfall plots a solid line indicates the regional mean monthly rainfall for the spring-summer period and a dotted line the equivalent value for the preceding autumn (after Smith, 1976).

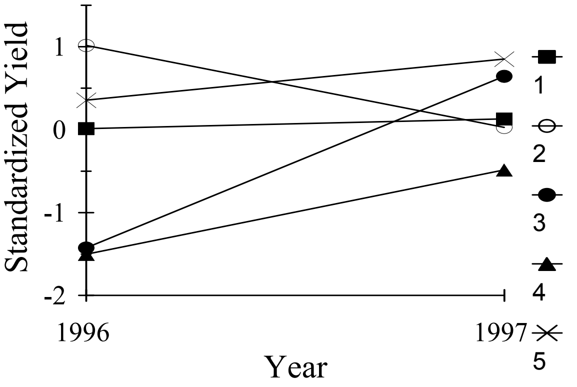
**Table 3.1 Weather Stations used.**

| Farm | Station |
|---|---|
| Shuttleworth | Silsoe |
| Chicksands | Silsoe |
| Houghton Conquest | Silsoe |
| Andover | Boscom |
| Cirencester | Oxford |
| Heydour | Cranwell |
| Boxworth | Cambridge NIAB |
| Whittlesford | Cambridge NIAB |
| Broom's Barn | Broom's Barn |
| Ixworth | Broom's Barn |
| Flawborough | Sutton |
| Crowmarsh Battle | Benson |
| Yokefleet | Hull Pearson's Park |

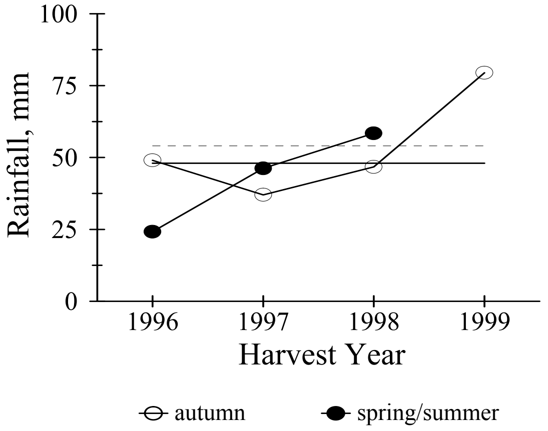
## 3.5 Class centres and weather data for intensively sampled fields.

### 3.5.1 Hall 8 Field, Ixworth.

**Class centres**

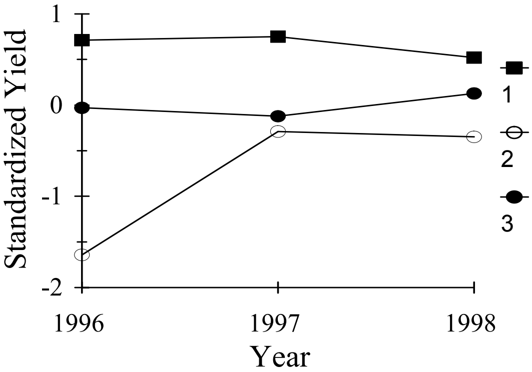


**Mean monthly rainfall**

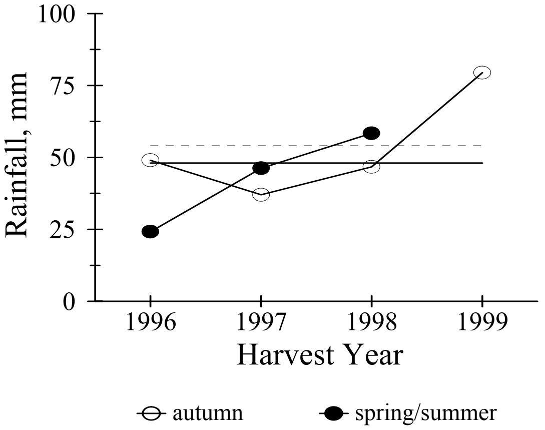


### 3.5.2 Little Lane Field, Broom's Barn.

**Class centres**

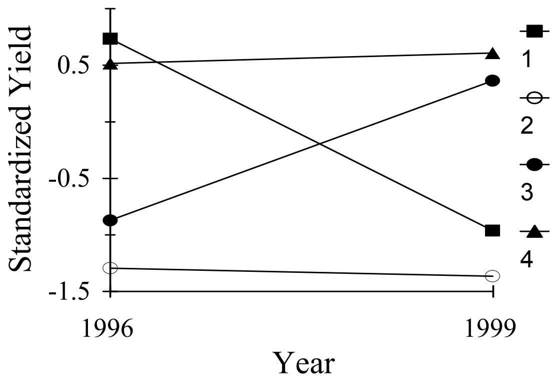


**Mean monthly rainfall**

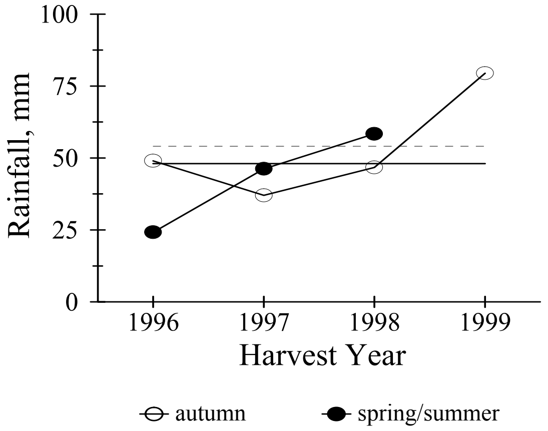


### 3.5.3 Brome Pin Field, Brooms Barn.

**Class centres**

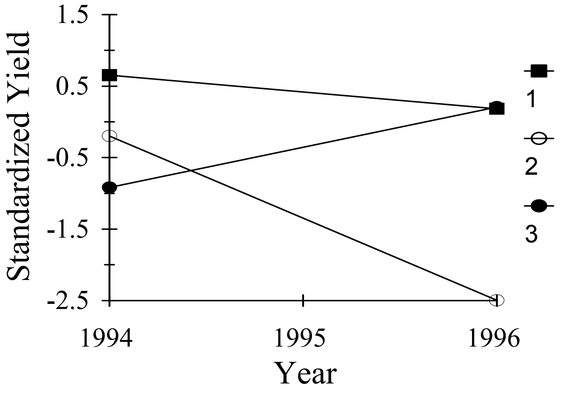


**Mean monthly rainfall**

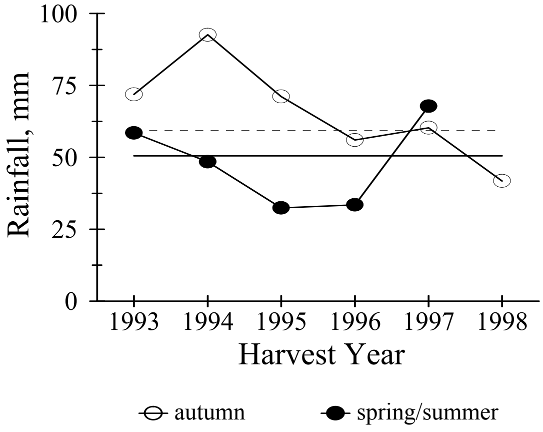


### 3.5.4 Commissioners Field, Yokefleet.

**Class centres**

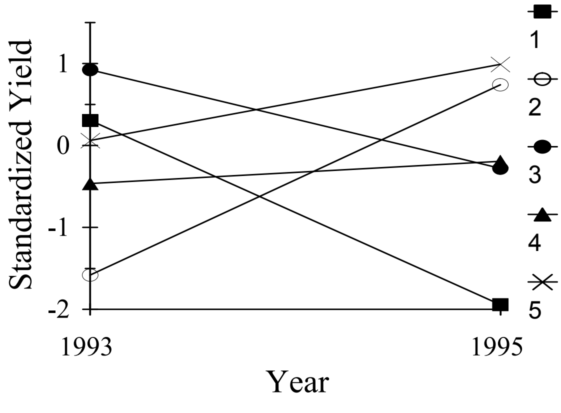


**Mean monthly rainfall**

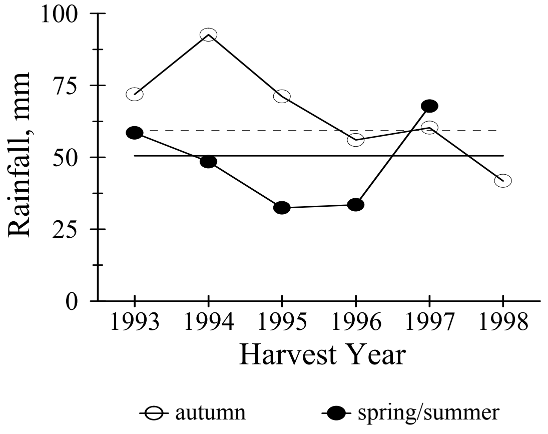


### 3.5.5 South Warpings Field, Yokefleet.

**Class centres**

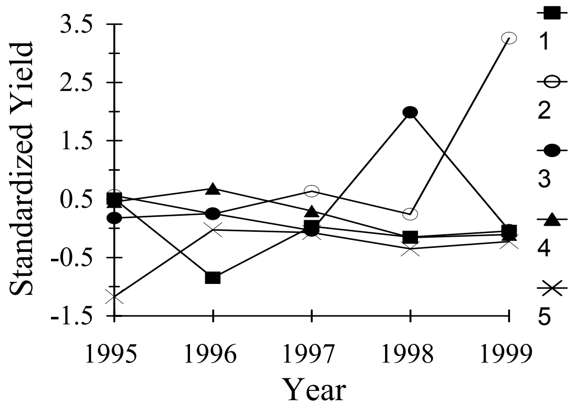


**Mean monthly rainfall**

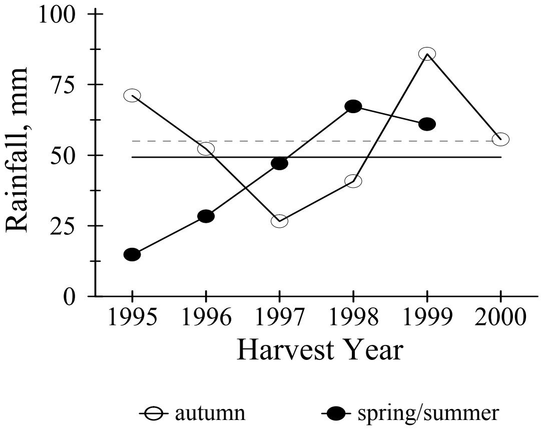


### 3.5.6 Field 2, Flawborough.

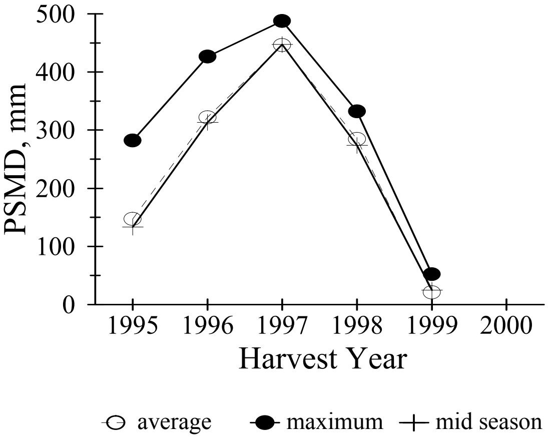
**Class centres**



**Mean monthly rainfall**

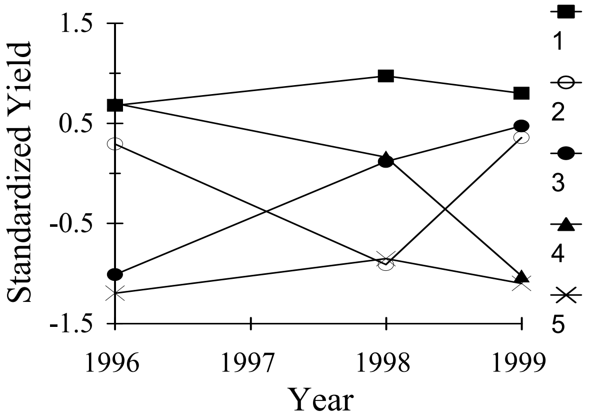


**Potential Soil Moisture Deficit**

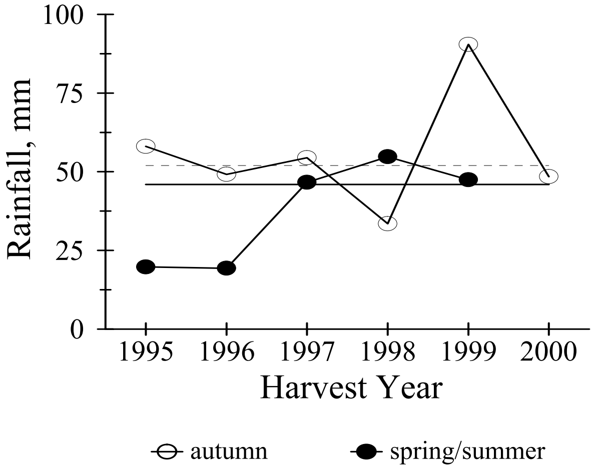


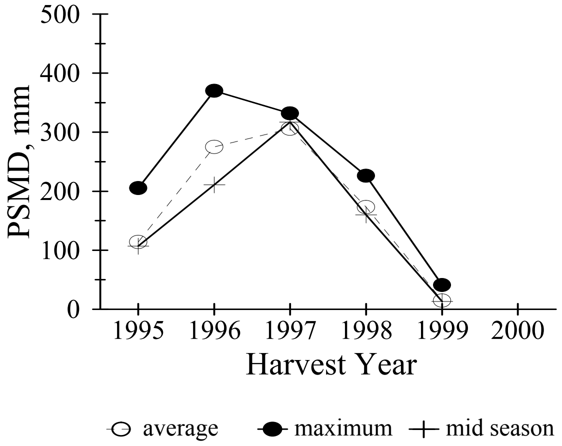
### 3.5.7 Gate Field, Whittlesford.

Class centres



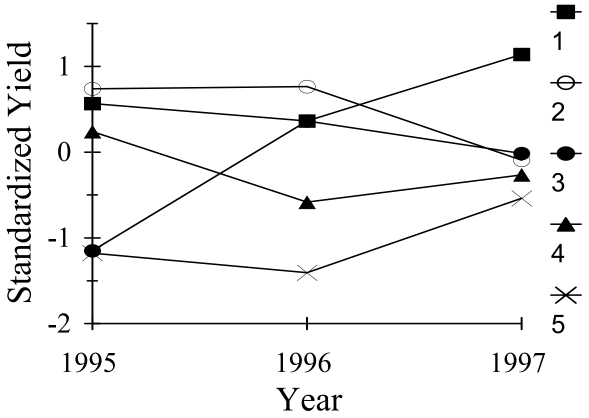
Mean monthly rainfall



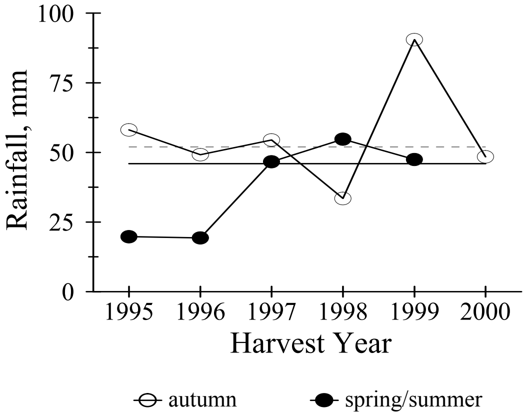


### 3.5.8 Top Pavements Field, Boxworth

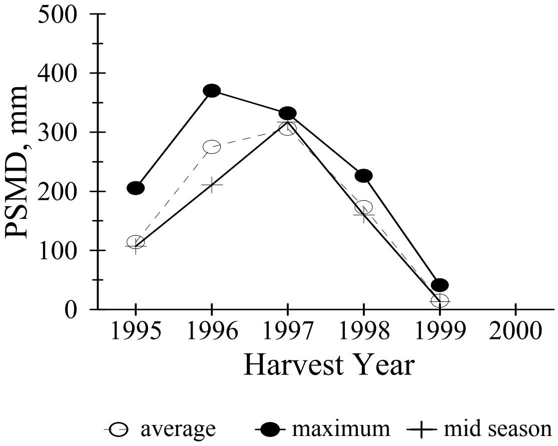
**Class centres**



**Mean monthly rainfall**

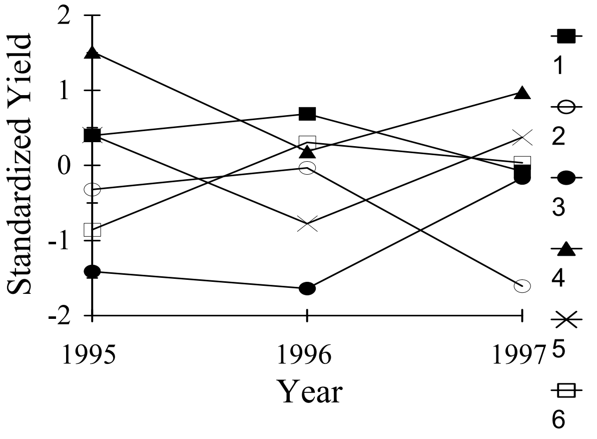


**Potential Soil Moisture Deficit**



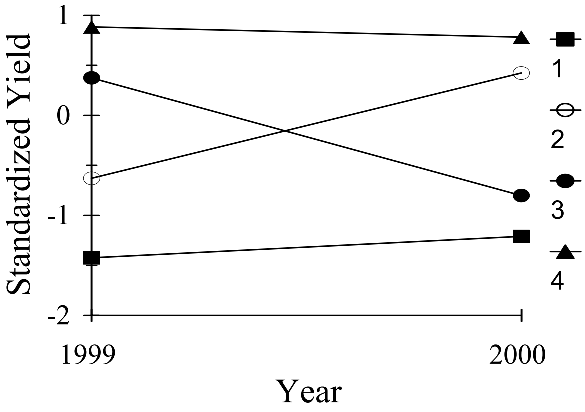
### 3.5.9 Knapwell Field, Boxworth.

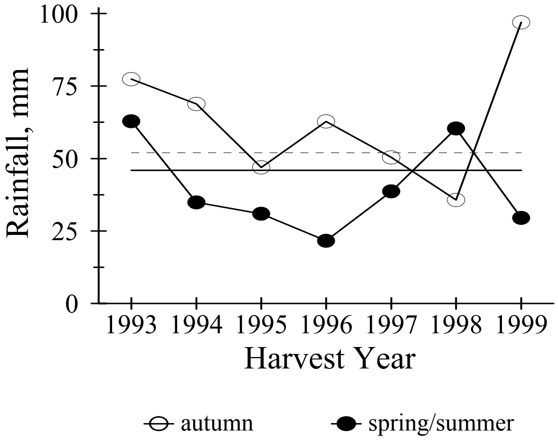
**Class centres** **Weather data as in 3.5.8**



### 3.5.10 Shagsby 4 Field, Chicksands.

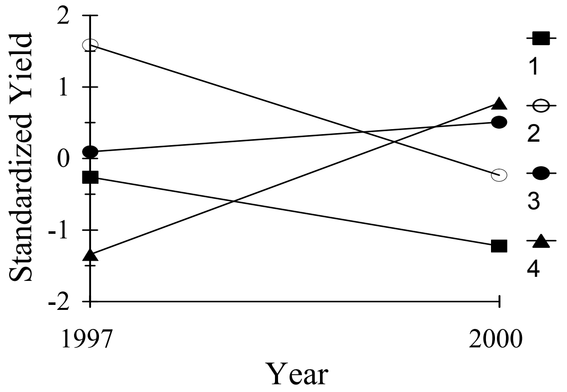
**Class centres** **Mean monthly rainfall**

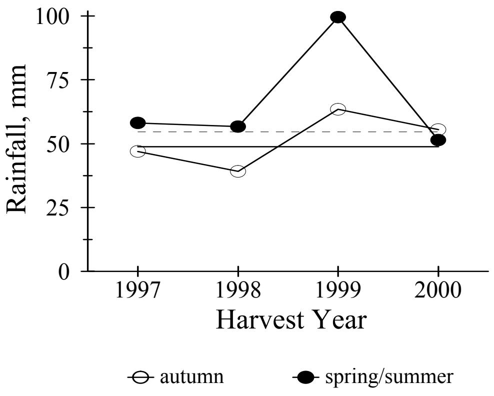




### 3.5.11 Field 107, Heydour.

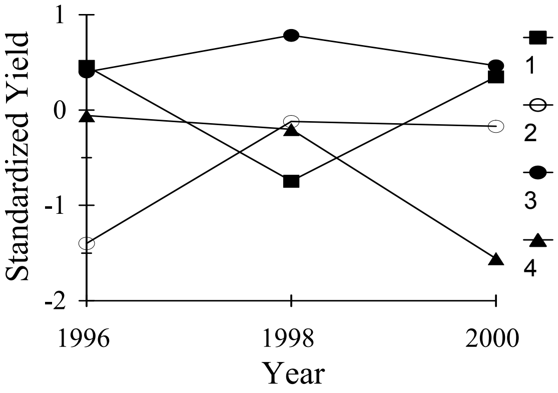
**Class centres** **Mean monthly rainfall**



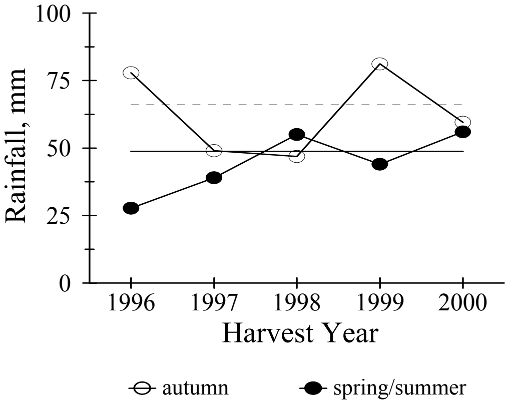


### 3.5.12. The Clays, Crowmarsh Battle.

**Class centres**

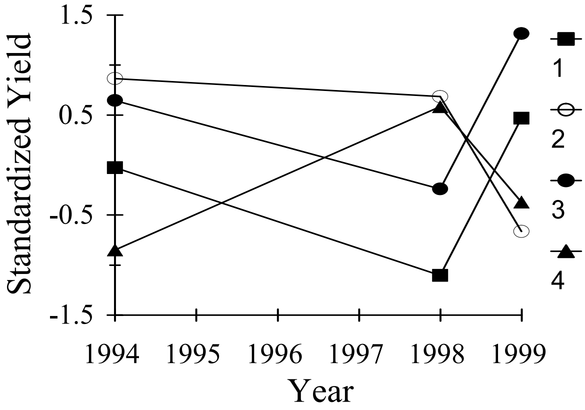


**Mean monthly rainfall**

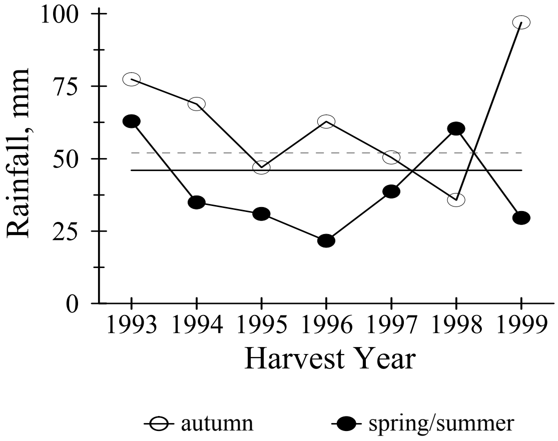


### 3.5.13. Football Field, Shuttleworth Farms.

**Class centres**

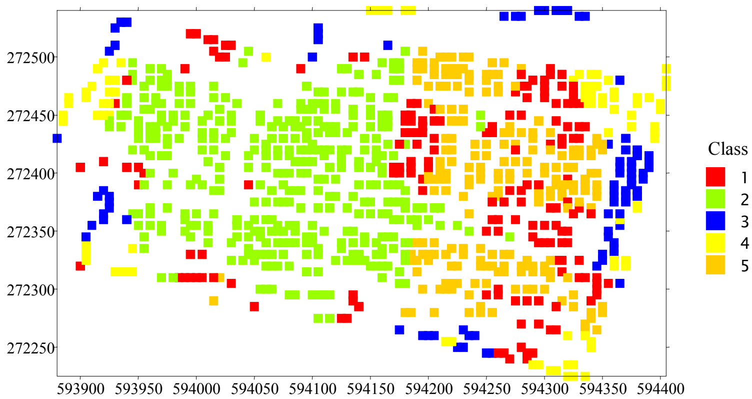


**Mean monthly rainfall**

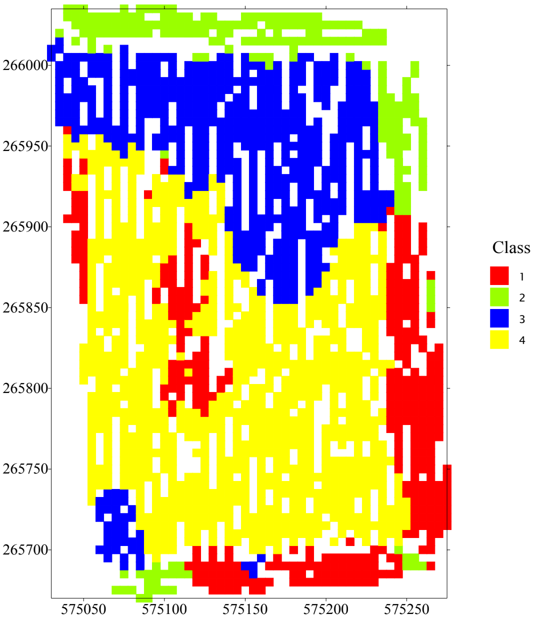


## 3.6 Class of maximum membership at sites across each field (corresponding to class centres in 3.5)

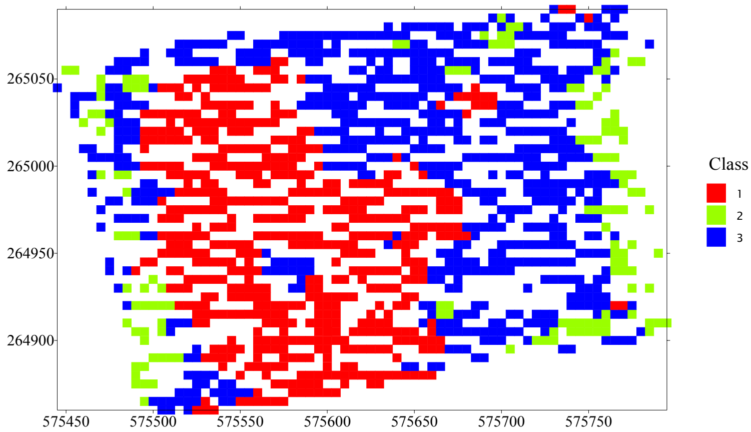
Hall 8 Field



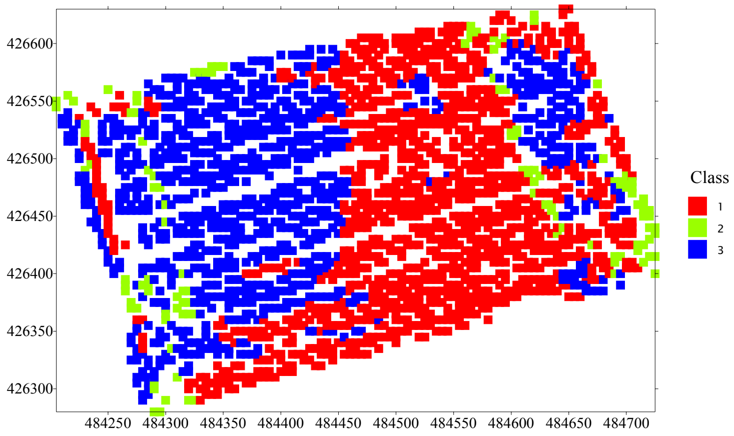
Brome Pin Field



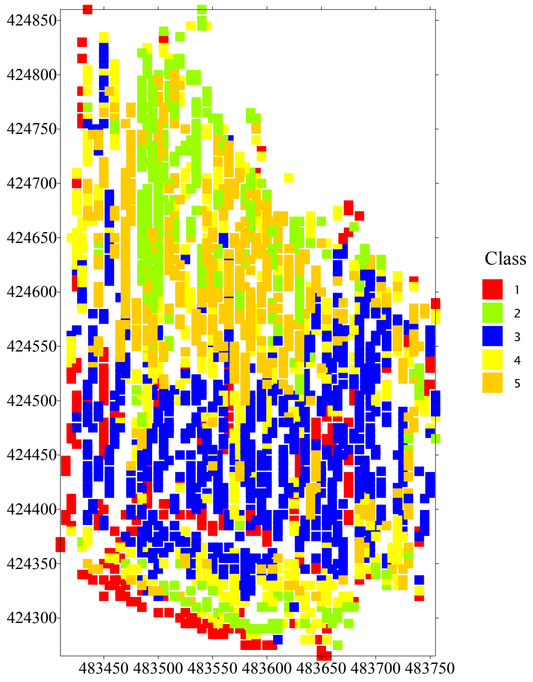
Little Lane Field



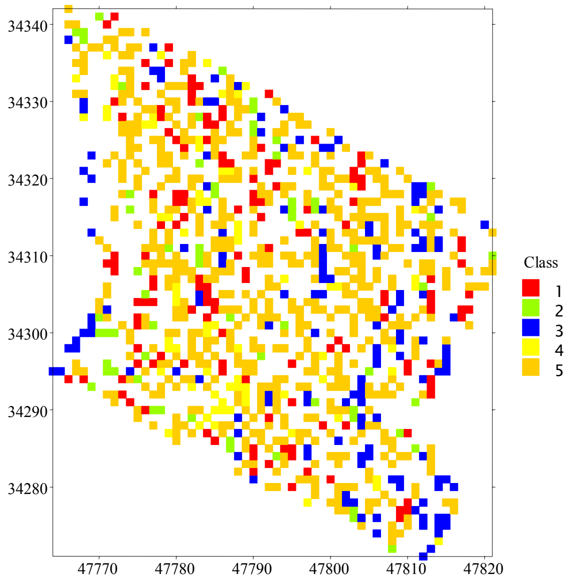
Commissioners Field



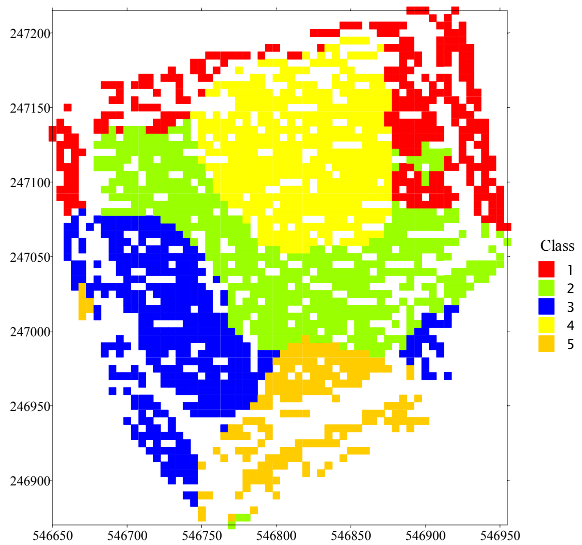
South Warpings Field



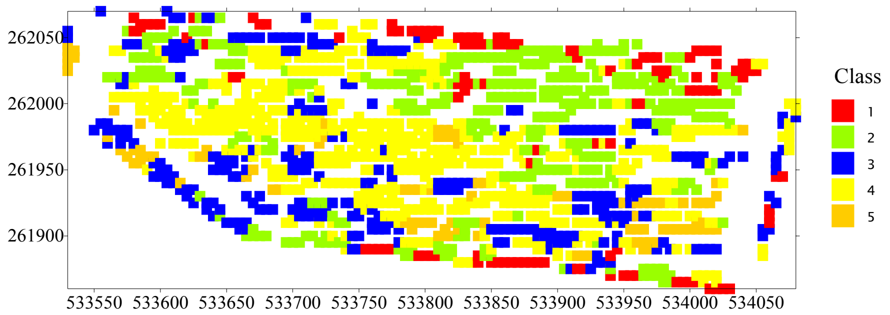
Field 2 Flawborough



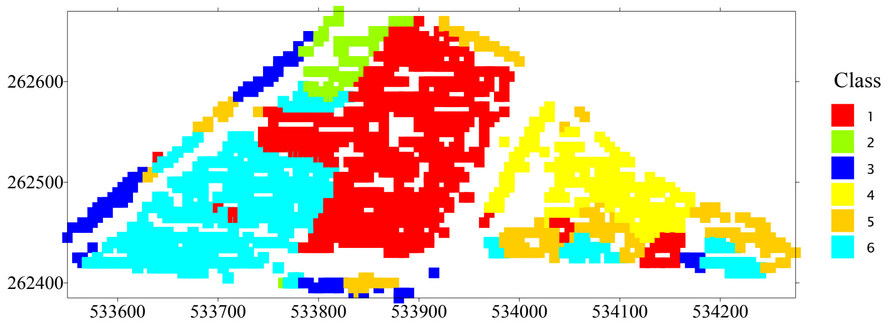
Gate Field, Whittlesford

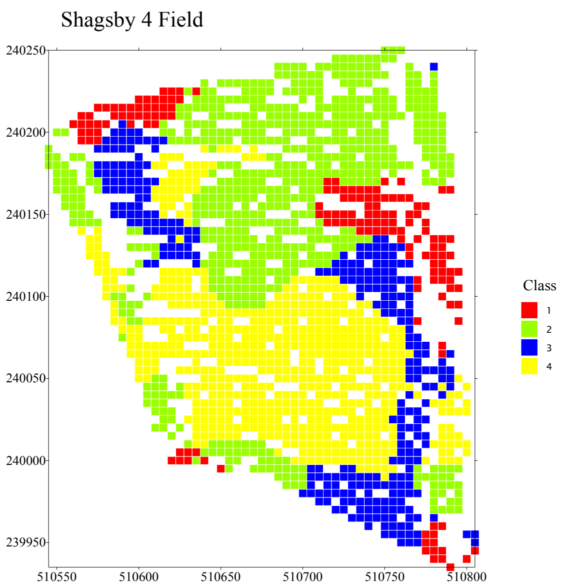


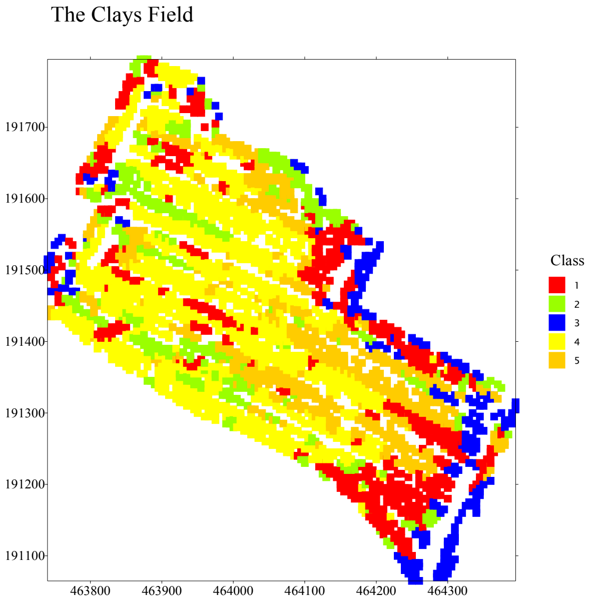
Top Pavements Field

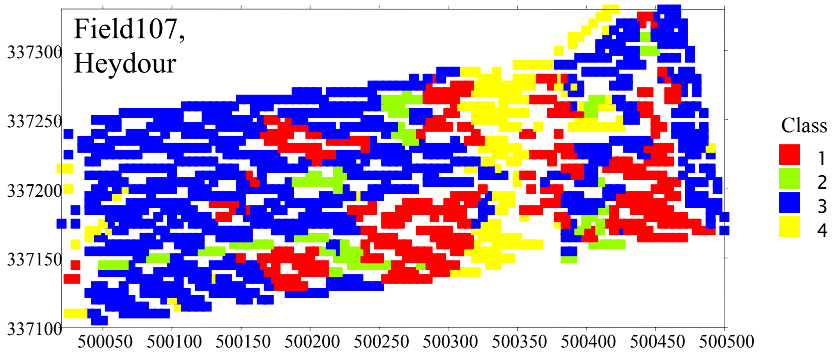


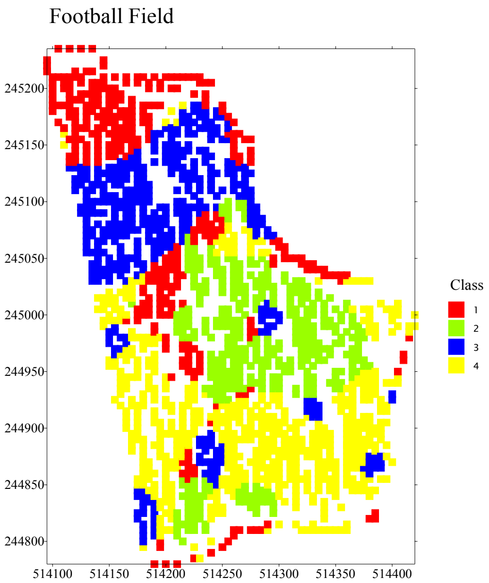
Knapwell Field




Shagsby 4 Field
240250
240200
240150
240100
240050
240000
239950
510550
510600
510650
510700
510750
510800
Class
1
2
3
4


The Clays Field
191700
191600
191500
191400
191300
191200
191100
463800
463900
464000
464100
464200
464300
Class
1
2
3
4
5


Field107, Heydour
337300
337250
337200
337150
337100
500050
500100
500150
500200
500250
500300
500350
500400
500450
500500
Class
1
2
3
4


Football Field
245200
245150
245100
245050
245000
244950
244900
244850
244800
514100
514150
514200
514250
514300
514350
514400
Class
1
2
3
4

# Chapter 4.  Assessement of potential for variable rate management

## Introduction

In Chapter 1 we recognized the need for a method to identify the scope for precision management of fields on the basis of minimum prior information.  In this chapter we test the hypothesis that this can be done using (i) generalized information on the local soil (as can be obtained from soil survey information across England and Wales), and (ii) this information in combination with yield maps of past cereal crops on the particular field.  First we describe the selection of fields for use in this work, and the way in which their variability was assessed.  We then describe the soil landscape information and the information obtainable from yield maps, and show how this information can be used to predict the scope for precision farming in a particular field.

## 4.1  Selection of fields for extensive study

The list of farms with fields that had been yield mapped for three or more years was used to chose suitable fields across arable England and Scotland (See Figures 2.2 and 2.3 and Table 4.1).  When selecting fields we ensured that there was a mix of yield patterns - simple and complex.  All farmers approached  agreed that their fields be mapped and their yield data be used in the project.  The yield maps were not available to the soil surveyor in the field.

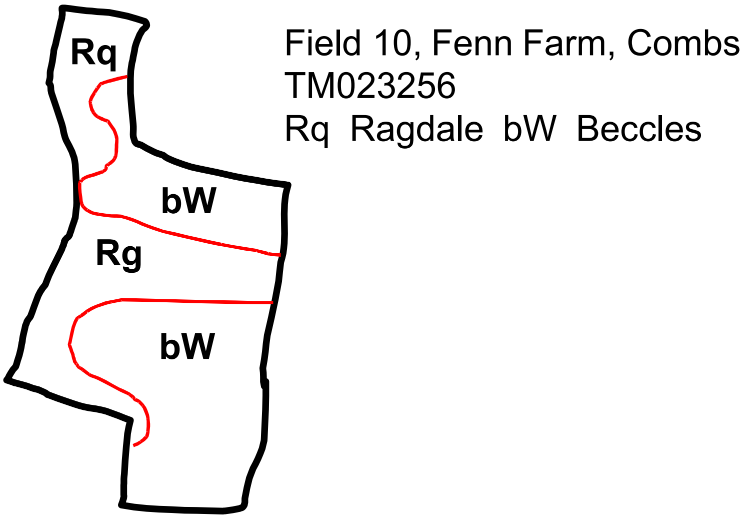
## 4.2  Field procedures for extensive study

Techniques for this exercise mirrored those of "traditional" soil survey.No more than one day was allocated to each field to cover travelling and mapping.  The aim was to make a minimum of one observation per hectare, but in many fields the density was closer to two per hectare.  In some cases the OS grid (at 100m spacing) was used to locate the sampling points, supplemented by additional points (for example Garden field, Doggetts), in others a 100 m grid was laid across the field in order to maximise the number of observations (for example Wellgrove field, Collins Green).  At each point the soil was described using a Dutch auger to 1.2 m (or shallower if rock or an impenetrable layer intervened).  Grid references were determined either by dead reckoning or using a Trimble GeoExplorer 3c GPS (reproducibility of position ± 2.5 m).  No samples were taken.  Soil boundaries were drawn on a base map to reflect the different soil series present (see Figure 4.1 for an example).

**Table 4.1 Location of extensively sampled fields. The NATMAP soil is the Association in the National soil map of England and Wales at 1:250,000 scale** (Hodge *et al.*, 1983).

| Field name | Farm | NATMAP soil |
|---|---|---|
| 2 | Flawborough | 431 - Worcester |
| 7 | Southern Green | 411d - Hanslope |
| 10 | Fenn | 711t - Beccles |
| 22 | South Fawley | 343h - Andover |
| 29 | Brooker | 814 - Newchurch |
| 30 | Welburn | 813d - Fladbury |
| 32 | Brooker | 813g - Wallasea |
| 99 | Yattendon | 582c - Hornbeam |
| 50, 92 & 93 | Bradenham | 572n - Burlingham |
| Burnby Moor | Throstle Nest | 711c - Brockhurst |
| Commisioners | Yokefleet | 532a - Blacktoft |
| Cowfield | Whitsome Hill | 575 - Whitsome |
| Creake Pasture | Barwick Hall | 581f - Barrow |
| Drive | Otley | 711t - Beccles |
| Garden | Doggetts | 571z - Hamble |
| Hall 11 | Manor | 343g - Newmarket |
| Jacobs | Galloway | 411d - Hanslope |
| Lucker Lee & Old Cow Pasture | Lucker | 542 - Nercwys |
| R4 | Hare Hill | 813g - Wallasea |
| Sacrewell | Sacrewell Lodge | 571u - Sutton |
| Wellgrove | Collin's Green | 572n - Burlingham |

**Figure 4.1 Example soil map for extensively sampled field**



### 4.3 Assessment of soil variability (pedological)

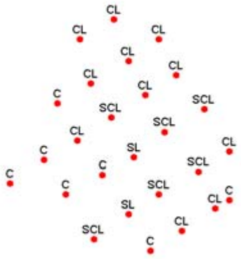
The data for each field were entered into a spreadsheet with the following properties recorded:

1) grid reference,
2) soil series,
3) wetness class, Hodgson (1997)
4) total available water capacity to 1 m depth*,
5) depth to clay/rock,
6) texture class for 0 to 30 cm, 30 to 60 cm and 60 to 90 cm,
7) stoniness class (Hodgson, 1977) for 0 to 30 cm, 30 to 60 cm and 60 to 90 cm (recorded in field but not included in the spreadsheet).

*Total available water capacity was calculated using APTAB a standalone program within LandIS, the NSRI Land Information System for soil and site data (Proctor *et al.*,1999). The calculations are based on Hall *et al.* (1977) plus more recent work by SSLRC, together with averages of available water capacity and easily available water capacity for almost 3600 soil horizons.

These data were then imported into the ArcView Geographical Information System (GIS). The overall variability of the soil properties was assessed using the data in the spread sheet. The spatial variability of the properties was assessed from a map in ArcView (see Figure 4.2 for an example).

**Figure 4.2. Texture classes at sample points in an exemplar field**



In Table 4.2 we present the scoring system by which the variability was assessed. The first six sets of scores refer to the **overall variability** of particular properties without reference to spatial structure. The final set of scores was used to assess the **spatial variability** of each of the pattern of soil series and of the particular properties. **Overall variability** was assessed by counting the number of points in each class and then allocating the field to a variability class. For example, in Figure 4.2, 19 sites (73%) were clay loam or sandy clay loam and 7 (27%) clays, placing the field in variability class 3 for topsoil texture. Boundaries were drawn manually around similar points to assess **spatial variability**. Again using Figure 4.2 as an example, most of the field is clay loam or sandy clay loam but the west/north west in predominantly clay. It was therefore allocated to spatial variability class 4 (speckled but with discrete map units with a minimum dimension 50 m and maximum more than 100 m). The distances in Table 4 were chosen with the distance between tramlines in mind, i.e. 25 m, as within field management can not take place within smaller widths. Table 4.3 shows the assessment for one particular field as an example.

## Table 4.2 Criteria for assessing within field variability

| | |
|---|---|
| SOIL SERIES | |
| 1 | 1 soil series in field |
| 2 | 2 soil series in field - closely related |
| 3 | 2 soil series in field - not closely related |
| 4 | 3 or more soil series in field - closely related |
| 5 | 3 or more soil series in field not closely related |
| DEPTH TO ROCK | |
| 0 | no rock in profile |
| 1 | all soils in 1 rock depth class |
| 3 | 2 depth classes |
| 5 | 3 depth classes |
| Depth classes: <40cm, 40 to 80cm, >80cm | |
| TEXTURE | |
| 1 | all soils in 1 texture class |
| 3 | 2 texture classes |
| 5 | 3 texture classes |
| Texture classes: sandy, light loamy, medium loamy, light silty, medium silty, clayey | |
| WETNESS CLASS | |
| 1 | all soils in 1 wetness class |
| 3 | 2 wetness classes |
| 5 | 3 wetness classes |
| DEPTH TO CLAY | |
| 1 | all soils in 1 clay depth class |
| 3 | 2 depth classes |
| 5 | 3 depth classes |
| Depth to clay classes: <40cm, 40 to 80cm, >80cm | |
| AVAILABLE WATER CAPACITY | |
| 1 | all soils in 1 available water capacity (AWC) class |
| 3 | 2 AWC classes |
| 5 | 3 AWC classes |
| AWC classes: 5-100 mm; 100-150 mm; 150-200 mm; 200-250 mm in 1 m of soil | |
| SPATIAL VARIABILITY | |
| 0 | no spatial variability - all one map unit |
| 1 | single map unit - some minor variability |
| 2 | speckled with no single map unit |
| 3 | speckled, but with discrete map units with maximum dimension less than 100m |
| 4 | speckled, but with discrete map units with minimum dimension 50m and maximum more than 100m |
| 5 | discrete map units dominated by single soil series |

**Table 4.3 Assessments of soil variability for extensively sampled fields**

Extract

| **Doggetts Farm, Rochford** | **Parent material:** Stoneless brickearth<br>**Association**: 571z Hamble 2 | | |
|---|---|---|---|
| **Garden field** | **Soil series** | | |
| | hL | Hamble | Silty, stoneless, well drained |
| | hK | Hook | Silty, stoneless, slight seasonal wetness with gleying above 70 cm depth |
| | Ssy | Sessay | Fine loamy, stoneless, slight seasonal wetness with gleying above 70 cm depth |

*Variability:*
Garden: moderate variations in topsoil texture silt loam to silty clay loam and depth to clay will affect management.

| | | |
|---|---|---|
| Soil series | Overall variability | 2 |
| | Spatial variability | 1 |
| Depth to rock or gravel | Overall variability | 0 |
| | Spatial variability | 0 |
| PROFILE texture | Overall variability | 2 |
| | Spatial variability | 3 |
| TOPSOIL texture | Overall variability | 1 |
| | Spatial variability | 2 |
| Depth to clay | Overall variability | 3 |
| | Spatial variability | 1 |
| Wetness class | Overall variability | 3 |
| | Spatial variability | 0 |
| AWC | Overall variability | 3 |
| | Spatial variability | 3 |

These reports and the field maps for each field were then passed to ADAS for agronomic assessment.

### 4.4 Agronomic assessment of the potential for variable rate management

The potential for variable rate management of crop inputs based on soil variability was assessed for each field using the classification scheme described in Table 4.2. Lime, NPKS fertiliser use, yield potential (influencing the offtake of P and K), residual herbicide use, ease of cultivation and risk of soil compaction were the main management decisions considered. The use of other agrochemicals (e.g. fungicides) were not considered since their use is primarily influenced by climate, crop type and management history rather than soil characteristics.

The assessments were carried out by one ADAS specialist based on the soils information provided by NSRI; the specialist did not visit the field and there was no access to any yield map information for the field. Each assessment was subjective based on a combination of research information and experience, but also taking account of the likely magnitude of the effects of variable rate application on crop and economic performance.

**Table 4.4. Classification scheme describing the potential for variable rate application of crop inputs.**

| Class | Description |
|---|---|
| 1 | No soil variability of practical significance |
| 2 | Some soil variability but probably not sufficient to justify variable rate management. Further investigations not worthwhile. |
| 3 | Some variability of one or more important soil properties. Further investigations or soil sampling needed to decide on potential for variable rate management. |
| 4 | Significant variability of one or more important soil properties on a manageable scale. Further investigations or soil sampling recommended. |
| 5 | Lots of soil variability that would probably be very worthwhile attempting to manage by variable rate application of one or more crop inputs. Investment in selected precision farming technologies worthwhile. |

These classifications were treated as ratings of each field by the potential for variable rate management. We refer to them below as *potential for variable rate management* (PVRM) *ratings*.

### 4.5 Prediction of the potential for variable rate mangement from soil parent material class.

In 1999 SSLRC (now NSRI) and ADAS compiled a report to MAFF entitled "*Towards identifying favourable soil types and associated cropping systems for precision farming*" (references CE0166 and CE0168). It provides a preliminary assessment of the soil (classified by parent material) and cropping systems where adoption of variable rate application of crop inputs within individual fields may be cost-effective and/or reduce the risk of pollution. The project output was regarded as the best assessment possible at the time, but it was based on relatively few observations so is subject to update as new research information and farm experience becomes available. The aim was to identify where variable rate application of crop inputs is most likely to provide farm or environmental benefits. Crop inputs were considered where there is a potentially predictable variation in input requirement. Winter wheat (for N, K, lime and herbicides) and potatoes (for irrigation) were the key crops for which assessments have been made. The use of fungicides, insecticides and plant growth regulators are primarily influenced by unpredictable seasonal or other factors and were not considered. Subjective experience and research results were considered. Soil type was the dominant factor used in the study because of its importance in both decision processes and crop performance. A key data source was the soil properties and national distribution information for soil series and soil associations held in the LandIS database. Simple classification schemes were developed to assess the potential of each soil association for variable rate application for each input. We refer to these as *opportunity classes.*

The fields studied in this project have been classified using the scheme developed in the MAFF project. Note that the classification for a field is based on the soil association (as identified for soils in England and Wales from the national soil map at 1:250,000 scale) and not on the detailed information on within-field variability. Table 4.6 explains the classifications derived from this scheme describing the opportunity for variable rate management of different inputs, and Table 4.7 shows the opportunity classes of each field in this project under this scheme.

We may now compare these opportunity classifications of the fields made on the basis of the local soil parent material class with the PVRM ratings made from specific field information. For this purpose we compare each of the opportunity classes (nitrogen, irrigation etc.) based on parent material class with (i) the five-level PVRM ratings (described in Table 4.4) and (ii) a simplified version of the PVRM rating. This is simply a generalization of the original rating into coarser categories with direct practical outcomes for the farmer making a decision about whether or not to take further steps towards precision management of a field. Table 4.5 below defines the simplified PVRM ratings.

**Table 4.5 Simplified PVRM ratings and their definition.**

| Original PVRM rating | Simplified PVRM rating | Definition of simplified PVRM rating |
|---|---|---|
| 1 | 1 | Further investigation not justified by variability. |
| 2 | | |
| 3 | 2 | Further investigation may be justified by variability. |
| 4 | 3 | Further investigation definitely justified. |
| 5 | | |

**Table 4.6 Classification schemes indicating the opportunity for adoption of variable rate application of different crop management inputs based on the soil parent material classification.**

**Nitrogen fertiliser**

| | |
|---|---|
| Class N1 (V.low) | No variation in soil type |
| Class N2 (Low) | 0-30% of soil association with a difference in recommended N of <30kg/ha N |
| Class N3 (Medium) | 0-30% of soil association with a difference in recommended N of 30-60kg/ha N<br>*or*<br>>30% of soil association with a difference in recommended N of <30kg/ha N |
| Class N4 (High) | 0-30% of soil association with a difference in recommended N of >60kg/ha N<br>*or*<br>>30% of soil association with a difference in recommended N of 30-60kg/ha N |
| Class N5 (V.high) | >30% of soil association with a difference in recommended N of >60kg/ha N |

**Irrigation**

| | |
|---|---|
| Class I1 (V.low) | No soil variation, or all soils with AWC difference of <15mm |
| Class I2 (Low) | 0-30% of soil association with AWC difference of 15-49mm |
| Class I3 (Medium) | 0-30% of soil association with AWC difference of 50-99mm<br>*or*<br>>30% of soil association with AWC difference of <49mm |
| Class I4 (High) | 0-30% of soil association with AWC difference of >100mm<br>*or*<br>>30% of soil association with AWC difference of 50-99mm |
| Class I5 (V.high) | >30% of soil association with AWC difference of >100mm |

**Weeds**

| | |
|---|---|
| Class W1 (Low) | Weed patches may be less likely (associations with no clay soils) |
| Class W2 (Medium) | Some weed patches likely (associations containing <30% clay soils) |
| Class W3 (High) | Lots of weed patches likely (associations containing >30% clay soils) |

**Residual herbicides**

| | |
|---|---|
| Class R1 (V.low) | No soils with sandy, organic or peaty topsoils |
| Class R2 (Low) | Uniform soils, all sandy topsoils |
| Class R3 (Medium) | Uniform soils, all organic or peaty topsoils |
| Class R4 (High) | Variation in topsoil type but containing <30% sand or organic or peaty topsoils |
| Class R5 (V.high) | Variation in topsoil type and containing >30% sand or organic or peaty topsoils |

**Phosphate and Potash fertiliser**

| | |
|---|---|
| Class K1 (Low) | No variation in soil type |
| Class K2 (Medium) | 0-30% of soil association with a difference in recommended K 50 kg/ha |
| Class K3 (High) | >30% of soil association with a difference in recommended K of 50kg/ha |

**Lime**

| | |
|---|---|
| Class L1 (Low) | Calcareous topsoils |
| Class L2 (Medium) | Non calcareous topsoils but no variation in lime requirement |
| Class L3 (High) | Associations containing topsoils of variable lime requirement |

Note: The descriptions ‘Very low’, ‘High’, etc. are undefined and are included for descriptive purposes only.

**Table 4.7 Opportunity classes based on parent material (from Table 4.6) for fields in this project.**

| Field name | Farm | Nitrogen Fertilizer | Irrigation | Weeds | Residual Herbicide | Phosphate and Potash | Lime |
|---|---|---|---|---|---|---|---|
| Garden | Doggetts | 1 | 1 | 1 | 1 | 1 | 2 |
| R4 | Hare Hill | 2 | 2 | 3 | 1 | 1 | 3 |
| Hall 11 | Manor | 3 | 2 | 1 | 4 | 2 | 1 |
| Creake Pasture | Barwick Hall | 2 | 3 | 1 | 4 | 2 | 2 |
| 22 | South Fawley | 3 | 2 | 1 | 4 | 2 | 1 |
| 99 | Yattendon | 1 | 2 | 1 | 1 | 1 | 2 |
| Wellgrove | Collin’s Green | 1 | 2 | 1 | 1 | 1 | 2 |
| 10 | Fenn | 1 | 2 | 2 | 1 | 1 | 3 |
| 30 | Welburn | 2 | 2 | 3 | 1 | 2 | 1 |
| Commisioners | Yokefleet | 2 | 2 | 2 | 1 | 2 | 1 |
| Football | Shuttleworth | 2 | 2 | 1 | 4 | 1 | 1 |
| Drive | Otley | 1 | 2 | 2 | 1 | 1 | 3 |
| Clays | Crowmarsh Battle | 2 | 2 | 1 | 1 | 1 | 1 |
| Knapwell | Boxworth | 1 | 1 | 3 | 1 | 1 | 1 |
| Gate | Whittlesford | 3 | 3 | 1 | 1 | 1 | 3 |
| Little Lane | Broom’s Barn | 2 | 2 | 2 | 1 | 1 | 3 |
| Shagsby | Lodge | 2 | 2 | 1 | 4 | 1 | 2 |
| 107 | Heydour | 3 | 3 | 2 | 1 | 2 | 1 |
| 2 | Flawborough | 1 | 1 | 3 | 1 | 1 | 2 |
| 3 | Flawborough | 2 | 2 | 3 | 1 | 1 | 2 |
| Lucker Lee, Old Cow Pasture | Lucker | 1 | 1 | 1 | 1 | 1 | 2 |
| Jacobs | Galloway | 1 | 1 | 3 | 1 | 1 | 1 |
| 7 | Southern Green | 1 | 1 | 3 | 1 | 1 | 1 |
| 50, 92 & 93 | Bradenham | 1 | 2 | 1 | 1 | 1 | 2 |
| 32 | Brooker | 2 | 2 | 3 | 1 | 1 | 3 |
| 29 | Brooker | 2 | 1 | 3 | 1 | 1 | 3 |
| Cowfield | Whitsome Hill | 1 | 1 | 1 | 1 | 1 | 2 |
| Sacrewell | Sacrewell Lodge | 1 | 3 | 1 | 1 | 1 | 3 |
| Burnby Moor | Throstle Nest | 3 | 2 | 3 | 1 | 1 | 2 |
| Top Pavements | Boxworth | 1 | 1 | 3 | 1 | 1 | 1 |
| Brome Pin | Boxworth | 1 | 2 | 1 | 1 | 1 | 3 |
| 12 acres | Hatherop | 1 | 2 | 3 | 1 | 1 | 1 |
| Trent | Andover | 3 | 2 | 1 | 1 | 1 | 1 |
| Short lane | Gamlingay | 1 | 1 | 3 | 1 | 1 | 3 |
| Sweet briar | Shuttleworth | 1 | 1 | 3 | 1 | 1 | 1 |
| Onion | Houghton Conquest | 1 | 1 | 3 | 1 | 1 | 3 |
| Parkway | Bunwell | 1 | 2 | 1 | 1 | 1 | 2 |
| Holly | Bunwell | 1 | 2 | 1 | 1 | 1 | 2 |

The comparison between each index for the potential for variable rate management of an input derived from the soil parent material class and the PVRM ratings (original and simplified) is made in a contingency table which shows how each of the fields for which the ratings were made are cross-classified. Thus a field in the first column and first row of the first subtable of Table 4.8 below is in class 1 for scope for variable nitrogen management under the soil parent material classification, and has a PVRM rating of 1. We would hope that the two ratings would be related, i.e. that the opportunity rating associated with a field given the soil parent material would be indicative of the scope for variable management made with field-specific information. To test this hypothesis we calculated a Chi-squared statistic for each table (by maximum likelihood using the CHISQUARE procedure in Genstat, Payne *et al.* 1988). From this we obtain a *p* value to test the null hypothesis that the soil parent material-derived classes and the field specific PVRM ratings are randomly associated. The results of these analyses are in Table 4.10 below.

**Table 4.10 Results of Chi-squared analysis for association of classification on potential for variable rate management of inputs based on soil parent material with original and simplified PVRM Ratings**

| Opportunity classification. | Chi-squared and *p* value | | | |
|---|---|---|---|---|
| | Original PVRM Rating | | Reduced PVRM Rating | |
| **Nitrogen** | 13.1 | 0.109 | 11.2 | 0.024 |
| **Irrigation** | 12.9 | 0.133 | 12.62 | 0.013 |
| **Weeds** | 12.6 | 0.127 | 6.03 | 0.197 |
| **Residual Herbicide** | 8.18 | 0.085 | 7.86 | 0.02 |
| **Phosphate and Potash** | 2.78 | 0.595 | 2.36 | 0.307 |
| **Lime** | 6.71 | 0.568 | 2.03 | 0.730 |

When a *p* value is larger than 0.05 we conventionally accept the null hypothesis which, in this case, is that the parent material categories and the field-specific PVRM ratings are randomly associated. Note that we must accept this null hypothesis in the case of the original PVRM ratings. However, there is evidence that the simplified ratings are positively associated with the parent material-based opportunity classes. This suggests that the original five-category assessment was too detailed to be predictable with any confidence from this very generalized information on the soil. However, the broader three-category rating made from field-specific information is associated significantly with the evaluations of opportunity for variable management of nitrogen, irrigation and residual herbicides based on parent material. This is encouraging, since it suggests that we can make practical predictions from general soil survey information (e.g. the soil association according to the 1:250,000 scale soil map of England and Wales) about how likely we are to find manageable variation within a field. The results presented above suggest that the classifications based on parent material are a useful indication of the potential for variable rate management of fields as assessed from field-specific information.

**Table 4.8 Association of opportunity classes for variable rate management of different inputs derived from soil parent material with PVRM rating (original 5-value scale).**

| | Nitrogen Class | | | Irrigation class | | | Weeds Class | | | Residual Herbicide Class | | Phosphate / Potash Class | | Lime Class | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PVRM rating** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **4** | **1** | **4** | **1** | **2** | **3** |
| **1** | 11 | 4 | 0 | 8 | 7 | 0 | 5 | 1 | 9 | 15 | 0 | 14 | 1 | 5 | 5 | 5 |
| **2** | 3 | 0 | 0 | 2 | 1 | 0 | 1 | 1 | 1 | 3 | 0 | 3 | 0 | 1 | 1 | 1 |
| **3** | 4 | 2 | 4 | 1 | 8 | 1 | 6 | 1 | 3 | 9 | 1 | 8 | 2 | 3 | 5 | 2 |
| **4** | 4 | 3 | 1 | 1 | 5 | 2 | 7 | 1 | 0 | 6 | 2 | 6 | 2 | 2 | 4 | 2 |
| **5** | 2 | 2 | 1 | 1 | 3 | 1 | 2 | 1 | 2 | 3 | 2 | 4 | 1 | 4 | 1 | 0 |

**Table 4.9 Association of opportunity classes for variable rate management of different inputs derived from soil parent material with simplified PVRM rating (3-value scale).**

| | Nitrogen Class | | | Irrigation class | | | Weeds Class | | | Residual Herbicide Class | | Phosphate / Potash Class | | Lime Class | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PVRM rating** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **4** | **1** | **4** | **1** | **2** | **3** |
| **1** | 14 | 4 | 0 | 10 | 8 | 0 | 6 | 2 | 10 | 18 | 0 | 17 | 1 | 6 | 6 | 6 |
| **2** | 4 | 2 | 4 | 1 | 8 | 1 | 6 | 1 | 3 | 9 | 1 | 8 | 2 | 3 | 5 | 2 |
| **3** | 6 | 5 | 2 | 2 | 8 | 3 | 9 | 2 | 2 | 9 | 4 | 10 | 3 | 6 | 5 | 2 |

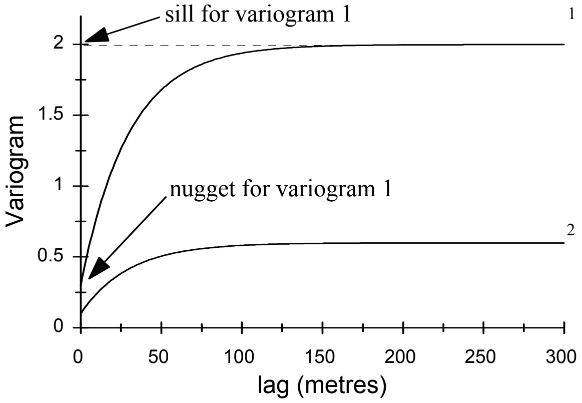
## 4.6 Measures of variability from yield maps for field-specific predictions of the scope for variable management.

Two or more yield maps were available for each field for which PVRM ratings had been derived from field-specific soil information. We now consider measures of variability which may be extracted from these yield data for predicting soil variability within the field.

The analyses of the yield data which we described in Chapter 3 generate two outputs. The first is a division of the field into zones within which there appears to be distinct patterns of season-to-season variation. We would expect such patterns to be very distinct in a field with much variability. A statistic which measures the 'distinctness' of the identified zones is the normalized classification entropy (NCE), which we mentioned in Chapter 4 because it is used to decide how many zones to define in each field. The smaller the NCE the more distinct the classes. NCE is therefore the first measure of variability which we consider for prediction of soil variability.

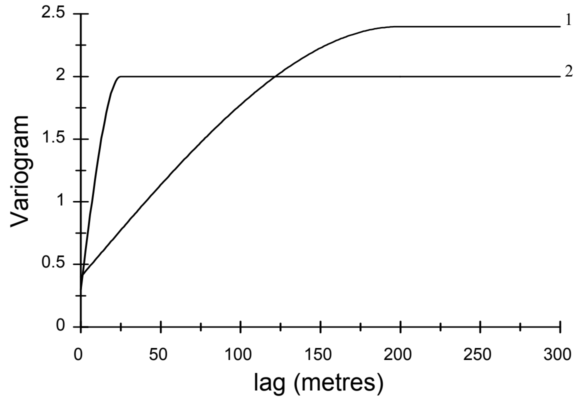
The other outputs of the yield map analysis are variograms for the yields in each season. Variograms may not be immediately meaningful to the non-specialist, but their size and shape are indicative of the scale and magnitude of variation. We now recall our explanation of the information contained in variograms which we gave in a previous report (Lark *et al*, 1998). We present three sets of hypothetical variograms of yield for various fields in Figures 4.3–4.5. The variogram is a plot of the variance of the difference between yield at two points in a field against the distance (or *lag*) between these points. In general as the lag distance between two points increases the variability of yield between them will increase, so variograms tend to increase with lag. The maximum value to which the variogram increases is called the *sill* , the intercept of the variogram (the apparent value at lag zero) is called the *nugget* (see Figure 4.3).

**Figure 4.3 Hypothetical variograms (i)**



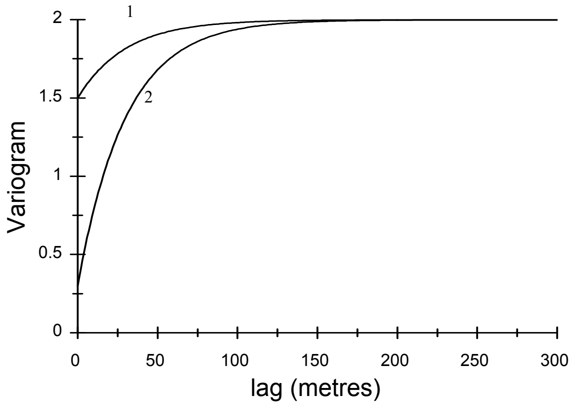
The sill of the variogram is a measure of the overall variability of yield in the field. Thus field 2 in Figure 4.3 is much more uniform with respect to yield than is field 1.

**Figure 4.4 Hypothetical variograms (ii)**



Yields vary because of factors which operate at different spatial scales. For example, there may be differences in the available water capacity of soil in a field which are due to changes in parent material which occur every 200 m on average. The important variations may occur over much smaller spatial intervals, for example, patchy deposits of sandy drift 25 m across on average. In the former case it may be feasible to map the underlying variation and respond to it. In the latter case it may not be possible. Figure 4.4 shows variograms like those which we might expect in the case of (1) yield variation dominated by long-range processes (e.g. differences between large contrasting zones in a field), and (2) variation dominated by short-range processes.

**Figure 4.5 Hypothetical variograms (iii)**



Yield variation may be influenced by processes at different scales in the same field. Of interest is the variation at such short range that the sampling has not resolved it. This is reflected in the nugget term of the variogram — the apparent value of the variogram at lag zero. If the nugget is large relative to the sill, then this implies that a lot of the variation in yield is happening at very short spatial scales. Thus, while both variograms in Figure 4.5 show the effects of a comparable long-range process, variogram (1) with a very substantial nugget actually reflects a situation in which much of the variability is extremely intricate, and so is probably not manageable in practice.

The challenge is to find a way of extracting automatically this information on variability from variograms of the yield data. This issue has also been addressed by McBratney *et al.* (2000), and our proposal draws on this work.

If a spatial variable, *z*, has a variogram, $\gamma(\mathbf{x})$, then we may compute the *dispersion variance* $\sigma^2{}_{\mathbf{B}}$ (Journel and Huijbregts, 1979). This is the variance of *z* within a region, **B**, of specified size and shape and it is obtained by evaluating the double integral of the variogram over all pairs of points, $\mathbf{x}_I$, $\mathbf{x}_j$ within **B:**

$$\sigma^2{}_{\mathbf{B}} = \iint_{\mathbf{B}} \gamma\left(\mathbf{x}_i - \mathbf{x}_j\right) \mathrm{d}\mathbf{x}_i \, \mathrm{d}\mathbf{x}_j \, . \qquad (4.1)$$

We follow McBratney *et al.* (2000) by computing the variance for a region excluding the nugget variance (i.e. the variance which is unstructured at the scale of sampling). Thus we may compare the variograms for a set of fields by computing the variance (less the nugget) which we would expect to see within a representative region. We computed this variance and then took its square root to obtain a standard deviation:

$$\sigma_{\mathbf{B}} = \sqrt{\iint_{\mathbf{B}} \gamma\left(\mathbf{x}_i - \mathbf{x}_j\right) \text{-}\, \mathrm{c}_{\mathrm{O}} \mathrm{d}\mathbf{x}_i \, \mathrm{d}\mathbf{x}_j} \; , \qquad (4.2)$$

where $\mathrm{c}_{\mathrm{O}}$ is the nugget variance of the variogram. This standard deviation will be large for variable fields and small for more uniform ones. Because it is computed for a region of standard size and shape it is comparable between fields in a way the ordinary sample standard deviation is not. Furthermore, by excluding the nugget we restrict the description of variation to that which is manageable.

A yield variation of 1 tonne/ha is more significant in a field where the mean yield is 5t/ha than it is in a field where the mean yield is 10t/ha. This is the reason for computing a coefficient of variation, the standard deviation of a data set expressed as a proportion (some authors use a percentage) of the mean. Here we computed the dispersion coefficient of variation by expressing the standard deviation in Equation (4.2) as a proportion of the mean yield. Thus if the average yield in a field was $\bar{z}$ and the dispersion standard deviation (less the nugget) for region **B** is $\sigma_{\mathbf{B}}$ then the coefficient of variation for region **B** , a notional representative region of the field, is

$$\mathrm{CV}_{\mathbf{B}} = \sigma_{\mathbf{B}} \ni \bar{z} \, . \qquad (4.3)$$

We computed $\sigma_{\mathbf{B}}$ from the variograms for the yield maps of all the fields for which agronomic variability ratings were available. In each case **B** was a square region, and we considered areas from 1ha to 25ha. In fact $\sigma_{\mathbf{B}}$ for these blocks of different area were strongly correlated (r>0.9) so we used a 1ha block as standard. The standard deviation $\sigma_{\mathbf{B}}$ and corresponding $\mathrm{CV}_{\mathbf{B}}$ was calculated for each yield map for a square 1ha block. We then computed the average of each of these statistics for all yield maps of each field.

A final statistic was extracted from the variogram of each yield map. This is the ratio of the dispersion variance for a 1ha square block (calculated using Equation 4.1, i.e. with the nugget included) to the variance of a 0.01ha square block. This ratio has a minimum value of 1 (when the variance is entirely unstructured over these scales). The larger the ratio the stronger the spatial structure of the variation in yield. This variance ratio statistic, VR, was computed from the variogram of each yield map and a field average was calculated.

In summary, for each field we have a normalized classification entropy (NCE) which measures the strength of evidence for distinctive season-to-season patterns of variation in the field. We also have mean values of the standard deviation and coefficient of variation for a 1ha sqaure block representative of the pattern of variation over the field as a whole ($\sigma_{\mathbf{B}}$ and $\mathrm{CV}_{\mathbf{B}}$ ) and the variance ratio of a 1ha and a 0.01ha square block (VR) which measures the extent to which the variation as summarized by the variogram appears to be spatially structured.

### 4.6 Predictive models for field variability.

The basic method used to predict the agronomic variability rating was the classification tree, as implemented in the statistical package S-PLUS (MathSoft, 1999). A classification tree is a set of rules for arriving at a prediction of the classification of an object by sequential decisions on the values of (categorical or continuous) predictor variables. In this regard a classification tree may be used just like a dichotomous key in biological identification.

A classification tree was used for two reasons. First, because it is appropriate for the prediction of a response which is a multinomial categorical variable (i.e. a set of more than two classes). We prefer to treat the agronomic variability ratings as classes because, although they are ordered (class *I* is more variable than class *I-1*) we have no reason to suppose that the class indices are linear (i.e. that the difference in variability between class *I* and class *I-1* is the same for any *I*). The classification tree method was also selected because it is a powerful technique for uncovering predictive rules within data sets without making any assumptions that the relationships between the classes we wish to predict and the predictor variables are linear or additive. We do not describe the statistical principles of classification trees in detail. Suffice it to say that any rule takes the form:

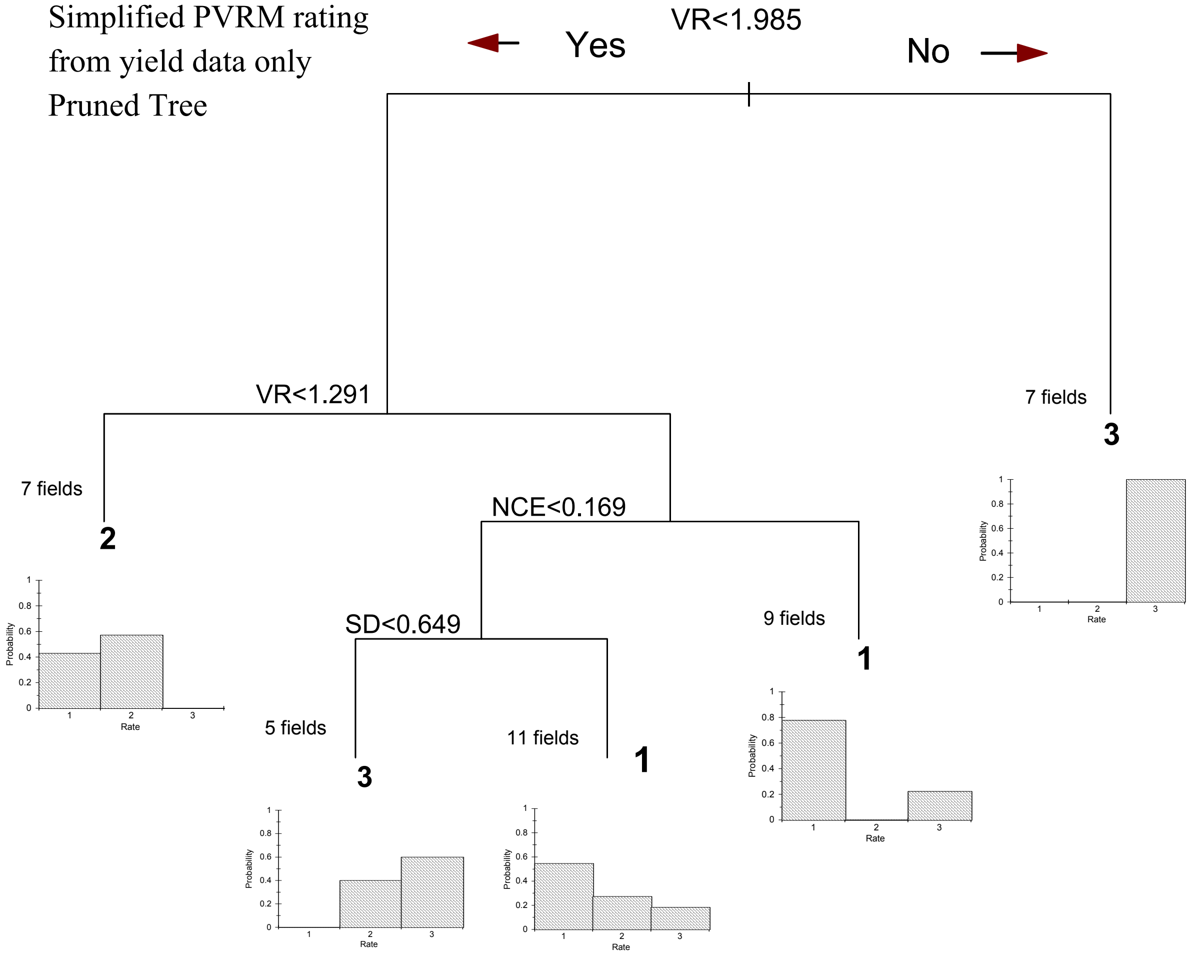
“**if** $x_1$ # 5.7 **and** $x_2$ 0 {a,b} **then** *y* is most likely to belong to class *I* ”

The best set of rules is found by a recursive search through the data and finding in succession the best set of rules to achieve the desired partition, only some of the available predictor variables may be used in the final tree. The resulting tree may be quite complex, after the algorithm is complete we may *prune* the tree by applying a *cost-complexity* measure which removes those rules which do not achieve a substantial improvement in the partition of the data.

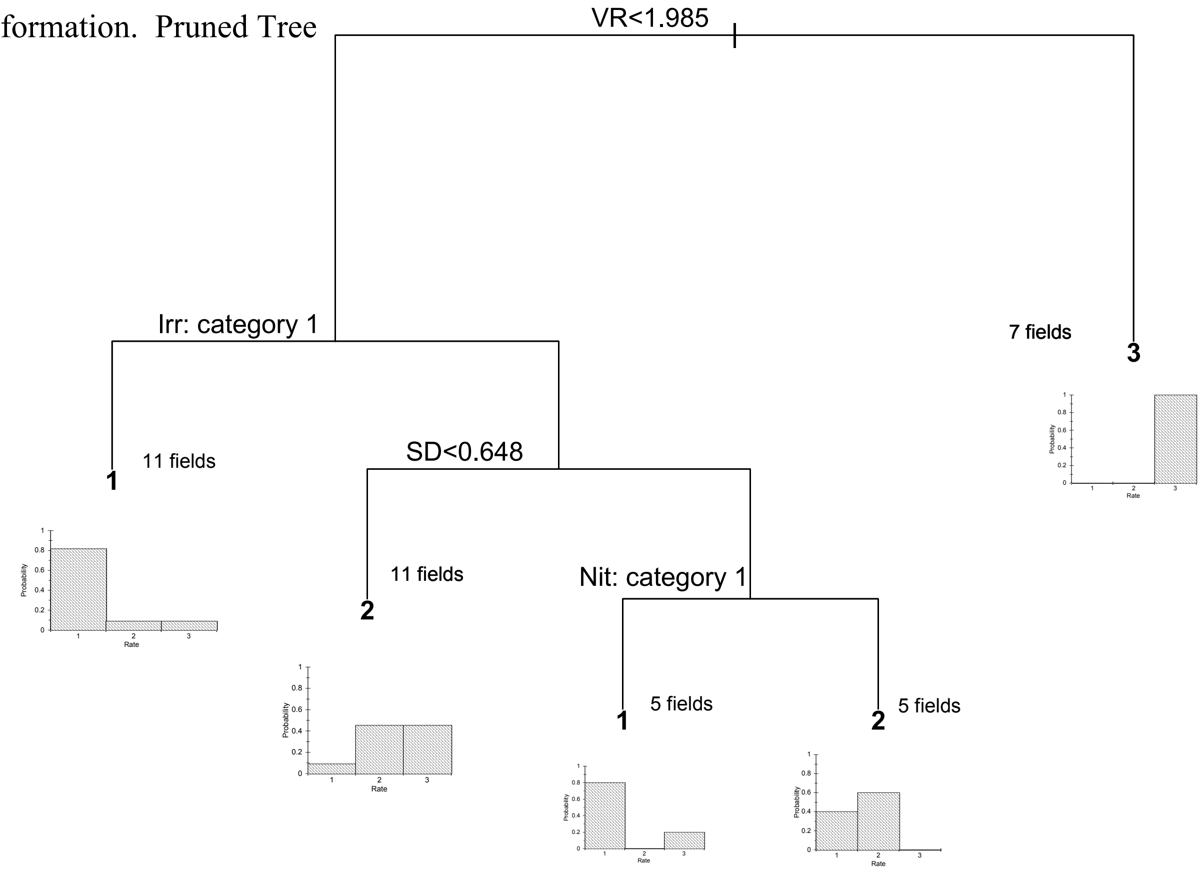
We present below two classification trees. The first one is for prediction of the simplified PVRM rating from the yield map variables described above. The second is for prediction of the same rating from the same variables along with the opportunity classes for management of all inputs obtained from the soil parent material classification.

The process of identifying the rating for a field begins at the top of the tree. The variables in the tree for prediction from yield data only are the 1ha:0.01ha variance ratio (VR), the standard deviation $\sigma_{\mathbf{B}}$ (called SD on the tree) and the Normalized Classification Entropy (NCE). The first question is whether VR < 1.985. If the answer is "yes" we proceed down the left-hand limb of the tree. If the answer is "no" then we immediately predict that the simplified PVRM rating is 3. This is the number at the end of the branch. We also show there the number of fields in our data set which are allocated to this terminal node, and the relative frequency of the different ratings among these fields (as a histogram). Thus, if VR <1.291 the predicted PVRM rating is 2, although we note that just over 40% of the 7 fields which occur at this node actually had rating 1. The histograms indicate the uncertainty of any particular predicted rating, and should be taken into account when decisions are made on the basis of this information. Fields where VR falls between 1.291 and 1.985 are then allocated on the basis of the NCE. If this is >0.169 (indicating that patterns of season-to-season variation are within the field are not very distinct) then rating 1 is predicted. Otherwise the rating depends on the $\sigma_{\mathbf{B}}$ . Counter-intuitively smaller values of this statistic are indicative of rating 3, but bear in mind that this decision follows on previous branches of the decision tree, this does not imply that, overall, small values of $\sigma_{\mathbf{B}}$ are indicative of scope for variable rate management.

Simplified PVRM rating
from yield data only
Pruned Tree



Simplified PVRM Rating
Yield and parent material
information. Pruned Tree



A similar tree incorporates information on the soil parent material class along with the field-specific information from analysis of the yield maps. Note that the pertinent classifications based on the soil parent material information are **Irr** — whether the possibility of variable rate irrigation management is in the lowest category (1, follow the left branch of the tree) or in any of the higher categories, and similarly **Nit:** the classification on scope for variable management of nitrogen. Here the most uncertain predictions are those of PVRM rating 2 (not unreasonably since this category is one where the scope for variable rate management was itself uncertain). However, if the prediction of rating 2 is made at the penultimate branching of the tree (SD<0.648) then the predicted rating is 2 but the uncertainty is skewed in the direction of a higher rating. If the prediction of rating 2 is made in the ultimate branching of the tree then the uncertainty of a prediction of class 2 (soil-landscape based assessment of the scope for variable nitrogen is rated higher than the lowest category) is skewed towards a lower rating.

In the tables below we show the fitted and true reduced PVRM rating for all 39 fields in the two trees. These tables are not necessarily a fair representation of the errors which we would see if we used the trees to predict variability ratings for new fields. For this reason we did a *jacknife estimation* (explained below) of the errors for the second tree (field-specific and soil-landscape information). This was done by randomly dividing the data into 8 validation sets (7 sets of 5 fields and 1 set of 4 fields). We then fitted a tree to each remaining subset of 34 data (or 35 in the last case), specifying only the predictor variables which were used in the illustrated tree, and specifying the same value of the cost- complexity for pruning it. Each tree was then used to allocate to a class the 5 (or 4) validation data *not* used to form it. The jacknife prediction is shown in the third table.

**Table 4.10 Fitted and actual simplified PVRM rating for Tree 1 (yield data only).**

| | **Reduced Agronomic rating (Fitted)** | | |
|---|---|---|---|
| **Reduced Agronomic rating (True)** | **1** | **2** | **3** |
| **1** | 13 | 3 | 0 |
| **2** | 3 | 4 | 2 |
| **3** | 4 | 0 | 10 |

**Table 4.11 Fitted and actual simplified PVRM rating for Tree 2 (yield and parent material information).**

| | **Reduced Agronomic rating (Fitted)** | | |
|---|---|---|---|
| **Reduced Agronomic rating (True)** | **1** | **2** | **3** |
| **1** | 13 | 1 | 2 |
| **2** | 3 | 5 | 0 |
| **3** | 0 | 3 | 12 |

**Table 4.12 Predicted (Jacknifing) and actual simplified PVRM rating for Tree 1 (yield and parent material information).**

| | Reduced Agronomic rating (Predicted) | | |
|---|---|---|---|
| **Reduced Agronomic rating (True)** | **1** | **2** | **3** |
| **1** | 13 | 2 | 1 |
| **2** | 1 | 6 | 2 |
| **3** | 2 | 1 | 11 |

Note that the overall percentage of fields for which the fitted rating is the correct one is 69% (yield information only) and 72% (yield and soil-landscape information. In the jacknifing 76% of the fields were correctly allocated. If fields were allocated among these categories at random (but with the correct overall proportion in each category) then the expected percentage correctly allocated is 35%. This indicates that the trees contain substantial information.

Following Lark's (1995) systematization of the accuracy measures which can be extracted from a cross-classification table note the following:

1) If a field is predicted to have rating 1, then the estimated probability that the true rating is 1 is 0.81.

2) If a field actually has rating 1 then the then the estimated probability that the predicted rating will be 1 is 0.81 (this is only coincidentally the same as the probability under (1)).

3) If a field is predicted to have rating 3, then the estimated probability that the true rating is 3 is 0.79.

4) If a field actually has rating 3 then the then the estimated probability that the predicted rating will be 3 is 0.79 (this is only coincidentally the same as the probability under (3)).

5) If the predicted rating is 3, and we therefore decide to spend resources on further information the probability that the variability will be found to be uninteresting is only 0.07 (7%).

6) If the predicted rating is 1, and we therefore decide NOT to spend resources on further information the probability that we will have missed an opportunity to pursue precision agriculture on the field is 0.125 (12.5%).

Note that in any decision making we will have further information on uncertainty to guide us from the histograms on the trees. Thus, for example, if in the first tree we predicted the rating of the field to be 3 at the first branch (i.e. VR>1.985) we can be more confident that there will be interesting variability in the field than we could if we arrived at this prediction at the final branching of the tree (VR>1.291, NCE<0.169, SD<0.649).

## 4.7 Conclusions

1) We have shown how the soil variability of a field may be assessed by an expert with a view to rating the potential for variable rate management of inputs to a crop.

2) We have shown further how the agronomic significance of this variability may be assessed in a rating scheme for potential for variable rate management.

3) We have shown that a general assessment of the opportunity for variable rate management of inputs based on the soil parent material is significantly associated with assessments of the potential for variable rate management based on field-specific information on the soil. This is both a validation of the tentative assessments in the original MAFF project which derived the opportunity indices and evidence that they may be a practically useful tool.

4) We have shown how the potential for variable rate management of inputs to a field may be predicted from measures of spatial variability extracted from yield maps. The classifications trees which we have produced and assessed by jacknifing constitute both a validation of our hypothesis that the variability of yield reflects the inherent scope for variable rate management of a field, and a tool for making such assessments in practice.

# Chapter 5. Differences between management zones with respect to soil properties, and the scope for predicting soil properties from yield and other ancillary data

## Introduction

Assume that a farmer has decided to investigate the soil variation within a field, perhaps using the decision making tools developed in the previous chapter. In this chapter we investigate how the management zones derived from analysis of yield maps (and the underlying membership values in the classes by which these zones are defined) may assist in this investigation. We therefore investigate whether the zones (and membership values) are significantly related to soil properties within the field, and we assess the predictive power of these relationships for mapping variation of these properties. We compare, at some fields, the predictive power of the zones and membership values with data from non-intrusive sensors.

This work required detailed soil information on several fields. We describe the data collection and then the analytical methods used before presenting our results.

## 5.1 Selection of fields for intensive study

The list of farms with fields that had been yield mapped for three or more years was used to chose ten suitable fields for intensive sampling across arable England that were, for logistical reasons, within 70 miles of Silsoe (Table 5.1). Selection was biased to ensure that there was a mix of yield patterns - simple and complex. Four of the fields chosen were also part of the associated HGCA funded project *Evaluation of non-intrusive sensors for measuring soil physical properties* (2243) and two were also part of project *To distinguish between the sources of spatial variation responsible for the variation in cereal yield that can and cannot be ameliorated as a means of optimising inputs, management and environmental sustainability* (2298). All farmers approached agreed to soil investigations on their fields and the use of their yield map data in the project. The yield maps were not available to the soil surveyor for in the field.

## 5.2 Field procedures for intensive study

No more than one day was allocated to each field to cover travelling and coring. The aim was to make a minimum of 50 observations using a coring device mounted on the back of an ATV (average was 57). The cores, approximately 90 cm long and 2.5 cm diameter, were collected in plastic tubes. They were taken across the fields in straight lines or along tramlines, depending on the season. Their positions were recorded using a Trimble GeoExplorer 3c GPS (reproducibility of position ± 2.5 m).

**Table 5.1 Location of intensively sampled fields. The NATMAP soil is the Association in the National soil map of England and Wales at 1:250,000 scale.** (Hodge *et al.*, 1983).

| Field Name | Farm | NATMAP soil |
|---|---|---|
| 3 | Flawborough | 813b - Fladbury |
| 107 | Heydour | 343a - Elmton |
| Gate | Whittlesford | 571k - Moulton |
| Brome Pin | Broom’s Barn | 571o - Melford |
| Clays | Crowmarsh Battle | 511g - Coombe |
| Football | Shuttleworth | 541A - Bearsted |
| Knapwell | Boxworth | 411d - Hanslope |
| Little Lane | Broom’s Barn | 572q - Ashley |
| Shagsby | Lodge | 541A - Bearsted |
| Top Pavements | Boxworth | 411d - Hanslope |

### 5.3 Soil analyses

On returning to the laboratory, a short description was made of the soil horizons of each core. The cores were divided into sub-samples and put in polythene bags. Samples were taken at 10 to 20 cm and 45 to 55 cm depths for chemical and physical analysis. The length of each sub-sample was measured as this, together with the diameter of the tubes, formed the basis of the bulk density calculation.

The distribution of particles in a soil sample between the standard size fractions — sand, silt and clay (Hodgson, 1977) — may be determined either by a laboratory method (Avery and Bascomb, 1982) or in the field by hand-texturing : working moist soil between the fingers and thumb. These methods are generally internally consistent, but are not necessarily comparable. The determination of particle size distribution is particularly difficult in soils derived from Chalk (as at The Clays, Crowmarsh Battle) where hand texturing underestimates sand fraction by comparison to laboratory methods.

Organic carbon content of samples was determined by the wet oxidation method described by Avery and Bascomb (1982). Bulk density was determined as the ratio of the mass of oven-dry soil to the field volume of the sample.

### 5.4 Statistical relationships between soil properties and classified yield data (and other possible predictors).

### 5.4.1. The available soil data.

Two sets of data were analyzed in this project. The first set is from the fields sampled as part of the project, as described in the previous sections of this Chapter. For each of the intensively sampled fields we have measurements of soil properties from grid samples, as described above. We also have some variables derived from these measurements using pedotransfer functions. These are either **AWC** (available water capacity derived by the APTAB program in LandIS (see section 4.3), *or* available water capacity (volumetric proportion) in the subsoil (**AWS**) or topsoil (**AWT**) *or* available water (mm available water) over the top metre of soil (**AW1m**) These three latter variables were determined from the measured soil properties using the pedotransfer functions of Mayr *et al.* (1999). These latter estimates were used in preference to the APTAB AWC where sufficient measurements of the necessary soil variables were available.

Two of these intensive fields were in common use for the present project and for the University of Reading's HGCA project (2298). At The Clays field, Crowmarsh Battle farms, the University of Reading analyzed soil samples from the current project as part of their work, so in addition to the physical properties measured at the other intensively sampled field we have information on available potassium and pH (topsoil and subsoil) and available phosphorous (topsoil). At Football field, Shuttleworth farms, we have measured data on topsoil potassium, phosphorus and pH along with the physical properties measured for other fields.

At these intensively sampled fields the soil series was also identified at each sample point.

In addition we have five data sets from work funded by BBRO where comparable information was available on soil physical properties and available phosphorous. The physical properties include particle size distribution of the soil, and available water capacity (expressed here as a percentage by volume) in the topsoil (**AWt**) or subsoil (**AWs**) which was obtained from laboratory estimates of the soil water characteristic curve. Yield maps of cereal crops, analyzed with the same methods described in Chapter 3, were available for these fields. Note that two of the BBRO fields were also used as intensively sampled fields in this current project.

Given these multiple sources of data we define the variables measured in each field at the beginning of the results section for each field.

### 5.4.2. Predictor variables.

The analysis of the yield data described in Chapter 3 generates a classification of sites across the field into *g* classes. Each site has a membership in each of the classes, and these memberships were smoothed after log-ratio transformation, the transformation being appropriate because the memberships for each site over all *g* classes are constrained to sum to 1. At each site the class of maximum membership may be identified.
As described in more detail below, we consider how much variation in the measured soil properties is accounted for by both the classification of the field (class of maximum membership) and the log-ratio

transformed memberships in the classes. This serves two purposes. First, it indicates how far the subdivision of the field into management zones by classification of the yield data has captured the key variations within the field. Second, it allows us to assess quantitatively how useful the class mean or a prediction from the class memberships will be for predicting the values of soil properties at points across the field on the basis of a small calibration sample.

Four of the fields intensively sampled as part of the present project were also in use as part of HGCA funded project *Evaluation of non-intrusive sensors for measuring soil physical properties* (2243). At each of these fields data on the apparent electrical conductivity of the soil (ECa) were made by electromagnetic inductance (EMI) using an EM38 sensor in horizontal (H) and vertical (V) mode. The measurements were made as detailed in the following table.

| Farm | Field | Date | Mode |
|---|---|---|---|
| Heydour, Grantham | 107 | 16.03.00 | H & V |
| | | 17.08.00 | H & V |
| Lodge Farm, Chicksands | Shagsby 4 | 11.02.00 | V |
| | | 23.08.00 | H & V |
| Shuttleworth | Football | 19.09.01 | H & V |
| | | 21.03.02 | V |
| Crowmarsh Battle | The Clays | 22.08.01 | H & V |
| | | 04.04.02 | H & V |
| | | 06.06.02 | H & V |

The analysis of these measurements is presented in full in the project report 2243. Here we compare the ECa measurement as a predictor of soil properties with the zone means and log-ratio memberships derived from the yield data. ECa was estimated at the soil sampling site by punctual kriging from measurements within 25 m. We generally used one or two of the data sets for this purpose. The choice was generally determined by (i) whether the coverage of the data set was good enough and (ii) whether the ECa measurements in the different data sets were correlated. We only used two sets of measurements (different modes or dates) when these were correlated with a correlation coefficient smaller than 0.9.

At Shagsby 4 field we also had spectral reflectance data in a visible red waveband (652.5 – 667.5 nm) measured on bare soil in the autumn of 2002 as part of project 2243. These values were kriged at the soil sampling site in the same way as the ECa data, then compared with the ECa and yield-derived predictors.

### 5.4.3. Assessing the value of predictor variables.

An objective of these analyses is to assess the usefulness of the analyzed yield maps, and the ECa data, for predicting the variations of soil properties at within-field scale, on the basis of some limited calibration data. Two general approaches to this problem of prediction are considered.

The first approach is to use the local class mean as a predictor. Analysis of our sample data allow us to estimate the mean value of a soil property, *z*, for sites where the class of maximum membership is class *i* :— $\bar{z}_i$. Thus, at any location **x** in the field where the smoothed membership in class *i* is larger than that of any other class the predicted value of property *z* is $\bar{z}_i$. The usefulness of this predictor will depend on the variability within the classes, if this is small by comparison to the overall variability of the soil property then the class mean will be a useful predictor. Another advantage of this approach is that for various soil properties the class mean may be estimated from a bulk sample collected from sites within the class and combined by physical mixing into a single sample for analysis. This is only suitable for those soil properties where the physical average obtained in this way is expected to equal the statistical average (Webster and Oliver, 1990). Where this assumption is reasonable then there is scope to reduce costs of analysis by using bulk samples to estimate the class mean.

The second approach is to use the smoothed log-ratio memberships as predictors of the soil properties in a multiple linear regression. This has been shown to be a powerful way of using results of continuous classification on ancillary variables for prediction of soil properties in other studies (Lark, 1999, 2000a). The same analysis was also used to test ECa and the visible red reflectance measurements as predictors of soil properties.

A useful measure of the value of a **point prediction** of a soil property is the prediction variance or mean squared error of prediction, *MSEP*. If a prediction of the value of the soil property at location **x** is $\hat{z}_{\mathbf{x}}$ then the *MSEP* is defined as;

$$MSEP = \mathrm{E}\left[\{\mathrm{z}_{\mathbf{x}} - \hat{\mathrm{z}}_{\mathbf{x}}\}^2\right], \qquad (5.1)$$

where $z_{\mathbf{x}}$ denotes the true value of the property, and E[] denotes the statistical expectation (or mean) of the term in brackets.

The analyses to test these approaches with the data from the intensively sampled fields, and to estimate and compare their *MSEP* are described below in more detail.

We also evaluated the scope for predicting the soil series at unvisited sites using the log-ratio memberships from the yield map analysis (in combination with sensor data where available). The classification tree method described in Chapter 4 was used for this purpose.

**5.4.1. Analysis**

### 5.4.1.2 General analysis

For each soil variable the mean and the sample variance were calculated. The first quartile of the data (Q1) was extracted (i.e. the value such that 25% of the observations are less than or equal to Q1) and the third quartile Q3 (25% of the observations are larger than to Q1). The minimum and maximum values were also extracted.

### 5.4.1.3 Estimating zone means and assessing their predictive power

The smoothed log-ratio memberships for all *g* classes were extracted for the yield site nearest to each soil sampling site. The class of maximum membership (the management zone) can therefore be identified. Because the sampling was not random the class means and the within-class variance were estimated by *restricted maximum likelihood* using the REML directive in Genstat (Payne *et al.*, 1988). This returns a *p* value for a test on the null hypothesis that the class means for the property of interest are all equal (based on the Wald statistic). If this null hypothesis is rejected then we cannot regard the management zones as different with respect to the particular soil property. The within-class variance, $\hat{\sigma}^2_{\mathrm{W}}$ measures the variability of the soil properties within the management zone. When we use the class mean as a predictor the *MSEP* will depend on the within-class variance. We may estimate the *MSEP* of the class mean, $MSEP_c$ as:

$$MSEP_c = \hat{\sigma}^2_{\mathrm{W}} \left\{ 1 + \frac{g}{n} \right\}, \qquad (5.2)$$

where *n* is the number of soil data available to estimate the class means. The second term in the braces allows for the estimation error of the class means.

### 5.4.1.4 Estimating regression models and assessing their predictive power

Regression analyses of the soil variables on all the log-ratio membership values (smoothed) and on a subset of these were computed. The subset consisted of the memberships in one to three classes, selected because these appeared to represent the dominant variation across the field.

It is important to note that the regression was not done by ordinary least squares (OLS). OLS regression is not suitable for the analysis of spatial data sets that have not been collected by random sampling. All these data were collected on systematic grids. In these circumstances OLS can seriously overestimate the significance of a regression relationship.

In the case of the data sets collected specifically for this project we used the REML directive in Genstat (Payne *et al.*, 1988) to estimate the predictive model. This returns an overall measure of the significance of the relationship (based on the Wald statistic) and an estimate of the error variance of the regression model $\hat{\sigma}_{\mathrm{R}}{}^{2}$. We used REML because in most of these cases the number of data available were relatively small (fewer than 70, often much fewer). REML is particularly desirable in such circumstances because the estimated variances are unbiased.

The *MSEP* for a regression predictor with *k* predictor variables fitted to *n* data, based on Shibata's (1986) final prediction error ($\alpha$ =2) is then estimated by :

$$MSEP_{\mathrm{R}} = \hat{\sigma}^{2}{}_{\mathrm{R}}\left\{1 + \frac{2k}{n-k}\right\}. \qquad (5.3)$$

We used this procedure to investigate the predictive power of the full set of log-ratio memberships in classes derived from the yield data. We also used a subset of the memberships, i.e. membership in fewer than *g* classes. The classes to use were selected by (i) identifying those which are class of maximum membership over a significant proportion of the field then (ii) computing the correlation between the log-ratio memberships and selecting (typically two or three) classes whose memberships are weakly correlated and so which provide independent information. We also used the procedure to estimate the *MSEP* for a regression predictor using ECa data, where these were available.

At the fields sampled in the BBRO project we used maximum likelihood regression. This was used because at all fields there were more than 80 data points. The ML regression procedure is described elsewhere by Lark (2000a, 2000b). The ML procedure also generates an overall measure of significance (the Wald statistic) and a statistic for comparing alternative predictors of the same variable (the Akaike information criterion). It also generates a variogram of the error variable in the model. In the case of the BBRO data we also estimated the variogram of each soil variable by maximum likelihood (Pardo-Iguzquiza, 1997) . This variogram may be compared with that of the error process to shed light on the sources of variation which the regression accounts for or has failed to explain.

These equations for estimating the *MSEP* are useful because they give a realistic measure of the utility of the fitted models for point prediction of soil properties. This is distinct from the scientific inference which the model allows us to make. Thus, we may conclude that the classes differ significantly with respect to a soil property (and so may represent a useful zonation of the field for management). At the same time, the *MSEP* may be large because a good deal of variation in the property of interest remains unexplained by the model.

### 5.4.2. Results

The results of these analyses are given in the following Tables and Figures. These include the summary statistics (and variogram in the case of the BBRO data) for each soil property in each of the fields, the estimated class means and the within-class variance for each property along with the *p*-value for the null hypothesis that the class means are the same. Conventionally we reject the null hypothesis if $p \leq 0.05$. The results of the regression analysis are also presented; in each case there is a regression analysis on the full set of smoothed log-ratio memberships and a regression analysis on a reduced subset. Results are also presented for regression on ECa data where these existed.

Note that the class centres which correspond to each of these analyses are presented in the same order as the results section of this chapter in Chapter 3 section 5.

Where the classes differ significantly the $MSEP_c$ is presented, calculated on the assumption that a total of 30 data are available to estimate the class means. The $MSEP_R$ is also presented for the best (as judged by the AIC) significant regression. The square root of the *MSEP* is also presented, the true value of a property should be within ± 2 *MSEP* of the prediction with 95% probability. We also present the root of MSEP as a percentage of the mean value of the property (% error).

As a measure of how good the *MSEP* for the different predictors are, we consider the case where no attempt is made to account for the spatial variation of the soil property and the overall field mean is used as the predictor. The *MSEP* of this predictor $MSEP_O$ is estimated from $\hat{\sigma}^2$, the sample variance of the soil property (or the variance determined by the maximum likelihood variogram estimator in the case of the BBRO data), using the following equation

$$MSEP_O = \hat{\sigma}^2 \left\{ 1 + \frac{1}{n} \right\} , \qquad (5.9)$$

where $n$ is the number of data used to estimate the field mean. The estimate of $MSEP_O$ and the equivalent root mean squared error and error % may be compared with those for the class mean predictors and the regression predictors.

In addition to these tables we present classified post-plots that show the spatial variability of the soil variables directly. We also present the overall variograms for soil properties along with the variograms of the error terms for the best significant regression model (where one existed). The best significant regression was the one with the smallest *MSEP*, as determined by Equation (5.3).

For those fields where the soil series was identified we present a confusion matrix which shows how the sample sites were allocated between series by the best classification tree. This shows which series can be effectively separated by prediction from yield or other data and which are confused.

The results for each section are followed by a short discussion.

#### 5.4.2.1 Hall 8 Field (BBRO project data). Manor Farm, Ixworth Suffolk.

### Summary Statistics, Hall 8.

AWt: Available water capacity (% by volume) in topsoil (5–200-mm)
AWs: Available water capacity (% by volume) in subsoil (300–450-mm)

| Soil Property | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|
| AWt | 16.0 | 1.70 | 15.3 | 16.9 | 12.6 | 18.8 |
| AWs | 15.8 | 4.18 | 14.8 | 17.5 | 10.2 | 19.7 |

### Variograms, Hall 8.

| Soil Property | Model type | C0 | C1 | A metres |
|---|---|---|---|---|
| AWt | Exp | 0.93 | 1.14 | 291 |
| AWs | Exp | 0 | 4.18 | 28 |

### Class means, Hall 8.

| Soil Property | Class | | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| AWt | 15.8 | 16.5 | 15.9 | 17.5 | 14.9 | 1.3 | <0.001 |
| AWs | 15.0 | 16.8 | 14.3 | 14.9 | 15.2 | 3.5 | <0.001 |

### Regression models, Hall 8.

Predictors: All log-ratio memberships

| Soil Property | Model type | *s* | *A* metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.3 | 95 | 9.5 | 0.09 | 1.3 | 22.3 |
| AWs | Exp | 1.0 | 23 | 13.0 | 0.02 | 3.4 | 97.2 |

Predictors: Class 2 and 5 log-ratio memberships*

| Soil Property | Model type | *s* | A metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.2 | 145 | 8.6 | 0.01 | 1.31 | 18.6 |
| AWs | Exp | 1.0 | 23 | 11.9 | 0.003 | 3.4 | 92.2 |

* Selected as these are the largest two classes and appear to account for most of the variation

**Mean Squared Error of Prediction. (assuming 30 calibration data) , Hall 8.**

AWt

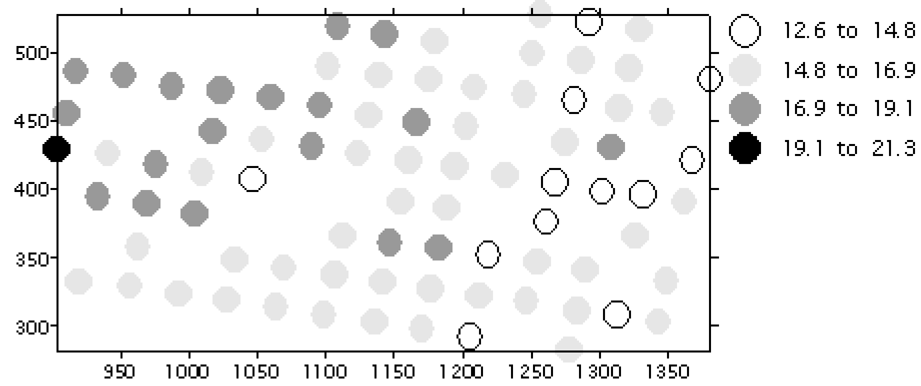
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 2.14 | 1.46 | 9.1 |
| Class mean | 1.50 | 1.20 | 7.7 |
| Best regression | 1.50 | 1.20 | 7.6 |

AWs

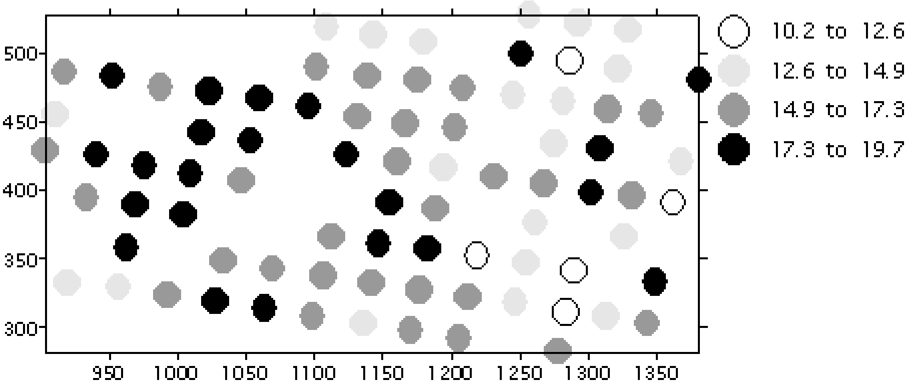
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 4.3 | 2.1 | 13.2 |
| Class mean | 4.1 | 2.0 | 12.8 |
| Best regression | 3.9 | 1.9 | 12.5 |

**Post plots of soil data, Hall 8.**

AWt (%).  Coordinates are in metres from a local datum

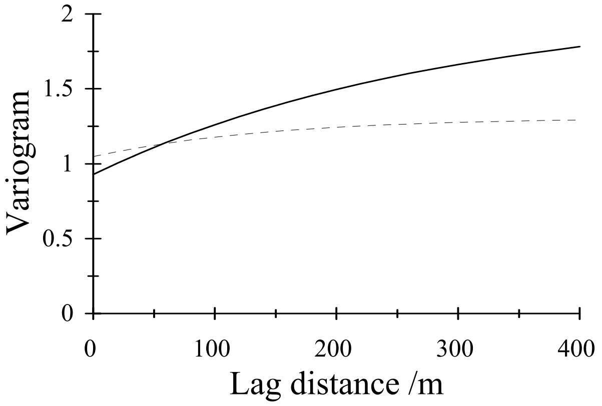


AWs (%). Coordinates are in metres from a local datum

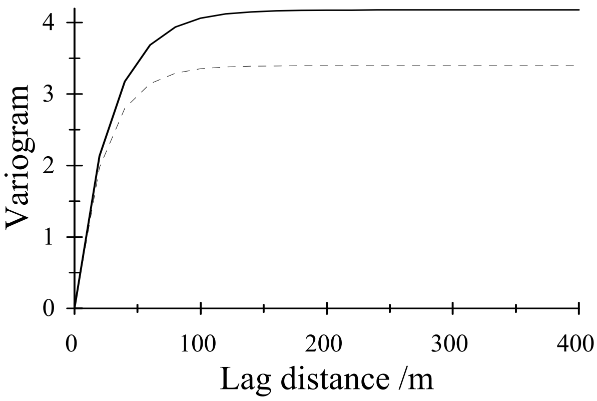


**Variograms, Hall 8.**

AWt Solid line: raw data, broken line error for best regression



AWs Solid line: raw data, broken line error for best regression



**Discussion.**

There are significant differences between the classes with respect to available water content at both depths. Classes 3 and 4, which are relatively sparse and which are confined to the margins of the field, particularly the eastern margin, have the smallest AWs, and were also low-yielding in the driest season (1996 harvest), this may be confounded with other field margin effects, although detailed field observations made during the BBRO project did confirm that the soil was notably drier at the eastern margin. The regressions of water content at both depths on memberships in the two dominant classes are significant, although the error variance is still large. The variogram for the error variance of the regression of AWt shows very little spatial structure, indicating that the regression accounts for most of the 'mappable' variation. The variogram for the error of the regression of AWs has more spatial structure suggesting that a better map of the AWs could, in principle, be made.

### 5.4.2.2 Little Lane Field (BBRO project data). IACR Broom's Barn Suffolk.

### Summary Statistics, Little Lane

| | |
|---|---|
| AWt: | Available water capacity (% by volume) in topsoil (5–200-mm) |
| AWs: | Available water capacity (% by volume) in subsoil (300–450-mm) |
| P: | Available P (mg/kg) |

| Soil Property | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|
| AWt | 17.7 | 1.20 | 17.0 | 18.6 | 15.0 | 20.5 |
| AWs | 17.8 | 1.81 | 17.7 | 18.7 | 13.2 | 51.0 |
| P | 43.1 | 868.1 | 17.2 | 65.8 | 9.8 | 118.2 |

### Variograms, Little Lane

| Soil Property | Model type | C0 | C1 | A metres |
|---|---|---|---|---|
| AWt | Nugget | 1.19 | - | - |
| AWs | Exp | 1.37 | 0.46 | 57 |
| P | Exp | 0.0 | 780.2 | 193 |

### Class means, Little Lane

| Soil Property | Class | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | | |
| AWt | 18.0 | 17.2 | 17.7 | 1.2 | 0.053 |
| AWs | 18.5 | 17.6 | 17.5 | 1.7 | 0.004 |
| P | 77.8 | 34.1 | 25.8 | 330.2 | <0.001 |

**Regression models, Little Lane**

Predictors:  All log-ratio memberships

| Soil Property | Model type | *s* | *A* metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.4 | 2.0 | 8.3 | 0.09 | 1.10 | 15.3 |
| AWs | Exp | 1.0 | 6.0 | 14.6 | 0.002 | 1.55 | 48.8 |
| P | Exp | 1.0 | 178 | 6.3 | 0.10 | 681.0 | 472.7 |

Predictors: log-ratio memberships of class 1 and 3*

| Soil Property | Model type | *s* | A metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Nugget | - | - | 7.2 | 0.03 | 1.10 | 14.2 |
| AWs | Exp | 1.0 | 7.0 | 13.7 | 0.001 | 1.55 | 47.1 |
| P | Exp | 1.0 | 173 | 6.29 | 0.04 | 663.0 | 470.8 |

* Selected as these are the dominant classes, class 2 represents the field margins

**Prediction variance (assuming 30 calibration data) , Little Lane**

AWt

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1.24 | 1.11 | 6.29 |
| Class mean | 1.32 | 1.15 | 6.49 |
| Best regression | 1.26 | 1.12 | 6.33 |

AWs

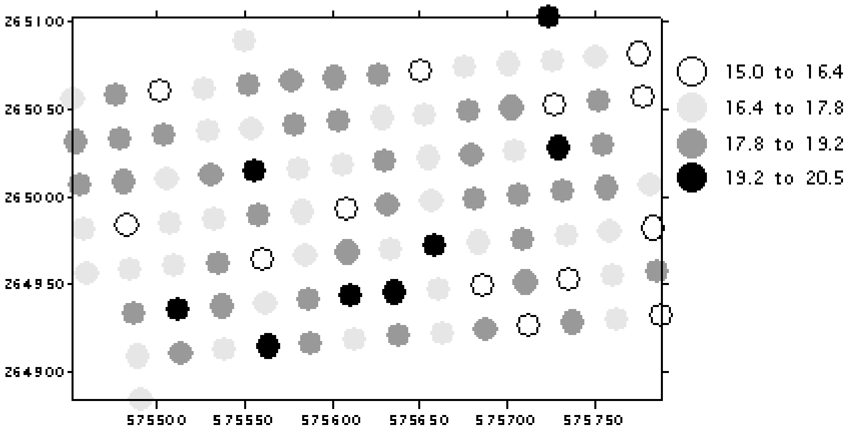
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1.89 | 1.38 | 7.73 |
| Class mean | 1.87 | 1.37 | 7.68 |
| Best regression | 1.77 | 1.33 | 7.48 |

P

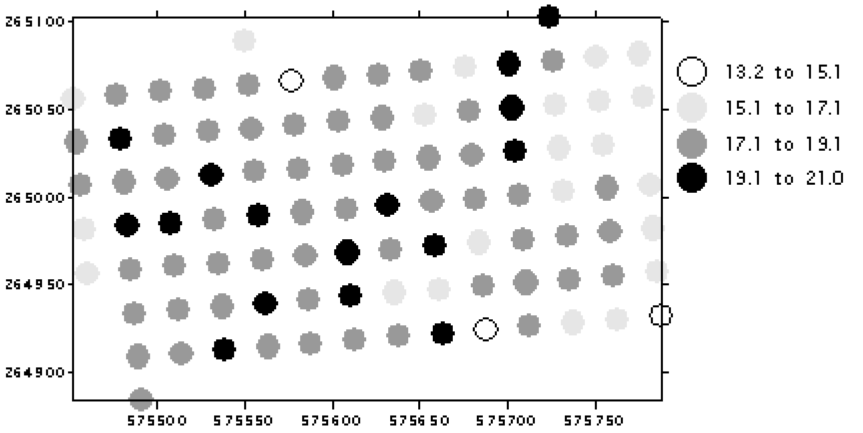
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 806.2 | 28.4 | 65.9 |
| Class mean | 363.2 | 19.0 | 44.2 |
| Best regression | 757.7 | 27.5 | 63.9 |

**Post plots of soil data, Little Lane**

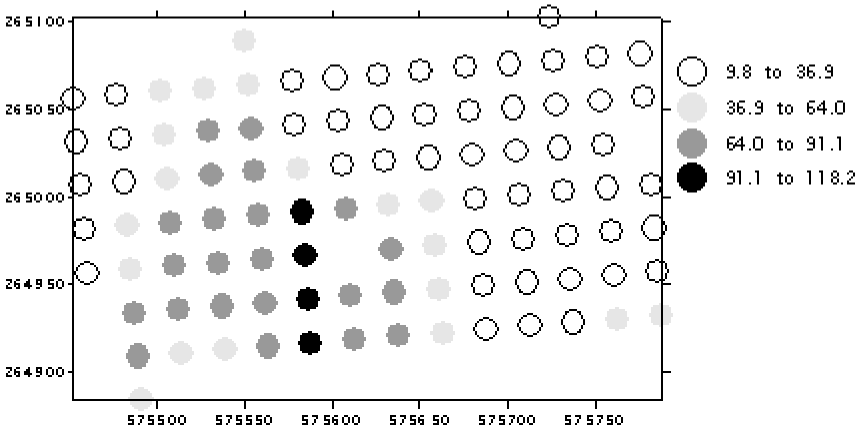
AWt (%). Coordinates are in metres



AWs (%)  Coordinates are in metres

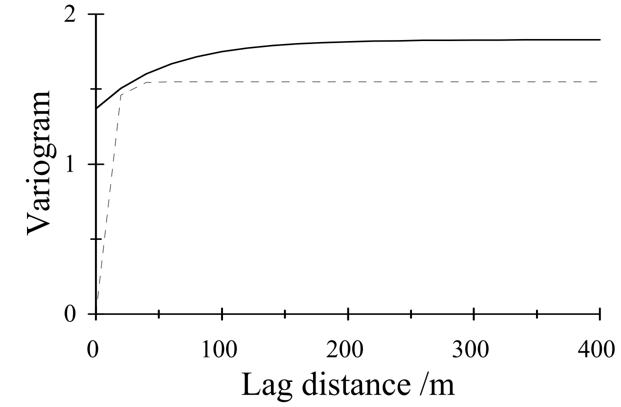


P (mg/kg)  Coordinates are in metres

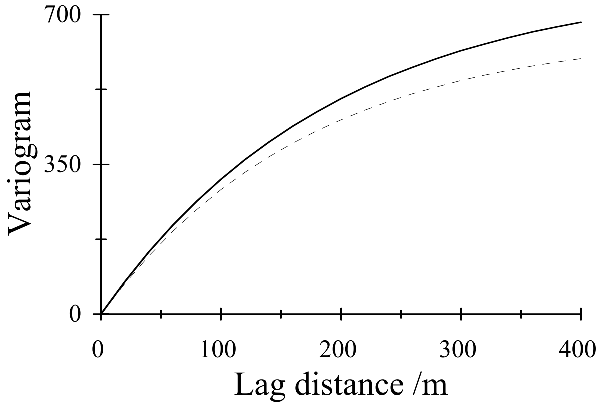


## Variograms, Little Lane

AWs



P



**Discussion.**

Class 1, with the largest yields in all three seasons also has the largest AWC and available P. Class 3, with yields consistently near average has the smallest available P. This is contrary to the hypothesis advanced by Dawson (1997) that consistently large yields, leading to larger offtake, will tend to be associated with smaller concentrations of relatively immobile nutrients. Class 2, predominantly near the field margins, shows smallest yields in the drier 1996 harvest season. AWC in the topsoil shows no spatial structure, so not surprisingly, although there is a significant regression on the memberships in classes 1 and 3, this does not give useful predictions (*MSEP* is larger for regression or the class means than it is for the field mean). The subsoil AWC is better related to the class memberships, but the *MSEP* by the regression is not much smaller than that of the field mean. The error variance of this regression is only structured at very short lags, suggesting that there is little scope for improvement. Prediction of available P by class means gives a substantial improvement over the field mean, and does better than the regression. The regression error variogram for available P is very similar to that of the original data.

**Little Lane Field (Data collected by NSRI for this project)**

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SiT (Silt content % in Topsoil), OCT (Organic Carbon % in Topsoil), AWT (Available Water content in Topsoil vol/vol), ClS, SiS, OCS, AWS are same for subsoil (45–55cm), AW1m is Available water in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 55 | 22.2 | 49.25 | 17 | 28 | 10 | 40 |
| SiT | 55 | 32.2 | 86.1 | 25 | 40 | 17 | 52 |
| OCT | 33 | 1.03 | 0.061 | 0.9 | 1.10 | 0.40 | 1.7 |
| AWT | 32 | 0.223 | 0.001 | 0.208 | 0.237 | 0.182 | 0.292 |
| ClS | 54 | 31.15 | 104.5 | 25 | 40 | 5 | 45 |
| SiS | 54 | 30.0 | 118.4 | 22 | 37 | 10 | 60 |
| OCS | 28 | 0.43 | 0.05 | 0.3 | 0.50 | 0.20 | 1.2 |
| AWS | 28 | 0.184 | 0.001 | 0.169 | 0.198 | 0.135 | 0.253 |
| AW1m | 27 | 196.1 | 354.0 | 184.0 | 207.0 | 155.0 | 245.0 |

**Class means**

| Soil Property | Class | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | | |
| ClT | 18.8 | 19.6 | 24.7 | 42.9 | **0.007** |
| SiT | 36.1 | 31.3 | 30.8 | 83.6 | 0.165 |
| OCT | 1.16 | 0.98 | 1.0 | 0.06 | 0.28 |
| AWT | 0.232 | 0.223 | 0.220 | 5.4E-4 | 0.54 |
| ClS | 29.3 | 30.8 | 32.24 | 106.9 | 0.66 |
| SiS | 28.8 | 28.0 | 31.4 | 121.0 | 0.645 |
| OCS | 0.40 | 0.40 | 0.46 | 0.053 | 0.81 |
| AWS | 0.183 | 0.193 | 0.181 | 6.4E-4 | 0.66 |
| AW1m | 197.7 | 200.2 | 194.0 | 376.5 | 0.80 |

**Regression models**

Predictors: All log-ratio memberships

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 9.58 | **<0.05** | 43.18 |
| SiT | 2.48 | >0.05 | 85.3 |
| OCT | 4.08 | >0.05 | 0.058 |
| AWT | 1.69 | >0.05 | 5.4E-4 |
| ClS | 2.61 | >0.05 | 103.3 |
| SiS | 2.30 | >0.05 | 117.7 |
| OCS | 0.59 | >0.05 | 0.052 |
| AWS | 0.54 | >0.05 | 6.5E-4 |
| AW1m | 0.21 | >0.05 | 380.1 |

Predictors: Membership in class 1 and 3

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 9.58 | **<0.05** | 43.18 |
| SiT | 2.48 | >0.05 | 85.3 |
| OCT | 4.08 | >0.05 | 0.058 |
| AWT | 1.69 | >0.05 | 5.4E-4 |
| ClS | 2.61 | >0.05 | 103.3 |
| SiS | 2.30 | >0.05 | 117.7 |
| OCS | 0.59 | >0.05 | 0.052 |
| AWS | 0.54 | >0.05 | 6.5E-4 |
| AW1m | 0.21 | >0.05 | 380.1 |

* Selected as dominant classes.

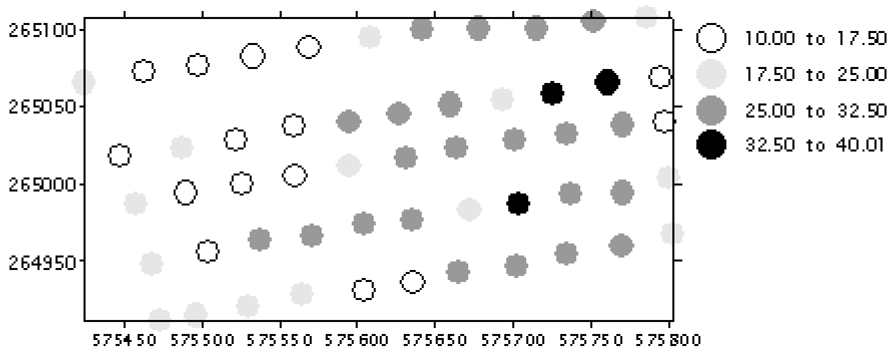
**Prediction variance (assuming 30 calibration data)**

ClT

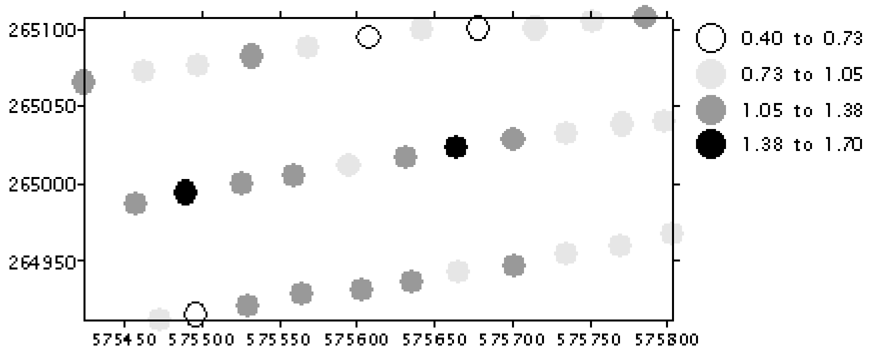
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 50.9 | 7.13 | 32 |
| Class mean | 47.2 | 6.87 | 31 |
| Best regression | 49.3 | 7.02 | 32 |

**Post plots of soil data**

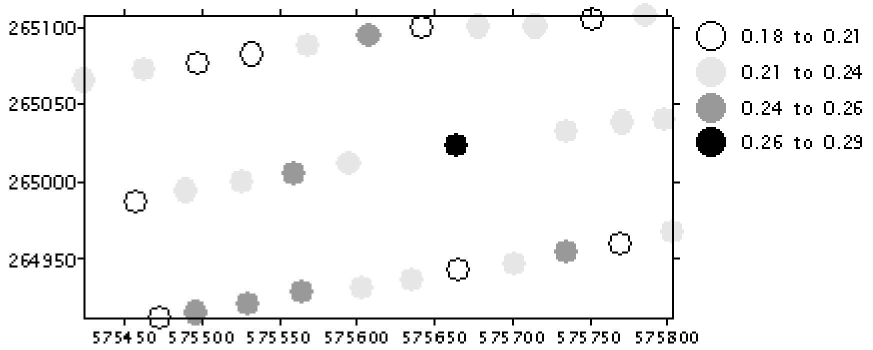
**Clay (%), Topsoil**



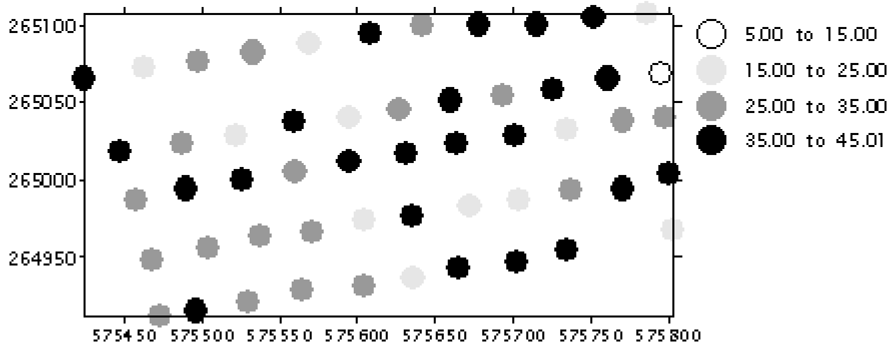
**Organic Carbon (%), Topsoil**



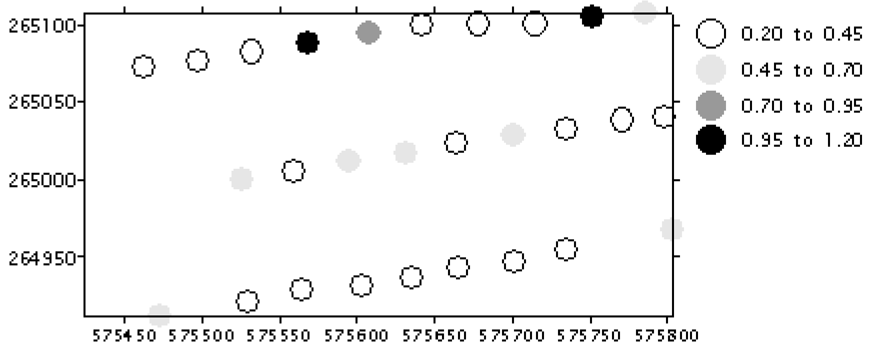
**Available Water (v/v), Topsoil.**



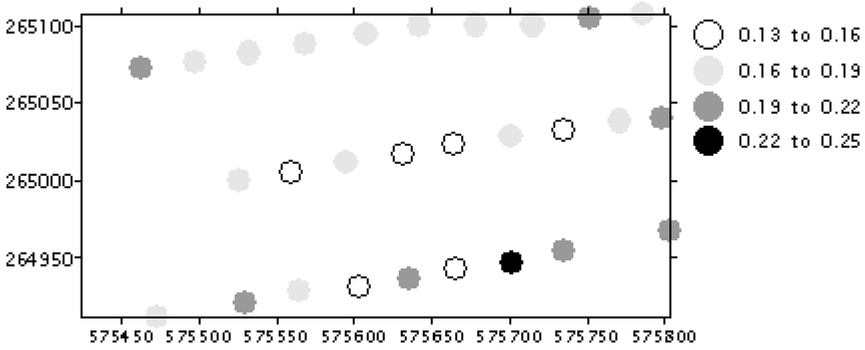
**Clay(%), Subsoil**



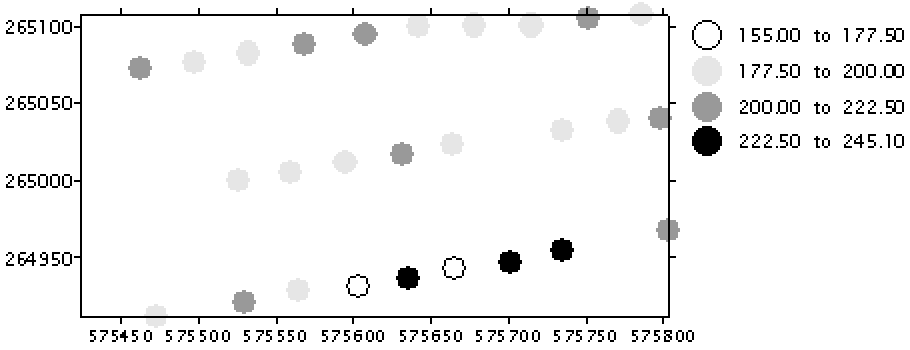
**Organic Carbon(%), Subsoil**



**Available Water (v/v), Subsoil.**



**Available Water, 1m (mm).**



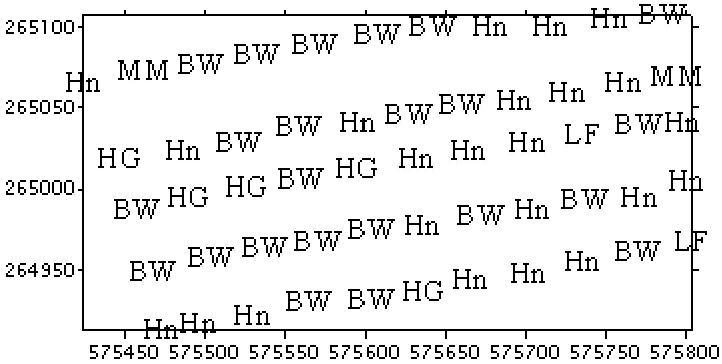
**Soil Series.**

A classification tree correctly partitioned 68% of the observations among the soil series (expected proportion correct under random allocation is 36%). The error matrix is shown below. The classification tree does not recognize the Honingham or Maplestead series at all.

**True Series**

| Predicted | BW | HG | Hn | LF | MM |
|---|---|---|---|---|---|
| **Burlingham, BW** | 21 | 4 | 9 | 0 | 0 |
| **Honingham, HG** | 0 | 0 | 0 | 0 | 0 |
| **Hanslope, Hn** | 1 | 1 | 14 | 0 | 1 |
| **Ludford, LF** | 1 | 5 | 0 | 2 | 1 |
| **Maplestead, MM** | 0 | 0 | 0 | 0 | 0 |

**Soil Series (true) at each sample location.**



## Discussion

In this instance the only soil property related to the yield classes and their memberships was clay content in the topsoil — largest in class 3 which has near-average yields throughout. The classification tree separates nearly 70% of the soil series correctly. The main confusion is where 9 instances (out of 23) of the Hanslope series are allocated to the Burlingham series. The Honingham series is not separated from the others (allocated to the Burlingham or Ludford series). Neither is the (rare) Maplestead series (allocated to Hanslope or Ludford). However, Burlingham and Hanslope differ only in texture, clay loam and clay respectively, they have the same wetness class and stoniness (chalky till). Honingham soils are coarse loamy over clayey till and Burlingham fine loamy. Ludford soils are also fine loamy but well drained in non-chalky drift. Hence the allocations are understandable in terms of series concepts and emphasize that soil series do not necessarily reflect the soil variations which are critical for crop performance.

### 5.4.2.3 Brome Pin Field (BBRO project data). IACR Broom's Barn Suffolk.

**Summary Statistics, Brome Pin**

| Soil Property | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|
| AWt | 15.5 | 0.80 | 14.9 | 16.1 | 12.9 | 17.9 |
| AWs | 14.7 | 1.87 | 13.8 | 15.5 | 11.1 | 18.5 |
| P | 24.7 | 70.45 | 17.6 | 30.2 | 11.9 | 47.0 |

AWt: Available water capacity (% by volume) in topsoil (5–200-mm)
AWs: Available water capacity (% by volume) in subsoil (300–450-mm)
P: Available P (mg/kg)

**Variograms, Brome Pin**

| Soil Property | Model type | C0 | C1 | A metres |
|---|---|---|---|---|
| AWt | Exp | 0.71 | 0.08 | 38 |
| AWs | Nugget | 1.85 | - | - |
| P | Exp | 11.4 | 102.6 | 158.7 |

**Class means, Brome Pin**

| Soil Property | Class | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| AWt | 15.8 | 15.8 | 15.4 | 15.4 | 0.79 | 0.19 |
| AWs | 14.9 | 14.6 | 14.7 | 14.6 | 1.89 | 0.67 |
| P | 23.0 | 36.0 | 28.8 | 22.6 | 61.61 | <0.001 |

**Regression models**

Predictors: All log-ratio memberships

| Soil Property | Model type | *s* | *A* metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 1.0 | 7 | 1.97 | 0.71 | 0.78 | -21.79 |
| AWs | Nugget | - | - | 5.34 | 0.25 | 1.77 | 76.28 |
| P | Exp | 1.0 | 40 | 19.1 | <0.001 | 55.2 | 423.7 |

Predictors: log-ratio memberships of class 1 and 3*

| Soil Property | Model type | *s* | A metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.1 | 36 | 0.66 | 0.79 | 0.79 | -25.1 |
| AWs | Nugget | - | - | 2.66 | 0.26 | 1.82 | 75.0 |
| P | Exp | 1.0 | 44 | 12.47 | <0.001 | 61.6 | 424.8 |

* Selected as these are the dominant classes, class 2 represents the field margins

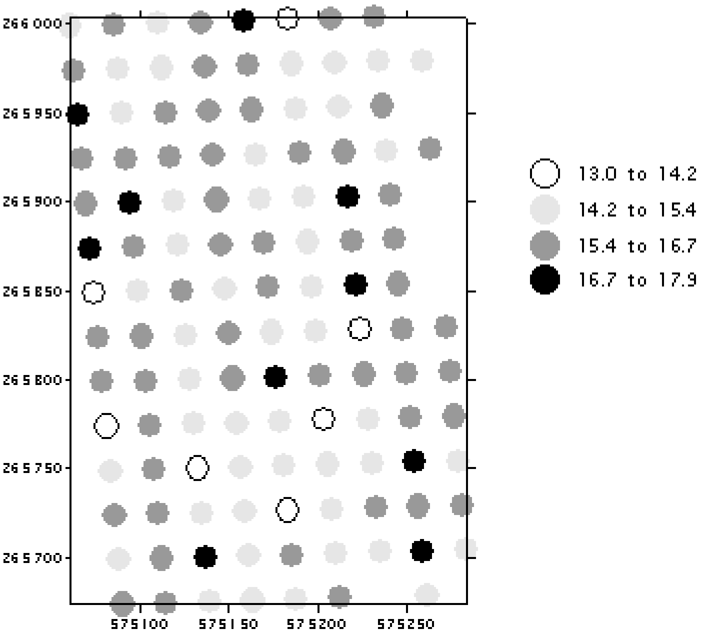
**Prediction variance (assuming 30 calibration data), Brome Pin**

P

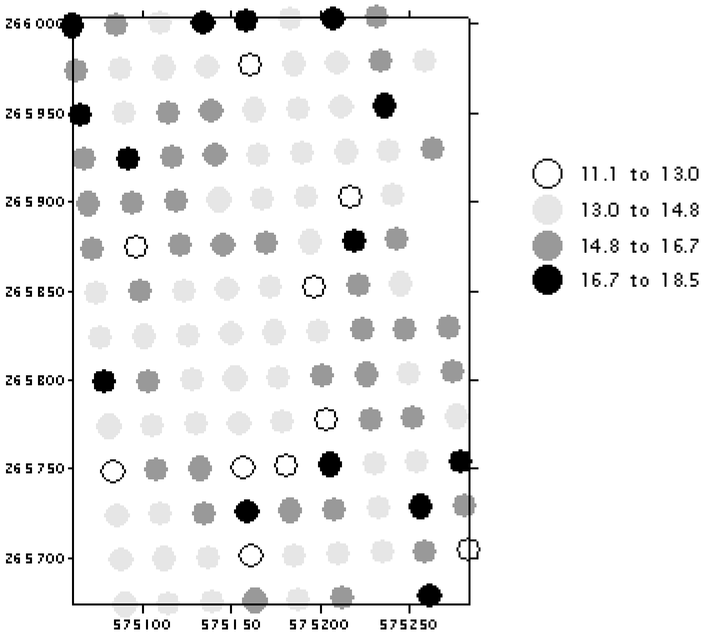
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 117.8 | 10.9 | 43.9 |
| Class mean | 69.82 | 8.36 | 33.8 |
| Best regression | 72.28 | 8.49 | 34.4 |

**Post plots of soil data, Brome Pin**

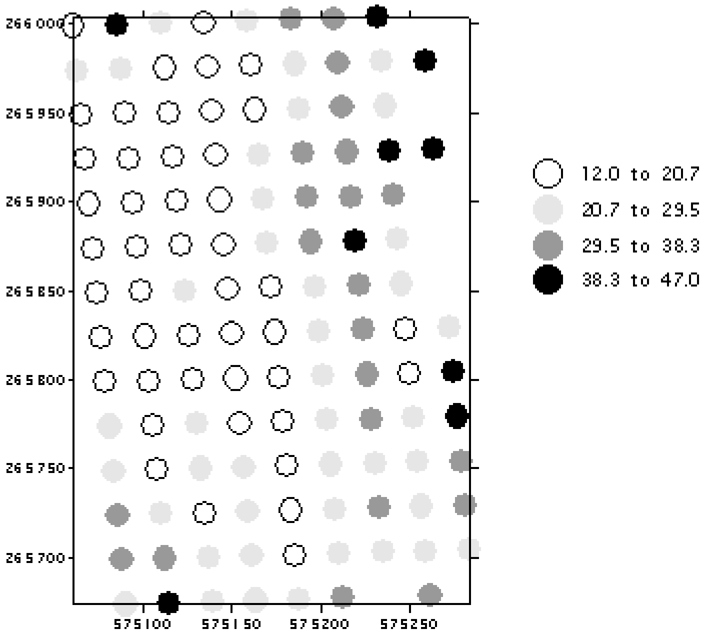
**AWt (%). Coordinates are in metres**



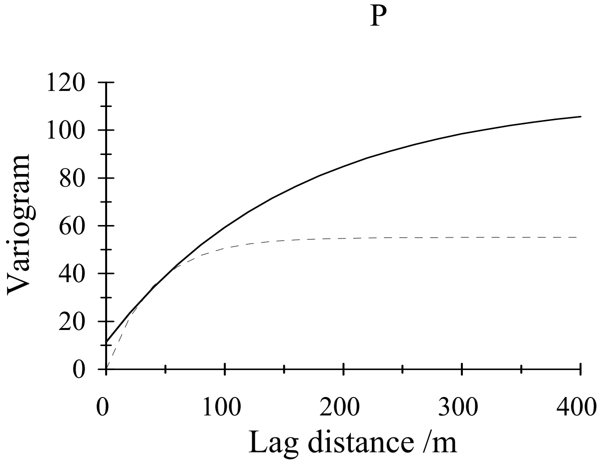
## AWs (%)  Coordinates are in metres



## P (mg/kg),  Coordinates are in metres



### Variogram, Brome Pin



**Discussion.**

Available water at both depths shows little or no spatial structure in this field with nugget and very short-range variograms, and does not differ significantly between the classes, and is not significantly related to the class memberships. Available P does differ significantly between the classes, and can be predicted with similar precision by the class mean or a regression on all the log-ratio memberships. The error variogram for this regression shows that it removes much of the spatially structured variation in available P in this field. Concentrations of available P are largest in class 2, which is the lowest yielding in both seasons, and are lowest in class 4, which is highest yielding. This does support the 'offtake' hypothesis referred to in the discussion of Little Lane field.

**Brome Pin, (Data collected by NSRI for this project).**

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), ClS, SaS,AWC is available water (APTAB) in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 61 | 15.2 | 17.16 | 12 | 15 | 10 | 30 |
| SiT | 61 | 23.49 | 18.55 | 20 | 25 | 15 | 35 |
| ClS | 60 | 16.77 | 57.1 | 11 | 20 | 5 | 40 |
| SiS | 60 | 19.65 | 39.55 | 15 | 25 | 5 | 30 |
| AWC | 61 | 144.2 | 131.5 | 148 | 148 | 119 | 155 |

**Class means**

| Soil Property | Class | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| ClT | 15.8 | 15.0 | 14.1 | 15.7 | 17.46 | 0.58 |
| SiT | 23.9 | 22.5 | 23.7 | 23.1 | 19.4 | 0.92 |
| ClS | 19.1 | 16.0 | 12.9 | 17.9 | 53.5 | 0.07 |
| SiS | 21.8 | 24.0 | 18.5 | 18.7 | 38.86 | 0.256 |
| AWC | 148.8 | 151.5 | 138.6 | 144.7 | 120.8 | **0.04** |

**Regression models**

**Predictors: All log-ratio memberships**

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 1.09 | >0.05 | 17.72 |
| SiT | 0.32 | >0.05 | 19.42 |
| ClS | 7.91 | >0.05 | 52.70 |
| SiS | 1.15 | >0.05 | 40.83 |
| AWC | 6.15 | >0.05 | 124.9 |

**Predictors: Membership in classes 3 and 4.**

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.63 | >0.05 | 17.50 |
| SiT | 0.16 | >0.05 | 19.14 |
| ClS | 7.9 | <0.05 | 51.91 |
| SiS | 1.09 | >0.05 | 40.17 |
| AWC | 6.26 | <0.05 | 122.8 |

* Selected as dominant classes.

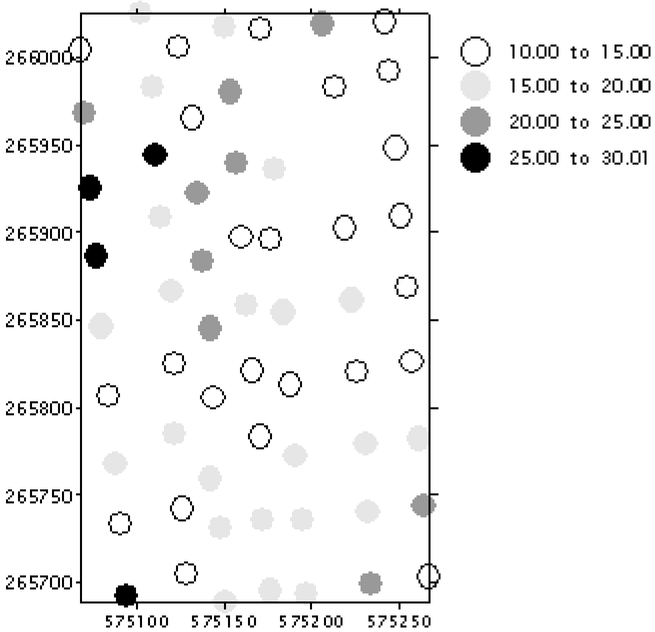
**Prediction variance (assuming 30 calibration data)**

**AWC**

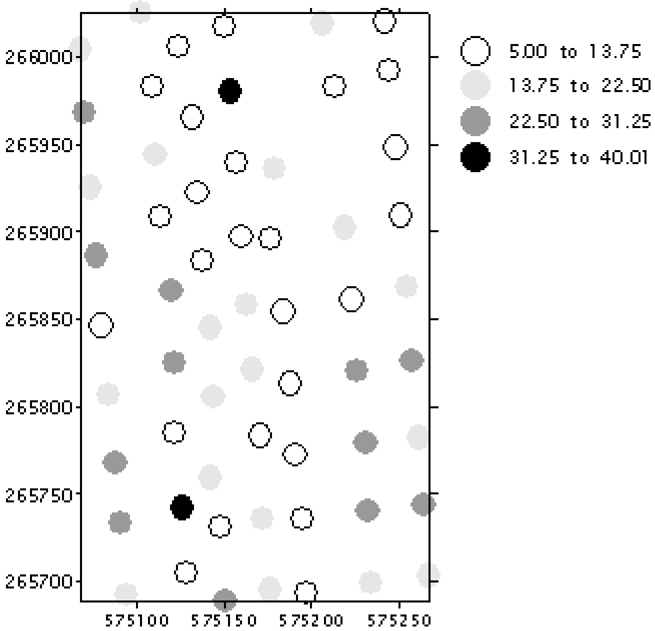
| **Predictor** | **MSEP** | **RMSEP** | **Error %** |
|---|---|---|---|
| Overall mean | 135.9 | 11.66 | 8.1 |
| Class mean | 136.9 | 11.70 | 8.1 |
| Best regression | 140.3 | 11.85 | 8.2 |

**Post plots of soil data**

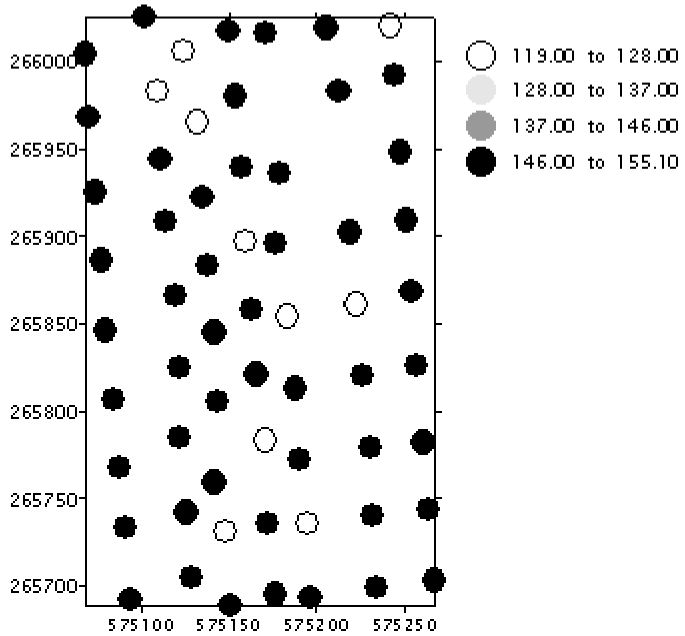
### Clay (%), Topsoil



### Clay (%), Subsoil



**AWC (mm)**



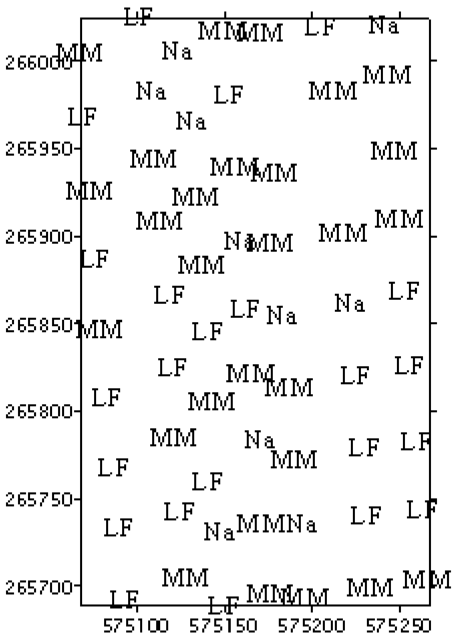
**Soil Series.**

A classification tree correctly partitioned 67% of the observations among the soil series (expected proportion correct under random allocation is 21%). The error matrix is shown below.

**True Series**

| **Predicted** | **LF** | **MM** | **Na** |
|---|---|---|---|
| ***Ludford, LF*** | 12 | 9 | 2 |
| **Maplestead, MM** | 2 | 25 | 1 |
| **Newport, Na** | 0 | 6 | 4 |

**Soil Series (true) at each sample location.**



**Discussion.**

AWC (APTAB) differs significantly, although not very markedly, between the classes.  It is smallest in class 3 (yielding relatively poorly in the dry 1996 harvest season) but largest in class 2 with the lowest yield in both seasons, so a simple interpretation is not possible.  All three soil series in this field are distinguished by the classification tree, using the class memberships as predictors.  The main errors are due to allocation of 9 instances (out of 40) of the Maplestead series to Ludford, and 6 to the Newport series. Newport, Maplestead and Ludford are similar soils, differing only in the proportions of loamy sand, sandy loam and clay loam layers (respectively).

#### 5.4.2.4 Commissioners Field (BBRO project data). Yokefleet farm, Humberside.

### Summary Statistics, Commissioners field

AWt: Available water capacity (% by volume) in topsoil (5–200-mm)
AWs: Available water capacity (% by volume) in subsoil (300–450-mm)
P: Available P (mg/kg)

| Soil Property | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|
| AWt | 21.4 | 0.64 | 21.0 | 22.0 | 18.7 | 22.5 |
| AWs | 21.1 | 1.32 | 20.5 | 22.0 | 17.8 | 22.5 |
| P | 13.9 | 14.06 | 11.4 | 16.0 | 7.8 | 26.6 |

### Variograms, Commissioners field

| Soil Property | Model type | C0 | C1 | A metres |
|---|---|---|---|---|
| AWt | Exp | 0.03 | 0.66 | 69.8 |
| AWs | Exp | 0.27 | 1.07 | 41.9 |
| P | Exp | 8.4 | 5.6 | 43.0 |

### Class means, Commissioners field

| Soil Property | Class | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | | |
| AWt | 21.8 | 21.2 | 21.1 | 0.55 | <0.001 |
| AWs | 21.6 | 20.8 | 20.8 | 1.18 | 0.003 |
| P | 14.1 | 14.5 | 13.8 | 14.4 | 0.9 |

**Regression models, Commissioners Field**

Predictors: All log-ratio memberships

| Soil Property | Model type | *s* | *A* metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.9 | 62 | 14.9 | 0.002 | 0.52 | -84.5 |
| AWs | Exp | 0.9 | 26 | 14.6 | 0.002 | 1.03 | 0.41 |
| P | Exp | 0.5 | 39 | 2.9 | 0.40 | 13.7 | 214.9 |

Predictors: log-ratio memberships of class 1 and 3*

| Soil Property | Model type | *s* | A metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| AWt | Exp | 0.8 | 70 | 12.9 | 0.002 | 0.52 | -83.9 |
| AWs | Exp | 0.9 | 24 | 14.1 | <0.001 | 1.02 | -0.55 |
| P | Exp | 0.4 | 49 | 1.01 | 0.60 | 13.9 | 213.9 |

* Selected as these are the dominant classes, class 2 represents the field margins

**Prediction variance (assuming 30 calibration data), Commissioners Field**

AWt

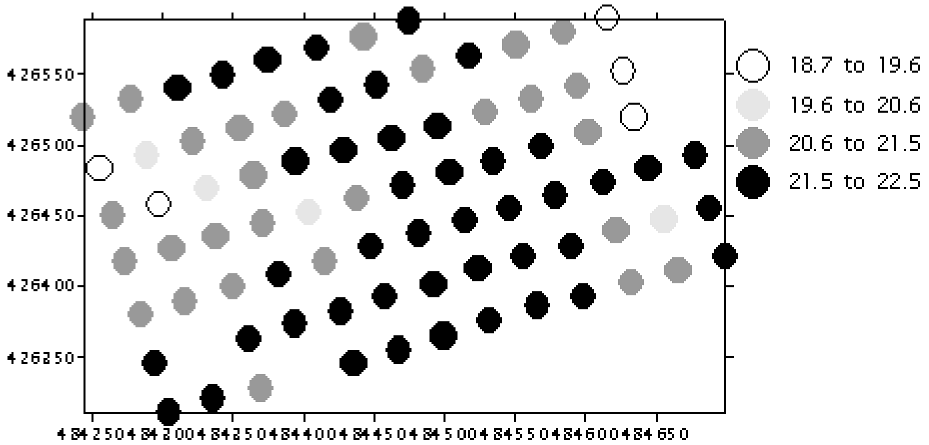
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.71 | 0.84 | 3.9 |
| Class mean | 0.61 | 0.78 | 3.6 |
| Best regression | 0.64 | 0.80 | 3.7 |

AWs

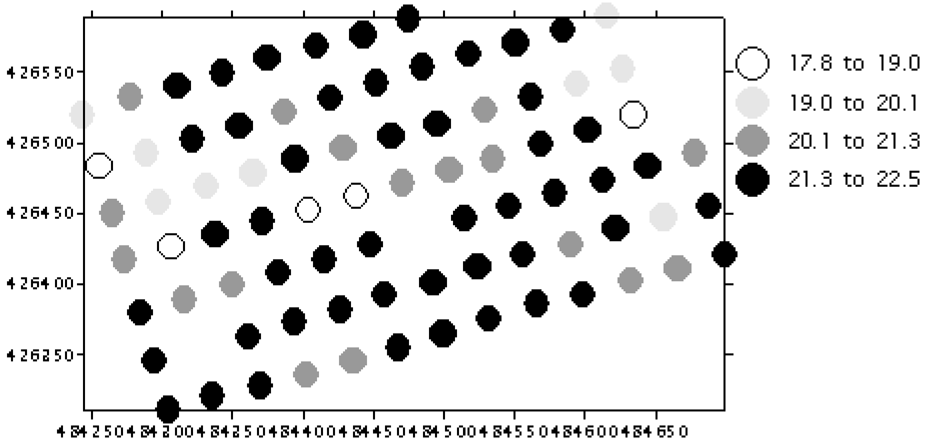
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1.38 | 1.18 | 5.58 |
| Class mean | 1.30 | 1.14 | 5.40 |
| Best regression | 1.25 | 1.12 | 5.29 |

**Post plots of soil data, Commissioners Field**

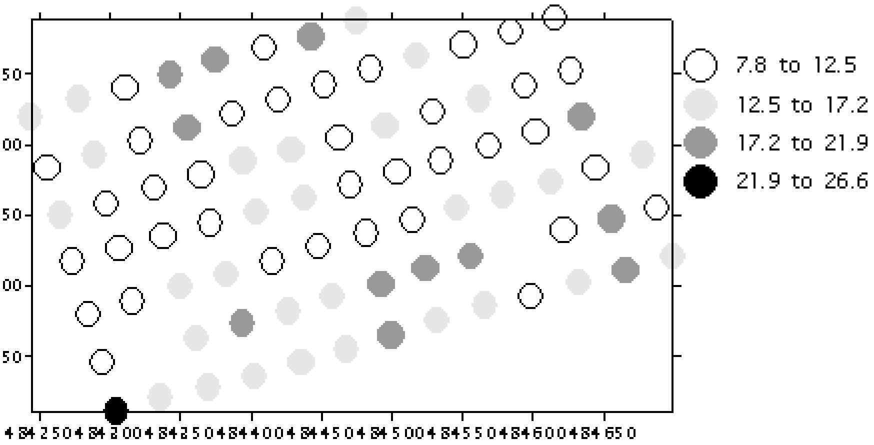
AWt (%). Coordinates are in metres



AWs(%) Coordinates are in metres

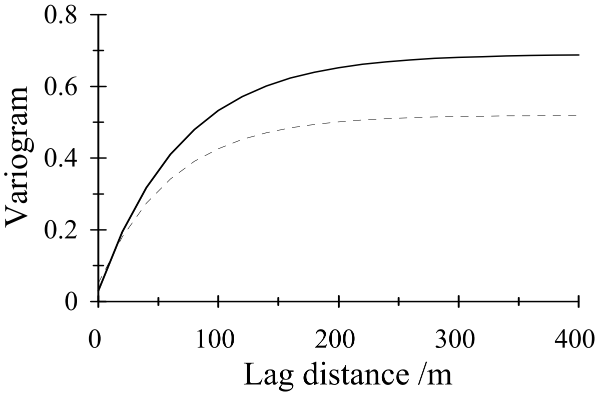


P (mg/kg), Coordinates are in metres

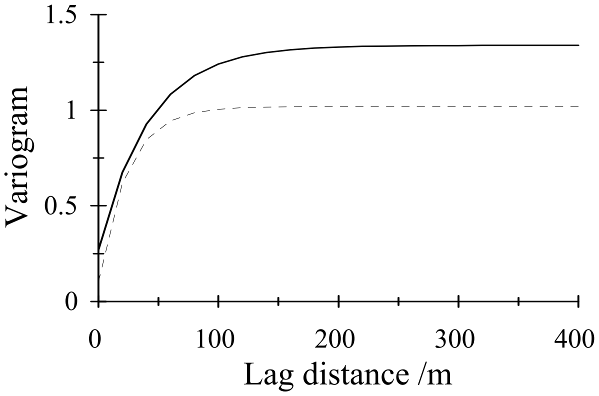


## Variograms, Commissioners Field

AWt



AWs



**Discussion.**

Only the available water data are significantly related to the information on yield in this field. However, at both depths the improvements in predictions over the field mean are relatively small and the error variograms for the best regressions show that a good deal of spatially structured variation remains unexplained.

### 5.4.2.5 South Warpings Field (BBRO project data). Yokefleet farm, Humberside.

**Summary Statistics, South Warpings Field**

P: Available P (mg/kg)

| Soil Property | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|
| P | 49.4 | 551.1 | 31.2 | 71.1 | 13.9 | 109.4 |

**Variograms, South Warpings Field**

| Soil Property | Model | C0 | C1 | A metres |
|---|---|---|---|---|
| P | Exp | 33.3 | 632.1 | 167 |

**Class means, South Warpings Field**

| Soil Property | Class | | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| P | 18.5 | 82.3 | 36.2 | 46.3 | 56.4 | 292.7 | <0.001 |

**Regression models, South Warpings field**

Predictors: All log-ratio memberships

| Soil Property | Model type | *s* | *A* metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| P | Exp | 0.9 | 143 | 14.3 | 0.014 | 435.7 | 573.16 |

Predictors: log-ratio memberships of classes 2, 3 and 5*

| Soil Property | Model type | *s* | A metres | Wald statistic | *p* | Error variance | *a* |
|---|---|---|---|---|---|---|---|
| P | Exp | 0.9 | 160 | 9.26 | 0.026 | 482.0 | 573.75 |

* Selected as these are the dominant classes.

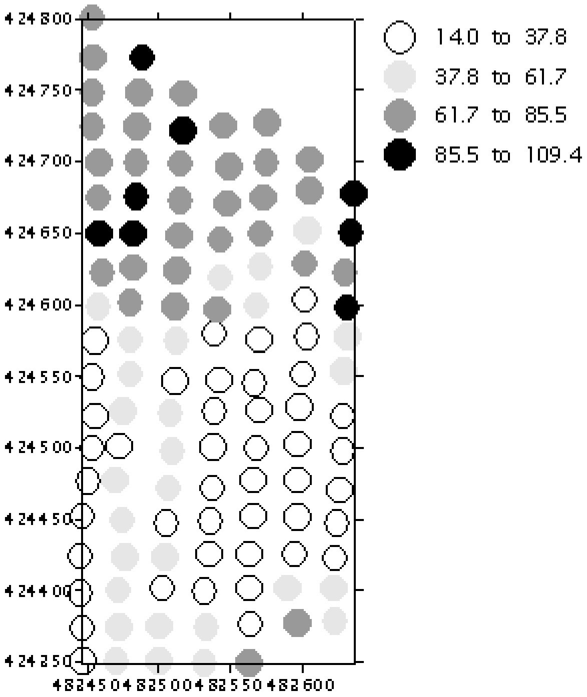
**Prediction variance (assuming 30 calibration data), South Warpings Field**

P

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 687.2 | 26.2 | 53.0 |
| Class mean | 341.5 | 18.5 | 37.4 |
| Best regression | 610 | 24.7 | 50.0 |

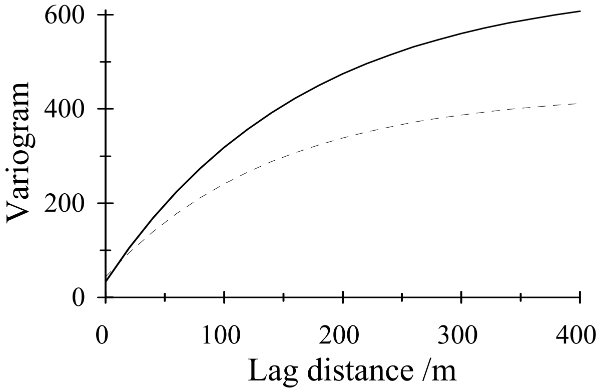
**Post plot of soil data, South Warpings Field**

P (mg/kg),  Coordinates are in metres



**Variograms, South Warpings Field**

P



**Discussion.**

The available P data are significantly related to the information on yield here, but all the zone means are above the values where a yield response to P would be expected.  As with the data on Little lane field the class means are the best predictors of this property, and the error variogram of the best regression shows substantial spatially structured variation that is unaccounted for.  In this field the yields vary markedly between seasons with no consistently high or low regions apart from the relatively rare class 4.

### 5.4.2.6 Field 2, Flawborough farm.

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), OCT (Organic Carbon % in Topsoil), AWT (Available Water content in Topsoil vol/vol), ClS, SaS, OCS, AWS are same for subsoil (45–55cm), AW1m is Available water in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 47 | 37.6 | 78.73 | 30 | 45 | 20 | 55 |
| SaT | 47 | 15.2 | 57.75 | 10 | 20 | 5 | 45 |
| OCT | 24 | 2.95 | 0.630 | 2.3 | 3.7 | 1.8 | 4.4 |
| AWT | 23 | 0.259 | 3.3E-4 | 0.246 | 0.270 | 0.23 | 0.306 |
| ClS | 49 | 35.6 | 353.9 | 15 | 50 | 2 | 57 |
| SaS | 49 | 27.8 | 770.4 | 7.5 | 56.2 | 2 | 85 |
| OCS | 26 | 0.71 | 0.111 | 0.40 | 0.90 | 0.30 | 1.60 |
| AWS | 24 | 0.207 | 0.001 | 0.191 | 0.226 | 0.137 | 0.241 |
| AW1m | 20 | 221.6 | 312.6 | 210.0 | 236.5 | 176.0 | 243.0 |

**Class means**

| Soil Property | Class |  |  |  |  | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 |  |  |
| ClT | 34.1 | 30.0 | 42.3 | 37.0 | 39.3 | 75.02 | >0.05 |
| SaT | 13.1 | 20.0 | 15.0 | 10.0 | 16.0 | 56.6 | >0.05 |
| OCT | 3.15 | 3.0 | 2.35 | 2.90 | 2.99 | 0.714 | >0.05 |
| AWT | 0.273 | 0.278 | 0.249 | 0.267 | 0.251 | 2.7E-4 | **<0.05** |
| ClS | 30.71 | 28.25 | 45.0 | 30.0 | 32.73 | 364.7 | >0.05 |
| SaS | 33.3 | 43.5 | 8.0 | 38.0 | 24.6 | 767.5 | >0.05 |
| OCS | 0.450 | 0.50 | 1.0 | 0.70 | 0.75 | 0.108 | >0.05 |
| AWS | 0.194 | 0.212 | 0.235 | 0.176 | 0.210 | 5.79E-4 | >0.05 |
| AW1m | 217.5 | 240.0 | 238.5 | 203.5 | 221.5 | 287.1 | >0.05 |

**Regression models**

Predictors: All log-ratio memberships

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 6.92 | >0.05 | 74.02 |
| SaT | 1.57 | >0.05 | 60.98 |
| OCT | 6.86 | >0.05 | 0.56 |
| AWT | 11.38 | **<0.05** | 2.6E-4 |
| ClS | 4.22 | >0.05 | 352.4 |
| SaS | 8.93 | >0.05 | 698.6 |
| OCS | 4.61 | >0.05 | 0.109 |
| AWS | 7.07 | >0.05 | 5.9E-4 |
| AW1m | 4.86 | >0.05 | 299.0 |

Predictors: Membership in class 4 and 5

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 1.21 | >0.05 | 80.1 |
| SaT | 0.40 | >0.05 | 59.84 |
| OCT | 6.38 | **<0.05** | 0.529 |
| AWT | 6.64 | **<0.05** | 2.86E-4 |
| ClS | 2.09 | >0.05 | 353.3 |
| SaS | 4.54 | >0.05 | 731.7 |
| OCS | 2.85 | >0.05 | 0.107 |
| AWS | 6.92 | **<0.05** | 5.5E-4 |
| AW1m | 5.50 | **<0.05** | 263.9 |

* Selected as dominant classes.

**Prediction variance (assuming 30 calibration data)**

AWT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.00034 | 0.0185 | 7.13 |
| Class mean | 0.00032 | 0.0177 | 6.85 |
| Best regression | 0.00033 | 0.0181 | 6.98 |

AWS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.0010 | 0.032 | 15.53 |
| Best regression | 0.0006 | 0.025 | 12.11 |

OCT

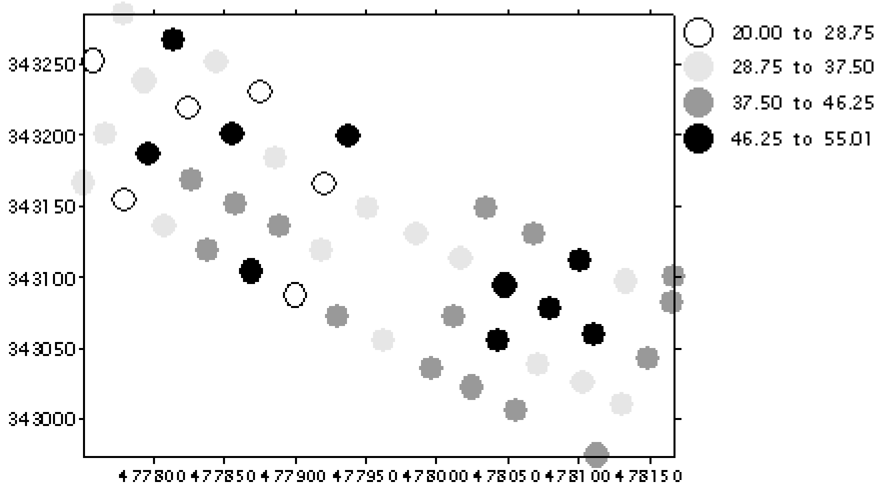
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.651 | 0.81 | 27.4 |
| Best regression | 0.605 | 0.78 | 26.4 |

AW1m

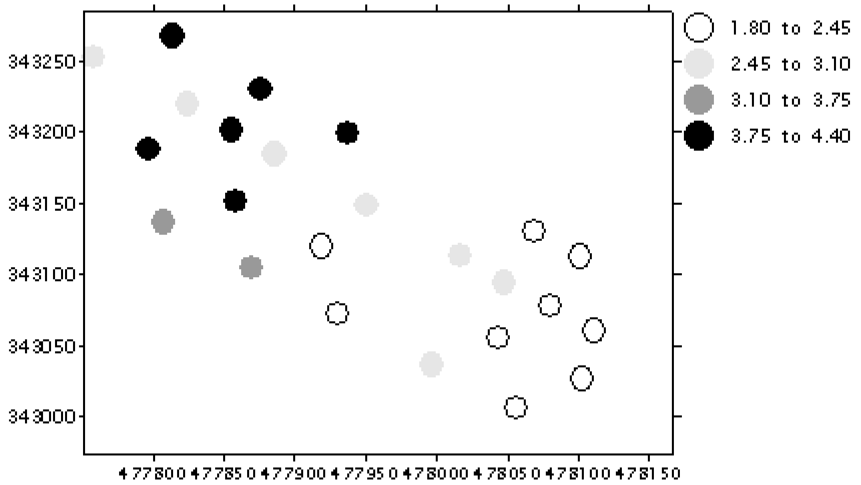
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 323.02 | 17.97 | 8.11 |
| Best regression | 301.6 | 17.37 | 7.84 |

## Post plots of soil data

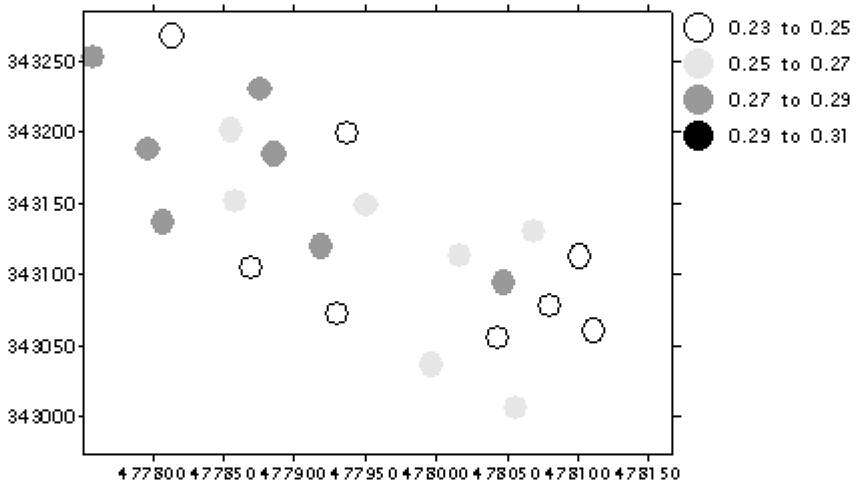
### Clay(%), Topsoil



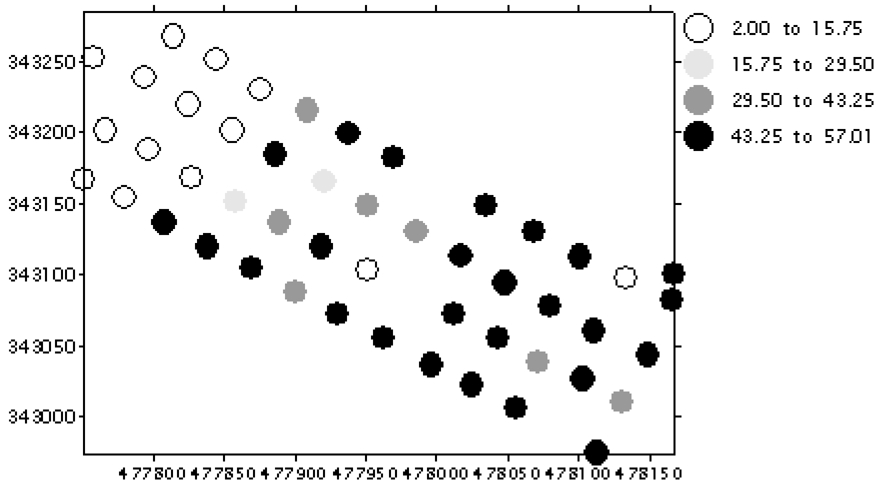
### Organic Carbon(%), Topsoil



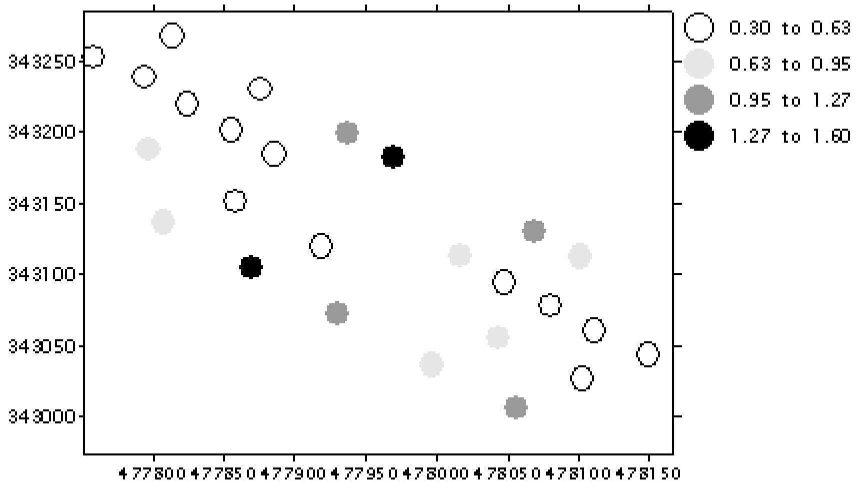
### Available Water(v/v), Topsoil.



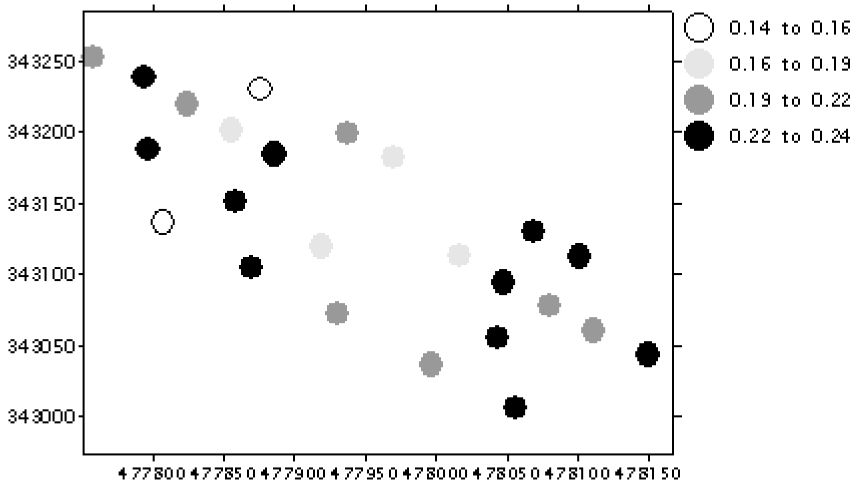
## Clay(%), Subsoil



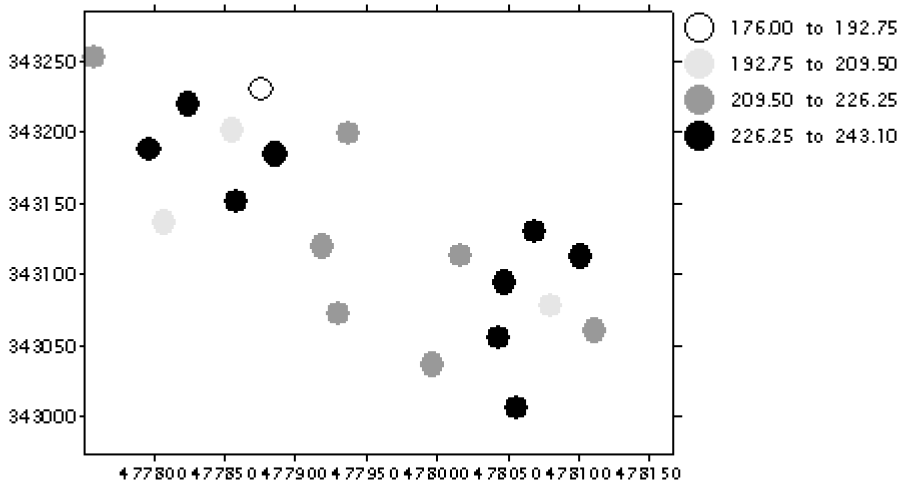
## Organic Carbon(%), Subsoil



## Available Water(v/v), Subsoil.



**Available Water, 1m (mm).**



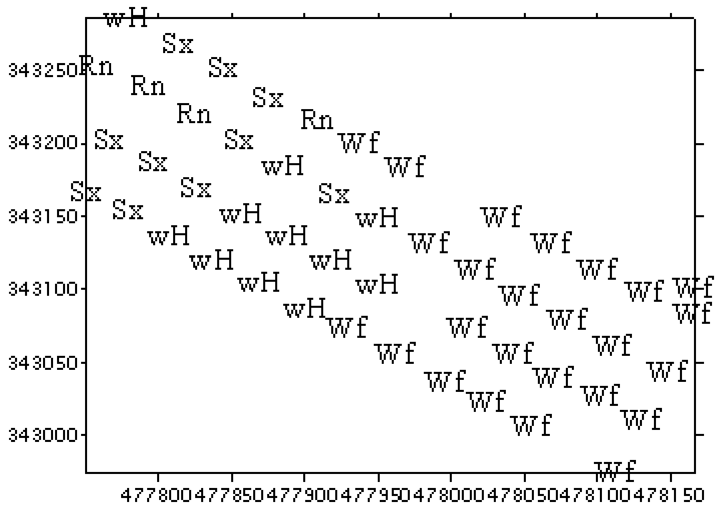
**Soil Series.**

A classification tree correctly partitioned 70% of the observations among the soil series (expected proportion correct under random allocation is 34%). The error matrix is shown below. Note that most of the observations are in the Worcester series. The classification tree does not recognize the Rheidol series at all.

**True Series**

| **Predicted** | **Rheidol Rn** | **Shipton Sx** | **WorcesterWf** | **Wyre wH** |
|---|---|---|---|---|
| ***Rheidol Rn*** | 0 | 0 | 0 | 0 |
| **Shipton Sx** | 0 | 6 | 0 | 2 |
| **WorcesterWf** | 3 | 1 | 24 | 4 |
| **Wyre wH** | 1 | 3 | 1 | 5 |

**Soil Series (true) at each sample location.**



## Discussion.

The mean available water capacity in the top soil for the five classes (maximum membership) derived from the yield data differ significantly, and this variable and available water capacity in the subsoil, the organic carbon (topsoil), and the available water capacity in the top metre are significantly related to the membership values in the two dominant classes. There is no simple interpretation of these differences with respect to the yield pattern which each class represents. The two dominant classes are 4 and 5, class 5 has the smallest yields in the dry harvest season of 1995 but has the larger available water capacity over the top metre. The best prediction by regression on the class memberships is of available water in the subsoil, but a good deal of variation remains unaccounted for in all cases.

The classes derived from the yield data do not show a very clear spatial pattern, unlike the soil series, but the classification tree using memberships as predictors achieves a reasonable separation of the soil series. The main confusion is where four (out of eleven) instances of the Wyre series are allocated to the Worcester series. Wyre and Worcester soils are all derived from reddish clayey parent materials, Wyre from river alluvium and Worcester from in situ mudstones.

### 5.4.2.7 Gate field, Whittlesford farm.

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SiT, (Silt content, % in Topsoil), OCT (Organic Carbon % in Topsoil), ClS, SiT,OCS are same for subsoil (45–55cm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 24 | 16.83 | 25.10 | 13.5 | 20 | 10 | 28 |
| SiT | 24 | 33.1 | 86.5 | 26.5 | 39 | 15 | 53 |
| OCT | 15 | 1.30 | 0.036 | 1.2 | 1.4 | 1.0 | 1.7 |
| ClS | 25 | 17.48 | 48.43 | 14.0 | 25.0 | 8 | 30 |
| SiS | 25 | 37.4 | 199.3 | 28.75 | 47.75 | 11 | 63 |
| OCS | 14 | 0.421 | 0.008 | 0.4 | 0.5 | 0.3 | 0.6 |

NB Since much of the field-work was conducted in an area adjacent to the yield mapped crop only 24-25 data are available on clay and 14-15 on carbon.

**Class means**

| Soil Property | Class | | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| ClT | 17.3 | 15.9 | 17.0 | 17.7 | 15.0 | 29.25 | 0.946 |
| SiT | 36.8 | 32.0 | 25.0 | 34.8 | 25.0 | 89.6 | 0.524 |
| OCT | 1.30 | 1.18 | 1.40 | 1.35 | 1.30 | 0.04 | 0.659 |
| ClS | 16.0 | 18.6 | 9.0 | 19.2 | 11.5 | 48.49 | 0.410 |
| SiS | 30.35 | 44.6 | 22.0 | 39.9 | 21.0 | 108.7 | 0.08 |
| OCS | 0.45 | 0.35 | 0.50 | 0.47 | 0.30 | 0.006 | **0.036** |

**Regression models**

Predictors:  All log-ratio memberships

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 2.17 | >0.05 | 28.61 |
| SiT | 2.88 | >0.05 | 95.53 |
| OCT | 0.42 | >0.05 | 0.048 |
| ClS | 3.84 | >0.05 | 50.86 |
| SiS | 5.24 | >0.05 | 197.3 |
| OCS | 14.55 | **<0.05** | 0.004 |

Predictors: Membership in class 4 and 5

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 1.16 | >0.05 | 26.04 |
| SiT | 2.32 | >0.05 | 85.33 |
| OCT | 0.27 | >0.05 | 0.041 |
| ClS | 0.43 | >0.05 | 51.81 |
| SiS | 2.94 | >0.05 | 191.8 |
| OCS | 10.55 | **<0.01** | 0.0048 |

* Selected as least-correlated memberships.

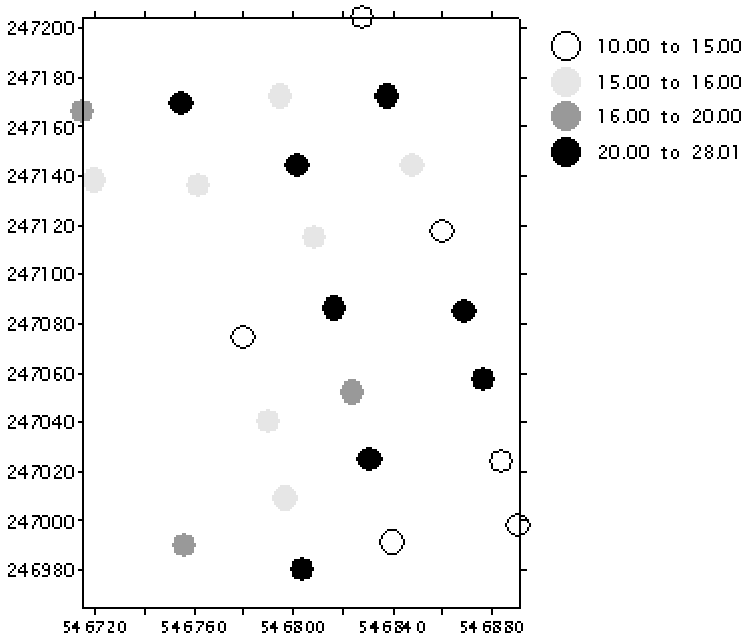
**Prediction variance (assuming 30 calibration data)**

OCS

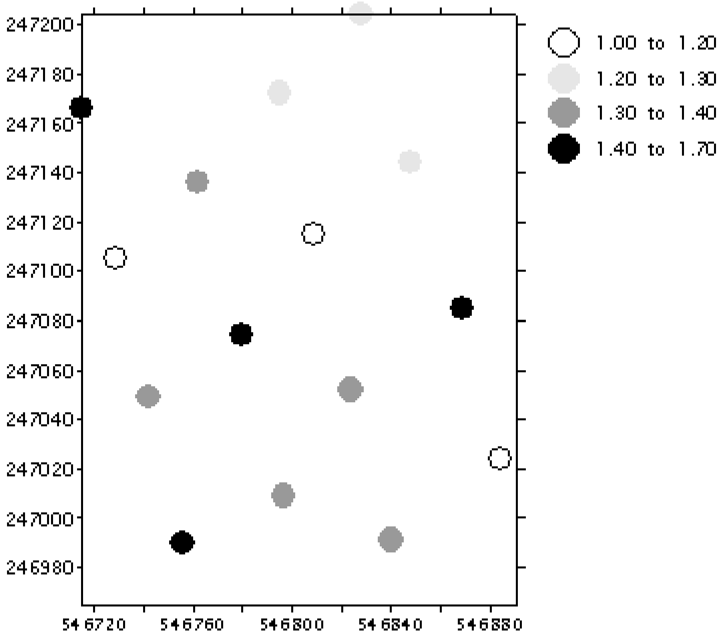
| **Predictor** | **MSEP** | **RMSEP** | **Error %** |
|---|---|---|---|
| Overall mean | 0.008 | 0.091 | 21.6 |
| Class mean | 0.007 | 0.084 | 19.9 |
| Best regression | 0.005 | 0.074 | 17.6 |

## Post plots of soil data

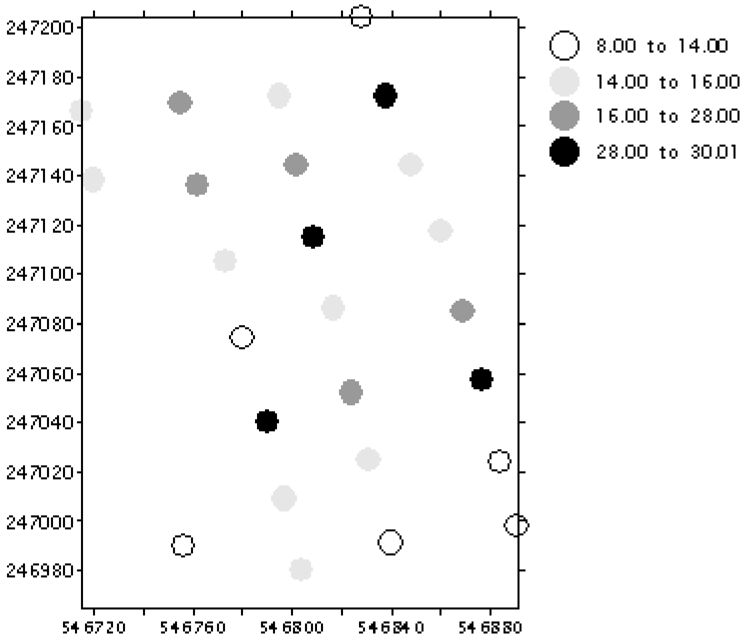
Clay(%), Topsoil



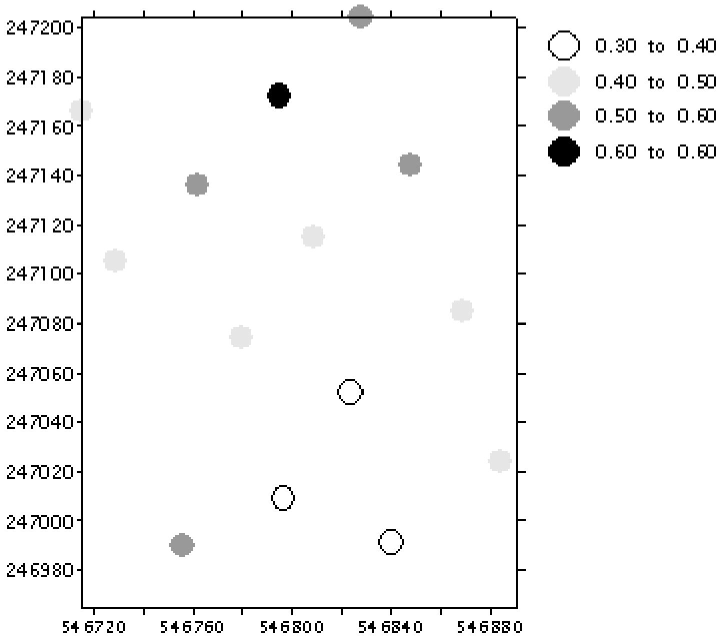
Organic Carbon(%), Topsoil



## Clay(%), Subsoil



## Organic Carbon(%), Subsoil



## **Discussion**.

The classes defined from yield data in this field show strong spatial structure and contrasting yield patterns - no part of the field yielded consistently above average over the period.  Unfortunately only 25 or fewer observations are available of soil properties because much of the field-work was conducted in an area which had not been mapped.  Not surprisingly evidence for differences between these classes is non existent, with the exception of the organic carbon content of the subsoil.

### 5.4.2.8 Top Pavements Field. ADAS Boxworth.

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), ClS, SaS, AWC is available water (APTAB) in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 42 | 46.3 | 54.67 | 45 | 50 | 30 | 55 |
| SaT | 42 | 6.1 | 4.31 | 5 | 5 | 5 | 10 |
| ClS | 42 | 50.9 | 19.8 | 50 | 50 | 40 | 60 |
| SaS | 42 | 9.6 | 10.2 | 10 | 10 | 5 | 15 |
| AWC | 42 | 140.7 | 7.2 | 138 | 144 | 135 | 144 |

**Class means**

| Soil Property | Class | | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| ClT | 44.5 | 48.6 | 50.6 | 44.1 | 40.0 | 49.5 | 0.08 |
| SaT | 6 | 5 | 6.9 | 6 | 8.3 | 3.9 | 0.07 |
| ClS | 51.5 | 49.1 | 50.6 | 53.5 | 48.3 | 18.5 | 0.142 |
| SaS | 9.0 | 8.64 | 12.5 | 8.5 | 11.7 | 8.48 | 0.01 |
| AWC | 143.2 | 140.3 | 139.4 | 140.1 | 140 | 5.75 | 0.006 |

Regression models

Predictors: All log-ratio memberships

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 4.01 | >0.05 | 53.4 |
| SaT | 6.2 | >0.05 | 4.09 |
| ClS | 7.78 | >0.05 | 18.13 |
| SaS | 4.47 | >0.05 | 10.12 |
| AWC | 19.31 | **<0.01** | 5.16 |

**Predictors: Membership in classes 2 and 4**

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 1.81 | >0.05 | |
| SaT | 6.17 | **<0.05** | |
| ClS | 6.64 | **<0.05** | |
| SaS | 1.15 | >0.05 | |
| AWC | 4.94 | **<0.05** | |

* Selected as dominant classes.

**Prediction variance (assuming 30 calibration data)**

**SaT**

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 4.45 | 2.11 | 34.6 |
| Best regression | 4.46 | 2.11 | 34.6 |

**SaS**

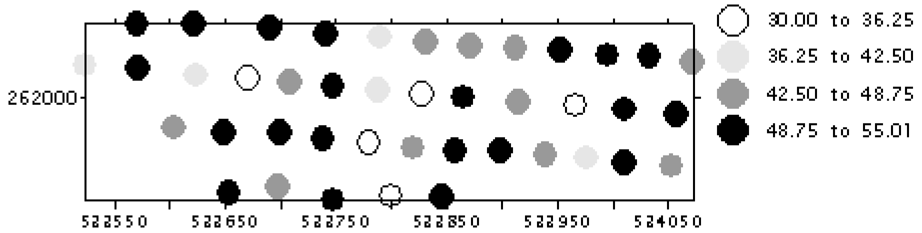
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 10.54 | 3.25 | 33.8 |
| Class mean | 9.89 | 3.15 | 32.8 |
| Best regression | 11.94 | 3.40 | 36.0 |

## AWC

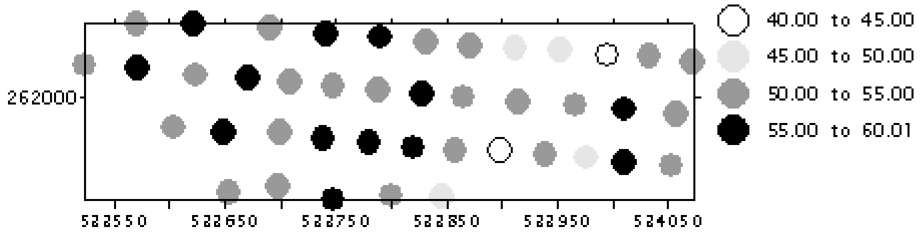
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 7.44 | 2.72 | 1.94 |
| Class mean | 6.71 | 2.59 | 1.84 |
| Best regression | 7.22 | 2.68 | 1.91 |

## Post plots of soil data

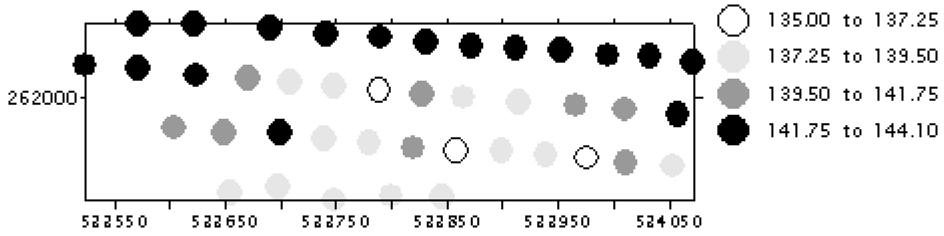
### Clay(%), Topsoil



### Clay(%), Subsoil



### Available Water (AWC) (mm).



### Soil Series

All sample sites in the field were identified as Hanslope series.

### Discussion.

The principal contrast in this field is between areas with maximum membership in class 2 (near or below average yield in all seasons) and class 4 (above average yield). There are significant differences among all the classes with respect to sand content in the topsoil and AWC (APTAB) but the differences between classes 2 and 4 are very small with much of the variation due to the smaller classes. Clay in the topsoil, sand in the subsoil and AWC are significantly related to the memberships in class 2 and 4, but point predictions of these properties have large associated errors.

#### 5.4.2.9 Knapwell Field, ADAS Boxworth

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), ClS, SaS,AWC is available water (APTAB) in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 59 | 48.2 | 40.31 | 45 | 53.75 | 30 | 60 |
| SaT | 59 | 11.4 | 14.70 | 10 | 15 | 5 | 20 |
| ClS | 56 | 52.5 | 29.34 | 50 | 55 | 30 | 60 |
| SaS | 56 | 9.4 | 35.51 | 5 | 10 | 5 | 45 |
| AWC | 59 | 141.8 | 4.1 | 141.0 | 144 | 137 | 144 |

**Class means**

| Soil Property | Class | | | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | | |
| ClT | 47.4 | 50.0 | 45.0 | 47.5 | 47.0 | 51.2 | 38.91 | 0.214 |
| SaT | 9.71 | 13.3 | 11.1 | 15.0 | 13.0 | 10.9 | 12.66 | **0.014** |
| ClS | 53.5 | 53.3 | 51.9 | 52.9 | 51.3 | 51.9 | 31.54 | 0.948 |
| SaS | 10.3 | 8.3 | 8.1 | 9.3 | 11.3 | 8.8 | 38.07 | 0.935 |
| AWC | 141.0 | 143.0 | 143.1 | 141.1 | 141.2 | 142.2 | 3.72 | 0.066 |

Regression models

Predictors: All log-ratio memberships

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 5.88 | >0.05 | 40.4 |
| SaT | 15.05 | **<0.05** | 12.71 |
| ClS | 2.56 | >0.05 | 31.3 |
| SaS | 2.14 | >0.05 | 38.2 |
| AWC | 13.01 | **<0.05** | 3.62 |

**Predictors: Membership in classes 1, 4 and 6**

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 2.34 | >0.05 | 40.8 |
| SaT | 11.94 | **<0.01** | 12.7 |
| ClS | 0.71 | >0.05 | 30.6 |
| SaS | 1.34 | >0.05 | 36.6 |
| AWC | 12.15 | **<0.01** | 3.51 |

* Selected as dominant classes.

**Prediction variance (assuming 30 calibration data)**

SaT

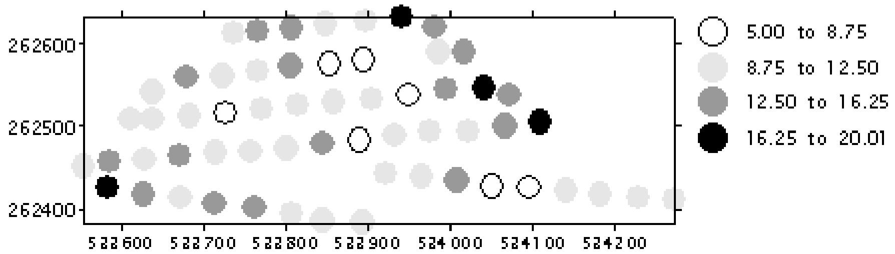
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 15.19 | 3.89 | 34.1 |
| Class mean | 15.19 | 3.89 | 34.1 |
| Best regression | 15.56 | 3.94 | 34.5 |

**AWC**

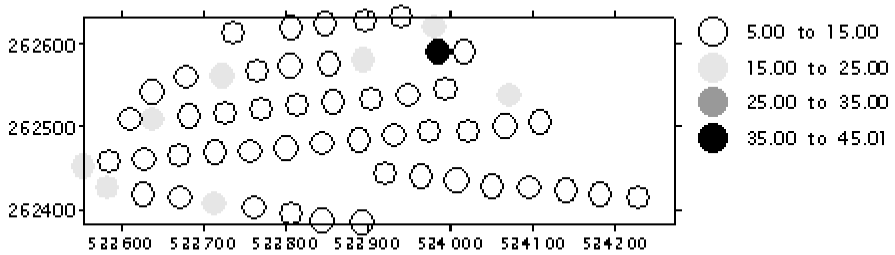
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 4.24 | 2.06 | 1.45 |
| Best regression | 4.29 | 2.07 | 1.46 |

**Post plots of soil data**

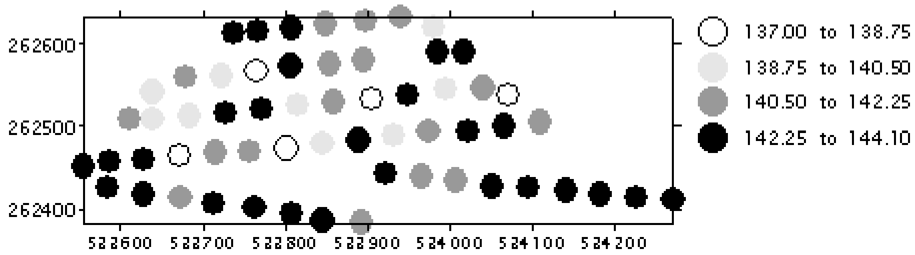
**Sand(%), Topsoil**



**Sand(%), Subsoil**



**Available Water,(AWC) (mm).**



**Soil Series.**

All sample sites in the field were identified as Hanslope series.

**Discussion**.

The dominant classes in this field are 1, 4 and 6. Classes 1 and 4 have near or above average yield in all seasons. Class 6 has low yields in the 1995 harvest, and near average yield in the others. The classes (maximum membership) differ significantly with respect to the sand content of the topsoil. Class 4 has the largest sand content in the topsoil, while that for class 6 is relatively small. The available water capacity (APTAB) is significantly related to the class memberships, although point predictions are poor, and the differences between the classes are small.

#### 5.4.2.10 Shagsby 4 Field, Chicksands.

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), OCT (Organic Carbon % in Topsoil), AWT (Available Water content in Topsoil vol/vol), ClS, SaS, OCS, AWS are same for subsoil (45–55cm), AW1ms Available water capacity in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 69 | 20.14 | 60.98 | 15 | 25 | 4 | 38 |
| SaT | 69 | 57.38 | 184.68 | 45 | 67.75 | 30 | 85 |
| OCT | 50 | 1.06 | 0.124 | 0.8 | 1.3 | 0.1 | 1.7 |
| AWT | 45 | 0.207 | 0.001 | 0.172 | 0.229 | 0.161 | 0.294 |
| ClS | 68 | 24.4 | 237.1 | 13 | 30 | 5 | 63 |
| SaS | 68 | 52.3 | 699.6 | 31.5 | 71 | 3 | 92 |
| OCS | 49 | 0.47 | 0.049 | 0.3 | 0.6 | 0.2 | 1.5 |
| AWS | 34 | 0.188 | 0.001 | 0.175 | 0.213 | 0.07 | 0.235 |
| AW1m | 34 | 196.1 | 928.7 | 186 | 221 | 110 | 232 |

**Class means**

| Soil Property | Class | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| ClT | 20.6 | 21.1 | 17.3 | 23.3 | 73.44 | >0.05 |
| SaT | 55.0 | 54.5 | 54.0 | 49.3 | 267.6 | >0.05 |
| OCT | 1.22 | 1.27 | 0.95 | 1.30 | 0.212 | >0.05 |
| AWT | 0.219 | 0.199 | 0.204 | 0.213 | 0.001 | >0.05 |
| ClS | 21.9 | 20.4 | 22.6 | 31.5 | 222.0 | >0.05 |
| SaS | 56.8 | 60.9 | 55.3 | 37.8 | 621.5 | **0.01** |
| OCS | 0.47 | 0.44 | 0.68 | 0.43 | 0.046 | >0.05 |
| AWS | 0.177 | 0.185 | 0.190 | 0.195 | 0.001 | >0.05 |
| AW1m | 193.0 | 191.4 | 195.4 | 203.7 | 988.2 | >0.05 |

**Regression models**

Predictors:  All log-ratio memberships

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.79 | >0.05 | 78.79 |
| SaT | 0.38 | >0.05 | 272.8 |
| OCT | 0.73 | >0.05 | 0.234 |
| AWT | 1.41 | >0.05 | 0.001 |
| ClS | 20.71 | **<0.001** | 187.5 |
| SaS | 25.80 | **<0.001** | 522.0 |
| OCS | 9.40 | >0.05 | 0.004 |
| AWS | 8.66 | >0.05 | 0.001 |
| AW1m | 7.99 | >0.05 | 806.6 |

Predictors: Membership in class 2 and 4

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.79 | >0.05 | 77.6 |
| SaT | 0.36 | >0.05 | 271.0 |
| OCT | 0.36 | >0.05 | 0.233 |
| AWT | 0.63 | >0.05 | 0.001 |
| ClS | 21.04 | **<0.001** | 184.6 |
| SaS | 25.66 | **<0.001** | 517.0 |
| OCS | 5.03 | >0.05 | 0.466 |
| AWS | 8.93 | **<0.05** | 0.001 |
| AW1m | 8.17 | **<0.05** | 782.6 |

* Selected as dominant classes.

Predictors: $EC_a$ (date 1)

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 25.19 | **<0.001** | 44.98 |
| SaT | 19.08 | **<0.001** | 145.9 |
| OCT | 9.93 | **<0.01** | 0.105 |
| AWT | 0.66 | >0.05 | 0.001 |
| ClS | 10.5 | **<0.01** | 207.6 |
| SaS | 8.88 | **<0.01** | 625.9 |
| OCS | | | |
| AWS | 5.32 | **<0.05** | 0.001 |
| AW1m | 4.26 | **<0.05** | 845.2 |

Predictors: Visible red reflectance

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 4.99 | **<0.05** | 72.01 |
| SaT | 3.22 | >0.05 | 254.0 |
| OCT | 2.36 | >0.05 | 0.219 |
| AWT | 1.28 | >0.05 | 0.001 |
| ClS | 0.34 | >0.05 | 239.4 |
| SaS | 0.00 | >0.05 | 710.2 |
| OCS | 6.22 | **<0.05** | 0.045 |
| AWS | 0.30 | >0.05 | 0.001 |
| AW1m | 0.54 | >0.05 | 941.9 |

**Prediction variance (assuming 30 calibration data)**

ClT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 63.01 | 7.93 | 39.4 |
| Regression on ECa | 48.08 | 6.93 | 34.4 |

SaT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 190.8 | 13.8 | 24.1 |
| Regression on ECa | 156.0 | 12.5 | 21.8 |

OCT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.128 | 0.36 | 33.8 |
| Regression on ECa | 0.112 | 0.34 | 31.6 |

ClS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 245.0 | 15.7 | 64.1 |
| Regression on Eca | 221.9 | 14.9 | 61.1 |
| Regression on 2 log-ratio memberships | 210.9 | 14.5 | 59.5 |

SaS

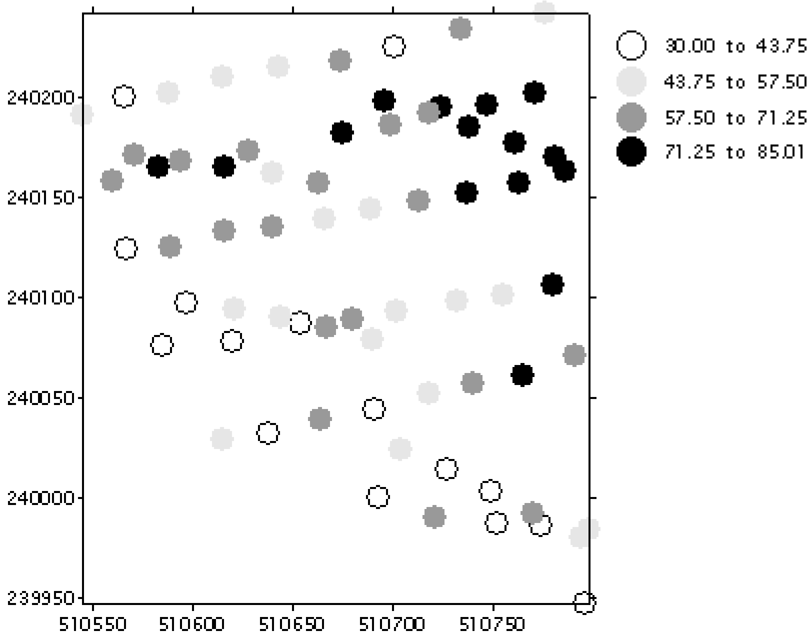
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 722.9 | 26.9 | 51.4 |
| Class means | 704.3 | 26.5 | 50.7 |
| Regression on Eca | 669.1 | 25.9 | 49.5 |
| Regression on 2 log-ratio memberships | 590.5 | 24.3 | 46.5 |

AW1m

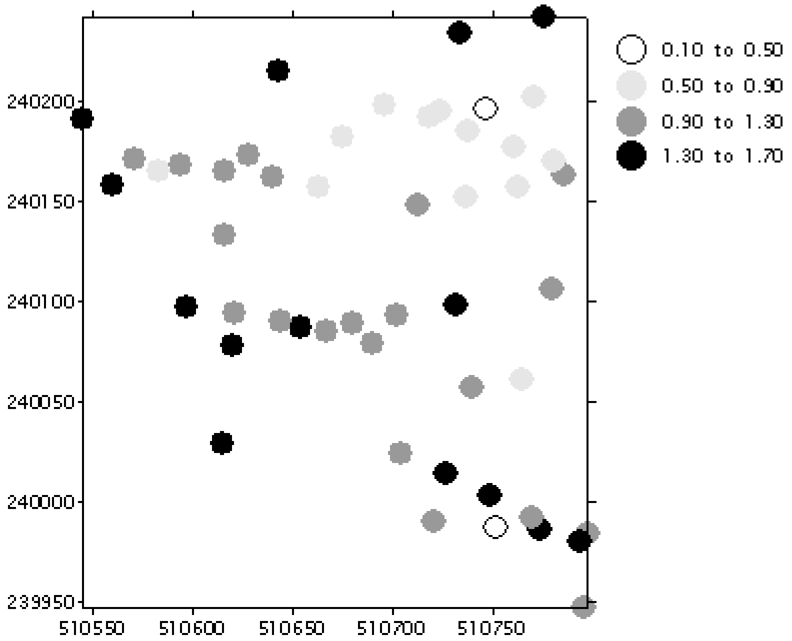
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 959.7 | 31.0 | 15.8 |
| Regression on Eca | 903.5 | 30.1 | 15.3 |
| Regression on 2 log-ratio memberships | 894.4 | 29.9 | 15.2 |

## Post plots of soil data

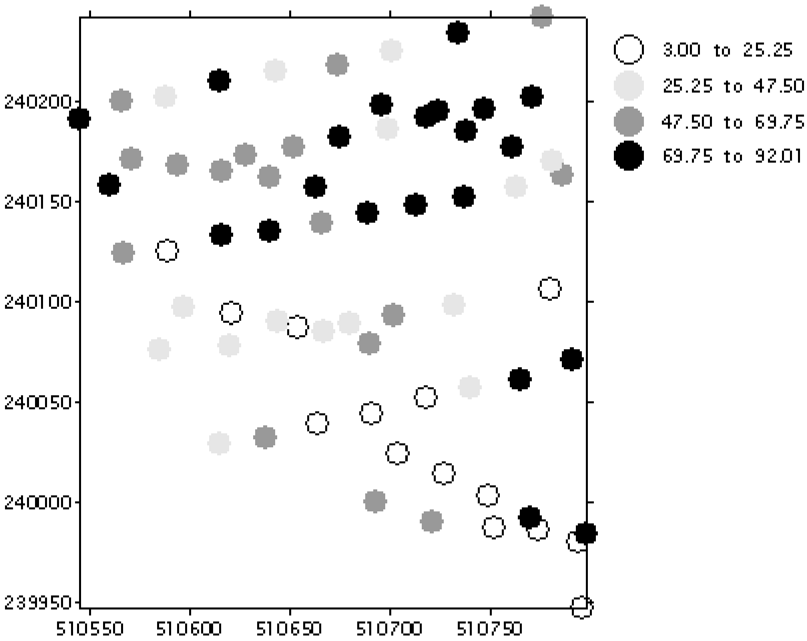
### Sand(%), Topsoil



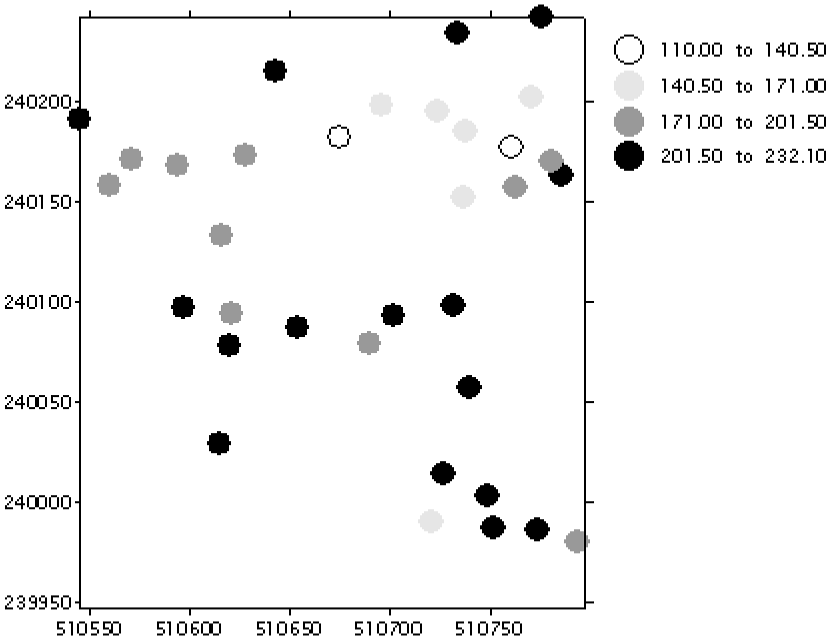
### OC(%), Topsoil



## Sand(%), Subsoil



## AW1m (mm)

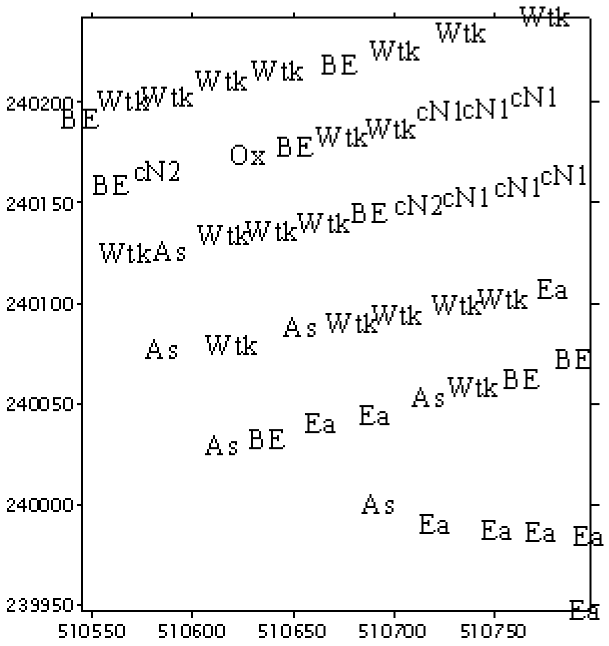


**Soil Series.**

In this case the log-ratio memberships, ECa (date1) and VR data were offered for the classification tree. Only log-ratio memberships (classes 1 and 4) and the VR data were used. The classification tree correctly partitioned 60% of the observations among the soil series (expected proportion correct under random allocation is 22%). The error matrix is shown below. The classification tree does not recognize the Bearsted, Cottenham (phase 2) or Oxpasture series at all.

| | True Series | | | | | | |
|---|---|---|---|---|---|---|---|
| **Predicted** | **As** | **BE** | **CN1** | **CN2** | **Ea** | **Ox** | **Wtk** |
| **Ashley, As** | 3 | 2 | 0 | 0 | 0 | 0 | 1 |
| **Bearsted, BE** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Cottenham 1, cN1** | 0 | 2 | 3 | 2 | 0 | 0 | 0 |
| **Cottenham 2, cN2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Evesham, Ea** | 1 | 0 | 2 | 0 | 6 | 0 | 0 |
| **Oxpasture, Ox** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Waterstock, Wtk** | 2 | 4 | 1 | 0 | 2 | 1 | 17 |

**Soil Series (true) at each sample location.**



**Discussion.**

The dominant classes here are 2 and 4, although there are substantial areas where the maximum membership is in the other classes. Class 2 corresponds to near or below average yields, while class 4 corresponds to the highest yields in both seasons. The classes (maximum membership) differ with respect to mean sand content of the subsoil, which is smallest in class 4, largest in class 2 and intermediate (and similar) in the other classes. Available water capacity in the subsoil, and available water capacity in the top metre are significantly related to the class memberships. This will reflect the generally larger available water capacity in class 4 and lower available water capacity in class 2.

Here we consider regressions of soil properties on the ECa data from the first date on which it was collected only since these data are strongly correlated (r=0.94) with the data available for the second date. The ECa data are significantly related to topsoil variables which the yield class memberships are not. However, the yield class memberships are better predictors of subsoil variables than are ECa data, and of available water capacity. This latter result is not necessarily surprising, since available water capacity is a function of soil physical properties *and* of plant physiological properties (the wilt point) so may be better reflected in the variation of yield than of a purely physical measurement. While point prediction of all these properties using ECa or yield memberships achieves significant reduction in variation compared to the overall mean, a good deal of unexplained variation remains.

Spectral reflectance in the visible red is weakly related to topsoil clay content, and to organic carbon content of the subsoil. The latter relation has no obvious explanation. The measurements were made on a seedbed, and variation in topsoil texture resulted in (visually apparent) variations in seedbed structure which will influence the overall bidirectional reflectance properties of the soil.

### 5.4.2.11 Field 107, Heydour farm, Grantham.

Note that a section of this field, at the easternmost end, was not yield mapped. The ECa data were much larger in this part of the field elsewhere. We therefore present results for predictions of soil properties by ECa for both the area which was yield mapped, and for the whole field.

**Summary Statistics**

ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), OCT (Organic Carbon % in Topsoil), AWT (Available Water content in Topsoil vol/vol), ClS, SaS, OCS, AWS are same for subsoil (45–55cm), AW1ms Available water capacity in top metre (mm).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 48 | 15.8 | 70.48 | 10 | 18 | 8 | 40 |
| SaT | 48 | 59.2 | 320.1 | 47.5 | 75 | 15 | 85 |
| OCT | 29 | 1.32 | 0.09 | 1.1 | 1.5 | 0.8 | 2.0 |
| AWT | 21 | 0.196 | 0.001 | 0.182 | 0.222 | 0.126 | 0.259 |
| ClS | 28 | 32.9 | 415.1 | 13.5 | 50 | 5 | 73 |
| SaS | 28 | 39.3 | 801.0 | 12 | 66 | 5 | 86 |
| OCS | 17 | 0.53 | 0.049 | 0.38 | 0.70 | 0.2 | 1.00 |
| AWS | 14 | 0.172 | 0.0015 | 0.158 | 0.200 | 0.089 | 0.224 |
| AW1m | 13 | 176.2 | 1250.9 | 150.5 | 200.0 | 102.0 | 230.0 |

**Class means**

| **Soil Property** | **Class** | | | | **Within-class variance** | ***p*** (H0, equal class means) |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | | |
| ClT | 15.1 | 15.0 | 167.0 | 14.0 | 74.03 | 0.862 |
| SaT | 361.4 | 40.0 | 55.9 | 69.3 | 304.6 | 0.145 |
| OCT | 1.5 | 1.5 | 1.29 | 1.26 | 0.09 | 0.325 |
| AWT | 0.227 | 0.227 | 0.200 | 0.176 | 0.0009 | 0.05 |
| ClS | 28.34 | 28.34 | 39.07 | 16.81 | 354.3 | 0.05 |
| SaS | Failed to | converge | | | | |
| OCS | 0.633 | 0.633 | 0.556 | 0.42 | 0.048 | 0.361 |
| AWS | 0.165 | 0.165 | 0.194 | 0.129 | 0.0007 | <0.001 |
| AW1m | 182.5 | 182.5 | 193.9 | 125.0 | 456.7 | <0.001 |

**Regression models**

Predictors: All log-ratio memberships

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.89 | >0.05 | 73.79 |
| SaT | 3.80 | >0.05 | 314.8 |
| OCT | 3.21 | >0.05 | 0.089 |
| AWT | 6.43 | >0.05 | 0.0009 |
| ClS | 3.07 | >0.05 | 414.3 |
| SaS | 5.79 | >0.05 | 716.1 |
| OCS | 1.32 | >0.05 | 0.054 |
| AWS | 12.25 | <0.05 | 0.00087 |
| AW1m | 14.51 | <0.001 | 638.5 |

Predictors: Membership in class 3 and 4

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.57 | >0.05 | 72.69 |
| SaT | 3.787 | >0.05 | 308.4 |
| OCT | 0.06 | >0.05 | 0.09 |
| AWT | 3.35 | >0.05 | 0.001 |
| ClS | 3.38 | >0.05 | 391.4 |
| SaS | 5.79 | >0.05 | 700.3 |
| OCS | 1.25 | >0.05 | 0.051 |
| AWS | 13.41 | <0.01 | 0.0008 |
| AW1m | 15.83 | <0.001 | 581.0 |

* Selected as dominant classes with weakly correlated memberships

Predictors: $EC_a$ (date 1)

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 0.00 | >0.05 | 72.01 |
| SaT | 0.92 | >0.05 | 320.6 |
| OCT | 0.22 | >0.05 | 0.093 |
| AWT | 1.35 | >0.05 | 0.001 |
| ClS | 0.53 | >0.05 | 836.7 |
| SaS | 0.86 | >0.05 | 794.8 |
| OCS | 0.17 | >0.05 | 0.051 |
| AWS | 0.22 | >0.05 | 0.002 |
| AW1m | 0.91 | >0.05 | 1260.0 |

Predictors: $EC_a$ (date 1) (whole field data including eastern points)

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 23.82 | <0.001 | 40.35 |
| SaT | 23.37 | <0.001 | 173.6 |
| OCT | 2.76 | >0.05 | 0.147 |
| AWT | Failed | | |
| ClS | 9.67 | 0.002 | 255.7 |
| SaS | 4.75 | 0.03 | 825.7 |
| OCS | 1.48 | >0.05 | 0.073 |
| AWS | 4.19 | 0.04 | 0.001 |
| AW1m | 2.22 | >0.05 | 1159.0 |

**Prediction variance (assuming 30 calibration data)**

AWT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.0010 | 0.032 | 16.4 |
| Class means | 0.0010 | 0.031 | 16.3 |

ClS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 428.9 | 20.7 | 63.0 |
| Class means | 401.5 | 20.0 | 60.9 |

AWS

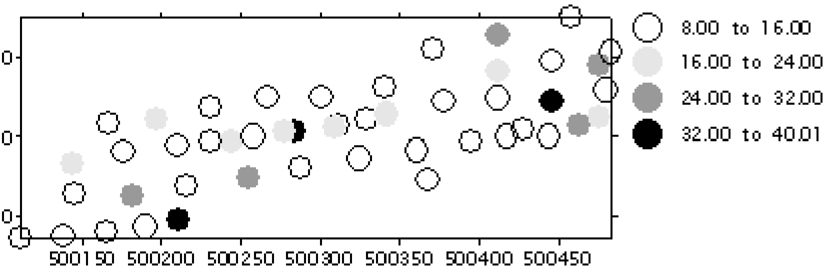
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.0016 | 0.039 | 22.9 |
| Class means | 0.0008 | 0.028 | 16.4 |
| Regression on 2 log-ratio memberships | 0.0009 | 0.030 | 17.6 |

AW1m

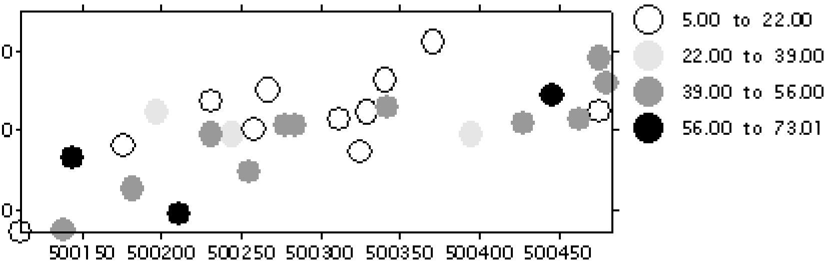
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1292.6 | 36.0 | 20.4 |
| Class means | 517.6 | 22.8 | 12.9 |
| Regression on 2 log-ratio memberships | 664.0 | 25.7 | 14.6 |

**Post plots of soil data**

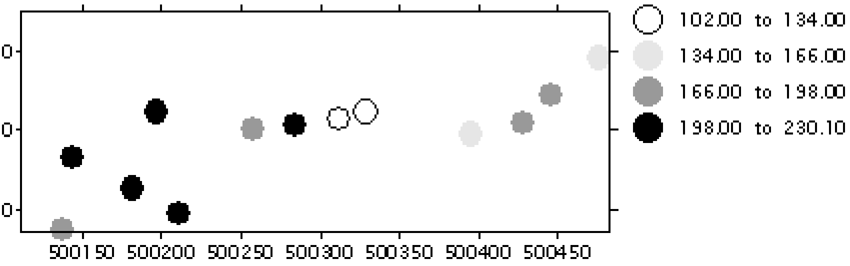
**Clay(%), Topsoil**



**Clay(%), Subsoil**



**Available water, 1m (mm)**



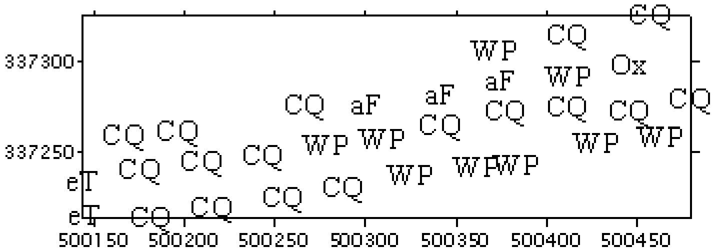
**Soil Series.**

In this case the best separation of the classes was obtained by a classification tree using log-ratio memberships only. The classification tree correctly partitioned 71% of the observations among the soil series (expected proportion correct under random allocation is 38%). The error matrix is shown below. The classification tree only recognizes the Cranwell and Wilsford series.

**True Series**

| Predicted | aF | CQ | eT | Ox | WP |
|---|---|---|---|---|---|
| ***Astrop, aF*** | 0 | 0 | 0 | 0 | 0 |
| **Cranwell, CQ** | 3 | 17 | 2 | 1 | 3 |
| **Elmton, eT** | 0 | 0 | 0 | 0 | 0 |
| **Oxpasture, Ox** | 0 | 0 | 0 | 0 | 0 |
| **Wilsford, WP** | 0 | 0 | 0 | 0 | 6 |

**Soil Series (true) at each sample location.**



**Discussion.**

Note that the yield data for this field were relatively poor. Not all of the field where ECa data were collected was yield mapped, and the yield mapped area excludes a very heavy textured section at the east of the field. The variation in ECa within the yield mapped area is much smaller than within the field as a whole. Our discussion here is limited to the mapped area. Only ECa data in vertical mode from the first date were used for prediction since these are strongly correlated (r>0.9) with all the other data.

There are three dominant classes derived from the yield data for this field. Class 1 has below average yields in both seasons. Class 3 has near or above average yields, and Class 4 changes from below to above average. The class means differ substantially with respect to both available water capacity in the subsoil and available water capacity over the top metre. Class 3 has the largest available water capacity and class 4 the smallest. This may be reflected in the larger than average yields in class 3 and the variability of yield in class 4. Prediction of available water, particularly within the top metre using the classified yield data shows quite a substantial improvement over the field mean, particularly prediction by the class mean. The ECa data show no significant relationship to soil properties within the yield mapped area.

Only two of the soil series in this field are separated, although overall over 70% of sites are correctly allocated since all instances of one dominant series (Cranwell) are correctly allocated. All instances of Astrop, Elmton and Oxpasture series and 3 (out of 9) of the Wilsford series are allocated to Cranwell series. All 6 sites allocated to Wilsford series actually belong to this series. There are three separate factors affecting crop response in this field - depth to rock (shallow to deep), texture (sand to clay) and wetness class (I to IV) (Hodgson, 1977). It is perhaps significant that the area of deep, naturally poorly drained clay soils in the east of the field is always managed separately from the sandy and loamy soils for which yield data were available.

### 5.4.2.12 The Clays field. Crowmarsh Battle farms.

**Summary Statistics**

Note. The particle size variables were all determined by hand-texturing at the sites used in this analysis. ClT (Clay content % Topsoil (5–15cm)), SaT (Sand content % Topsoil), AWT (Available Water content in Topsoil vol/vol), ClS, SaS, AWS are same for subsoil (45–55cm), AW1ms Available water capacity in top metre (mm). KT and KS are available potassium (topsoil and subsoil), PT is available phosphorous, pHT pHS are pH (topsoil and subsoil).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 90 | 38.8 | 51.47 | 30 | 45 | 25 | 50 |
| SaT | 90 | 10.61 | 62.26 | 5 | 10 | 5 | 60 |
| OCT | 120 | 4.37 | 1.81 | 4.29 | 5.21 | 1.30 | 6.29 |
| AWT | 89 | 0.206 | 0.001 | 0.191 | 0.229 | 0.044 | 0.278 |
| ClS | 74 | 37.12 | 88.88 | 30 | 45 | 25 | 55 |
| SaS | 74 | 10.41 | 79.29 | 5 | 15 | 5 | 35 |
| OCS | 113 | 2.0 | 0.69 | 1.70 | 2.55 | 0.20 | 3.71 |
| AWS | 73 | 0.193 | 0.0006 | 0.176 | 0.208 | 0.127 | 0.253 |
| AW1m | 71 | 198.4 | 471.2 | 181.2 | 214.3 | 148.5 | 249.8 |
| KT | 100 | 249.0 | 4236.4 | 194.7 | 296.7 | 137.0 | 426.7 |
| KS | 95 | 91.0 | 668.6 | 73.5 | 100.8 | 48.2 | 199.1 |
| PT | 100 | 16.4 | 53.75 | 12.1 | 18.9 | 3.2 | 58.1 |
| pHT | 100 | 7.89 | 0.032 | 7.77 | 8.03 | 7.48 | 8.22 |
| pHS | 95 | 8.29 | 0.056 | 8.14 | 8.48 | 7.68 | 8.72 |

**Class means**

| **Soil Property** | **Class** | | | | | **Within-class variance** | ***p*** (H0, equal class means) |
|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | | |
| ClT | 36.67 | 39.10 | 41.86 | 39.81 | 36.59 | 50.87 | 0.281 |
| SaT | 18.33 | 12.50 | 8.57 | 9.05 | 10.23 | 56.87 | 0.014 |
| OCT | 3.77 | 4.59 | 4.09 | 4.35 | 4.89 | 1.74 | 0.07 |
| AWT | 0.186 | 0.203 | 0.206 | 0.202 | 0.219 | 0.001 | 0.184 |
| ClS | 37.67 | 39.22 | 36.43 | 37.54 | 35.24 | 92.41 | 0.877 |
| SaS | 19.17 | 10.56 | 11.43 | 9.00 | 9.71 | 75.98 | 0.127 |
| OCS | 1.62 | 2.13 | 1.68 | 2.04 | 2.30 | 0.663 | 0.05 |
| AWS | 0.166 | 0.191 | 0.205 | 0.193 | 0.198 | 0.0005 | 0.038 |
| AW1m | 175.4 | 200.1 | 205.3 | 196.5 | 206.3 | 437.9 | 0.053 |
| KT | 210.1 | 269.2 | 219.9 | 272.2 | 232.3 | 3761.0 | 0.002 |
| KS | 80.5 | 98.0 | 77.7 | 92.1 | 97.0 | 650.4 | 0.157 |
| PT | 20.5 | 18.6 | 17.5 | 15.4 | 14.7 | 51.7 | 0.094 |
| pHT | 7.99 | 7.74 | 7.96 | 7.84 | 7.79 | 0.026 | <0.001 |
| pHS | 8.30 | 8.31 | 8.32 | 8.30 | 8.25 | 0.058 | 0.945 |

**Regression models**

Predictors:  All log-ratio memberships

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 4.03 | >0.05 | 51.46 |
| SaT | 7.76 | >0.05 | 59.73 |
| OCT | 14.31 | **<0.05** | 1.67 |
| AWT | 9.83 | >0.05 | 0.0011 |
| ClS | 0.88 | >0.05 | 92.83 |
| SaS | 4.72 | >0.05 | 78.50 |
| OCS | 12.02 | **<0.05** | 0.648 |
| AWS | 7.51 | >0.05 | 0.0006 |
| AW1m | 6.22 | >0.05 | 456.7 |
| KT | 19.61 | **<0.01** | 3659.0 |
| KS | 9.03 | >0.05 | 634.6 |
| PT | 12.97 | **<0.05** | 49.28 |
| pHT | 23.03 | **<0.001** | 0.027 |
| pHS | 0.50 | >0.05 | 0.058 |

**Predictors: Membership in class 1,4 and 5**

| **Soil Property** | **Wald statistic** | ***p*** | **Error variance** |
|---|---|---|---|
| ClT | 4.06 | >0.05 | 50.86 |
| SaT | 7.41 | >0.05 | 59.32 |
| OCT | Failed | | |
| AWT | 7.91 | **<0.05** | 0.0012 |
| ClS | 0.56 | >0.05 | 91.95 |
| SaS | 4.65 | >0.05 | 77.54 |
| OCS | 8.07 | **<0.05** | 0.665 |
| AWS | 4.08 | >0.05 | 0.0006 |
| AW1m | 5.21 | >0.05 | 456.8 |
| KT | 17.36 | **<0.001** | 3700 |
| KS | 6.67 | >0.05 | 643.5 |
| PT | 10.56 | **<0.05** | 49.93 |
| pHT | 21.95 | **<0.001** | 0.027 |
| pHS | 0.51 | >0.05 | 0.058 |

* Selected as dominant classes least correlated.

**Predictors: $EC_a$ (date 3, V and H)**

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 0.79 | >0.05 | 52.18 |
| SaT | 74.44 | **<0.001** | 34.33 |
| OCT | 13.85 | **<0.001** | 1.645 |
| AWT | 12.49 | **<0.001** | 0.0011 |
| ClS | 2.98 | >0.05 | 87.71 |
| SaS | 20.71 | **<0.001** | 63.11 |
| OCS | 7.97 | **<0.05** | 0.660 |
| AWS | 7.94 | >0.05 | 0.0006 |
| AW1m | 7.59 | >0.05 | 436.4 |
| KT | 26.25 | <0.001 | 3403 |
| KS | 6.20 | <0.05 | 640.0 |
| PT | 21.66 | **<0.001** | 44.84 |
| pHT | 1.20 | >0.05 | 0.032 |
| pHS | 13.1 | >0.05 | 0.056 |

**Prediction variance (assuming 30 calibration data)**

AWT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.0013 | 0.04 | 17.79 |
| Regression on ECa | 0.0013 | 0.04 | 17.21 |

SaT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 64.34 | 8.02 | 75.6 |
| Class means | 66.35 | 8.12 | 76.8 |
| Regression on ECa | 39.23 | 6.26 | 59.0 |

OCT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1.87 | 1.37 | 31.3 |
| Regression on ECa | 1.89 | 1.37 | 31.4 |

SaS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 81.93 | 9.05 | 87.0 |
| Regression on ECa | 72.13 | 8.49 | 81.6 |

OCS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.71 | 0.84 | 42.2 |
| Regression on ECa | 0.75 | 0.87 | 43.4 |

AWS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.0006 | 0.02 | 12.9 |
| Class means | 0.0006 | 0.02 | 12.5 |

KT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 4377 | 66.2 | 26.6 |
| Class means | 4388 | 66.2 | 26.6 |
| Regression on ECa | 3889 | 62.4 | 25.1 |

PT

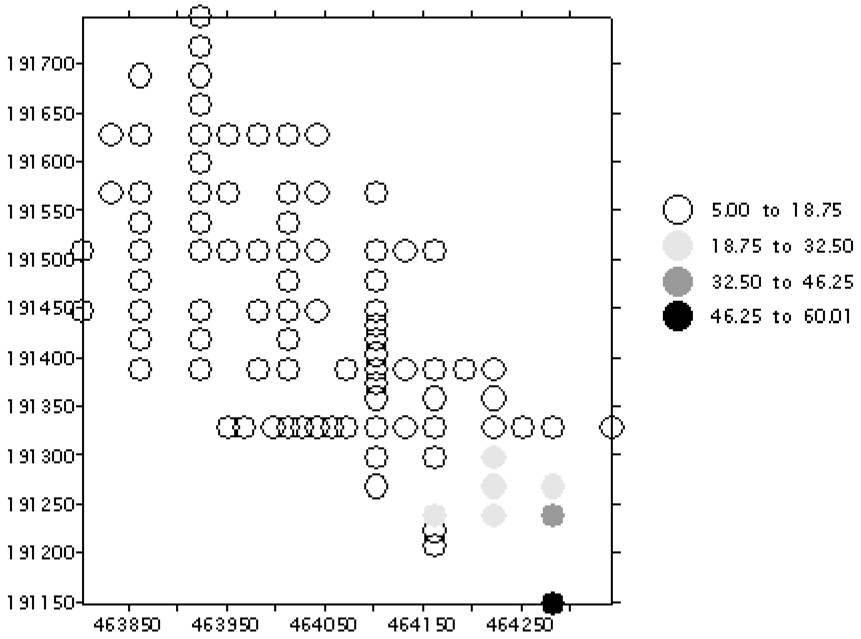
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 55.5 | 7.5 | 45.4 |
| Regression on ECa | 51.3 | 7.2 | 43.7 |

pHT

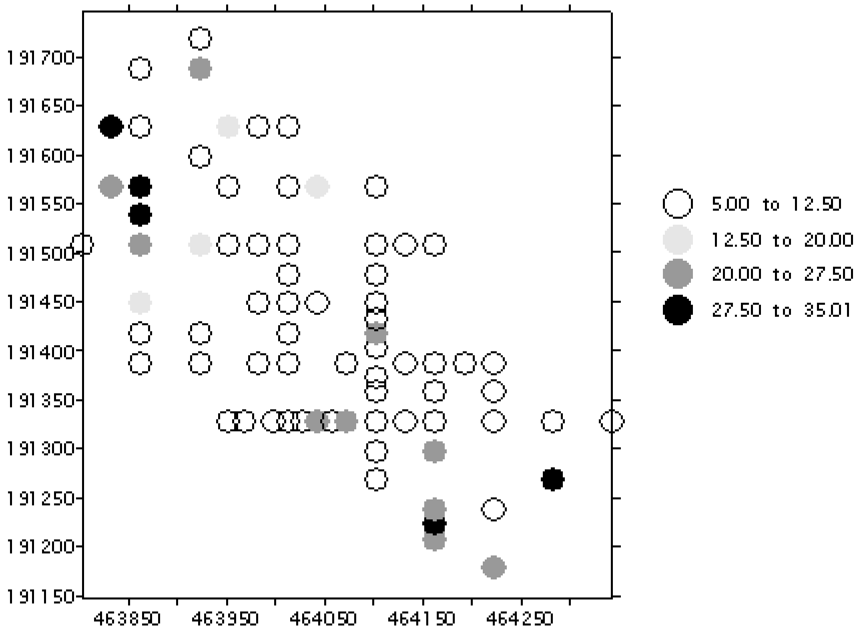
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.03 | 0.18 | 2.3 |
| Regression on 3 log-ratio memberships | 0.03 | 0.18 | 2.3 |

**Post plots of soil data**

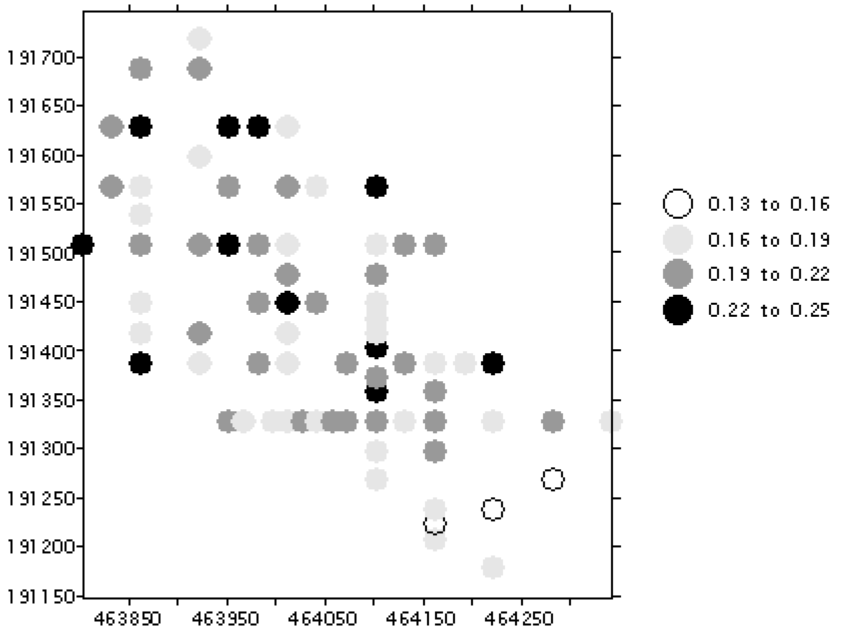
**Sand(%), Topsoil**



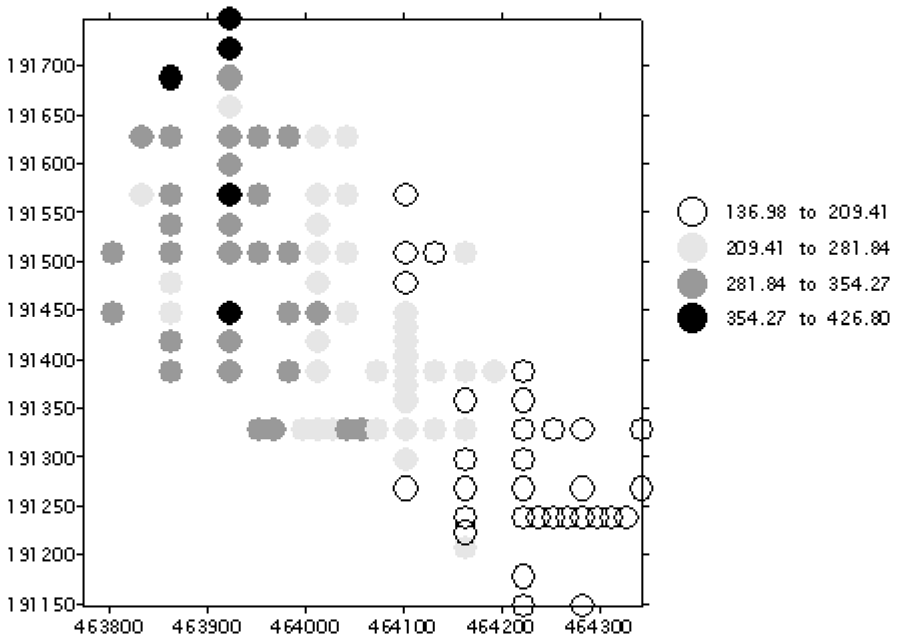
## Sand(%), Subsoil



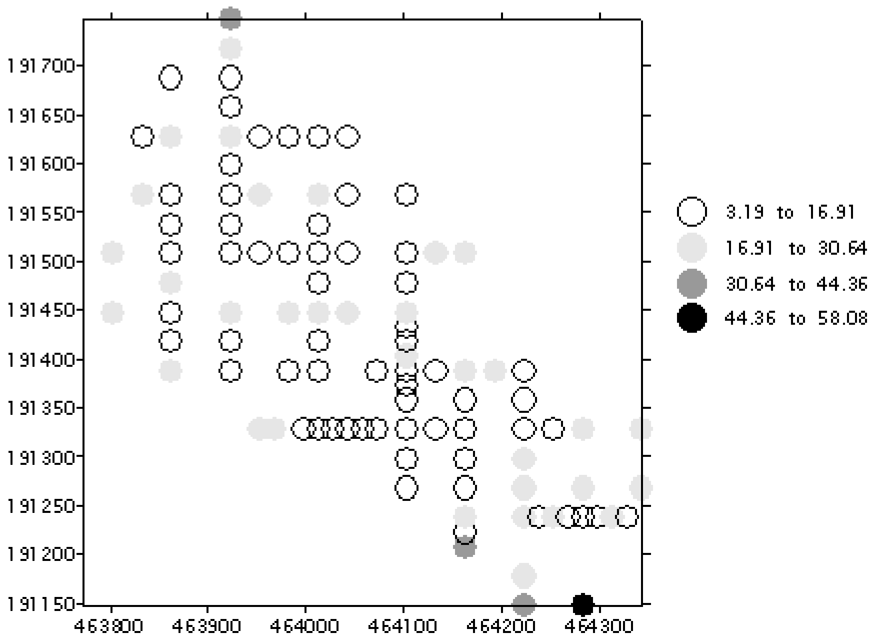
## AWS (v/v)



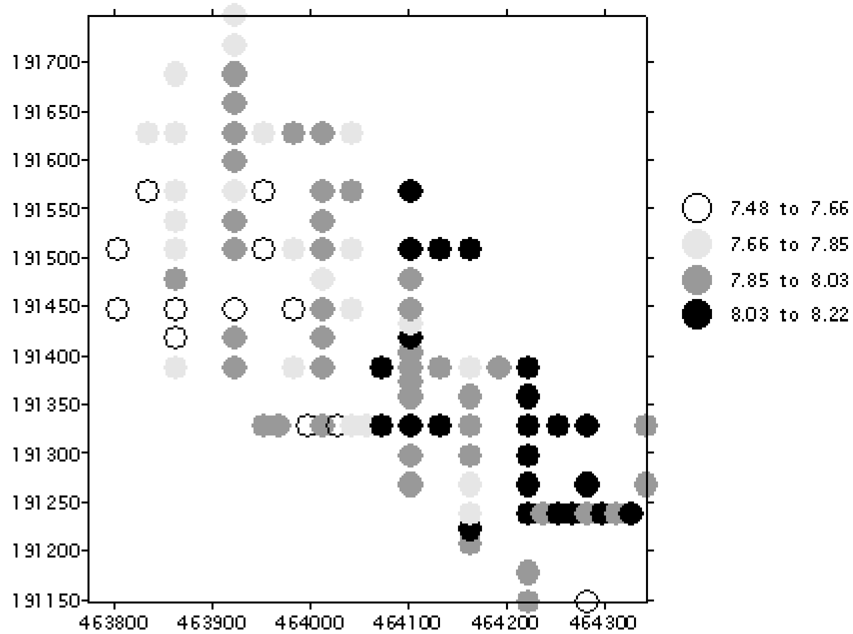
## K (mg/kg), Topsoil



## P(mg/kg), Topsoil



**pH, Topsoil**



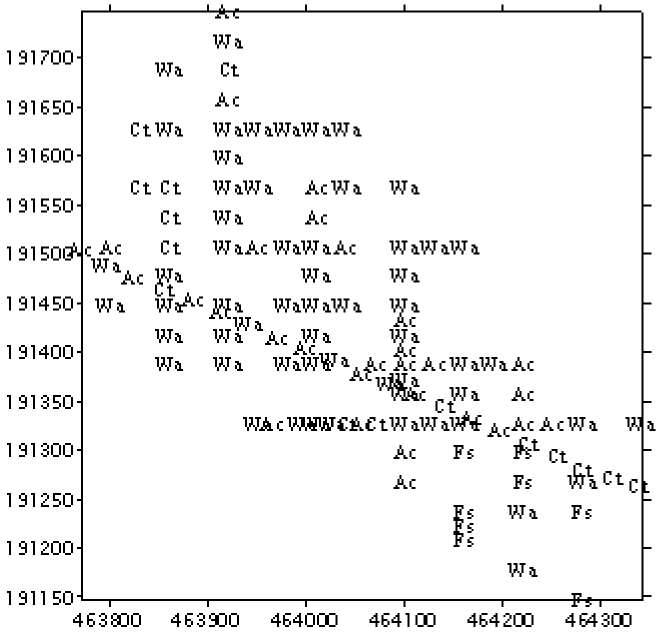
**Soil Series.**

The best separation of the soil series was achieved by a classification tree using the log-ratio memberships only. The classification tree correctly partitioned 70% of the observations among the soil series (expected proportion correct under random allocation is 37%). The error matrix is shown below. The classification tree does not recognize the Frilsham series.

**True Series**

| Predicted | Ac | Ct | Fs | Wa |
|---|---|---|---|---|
| **Andover, Ac** | 20 | 1 | 2 | 7 |
| **Coombe, Ct** | 1 | 7 | 3 | 0 |
| **Frilsham, Fs** | 0 | 0 | 0 | 0 |
| **Wallop, Wa** | 9 | 7 | 3 | 52 |

**Soil Series (true) at each sample location.**



## Discussion.

Of five classes derived from yield data for this field three have maximum membership over much of the field. These are classes 1, 4 and 5. Classes 4 and 5 have above average yield in all seasons. Class 1 has below-average yield in the 1996 harvest and close to average yield in the later two seasons.

Sand content of the topsoil, organic carbon and available water capacity in the subsoil and potassium and pH in the topsoil differ significantly between the classes (maximum membership) and phosphorous in the topsoil is also significantly related to the class memberships. Note that potassium concentrations are large in class 4 and small in class 1 (i.e. not reflecting the effect of offtake proposed by Dawson, 1997). The differences in pH between the classes, while statistically significant, are relatively small. Phosphorous concentration in the topsoil is large in class 1 and small in classes 4 and 5. This is consistent with the hypothesis that variation is driven by offtake. The sand content in the topsoil is highest, and available water capacity in the subsoil is lowest in class 1 with low yields in the dry 1996 harvest season.

ECa data collected on the third date were used since these have the most comprehensive coverage of the field. The vertical and horizontal mode data were not strongly correlated, so they were considered as joint predictors. ECa data were the best predictors of available water in the topsoil, sand content at both depths and potassium and phosphorous in the topsoil. Presumably these latter effects are because variables such as clay content of the soil, cation exchange capacity and iron oxide content which will influence the availability of these nutrients also influence the dielectric properties of the soil.

Despite these relationships the precision of point predictions using regressions on ECa measurements is poor with little improvement over the field mean, with the exception of sand content, particularly in the topsoil.

The best classification tree for separating the soil series used the membership values only and 70% of the sites were correctly separated. One series (Frilsham) was not recognised with instances partitioned more or less equally between the Andover and Coombe series. The main confusion was 9 instances (out of 30) of the Andover series assigned to the Wallop series and 7 (out of 59) of the Wallop series assigned to Andover. Wallop and Andover soils differ in their topsoil textures, being silty clay and silty clay loam respectively.

### 5.4.2.13. Football Field, Shuttleworth farm.

**Summary Statistics**

ClT (Clay content % Topsoil) SaT (Sand content % Topsoil), ClS, SaS are same for subsoil (30), AW1ms Available water capacity in top metre (mm). KT, PT ande pHT and KS are available potassium, available phosphorous and pH (topsoil).

| Soil Property | n | mean | variance | Q1 | Q3 | min | max |
|---|---|---|---|---|---|---|---|
| ClT | 244 | 10.4 | 23.0 | 7 | 11 | 4 | 26 |
| SaT | 244 | 69.1 | 133.9 | 66 | 76 | 35 | 87 |
| ClS | 112 | 12.2 | 30.8 | 9 | 14.5 | 5 | 28 |
| SaS | 112 | 70.3 | 108.3 | 64.5 | 78 | 42 | 85 |
| KT | 244 | 166.2 | 1636.1 | 139.7 | 186.8 | 49.9 | 356.7 |
| PT | 244 | 40.2 | 254.8 | 28.3 | 49.6 | 11.9 | 124.5 |
| pHT | 244 | 6.56 | 0.138 | 6.33 | 6.80 | 4.99 | 7.59 |

**Class means**

| Soil Property | Class | | | | Within-class variance | *p* (H0, equal class means) |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | |
| ClT | 14.9 | 7.9 | 11.9 | 8.8 | 16.58 | <0.001 |
| SaT | 58.4 | 73.7 | 63.5 | 74.6 | 90.14 | <0.001 |
| ClS | 15.7 | 10.7 | 15.1 | 10.7 | 26.7 | <0.001 |
| SaS | 64.2 | 71.9 | 63.1 | 74.9 | 85.65 | <0.001 |
| KT | 171.2 | 170.2 | 158.0 | 165.8 | 1634 | 0.343 |
| PT | 34.4 | 45.2 | 34.2 | 43.3 | 234.7 | <0.001 |
| pHT | 6.51 | 6.48 | 6.87 | 6.47 | 0.116 | <0.001 |

**Regression models**

**Predictors: All log-ratio memberships.** Since all the classes are relatively frequent and the memberships are weakly correlated, no subset was considered

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 98.32 | <0.001 | 16.5 |
| SaT | 127.0 | <0.001 | 88.6 |
| ClS | 20.3 | <0.001 | 26.0 |
| SaS | 36.1 | <0.001 | 83.4 |
| KT | 7.1 | >0.05 | 1609 |
| PT | 20.4 | <0.001 | 237.7 |
| pHT | 73.8 | <0.001 | 0.107 |

**Predictors: $EC_a$ (date 1, V and H)**

| Soil Property | Wald statistic | *p* | Error variance |
|---|---|---|---|
| ClT | 36.6 | <0.001 | 20.1 |
| SaT | 76.0 | <0.001 | 102.6 |
| ClS | 13.7 | <0.01 | 27.8 |
| SaS | 30.95 | <0.001 | 85.9 |
| KT | 12.46 | <0.01 | 1569 |
| PT | 13.76 | <0.01 | 243.0 |
| pHT | 43.0 | <0.001 | 0.118 |

**Prediction variance (assuming 30 calibration data)**

ClT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 23.77 | 4.88 | 46.88 |
| Class means | 18.79 | 4.33 | 41.68 |
| Regression on 4 memberships | 21.6 | 4.65 | 44.69 |

SaT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 138.36 | 11.76 | 17.02 |
| Class means | 102.16 | 10.11 | 14.63 |
| Regression on 4 memberships | 115.86 | 10.76 | 15.58 |

ClS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 31.83 | 5.64 | 46.24 |
| Class means | 30.26 | 5.5 | 45.09 |
| Regression on ECa | 31.82 | 5.64 | 46.23 |

SaS

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 111.91 | 10.58 | 15.04 |
| Class means | 97.07 | 9.85 | 14.01 |
| Regression on ECa | 98.15 | 9.91 | 14.08 |

KT

| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 1690.64 | 41.12 | 24.74 |
| Regression on ECa | 1793.14 | 42.35 | 25.48 |

PT

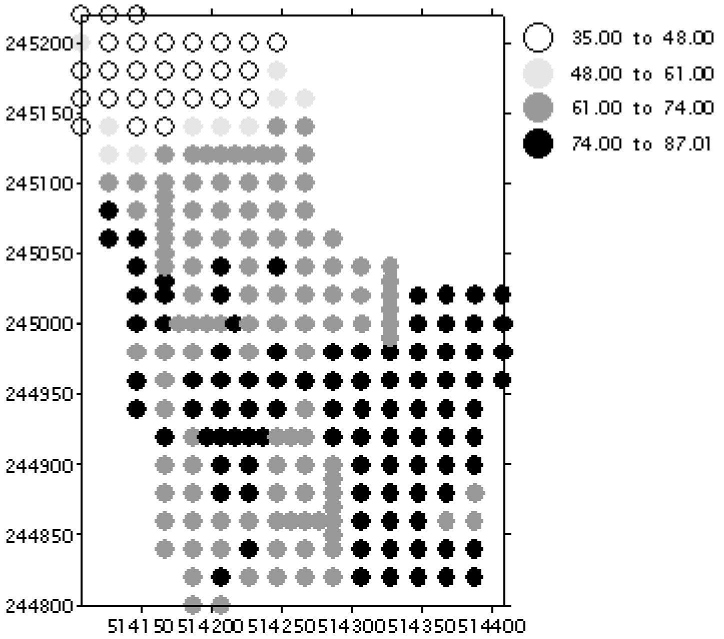
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 263.29 | 16.23 | 40.38 |
| Class means | 265.99 | 16.31 | 40.59 |
| Regression on ECa | 277.71 | 16.66 | 41.48 |

pHT

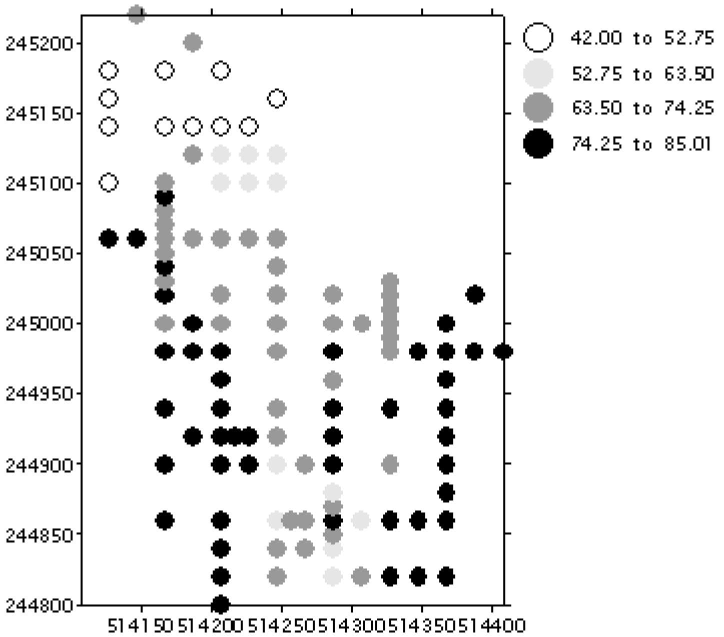
| Predictor | MSEP | RMSEP | Error % |
|---|---|---|---|
| Overall mean | 0.14 | 0.38 | 5.76 |
| Class means | 0.13 | 0.36 | 5.53 |
| Regression on ECa | 0.13 | 0.37 | 5.6 |

**Post plots of soil data**

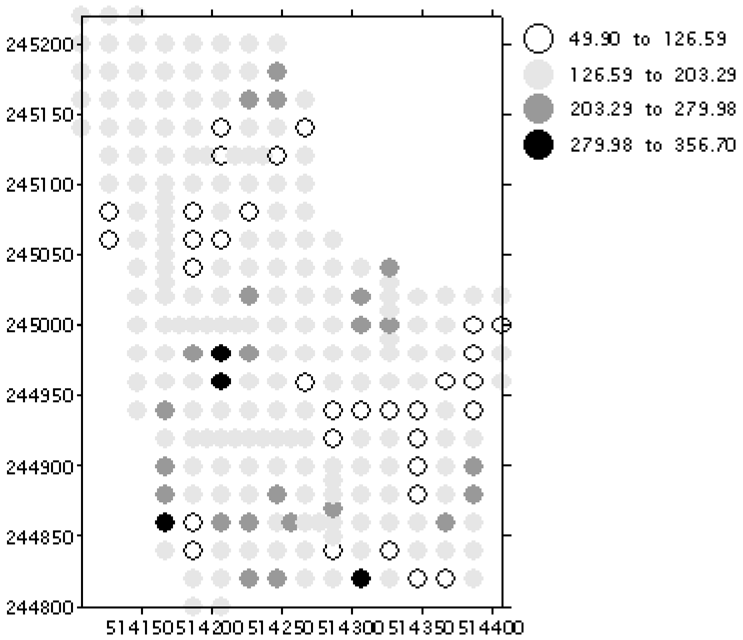
### Sand (%), Topsoil



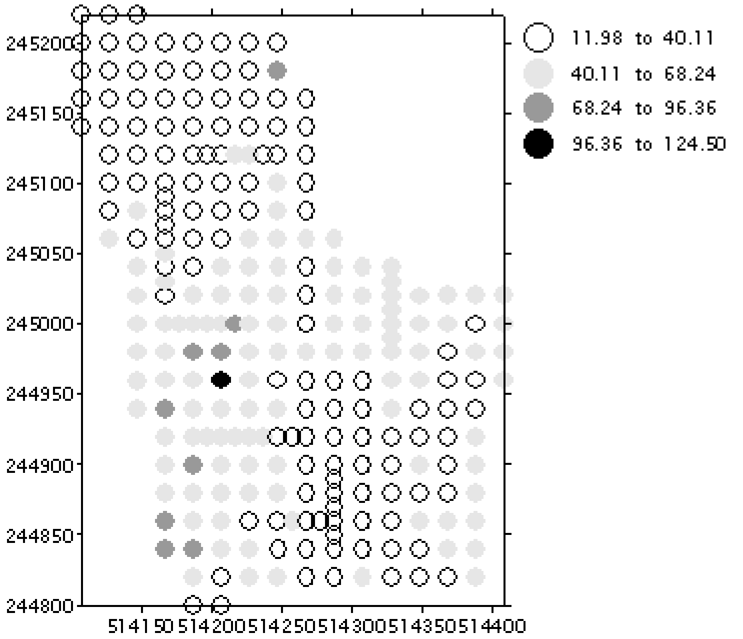
### Sand (%), Subsoil



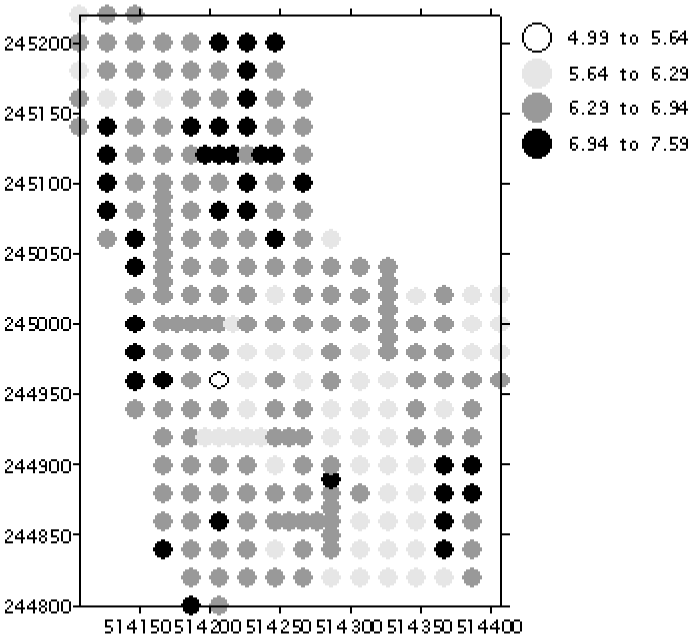
## K(mg/kg), Topsoil



## P(mg/kg), Topsoil



**pH, Topsoil**



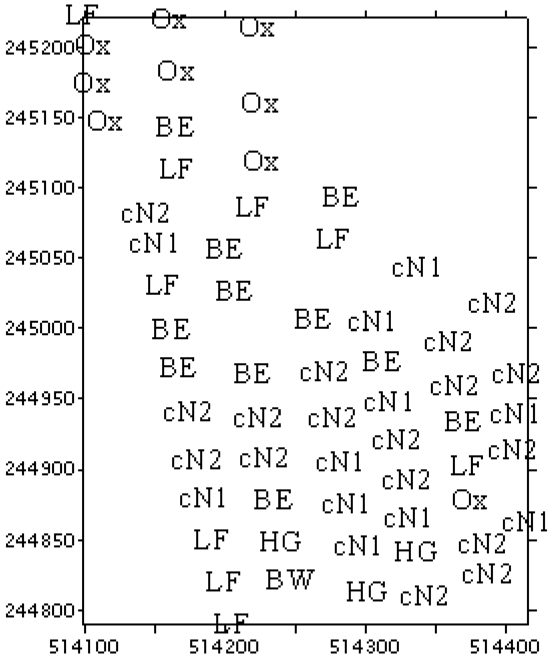
**Soil Series.**

The best separation of the soil series was achieved by a classification tree using the log-ratio memberships and the EC data. The classification tree correctly partitioned 66% of the observations among the soil series for which all predictor variables were available (expected proportion correct under random allocation is 25%). The error matrix is shown below. The classification tree does not recognize the Burlingham, Honingham and Oxpasture series.

**True Series**

| **Predicted** | **BE** | **BW** | **cN1** | **cN2** | **HG** | **LF** | **Ox** |
|---|---|---|---|---|---|---|---|
| ***Bearsted, BE*** | 3 | 0 | 2 | 1 | 0 | 0 | 0 |
| **Burlingham, BW** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Cottenham 1, cN1** | 1 | 0 | 6 | 1 | 2 | 0 | 1 |
| **Cottenham 2, cN2** | 1 | 0 | 0 | 13 | 1 | 1 | 0 |
| **Honingham, HG** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Ludford, LF** | 0 | 1 | 1 | 0 | 0 | 3 | 0 |
| **Oxpasture, Ox** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Soil Series (true) at each sample location.**



## Discussion.

The analysis of the yield data for this field reveals four basic patterns of season-to-season variation in yield. Each class is the class of maximum membership over a substantial proportion of the field. However, the yield patterns are complex, no class has consistently above or below average yields. This may reflect the differences in rainfall between the three seasons, but there is no obvious interpretation of these patterns when the soil conditions within the classes are considered.

The classes (maximum membership) differ significantly with respect to all properties apart from the potassium content of the topsoil. Similarly the ECa data (vertical and horizontal mode on the first date) are significantly related to all properties. However, the prediction error is smallest for the class mean than for a regression on ECa data in all cases (apart from potassium). However, the errors of point predictions are all not much smaller than the error for using the field mean.

The classification tree correctly separated 66% of the sites between series, although three were not distinguished. Sites in the Honingham series were allocated to phases of the Cottenham series.

### 5.5 Conclusions.

We have discussed the results for each field in turn, considering the evidence that the classification analysis of the yield data has captured information on key soil variations in each field, and assessing the scope for predicting soil properties using this information, and sensor data where available. We now summarize the key conclusions which these results support.

1) In all fields at least one soil property, physical or chemical, was significantly related to the classification on the yield data, suggesting that this analysis is capturing some important within-field variations in most cases. This supports the second principal hypothesis of this project. It also suggests that the zones derived from the classification analysis on the yield data may provide a useful basis for managing this within-field variability, perhaps by variable rate management of inputs.

2) Nutrients (P or K) were often significantly related to the classification. In some cases the nutrients were more strongly related to the zones or membership values than were physical properties of the soil, which could suggest that there is a causal link (either way) between yield patterns and spatial variation of the nutrients. In general, however, we expect that these relationships are attributable to both yield and nutrient concentrations being correlated with physical properties.

3) Interpretation of the season-to-season class patterns is difficult. For example, the hypothesis that higher yielding areas of the field will have lower concentrations of P and K has not been generally validated; in some cases this is true, in other cases the converse is true. Analysing a bulk sample from within each class for soil nutrients will be essential.

4) With a few exceptions the point predictive power of class means or regressions is not very good. That is to say our prediction of the value of a soil property at a site from the class mean or regression on class memberships or ECa has a large error, not much smaller than the error of using the field mean as the predictor. The variograms obtained for the BBRO data suggest that the unexplained variation may be spatially uncorrelated (and so un-mappable in principle)— e.g. Hall 8 field available water in the subsoil, but this is not always the case— eg. Hall 8 available P.

5) The relationships between soil properties and ECa data were not consistently weaker or stronger than the relationship between soil properties and the yield classes and memberships. This suggests that both sources of information are likely to be useful. This is not surprising. The ECa measurement is purely physical, while concepts such as available water have a physiological component which yield response may better detect.

6) From 1 and 4 above we may conclude that the analysis of yield data (and ECa data) is unlikely to substitute for more intensive soil sampling and geostatistical analysis if what we need are point predictions

of soil properties.  However, the division of a field into management zones by analysis of these variables is likely to capture important within-field variations in most instances.

## Chapter 6. The usefulness of classified yield data for planning sampling

In the previous chapter we showed that the variation of soil properties within fields is reflected in the variation of yield data, captured by the classification analysis to define management zones. One way in which this might be exploited is to target sampling in the field, using the information extracted from yield data to characterize the key variation in the field more efficiently both by directing the sampling to informative sites and by interpreting the information from sampling and extending it to the whole field. In this chapter we report an exercise to test the scope for this approach using results from intensively sampled fields.

### 6.1 Test fields

Four fields were chosen from the intensive sampling exercise: Knapwell and Top Pavements fields at Boxworth and Brome Pin and Little Lane at Broom's Barn. Each field had been mapped in the past by soil surveyors at 1:10,000 scale (Figures 6.1, 6.2 and 6.3) and those maps have been compared with the intensive data collected in this project.

**Brome Pin field**

The published soil map for Broom's Barn (Hodge 1991) suggests that the north east corner of the field is a complex of Swaffham Prior and Moulton soils with the remainder mapped as Barrow. However, the field notes for the farm (unpublished, dating from 1969) (Figure 5.1) suggest that the most common soil in the north east corner is Maplestead. There were 10 observations in the field. Barrow and Maplestead are related in that any clay layer occurs further down the profile (below 80 cm depth) in the latter.

**Figure 6.1 Soil map of Brome Pin (from unpublished farm map dated 1969).** Bar: Barrow series, MM: Maplestead series.

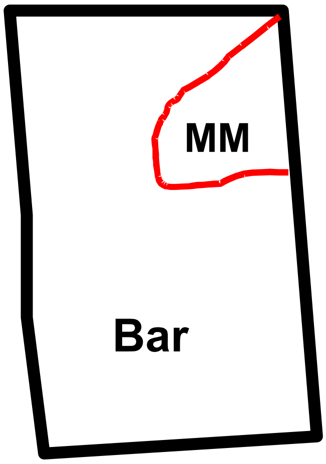
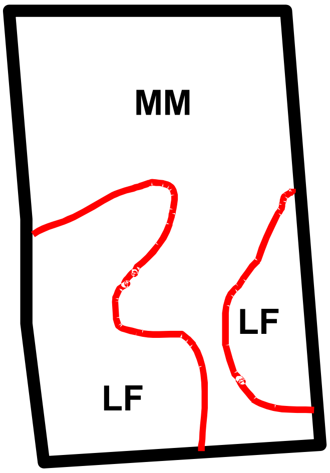


**Figure 6.2 Soil map based on intensive sampling of Brome Pin field.** MM: Maplestead series, LF: Ludford series.



**Little Lane field**

The published soil map for Broom's Barn (Hodge 1991) suggests that there are five soil map units in the field, Melford, Eyeworth, Block, Rudham and Dullingham (from west to east). However the field notes for the farm (unpublished, dating from 1969) (Figure 5.3) suggest a more simple pattern with Ashley as the most common soil in the north west corner and the remainder of the field Stretham. There were 15 observations in the field. Eyeworth, Stretham and Hanslope soils are closely related in texture profile and drainage and differ mainly in the lower clay in the topsoil of the Eyeworth and the marginally freer drainage of the Stretham soils (wetness class I). Ashley soils are non-calcareous above chalky boulder clay with wetness class II.

**Figure 6.3 Soil map of Little Lane (from unpublished farm map dated 1969)** As: Ashley series. St: Stretham series

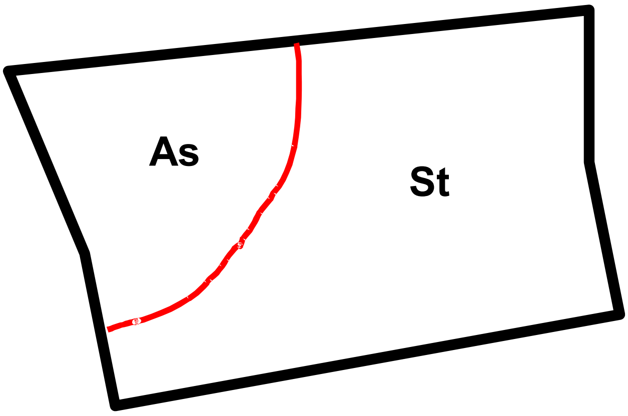
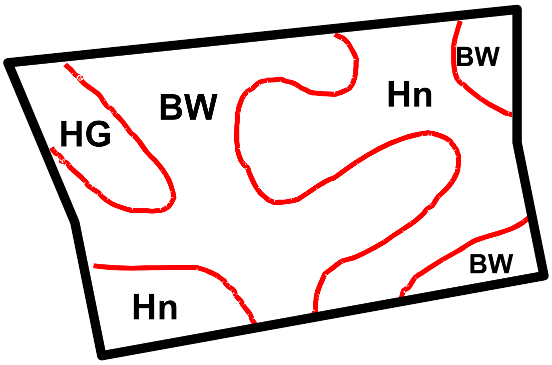


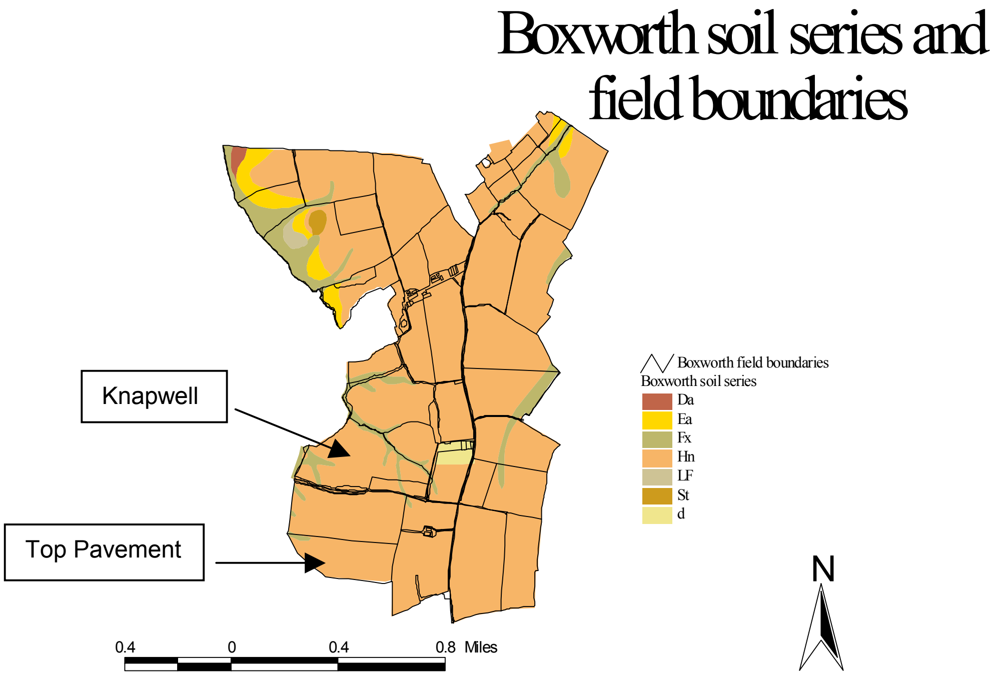
**Figure 6.4 Soil map based on intensive sampling of Little Lane field**. BW: Burlingham series, Hn: Hanslope series, HG: Honingham series.



**Top Pavements and Knapwell Fields at Boxworth**

The farm map for Boxworth (Figure 6.5) suggests Knapwell field is all Hanslope series, except for "tongues" of Folksworth series in the west and east of the field, where the dominant stones are flints rather than chalk. In Top Pavements, again most of the field is Hanslope with a small area in the north west corner of Folksworth soils. The surveyor used slight changes in the landscape shape to delineate these areas.

**Figure 6.5 Soil map of Boxworth farm** Hn: Hanslope, Fx: Folksworth.



### 6.2 Analyses.

The soil sample sites were plotted over the map of zones derived from the yield data and the zone corresponding to each sample site was identified. We then considered how the sampling intensity could be reduced to 50% or 25% of the original intensity or to a typical "extensive" sampling intensity of one sample per ha. This "thinning" of the original sampling grid was done without reference to the soil series at each point (since these are unknown when sampling the soil) but did ensure reasonable representation of all the zones (since these can be obtained before field-work).

If the management zones reflect the underlying pattern of soil variation as expressed by the soil series then we should be able to identify the soil series (one or more) associated with each zone by limited sampling, and then assume that this/these series are associated with other locations within the same zone. In the results below we show the soil series associated with each zone at different intensities of sampling.

The approach above assumes that there will be a relatively simple relationship between the zones derived from yield data and the soil series. It may be that the yield data contain information about the soil variation which is related to the pattern of soil series in more complex ways. For this reason we considered the possibility of using observations from a reduced sampling intensity to derive a classification tree (as used in Chapter 4 section 6) to predict the soil series from the (log-ratio transformed) memberships in the classes derived from yield data which give rise to the zones.

To test this hypothesis we used the subset of data obtained by thinning the original intensive sample to 50%. This was done for the Brome Pin and Little Lane data sets only. The soil series at each sample site was noted and the membership values in all classes recognized in the yield data at the nearest yield data point were extracted. A classification tree was then derived to predict the soil series. In fact the terminal nodes of the classification tree did not necessarily correspond to a single soil series so the classification tree sometimes predicted a complex soil map unit. We used the tree to predict the soil map unit (simple series or complex) at each yield data point in the field (thus forming a soil map of the field), and at each of the remaining intensive sample points not used to form the tree (allowing us to test the predictions directly against the observed soil series).

## 6.3 Results

### 6.3.1 Brome Pin, Broom's Barn

Distinct patterns of closely related soils. There are four zones defined from the yield maps with distinct spatial structure. Sampling intensity: 7 cores per ha

**Soil series corresponding to each zone at each sampling intensity.**

**All observations**
Zone 1: MM 4 (40%) LF 6 (60%)
Zone 2: MM 5 (46%) LF 2 (22%) Na 2 (22%)
Zone 3: MM 7 (50%) LF 3 (22%) Na 4 (28%)
Zone 4: MM 8 (36%) LF 10 (45%) Na 4 (19%)

MM = Maplestead, LF = Ludford; Na = Newport

**50 % of observations**
Zone 1: MM 2 LF 2
Zone 2: MM 2 Na 2
Zone 3: MM 3 LF 1 Na 3
Zone 4: MM 5 LF 6

**25 % of observations**
Zone 1: LF 2
Zone 2: MM 1 LF 1
Zone 3: MM 2 LF 2
Zone 4: Na 2 LF 3

**1 per ha**
Zone 1: LF 2
Zone 2: MM 1
Zone 3: MM 2 Na 1
Zone 4: MM 1 Na 1 LF 1

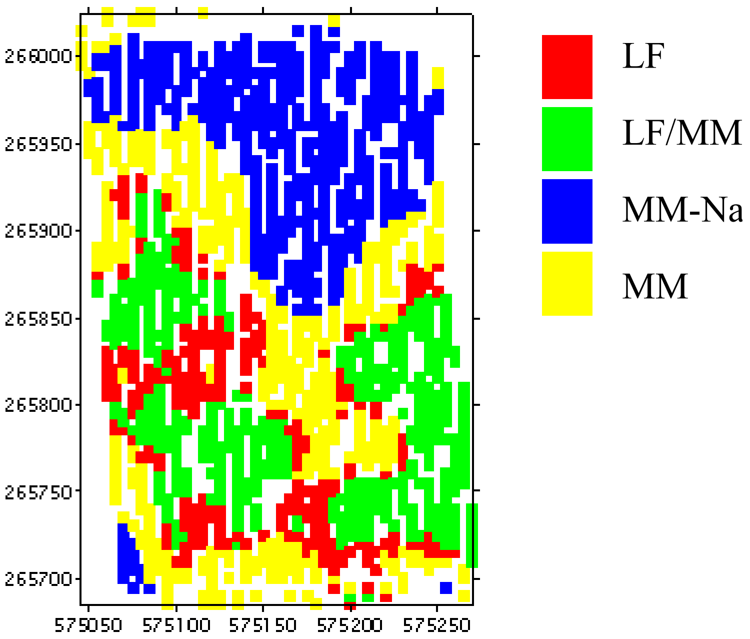
The classification tree, formed from the 50% reduced data set, defined 4 soil map units:
LF (Ludford series)
LF/MM (Predominantly, LF series with 40% inclusions of Maplestead (MM))
MM-Na (MM and Newport (Na) in equal proportions)
MM

When the classification tree was applied to the remaining sample sites (i.e. those not used to define it) these were allocated among the map units as shown in Table 6.1 below. The series identified at nine of the test sites were not in the definition of the corresponding map unit. That is to say, at 75% of the test sites the identified series WAS one of those included in the definition of the map unit. The predicted map unit at all yield data points in the field is shown in Figure 6.6

**Table 6.1 Observed series and predicted map unit at test sites on Brome Pin**

| | *Observed Series* | | |
|---|---|---|---|
| ***Map Unit*** | **LF** | **MM** | **Na** |
| **LF** | 2 | 0 | 1 |
| **LF/MM** | 7 | 2 | 2 |
| **MM-Na** | 3 | 4 | 2 |
| **MM** | 2 | 9 | 1 |

**Figure 6.6 Predicted soil map unit at all yield data points in the field**



**Discussion.**

Using all the cores characterises the four zones as predominantly LF, MM, MM and LF respectively. Reducing the number of observations by targeting gives similar results for zones 1, 2, and 3 but the interpretation of zone 4 is more uncertain. This suggests that the zones do reflect some differences in soil type but that other agronomic factors, including perhaps soil factors which the series do not represent, also determine the pattern of zones. The soil series and zones are clearly related but the boundaries drawn between the soil types is different on the two soil maps and the interpretation of the zones from yield data. This is not an unusual situation in a relatively flat field with few clues from the landscape to indicate the position of soil boundaries.

The classification tree generates map units which give a reasonable prediction of the soil type at the test sites and the soil map in Figure 6.6 clearly describes a similar pattern to the detailed map of the field in Figure 6.2.

### 6.3.2. Little Lane, Broom's Barn

Distinct pattern of related soils; there are three zones (2,3 and 5) which form distinct spatial units, the rest form a fine-scale pattern or 'speckle'. Sampling intensity: 6.4 cores per ha

**Soil series corresponding to each zone at each sampling intensity.**

**All observations**

Zone 2: BW 2 Hn 2

Zone 3: BW 7 HG 1

Zone 5: BW 2 Hn 8 LF 1

BW = Burlingham; Hn = Hanslope; HG = Honingham; LF = Ludford

**50 % of observations**

Zone 2: BW 1

Zone 3: BW 4 HG 1

Zone 5: BW3 Hn 4

**25 % of observations**

Zone 2: Hn 1

Zone 3: BW 2

Zone 5: BW 1 Hn 1

**1 per ha:**

Zone 1: BW 1 Hn 1

Zone 2: BW 1 Hn 1

Zone 3: BW 2

Zone 5: BW 2 Hn 1

The classification tree, formed from the 50% reduced data set, defined 2 map units:

BW (Burlingham series)

Hn (Hanslope series)

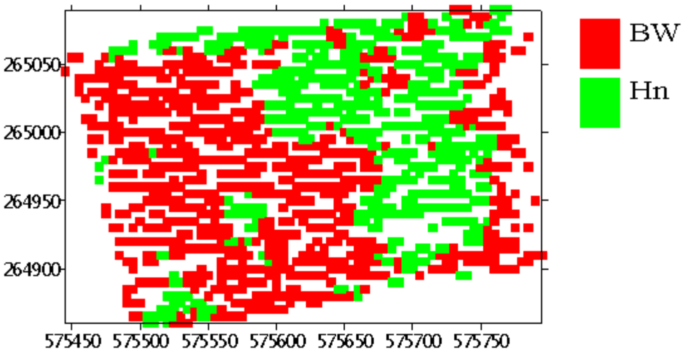
When the classification tree was applied to the remaining sample sites (i.e. those not used to define it) they were allocated among the map units as shown below. The series identified at 27 of the bores was not in the definition of the corresponding map unit. That is to say, at only 36% of the test sites was the identified series was one of those included in the definition of the map unit. This is poor separation, probably in part because the training data set was too small at this intensity of sampling.

**Table 6.2 Observed series and predicted map unit at test sites on Little Lane**

| | *Observed Series* | | | | |
|---|---|---|---|---|---|
| ***Map Unit*** | **BW** | **HG** | **Hn** | **LF** | **MM** |
| **Bw** | 8 | 3 | 11 | 1 | 2 |
| **Hn** | 7 | 1 | 7 | 1 | 0 |

Figure 6.7 shows the predicted map units at all yield data points across the field.

**Figure 6.7 Predicted soil map unit at all yield data points in the field**



**Discussion**

Using all the cores characterizes zones 2, 3, and 5 as complex, Burlingham (BW) and Hanslope (Hn) respectively. There are not enough cores completely within the remaining units to characterise them. Reducing the number of observations by targeting leads to the conclusion that all the units are complexes of soil series. There is a better match between the farm map and the core data than with the published ("district") map when considering the broad soil types. Although the point assessment of the mapunits predicted by the classification tree suggested that the predictions are poor it is clear that the map in Figure 6.7 reflects a pattern of soil variation which is related to the detailed soil map in Figure 6.4

### 6.3.3. Knapwell and Top Pavements, Boxworth

Sampling intensity was: Knapwell 4.3 cores per ha; Top Pavements 4.4 cores/ha.

While the analysis of yield data for both fields generated two or more distinct zones on this field all sample sites were identified as the Hanslope series.

**Discussion**

The results indicate that the yield variations identified by the zones in both fields are not associated with differences between soil series. The limited extent of the Folksworth series in the field is not large enough to be detected by the intensive sampling.

The results in section 5.4.2.8/9 showed that there are statistically significant differences between the zones with respect to sand content and available water capacity. In some parts of East Anglia topsoil and subsoil phases of the Hanslope series have been mapped, for example, heavy clay loam as opposed to clay topsoils. AWC reduces down the profile as compaction increases and such compaction is likely to impinge on root development to exploit the water that is available.

Having said that, it is not expected *a priori* that these Hanslope soils are prime targets for variable rate management, and it should be noted that in the jacknifing assessment of the classification tree for predicting the simplified PVRM rating (Chapter 4 section 6) both these fields were allocated to rate 1, the PVRM rating they had been assigned originally.

## 6.5 Conclusions

While the management zones account for significant variation in many soil properties the membership values in the classes derived from the yield data contain considerable additional information which is lost when we simply identify the class of maximum membership at a site. Hence the classification trees using these membership values as predictors allow us to map soil variation in terms of soil series rather better than assuming that the zone (class of maximum membership) corresponds directly to a soil map unit. This may be the best way of using the classification tree to map within-field variation from a limited set of soil samples.

It is clear that soil series are useful ways of summarizing soil variability but do not necessarily capture all the features of soil variation which may drive yield variation and so the delineation of management zones from yield data. The simple pattern of zones may therefore not of itself reveal all the variability that a pedologist would recognize in a field, but it may well be that it ‘filters out’ a good deal of the variability which is not pertinent to crop performance. Having said that, the soil maps produced by classification trees from the soil sampling at reduced intensity captured many of the key features of the soil variation identified in Brome Pin and Little Lane field by soil surveyors, particularly in the former case.

Although distinct zones were identified in the two fields from Boxworth the variability of these fields is unlikely to warrant variable rate management of inputs. This would have been correctly determined using one of the classification trees developed in Chapter 4. This emphasizes that the appearance of spatially distinct zones in the analysis of the yield data is not of itself evidence that the soil variations within a field are worth managing.

In summary, the results presented in this chapter indicate that the memberships derived in the classification of yield data may be useful for extending a limited number of observations of the soil series to a soil map of the field — in those fields where we have reason to believe that there is substantial soil variation which may warrant variable rate management of inputs.

## 7. Acknowledgements.

This research was funded by the Home-Grown Cereals Authority, and we are grateful to the Cereals Committee of HGCA for their support. We are also grateful to the R&D department staff at HGCA for their assistance.

We are particularly grateful to those farmers who made their yield data available to this project, and in particular to the following who allowed field-work on their land. A Cragg, C Allhusen, Christopher Wilson, D N Parkes, D Niven, Dan Squier, Graham Allen, J Dingemans, J Hawthorne, James Moldon, John Ward , M Harrold, Mr Robson, Mr Van de Peer, N Walker, P B Thomas, P King, Paul Cripsey, Paul Truepenny, Philip Chamberlain, Robert Pask, Steven Goward (Broom's Barn), Stuart Galloway and Stuart Scarff.

As noted in the report we have made use of soil data from previous work funded by the British Beet Research Organization (formerly SBREF) in which SRI was a participant. We are grateful to Dr Keith Jaggard of IACR-Brooms Barn for making these data available.

Silsoe Research Institute is sponsored by the Biotechnology and Biological Sciences Research Council (BBSRC) and so by the British taxpayer. The BBSRC's core grant to the Institute has allowed the development of many of the methods used in this research.

## 8. References.

Aitchison, J., (1986) *The statistical analysis of compositional data.* Chapman and Hall, London.

Akaike, H., (1973). Information theory and an extension of the maximum likelihood principle. In Petov, B.N. and Csaki, F., (Eds). *2nd International Symposium on Information Theory*. Akademia Kiado, Budapest. pp. 267–281.

Avery, B.W. and Bascomb, C.L. (1982). *Soil Survey Laboratory Methods.* Soil Survey Technical Monographs No. 6.

Bourgault, G. and Marcotte, D., (1991) Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology* **23**, 899–928.

Dawson, C.J. (1997). Management for spatial variability. *Proceedings of the 1st European Conference on Precision Agriculture.* pp. 45–58.

Godwin, R.J., Earl, R., Taylor, J.C., Wood, G.A., Bradley, R.I., Welsh, J.P., Richards, T., Blackmore, B.S., Carver, M. and Knight, S. (2002). *'Precision farming' of cereal crops: a five year experiment to develop management guidelines.* Project Report No. 264e. HGCA, London.

Hall, D.G.M., Reeve, M., Thomasson, A.J. & Wright, V.F. (1977*). Water retention, porosity and density of field soils*. Soil Survey Technical Monograph No. 9.

Harding, S.A. and Webster, R. (1995) Procedure MVARIOGRAM. In: *Genstat® 5 Procedure Library Manual Release 3[3]* (eds R.W. Payne and G.M. Arnold), pp. 267–269. Lawes Agricultural Trust (Rothamsted Experimental Station), Harpenden.

HGCA (2002). *'Precision farming' of cereals. Practical guidelines and crop nutrition.* HGCA, London.

Hodge, C.A.H. (1991) Soils in Suffolk I: sheet TL76E/86W (Risby). Soil Survey Record No. 107.

Hodge, C.A.H., Burton, R.G.O., Corbett, W.M., Evans, R. and Seale, R.S. (1983). Soils of England and Wales: sheet 4: Eastern England. 1:250,000 scale map. Soil Survey of England and Wales, Harpenden.

Hodgson, J.M.(ed.) (1997). Soil survey field handbook. Soil Survey Technical Monograph No.5.

Journel, A.G. and Huijbregts, Ch.H. (1979). *Mining Geostatistics.* Academic Press.

Lark, R.M. (1995). Components of accuracy of maps with special reference to discriminant analysis on remote sensor data. *International Journal of Remote Sensing,* **16,** 1461-1480.

Lark, R.M. (1998) Forming spatially coherent regions by classification of multivariate data. *International Journal of Geographical Information Science.* **12,** 83–98.

Lark, R.M. (2000a). The use of ancillary data in field investigations for site-specific agriculture. *Proceedings of the 4th International Symposium on Spatial Accuracy Assessment*, pp 397–404.

Lark, R.M. (2000b) Regression analysis with spatially autocorrelated error: examples with simulated data and from mapping of soil organic matter content. *International Journal of Geographical Information Science.* **14**, 247–264.

Lark, R.M. (2001) Some tools for parsimonious analysis and interpretation of within-field variation. *Soil and Tillage Research*. **58**, 99–111.

Lark, R.M. (2002). Estimating within-field variation of crop responses to inputs. *Proceedings of the XXI International Biometrics Conference* Freiburg. pp 265–278.

Lark, R.M. and Stafford, J.V. (1997) Classification as a first step in the interpretation of temporal and spatial variability of crop yield. *Annals of Applied Biology.* **130,** 111–121.

Lark, R.M., Bolam, H.C., Mayr, T., Bradley, I., Burton, R.G.O. and Dampney, P.M.R. (1999). Analysis of yield maps in support of field investigations of soil variation. *Proceedings of the 2nd European Conference on Precision Agriculture*. pp 151–161.

Lark, R.M., Bolam, H.C., Mayr, T., Bradley, R.I., Burton, R.G.O. and Dampney, P.M.R. (1998). *The development of cost-effective methods for analysing soil information to define crop management zones*. Project Report No. 171. HGCA, London.

MathSoft (1999). *S-PLUS 2000 Guide to Statistics, Volume 1.* Data Analysis Products Division, MathSoft, Seattle WA.

Mayr, T., Jarvis, N. and Simota, C. (1999). Pedotransfer Functions for Soil Water Retention Characteristics. *Proceeding of the International Workshop Characterisation and Measurement of the Hydraulic Properties of Unsaturated Porous Media*, Riverside, CA. 22-24 October 1998.

McBratney, A.B. and Moore, A.W. (1985) Application of fuzzy sets to climate classification. *Agricultural and Forest Meteorology*. **35,** 165–185.

McBratney, A.B. and Pringle, M.J. (1997). Spatial variability in soil - implications for precision agriculture. *Proceedings of the 1st European Conference on Precision Agriculture. Volume I: Spatial Variability in Soil and Crop* pp. 3–31. Oxford, BIOS Scientific Publishers Ltd.

McBratney, A.B., Whelan, B.M., Taylor, J.A. and Pringle, M.J. (2000). A management opportunity index for precision agriculture. In: (P.C. Robert, R.H. Rust and W.E. Larson, eds.) *Proceedings of the 5th International Conference on Precision Agriculture and Other Resource Management.* July 16th -19th 2000. Radisson Hotel South, Bloomington, Minnesota, USA.

Pardo-Iguzquiza, E. (1997). MLREML: A computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. *Computers and Geosciences.* **23,** 153–162.

Payne, R.W. *et al.* (1988). *Genstat 5 reference manual.* Clarendon Press, Oxford.

Pedersen, S.M., Ferguson, R.B. and Lark, R.M. (2001). A multinational survey of precision farming early adopters. *Farm Management.* **11**, 147–162.

Proctor, M.E., Siddons, P.A., Jones, R.J.A., Bellamy, P.H. & Keay, C.E. 1998. *LandIS - a land information system for the UK.* In: *Land Information Systems: Developments for planning the sustainable use of land resources*. (eds H.J. Heineke *et al.*), pp. 219-233. European Soil Bureau Research Report No. 4, EUR 17729 (EN), Office for Official Publications of the European Communities, Luxembourg.

Shibata, R. (1986) Regression variables, selection of. In: *Encyclopedia of Statistical Sciences* Vol 7. (Eds. S. Kotz and N.L. Johnson). pp 709–714. Wiley, New York

Smith, L.P. (1976) *The Agricultural Climate of England and Wales.* MAFF Technical Bulletin 35. London.

Tukey, J.W. (1977) *Exploratory data analysis.* Addisson Wesley, Reading MA.

Webster, R. and Oliver, (1990) *Statistical methods in soil and land resource survey*. Oxford University Press.