Feasibility study into combining data resources to explore the drivers of crop yield in winter wheat and oil seed rape

James Rainford
Glyn Jones
Roy Macarthur
David Garthwaite

October 2019

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Contents

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Acknowledgements

Vahid Mojtahed, Yi Lu and Giles Budge (FERA Science Limited) were involved in developing an earlier version of this report which established many of the datasets, conventions and representations used below. This work builds on these foundations with a more aggregated and integrated approach to merging the compiled datasets, and in the use of alternative methods and processes to support analysis.

The authors would like to thank Professor Dionysis Bochtis of the Institute for Bio-economy and Agri-technology (IBO), Centre for Research and Technology Hellas (CERTH) for helpful comments in the development of this work.

This work was conducted on by FERA Science Limited on behalf of the Department for Environment Food & Rural Affairs, project PS2822.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Executive Summary

Analysis of the drivers of yield

This work represents an initial exploration of fusing several large-scale datasets relating to UK agricultural practice to examine the drivers of farm level yields of focal crops. The following datasets were identified for combination, as relevant to understanding crop yields in the UK:

- FERA pesticide usage survey (PUS) aggregated to a farm level
- June farm survey
- DEFRA crop pest survey
- Weather data from the Meteorological office
- Soil type information from the National Soil Map
- Annual economic information from the John Nix farm management pocketbook

Datasets were combined for the biennial period between 2000 and 2010, based on the associated labelling information and geolocation via the June survey. The combined data represents an analytical dataset through which the potential drivers of winter wheats and oilseed rape (OSR) could be examined.

The analytical approach undertaken combines parametric statistical analysis with the use of high-flexibility machine learning techniques (Random Forest) to provide hypothesis-driven and heuristic approaches to revealing the key factors that may be used to predict yield. We defined a core feature set (a limited number of factors that we expected to be important based on existing literature), and a set of expanded features (i.e. all the relevant variables within the study). We undertook statistical hypothesis testing on the core feature set, with machine learning to describe the potential shapes and observed importance of the identified significant drivers. Machine learning was also used to explore the expanded feature set to identify factors that may have been missed in the definition of the core set and which might warrant further investigation into their impacts on yield.

Wheat varieties were divided into two categories (bread wheats and feed wheats) based on the grouping system provided by the National Association of British and Irish Flour Millers ("nabim"). For bread wheats, the most important predictors from core features set include

- the major variety on the holding,
- the area of the holding (larger holdings having higher yield),
- proportion of own seed used (lower yields for more own seed),
- increased yields under dry conditions during the pre-frost period (July to December of the year prior to harvest),
- number of unique active compounds applied, and total mass of pesticide associated with the holding (both having positive effects on yield).

Machine learning using the expanded feature set reinforced these conclusions. In addition, a strong positive signal associated with the application of growth regulators was observed. In machine learning a stepwise relationship between the yield and the number of compounds was observed: applying 12 or more active compounds was associated with a step up in expected yields.

Different factors were associated with yield change in feed wheats compared with bread wheats:

- The strongest yield effect was associated with by diversity of applied compounds (positive effect);
- conditions during the pre-frost period (low rainfall and high humidity being associated with increased yield);

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

- small effect of the number of spray rounds conducted replacing the effect of total pesticide load (positive effect)
- an *absence* of clear effects associated with primary variety

Machine learning analysis of the complete feature set gave, in common with bread wheat, an observation that growth regulators are associated with yields. The recorded proportion of land set aside as grassland was also highlighted for further investigation.

For OSR the key drivers in the statistical analysis include, the diversity of compounds applied and the year of sampling (overall upward trend in yields during the period, with lower values reported in 2004 and 2008)By far the largest and most important identified driver was latitude, with a noticeable stepped effect of increasing yield north of the 52 degrees north (approximately the latitude of Ipswich). The causes of this are unclear and may relate to the distribution of economically important pest species. This interpretation is reinforced by analysis of the expanded feature set which also revealed the quantity of fungicide applied on a holding as a key determinate of yields in OSR crops.

The applied modelling indicated large amounts of (apparently) random variation in yields which could not be accounted within the fitted functions. This is likely the result of a combination of intrinsic noise in the combined datasets, and limited predictive power in the included variables. Confirmatory studies are advised to further investigate the effects identified and their impacts on crop yield across holdings.

We found that, for all three crops, there was a potential sub-population of unusually low-yielding sites compared with their predicted yield. These might represent either unidentified failed crops (with resulting differences in farmer behaviour) or holdings where there are specific socio-economic factors which are associated with reduced yield values (e.g. failure to incorporate innovation). None of the examined drivers, including the expanded feature set, showed useful correlation with the observation of low yield suggesting a role for other potential drivers outside of the scope of this analysis.

In conclusion, our results suggest that diversity of agrochemical inputs is a key component of observed landscape level yields in both wheats and OSR. We also show that climatic conditions during the growth year are key predictors for wheat while OSR appears largely driven by latitude. A less consistent effect of total pesticide load is estimated (positive in bread wheats, non-significant for feed wheats, and negative for OSR). We didn't find evidence of a relation between soil types or disease prevalence and yield. This may be because our measures of disease prevalence were not suitable for finding a relationship. The use of growth regulators in wheats and fungicides in OSR were identified for potential further study particularly in the context of standardised field trials.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Recommendations for future work

The modelling conducted here is preliminary and subject to constraints relating to the content and structure of the underlying datasets. Areas of concern include the representativeness of the yield data available in the PUS (many farms do not provide yield estimates and the remainder may not be fully representative of variation across the UK), as well as the representation of some of the variables, particularly for soil type and disease prevalence (which were subject to sampling and aggregation constraints within this study). Some of these issues could be minimised by combining similar analytical approaches with data from standardised plots such as those run by the AHDB for variety level yield assessment. These static sites, with standardised input regimes across years, may help in characterising the abiotic components of crop yield (e.g. weather, soil type and latitude).

Modelling the sub-population of low yielding sites discussed above, remains an outstanding challenge for statistical analysis. Hurdle modelling, based on whether localities achieved a commercially viable yield, or similar techniques, may help to refine the models fitted here.

Alternatively, increased understanding of how this population might be characterised, e.g. based on socio-economic factors, may provide insight into how they should be represented in any future modelling and the consequences for policy decisions around yield

This work highlights the opportunities and challenges that arise from combining related datasets in the agrarian sector. We provide a framework and discussion relating to the overlapping use of statistical and machine learning based techniques in the context of the numeric analysis of fused datasets. To summarise, statistical procedures are based on an explicit model of the system under study, that are dependent on knowledge regarding the relation between the studied factors and the way in which observations may vary. This provides greater power and interpretability when testing hypotheses about the workings of the system, where the model is judged to be adequate.

The flexibility of machine learnings makes it better able to reflect complex relationships that may be present within data (e.g. for forecasting future states) and which may reflect system processes. However, this same flexibility can lead to over reliance on random variation in the data used to fit the function and undermine the generality of the resulting model. These contrasting strengths represent important considerations in how methods are to be used in numeric analyses, and how to structure similar studies in other relevant policy areas.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

1. Background

Crop yields are one of the key measures within agrarian systems and have been widely studied in a range of contexts, particularly in relation to food security and changes in policy frameworks. While considerable work has focused on overall trends in yields, information relating to specific holdings and the meso-scale processes that generate yields is often highly fragmented, with limited attention given to understanding common processes and factors that may be associated with high yielding holdings (1). In this study we examine yield values for the most important grain and fodder crop within the UK, winter wheat, and its primary break crop, oilseed rape (henceforth OSR). Our focus is on yields during 2000 to 2010 based on biennial yield data collected as part of the arable pesticide usage survey (conducted every two years by FERA on behalf of the Chemicals Regulation Division). The aim will be to explore farm-level factors associated with yield values in these key products based on broadscale standardised datasets. The aim is to explore both how combining data sets improves understanding of yield drivers and to identify possible policy areas which may warrant further analysis in the context of the UK agricultural system.

Aggregated wheat yields across the UK are associated with a series of well publicised trends over recent decades. Following significant technological and variety improvements in the 70's and early 80's a gradual upward trend in year on year improvements in wheats yields was observed up to the mid 90's followed by a long term plateau up-to at least 2011 (2)[1]. The causes for this relative stasis are incompletely understood. Evidence suggests a mix of agronomic effects particularly around nutrition and fertiliser usage, trends towards earlier sowing and reduced tillage, as well as various weather related impacts may collectively offset the increased potential of new varieties as they enter the market (2, 3). Comparisons across Europe indicate a non-linear response in yields of winter wheats to seasonal temperature variation (4). Recently there has also been discussion of the impacts of climate change, with a particular emphasis on the role of CO_2 fertilisation, and concerns around the impacts of summer drought (5, 6).

Over the same period, OSR yields underwent a notable decline during the late 80s to early 90s before stabilising and trending upwards since around 2002. This pattern has been attributed to changes in agronomy (particularly the use of sulphur fertilisers) and a limited uptake in high yielding varieties prior to the early 2000's (2) This periods is also associated with notable increases in typical pesticide inputs associated with OSR and the transition towards being considered an intensively treated crop. Recent work has also strongly indicated an impact of early winter temperatures on OSR yield which may also contribute to these longer term trends (7). However, the specific relations in the UK remain unclear and there is limited discussion of farm level variation.

While these reviews discuss various causes of the trends in the overall yields in the focal crops, information on between-farm variation and its potential causes is much more incomplete, or is based on outdated data (e.g.(8, 9)). Vigani, Cerezo (1) present information for wheat farmers in France and Hungary based on a broad-scale survey which concluded that there were significant differences in practice, preferences and perceived cost-benefits between countries that limited the potential for gaining information about drivers of yield that could be applied generally. Likewise, informal discussion within the industry and elsewhere, assumes that a subset of low yielding, presumed to be highly conservative holdings exist, although with little discussion of how these might be characterised and their effect on overall wheat yields. Also absent, at least within the UK, is consideration of the spatial structure underpinning yield. This is surprising given the oft cited relevance of climatic and soil

[1] There is some dispute as to whether the trend persists in more recent data, which lies outside of the scope of this analysis

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

related factors (e.g.(1)), which would be expected to generate strong spatial clustering in yields. In the UK it is generally acknowledged that the highest yields tend to be associated with northerly latitudes. This is believed to be primarily associated with the distribution of economically important pests but the extent to which this generalises across holdings and the relative size of the effect remains incompletely understood.

This work originated from, and builds on, efforts to synthesise combinations of different agronometric datasets relevant to patterns in UK crop yields. UK agricultural production is a highly monitored system with numerous standardised datasets being compiled to represent specific aspects relevant to crop yield. Notable examples used in this study include:

- The pesticide usage survey [PUS] (managed by FERA on behalf of the Chemicals Regulation Division [CRD]) which surveys the usage of pesticides and related agri-compounds on a sample of locations (stratified by farm size and region) across the UK. The data here is taken from the arable pesticide use survey, conducted every two years during the period examined.
- The DEFRA 'June Survey' which is an annual survey of between 30,000 and 70,000 holdings each year, with an emphasis on understanding patterns of behaviour in the agrarian sector, the impacts of policy change and calculating national inventories of various relevant products.
- The DEFRA Crop Disease Survey, which is conducted annually on a stratified sample of localities, with an emphasis on monitoring the presence of certain well-known diseases on several important crops including winter wheat and OSR.

Following discussion with DEFRA and CRD these three datasets, in combination with other information such as: geographic location[2], climatic data[3], information about soil types[4], and some relevant economic variables[5] were identified as potentially relevant to understanding UK yields with respect to the focal crops. Here we present an analysis of this combined dataset focusing on the biennial sequence of yield values during the period 2000 to 2010[6]. Our focus here is on understanding yields based on a simplified framework of measures arising from the combined data (the 'feature set'). The construction of the combined data set is described below, and a copy of the code used in assembly (R version 3.5.1 (10); various packages) is provided in the attached code appendix.

[2] taken from the farm centroids in June survey
[3] taken from the Meteorological Office
[4] adapted from National Soil Map for England
[5] taken from the John Nix farm management pocketbook
[6] Following 2010 the structure of the PUS survey, from which the yield data are derived, underwent major revisions in the data collection process. One of the consequences of this is that yield information after this date ceases to be directly comparable with that generated during the preceding period. The presented analyses are restricted to the period where comparable data was available across all studied datasets.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

2. Conceptual approach

2.1. Statistical analysis and machine learning; an overview

Insights from the combined dataset were developed using a combination of both statistical modelling and machine learning. In the context of this study, statistical modelling is defined as the application of classical parametric linear models fitted under ordinary least squares. The modelling approach is based primarily on the application of linear mixed effect models based on the implementation provided in the R package *nlme* (11). Full details on the implementation are provided in the descriptions of the various analyses and in the attached code appendix.

Machine learning is a somewhat more ambiguous term, the use of which warrants discussion in the context of this study. Here, the focus is on one of the two major classes of machine learning problems, the so called 'supervised' problem wherein the objective is to use an algorithmic procedure (as opposed to a human/knowledge driven process of model selection) to identify [under some criteria] an optimal mathematical object (or set of objects) (henceforth 'the model') to the relationship between a set of variables (henceforth the 'feature set') and the value of some dependent measurement ('the endpoint'). Classic examples of supervised problems include classification, where the end point is membership of some predefined grouping and regression, where the end point is the value of some continuous variable (such as yield).

There is conceptual similarity between numeric analysis using supervised ML and statistical modelling. Both fit some sort of descriptor to feature set based on the values of the endpoint and in both cases the parameters of the description are optimised to minimise the error in the representation of the end-point values. However, there are some conceptual distinctions which serve to separate statistical modelling from the majority of commonly implemented machine learning approaches, and which define a distinct and non-overlapping role within numerical analysis.

The first important difference is that many machine learning procedures combine within a single algorithm two different stages of the modelling process, feature selection (i.e. the choice of which elements of the feature set are relevant for inclusion in the generated description) and parameter optimisation (i.e. setting the values within the model object which link elements of the feature set to the end point). In a statistical modelling procedure these would typically be discreet steps, with manual feature selection based on expert theoretical understanding of the system being modelled, followed by parameter estimation using an optimisation algorithm. Typically, this will be followed by a model comparison process undertaken by a statistician, wherein the fit to the data is compared under different versions of model, to identify significant parameters and to examine whether the assumptions underlying the model are being met. In most common forms of ML these steps are treated as a single process wherein the feature selection and parameter values are jointly estimated while fitting the model object. Hence, concepts relating to formal model comparison are largly irrelevant in an ML context, with less emphasis placed on comparing alternative representations of the feature set. There are automated methods of feature selection applicable to what is otherwise conventional statistical models (the most well-known being LASSO or Ridge regression (12)) but in general this is an important distinction which shapes the utility of different methods to addressing different questions.

Another very important difference between most common ML procedures and statistical modelling is related to treatment of the variation in observed endpoints which is not explained by the fitted model. In general models which are perfect descriptions of the endpoint are undesirable (for reasons relating to 'overfitting', see below). Hence, there is typically a component of the variation which remains

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

unexplained even after the optimal parameters have been estimated. This is formally termed the 'error' or 'residual' on the model and its treatment has important consequences for interpretation of the findings.

A key property of parametric statistics, and one which is largely responsible for the utility and power of these methods, are explicit assumptions in the model fitting regarding the distribution of the error on the model. In simple terms, when fitting a parametric statistical model, we assume that the values of the error are drawn from an unbiased normal distribution. This property allows us to describe the 'fit' of a model to a dataset in a rigorous and mathematically well-defined way (e.g. via maximum likelihood or Akaike's information criterion [AIC]) In addition, it is by making assumptions regarding the distribution of the error that we can calculate the uncertainty and confidence intervals around the model parameters, which can often be vital in of our interpretation of the model effects. However, this limits the application of statistical modelling to cases where the assumption is (approximately) true.

By contrast most machine learning methodologies do not have a formal concept of error structure independent of the algorithm used to fit the model. Optimality in these methods is defined relative to the fitting algorithm and (usually) does not make the same sort of stringent assumptions about the distribution of error around the fitted model object. This can be very freeing in terms of application; in that it means that many ML techniques can be applied to data which fails to meet the assumptions necessary for statistical analysis. However, by the same token most ML approaches cannot benefit from many of the advantages which the assumed error structure provides in statistical modelling, e.g. simple methods for objectively choosing the best model among alternatives, and it is often challenging to estimate appropriate confidence intervals and uncertainty around estimates arising from ML approaches. This is one of the major reasons why of the two approaches, classical statistical modelling is often more appropriate when the question of interest relates to hypothesis testing (i.e. "Does this factor cause our endpoint to change?") while ML is most powerful when applied to forecasting (where the formal structure of the model is less important than the quality of the fit to the data) and/or exploratory analysis (where the benefits of flexibility in representing the feature set may outweigh issues around model interpretability).

One of the key properties of an informative model is that it should provide an acceptable description of the underlying system beyond the specific information used to generate the fit. This is perhaps most obvious when considering a forecasting framework, where conceptually a model might be trained on one data set (e.g. the yield values for a particular year henceforth the 'training data') and then applied to different but similarly structured dataset in order to generate an output (e.g. forecasting the next years yield values based on new values for the same features; henceforth the 'use data'). The extent to which a ML model trained on one data set is applicable to other sets is described as generalisability.

One of the key challenges in appropriate use of ML techniques is the issue that, because these techniques are so flexible in how they represent the relationships between feature set and end-point it is often the case that the resulting objects can place high weight on idiosyncrasies within the training data which are atypical of the use data. Where this occurs (termed 'overfitting'), it can result in incorrect predictions from the ML model when applied outside of the training data. One way to examine for overfitting is to designate some part of the known data as 'test' data, which is excluded from training the model. This provides a test of how the model preforms on new data. The extent to which this provides a realistic test depends on the relation between the test data and fitted data being sufficiently close to the relation between use data and fitted data. Ideally the training-test and training-use data should, as far as possible have identical relations, although in practice this is rarely possible. The emphasis is typically on avoiding systematic bias which might invalidate the test process.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

In the case of well-trained but not overfitted model there should be consistent high levels of fit (e.g. high correspondence between the fitted and predicted values) in both the training [which shows that the algorithm has fitted well to the provided data] and testing case [the latter showing that the fit is general and encompasses never-before seen data. This provides assurance that the ML model reflects some real physical process within the system]. This framework of training verses test datasets is also one of the only ways in which different methods for machine learning (in this case including statistical modelling) may be directly compared as improved fit on the test data is an index of improved model performance.

To summarise, machine learning methods are very effective for fitting complex non-linear functions to relate a feature set and an endpoint in a highly optimised and systematised way. ML can often outperform equivalent statistical approaches in terms of their representation of the shape of relationships within the training data. However, the lack of an explicit error structure and the tendency towards overfitting means that the outcomes of such methods need to be interpreted with care (although this can be moderated somewhat by use of a test dataset). ML faces greater challenges around hypothesis testing and the scientific understanding of a complex system, as these situations typically rely on a more comprehensive understanding of the error structure within the data.

2.2. Applications to this study

For this study, due to the general lack of knowledge regarding the potential drivers of yield within our candidate datasets we have adopted a hybrid approach whereby we attempt to make use of the strengths of both statistical methods and ML for different purposes within the analysis. We began our investigation with a statistical analysis based on a restricted set of probable drivers which we refer to as the 'core feature set'[7]. This is our initial hypothesis driven approach aimed at identifying which of our candidate features are potentially relevant in the context of yield and the extent to which we can fit adequate models with the restrictive assumptions of linear models.

To support the interpretation of these models we have also fitted a ML algorithm (random forest) to the same data set with the intention of looking for structural elements, such as non-linearity and interactions which might be impacting on the fit of our linear models. Finally, we expand the feature set to include a wider array of possible predictors (the 'Expanded feature set') and repeat the ML analysis looking for any relevant predictors potentially absent from our analysis of the core feature set and to examine the impacts of this wider array of predictors on the conclusions about the identified drivers

2.3. Random Forests

The machine learning procedure used throughout this study is the widely used ensemble learner known as Random Forest (13-16). Ensemble learners are a family of techniques based on the principal that combining a set of different model objects (often trained on subsets of the original data) can generate a combined prediction which is much more powerful than any such object treated individually. In the context of random forest the individual objects are decision trees, (sometimes also referred to as Classification And Regression Trees; CART), a form of classification algorithm which identifies splits based on the input feature set which can be used to divide subpopulations which differ in their end point value (in this case yield).

[7] A core set was defined to avoid a weakness of complex linear models whereby they tend to become mathematically undefined if large numbers of parameters are fitted simultaneously, a feature which ML avoids through integrated feature selection

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Individually, CART have a strong tendency toward overfitting however this can overcome by using a subsampling procedure such that each of the set of CART contributing to the forest is estimated on only a sample of the training set (a technique known as 'bagging') and a random sample of the available features (13). The resulting collection of trees (known as the 'forest') can be used both for joint prediction based on the feature set (e.g. from test data) and has informative properties that relate to the distribution and relative 'importance' of the features being fitted (variable importance).

Variable Importance in a random forest (related, but not equivalent to, significance in a statistical model) can be measured in several ways focusing on different elements of the model fit. An approach used here is based on the change in the mean squared error on the predicted values (MSE) following the exclusion of the focal variable. This can be expressed either in terms of the overall model prediction or the so called 'out of the bag sample' [OOB], which is the set of data generated for each tree using the rows that are not included for estimating the CART function. Another approach, more suitable to categorical factors, is the node purity increase associated with including a particular measure (i.e. the extent to which the inclusion of a measure causes the outputs to group more homogeneously based on the decreased sum of square deviation within terminal groups). The final measure used is based on applying a binomial test of the hypotheses that inclusion of splits based on a variable within the set of calculated trees is a predicted function of the number of times that variable was included in the samples used to estimate the trees. Deviation from this hypothesis can be taken as statistical evidence than a variable is identified in the decision tree algorithm more than would be expected by chance, which is evidence for increased importance (the methodology follows that of the R package *RandomForestExplainer* (17)).

In addition to these various measures of importance we will also explore the shape of the marginal effect of variables on the predicted outcome of the random forest using partial dependence plots. These one-dimensional plots display the estimated relationship of that feature of the value of the end point under the assumption of independence.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

3. Dataset construction

The selection of datasets for inclusion in this analysis is inherited from a previous phase of analysis and based on discussion with DEFRA and CRD. It includes the PUS, DEFRA June survey, DEFRA crop disease survey, data from the Meteorological Office, soil types from the National Soil Map for England and economic variables from the John Nix farm management for the biennial period from 2000 to 2010[8] .

The aim of the data assembly was to provide values for each of the factors taken from each of the six data sources that could be applied to each holding defined by a county parish holding (CPH) number and the year of sampling. The assembly of the data set was based on combining observations about holdings based on CPH number with explicitly geo-located observations using easting and northing map references for the centroids of the holdings. Internal holding codes for the PUS surveys for each year were mapped to CPH based on information provided by the PUS team and thus associated with records from the June survey. Information from the crop disease survey and met-office weather data were mapped via the recorded easting and northing references as outlined in Appendix 1: Dataset construction. Eastings and northings were used to estimate postcode districts (based on data provided from post office records), which were used to associate the extracted (raw information on soil types; prior to calculation of principal components). Economic variables were mapped to the year of survey. Information explicitly identifying individual holdings was not retained during analysis and all visualisations are aggregated to prevent the identification of any specific holdings.

Factors judged likely to be most important in driving yield by agronomy specialists at FERA were put into a *core feature set*. In order to produce the core feature set some individual factors were aggregated into values that were considered to be likely to be related to yield, for example: total quantity of pesticide applied, number of different pesticides applied. Remaining factors were assigned to an *expanded feature set.*

A full outline of the aggregation and mapping of the various datasets is provided in Appendix 1: Dataset construction and in the attached R code appendix.

[8] Yield data from the PUS is available for every two years within the studied period. All other datasets are matched to these intervals

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4. Modelling and Analysis

4.1. High-quality bread wheats (nabim One group)

Modelling consisted of applying hypothesis driven statistical modelling to the core feature set before using the more flexible machine learning procedures to

a) describe the shape of the relationship between important variables and yield (information which is not obtainable from the linear modelling used in statistical analysis)

b) to investigate any potential interactions of interest between parameters, and finally

c) to expand the scope of analysis beyond what is feasible for statistical procedures. This allows us to identify further non-core candidates for more in depth analysis.

Prior to modelling we explored the potential for spatial structure in the yield values independent of the fitted variables. Spatial structuring in yield values reflects how similarity in yields between holdings varies as a function of distance. This is of interest in our modelling of yields because many of the potential predictors: day length, rainfall etc, are expected to have or reflect spatial gradients. Hence, we may expect spatially correlated yield values, which has implications for model fitting and representation of the error structure in the dataset.

For hypothesis testing we adopted a step-down modelling approach whereby an initial model consisting of all core-features was fitted, and then non-significant parameters were sequentially removed until an optimal model was identified. Spatial locations of sites were represented as the latitude and longitudinal equivalents to the reported easting and northing values for the centroids of the CPH numbers associated with the holdings in a particular year (see Dataset construction).

4.1.1. Spatial structure assessment

In order to describe the spatial structure we fitted a least squares estimate of the change in yield values between pairs of holding over varying distance using the R package *spatial* (18). The results of this estimation can be viewed as a plot (known as a semi-variogram) which summarised the expected similarity between pairs of sites across varying distance. The semi-variogram for yields of the nabim One varieties of wheats is shown in Figure 1. We can see that holdings which are very close together tend to generate similar yield values (and hence show low variation with respect to the semi-variogram). This effect rapidly drops off above around 0.1 degrees of separation and the variation then becomes largely unstructured. This pattern is strongly suggestive that if there are any spatial structuring elements present within the data set there are operating at highly localised scales, as opposed to following large scale climatic or geological gradients.

We can assess the significance of the potential spatial structuring using the Moran I statistic, which is a measure of the relative contribution of spatial autocorrelation to the overall variance(19, 20). The implementation used is taken from the R package *pgirmess* (21). The most significant spatial structuring is observed at the very smallest spatial scales (so small that these may in fact reflect similarities between the limited number of holdings which are sampled in multiple years), with another marginally significant effect observed around 2 degrees of distance (just over 220 kilometres separation). On the basis of these observations we have elected to explicitly include an element of spatial structure in our model of nabim One wheat yields, which we have represented as an exponential correlation based on distance. This function models an expected structure where there

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

should be high similarly at very close distances with a rapid decline to zero correlation, and hence is the closest approximate match to what is observed in the empirical variogram.

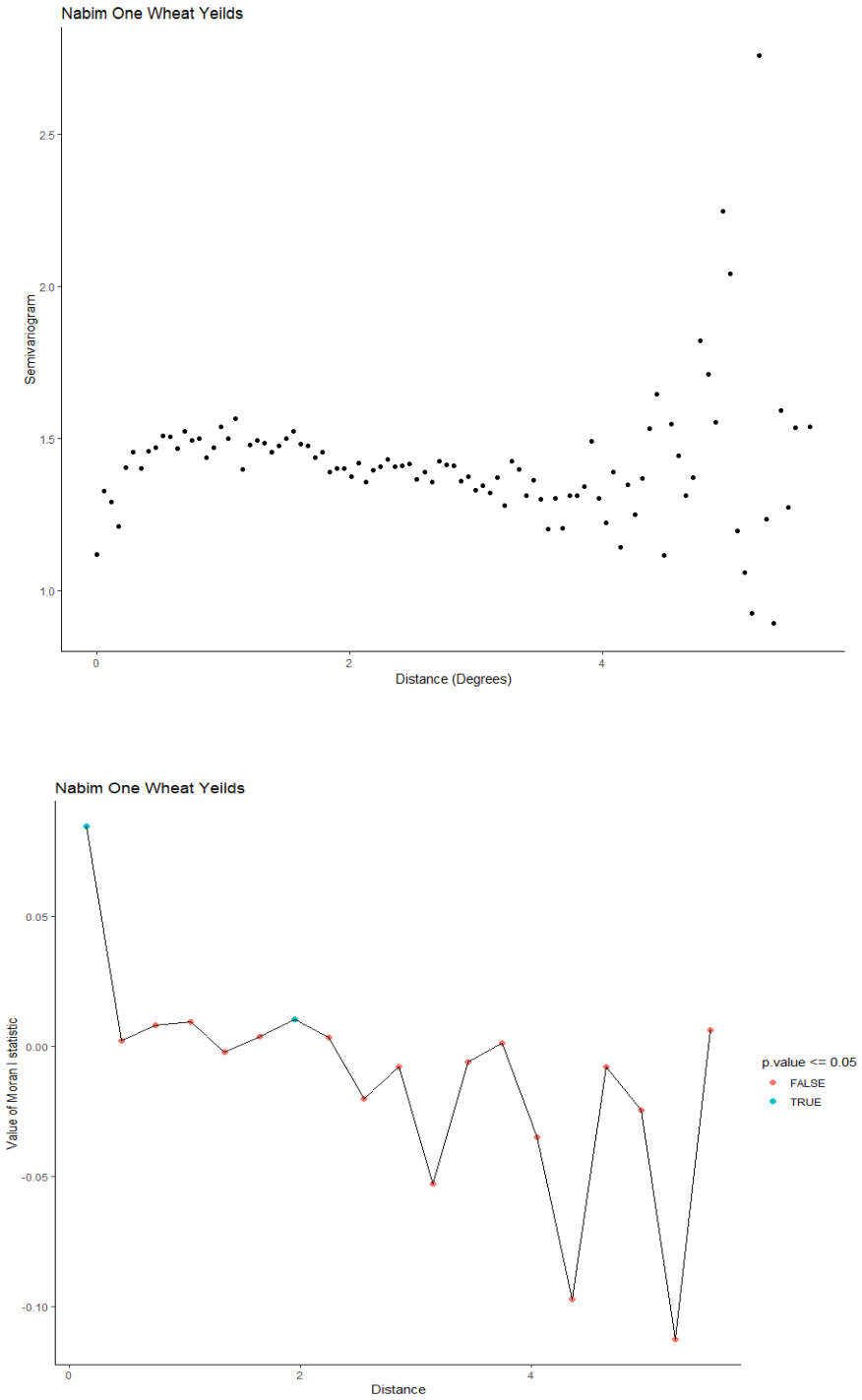


Figure 1 Visualisation of the empirical spatial structure associated with yields of nabim One wheat varieties. Upper: Empirical semi-variogram based on estimated least squares surface; Lower: results of simulation-based testing of the value of Moran's I (an index of spatial autocorrelation) and varying distance, points shown in blue represent greater than expected spatial autocorrelation within the sample. Distances are calculated based on latitude and longitude coordinates using linear approximation. See text for discussion.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

In total the nabim One data set included 615 records representing 589 unique CPH numbers. Throughout modelling, here and elsewhere, any missing data were fixed to the mean value of the relevant column. To reduce the effect of outliers on parameter estimates we restricted the analysis of the nabim One yields to only include values within a 50% window around the overall population mean (approximately 4.2-12.6 tonnes per hectare; shown by the red lines on Figure 2).

The core feature set applied to wheat is shown below. It includes a three-way interaction between our aggregated descriptions of pesticide usage (Mean_Spray_Rounds*Count_Compounds *Total_Pestcide)[9], which was one of the key parameters of interest in terms of the effect on yield. Names given refer to those in Appendix 1: Dataset construction

nabinOne_Yield~
log(Area) [natural log],
Prop_Own_seed,
Primary_Variety
Mean_Spray_Rounds*Count_Compounds*Total_Pestcide [interaction],
log(Land_Farmed_by_Farmer_June) [natural log],
mean.wind.preharvest
mean.RH.prefrost,
mean.temp.prefrost
mean.rain.prefrost,
Soil_PCA_ScorePC1
Soil_PCA_ScorePC2
Average_Corrected_Tritici_index
Average_Corrected_brust_index
Average_Corrected_yrust_index,
Year [as a multilevel factor]
latitude+ [Exponential spatial auto correlation]

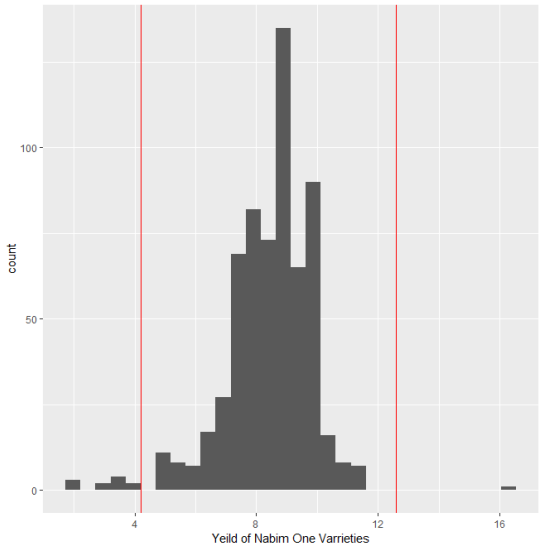


Figure 2 Distribution of Yield values associated with nabim One varieties of wheat in the analyses data set. The red lines show the cut off thresholds for the data analyses in this study corresponding to a 50% window around the overall population mean

[9] Syntax used to denote a three-way interaction between the number of spray rounds undertaken, the number of unique active compounds applied, and the total mass of pesticide applied. See Appendix 1: Dataset construction for discussion.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Table 1 Summary of the significant parameters identified from the analyses of the core feature set under statistical modelling to the yield values for nabim One varieties of wheat. Columns list the estimated parameter, standard error and the results of model testing to compare the relative fit to models where the parameter is removed. Names given reference those in the Dataset Construction section. See text for discussion

Parameter	Estimate	Standard Error on the Estimated Value	AIC of favoured model (df)	AIC of favoured model excluding parameter (df)	Likelihood ratio value of model test Ratio	P value of model comparison test
Primary Variety	see Figure 3		1809.4 (29)	1902.3 (11)	128.9	<.0001
Log (Area)	0.333	0.0536		1846.2 (28)	38.86	<.0001
Log(Land_Farmed_by_Farmer_June)	-0.1803	0.0735		1813.6(28)	6.24	0.012 [Marginal]
Prop_Own_seed	-0.441	0.1057		1825.2(28)	17.89	<.0001
Count Compounds	0.0379	0.0124		1816.8 (28)	9.484	0.0021
Total Pesticide	0.1033	0.0240		1826.4 (28)	19.04	<.0001
mean.rain.prefrost	-0.2927	0.0784		1821.2 (28)	13.81	<.0001
latitude	0.1057	0.0529		1811.51 (28)	4.149	0.0416 [Marginal]

4.1.2. Statistical modelling (Core feature set)

Following step down simplification the resulting optimal model for the nabim One varieties of wheat is described in Table 1. Drivers identified include:

- **Primary Variety**[10]. Evidence for significant impacts of the primary variety can be broken down to represent specifically high yields associated with holdings where the variety 'humber' is the predominant cultivar, and low yield values associated with the 'claire', 'hereward' ,'malacca' and 'soissons'(Figure 3).The challenge for interpreting these findings is a) resolving the relatively small sample sizes (minimum of 20 holdings) associated with each primary variety, and b) resolving the fact that many of these primary varieties are predominantly used as feed wheats. Hence it may be management practices on the holding, as opposed to the genetic variation in the crop grown, which are identified as driving yield values. This ambiguity is one of the drawbacks of the aggregation of the PUS data to farm level. More highly resolved data would be required to tease out effects[11]. Nevertheless, the observation that different major varieties are associated with systematic differences in yield over the decadal timescale considered here may warrant further investigation particularly if it is aligned to standardised potential yield data provided annually by the AHDB.
- (log) **Area of the holding**. Larger holdings having higher yield per area values. This is likely to be a reflection of the impact of economies of scale on several potential drivers of yield, e.g. mechanisation and agrochemical input.
- **Proportion of own seed**. This measure was included as a proxy for secondary sowing of wheat crops. Second wheat crops are often used as part of a cultivation cycle, and which often include a high proportion of the farmers own seed. Such sowings are widely recognised as

[10] The majority variety on the holding, not necessarily the majority nabim One variety

[11] Resolving PUS data to variety level is not possible due to single yield values being reported for fields containing multiple varieties. Resolving to field level introduces other issues, as individual fields cannot be geo-located for combination with soil, pest occurrence or weather data under the system outlined here.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

producing lower yield values although, due to reduced costs, the economic margins may remain viable. This effect is estimated as the single largest in the optimal model (discounting the fitted spatial structure).

- **Total volume and diversity of pesticides applied**. Both the total volume of pesticide usage and the diversity of active compounds applied are observed to have positive impact on yield. Interestingly these is no evidence to suggest that the measures used to represent the pest distribution from the crop disease survey have significant effect in the model of yield, which may indicate that the primary threat for which agrochemicals must be applied is not one of those measured in the crop disease survey (which is primarily focused on fungal pathogens). Alternatively, it may be that a large proportion of the recorded inputs are precautionary, and aimed at minimising risk of exposure, as opposed to being responses to pest outbreaks. It should also be noted that other factors, such as genetic variety and recent weather conditions can play a role in shaping regimes of pesticide application, which may weaken the links between pest occurrence and overall application.
- **Mean rainfall during pre-frost period**. One of the most surprising effects identified in the model of yield is a negative impact of rainfall during pre-frost. More detailed examination of the shape of the effect indicates that this is driven primarily by several very dry sites which are associated with high yield values, rather than being reflective of a clear overall trend. As a result this parameter should be interpreted with caution.

Other less significant potential drivers include that of the area farmed by the farmer (a negative trend largely driven by extreme outliers associated with the very largest holdings) and very small marginally significant latitudinal gradient, with very slight increases in yields at high latitudes. While statistically significant drivers of yield were found, the proportion of variation in yield described by the model is small (r-squared 0.23). This means that the great majority of variation in yields is either driven by factors that are not included or is largely random with respect to the fitted model. Primary variety is estimated to lead to a potential expected difference of 2.5 T/Ha between the highest and lowest yielding varieties; other factors may lead to an expected difference of approximately 1T/Ha. (see Appendix 2: Visualising the shape of numeric parameter effects in statistical modelling; Figure 31)

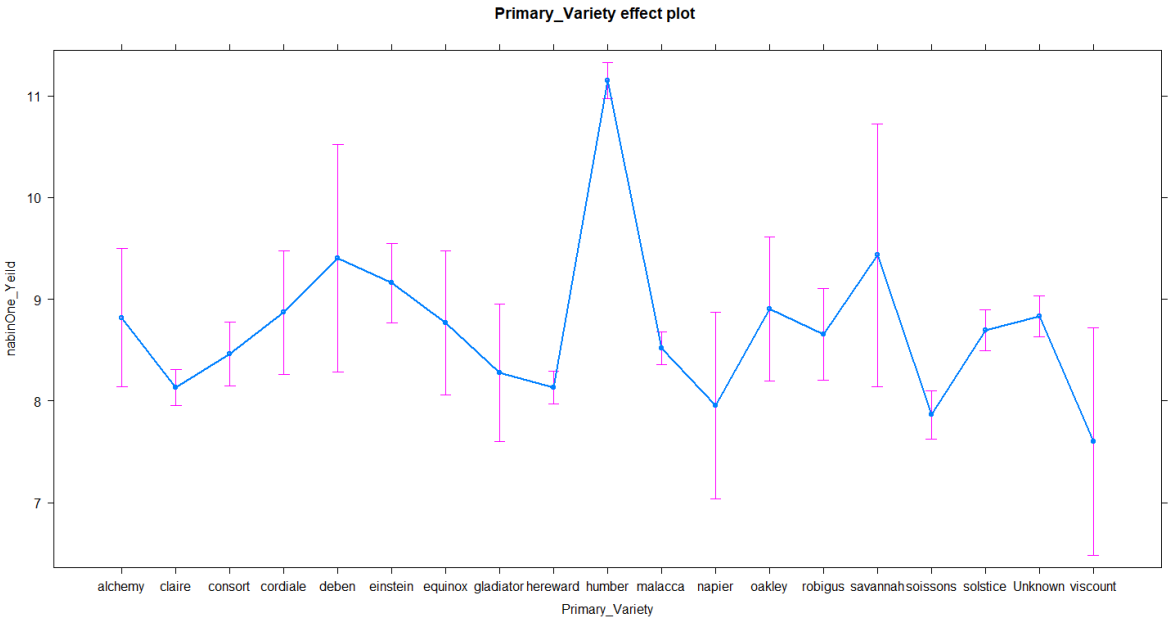


Figure 3 Estimated effects (model estimate and confidence interval) for the levels of Primary Variety in the optimal model for nabim One wheats.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

In statistical modelling, our estimates of effect sizes are only reliable so long as the model fits the observations and our assumptions about the form of the residual variation around the fit are adequately met. Examining whether the residual variation conforms to expectations is an important and informative part of statistical modelling. One way to visualise the variation of observations around fitted values is the quantile-quantile plot (Figure 4; Left). Here empirical quantiles of residual variation around the model are plotted against the theoretical quantiles of the assumed distribution (an unbiased normal distribution). A good fit produces residuals that fall close to the expected straight line. Examining this plot shows that at the low end of the distribution there are points which show a poor fit. Observations are plotted against fitted values (Figure 4; Right) we can see that the non-conforming data tends to cluster at the low end of the distribution and in particular all points where the observed yield is less than 6 tonnes per hectare are poorly described in terms of the fitted model. This a) raises issues in terms of the interpretation of the fitted model and b) represents an interesting divergence from observations in previous studies on the distribution of farm level yields in wheats (22). Non-normality in yield distributions has been previously reported (e.g. (23)), although correcting for this in modelling requires specialist tools beyond the scope to this analysis, particularly where geographic structuring/ correlation is also of interest within the study (24, 25).

One interpretation for the observed difference from the expected error structure is the presence of an unrecognised population of holdings with systematically lower than expected yields. This impression is reinforced by examining the prediction interval around of the some of the key parameters of the model and in particular the failure of the low yielding site to fall within the envelope of the predicted model parameters (Appendix 2: Visualising the shape of numeric parameter effects in statistical modelling). This suggests that underlying processes driving yield at these sites includes factors that do not drive yield in most sites. Possible reasons for this are:

- These records could represent errors in the yield data collected for the PUS. Systematic bias in data collection within the PUS is considered unlikely due to the structure of the survey. However, other issues such as non-comparable cropping data, issues around mixed fields and /or re-sown crops could contribute to errors and would account for some of the very low values reported for problematic sites.
- These records may represent failed crops (for whatever reason) which caused changes in behaviour, such as differences in pesticide usage.
- These records maybe evidence for a subpopulation of relatively inefficient farms, where low yield is a function of some unobserved biological (e.g. a sporadic pest outbreak) or social factors (e.g. failure to uptake relevant technologies). The existence of a low efficacy population with UK wheat farms has previously been speculated [e.g. (26)], although there is limited information on their characterisation. A potential extension of the work conducted here could be to examine the extent to which other factors, including social factors, are associated with wheat yields.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

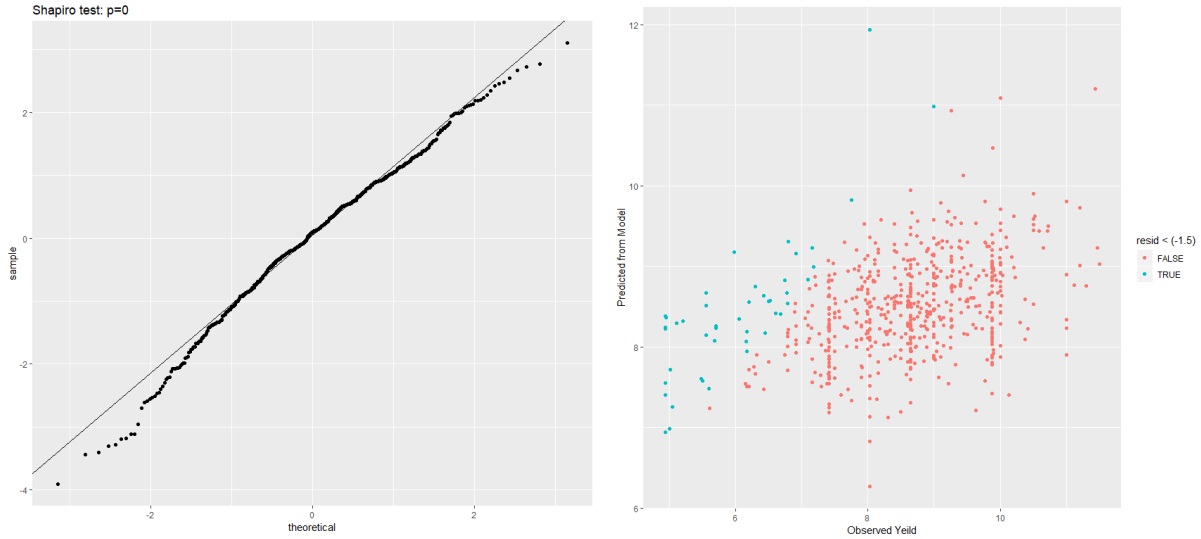


Figure 4 Diagnostic plots for the relative fit of the optimal model identified for nabin One wheats relative to the assumed error structure for a linear model. Left; the qqnorm plot (described in text; the line represents the ideal assumed error structure and the points represent that observed); Right Observed (x axis) and predicted values from the model (y axis) points associated with large negative residuals are highlighted in blue

4.1.3. Machine learning (Core feature set)

After fitting a suitably tuned random forest to the dataset the parameters of highest importance largely correspond to those identified in the statistical analysis. There are some notable differences in the relative impact as represented by estimated importance which combines concepts of significance and effect size (Figure 5). By far the most important drivers identified are the area of the holding and the diversity of the applied active ingredients ('Count Compounds'). Also identified as being of high importance is the Primary variety, with a cluster of other measures including latitude, total pesticide usage and the number of spray rounds appearing a group of somewhat lower relevance; Figure 5.

Focusing on the two most important identified measures, the partial dependence plot of the diversity of compounds applied is striking in that it implies a stepped function in relationship between this measure and yield (Figure 6; Upper Left). The impact of a greater diversity of applied compounds has a positive impact on yield after a minimum of eight actives are included before levelling out after around 17 compounds applied. At 12 active ingredients however, there is a major transition associated with greatly enhanced expected yield values above this threshold, which may imply structural differences in the profile of spraying regimes as a potential component of predicted yield.

Farm area shows a very simple relationship with yield values: yields are predicted to be extremely low on the very smallest holdings, possibly due to the effects of economies of scale, and then rapidly level out such that there is very little impact of increasingly large holdings sizes (Figure 6; Upper Right)[12]. Brief investigation reveals limited evidence of structured interactions between the two parameters with the exception that they are strongly reinforcing for small area and low diversity of compounds.

[12] Although it should be noted that the area of an individual holding is not always indicative of scale of the parent business, as some of the farms sampled within the PUS are managed as part of larger collective units.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

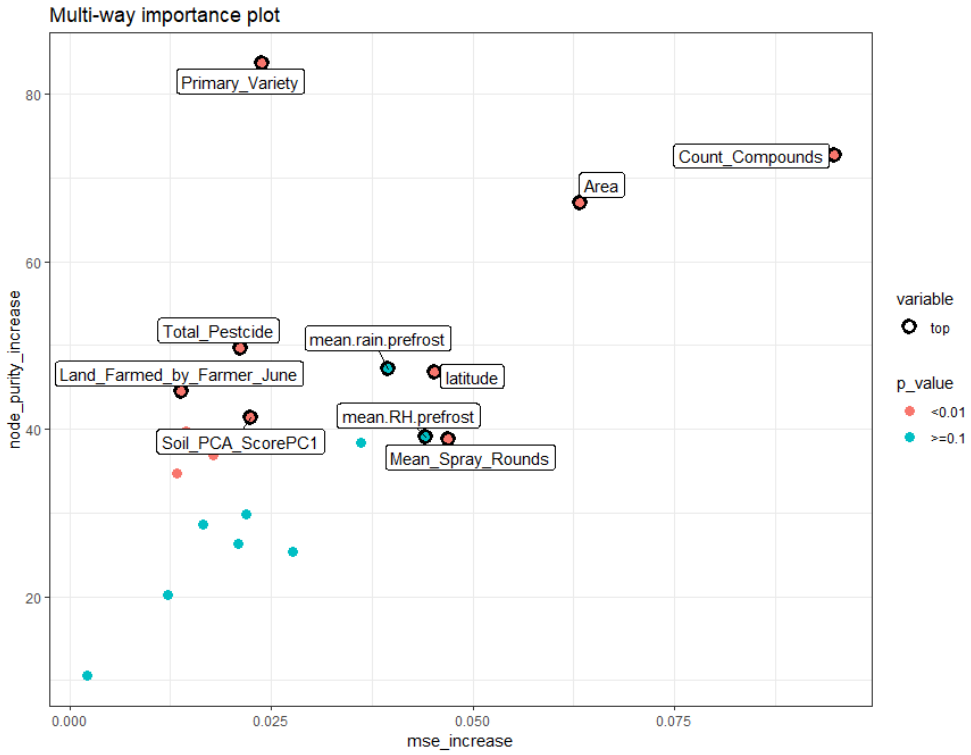


Figure 5 Scores in relation to various measures of importance for variables in the random forest of the core feature set as applied to nabim One wheats. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

While the overall measure importance from the random forests largely reinforce the conclusions generated in the statistical modelling, it should be noted that there is evidence for over-fitting. If the forest is recalculated using a random sample of 80% of the original dataset, the resulting fitted model is similar to that described above: the fit appears to be good (R-squared is 0.85). However, upon application to the remaining 20% of data (unseen during training), the resulting fit is very poor (R-squared of predicted verses observed values in the test set is 0.26). This is evidence that the resulting model does not provide a useful way of predicting yield. This is of concern, because if our model was truly reflective of physical processes driving yield, we would expect similar performance on both training and test data. This implies we should be sceptical of over-interpretation of the resulting model fit.

Comparable values from the statistical models are 0.162 with respect to the training data and 0.060 to testing data, hence the forest is doing better job of representing the data than is the case for the linear models, but it currently neither of the models appear to be making *reliable* estimates of yield . This can be considered evidence that there is substantial variation in yield which is not captured in the data set analysed, suggesting that either the representation of the information is inadequate to accurately reflect the relationships with yield or, far more likely, that there is insufficient power in the included measures to predict yield values at a the level of individual holding.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

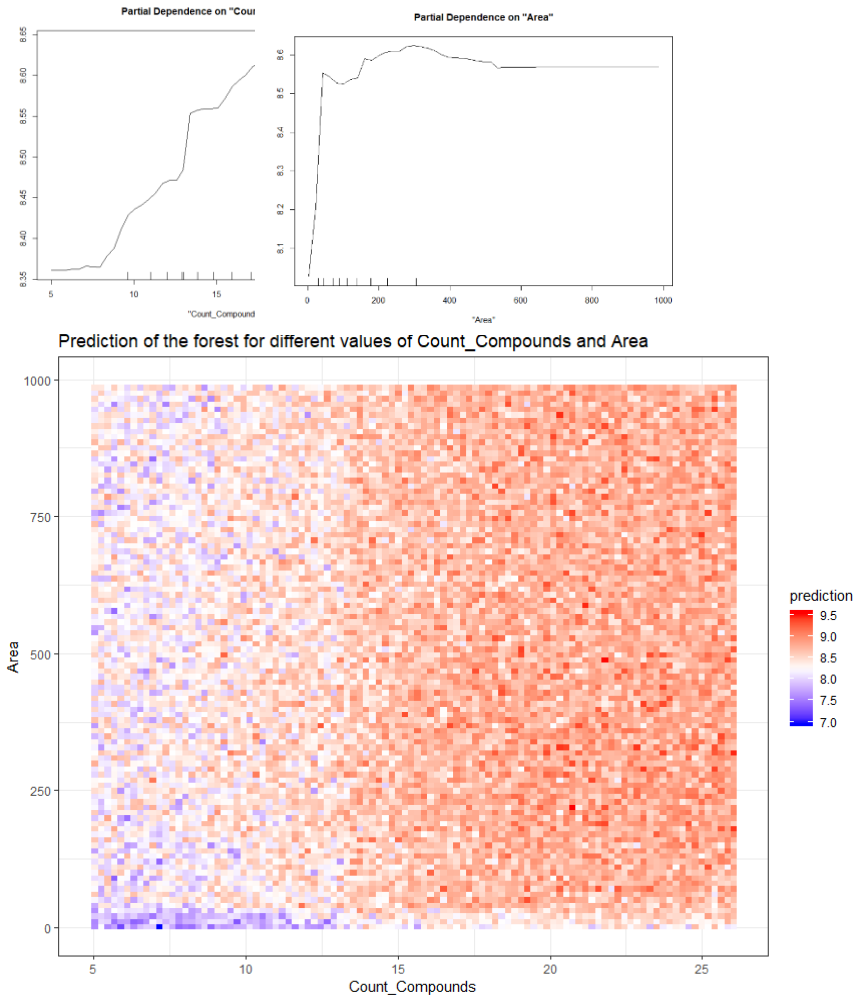


Figure 6 Predicted model effects for selected measures in the random forest of the core feature set as applied to nabim One wheats. Upper left and right; partial dependence plots of the relationship between predicted yield and the diversity of actives and the area of the holding respectively. Lower plot shows the interaction between these measures in terms of predicted yield values

4.1.4. Machine Learning (Expanded feature set)

The application of ML to the expanded feature set attempts to identify parameters that may have been missed in the core feature set which can be used to describe the drivers of yield in nabim One wheat. The importance of parameters based on a random forest of the expanded parameter set (see Dataset construction) are shown in Figure 7. Once again, under the majority of the measures considered, the diversity of compounds (with a similar stepped relationship as in the core feature set) is identified as notably more important than any other measure in the data set, followed by area and primary variety.

The most notable parameter identified as important but not included in the core feature set is the mass of growth regulator applied on a holding; Figure 8. This is only apparent in a subset of the included importance measures but may be important given existing research into the effect of these compounds on yield values. Growth regulators are compounds which actively manipulate the growth patterns within the wheat crop, to generate shorter plants (internodes), thicker stems, thicker stem cell walls and less lodging. Previous research has indicated both positive and negative impacts on crop yield (9, 27) although there is limited evidence associated with trials in multi-input systems (28). There

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

is also data uncertainty within our analysis in relation to this parameter, as there are many “not known” values associated with the application of growth regulators (reflecting decisions in the data assembly process) which may skew the measure effect. Overall, the shape of the relationship between yield and growth regulators appears stepped with a yield increase associated with non-zero inputs followed by a further increase at around 1.5 tonnes per hectare.

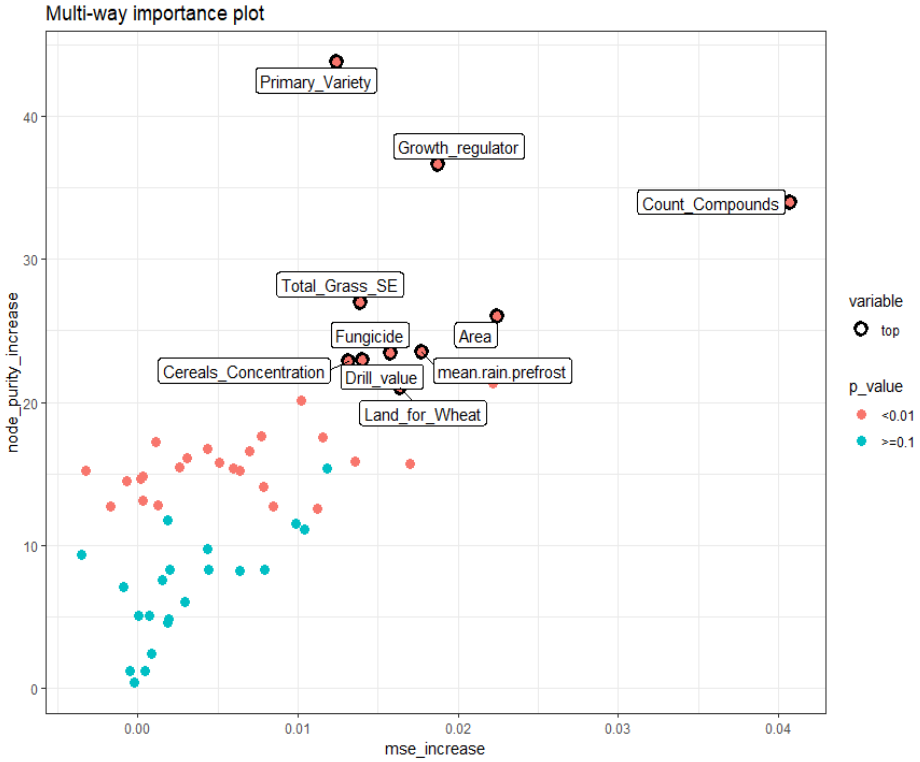


Figure 7 Scores in relation to various measures of importance for variables in the random forest of the expanded feature set as applied to nabim One wheats. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

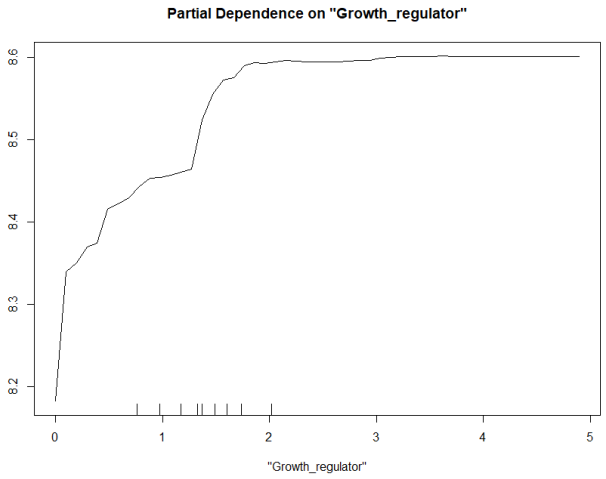


Figure 8 Predicted model effects for selected measures in the random forest of the expanded feature set as applied to nabim One wheats. Partial dependence plots of the relationship between predicted yield and the application of growth regulators

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.2. Feed wheats (nabim Four and unclassified wheat varieties)

A challenge for assessing the potential drivers of yield in feed wheats is that, in contrast with the milling wheats described above, this group does not have a clear delimitation to provide a basis for selecting comparable crops. Many varieties otherwise used as milling wheats will, upon failing to meet quality criteria, be converted to feed, which makes a variety driven classification challenging to definitively interpret. The definition used here is based on combining the set of known feed wheats, listed under group four of the nabim classification with the group of unclassified varieties which are assumed but not demonstrated to be predominately feed wheats (which make up the vast majority of the overall wheat crop). We considered this to be the best possible compromise which serves to maximise the data sample available for model fitting from the limited information available.

4.2.1. Spatial structure assessment

The approach to modelling used for the feed wheats largely mirrors that of the milling wheats outlined above. One important difference between the two groups lies in the empirical evidence of their spatial structure revealed in their respective semi-variograms (Figure 9). Where for the nabim One wheats there is a clear signal of increased similarity between holdings at short distances, this is almost completely absent from the patterns observed in the feed wheats. Rather, there is puzzling 'hump' in the difference between sites separated by a distance of greater than 3 degrees which is difficult to account for in any simple model for spatial structure. This may reflect some larger scale process within our restricted geographic sample (e.g. the distribution of major agricultural regions within the UK). In the absence of a clear function to be used to represent the spatial structure the fitted models for the feed wheats are calculated without any explicit correlation function to represent spatial autocorrelation, with the parameters included in the model being otherwise identical to that outlined for the nabim One wheats (see above).

The data set for feed wheats includes 903 records representing 873 unique CPH numbers. As with the milling wheats the model sample is restricted to a ±50% envelope around the mean (approximately the range 4.31-12.9 tonnes per hectare) to reduce the potential impact of extreme outliers (Figure 10). Note also that compared to the milling wheats the distribution of yield values for feed wheats is more bimodal with a noticeable secondary peak at around 7 tonnes per hectare.

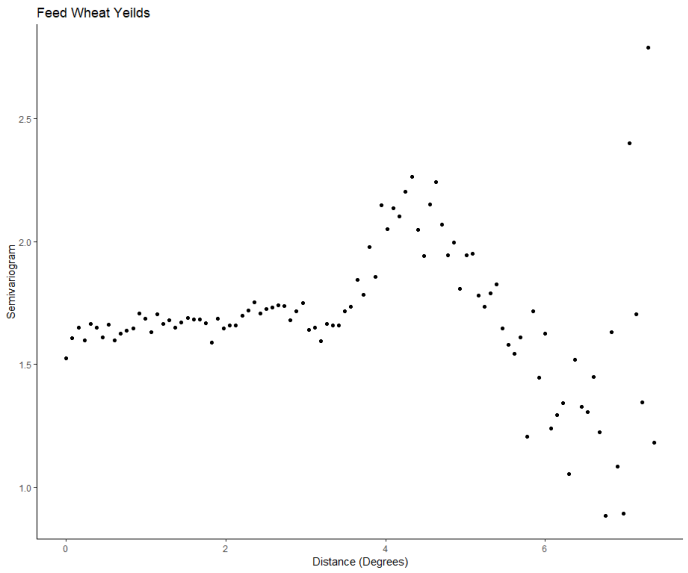


Figure 9 Empirical semi-variogram based on estimated least squares surface of yield in feed wheat varieties, see discussion in text

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

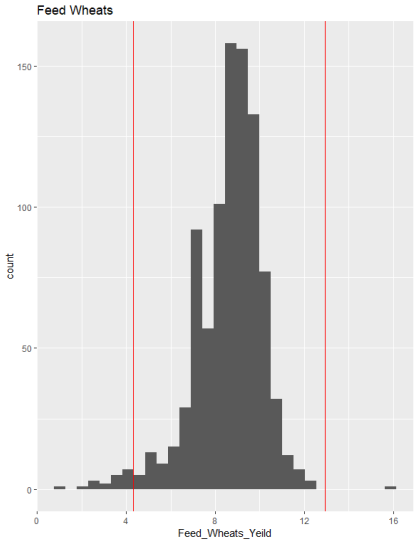


Figure 10 Distribution of yield values for feed wheat varieties, Red lines indicate the envelope included in the presented analyses

Table 2 Summary of the significant parameters identified form the analyses of the core feature set under statistical modelling to the yield values for feed wheat varieties Columns list the estimated parameter, standard error and the results of model testing to compare the relative fit to models where the parameter is removed. Names given reference those in the Dataset Construction section. See text for discussion

Parameter	Estimate	Standard Error on the Estimated Value	AIC of favoured model (df)	AIC of favoured model excluding parameter (df)	Likelihood ratio value of model test Ratio	P value of model comparison test
log(Area)	0.2007	0.0411	2957.1 (10)	2978.8 (9)	23.74	<.0001
Prop_Own_seed	-0.320	0.107		2964.1 (9)	9.061	0.0027
Mean_Spray_Rounds	0.0677	0.033		2959.1 (9)	4.04	0.0455 [Marginal]
Count_Compounds	0.0564	0.021		2962.0 (9)	6.957	0.0087
mean.RH.prefrost	0.0671	0.017		2969.6(9)	14.48	0.0002
mean.rain.prefrost	-0.353	0.072		2978.6 (9)	23.61	<.0001
Soil_PCA_ScorePC2	0.266	0.124		2959.7 (9)	4.67	0.0316[Marginal]
Latitude	0.1057	0.1057		2960.9 (9)	5.83	0.0163

4.2.2. Statistical modelling (Core feature set)

The optimal set of model parameters for feed wheats after model simplification shows notable differences compared with that inferred for the nabim One wheat group. Conspicuously absent from the optimal model is evidence for a significant impact of primary variety (the majority variety on the farm), which may provide evidence for a reduced role of farm management practices in the yield of the feed varieties. Also notable is the lack of evidence for total pesticide load as a predictor of yield. Instead the diversity of compounds and the number of spray rounds are recovered as the only significant predictors associated with agrochemical inputs. We can hypothesise that this may be

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

related to variation in the spray regime between holdings or be evidence of saturation wherein the coverage of applied actives is of greater significance than the total pesticide application. Teasing out the specific effects of spray regime is challenging with the current dataset; more high resolution and standardised approach to data collection (to resolve some of the confounding factors) may be required (see Further Work).

Similarities between the feed and milling wheats results include continued evidence for efficiencies of scale (log Area), reduced yield associated with secondary sowing (Prop_Own_seed) and a weak effect of increasing yield at higher latitudes. As previously, low rainfall in the pre-frost period is recovered as a significant factor in predicting yield values, although many of the same caveats for milling wheats apply here. Novel to this analysis, is evidence for a positive impact of elevated humidity during the pre-frost period. However, it is unclear to what extent these interact and/or reflect a common underlying process. There are marginal impacts of soil type: specifically, the second principle component associated with conditions of impeded drainage. This may be associated with the same phenomenon, given the role of soils as mediating water availability particularly during the pre-frost period.

In common with nabim one wheats, while statistically significant drivers of yield were found, the proportion of variation in yield described by the model is small (r-squared 0.19). This means that the great majority of variation in yields is either driven by factors that are not included in the model or occurs randomly.

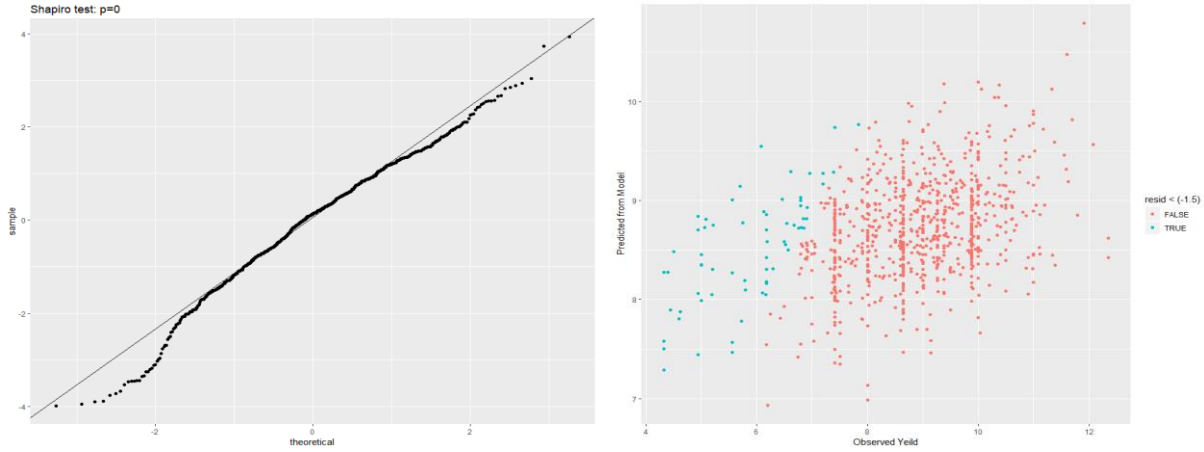


Figure 11 Diagnostic plots for the relative fit of the optimal model identified for feed wheats relative to the assumed error structure for a linear model. Left; the qqnorm plot (described in text; the line represents the ideal assumed error structure and the points represent that observed); Right Observed (x axis) and predicted values from the model (y axis) points associated with large negative residuals are highlighted in blue

When the model diagnostics are considered, the extent of deviations from the assumptions of normality can be observed in the feed wheats, that may be more serious than those previously discussed. Once again there is a population of low yielding sites which do not appear to conform to the assumptions underlying the fitted model and which dominate at the lowest end of the reported yields (Figure 11). Plotting the most divergent site reveals no obvious spatial or temporal pattern to their occurrence[13]. This may provide evidence of social factors in driving non-compliance (although this would require additional datasets to confirm). Exploring these unexplained population level differences is a major potential area of study following on from this initial analysis.

[13] Figure not shown due to confidentiality requirements around the use of PUS and June survey datasets

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.2.3. Machine Learning (Core feature set)

Based on the Random Forest ML applied to the core feature set we observed that, as with the milling wheats, the count of unique active ingredients applied is identified as the single most important factor for the prediction of yield (Figure 12). The number of spray rounds undertaken is also given high importance and other variables most noticeably longitude also feature, having not previously played any major role is determining wheat yields. Primary variety continues to be identified as contributing node purity, but not to MSE reduction (which may align with it not being a significant factor in statistical modelling). Area of holding continues to be a relevant measure, although noticeably less so than was observed previously, particularly given the increase relevance placed on the number of spray rounds undertaken.

The shape of the relationship with the diversity of actives is broadly similar to that observed in the milling wheats, although the step change at 12 active compounds is perhaps more pronounced (Figure 13). The shape the response to the number of spraying rounds is largely reflective of a continuous trend of increasing yields over the interval between 5 and 10 spray rounds (containing the vast majority of the recorded data) with levelling off at the extremes of the distribution. There is limited evidence for structured interactions between these two measures. However, below around 7 rounds the impact of low compound diversity appears particularly pronounced.

As with the milling wheats, there is evidence for lack of generality /overfitting in the application of the Random forest, with R-squared with respect to the training (0.842) and testing (0.198) datasets again differing widely, which may indicate issues in the generality of the fitted function.

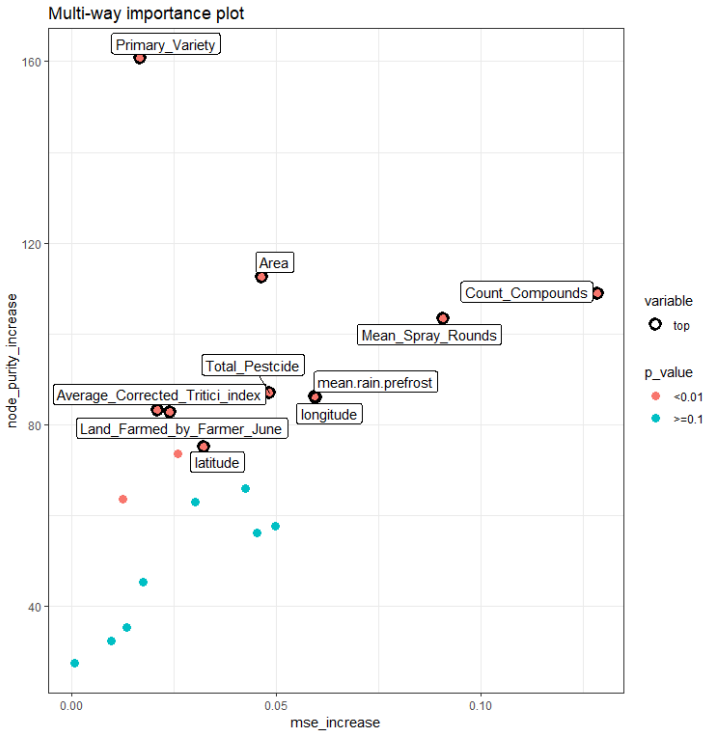


Figure 12 Scores in relation to various measures of importance for variables in the random forest of the core feature set as applied to feed wheats. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

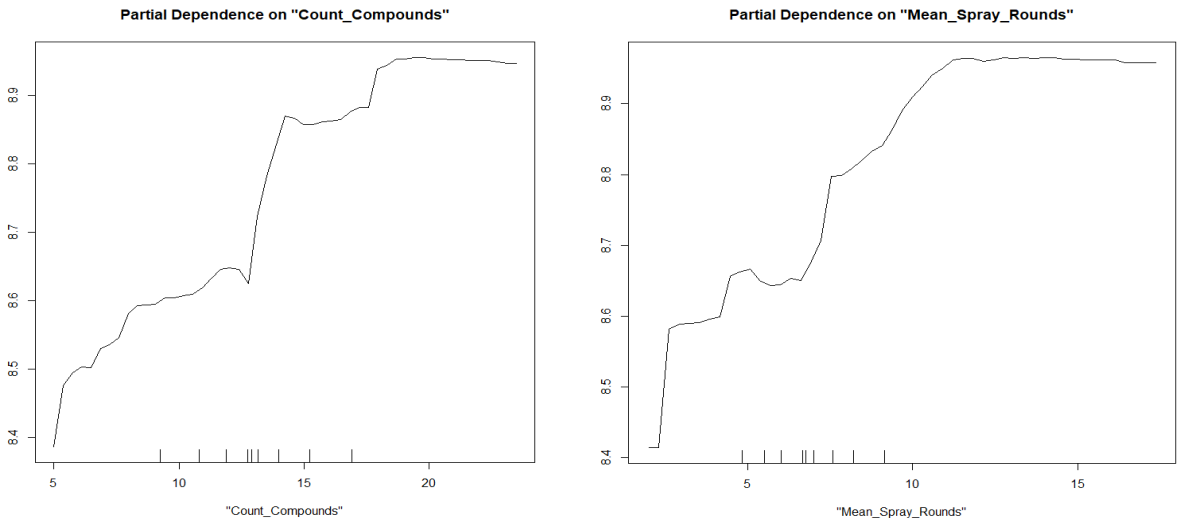


Figure 13 Predicted model effects for selected measures in the random forest of the core feature set as applied to feed wheats. Left and right; partial dependence plots of the relationship between predicted yield and the diversity of actives and the number of spray rounds respectively.

4.2.4. Machine Learning (Expanded feature set)

The expanded feature set reveals several additional measures potentially of interest beyond the core feature set. As in milling wheats, growth regulators are identified as a factor affecting yield values, with a stepwise structure in the associated response curve: yields increasing from around 1.2 kg per hectare. Also of interest is a negative association with the proportion of set aside grassland (Total_Grass_SE; expressed as a proportion of the total area farmed by the farmer see above), which may be evidence for higher yields being associated with more intensive agricultural practice. This is of particular interest given the association with the agricultural payment schemes and farmer incentives which may provide tools to shape yields and behaviour across holdings. More in depth analysis of the potential impacts of set aside on wheat yields require information not currently included in the study sample but which could in principal be obtained in follow on work.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

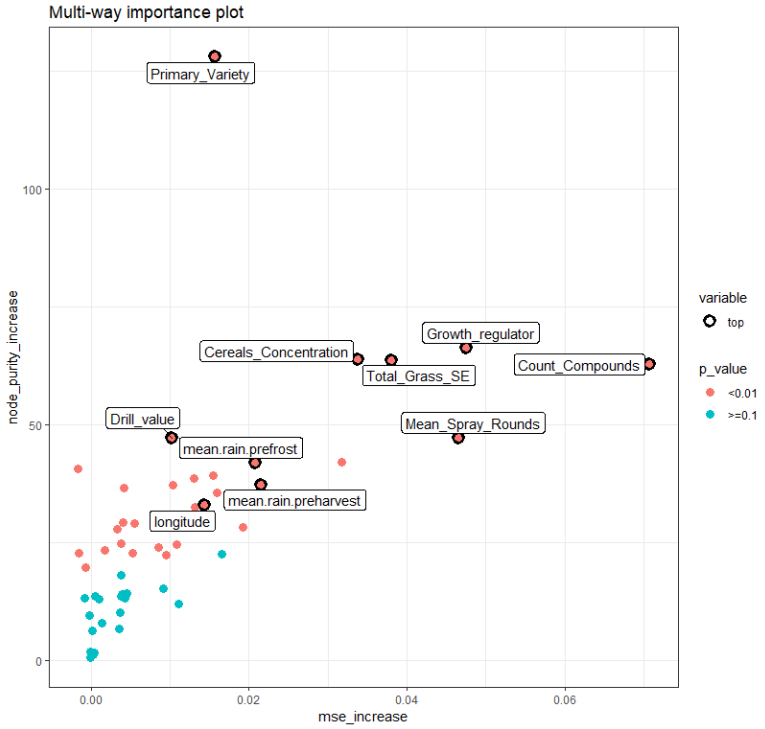


Figure 14 Scores in relation to various measures of importance for variables in the random forest of the expanded feature set as applied to feed wheats. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

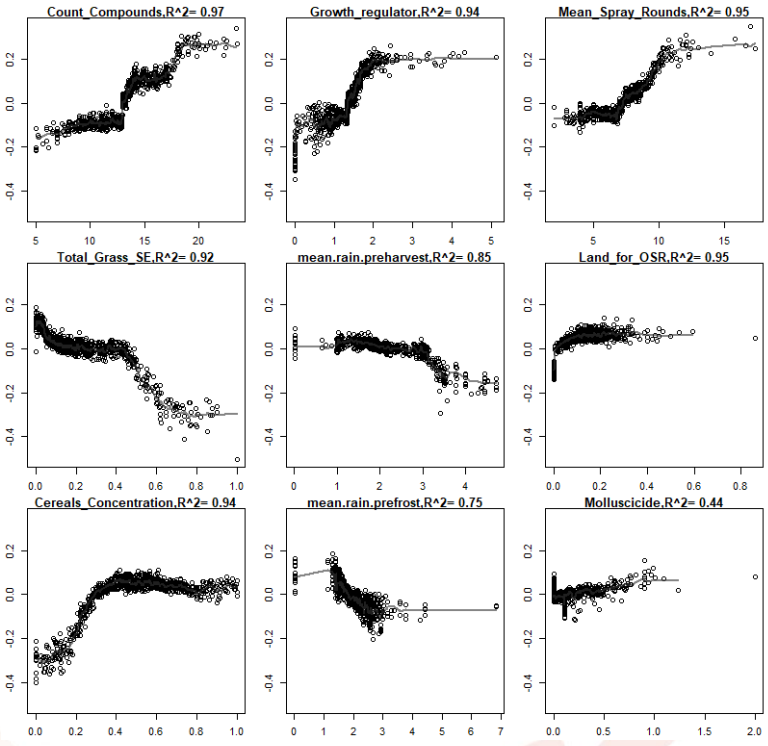


Figure 15 Partial dependence curves of yield in feed wheats with various important measures from the expanded feature set. Plots are based on the cross validated feature contribution concept from the forestFloor R package (29) and are shown in decreasing order of importance (mean squared error on the prediction of the OOB error) from the upper left.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.3. Oilseed rape

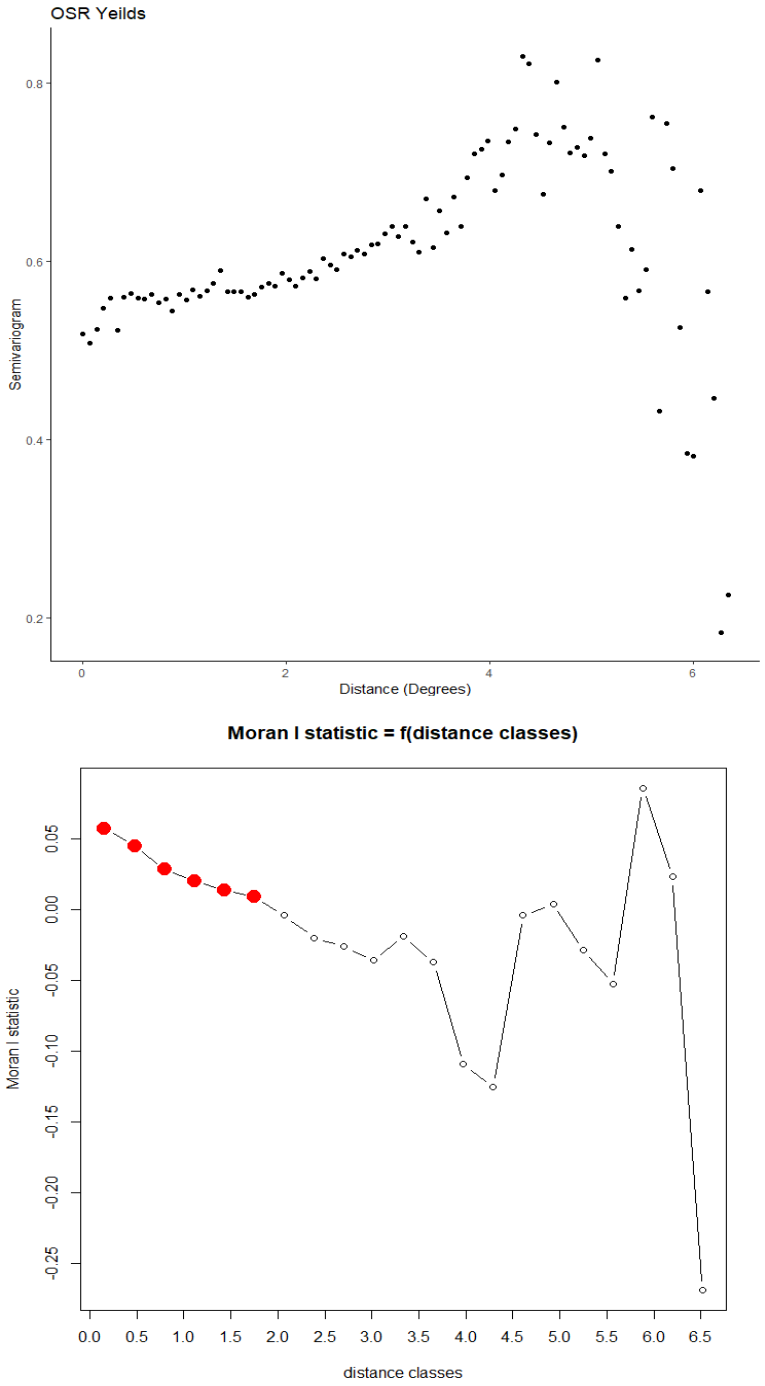


Figure 16 Visualisation of the empirical spatial structure associated with yields of oilseed rape. Upper: empirical semi-variogram based on estimated least squares surface; Lower: results of simulation-based testing of the value of Moran's I (an index of spatial autocorrelation) and varying distance, points shown in blue represent greater than expected spatial autocorrelation within the sample. Distances are calculated based on latitude and longitude coordinates using linear approximation. See text for discussion.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.3.1. Spatial structure assessment

The empirical spatial structure associated with oilseed rape (OSR) shows a trend of increasing variation between holdings of increasing distance which is quite distinct from that of either of the two wheats (Figure 16). Unlike the nabim One wheats there is no rapid drop off in similarity associated with closely adjacent sites. Rather, greater than expected similarity (based on the Moran I statistic) is recovered continuously over the range 0 to 1.5-degree distance and qualitatively up to 4 degrees distance. Above around 3.5-degree separation the similarity becomes largely unstructured, with the apparent hump likely being driven by change variation in a small number of pairs of holdings. The observed distribution suggests a more gradual correlation function than the exponential model used above. Hence, we have applied a normally distributed correlation structure to our statistical model to represent a somewhat gradual decline in similarity between sites at increasing separation (mathematically, if holdings are a distance r apart, their expected correlation due to distance (r) on a variable with range d is $\exp(-(r/d)^2)$.

The OSR dataset included 1034 records representing 981 unique CPH values. In contrast with the wheat data the number of outliers in the OSR yields were relatively small (Figure 17). Hence we have elected to not restrict the analysis window for this dataset. The core feature set (prior to simplification) used to model OSR is as follows:

OSR_Yeild~
- log(Area)
- Prop_Own_seed,
- Primary_Variety
- Mean_Spray_Rounds*Count_Compounds*Total_Pestcide [Indicating an interaction between these parameters],
- log(Land_Farmed_by_Farmer_June),
- mean.wind.preharvest
- mean.RH.prefrost,
- mean.temp.prefrost
- mean.rain.prefrost,
- Soil_PCA_OSR_ScorePC1,
- Soil_PCA_OSR_ScorePC2,
- Soil_PCA_OSR_ScorePC3,
- Average_Corrected_lls_index,
- Average_pod_alternaria,
- Average_stems_with_alternaria,
- Average_stems_sclerotinia,
- Year (as multilevel factor)
- latitude+ [Normal spatial auto correlation]

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

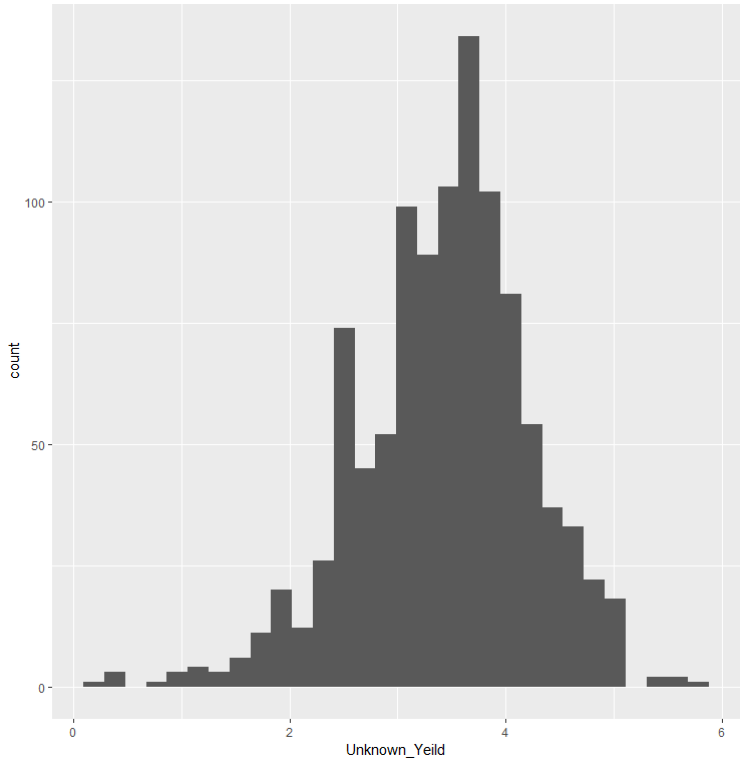


Figure 17 Distribution of yield values for oilseed rape

Table 3 Summary of the significant parameters identified form the analyses of the core feature set under statistical modelling to the yield values for oilseed rape. Columns list the estimated parameter, standard error and the results of model testing to compare the relative fit to models where the parameter is removed. Names given reference those in the Dataset Construction section. See text for discussion

Parameter	Estimate	Standard Error on the Estimated Value	AIC of favoured model (df)	AIC of favoured model excluding parameter (df)	Likelihood ratio value of model test Ratio	P value of model comparison test
Prop_Own_seed	-0.1383	0.0617	2302.4 (26)	2305.5 (25)	5.127	0.023
Primary Variety	Figure 19			2369.14 (14)	21.54	0.0429 [Marginal]
Year	Figure 18			2463.33 (21)	67.14	<.0001
Soil_PCA_OSR_ScorePC3	0.2828	0.1072		2307.5 (25)	7.09	0.007
Latitude	0.1678	0.0232		2352.3 (25)	51.97	<.0001
Total Pesticide	-0.0474	0.0194		2306.4 (25)	6.052	0.013
Count Compounds	0.066	0.0106		2994.4 (25)	693.99	<.0001

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.3.2. Statistical modelling (Core feature set)

The optimal parameter set identified under stepwise simplification for OSR yields are shown in Table 3. In marked contrast with wheats the most important parameters include the year of sample, indicative of pronounced a trend in increasing yields observed through the dataset (consistent with previous observations for OSR yields during this period (2)). There is some evidence from the parameter estimates for a degree of periodicity in OSR yields (with yields in 2004 and 2008 being noticeably lower than for the surrounding intervals). This may reflect elements of cyclic land usage (i.e. between OSR and wheat) across the sampled holdings although the interval under study is too restricted for this to confirmed. Variation associated with the primary variety was detected, although with marginal significance. The impact of the major crop on a holding appears to be small, with large uncertainties on the estimates of the parameters, although there may be some difference between (relatively) low yielding sites where 'expert' is the major variety when compared with higher yields associated with 'castille ' and 'escort' (Figure 19).

Of the numeric predictors identified the most important effects are associated with latitude and the diversity of applied active compounds. The wheat data had shown a small and statistically marginal impact of increasing latitude on increased yields, however, for OSR, this effect is larger and more robustly supported by the model. Note that this is in addition to the explicit spatial structure applied to the modelling. Hence high latitude sites are associated with increased yields above the level which might be expected given the implied spatial structure. Potential causes for this trend include the impact of day length or other climatic factors strongly correlated with latitude, or potentially the distribution limits of key pest species.

The trend of increased yield with increased diversity of applied actives is one observed throughout this analysis and is perhaps the single most well supported conclusion of the applied modelling. Interestingly for OSR there is evidence of a negative impact of total mass density of pesticide use. The causality is unclear: it may be that increased pesticide loads are associated with holdings where pest outbreaks were observed.

Of the remaining measures, second sowing of OSR is relatively rare, and hence Prop_Own_seed likely reflects differences in agronomy practice and the viability of stored seed, as opposed to the repeated sowing observed in wheat. In contrast with wheats, we also identify a small effect of soil types on OSR yields, in particular the distinction between naturally wet and loamy soils captured in the third principal component of the calculated PCA. Given the relatively small proportion of variance that this component represents and the failure to associate significance with the other soil measures the relevance of this factor for OSR yields is unclear and may warrant a more detailed investigation.

In common with the wheats, while statistically significant drivers of yield were found, the proportion of variation in yield described by the model is small (r-squared 0.15). This means that the great majority of variation in yields is either driven by factors that are not included in the model or occurs at random.

Examination of the model's error structure again reveals that particularly at the low end of yield values there are deviations from the assumed error structure (Figure 20). When compared with the wheats the distribution of poorly predicted sites is less clustered at the extreme low end of the distribution (although all sites where yields less than 1 tonne per hectare were reported are included) which may indicate that this issue in term of model fit is a more dispersed one that is less clearly associated with a discrete sub population of sites.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

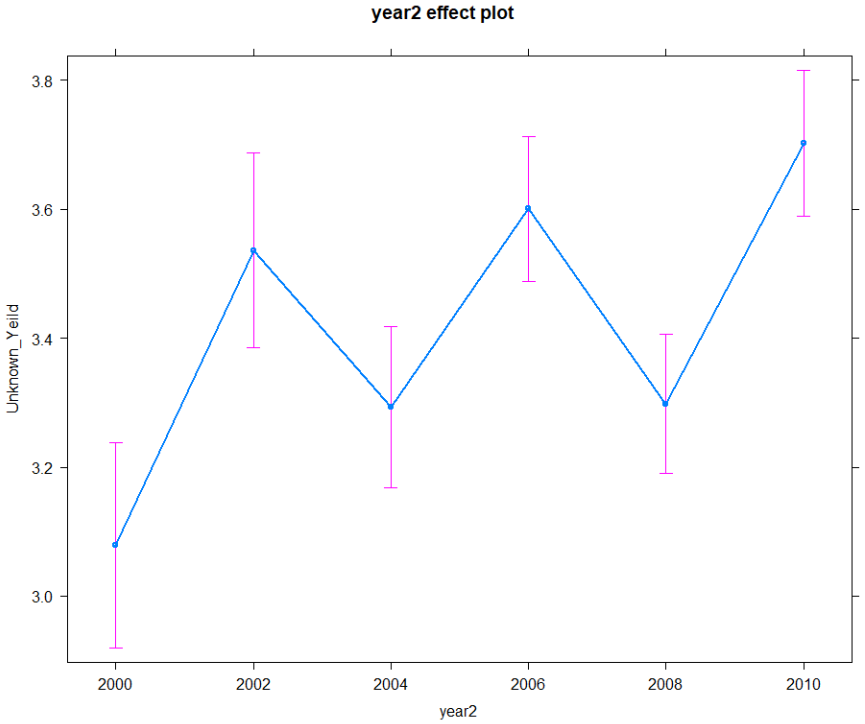


Figure 18 Estimated effects (model estimate and confidence interval) for year in the optimal model for oilseed rape.

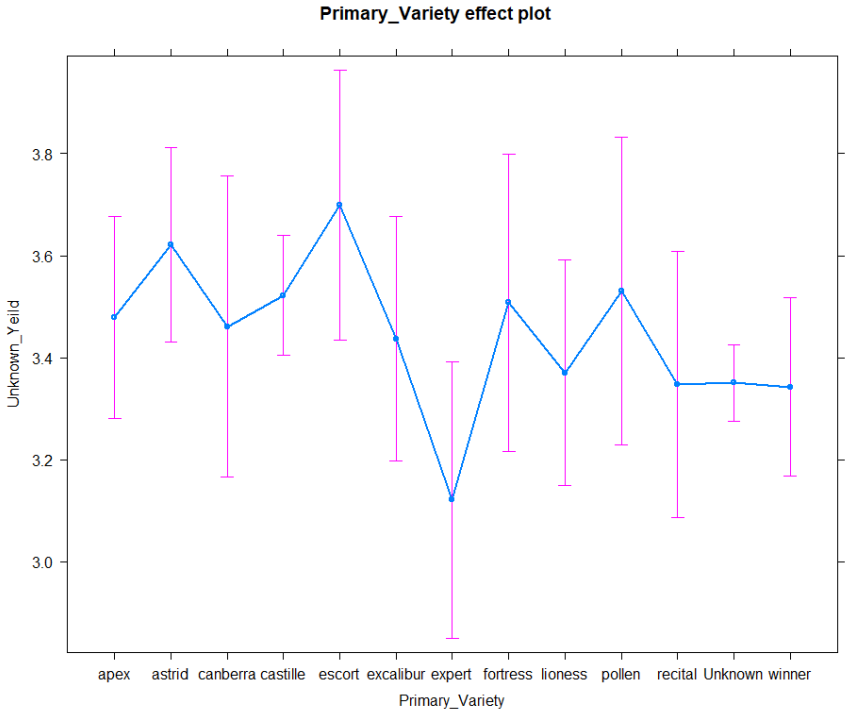


Figure 19 Estimated effects (model estimate and confidence interval) for primary variety in the optimal model for oilseed rape

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

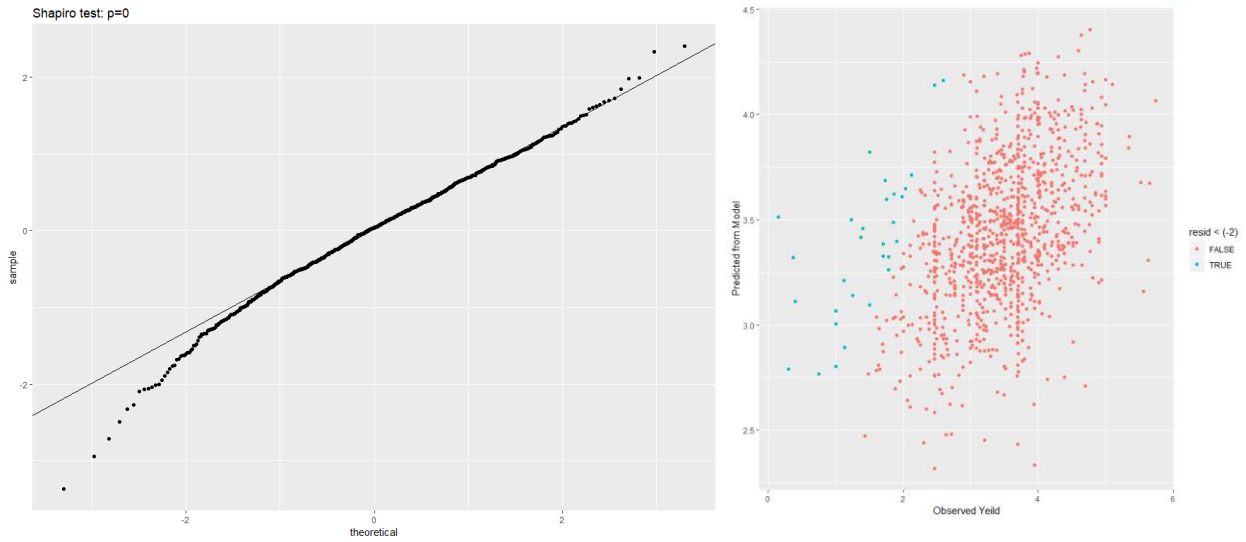


Figure 20 Diagnostic plots for the relative fit of the optimal model identified for OSR relative to the assumed error structure for a linear model. Left; the qqnorm plot (described in text; the line represents the ideal assumed error structure and the points represent that observed); Right Observed (x axis) and predicted values from the model (y axis) points associated with large negative residuals are highlighted in blue

4.3.3. Machine Learning (Core feature set)

The application of random forest machine learning to the core feature set for OSR largely reinforces effect of latitude as the most important parameter. (Figure 21). Other important parameters include primary variety, year and the diversity of compounds although in general each of these is only identified as of major importance on a subset of calculated measures.

The shape of the relationship between yield and latitude shows a stepped function wherein the boundaries between 52 degrees north (approximately the latitude of Ipswich) and just south of 54 degree north (approximately Leeds) are associated with transitions in the estimated effect on yield (Figure 22). Precisely what causes these transitions is unclear, it may be that they represent the distribution limits of some pest organism or pathogen or that they reflect areas of common agricultural practice. There is evidence that the southerly and more influential step, appears to be common across different years and primary varieties (Figure 23) although with some variation) which may indicate that an administrative/behavioural cause is more likely (as a biological effect would be expected to show inter-year variation). Alternatively, the crop disease survey identifies outbreaks of *Alternaria* ('pod spot') as having a northerly limit approximating the observed transition in yield, although in this case it is unclear why the measures of disease prevalence do not feature more prominently in the model. Likewise, another major economic pest of OSR, cabbage stem flea beetle (*Psylliodes chrysocephala L.*) also has a southerly distribution within the UK and could play a major role in influencing patterns of yield, although testing this would lie outside of the scope to the reported analyses[14].

In contrast with wheats, the shape of the relationship between yield and the diversity of active ingredients applied does not show a sharp transition; instead it shows a general increase in yield response levelling out at around 13 compounds applied.

[14] Data on the occurrence of cabbage stem flea beetle was not available within the studied datasets. FERA has access to datasets relating to the occurrence of this pest which could be investigated for use in further work.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

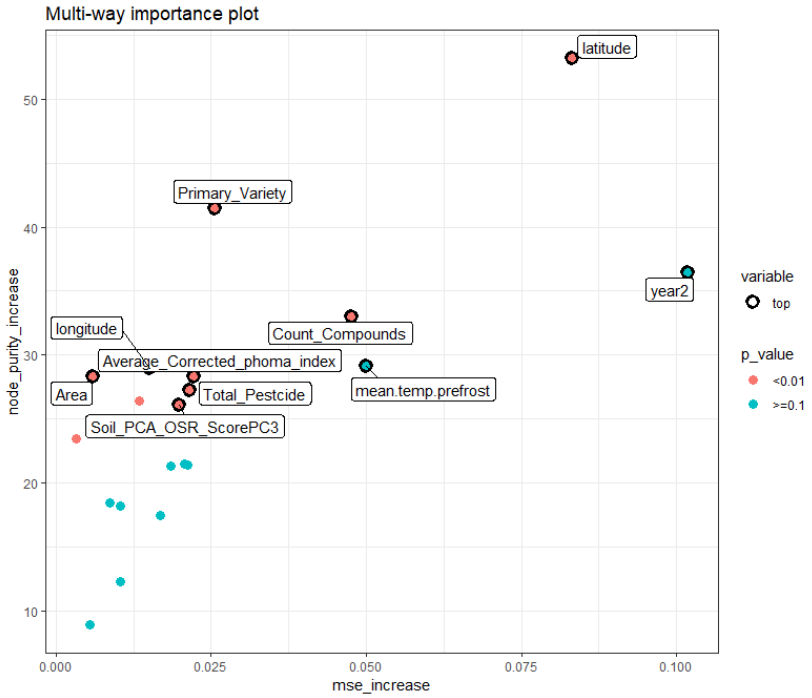


Figure 21 Scores in relation to various measures of importance for variables in the random forest of the core feature set as applied to OSR. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

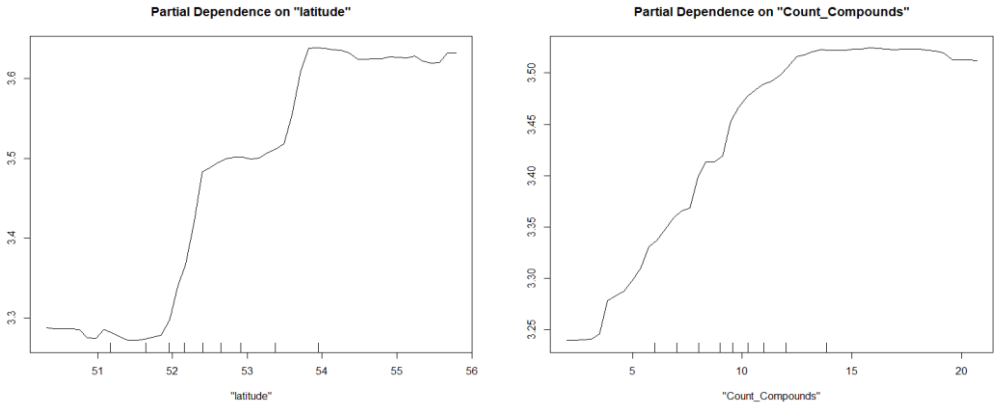


Figure 22 Left and right; partial dependence plots of the relationship between predicted yield and latitude and the diversity of applied compounds respectively in the random forest of the core feature set as applied to OSR.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

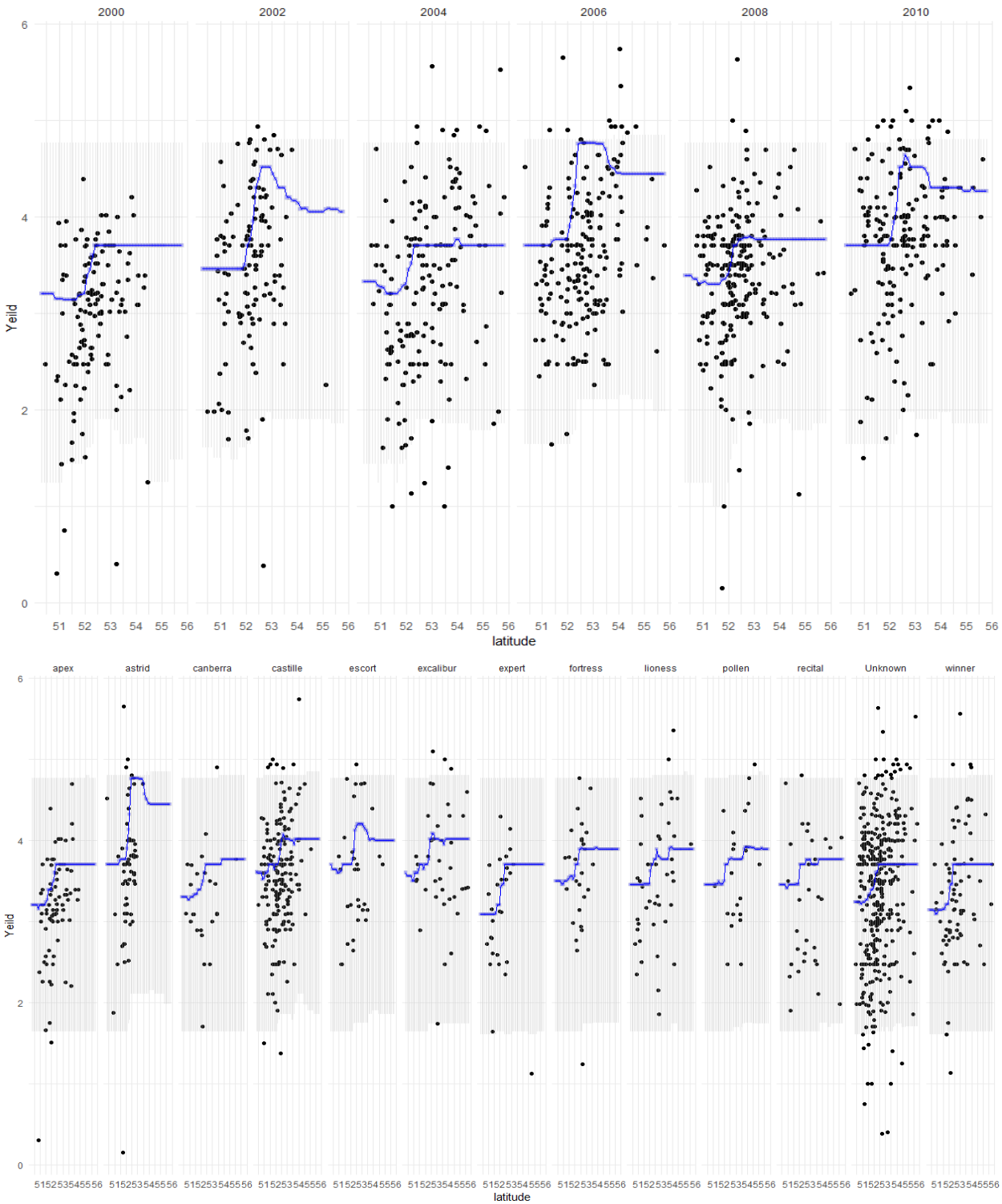


Figure 23 Predicted function for the relationship between yield and latitude from the random forest of the core feature set applied to OSR. Upper by Year, Lower by Primary variety. Observed values are shown as points.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

4.3.4. Machine Learning (Expanded feature set)

Random Forest of the expanded feature set with respect to OSR yield reveals that in addition to latitude the most important measure in predicting yield values is the volume of Fungicide[15] applied (Figure 24). As shown by the disease surveyed under the crop disease survey, OSR is subject to a wide range of economically important fungal diseases and it may be that control of these diseases is the major driver behind the relative high importance assigned to Fungicide usage. It is notable that in the context of the expanded dataset the diversity of compounds applied has a relatively reduced importance, strongly suggesting that it is fungicides, and possibly the diversity thereof, that are the key component associated with systematic yield change in OSR. In terms of the shape of the relationship, the volume of fungicide application shows a clear tendency towards diminishing returns above around 0.8 kilograms per hectare and interactions with latitude are minimal; the expectation is that very low fungicide usage has a severe impact on yield regardless of position within the latitudinal gradient (Figure 25)

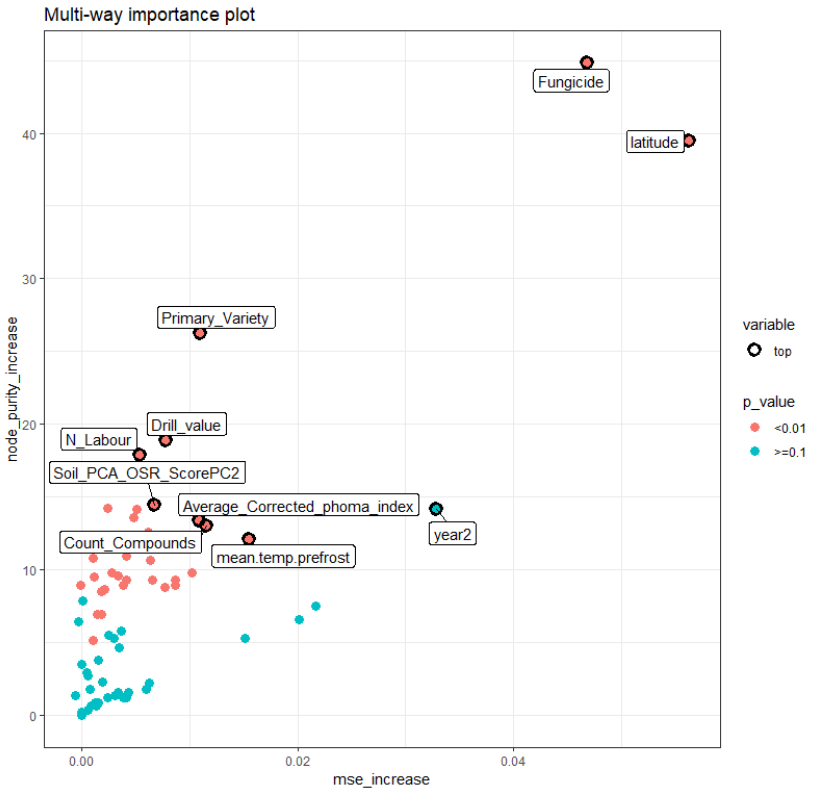


Figure 24 Scores in relation to various measures of importance for variables in the random forest of the expanded feature set as applied to OSR. X-axis; the increase in the estimated measure square error on the predicted values associated with the loss of a parameter. Y-axis; the increase in node purity (relative similarity of points clustered in the underlying decision tree); Colour; a binomial test for if the variable is used to subdivide data in the underlying decision tress more then would be expected by chance. For clarity only the top ten most important measures are named.

[15] Note that some common fungicidal compounds used with OSR, such as conazoles, are sometimes also used and marketed as growth regulators.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

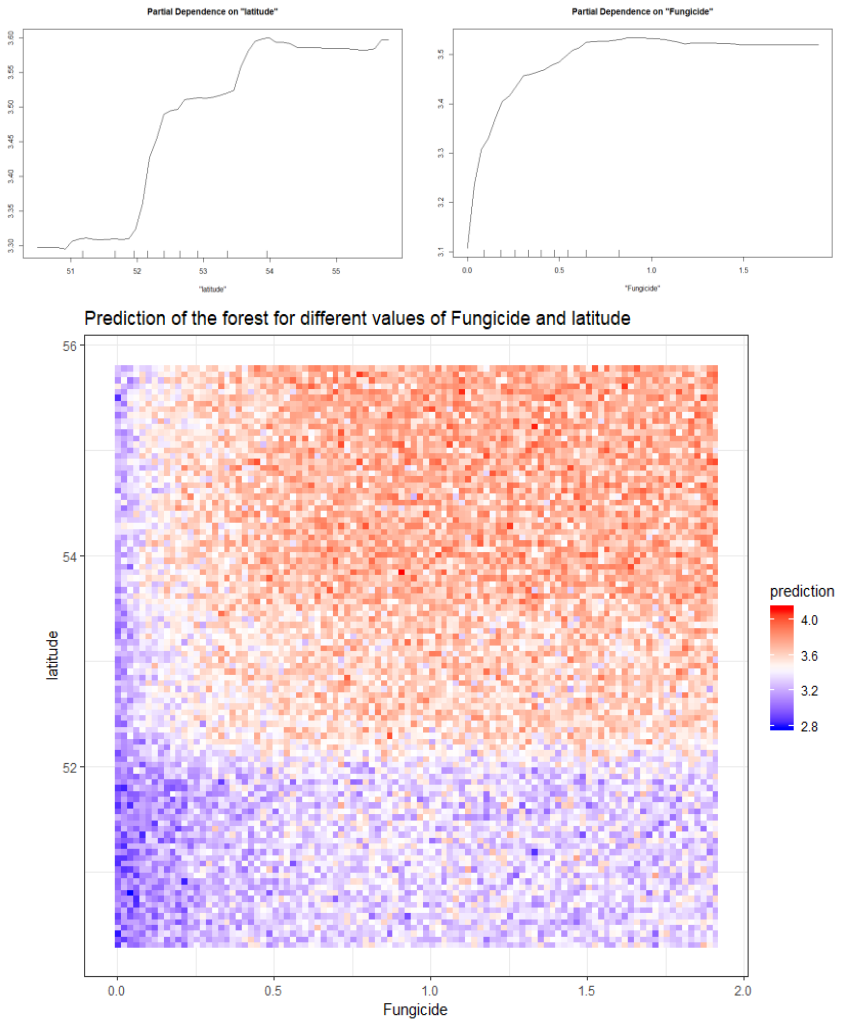


Figure 25 Predicted model effects for selected measures in the random forest of the expanded feature set as applied to OSR Upper left and right; partial dependence plots of the relationship between predicted yield and latitude and application of fungicide respectively. Lower plot shows the interaction between these measures in terms of predicted yield values

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

5. Conclusions and Discussion

The focus of the modelling conducted here has been to take a large number of potential drivers of yield and to use statistical and machine learning approaches to describe the subset that display a significant correlation with observed yields. This should be interpreted to be primarily an exercise in hypothesis generation based on the identification of areas for further research. This work is one of a minority of studies into crop yields based on data from real holdings and combines data from many different sources originally collected for different purposes and spatial scales. As a result, there is a large amount of noise and other confounding effects in the data collection process which restrict our ability to clearly distinguish the processes driving yield. Confirmatory investigation, e.g. based on more standardised and focused data collection, is thus advised in exploring the features identified here and in understanding their significance within the wider UK farming infrastructure and policy environment.

Changes in yield associated with the factors examined in this study were generally small compared with random variation in yield observed between holdings. The most consistent effect identified across the three crops considered was evidence that the diversity (as opposed to the total mass) of pesticides is one of the key factors associated with change in yield in the three systems. We estimated that this may lead to an expected change in yield of approximately 1T/Ha when comparing the lowest to highest diversity (see Appendix 2: Visualising the shape of numeric parameter effects in statistical modelling). That diverse agrochemical profiles, particularly in relation to pesticides, should be associated with higher yields is not in itself surprising, because it indicates that a crop has been protected against a range of potential threats, including both a wide range of pathogens and, potentially, issues around resistance. The stepped profile associated with the wheat yields is somewhat more surprising. The implication is that there is an optimal spray profile (including over 12 compounds) above which further increases in expected yields are not seen. What is less clear from the current analysis whether this effect is related to the diversity of compounds per se or if the effect can be attributed to the inclusion within spray programmes of particular compounds or combinations of compounds which are only included in high diversity regimes. This is a potentially interesting area of further research but one which lies outside if the scope of the current analysis and requires a redefinition of how PUS dataset is represented in order to fully address.

The observed importance of seasonal weather conditions aligns well with the perceptions of farmers (1). However, due to a number of data issues outlined above, it is unclear whether differences between nabim One wheats (where only rainfall during the pre-frost periods is identified) vs feed wheats (where the humidity is also identified as a key driver) are driven by differences between varieties. Ideally this would feed into discussion on soil moisture levels, particularly given how soil type may interact with water retention and the economics of irrigation (30). However, it is unclear if there are adequate proxies in the data currently available or if this would require expansion of the source datasets.

Also, in wheat, evidence for economies of scale, with larger holdings generating disproportionately high yield values has implications for the structuring of Agri-payment systems particularly given the potential for EU exit. The recorded negative impact of the proportion of land 'set aside' as grassland within the feed wheats suggests that there as might be trade-offs to consider in terms of the goals and targeting of payment schemes. It could be interesting to consider how information on biodiversity and other externalities could be incorporated into the dataset, as well as more direct consideration of payments received, as previous work has suggested some association between Common Agricultural Policy (CAP) payments and yield (1) which is a leading candidate for explaining the subset of low yielding localities.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

With respect to OSR the effect of the latitudinal gradient on yield and its stepwise profile was unexpected, and its causes are not fully understood by the investigator. Some very recent work has indicated that low temperatures particular in December are associated with elevated yield in OSR(7) but it not immediately apparent how this translates to a latitudinal effect . Of the major pests monitored as part of the crop disease survey *Alternaria* ('pod spot') has a southerly distribution approximately matching the stepped profile observed, while incidence of Light Leaf Spot may account for the apparent levelling off observed at the highest latitudes within England and Wales. Why these disease profiles would manifest as a latitudinal gradient as opposed to the included crop disease measures is unclear and may suggest that there are issues in the calculated index used here to indicate regional prevalence. As noted previously, cabbage stem flea beetle is another potential candidate for driving latitudinal gradients in yield but one for which we currently lack currently lack adequate characterisation within the analysed dataset.

The two measures singled out for further investigation are growth regulators in the context of wheats and fungicide usage in the context of OSR. Evidence for increase yield with the use of growth regulators is surprising as it conflicts with published findings (31) and may be associated with data issues in within the presented analysis. The importance of fungicide usage in OSR supports the idea that the latitudinal gradient may be associated with pest distributions and may have interesting interactions with the significance of pesticide diversity within this crop.

In interpreting this work it is important to recognise that there has been no attempt (beyond the classification of varieties) to describe the ***quality*** of the yields obtained from the sampled holdings. No quality information is available in the PUS dataset and any discussion would necessitate expanding the collection of datasets under study. Also not discussed in detail are the potential economic drivers, especially the differences between economically viable yield and failed crops which may vary from location to location and have an impact on, for example decisions around pesticide usage. In this analysis we have minimised the role of the economic data from the John Nix handbook as, being a set of simple yearly averages, lacked any sort of spatial resolution to distinguish between holdings. As a result, these measures proved unimportant in the fitted models. Were more high-resolution data around economic conditions and decision making to be available this could have a major impact on the structure of the analysis and the conclusions presented.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

6. Further Work

6.1. Combination and representation of variables of interest

The core task outlined in this project was to provide insights into the patterns of attained yield based on a subset of complied data regarding UK agriculture. The datasets used in the study are largely selected for ease of availability and do not necessarily reflect current theoretical understanding of crop yields based on standardised plots. Inevitably there are gaps in the coverage as well as outstanding questions regarding how some of the information is used, that remain to be addressed in follow on work. Some of these relate to representation and the methodology used for analysis but there are also questions around the definition of the questions under study and how they tie into policy objectives around crop yield.

The present analysis is but one of many possible representations of the information available in the included dataset and these are by no means comprehensive in terms of what is known regarding the UK agricultural system. Interesting data gaps include knowledge of farm practice, e.g. in relation to land use payment schemes such as CAP or Environmental Land Management Schemes, as well as many of the social-economic aspects and measures of farmer behaviour (for example the profit point associated with the yields of focal crops). Some of these can be obtained implicitly by considering proxies from existing data (e.g. using the proportion of own seed as a proxy for secondarily sown crops) but others require information that has not been complied within this study. Of interest to the results represented here are any measures which based on previous analyses of the system might be associated with defining the correlates of the low yield population identified in the statistical modelling.

Another area of potential development is in the representation of some of the parameters already included in the study. Currently there are several measures, most obviously those arising from the crop disease survey and met office, the inclusion of which have necessitated specific decisions about how the measures are calculated and how they are associated with the localities under study. It may be that revisions to these decisions e.g. based on a more detailed understanding of the relevant climatic conditions associated with the focal crops may result in features which better encompass the key drivers and lead to greater insights into the distribution of yield on the landscape. Related to this is the observation that, particularly for the weather data the measures under study were developed with information around drivers in wheat and may have limited relevance when applied to OSR. The decisions made here and outlined in the data are largely either driven by a previous round of analysis or represent the lead authors best understanding of the farming data landscape and may be revisited when on building on the foundations provided here.

An obvious data extension to be considered is the expanding the data sample beyond the period 2000-2010. Extending the temporal window makes any modelling approach more powerful by reducing the impact of within sample noise, and potentially opening new possibilities in time series and other related analyses. The current restrictions are largely around access and association of the PUS dataset, the structure of which was revised after the studied period, however in principal these could be overcome, and an expanded set made available (with caveats regarding coverage and comparability to the reported period).

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

6.2. Validating and expanding on the effects observed in this study

This study represents one of a small number of examples that attempt to understand landscape level yield values from the reported yields of real holdings as opposed to standardised plots. This has advantages in terms of the relevance of findings to the agricultural industry, and disadvantages in the introduction of a large set of confounding variables, data uncertainty and other sources of noise. Because of these intrinsic issues in data representation there is significant value in understanding the extent to which identified effects on yield are replicated under the more controlled conditions of standardised plots. The key findings described here; in relation to the impact of the diversity of applied active compounds and the local climatic conditions (with respect to wheats) and/or the apparent latitudinal gradient (with respect to OSR) could be replicated at least in part by existing monitoring schemes (e.g. that run annually by the AHDB to generate the set of recommended nabim varieties). Integration of information from such trials provides an important validation to the findings discussed here and may also help to shed light on differences in the drivers of yield between the highly controlled field trails and less structured real holdings, particularly in relation to economic decisions such as pesticide application.

One of the main challenges associated with the statistical interpretation of the fitted models is the evidence for a non-conforming population clustered at the low end of the yield distribution. This implies a failure of the fitted models to correctly represent the error structure of the dataset and may be evidence for heterogeneity in the processes generating the yield values. Given that spatial correlation does not appear to be the driver (and there are no clear correlations with any of our included measures) we are faced with a choice between excluding the low yielding sites (which represents a major loss of information, particularly given that our focus is understanding the causes of low yield), or attempting to fit a more complex statistical model which accounts for different processes at the extremes of the distribution. One possible framework, although little explored in the context of modelling yields, is the idea of censored or truncated regression (sometimes also called hurdle models, see (32)) wherein a secondary model process is fitted based on whether or not values exceed some predefined threshold (e.g. the average yield required for profitability in a given year) and measures can be represented by their effect on either the probability of exceeding the threshold or the impacts on the yield once the threshold is overcome. Fitting such an approach is complex and experimental and thus considered outside of the scope of the current work but it may represent a valuable potential avenue in further understanding the impacts on yield across the UK landscape.

6.3. Applications of machine learning to agronomic systems, method selection and interpretation

Above we discuss in detail the differences between statistical modelling and common applications of machine learning in terms of how they reflect different components of numerical analysis. To summarise classical statistics makes restrictive assumptions regarding the relationship between the model object and the data (particularly relating to the error structure) but because of this is better able to reflect uncertainty and incomplete information about the underlying population. Likewise machine learning, at least as commonly applied, is much more flexible in terms of fitting the form of the training dataset with respect to the modelled endpoint but consequently can be much more challenging to extract knowledge of the system, particularly where the models are fitted as a 'black box'. Moving forward we need to recognise that these different strengths imply differences in the use case and sorts of insights which can be generated. There are a number of other areas within agronomy data where the techniques and approaches outlined here may serve to provide useful insights for

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

policy (particularly in relation to composition and impacts of different spray regimes) but the key remains a clear vision of the study outcomes, strengths of the methods used, and how this ties into any proposed intervention or policy change

References

1. Vigani M, Cerezo ER, Barbero MG. The determinants of wheat yields: the role of sustainable innovation, policies and risks in France and Hungary. Joint Research Centre, the European Commission. 2015(EUR 27246 EN).
2. Knight S, Kightley S, Bingham I, Hoad S, Lang B, Philpott H, et al. Desk study to evaluate contributory causes of the current yield plateau in wheat and oilseed rape. Crop Improvement & AgronomyCentral Faculty ManagementSoils & Systems; 2012.
3. Petersen J, Haastrup M, Knudsen L, Olesen JE. Causes of yield stagnation in winter wheat in denmark. Faculty of Agricultural Sciences, Aarhus University; 2010.
4. Peltonen-Sainio P, Jauhiainen L, Trnka M, Olesen JE, Calanca P, Eckersten H, et al. Coincidence of variation in yield and climate in Europe. Agriculture, Ecosystems & Environment. 2010;139(4):483-9.
5. Cho K, Falloon P, Gornall J, Betts R, Clark R. Winter wheat yields in the UK: Uncertainties in climate and management impacts Climate Research. 2012;54(08):49-68.
6. Semenov M. Impacts of climate change on wheat in England and Wales. Journal of The Royal Society Interface. 2008;6(33):343-50.
7. Brown JKM, Beeby R, Penfield S. Yield instability of winter oilseed rape modulated by early winter temperature. Scientific Reports. 2019;9:Article number: 6953.
8. Church BM, Austin RB. Variability of wheat yields in England and Wales. The Journal of Agricultural Science. 1983;100(1):201-4.
9. Gales K. Yield variation of wheat and barley in Britain in relation to crop growth and soil conditions—A review. Journal of the Science of Food and Agriculture. 1983;34(10):1085-104.
10. R Core Delevopement Team R: A language and environment for statistical computing. Vienna, Austria: R Foundation For Statistical Computing; 2016.
11. Pinheiro J, Bates D, DebRoy S, Sarkar D, Team} RC. {nlme}: Linear and Nonlinear Mixed Effects Models. R package version 3.1-137 ed2018.
12. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(1):267-88.
13. Breiman L. Random Forests. Machine Learning. 2001;45(5):5-32.
14. Dietterich TG. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning. 2000;40(2):139-57.
15. Genuera R, Poggiab J-M, Tuleau-Malotc C. Variable selection using random forests. Pattern Recognition Letters. 2010;31(14):2225-36.
16. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18-22.
17. Paluszynska A, Biecek P. randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. R package version 0.9 ed2017.
18. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth Edition ed: Springer; 2002.
19. KelejianIngmar HH, Prucha R. On the asymptotic distribution of the Moran I test statistic with applications. Journal of Econometrics. 2001;104(2):219-57.
20. Moran P. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1-2):17–23.
21. Giraudoux P. pgirmess: Spatial Analysis and Data Mining for Field Ecologists. R package version 1.6.9 ed2018.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

22. Claassen R, Just RE. Heterogeneity and Distributional Form of Farm-Level Yields American Journal of Agricultural Economics. 2011;93(1):144–60.
23. Just R, Weninger Q. Are Crop Yields Normally Distributed? American Journal of Agricultural Economics. 1999;81(2):287-304.
24. Harri A, Erdem C, Coble KH, Knight TO. Crop Yield Distributions: A Reconciliation of Previous Research and Statistical Tests for Normality. Applied Economic Perspectives and Policy. 2009;31(1):163–82.
25. Norwood B, Roberts MC, Lusk J. Ranking Crop Yield Models Using Out-of-Sample Likelihood Functions American Journal of Agricultural Economics. 2003;86(4):1032-43.
26. Wilson P, Hadley D, Asby C. The influence of management characteristics on the technical efficiency of wheat farmers in eastern England. Agricultural Economics. 2005;24(3):329-38.
27. Shekoofa A, Emam Y. Effects of Nitrogen Fertilization and Plant Growth Regulators (PGRs) on Yield of Wheat (Triticum aestivum L.) cv. Shiraz. Journal of Agricultural Science and Technology. 2008;10:101-8.
28. King C. Plant growth regulators for wheat. Top crop manager 2015.
29. Welling SH, Refsgaard HHF, Brockhoff PB, Clemmensen LH. Forest Floor Visualizations of Random Forests. ArXiv. 2016;e-prints.
30. Chami DE, Knox JW, Daccache A, Weatherhead EK. The economics of irrigating wheat in a humid climate – A study in the East of England. Agricultural Systems. 2015;133:97-108.
31. Zhang Y, Su S, Tabori M, Yu J, Chabot D, Baninasab B, et al. Effect of Selected Plant Growth Regulators on Yield and Stem Height of Spring Wheat in Ontario. Journal of Agricultural Science. 2017;9(11):30.
32. Messner JW, Mayr GJ, Zeileis A. Heteroscedastic Censored and Truncated Regression with {crch}. The R Journal. 2016;8(1):173--81.
33. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Journal of Statistical Software. R package version 1.0.5 ed2017.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Appendix 1: Dataset construction

Pesticide Usage Survey

The primary function of the PUS is consistent monitoring of agrochemical inputs onto the UK landscape (here we only use data for England and Wales) with a focus on pesticide application. However, due to the way in which the data is compiled, incidental information on yield values is also collected at a field level resolution and it is these values (after aggregation, see below) that represent the end point for the conducted analyses. It should be noted that there are a large number of samples in the PUS which lack recorded yield information (in 2010 this was nearly 30% of farms surveyed [168 of 564]), which were excluded from the study, and we currently have no way of assessing the extent to which this data loss is non-random. Examining only the data from the crops 'wheat' and 'osr', information was compiled In the case of wheat the sample was restricted to the sum of fields explicitly label as winter wheat or which have sowing dates during the period September to February (based on discussion with the PUS team).

Data (originally collected at field level resolution) was aggregated to the to the level of the holding, so as to be consistent with the geolocation information outlined below. In general, unless otherwise noted, measures taken from the PUS are the means for a holding after excluding any inapplicable data. Farms within the PUS are identified by a unique hold number assigned for the year of survey. These were mapped to the country parish holding scheme (the basis of the June Survey) using supporting information provided by the PUS team.

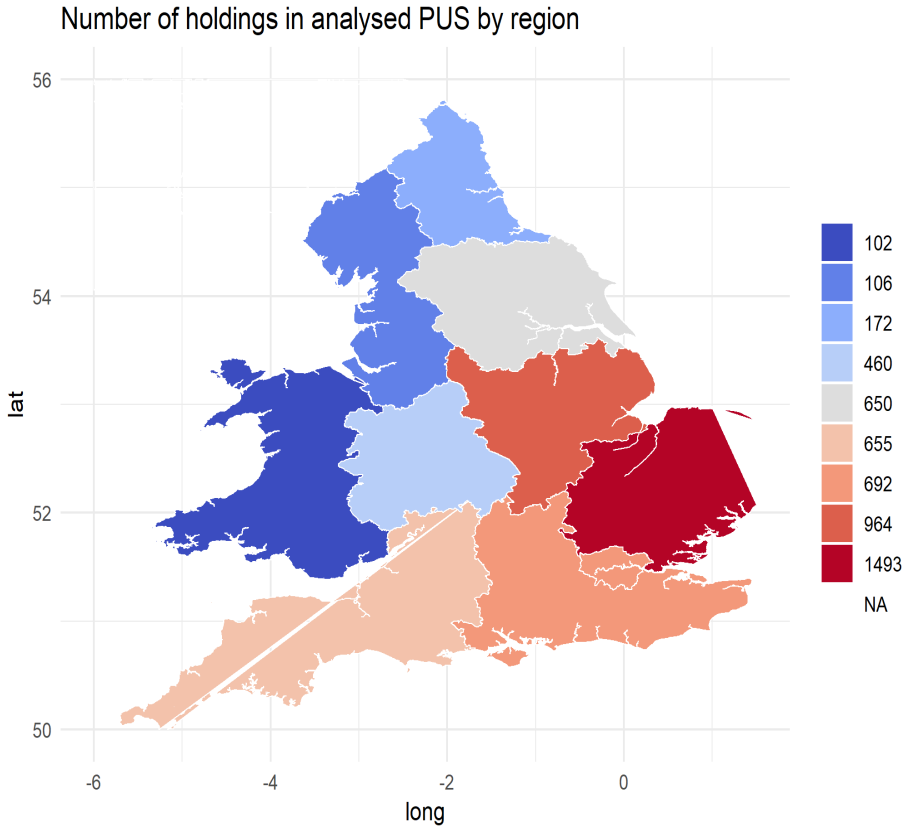


Figure 26. Geographical distribution of farms surveyed in the source PUS data (all years; only sites where yield values of focal crops were reported are used in analyses).

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Additional to previous analyses, we have grouped wheat varieties based on the usage classification provided by the National Association of British and Irish Flour Millers (henceforth nabim). This industry body is largely responsible for defining the economic usage of wheat varieties in a particular year. For simplicity we assessed the historic record of nabim classification during our period of interest and then assigned a single classification to each named variety based on the highest quality class (rated 1 to 4) obtained during the period. When discussing 'high quality bread wheats' we explicitly refer to those wheats classified as group 1 under the nabim scheme. When we refer to 'feed wheats' this is considered to include the sum of varieties listed as nabim group 4 and those which are unclassified under the nabim scheme (on the assumption that feed wheat is by far the majority of UK production). While initially classified at a variety level for aggregation purposes any mixed fields were assigned to the lowest applicable nabim class (as it is not possible within PUS to resolve yield values below the level of a single field). Yields are calculated as the division of the total quantity (yield*area) for a nabim group divided by the total area field growing the group for a holding.

Within the PUS applied agrochemicals are broadly classified into usage types, the structure of which we follow here (see below). For our core feature set we have also calculated additional aggregated statistics as summaries of the overall patterns of pesticide usage. These include, the total (rate of) pesticide application, the average number by field of unique active ingredients applied and the average number of distinct spray round undertaken, with a round defined as a unique combination of a spraying date, method and area of application. Measure included from the PUS in the expanded feature set are named and defined as follows:

- nabinOne_Yield/Feed_wheat_Yield/Unknown_Yield. Average yields for crops belonging to the different nabim groups (wheat only). All OSR is collected under Unknown_Yield (a convention for the way classes were assigned). Missing data and 0 values are excluded from models
- Area; Total area in ha of fields in PUS. Used to standardise the yield and pesticide use values. Usually logged.
- Drill_value; the mean drilling rate across included fields [Values over 1000 (wheat) or 100 (OSR) are excluded]
- Prop_Own_seed; the average proportion of Home saved seed sown
- Mean_Sowing_diff; Average across fields of the difference in days between the recorded date of sowing (where given) and the overall average for the crop in that year; default 0 if no sowing data is recorded
- Mean_Harvest_diff; As Mean_Sowing_diff
- Primary_Variety; The variety occupies the largest area on the holding (Varieties present at less than 20 sites are listed as Unknown)
- Adjuvant/Desiccant/Foliar feed/Fungicide/ Fungicide/Growth regulator/Herbicide/Insecticide/Molluscicide/Repellent/Seed treatment/Sulphur/Desiccant- The total amounts of types of pesticides applied on a farm, standardised by Area.
- Total_Pestcide; sum of the (standardised) pesticides used
- Mean_Spray_Rounds; the average across fields of the number of spray rounds undertaken.
- Count_Compounds: the average across included fields of the number of distinct active ingredients in the applied pesticides
- Count_Inseticide_Rounds: the average across included fields of the number of distinct spray rounds including at least one member of the group Insecticide

DEFRA June Survey

The June survey assesses (among other measures) land use, livestock numbers and the agricultural workforce for a sample of English and Welsh holdings each year. These samples are typically

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

interpreted in the context of a decadal full census of land use, which for our dataset occurred in 2000 and 2010. The June survey is particularly relevant as the source of the geolocatable information (presented as eastings and northings of the centroids of the holding) which were used to fit the spatial component of the models outlined below. Due to subtle variations in the structure of the survey year on year extracted information the combinations of questions used to construct different measures vary across the dataset [in brackets below]. Before inclusion in modelling all variables, apart from Cereals_Concentration were standardised to the value of the Land_Farmed_by_Farmer_June. Include measures are defined as

- Land_for_Wheat; Land cultivated by wheat [a1]
- Land_for_OSR: Land cultivated by winter oil seed rape [a24, sum(a241,a242)]
- Total_Open_Veg; Total Vegetables and Salad Grown in the Open [b99]
- Total_NS_Bulbs_Flowers; Total Hardy Nursery Stock and Bulbs and Flowers[d99]
- Total_Grass_SE; Total grassland, Set-aside and Other types of land [g99, g98]
- Area_Holding_June; Total area of Holding [h1, sum(h2, h3,h4,h5,h11)-sum(h8,h9)]
- Land_Owned_June; Land Owned [h2]
- Land_Farmed_by_Farmer_June; Area Farmed by Farmer [h10]
- Land_Rented_June; Area rented for 364 days or less [h11, missing from 2000 and 2002]
- Land_Letout_June Area you let out for 364 days or less [h12, missing from 2000 and 2002]
- N_Pigs; Number of Pigs [l98]
- N_SheepLambs Number of sheep/lambs [m98]
- N_Labour Number of labourers employed[q98]
- N_CowsCalves Number of cows/calves [k98, k299]
- Cereals_Concentration The proportion of the total holding given over to production of wheat or osr [(a1+a24)/h10]

DEFRA Crop disease Survey

The crop disease survey, of which oilseed rape and wheat are components, is undertaken annually for the purposes of assessing the prevalence of various commercially sensitive pests across the UK. For winter wheat, data is collected from approximately 300 crops sampled randomly from a geographically stratified list of sites during the medium milk development stage. Each sample represents the average of 25 tillers and assessment is focused around widespread fungal diseases e.g. brown rust (*Puccinia triticina*), *Septoria tritici*, and yellow rust (*Puccinia striiformis*). Oils seed rape values are calculated from a sample of 100 crops, spatially stratified, and with each record comprising 30 crops sampled at three points within the season; mid-leaf production, early stem extension, and at pod ripening. The key focus of this survey is *Alternaria, Phoma, Sclerotinia*, and light leaf spot (*Pyrenopeziza brassicae*).

One of the challenges for incorporating the crop disease survey information into our combined dataset is that there is very limited overlap between the sites sampled for the PUS and/or June survey and those used for the crop disease sample. Therefore, is it necessary to construct an index reflecting the regional prevalence of the disease (a potential driver of yield), based on the information available in the crop disease survey. To further add to the complexity, different varieties sampled in the crop disease survey have known differences in their resistance to the various crop diseases and so provide different levels of information regarding the regional prevalence of the pest organism. In the absence of clear guidelines, our approach here has been largely experimental and may be revised in subsequent work. We have taken information on the definition of variety resistance from the publications of the AHDB and used this to approximate the log linear relationship between disease

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

prevalence and the assigned resistance category (approximate slope of 0.2 disease units)[16].Using this relationship we have then standardised the reported prevalence (taken that the mean of prevalence of the disease at different growth stages or on different leaves if applicable) in the samples taken in the crop disease survey based on the reported crop resistance level and used this as the basis for a regional index of disease prevalence. To map this index to our sites as identified in the PUS our dataset includes two different approaches, firstly assigning each site an index value based on the crop disease sample with the shortest strait line distance (ties resolved arbitrarily; this methodology follows that established in the previous round of analysis), and secondly by assigning the index value of a site as the average of all crop disease samples with a radius of 50km (this our preferred approach and the basis of the disease prevalence measures included in the core feature set).

Variables as named in dataset (referenced are column heading in the original crop disease survey dataset supplied):

- Wheat
    - Mapped to closest sample:
        - "corected_yrust_inc", mean of yrust_lf1_inc and yrust_lf2_inc scaled to the varietal yrate
        - "corected_brust_inc", mean of brust_lf1_inc and brust_lf2_inc scaled to the varietal brate
        - "corected_trrust_inc", mean of tritici_lf1_incidenc and tritici_lf2_incidence scaled to the varietal trate
    - Mapped as mean within 50km (all likewise scaled):
        - 'Average_Corrected_yrust_index',
        - 'Average_Corrected_Tritici_index',
        - 'Average_Corrected_brust_index'
- OSR
    - Mapped to closest sample:
        - corected_phoma_inc; mean of incidence of phoma_summer and phoma_spring scaled (as above) to the varietal res_sk
        - corected_lls_inc; mean of incidence of plants_spring_lls scaled to the varietal res_lls
        - pod_Alternaria*
        - stems_with_alternaria*
        - stems_sclerotinia*
    - Mapped as mean within 50km(all likewise scaled)::
        - Average_Corrected_phoma_index',
        - 'Average_Corrected_lls_index',
        - 'Average_pod_alternaria',*
        - 'Average_stems_with_alternaria',*
        - 'Average_stems_sclerotinia*

*There is no resistance rating reported for these measures and hence they are reported as their raw values

[16] This value was estimated based on the relationship function given for wheat resistance and observed disease prevalence in https://cereals.ahdb.org.uk/media/408807/Understanding-RL-disease-ratings.pdf

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Meteorological Office Climatic data

Climatic data is taken from geo-located weather stations across the UK and consists of daily records of average temperature, rainfall, wind speed and humidity. For each weather station, data were aggregated to the 'pre-frost' (defined as the period as the period from July to December of the year prior to harvest for the purposed on rainfall and humidity or as July of the year prior to harvest to February of the year of harvest in the case of temperature) and 'pre-harvest' periods (the latter defined as the period from June to September in the year of harvest) and average values calculated[17]. The variables included in the modelling are based on previously identified potential drivers of wheat yields (see previous phase of analysis) and are named below. For assembly purposes, the values associated with a holding were taken to be those of the closest weather station based on strait line distance calculated on their respective eastings and northings.

- mean.rain.prefrost
- mean.temp.prefrost
- mean.RH.prefrost
- mean.wind.preharvest
- mean.rain.preharvest

National Soil Map

During the calculation of the reported analysis, administrative constraints resulted in a lack of direct access to data previously used to define soil types based on the DEFRA national soil map. The presented values are therefore approximations based on the records constructed as part of the previous round of analysis. Underlying the presented measures are nice soil measurements representing fertility ["low", "moderate", "moderate to high"], drainage ["freely draining", "impeded drainage" ,"naturally wet"] and soil texture ["loamy", "peaty", "sandy"]. These are presented as the unique values by postcode district and assembled to the dataset via postcode information provided from the previous round of analysis.

Compared to other variables included in this study the various soil measures are expected to be highly correlated with one another, reflecting for example the tendency for impeded drainage and low fertility in peaty soil types. To address this issue Principal component Analysis (PCA) was applied to the soil data prior to modelling, thus reducing the number of included variables to a set of uncorrelated linear axes representing successively smaller quantities of the underlying variation in the dataset.

PCA analysis was conducted individually on the wheat and oil seed rape datasets and the resulting axes are visualised on Figure 28 and Figure 30 respectively (visualised using R package *factoextra* (33)).

[17] This approach originates from preliminary work by Mojtahed and Budge and reflects discussions regarding the relevant periods for crop growth for wheat. Alternative schemes based on, for example on sowing date, are possible in the context of the compiled data but have not been explored within the present analysis.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

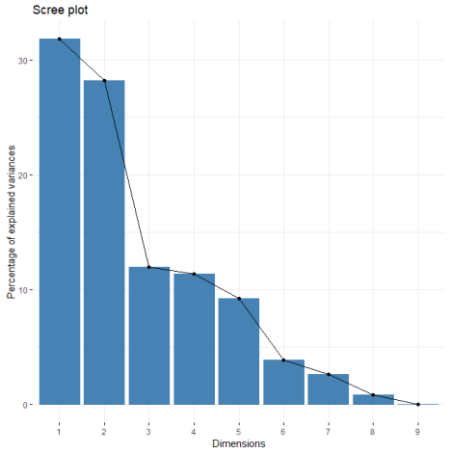


Figure 27 Scree plot of the proportion of variance associated with each calculated axis of the PCA of soil variables applied to wheat

Figure 28 PCA of Soil variables for wheat displayed as a biplot. The alignment of the underlying measures with respect to the calculated principal components are shown by the blue arrows. The black dots represent the position of the recorded data with respect to the calculated axes. Plot shown is for PC1 (x axis) and PC2 (y-axis).

For wheat the first principal component (comprising 32% of observed variance) is primarily associated with the transition from freely draining (low scores) verses both classes of wet soils (high scores), with also a strong contribution for scores along the low fertility axis and the distinction between loamy (low scoring) as opposed to peaty soils (high scoring). PC2 (28% of variance) is most closely correlated with the trend towards sandy soil textures (high scoring) with also a strong contribution in the distinction between soils with impeded drainage and moderate fertility (both associated the lower scoring localities). For ease of use we have restricted the core data set to only include these first two components which are by far the most important in terms of explaining overall variance (Total 60% of the overall variance in soil types is explained by this representation; Figure 27)

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

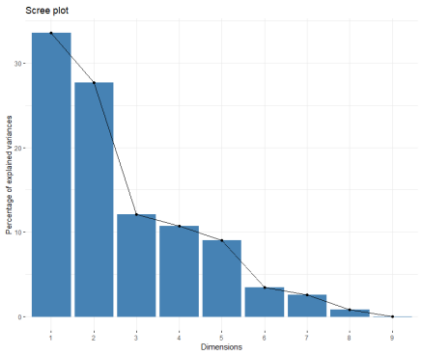


Figure 29 Scree plot of the proportion of variance associated with each calculated axis of the PCA of soil variables applied to oil seed rape

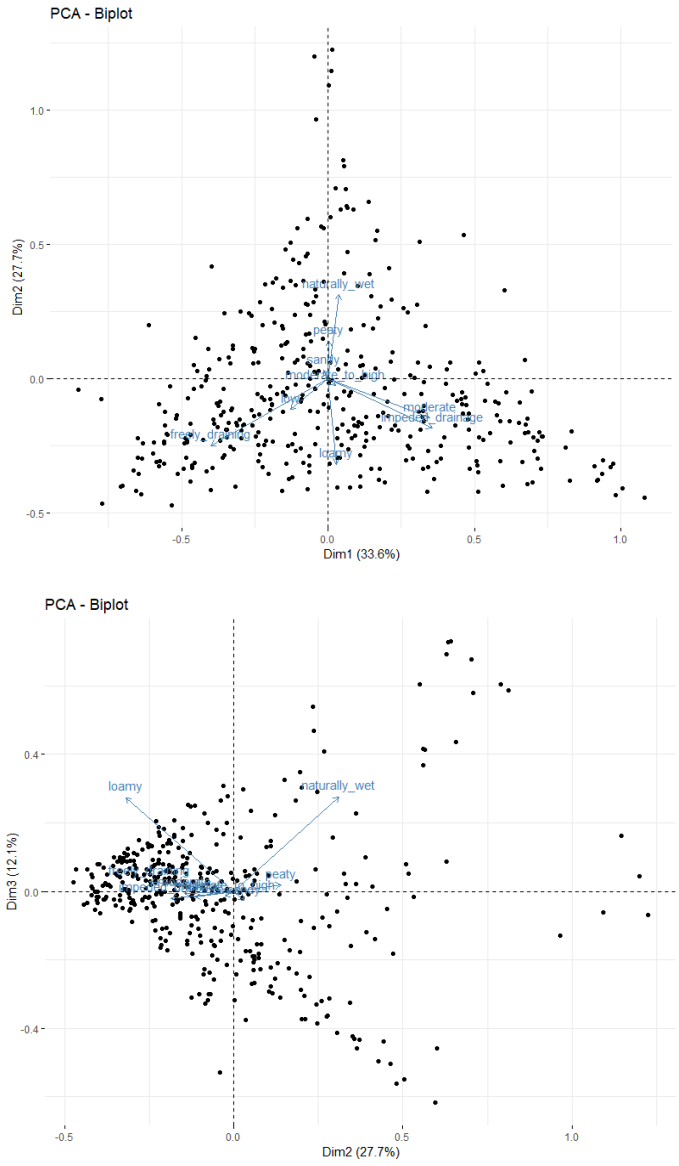


Figure 30 PCA of Soil variables for oil seed rape displayed as a biplot. The alignment of the underlying measures with respect to the calculated principal components are shown by the blue arrows. The black dots represent the position of the recorded data with respect to the calculated axes. PC1 (x axis) and PC2 (y-axis) are shown on the upper plot, while PC2 (x-axis) and PC3 (y-axis) are shown in the lower plot.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

For OSR records the results of PCA show subtle distinctions from that of wheat, particularly in the orientation of some variable with respect to PC2. Here PC1 (34% of variance) can largely be interpreted as the dissention between low fertility and freely draining soils on the one hand (low values) and moderate fertility soils with impeded drainage on the other. PC2 (28% of variance) is closely aligned to the distinction between peaty and loamy soils with naturally wet soils being strongly associated with high peat content. Also included in the models for this crop PC3 (12% of variance) can be interpreted as a continuum between naturally wet and loamy soils (high scoring) and relatively dry soils with a high sand content. Together these variables represent approximately 73% of the variation observed in the underlying dataset.

Economic Data

The economic data was extracted from John Nix farm management pocketbooks from 2000 to 2010 on a biennial basis matching the years of the arable pesticide use survey. The main economic variables are the prices of crops and costs of production such as pesticide costs and land rent. The data complied is not associated with localities and instead is reflective of overall average values across a year. Due to this low-resolution incorporation of the economic measures, use of this data was restricted to the Expanded feature set. Measures included are:

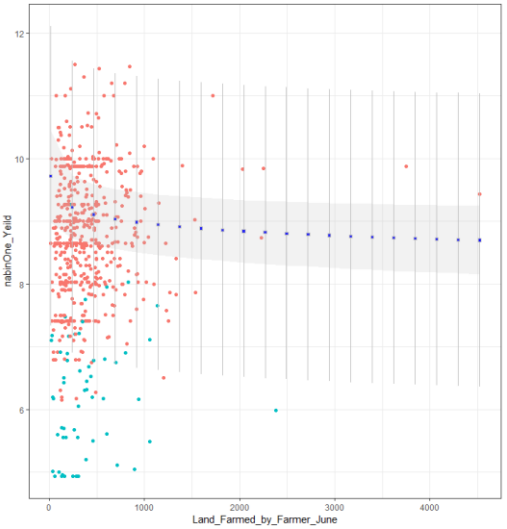
- average_Winter_Wheat_Price_feed; Winter wheat price for feed
- average_Winter_Wheat_Price_milling; Winter wheat price for milling
- average_Winter_Wheat; Average between milling and feed price
- average_OSR_price; Average oilseed rape price
- heribicide_price_cereal ; Average among several active ingredient types (e.g. general, under-sown crops, blackgrass, cleavers, wild oats, etc.)
- growth_regulator_price_cereal ; Average among several active ingredients (e.g. Chlormequat, Choline Chloride, Imazaquin, Mepiquat Chloride, etc.)
- fungicide_price_cereal; Average among several active ingredients (e.g. Azoxystrobin, fenpropimorph, epoxiconazole, tebuconazole, etc.)
- Aphicide_price_cereal; Average among several active ingredients (e.g. Cypermethrin, deltamethrin, pirimicarb, chlorpyriphos)
- slug_killer_price_cereal ; Average among several active ingredients (e.g. metaldehyde, methiocarb)
- heribicide_price_OSR; Average among several active ingredients (e.g. propyzamide, trifluralin, metazachlor)
- insecticide_price_OSR; Average among several active ingredients (e.g. deltamethrin, triflularin, metazachlor)
- fungicide_price_OSR; Average among several active ingredients (e.g. Improdione, Tebuconazole, Flusilazde, metconazole)
- dessicant_OSR ; Glyphosate
- land_prices_ha;Average price between 4 quarters of CALP/RICS farmland price index-Covering sales of 5ha and above
- rent_ha_full_agri; Full Agricultural Tenancies
- rent_ha_farm_business ; Farm business Tenancy rent (for 1 year and over)
- average_Rent; Average between Farm Business Tenancy and Full Agricultural Tenancy

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

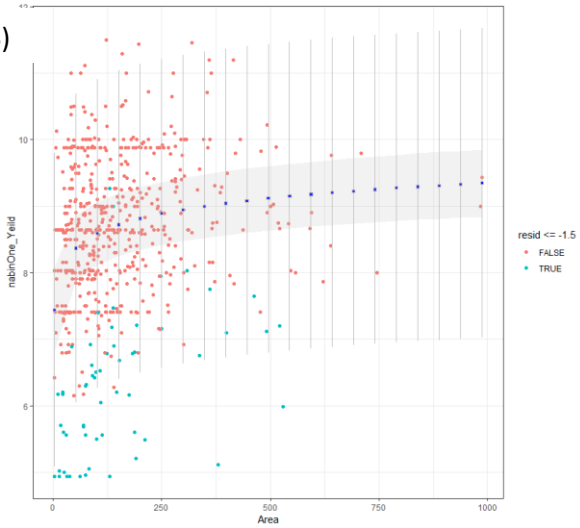
Appendix 2: Visualising the shape of numeric parameter effects in statistical modelling

Relationships from statistical modelling for significant parameters of core feature set of nabim One varieties of wheat

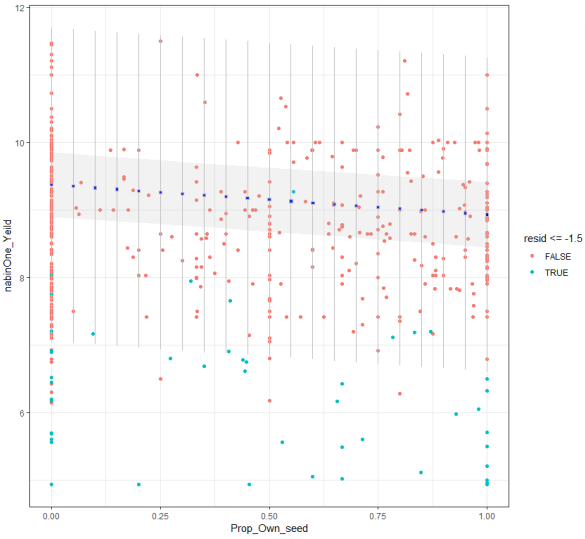
A)



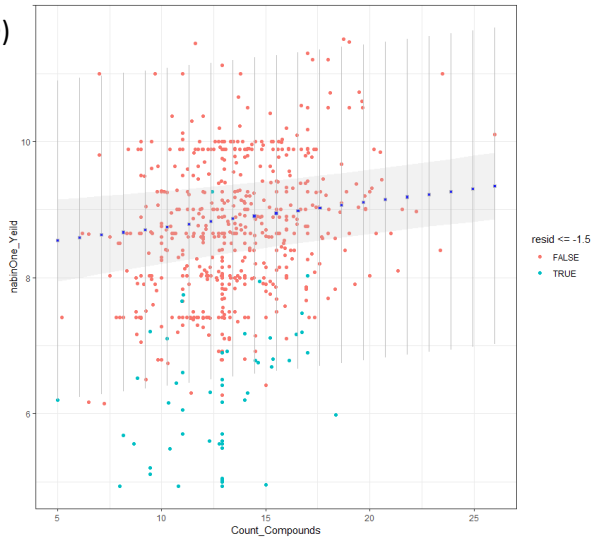
B)



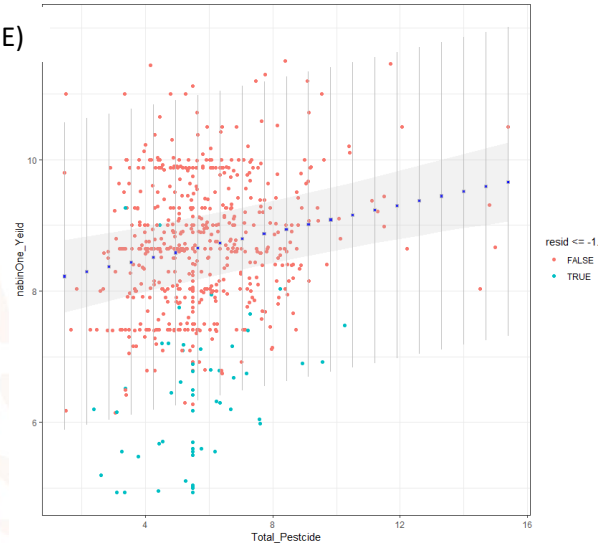
C)



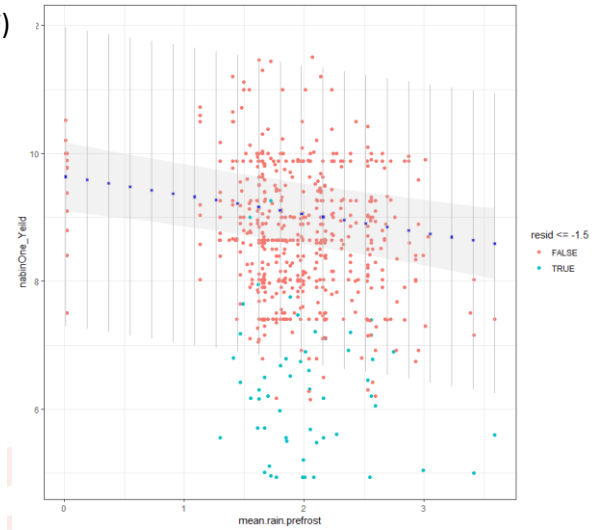
D)



E)



F)



Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

G)

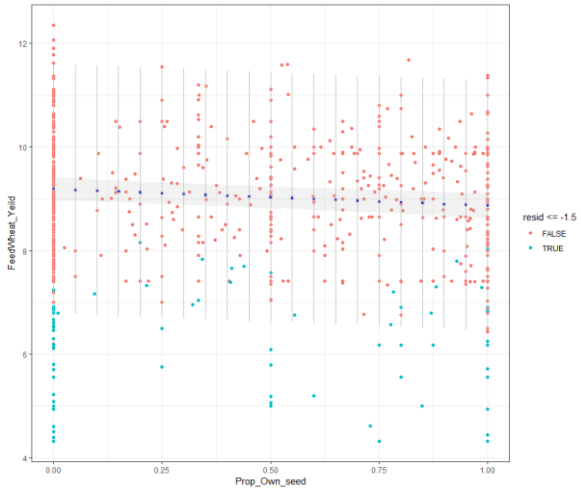
Figure 31 Plot of the prediction interval for the numeric significant parameters identified from statistical modelling of the core feature set for nabim One varieties of wheat. Dark points are the mean effect of the parameter estimated from the model with the shaded region showing the (95%) confidence interval on this estimate. The area within the lines represents the 95% prediction interval given the overall error on the model (i.e. the area in which 95% of the true values are expected to lie). The observed data is shown as points, with values associated with a large negative residual (subpopulation of non-conforming localities) shown in blue (see Figure 9). Effects are shown in the following order (matching that in Table 1) A) Area, B) Land farmed by farmer, C) Proportion own seed, D)Count compounds, E) Total pesticide, F) Mean rain pre-frost, G) Latitude.

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Relationships for significant parameters of core feature set from statistical modelling of feed varieties of wheat

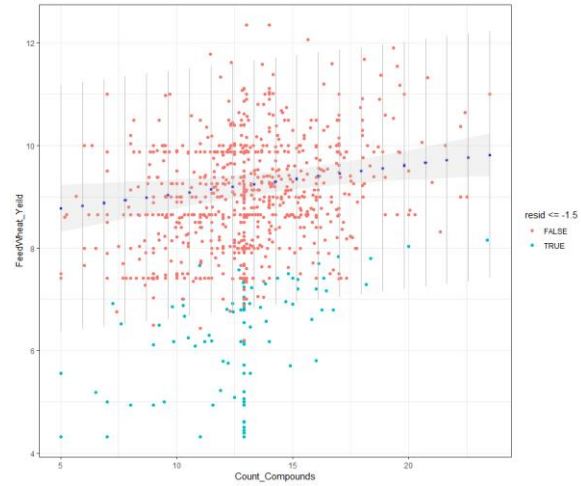
A)

B)

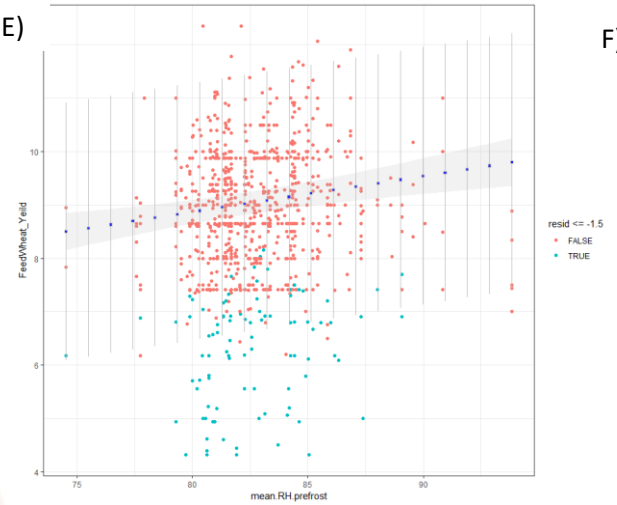


C)

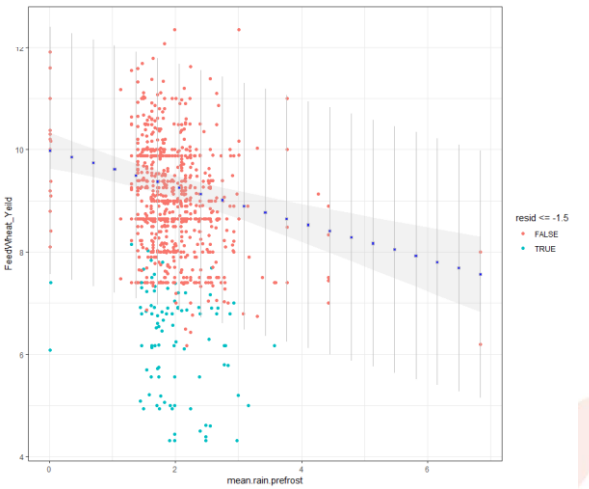
D)



E)

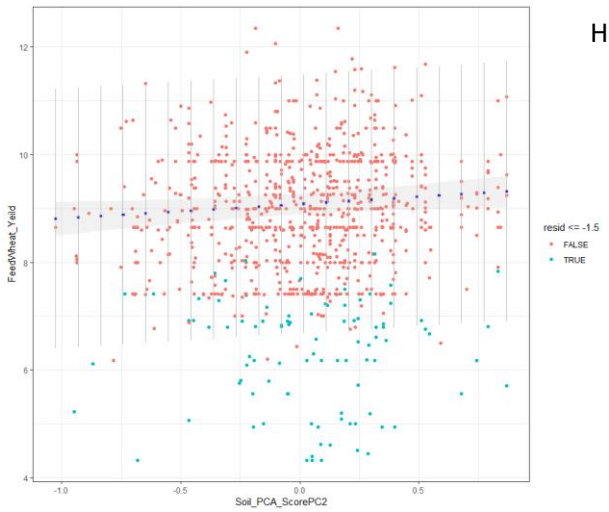


F)



Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

G)



H)

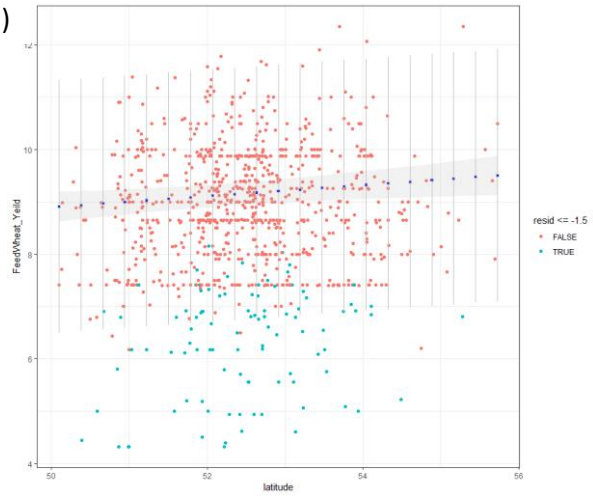


Figure 32 Plot of the prediction interval for the numeric significant parameters identified from statistical modelling of the core feature set for feed varieties of wheat. Dark points are the mean effect of the parameter estimated from the model with the shaded region showing the (95%) confidence interval on this estimate. The area within the lines represents the 95% prediction interval given the overall error on the model (i.e. the area in which 95% of the true values are expected to lie). The observed data is shown as points, with values associated with a large negative residual (subpopulation of non-conforming localities) shown in blue (see Figure 16). Effects are shown in the following order (matching that in Table 2) A) Area, B) Proportion own seed, C) Mean spray rounds, D)Count compounds, E) Mean humidity pre-frost, F) Mean rain pre-frost, G) Second principal component of soil factors H) latitude

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk

Relationships for significant parameters of core feature set from statistical modelling of OSR

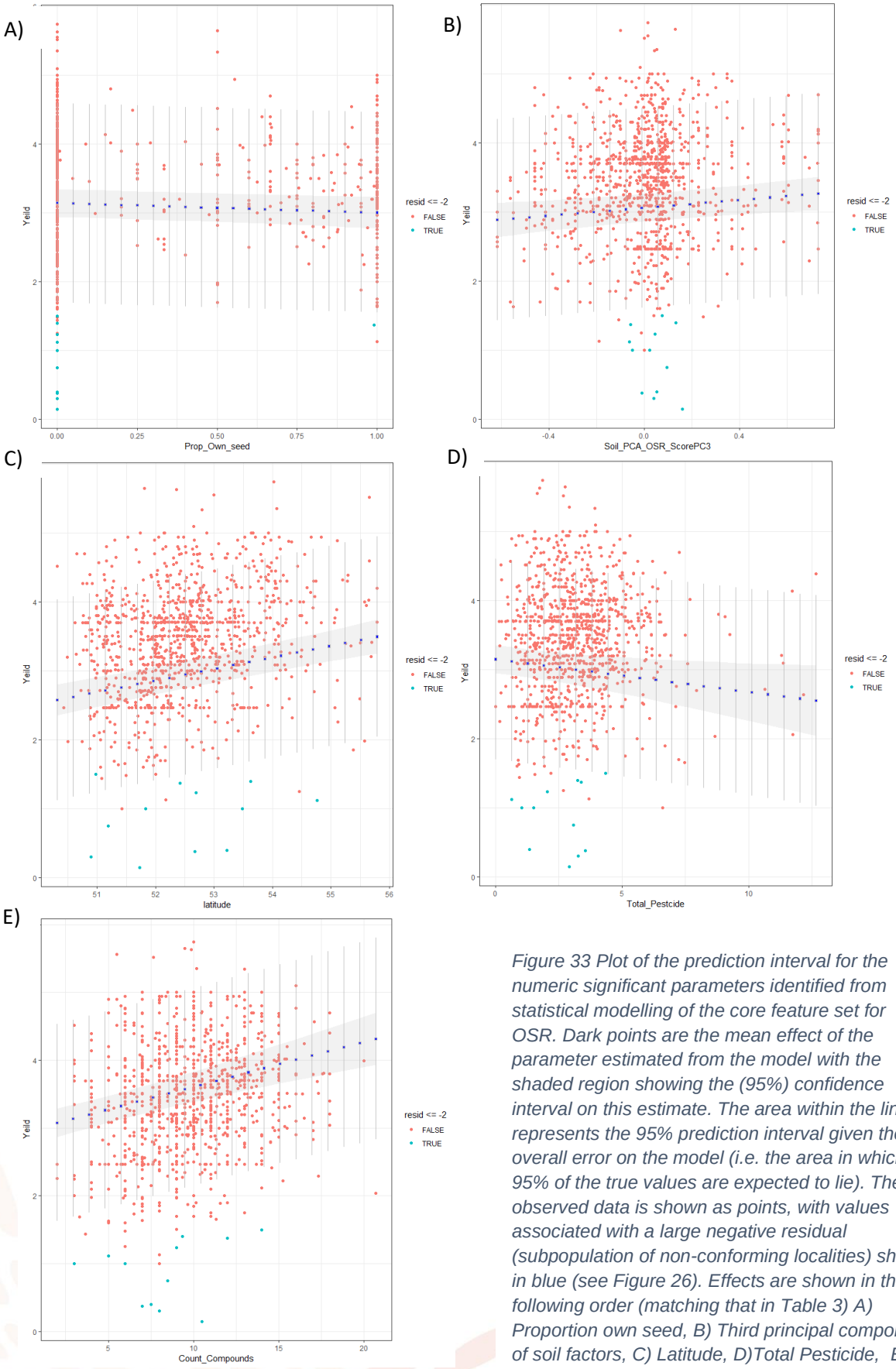


Figure 33 Plot of the prediction interval for the numeric significant parameters identified from statistical modelling of the core feature set for OSR. Dark points are the mean effect of the parameter estimated from the model with the shaded region showing the (95%) confidence interval on this estimate. The area within the lines represents the 95% prediction interval given the overall error on the model (i.e. the area in which 95% of the true values are expected to lie). The observed data is shown as points, with values associated with a large negative residual (subpopulation of non-conforming localities) shown in blue (see Figure 26). Effects are shown in the following order (matching that in Table 3) A) Proportion own seed, B) Third principal component of soil factors, C) Latitude, D)Total Pesticide, E) Count Compounds

Fera Science Ltd.
National Agri-food Innovation Campus,
Sand Hutton, York, YO41 1LZ
www.fera.co.uk
T: +44 (0)300 100 0321
E: sales@fera.co.uk