**PhD Summary Report No. 15**
**October 2009**
**Project No. RD-2005-3208**

# SPATIAL/TEMPORAL MODELLING OF CROP DISEASE DATA USING HIGH-DIMENSIONAL REGRESSION

by

**Weiqi Luo**

The Food and Environmental Research Agency
And The University of Leeds

September 2005- October 2008

# CONTENTS

## ABSTRACT

Septoria leaf blotch, caused by the fungus *Septoria tritici*, is one of the most serious foliar diseases of winter wheat across England and Wales, causing considerable reduction in yield quality and quantity. There are increasing pressures (e.g., legislative, environmental protection, public awareness) to control such diseases in a responsible and sustainable fashion. Disease forecasting systems may offer the potential to meet such an aim but previous work has struggled to find a reliable scheme that growers can adopt easily.

This project aims to develop new approaches and methodologies for developing risk prediction schemes, which are reliable and commercially practicable. To achieve this the project attempts to improve the reliability of forecasts for *S. tritici* through the development of novel methods for analysing weather records and associated historic disease data. The two major advances achieved are the development of high dimensional statistics and the use of a clustering method to define disease forecast zones. Firstly, new regression techniques were developed for the analysis of data where there are many more potential predictors than observations, so called "fat data". Next, two innovative procedures for selecting the significant predictors in high dimensional models were introduced which included a new methodology to handle data with correlated errors. In addition, the high dimensional models were applied to quantify important environmental factors influencing *S. tritici* development. This led to the use of cluster analysis to define a set of forecast zones across England and Wales which demonstrated an improvement in prediction accuracy at an early stage in the growing season.

The project results show that epidemics have intrinsic temporal and spatial scales that must be matched by control strategies if they are to be both effective and efficient. In addition to the current application on *S. tritici*, the high dimensional models developed can be extended to other wheat, barley and arable crop foliar diseases.

# INTRODUCTION

*Septoria tritici* is one of the major foliar diseases of winter wheat crops and often causes significant yield loss. Although there is clearly a link between the weather and disease severity, the relationship is not understood quantitatively. Various attempts have been made in Britain and elsewhere to produce forecasting systems to model this relationship and act as management decision aids. Grower adoption of forecast models that give early warning of disease risk with sufficient lead time to allow effective treatments could optimise crop inputs and improve profitability. However, forecasting schemes are difficult to implement widely due to the challenges of gaining weather observations at site-specific scales. In-field weather data collection (using weather sensors and data loggers installed in crop canopies) could be applied to improve disease forecast implementation, but this approach is unrealistic for many growers as it can be costly, laborious and unreliable. Alternatively, a more practical way to acquire local weather information is to estimate from adjacent weather station networks. However, it is uncertain whether the estimated weather conditions are sufficiently accurate for disease forecasting systems due to additional errors from spatial interpolation for a site specific location.

With those intrinsic limitations, we believe that predicting disease pressure accurately for all farms at a field scale is impracticable for commercial management of arable crops. To free wheat growers and crop consultants from collecting in-field meteorological data and to speed up the implementation of disease forecasting, we propose a possible framework for disease modelling at larger scales, which requires less rigorous weather inputs. Cluster analysis has the capability to aggregate homogeneous farms together to allow modelling at a large scale. After identifying the suitable predictive scale, several high dimensional regression techniques can be applied for disease modelling.

The main objectives of the project were to:
(1) Evaluate the performance of various cluster analysis algorithms;
(2) Test whether the identified cluster regions can be used as an informative predictive scale for disease forecasting;
(3) Investigate the effects of key meteorological parameters such as temperature and rainfall on the development of *S. tritici*;
(4) Quantify the important weather variables for disease forecasting systems using high dimensional regression techniques and test the new system with historical disease and weather data.

# MATERIALS AND METHODS

## Clustering Methods

Instead of providing wheat growers with more site specific disease risk information, forecasting models applied at a larger geographic scale could offer several important benefits:
(1) Regional climatic data can be used, so growers and crop consultants are not burdened with the expense and maintenance of in-field weather stations.
(2) Ability to give more reliable estimates for the occurrence and severity of disease outbreaks at a larger scale by reducing the errors caused by local disease fluctuations.
(3) Aggregated regions allow the history of disease severity to be traced with more consistency when the sampled locations (i.e., wheat fields) change from year to year.

There has been an increasing interest in aggregating sets of sites that together represent the overall disease situation of the relevant area. Simple aggregation for disease samples can be done by using landscape patches or arbitrary county boundaries. However, the aggregated areas do not always match the phenomenon under investigation, as the spatial patterns for disease are not taken into account when defining arbitrary boundaries. Identifying networks of sites or potential large regions that efficiently represent the overall disease situation of relevant areas is a prerequisite for effective disease modelling.

The objective of cluster analysis is to find a classification in which the items of interest are sorted into a small number of homogeneous groups or clusters. More generally, cluster analysis endeavours to maximise the similarity within the members of each group and to minimise the similarity between groups. As a result, the derived classification scheme may provide a convenient way to describe the patterns of similarities and differences in the data. The resulting clusters can be considered as forecast zones where the risk of disease will be consistent for all farms within the zone boundary.

In this study, attention is restricted to four hierarchical (single linkage, complete linkage, average linkage and Wards' algorithm) and two partitional (recursive partitioning and K-means) clustering approaches for sample field aggregation. The cluster analysis methods were carried out for classification of a large, irregular dataset of approximately 3080 disease samples collected from commercial crops between 1991-1998. The distribution of the sites indicates a reasonable coverage for the arable parts of England and Wales (Figure 1). The majority of the samples were densely located in Eastern and Northern regions but sparsely located in the Southwest and Wales. Mean severity of the top two leaves was extracted from each site. Prior to mapping, these data underwent extensive quality-control inspections, ensuring the mean value was calculated through a constant number (25) of sub-samples at a valid location. The historical range of mean severity over 7 consecutive years was 0–61%. The distribution of mean severity is heavily skewed towards low values as more than 54% of the sites measured no disease.

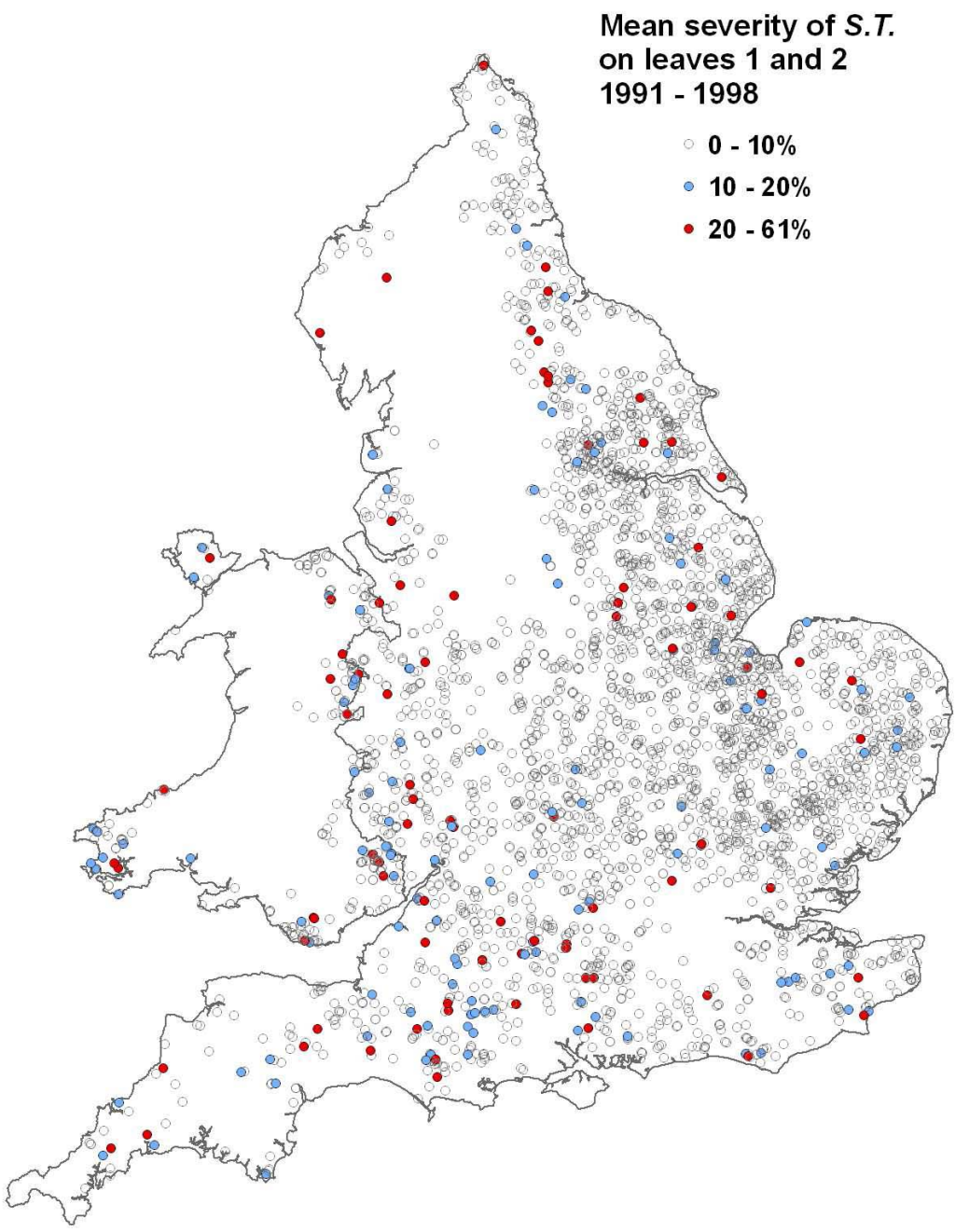


**Figure 1** Location of sample sites of winter wheat, 1991–1998.

Based on the spatial distribution of disease sample sites, the goal of cluster analysis is to group together localities that are geographically adjacent and whose populations are similar in the characters of interest: in this case disease severity. With these concepts borne in mind, a two stage clustering process was designed to analyse the selected data.

The first stage focuses on spatially distinguishable groups, and therefore the geographical distance between pairs of sites is used for the detection of distinct clusters. The clustering map from each method is then evaluated to determine the final cluster structure. The two stage cluster process aims to aggregate disease samples with as much homogeneity as possible and to enable the investigation of the relative importance of geographical area and meteorological conditions in determining disease progress. Ultimately, the application of a two stage cluster analysis can support improvement of disease management by providing an appropriate cluster level for disease forecasting.

The distribution for the site-specific disease severities are temporally and spatially irregular. To investigate the long-term disease pattern, we summarised them to construct 7 year annual means in 953 uniform grid cells of 10 km × 10 km covering most areas where wheat is grown (Figure 2a). Grid cell size less than 10 km are not recommended, because at smaller scales a number of grid cells only cover one farm. Gridding increases the accuracy in representing historical disease pattern, but it leads to degraded resolution. For the purpose of better visualisation, a smooth map was produced through spatial interpolation (ordinary kriging) with a high spatial resolution of 1 km (Figure 2b). Although the interpolated severity sacrifices some site specific information that might be valuable in judging the clusters, it was done primarily for two reasons: (a) potential outliers and unusual severity values tend to be "absorbed" into the interpolated value, which more reliably represents the historical disease pattern, and (b) local disease fluctuations are reduced substantially, thus allowing large scale clustering to be observed. As noted previously, the disease severities were interpolated based on a fine resolution so the majority of the local disease identity is still maintained. In general, the most severe cases of *S. tritici* were found in the Southwest and Welsh border regions.

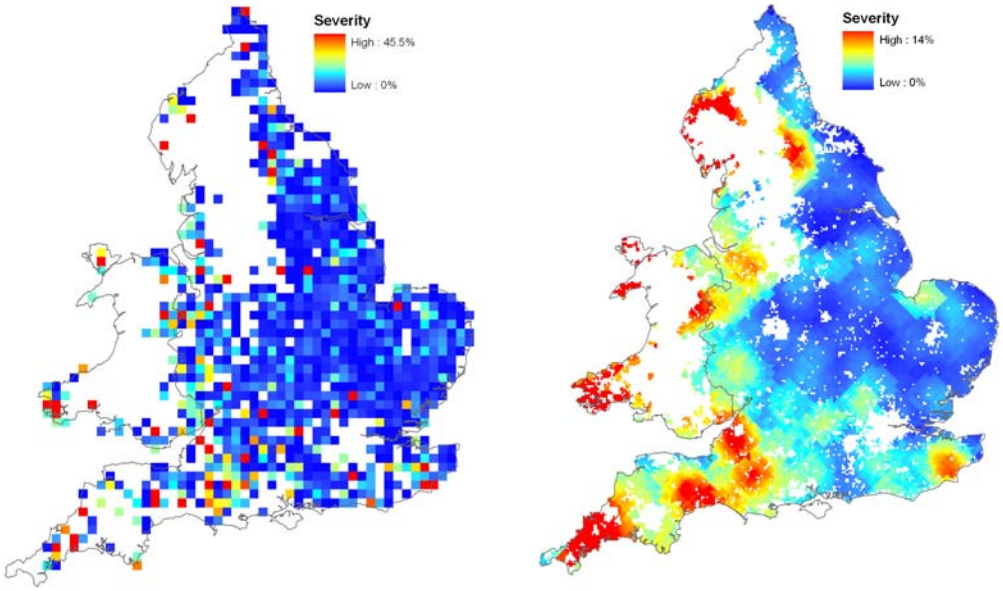


**Figure 2** (a) Summarised disease severity on 10 km × 10 km grid; (b) the corresponding smooth map for historical disease severity, 1991–1998.

Evaluation of how well the results of a cluster analysis fitted the data is referenced to external character differences using the historical disease severity on the top two leaves. The historical disease map indicated that retaining 2–30 clusters should be a more than sufficient range to reflect the disease fluctuation without losing any important broad scale information. As a second stage, we calculated the total historical disease severity variance between clusters, based on classification of 2–30 clusters for each year, and used the mean value as a measure for cluster effectiveness. The purpose was to determine a parsimonious set of discrete, possibly asymmetrically spaced, clusters that best summarise the disease situation.

## Cluster scale models

The historical disease data for *S. tritici* were obtained from the annual winter wheat disease survey in England and Wales during the period from 1998 – 2001. These disease data were used with frequently recorded weather data collected from the same period. The distribution of disease observations, with references labeling potential disease clusters, contains more than 1700 individual farm observations (Figure 3). The number of farms varied from 420 to 437 across the

years. Disease severity at the milky ripe stage (GS 75) was selected, because it is known to correlate closely with losses in yield quantity and quality.

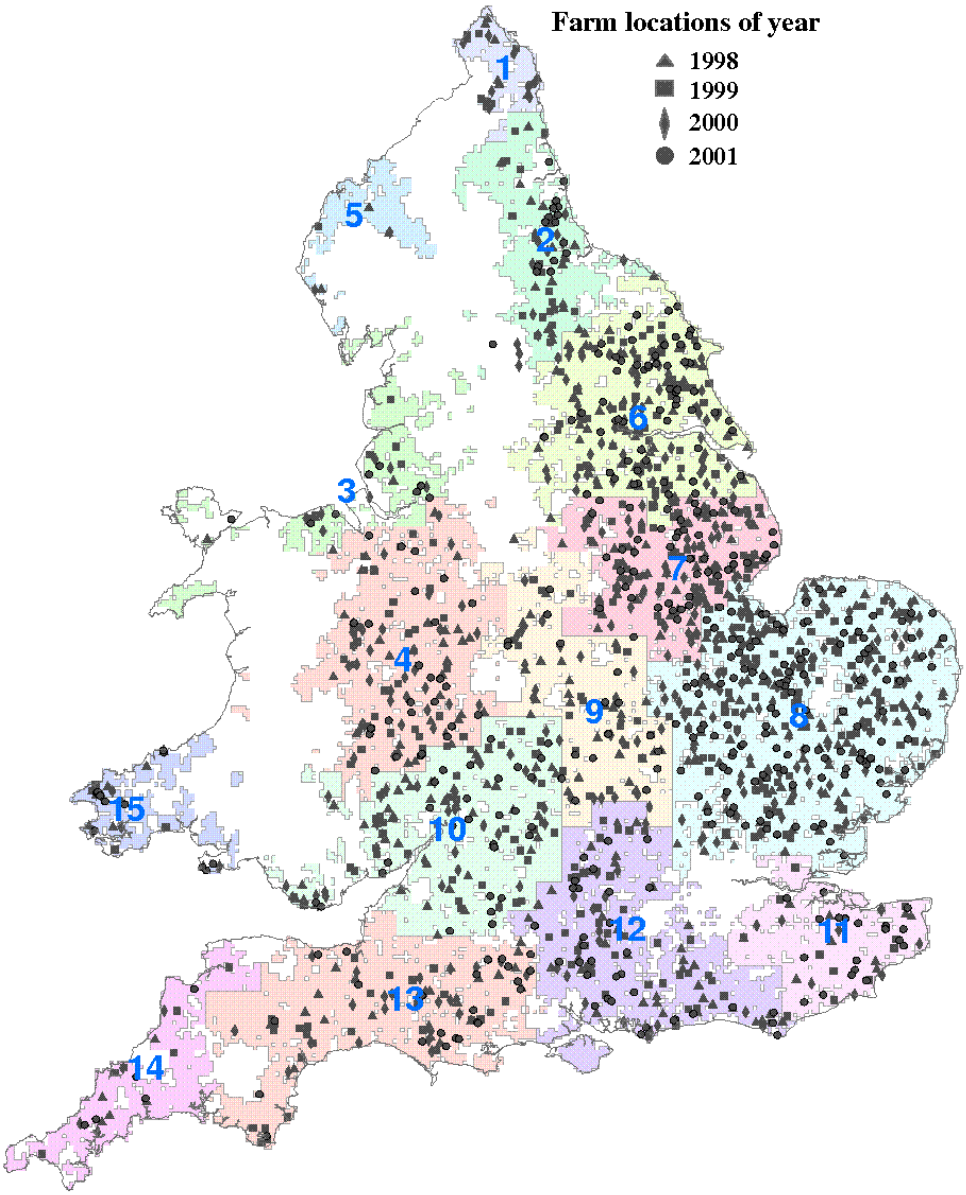


**Figure 3** The distribution of disease observations during the period from 1998 to 2001. The 15 potential clusters are labelled and displayed with different colours.

An aim of this study was to derive a practical model for predicting disease severity for the purpose of informing crop management decisions; final disease severity was therefore calculated across the top two leaves. As disease severity was determined based on visual assessments of symptomatic leaf area, accurate measurement for different leaf sizes between leaf layers was impossible. Mean

disease severity was therefore calculated by taking the average of the upper two leaf severities, rather than calculation based on absolute leaf area. For each disease cluster and for each year, the mean severity and standard error was then calculated. Great spatial variability exists between clusters, with disease levels tending to be higher in the southwest of England and Wales, and lower towards the north and east. To assure the reliability of the averaged disease severity, any cluster with fewer than three samples for a year was removed from the analysis.

Apart from weather conditions, a range of agronomic factors, such as cultivar, sowing data, and fungicide applications, are also associated with the severity of *S. tritici*. Farms with incomplete agronomic information and the small number sown after 31 December were excluded from the analysis. Less than 15% of the farm observations were omitted from the analysis. For winter wheat, a wide variety of cultivars are available. All of the cultivars contained in the dataset had a resistance rating in the range 3 (susceptible) to 7 (resistant). Resistance rating, to some extent, represents the individual effect of cultivar, although it is not so flexible as cultivar alone. A more parsimonious classification is defined by dividing resistance ratings into two groups, susceptible (rating 3,4,5) and resistant (rating 6,7). Only limited information is lost compared with using the full scale. Instead of calculating the mean resistance rating for each cluster in each year (which is inappropriate for a nominal variable), the proportion of farms for each re-classified resistance rating was calculated. Application of foliar fungicide is another effective method for disease control. The number of fungicide sprays applied before GS 75 was used to investigate the impact of fungicide application. Spray programmes with four or more applications were grouped as one category ($F_4$) in order to increase the number of observations for between group comparisons. One spray was defined as one foliar application regardless of dose or active ingredient applied. The sowing date can be transformed to a continuous variable by counting the number of days from 1 September. It is possible to have a negative value (by counting backward) if the crop was sown before September.

## Meteorological data

An entire set of in-field weather conditions were interpolated by appropriate methods using the UK met station network. Eight daily meteorological variables were analysed in this study: minimum (Tmin; ºC), maximum (Tmax; ºC) and mean (Tmean; ºC) air temperature, Rainfall (Rain; mm), minimum (Hmin; %), maximum (Hmax; %) and mean (Hmean; %) relative humidity, and mean wind speed (WS; m/s). No missing data were observed due to the benefits of using spatial interpolation, but the precision of the estimated data may vary because of different network coverage across time and space. An aggregation by mean across all farm weather data in each cluster was used to provide a representative weather situation for each cluster-year. These aggregated weather data form the foundation for further statistical analysis.

## Data analysis and high dimensional regression modeling

It is highly unlikely that weather conditions on a single day will explain fully development of disease in a crop or region. Consequently, an iterative summarising algorithm, window pane, was adopted to generate meaningful weather predictors. Window pane algorithms, described previously by several authors, have been used to search for correlations between the weather functions (of certain durations) and observed disease severity. In this study, the window pane algorithm was used to summarise weather variables into potential predictors for the period between 1 January and 30 June. The algorithm was implemented by searching with a time lag (defined as number of days counted backward from 30 June) and window length (defined as the number of days counted forward from the start of the time lag) of changeable size (Figure 4). Time lags varied from 180 to 30 days, and were moved in steps of 10 days. For each time lag, the window lengths were allowed to vary between two limits (the upper limit was equal to the current time lag and the lower limit was fixed as 30). The procedure started summarising with the largest window length and subsequently considered the smaller window lengths in decrements of 10 days. A comparatively large step length was chosen for both time lag and window length, to alleviate the problems of colinearity between summarised weather variables and hence reduce the spurious correlations with disease that might occur by chance. For the 4-year period (1998–2001) the approximate average spray date of T1, T2 and T3 is 20 April, 22 May and 15 June respectively.

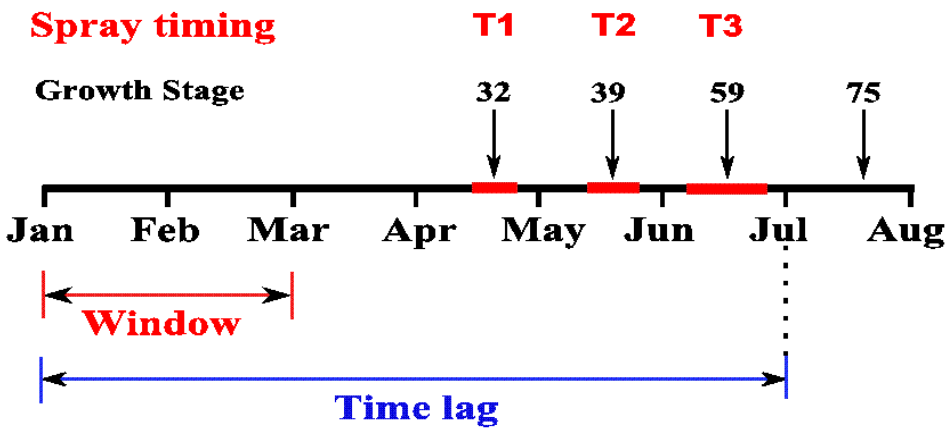


**Figure 4** Description of window pane algorithm. The search period was fixed between 1 January and 30 June. In order to balance between the precision of the analysis and the number of search windows, a moderate interval of 10 days was used for both the time lags and window lengths.

The weather functions summarise the daily weather variables within each searching window under certain criteria. Four commonly used weather functions were applied in this study (**Table 1**). With respect to the minimum temperature ($\mathrm{Tmin}$), it would not be surprising to find that the data from several weather functions are highly correlated. For example, little difference would be expected for the weather data derived from $\mathrm{Tmin}_{nod;>1}$ and $\mathrm{Tmin}_{nod;>2}$. The same issue applies to the other weather variables under examination, therefore determination of an "optimal" summarising weather function for each type of weather variable (e.g. $\mathrm{Tmin}$, $\mathrm{Tmax}$, $\mathrm{WS}$, etc.) is crucial.

**Table 1** Description of weather functions applied to each weather variable. For each weather variable we used a range of possible values as thresholds. For instance, the variable, $Tmin_{cnod;>7}$, represents the number of consecutive days with minimum temperature higher than 7ºC in a window.

| Weather function | Description | Thresholds |
|---|---|---|
| $V_{nod;>i}$ or $V_{nod;<i}$ | Number of days the weather variable (V) above (>) or below (<) a threshold ($i$) in a window | Tmin: $i = -2, \ldots, 7$; Tmax: $i = 20, \ldots, 25$; Tmean: $i = 7, \ldots, 14$; Rain: $i = 0, \ldots, 9$; Hmin: $i = 50, 60, 70$; Hmax: $i = 92, 95, 99$; Hmean: $i = 70, 80, 90$; WS: $i = 3, 5, 7$; |
| $V_{cnod;>i}$ or $V_{cnod;<i}$ | Number of consecutive days above or below a threshold | |
| $V_{acc;>i}$ or $V_{acc;<i}$ | Accumulation above or below a threshold | |
| $V_{avg}$ | Average value of the weather variable | |

We used predictive ability as a primary criterion to measure the performance of each summarising weather function. According to the settings of window pane, 136 variables were generated for a weather function based on specific settings, which is much larger than cluster-year combinations (56 observations). In the absence of sufficient observations, it is difficult to evaluate the predicative ability using ordinary linear regression. This study used high dimensional regression methodology, partial least squares regression, to address this problem. The predictive ability of a putative model is determined through a 'leave one out' cross validation (CV) procedure. In CV, the first observation of n data points is deleted and the PLSR model is fitted using the remaining n−1 observations. This model is used to predict the withheld observation and the prediction error is determined. This process is repeated for all other observations. The value for CV was compared with the standard deviation of the original data to identify the "optimal" weather function. The "optimal" weather function should have the lowest ratio, which varies between 0 and 1 (the smaller, the better). The correlation between the "optimal" generated weather variables and final disease severity was calculated. Level plots of correlation by time lags and window length were used to investigate the relationship between weather and disease severity.

Having determined the best summarising functions for the weather variables, a method was needed for combining this and identifying the key periods of influence. This project developed a new statistical approach (Stepwise – Partial Least Squares Regression) to solve this problem. This novel, high dimensional regression model could work with any number of predictors, but involving many that are unnecessary (i.e. uncorrelated or weakly correlated with response) may deteriorate the performance. To solve this, we only considered summaries of weather variables (derived earlier by PLSR) having a significant correlation ($p < 0.001$) with the response variable. All 56 combinations of clusters and years were taken together, and disease severity was analysed as a continuous response variable. Including agronomic factors as extra predictors altered some of the significance levels of other weather variables, so results are presented both from the model without agronomic factors (referred to as Model 1) and with agronomic factors (Model 2).

The ultimate aim of this research was to construct disease-forecasting models, which can provide guidance for farms to optimise their use of fungicides. Three individual predictive models before each key spray date were therefore derived, which would provide farmers with information useful for adjustment of fungicide applications. The CV statistics were also calculated for all formulated models as an additional test, to allow to objective judgement of the disease forecasting performance.

# RESULTS

## Clustering Methods

Selection of the most suitable clustering method is fundamental to obtaining meaningful results, especially when assessing a large and irregular disease dataset. To compare the performance of the clustering algorithms objectively, one decision must be made concerning the number of clusters to be retained for each method. If the chosen clustering accounts for an insufficient fraction of the variability present in the data, some clusters may be combined inappropriately. There is a clear trade-off between too few clusters and too many. The variance

curves for all hierarchical methods and recursive partitioning increase as the number of clusters also rise (Figure 5).

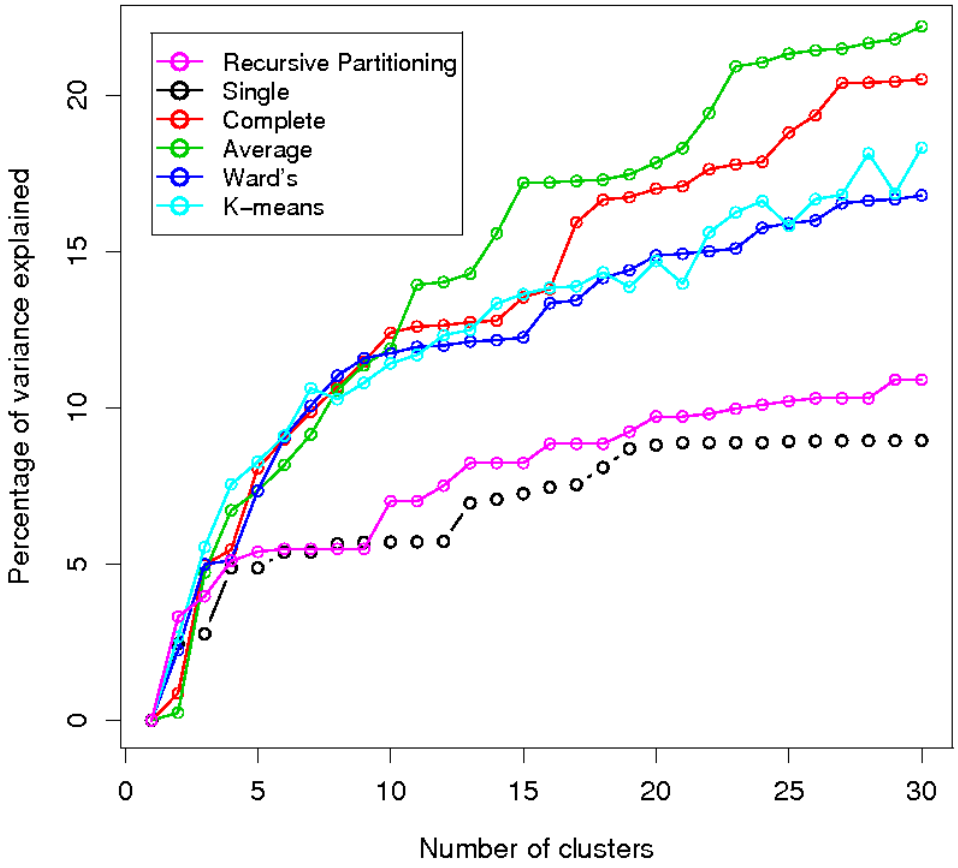


**Figure 5** The effectiveness of various cluster analysis algorithms in representing the historical disease situation for a range of cluster levels.

Based on the form of the curves in Figure 5, single linkage clustering gave the most unsatisfactory result. Although recursive partitioning generally outperformed single linkage it was still unacceptable due to the low percentage of disease variance explained. K-means and Ward's method, were found to reach a slightly higher level of accuracy at a lower number of clusters compared with other methods. The overall effectiveness of average linkage was greater than other cluster methods when more than 10 clusters were created, indicating that average linkage clustering provided a substantially better representation of the historical disease pattern. It is often possible to determine the appropriate number of clusters by visual inspection of the variance curve. For instance, the variance curve of average linkage exhibits clear breaks at 15 and 23 clusters,

while the specific clustering levels advised for complete linkage was either 10 or 18. We believe that between two and nine clusters sacrifice too much detail, and generate inadequate observations for disease modelling. However, we consider the 15 cluster classification to be suitable for evaluating the performance of the clusters for providing regionalisation of disease severity (cf. risk). To avoid repeated analysis of spatial patterns at various clustering levels, the 15 cluster solutions are presented to compare the capability of each method to accomplish 'regionalisation' (Figure 6a-f). The 15 cluster regionalisation is a consequence of both geographical distribution and historical disease pressure over the observed sample farms in England and Wales.

There is potential for problematic solutions from recursive partitioning, which produce unrealistic 'stripe like' clusters (Figure 6a). Such abnormal clusters are mainly located in the border between England and Wales. An extreme example of an algorithm-specific clustering pattern is the case of single linkage clustering (Figure 6b), which yielded very poor regionalisation as 'chaining effects' were very significant. The phenomenon of chaining refers to the tendency to incorporate intermediate points into an existing single cluster rather than initiating a new one.

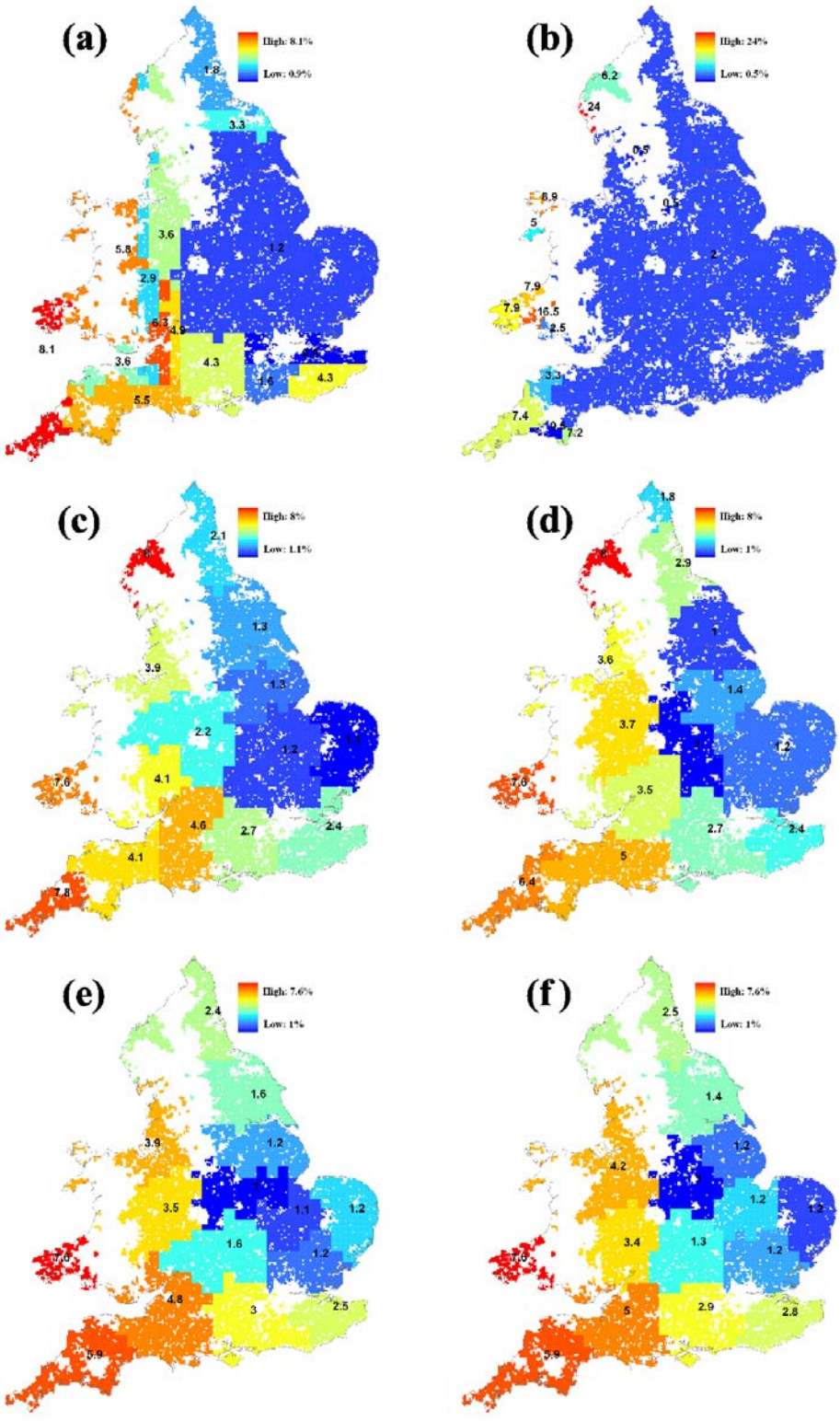


**Figure 6** Regionalisations of England and Wales for historical disease severity (91-98) on the basis of 5 clustering algorithms: (a) Recursive partitioning; (b) Single linkage; (c) Complete linkage; (d) Average linkage; (e) Ward's method; (f) K-means. The number of clusters was held fixed at 15. Each constructed cluster is labelled with the mean values of the historical disease severity on sampled crops within its boundary.

The spatial patterns from complete and average linkage exhibit much less of a chaining effect. When the sample farms are compact and well separated these two methods produced similar clusters (Figure 6c & d). As separation between sample farms decreases the likelihood of different outcomes from the clustering algorithms increases. More common situations consisting of fairly poorly inter-separated samples will have a substantial impact on the clustering outcomes. The greatest discrepancy can be found in the Eastern region. For average linkage, the Eastern region is composed principally of one spatially extensive cluster. In contrast, complete linkage identified several clusters in this region. However, subdivision of the Eastern region does not appear to be helpful to the intended purpose because the observed disease situation in this region is fairly stable and consistent. Different clustering outcomes from complete and average linkage also exist in the Midlands and central Southwest areas in terms of cluster membership size and shape. It is of interest to observe that one particularly small cluster (covering certain portions of northern Northumberland) was constructed by average linkage.

Ward's method and K-means are versatile techniques for cluster analysis. In general, the spatial patterns of Ward's method were in reasonable agreement with those obtained from K-means (Figure 6e & f). Actually, two groups of identical clusters are generated, one in the southwest of England and another in Wales. Minor differences were observed with clusters in the Midland areas. Ward's algorithm yielded patterns where two clusters (western and southern part of the Midlands) were a little larger and one (the farms located in the upper north of the Midlands) was smaller. Of most concern, both methods were shown to incorrectly place geographically separated portions of the domain into a single cluster (the north of England with mean historical severity of 2.4 and 2.5, respectively). They both failed to isolate high disease risk areas in west Cumbria with relatively small sample sizes. In addition, both algorithms subdivided the homogenous Eastern region into three spatially contiguous clusters, which tended to be of a more similar size. No major differences for the mean historical severity among these three clusters could be recognised.

Statistical tests are needed to evaluate the performance of the various cluster analysis methods. The disease situations (also measured on the top two leaves) from 1998 to 2007 were used as independent test data to examine the reliability of the constructed clusters. We were interested in comparing the effectiveness of constructed clusters in classifying the disease pattern. For this purpose, the percentage of disease variance explained for each test year was calculated and presented using a boxplot (Figure 7). The constructed clusters from the single linkage technique still had the worst performance. The variances explained were all less than 7%, indicating little correspondence to the actual disease pattern. The performances of clusters constructed by other methods were comparable but with different medians and ranges. The difference in performance of these constructed clusters ranged from 0.1% to 6.6%, depending on the test year. In at least 9 out of 10 test years average linkage clusters represent the disease pattern more accurately than other methods ( $p < 0.05$ ). Complete linkage and Ward's algorithms resulted in intermediate performance.

## Disease and weather correlation

Before deriving the different predictive models from our data set, the "optimal" weather function for each weather variable was determined. Several statistics were included to investigate the corresponding performance of each weather predictor. The predictive ability was used as the main judging criterion, while the number of significant correlations was used as an extra criterion to measure how well the disease data correlated with the weather predictors. For a genuine relationship between disease and the derived weather variable (i.e., $\mathrm{Tmin}_{avg}$ ), it is likely that many significant correlations will be found around the optimum window. If only a small number of significant correlations are identified, it is unlikely a genuine relationship for such weather predictors exists (e.g., $\mathrm{WS}_{avg}$ ). The most important summary weather variables influencing disease severity were $\mathrm{Tmin}_{avg}$ and $\mathrm{Rain}_{nod;>3}$ . Many significant correlations were observed for both of these weather predictors.

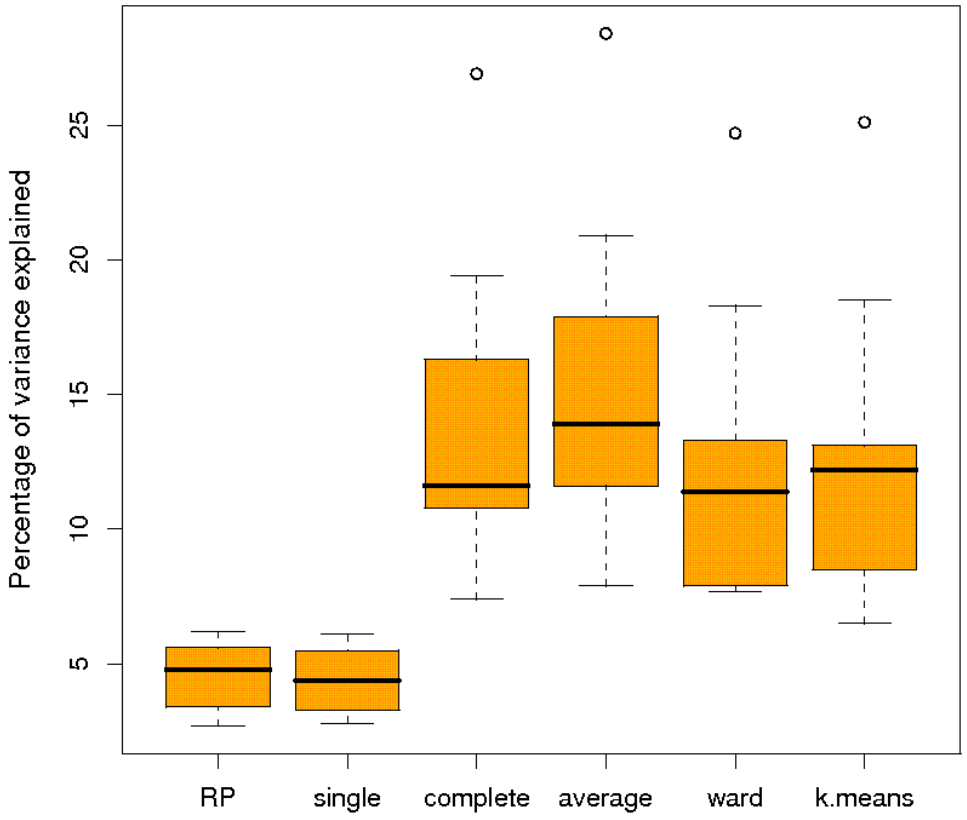


**Figure 7** The percentage of disease variance explained for each independent test year (1998–2007) with correspondence to different constructed clusters.

Further correlation analysis was implemented by screening disease severity with derived weather predictors, the pattern of correlation of each weather predictor by time lag and window length was illustrated by level plots (Figure 8). The level plots provide the best graphical representation for the effect of each individual weather predictor and allow the correlations to be investigated systematically. Positive correlations were found for weather predictors $\mathrm{Tmin}_{avg}$. The greatest correlations to final disease severity were found for the periods between early January and March. Weak correlations were observed for any minimum temperature after March. Such weather predictors will have a small impact on disease modelling. The difference in correlation between $\mathrm{Tmax}_{avg}$ and $\mathrm{Tmean}_{avg}$ is not obvious. The average maximum temperature with time lags less than 50 days was negatively correlated with disease development. Whilst the exact explanation for this is unclear, it is probably because hot conditions can be unfavourable for *S. tritici* development: high temperatures are often associated

with periods of low rainfall. The main correlations found with rain focused on the period from the beginning of May to the end of June. Significant correlations were also expected for any longer search window overlapping with this period. Similar scenarios were presented for all three relative humidity predictors ($\mathrm{Hmin}_{avg}$, $\mathrm{Hmax}_{avg}$, $\mathrm{Hmean}_{avg}$). The humidity predictors for May and June were best correlated to disease data. For weather predictors of $\mathrm{WS}_{avg}$ the majority of the correlations were positive. However only one particular window was identified with a correlation larger than 0.5, which is likely to be spurious.

## Forecasting models

Those weather predictors with a comparatively strong correlation with disease severity are those most likely to describe meteorological scenarios related to *S. tritici* infection and development. Hence, the weather predictors with highly significant correlations ($p < 0.001$) form the input to the disease predictive models developed here. To examine the overall quality of the predictive models, scatter plots were made of cross-validation estimation against disease observation (Figure 9). Derived by S-PLSR, the best equation describing the relationship between weather and disease severity at GS 75 was formed with two weather functions $\mathrm{Tmin}_{avg}$ and $\mathrm{Rain}_{nod;>3}$. Only one window (approximately February) was chosen for $\mathrm{Tmin}_{avg}$, while two important windows (approximately 20 February to 10 June and 20 April to 10 June) were defined for $\mathrm{Rain}_{nod;>3}$. The adjusted $R^2$ indicated that the weather-based predictive Model 1 can explain 68% of the variance in disease severity at GS 75.

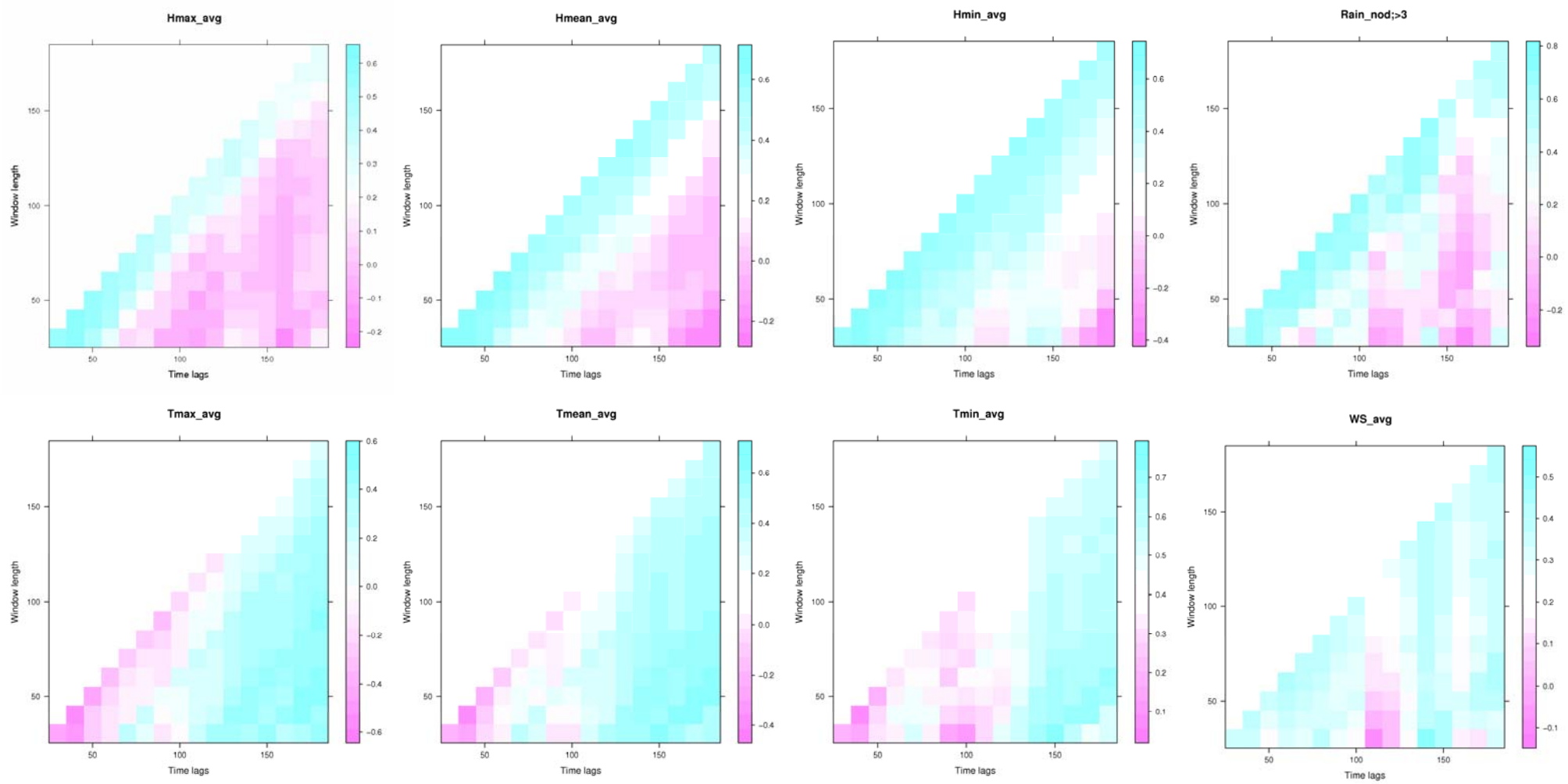


**Figure 8** The pattern of correlation for all weather predictors (the corresponding correlation range for each weather predictor is divided into 20 levels for graphical representation).

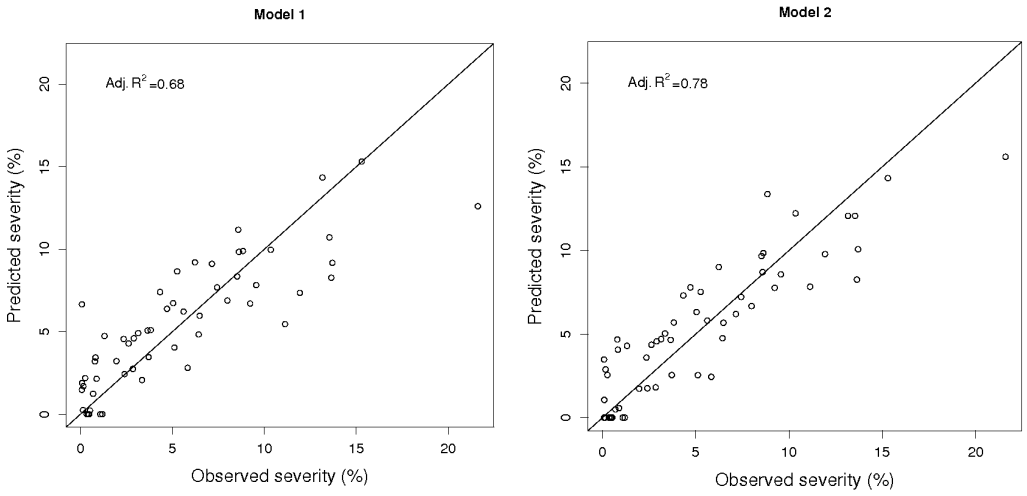


**Figure 9** Scatter plot of the cross-validation estimates against the observed severity of Septoria tritici at GS 75. Model 1 represents the regression with only meteorological predictors, and Model 2 includes the extra effect of agronomics. The Adj. $R^2$ is applied for the disease variance explained at cluster level, not representing the farm scale.

After quantifying the significant weather predictors, the next logical step is to predict disease severity more accurately with agronomic factors. For this purpose, the additional agronomic predictors were employed in a multiple linear regression analysis with the identified weather predictors. Only the agronomic factors with a significant improvement to the regression were kept in the result. Although the coefficients for weather factors were influenced by involving agronomic predictors, they balance each other and significance levels of weather factors were generally unaffected in the full Model 2. Apart from the meteorological scenarios, the estimate of final disease severity was also significantly affected by cultivar resistance ( $p < 0.01$ ) and the number of fungicide sprays ( $p < 0.01$ for a spray scheme with 3 applications). Surprisingly, no indication of a significant effect for sowing date was found. One possible explanation is that sowing date interacts with other agronomic factors, which are more influential in determining disease risk. On average, cultivar reduced disease severity by 5.51%. The effects on disease severity from various fungicide schemes were -9.29% ± 5.11 for 2 sprays, -11.72% ± 3.88 for 3 sprays, and -10.21% ± 4.6 for at least 4 spray

applications (effects are presented as mean ± standard error of the mean). Model 2 had an increased adjusted $R^2$ of 78%, which confirms the improvement of the model's predictive ability by consideration of agronomic effects. Furthermore, it allows users of the model to estimate the value of differing management scenarios they might be considering.

**Although the constructed predictive model explained much of the variability in severity of *S. tritici* at GS 75, the model fails to estimate disease risk before the final fungicide spray (T3). Therefore, empirical models were constructed using only weather variables available before each key spray timing is considered. Because of the exclusion of some important weather windows, the performance of the constructed forecasting models targeting each spray timing is certainly poorer than the full model with all significant predictors. The accuracy of the estimates also decreases with the days we intend to predict forward. The adjusted $R^2$ range from 54% to 65% (**

Table 2).

Although different models were produced for each spray timing, some similarities were shown for identified weather and agronomic predictors. $\mathrm{Tmin}_{avg}$[150,30] is constantly selected by all the models, and determined as the most significant predictor. A weather based prediction model before T1 including only minimum temperature, had an adjusted $R^2$ up to 54%. Despite only partial weather information, the number of days with rainfall above 3 mm is the most significant term for predictive models before T2 or T3. Positive relationships were identified for both models with rainfall windows summarised after late February.

**Table 2** Disease forecasting models targeting spray timing T1, T2 and T3, respectively. The results are shown for model without (Model 1) and with (Model 2) agronomic factors. $C_r$ proportion of farms with resistant cultivar; $F_3$ proportion of farms with 3 sprays.

| Spray | Model 1 & 2 | $R^2$ | Adj.$R^2$ |
|---|---|---|---|
| T3 | **1:** $2.34\text{Tmin}_{avg}[150,30]+0.26\text{Rain}_{nod;>3}[130,100]-7.55$ | .65 | .61 |
| | SE: (0.34) (0.11) (2.41) | | |
| | **2:** $2.14\text{Tmin}_{avg}[150,30]+0.41\text{Rain}_{nod;>3}[130,100]-5.55C_r-6.71F_3-4.72$ | .73 | .65 |
| | SE: (0.31) (0.10) (1.98) (2.73) (2.34) | | |
| T2 | **1:** $2.55\text{Tmin}_{avg}[150,30]+0.19\text{Rain}_{nod;>3}[130,80]-5.79$ | .62 | .58 |
| | SE: (0.33) (0.11) (2.48) | | |
| | **2:** $2.28\text{Tmin}_{avg}[150,30]+0.32\text{Rain}_{nod;>3}[130,80]-4.84C_r-6.65F_3-2.08$ | .69 | .61 |
| | SE: (0.33) (0.12) (2.14) (2.91) (2.26) | | |
| T1 | **1:** $2.69\text{Tmin}_{avg}[150,30]-1.89$ | .57 | .54 |
| | SE: (0.33) (0.98) | | |
| | **2:** $2.68\text{Tmin}_{avg}[150,30]-6.67F_3+0.49$ | .59 | .56 |
| | SE: (0.32) (3.08) (1.45) | | |

Once the models have adjusted for the large impact of weather, the true relationship between disease severity and agronomics is revealed. The proportion of resistant cultivar had a significant effect on reducing disease severity, but only shows for predictive models before T2 or T3. Most farms had received between one and four fungicide sprays. Less than 1% of farms received more than four sprays (range five to seven applications). *Septoria tritici* was not found in the two farms that received six or seven sprays. After allowing for the effects of other explanatory variables, the predicted models showed a strong decline in severity for the proportion of farms with three spray applications.

## DISCUSSION

The important weather variables, time windows and agronomic factors related to disease severity have been identified by S-PLSR. These

relationships can be used to formulate disease warning systems. We first discuss the most important weather relationships with comparisons to previously identified epidemiology and the life cycle of *S. tritici*, in order to consider whether these factors are biologically realistic. Then the agronomic effects, and finally the performance of the predictive models is considered.

## Weather Effects

TEMPERATURE

In general, *S. tritici* epidemics from 56 cluster-years evaluated in this analysis were constantly associated with a particular period of temperature. A warmer temperature was correlated with a greater disease severity, and all relationships suggest that the average minimum temperature in February is the most influential temperature factor in *S. tritici* development. The effect of temperature before January is ignored in the analysis, but is still worth some attention. A better understanding of how temperature affects latent period would be advantageous in providing guidance for disease prediction. According to correlation analysis (Figure 8), we also find an interesting relation with maximum temperature in June limiting epidemic development, although this factor is not included in the final model. This is probably because either high temperature reduces leaf wetness or periods of high temperature are correlated to periods of low rainfall.

RAINFALL

A key requirement for the development of *S. tritici* is the availability of rainfall, which aids the splash dispersal of spores on to the upper leaves. Our results are in broad agreement with past observations of rainfall associated with epidemics. In our results the number of days with rainfall above 3 mm was associated with the likelihood of subsequent serious disease severity. *Septoria tritici* is dependent on rainfall throughout the period from late February to early June with emphasis on the sub-window of the last 50 days (20 April to 10 June). This supported that rainfall shortly after the emergence of the upper leaves contributes more to disease development. Earlier precipitation events, before mid February,

lacked epidemiological significance because other meteorological variables, most likely temperature, are more limiting to infection and development during winter. The rainfall events after 10 June are still positively correlated with disease, but are not likely to contribute a lot to yield loss because by then top leaves are senescing naturally and the grain is in the soft dough stage.

RELATIVE HUMIDITY

High humidity favours the release of inoculum and facilitates infection. Conditions of continuously high humidity during an epidemic also enable greater success for the pathogen to complete successive cycles. Under controlled environment experiments, spore release is twice as rapid at 100% relative humidity as at 98%, and at 100% relative humidity, 11 to 23 times the number of spores are produced as at 86%. The correlation investigation showed that some relative humidity predictors are highly correlated with severity of *S. tritici* at GS 75. However, no significant predictor was found with relative humidity through S-PLSR, probably due to inter-correlation with rainfall variables. Relative humidity alone can increase infection risk on single leaves, but splash may be more important for an effective spread of disease from infected leaves to healthy leaves. In addition, it is rational to assume the pathogen responds to the humidity pertaining to its environment. Thus the humidity (commonly associated with leaf wetness) in proximity to the infected leaf, and not air humidity, is most likely to affect pathogen growth. It is obvious that the internal wetness of a leaf is not the same as air humidity. However, the UK meteorological network is not equipped with any direct means of measuring leaf wetness.

WIND

No significant negative correlation was found for wind speed in our analysis which immediately ruled out the possibility that wind decreases disease development through reduction of leaf wetness. Wind may also play an important role in the dispersal of *S. tritici* within a farm, affecting the abundance of inoculum. Although many positive correlations between wind and disease severity were observed, they were not strong enough to

be selected by regression modelling. Correlation with wind speed is often weak because average daily wind speed measured in a meteorological station might inaccurately represent what is actually happening in a crop canopy.

## Agronomic Effects

From the analysis presented, there was clear evidence that the proportion of resistant cultivars affected the severity of *S. tritici* epidemics, which is consistent with previous studies. The final disease severity is expected to be reduced by 5.5% if resistant cultivars were sown for an entire disease cluster. For years 1998–2001, the farms with susceptible cultivars in the south western part of England developed much greater disease severities than those planted in the eastern part of England. This difference was mainly related to the weather conditions in southeast and east England. Resistant cultivars are recommended in those high risk disease clusters to produce an immediate benefit in suppressing pathogen development and inoculum production. In addition, selection of more resistant cultivars may become more important for farmers in relation to the increasing cost of fungicides and public concerns over pesticide use.

The comparatively low average disease severity values for cluster-year reflected the widespread use of foliar fungicides on the commercial crops surveyed. Almost all commercial wheat crops (97.2%) are treated with at least one fungicide spray. The majority of the sprays (84.1%) were applied between GS 31 and GS 59, the period when fungicides are most likely to reduce disease severity on the final three leaves. Spray decisions are often made according to a precautionary strategy based on the crop's growth stage, though leaf emergence is a better guide for decisions on spray timing. The effects of different dates for spray application were not examined, as they are too complicated to be used as an alternative predictor combined with the number of sprays. According to the regression results, the number of fungicide sprays applied clearly makes a substantial contribution to the control of *S. tritici*. The spray scheme with three applications is most effective ( $p < 0.01$ ) against epidemics. The

difference was not significant between spray schemes applying three or four applications, suggesting the potential to eliminate unnecessary fungicide applications after three sprays as no substantial improvement was offered.

## Predictive models

The level plots (Figure 8) provide a systematic overview of the correlations between disease severity and summarised weather factors. Because of the iterative process used to look for correlations, some would always be found to be significantly large, even in the absence of genuine relationships. Instead of relying entirely on individual p values of correlation to select significant predictors, a qualitative selection criteria from discriminant analysis was used to distinguish genuine correlations from spurious ones. While this was a significant improvement, additional research is needed to validate the determined predictors. In this study, PLSR was first applied to solve collinearity problems caused by iterative searching of the weather variables then S-PLSR was used to further select the important weather predictors. S-PLSR was strong in extracting underlying weather predictors that account for most of the variance in observed disease severity. It only accepted the relationships that were strong enough to be reliable through a model selection process. In addition, it also quantifies weather predictors with overlapping windows (e.g. $\mathrm{Rain}_{nod;>3}$ [130,110] and $\mathrm{Rain}_{nod;>3}$ [70,50]). For interpretation of the result, it emphasises that rainfall for specific periods will play a different role in disease development and infection.

**The applications of constructed predictive models, like many weather-driven prediction systems, driven prediction systems, will depend on the availability, resolution and reliability of weather reliability of weather data. The full predictive model that combined weather and agronomic weather and agronomic factors had a comparatively high prediction accuracy of 78%. accuracy of 78%. Variability in disease assessment, unmeasured favourable weather events or favourable weather events or influence of cluster-scale environment and initial inoculum levels initial inoculum levels may account in part for the other unexplained disease variance. The disease variance. The spatial-temporal pattern of disease severity at cluster level is successfully**

**cluster level is successfully accounted for by significant weather and agronomic factors. In the regression results, the identification of appropriate searching windows for weather variables is determined by calendar date, which is easy to adopt by farmers. A potential limitation of the full predictive model was the dependence on weather information after most routine fungicide applications have been made. The period of observation before a spray is too short to include all meteorological scenarios that were, in fact, associated with disease risk. While it may be possible to overcome this limitation by using weather forecasts, the uncertainty of the predicted weather variables may reduce model prediction accuracy dramatically. In contrast, regression models using weather information prior to each of three key spray applications (T1, T2 and T3) were constructed. They had 13% − 24% lower prediction accuracy compared with the full predictive model, but were free from potential limitations associated with using forecasted weather variables. Attempts to incorporate agronomic factors that had potential to improve prediction accuracy, such as cultivar resistance and number of fungicide applications (**

Table 2), did improve the accuracy of models for T2 and T3 (no noticeable effect of sowing date was found).

The direct application of these models is to guide fungicide treatment decisions on a cluster basis to protect the upper leaves that are crucial to yield information. It provides evidence to farmers that can inform the timing and dose of their fungicide applications and eliminate unnecessary within-season sprays. The optimum T1 spray will give maximum disease control on leaf 3, and provide some protection for leaf 2. Triggered by warm winter temperatures, especially during the February period, an increased dose of active ingredient is recommended for T1 to provide long-term disease control and to reduce the reliance on curitative efficacies from subsequent treatments. T2 is the most important spray, and it gives maximum disease control on the flag leaf and eradicates latent infections on leaf 2 that have escaped earlier sprays, assuming the flag leaf has emerged fully by the time of fungicide application. Disease prediction could be updated with models for T2 as more real-time weather information becomes available after the T1 spray. If frequent and heavy rainfall splash occurs from late February, the application of a robust T2 spray is necessary with the intention to avoid the danger of favourable weather leading to widespread disease. T3 spray targeting ear diseases,

also gives additional control of disease on the top two leaves, especially important under high disease pressure. However, in resistant cultivars, an ear spray may not be necessary. T3 sprays may be adopted according to the last meteorological event, especially in May. With long periods of dry conditions, the use of T3 fungicide could be saved or a reduced spray dose applied without subsequent treatments.

## CONCLUSIONS

The project describes preliminary work on a new approach to define disease clusters based on cluster analysis. Of all algorithms compared, the constructed 15 clusters from average linkage are identified as the most appropriate aggregation scale for disease severity. Accurate definition of disease clusters is a necessary first step to exploration of the relationship between disease development and environmental factors. The identified disease clusters have been implemented in conjunction with disease analysis and modelling. A reasonable disease forecasting scheme has been established and made available for use.

Despite satisfactory performance of the constructed predictive models, there is still scope for improvements. One limitation is that predictive models constructed from regression surfaces assume a linear relationship between the observed data and the data to be predicted. In principle, the weather predictors could be transformed with nonlinear functions being employed in the regression. However, it is difficult to find suitable transformations with obvious mechanistic explanation to build the weather predictors. Given the importance of the initial inoculum level in disease development, it may be possible to improve model performance through additional variables containing information about local inoculum sources (disease lesions on lower leaves are the most common source of inoculum for upper leaves emerging during spring and summer). Unfortunately, this information was unavailable in this analysis. Although inferring farm level relationships from grouped data is also important in practical disease management (because the farm is the object of study and hence the target of inference), the work presented here is not attempting to further

distinguish individual-level effects from cluster-level effects. To achieve accurate site-specific forecasts, it is important to improve the accuracy and reliability of weather variables within farms, which need to be addressed in future analyses.

## KNOWLEDGE TRANSFER

Due to their satisfactory performance, the disease forecasting models in this report are implemented and available at the Crop Monitor (http://www.cropmonitor.co.uk) web site. The Septoria warning page shows predicted mean, maximum and minimum severities for each cluster (when clicked) and gives an indication of the reliability of the predictions based on the number of weather stations available to the analyses (Figure 10).

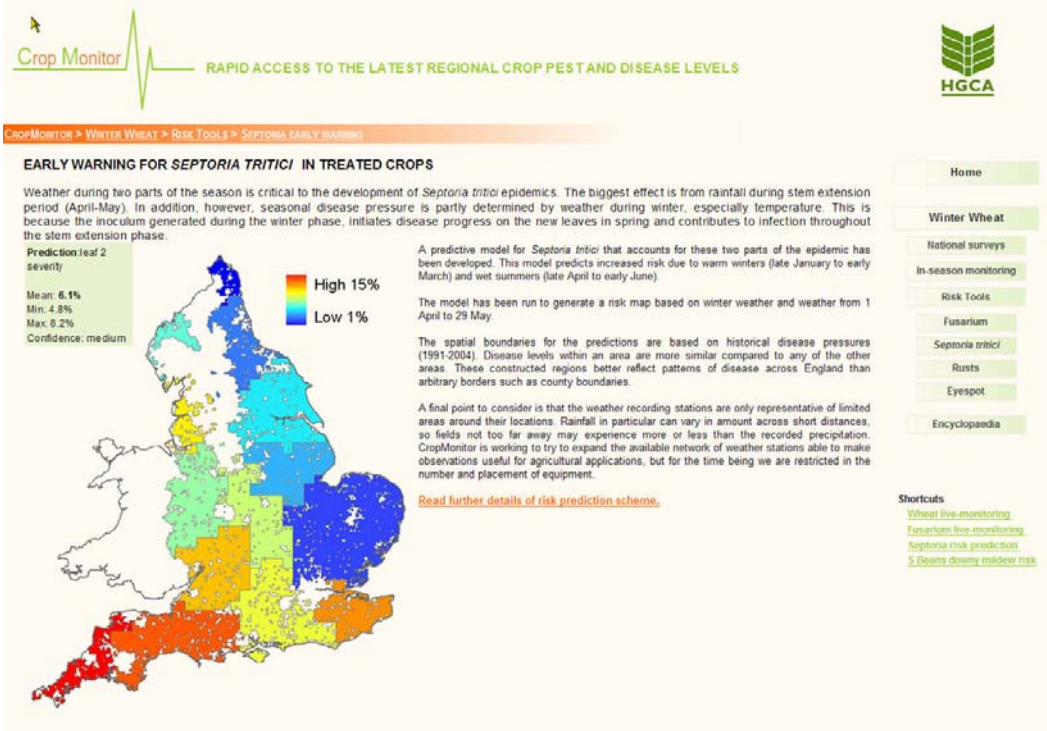


**Figure 10** Interactive web page illustrating the S. tritici warnings for the 2008 harvest season.