Final Report

Project No. 11120032

Spatiotemporal analyses of potato late blight outbreaks in Great Britain

James Hutton Institute

Report Authors:
Matt Aitkenhead
Malcolm Coull
Mads Troldberg
David Cooke
Peter Skelsey

Table of Contents

Summary

The GB potato industry requires high quality, robust data on the spatial and spatiotemporal risks of late blight to benefit decision-making and long-term strategic planning. In this project, data from late blight outbreak sampling by the AHDB Potatoes Fight Against Blight programme (2003–2018) were analysed together with environmental variables to provide new information on the spatial epidemiology of late blight in GB. The results were used to produce visual aids to facilitate improved decision-making.

Late blight severity is highly dependent on the local weather, so the timing, number and spatial pattern of sampled outbreaks varied from one season to the next. To reveal trends, the outbreak data collected over 16 seasons were analysed using the ESRI ArcGIS Pro 2.4 geospatial analysis platform. This revealed statistically significant patterns in space and time for early outbreaks, overall incidence, and the distribution of various pathogen genotypes. Together, these analyses provide valuable information on the variable risk posed by late blight across the potato production areas of GB.

ArcGIS was also used to calculate the velocity of spatial spread for two recently emerging, aggressive clones in GB (36_A2 and 37_A2). The results provide hitherto unseen information on the speed at which newly introduced clones can spread across a country.

Machine learning techniques (neural network models) were used to analyse weather, soil, geology, and topography data to identify the principle factors associated with late blight occurrence.

The key findings were:

Early outbreaks

- *Spatial analyses*: Averaged over all years, the statistically significant hot spots (clusters of high incidence) of early outbreaks of disease were generally found in the south of England and Wales, particularly near the coast, whereas a large cold spot (clusters of low incidence) extends across the production regions of Scotland. This supports the role of climate in the earliest occurrence of outbreaks.
- *Spacetime analyses*: When analysed as a time series, the spatial hot spots of early disease tended to be sporadic: i.e. locations that are on-again off-again hot spots. There were also some sporadic and consecutive hot spots (a run of recent hot spots) of early outbreaks in the Angus / Aberdeenshire regions. This may relate to a warming climate resulting in an increased frequency of early outbreaks in Scotland in recent years.
- *Driving factors*: As expected, the date of first outbreak was later in the north of GB, but only by a matter of weeks. Machine learning was used to develop a model that was 91.2% accurate in predicting low and high levels of incidence of early outbreaks. It identified temperature and precipitation as the most important predictors of early outbreak incidence.
- *Visual aids for decision-making*: Colour-coded maps were produced to show the overall risk of early outbreaks by postcode district, and the week of the year these were most likely to occur.

Spatial spread of disease across the whole season

- Risk of spread of disease among neighbouring postcode districts was highest in the potato growing regions of Tayside, Fife, Lothian, and East Anglia.
- The velocity of spatial spread was calculated from early foci of genotypes 36_A2 and 37_A2 and ranged from 3–17 km per week.

- *Visual aids for decision-making*: A colour-coded map was produced to show the risk of spread of disease among postcode districts.

Overall incidence, 2003-2018

- *Spatial analyses*: Statistically significant hot spots of sampled late blight incidence were found in the Angus, Tayside and Fife regions of Scotland, and in East Anglia, Kent and East Sussex in England. No cold spots were identified.
- *Spacetime analyses*: There were three types of temporal trend identified in the hot spots of incidence: consecutive, sporadic, and new (appearing in the final year). The lack of any persistent hot spots is a consequence of large inter-annual variation in the distribution of disease.
- *Driving factors*: A model was developed that can estimate outbreak risk based on environmental factors. The model revealed a strong positive relationship between density of potato cropping and late blight outbreak risk. Weather also had a strong impact on outbreak risk, particularly temperature, humidity, rainfall and windspeed. Topography (elevation, slope, aspect) had a large impact on outbreak risk, and there was also evidence of an association with soil conditions and geological type.
- *Visual aids for decision-making*: A colour-coded map was produced to show the overall risk of late blight by postcode district.

Pathogen genotypes

- *Spatial analyses*: The mean spatial patterns of the genotypes 13_A2, 6_A1, 8_A1, 37_A2, and 36_A2 were analysed and differences in the central tendency, dispersion, and directional trends of these distributions were observed. The pattern of hot and cold spots varied markedly for each genotype, and some opposing patterns suggested competition and displacement.
- *Spacetime analyses*: Most of the spatial clusters of the pathogen tended to be consecutive hot spots (a run of recent hot spots). The lack of any persistent (long-term) hot spots or cold spots of any type was due to a large degree of inter-annual variation in genotype distributions. Several new hot spots (hot spots in the final year of analysis) were found for genotypes 36_A2 and 37_A2, in line with their recent invasive spread.
- *Driving factors*: A model was developed to predict the dominant genotype in each postcode district. It identified precipitation and humidity as the most important predictors, suggesting moisture plays an important role in competition among genotypes. The emergence and rapid spread of genotypes 13_A2 and 6_A1 in 2006-2007 was clear but there were few clear patterns to their distribution in the subsequent years. One obvious feature was the local spread of 6_A1 in eastern Scotland in 2011 followed by its dominance in the subsequent years. This indicates the significance of local sources of primary inoculum carried over from the previous season. However, the drivers of the large variation in other genotype distributions between years were not clear and require further investigation.
- *Visual aids for decision-making*: A video was produced to show the changing pattern of genotype distributions each year.

Introduction

Although current methods of late blight management are effective, there is clear scope for improvement as there remains a heavy reliance on calendar spray regimes and challenges relating to the timing of the first spray and the position in the spray programme for the chemistry that offers ‘premium’ blight control. The primary aim of the proposed research is to

empower agronomists and growers with a new understanding of the epidemiology of late blight in GB, and new visual aids to facilitate improved decision-making:

Agronomists and growers need to know if their crops are at risk from early outbreaks of late blight due to factors such as: carryover of inoculum (intense blight pressure) from the previous season, intensity of potato production, and the presence of aggressive pathogen lineages in their own and surrounding postal districts.

Agronomists and growers need to know if late blight is likely to spread from neighbouring postal districts to theirs, and how quickly.

Agronomists and growers would benefit from an exact quantification of the historical risk of late blight occurrence in their own *and* neighbouring production areas to better plan for the growing season ahead.

The industry as a whole needs to understand the drivers of past change in the GB blight population in order to predict changes in the distribution of new aggressive lineages

Previous AHDB Potatoes-funded work, using samples submitted by Fight Against Blight (FAB) scouts from 2006 to 2018 had shown that British populations of the potato late blight pathogen, *Phytophthora infestans*, had changed markedly over time (Cooke et al., 2009; Cooke et al., 2013; Cooke, 2019). Subsequent AHDB Potatoes-funded work that examined this change using controlled environment experiments to investigate differences among the dominant genotypes in their response to weather conditions (Chapman 2012), although valuable, yielded conflicting results. No single genotype consistently showed dominance in terms of weather-dependent infection criteria, or in terms of competition when multiple isolates infected a plant at the same time. In further AHDB Potatoes-funded work (Dancey et al., 2017; Dancey, 2018;R473 Late Blight Models), a different approach was adopted whereby experimental work was combined with mapping and statistical modelling of existing AHDB Potatoes FAB late blight outbreak data to develop a new national warning system for late blight; the Hutton Criteria. This approach proved successful, leading to significant improvements in the performance of the national warning system, whereas previously the Smith Period showed great spatial variation in predictive power and performed poorly in some regions. The Hutton Criteria, currently deployed by BlightWatch (https://blightwatch.co.uk/) plays an important role in determining when to start the blight management programme. Although optimal fungicide application timing is important throughout the season, the early sprays are especially critical. Most fungicide chemistry relies on a protectant effect and managing an established infection is thus extremely challenging (Cooke et al., 2011). In this project, we build upon the success of this mapping and modelling work, and extend our analyses of these existing data to answer key epidemiological questions, derive new epidemiological parameters, and produce visual aids that will support informed decision making and long-term strategic management of late blight, and fungicide and resistance resources.

All infection obviously requires propagules of primary inoculum which, in the case of *P. infestans*, may be asexual or sexual spores carried over from the previous season. Previous work has shown the domination of clonal pathogen lineages within GB potato crops indicating that asexual inoculum carried over on infected tubers is the most prevalent source (Cooke, 2019). Potato seed, volunteer plants in neighbouring fields and plants growing on discard piles are the three main sources of such infection and their management is a key part of blight control strategies (Cooke et al., 2011). The objectives of this project were:

Determine whether the pattern of the earliest outbreaks of disease within each growing season is random, regular, or aggregated, and the spatial scales at which these patterns occur.

Identify geographic areas where late blight persistently occurs early in the season. Identify the principle factors associated with early outbreaks of disease.

Track the spread of late blight between postal districts over the course of each growing season to determine whether the pattern of outbreaks is random, regular, or aggregated, and the spatial scales at which these patterns occur. Determine the likelihood of spread of late blight among the postal districts of GB. Derive the rate of spatial spread across the landscape.

Identify geographic areas where incidence is consistently low or high over multiple growing seasons. Identify the principle factors associated with persistently low and high incidence of disease. Determine the relative risk of late blight occurrence within each postal district.

Track the change in the spatial distribution of pathogen genotypes over the duration of the study period. Identify the principle drivers of this change.

Methods

Datasets

The late blight outbreak data spanned a 16-year period (2003–2018) and consisted of the date, coordinates (UK postcode district centroid), potato variety, pathogen genotype and 'stage of outbreak' of 2518 late blight outbreaks sampled from across GB. These data are collected routinely each year by blight scouts as part of the Agriculture and Horticulture Development Board (AHDB) Potatoes 'Fight Against Blight' service that has been surveying and reporting on late blight incidence since 2003. In the UK, there are 127 postcode areas (AB, AL, B etc.) and around 3000 postcode districts (AB10, AB11, etc.).

Data on the spatial distribution of potato cropping throughout GB was also provided by AHDB. This provided information on the density and spatial distribution of the host crop in each year. Foliar and tuber blight resistance data for potato varieties were obtained from the AHDB Variety database (http://varieties.ahdb.org.uk/) and integrated with the crop data.

Hourly weather data corresponding to every outbreak location were provided by the UK Met. Office (UKMO) (2011−2018).

Spatial datasets of a large number of soil, geology, and topography (elevation, slope, aspect) variables were generated at a 100 metre resolution. These data were acquired from multiple sources (FAO WRB global soil dataset, OneGeology national geological maps, Shuttle Radar Topography mission global elevation dataset), all of which are available freely online. These latter datasets were acquired for other work prior to this project and had already been organised into data formats suitable for this project.

Data pre-processing

Preparation and integration of the datasets was necessary for the work to be carried out. Primary copies of all datasets were saved to a secure network location, then local copies were saved prior to the work.

The AHDB FAB data was standardised across the multiple variables included, with variety names and date formats corrected. Observations made over multiple years had resulted in a dataset that required a significant effort to clean, but which has provided us (and AHDB in future) with an optimised dataset that is of greater utility in the long-term.

Cleaning of the outbreak data was carried out using a combination of methods:

- Individual Excel worksheets were saved to separate text files where regular expressions could be used to standardise data fields, particularly date fields which were sometime recorded as DD-MMM-YYYY format, and other times as YYYY-MMM-DD, and where spaces/non printing characters/special characters had to be removed.
- For years where grid references were present, data was loaded into GIS software (ArcGIS) and plotted spatially; data points with invalid grid references were removed (no effort was made to investigate cause of miscoding due to size of dataset/time constraints).

Files were then imported into a database (PostgreSQL) as individual tables, and SQL code used to select out a consistent set of data fields into a single 'master' table. These data were then exported back to the network location where other members of the project team could use them for analysis and modelling.

UKMO Weather station data was provided with a single file for each day across all weather stations. This data was converted into a 'one file per weather station' format so that time-series weather for each outbreak/planting location could be determined.

Although data preparation and pre-processing took a significant amount of time, the project progressed rapidly with the analyses once this was complete.

ArcGIS analyses

We used ArcGIS Pro 2.4 (Esri (UK) Ltd, Aylesbury) and its geoprocessing tools to analyse space and spacetime patterns of late blight incidence. All coordinates and map outlines are shown projected to the British National Grid, with measurement units in metres. Analyses were performed on the outbreak data as a whole (total incidence) and various subsets of the data: the 10^{th} and 20^{th} percentiles of late blight outbreaks by reporting date (i.e. early outbreaks, henceforth referred to as early10 and early20), and outbreaks corresponding to pathogen genotypes 13_A2, 6_A1, 8_A1, 36_A2, and 37_A2. Spatial patterns were evaluated using choropleth maps (colour-coded maps showing the count per postcode district), the Optimized Hot Spot Analysis tool, and the Kernel Density tool. Risk of spatial spread of disease was evaluated using the Optimized Outlier Analysis tool, and the velocity of spatial spread using the Standard Distance tool. Spacetime patterns were characterised using the Emerging Hot Spot Analysis tool.

Optimized Hot Spot Analysis (OHSA)

The OHSA tool from the ArcGIS Spatial Statistics Toolbox was used to identify statistically significant spatial clusters of high (hot spots) and low incidence (cold spots) in various subsets of the late blight outbreak data: all outbreaks, early outbreaks, and genotypes 13_A2, 6_A1, 8_A1, 36_A2, and 37_A2. Data from the entire study period (2003–2018) were analysed together to produce maps showing the overall risk across GB.

The tool identifies statistically significant hot- and cold spots using the Getis-Ord Gi* statistic. The outputs of OHSA are z-scores (Gi*) and p-values. High z-scores indicate statistically significant spatial hot spots, and low z-scores indicate cold spots. The p-value is the probability that a random process formed the observed spatial pattern. When p-values are smaller than the required level of significance, the null hypothesis, which is complete spatial randomness, can be rejected.

The OHSA can be used to show where postcodes with a high or low number of sampled outbreaks cluster spatially. A postcode with a high number of sampled blight outbreaks is interesting but may not be a statistically significant hot spot. For a postcode to be a statistically

significant hot spot, the postcode will have a high number of observations and be surrounded by other postcodes that also have high values. The local sum of observations for a postcode and its neighbouring postcodes is compared proportionally to the sum for all postcodes. If the local sum is very different from the expected local sum, and if that difference is too large to be the result of random chance, it is considered statistically significant. Postcodes with large positive and statistically significant local Gi* values are hot spots, whereas postcodes with large negative and statistically significant values are cold spots.

The OHSA tool interrogates the data in order to determine settings that will produce optimal hot spot analysis results; it automatically aggregates incident data into weighted features, identifies an appropriate scale of analysis, and corrects for both multiple testing and spatial dependence. The output maps produced when performing OHSA with point data show every late blight outbreak location in the FAB dataset, with individual points colour-coded according to the level of clustering for the subset of data under analysis. Note that the OHSA for total incidence (all outbreaks) was performed by aggregating the point data into counts per postcode district containing >1ha of potato grown commercially. This was necessary as the tool requires variation in the values under analysis, i.e., if all points are being analysed, they must be aggregated into counts or analysed with respect to some other covariate.

Kernel Density Estimation (KDE)

Kernel density estimation (KDE) was used to provide a smooth geographic interpolation of the distributions of genotypes 13_A2, 6_A1, 8_A1, 36_A2, and 37_A2 in various growing seasons (2006-2017). The sampling intensity was insufficient to include 2003-2005. The resultant maps facilitate a visual analysis of change in the spatial distribution of pathogen genotypes from one growing season to the next and highlight areas where competition among genotypes, or other factors, results in a change in the predominant lineage.

KDE is a well-established non-parametric method of estimating the probability density function (PDF) of a finite dataset. It is non-parametric because it does not assume any underlying distribution for the data. KDE calculates the density of events around each observation by assigning a kernel function to every datum that weights the distances to other points in the feature space. The result is a smoothly tapered surface fit to each point. Each kernel has a bandwidth (smoothing) parameter that controls the size of the neighborhood around each datum. Larger values produce a smoother, more generalized density raster whereas smaller values produce a raster that shows more detail. We used a value of 75 km for all KDE analyses to produce smoother, generalized patterns. The PDF is then produced by summing the local contributions of the kernels and dividing by the number of observations to ensure that it satisfies the required properties of a PDF.

Optimized Outlier Analysis (OOA)

This tool also identifies statistically significant hot spots and cold spots, but unlike the OHSA it also identifies high and low outliers within the data. Clustering is assessed using a different statistic than the OHSA, namely Anselin Local Moran's I index, and z-scores and p-values are produced as described above. Whereas the OHSA identifies three types of geographical class, i.e. hot spots, cold spots, and insignificant spots, this analysis identifies five types of geographical class. On the one hand, it identifies postcode districts that have either high or low values of incidence in concordance with their surroundings (high-high, low-low). On the other, the analysis identifies anomalous areas where a postcode district has a value that is very different from its neighbours, whether much higher (high-low) or lower (low-high). There are also cases where no associations can be made. The OOA was performed on the entire dataset (all outbreaks, 2003-2018) using postcode districts as the basic geographic unit.

Postcode districts containing no commercial crops were excluded from the analyses to confine results to potato production areas only. The resultant map is used to assess the overall risk of spatial spread of late blight among postcode districts containing >1ha of potato grown commercially.

Standard Distance and Directional Distribution tools

The standard distance is a useful statistic as it measures the compactness of a distribution and provides a single value representing the dispersion of features (outbreaks) around the mean centre of the distribution. The value is a distance, so the compactness of a set of outbreaks can be represented on a map by drawing a circle with the radius equal to the standard distance value. The user can specify the desired radius (in standard deviations) of the standard distance circle. We used three standard deviation polygons so that the circles would cover approximately 99 percent of the outbreaks.

In this study the standard distance is used to calculate a proxy measure for the velocity of spatial spread for the two newly introduced clones 36_A2 and 37_A2. These were chosen as they are of great concern to the potato industry due to their aggressiveness (36_A2) and insensitivity to fluazinam (37_A2). Furthermore, there is a defined focal type spread over the last few growing seasons, as opposed to other genotypes that are well established and widely dispersed across GB. This allowed us to track their initial appearance and subsequent recent expansion into surrounding potato areas.

The difference in standard distance values (three standard deviation circle) for the distribution of outbreaks at the beginning and end of a growing season, divided by the time elapsed, was used as a proxy measure for the velocity of spatial spread:

$$vel = \frac{SD_{\text{last}} - SD_{\text{first}}}{t_2 - t_1} = \text{km wk}^{-1}$$

where *SD* is standard distance, and *t* is time in weeks. Note that the Standard Distance tool requires a minimum of three points, therefore SD_{first} and t_1 were calculated using the first three outbreaks reported, and SD_{last} and t_2 using the final distribution of outbreaks for that growing season. Further note that *SD* is the radius of a circular polygon that is three standard deviations of the standard distance circle, thereby covering approximately 99 percent of the outbreaks in each distribution.

A similar tool called the Directional Distribution (Standard Deviational Ellipse) was used to summarise the central tendency, dispersion, and directional trends of the distributions of genotypes 13_A2, 6_A1, 8_A1, 36_A2, and 37_A2 over the whole study period. It differs from the Standard Distance tool in that it calculates the standard distance separately in the x- and y-directions. This produces an elliptical as opposed to circular polygon that allows you to see if the distribution of outbreaks is elongated and hence has a particular orientation.

Emerging Hot Spot Analysis (EHSA)

The emerging hot spot analysis (EHSA) is similar to the OHSA but is used to identify temporal trends in the clustering of incidence (i.e., spacetime patterns). It finds new, intensifying, diminishing, and sporadic hot and cold spots. First, the outbreak data were organised into a spacetime cube (STC). The STC aggregates the data into location and time step bins. For the work here, the location bins were defined by the postcode districts (polygons containing >1ha of potato grown commercially), whereas the time step was chosen as 1 year. This means that each cell in the STC represents the number of sampled outbreaks (or outbreaks by genotype) for a given postcode in each year. As there are late blight observation data from 2003-2018

(i.e. for 16 years) and there are 2736 postcode districts in the UK, the total number of bins in the STC is 43776.

Whereas the OHSA automatically creates weighted features from incident counts, here the weights depend on how the spatial and temporal relationships are defined. The parameter values for Neighbourhood Distance and Neighbourhood Time Step define the extent of each bin's neighbourhood in space and time. For this study, a fixed neighbourhood distance band and a fixed neighbourhood time step were assumed. The fixed distance band values were calculated using the same method employed by the OHSA tool. First, the Incremental Spatial Autocorrelation tool was used to perform the Global Moran's I statistic for a series of increasing distances, measuring the intensity of spatial clustering for each distance. At some particular distance the intensity of clustering typically peaks, reflecting the distance where the spatial processes promoting clustering are most pronounced. If no peak distance were found, the spatial distribution of outbreaks was analysed to compute the average distance that would yield *K* neighbours. *K* was computed as 0.05 * *N*, where *N* is the number of outbreaks being analysed. *K* was adjusted so that it was never smaller than three or larger than 30. If the average distance that would yield *K* neighbours exceeded one standard distance, the scale of analysis was set to one standard distance. This produced distance bands of 36.9 km for early10, early20, total incidence, and genotypes 13_A2, 36_A2 and 37_A2. The distance bands for 6_A1 and 8_A1 were both 44.8 km. Although the primary interest was in the temporal trend of incidence clustering between growing seasons (a 1 year time step), preliminary analyses showed that a neighbourhood time step of 2 years gave superior results for all outbreaks, early10, early20, and genotype 6_A1. This was due to large inter-annual variation in outbreak distributions across GB, e.g., a 'good blight year' followed by a 'bad blight year'. A neighbourhood time step of 6 years was used for genotypes 8_A1 and 13_A2, due to the relatively low frequency of 8_A1 throughout the study period, and of 13_A2 in the latter half of the dataset. A neighbourhood time step of 3 months was used for genotypes 36_A2 and 37_A2, as their outbreak distributions were relatively concentrated in space and time.

Once the STC is created and the spatial and temporal relationships defined, the EHSA is performed using a combination of two statistical measures: (1) the Getis-Ord G_i^* statistic is used to evaluate the location and degree of spatial clustering (similar to the OHSA); and (2) the Mann-Kendall test evaluates the temporal trend in that clustering over time.

First, the Getis-Ord G_i^* statistic is calculated for each spacetime bin. Note that whereas the hot spot analysis above only considered spatial neighbours for the calculation of the G_i^*, the emerging hot spot analysis considers neighbouring bins in both space and time when calculating the G_i^* statistic. The computed G_i^* statistic can again be interpreted as Z-scores and p-values for each bin that tell you whether the number of outbreaks in a given bin is statistically clustered compared to the number of outbreaks in the neighbouring bins (in space and time). Positive Z-scores above 1.96 correspond to statistically significant hot spots and negative z-scores below -1.96 correspond to statistically significant cold spots.

Secondly, the Mann-Kendall statistic is used for each *location with data* to test whether a statistically significant temporal trend (and what type of temporal trend) exists over the whole 16-year (2003-2018) time series of z-scores from the Getis-Ord calculation above. Based on the resulting temporal trend z-scores and p-values, and the hot spot z-score and p-value for each bin, each postcode district is categorised as follows (in each case, the Hot (Cold) notation means it refers to either type of spot):

- *No Pattern Detected*: Does not fall into any of the hot or cold spot patterns defined below.

- *New Hot (Cold) Spot*: A location that is a statistically significant hot (cold) spot for the final time step and has never been a statistically significant hot (cold) spot before.
- *Consecutive Hot (Cold) Spot*: A location with a single uninterrupted run of statistically significant hot (cold) spot bins in the final time-step intervals. The location has never been a statistically significant hot (cold) spot prior to the final hot (cold) spot run and less than ninety percent of all bins are statistically significant hot spots.
- *Intensifying Hot (cold) Spot*: A location that has been a statistically significant hot (cold) spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering of high (cold) counts in each time step is increasing overall and that increase is statistically significant.
- *Persistent Hot (Cold) Spot*: A location that has been a statistically significant hot (cold) spot for ninety percent of the time-step intervals with no discernible trend indicating an increase or decrease in the intensity of clustering over time.
- *Diminishing Hot (Cold) Spot*: A location that has been a statistically significant hot (cold) spot for ninety percent of the time-step intervals, including the final time step. In addition, the intensity of clustering in each time step is decreasing overall and that decrease is statistically significant.
- *Sporadic Hot (Cold) Spot*: A location that is an on-again then off-again hot (cold) spot. Less than ninety percent of the time-step intervals have been statistically significant hot (cold) spots and none of the time-step intervals have been statistically significant cold (hot) spots.
- *Oscillating Hot (Cold) Spot*: A statistically significant hot (cold) spot for the final time-step interval that has a history of also being a statistically significant cold (hot) spot during a prior time step. Less than ninety percent of the time-step intervals have been statistically significant hot (cold) spots.
- *Historical Hot (Cold) Spot*: The most recent time period is not hot (cold), but at least ninety percent of the time-step intervals have been statistically significant hot (cold) spots.

An EHSA was performed on the entire dataset, early outbreaks, and outbreaks corresponding to genotypes 13_A2, 6_A1, 8_A1, 36_A2, and 37_A2.

Modelling

Model analysis of early outbreaks

A suite of 24 machine learning classification techniques were developed and tested for their ability to predict early outbreaks of disease, using MATLAB Version R2020a. The goal was to develop an accurate model and use it to identify the most important variables for prediction, i.e., the principle driving factors associated with the occurrence of early outbreaks. Classification is a technique where observational data are categorised into a given number of classes, and the main goal is to identify the category/class of new, unseen data. Classification algorithms learn from the input values given for training and automatically generate a model to predict the class labels/categories for the new data. The suite of 24 algorithms included numerous types of Decision Tree, Support Vector Machines, Naïve Bayes algorithms, K-Nearest Neighbour classifiers, and various ensemble methods, e.g., Bagged Trees and Boosted Trees. A description of these techniques is beyond the scope of this report, but the

interested reader can download an excellent reference ebook entitled “Machine Learning Yearning” by Andrew Ng (https://www.deeplearning.ai/machine-learning-yearning/)

The early outbreak class labels were provided by the Optimised Hot Spot Analysis (OHSA) described above. OHSA was used to categorise each early outbreak (and other subsets of the FAB data) as belonging to one of seven different classes of spatial cluster: -3 (Cold Spot – 99% Confidence), -2 (Cold Spot – 95% Confidence), -1 (Cold Spot – 90% Confidence), 0 (non-significant), 1 (Hot Spot – 90% Confidence), 2 (Hot Spot – 95% Confidence) or 3 (Hot Spot – 99% Confidence). The model inputs were weather data derived from the HADUK-Grid dataset of key UK climate variables (Hollis, et al. 2019). The variables temperature, relative humidity, sunshine duration (hours), wind speed and precipitation at 1km resolution were extracted and averaged over the potato growing season (May 1-October 31) and then averaged again over the duration of the study period (2003-2018). The mean value of each variable was then calculated per postcode district. The aim was therefore to develop a machine learning algorithm that can predict patterns of low or high incidence of early outbreaks from geographic variation in climate.

The data was split into training (80%) and test (20%) data using stratified sampling. Bayesian optimization was combined with a 5-fold cross-validation technique to train and tune the models. The optimal (tuned) models were then tested for their ability to predict the classes of the (unseen) held-out test data. Only the results of the most accurate algorithm are reported. The importance of each weather variable for prediction was estimated using the MATLAB procedure PREDICTORIMPORTANCE.

Model analysis of total incidence

A neural network model was developed to estimate risk of blight under different conditions. This kind of model can be used to estimate local environmental conditions based on spatially variable factors such as topography and climate (e.g. Aitkenhead et al., 2013; Aitkenhead & Coull, 2016; Aitkenhead & Coull, 2019). The model inputs included the weather, topographic, soil and geological data along with planting and local outbreak data. The model output was a single value ranging from 0 (no outbreak) to 1 (outbreak). ‘No outbreak’ data was produced using the planting data examples where no outbreak occurred, with random points during the growing season selected to provide a range of weather conditions. This was done because to train a predictive model of any kind, negative examples are needed in addition to positive ones (otherwise the model simply assumes that an outbreak will always occur).

Training was carried out using the backpropagation neural network approach, by which the model is repeatedly given randomly selected examples from the training dataset, and its output compared to the actual outbreak value (0 or 1) for that example. An error minimisation approach adjusted network connection weights based on the difference between ‘target’ and ‘actual’ model outputs, to reduce this difference. After 100,000 training examples, error minimisation reached a flat line, indicating no further possible improvement to the model. The trained model could then be used as predictive tool to estimate blight risk for any given set of input data.

Model training also included a 10-fold cross-validation approach, in which the following steps were applied:

1. The data was split into 13 equally sized subsets, with outbreak/no outbreak data selected at random for populating these subsets.
2. Three of these subsets were ‘held back’ for final model validation.
3. For each of the 10 remaining subsets, a neural network model was trained using 9 subsets, with the 10th subset (which varied for each of the ten models) used for testing data.

4. Once all 10 neural network models were trained, they were used as a 'consensus' model – the outputs of all 10 models were used to produce an average (mean) output. This approach has been demonstrated in machine learning to produce more robust and accurate estimation models that using a single-model approach.

The model was validated using data in the 3 subsets held back from the training process, providing a statistically reliable accuracy assessment. The model also incorporated information about local and neighbouring outbreaks over the prior 18 months, and so can adjust to information over the growing season if there are nearby outbreaks. Input variable values were normalised prior to modelling, so that the range presented to the model during training was in the range [0, 1] for all variables. This was done to prevent input variables with a larger range of values having more apparent importance in the model (i.e. elevation ranges from 0 to 1400 metres, while slope varies from 0° to a theoretical maximum of 90°).

Once the model was trained and tested, a sensitivity analysis was performed to determine the impact of individual input variables and assess their relative importance as drivers of late blight outbreak risk. This was done by calculating the rate of change of the output variable, in proportion to changes in each input variable by a small amount when all other inputs are fixed. Rates of change can be positive or negative.

Each neural network model was activated once for each training data point and the RMS value of rate of change determined across all training data points. The consensus model (all 10 neural networks) was used for this, instead of evaluating each model separately. With 118 input nodes (1 for each input variable), this gave 118 sensitivity relationships. These are presented in graphs that give the sensitivity for each type of input data. It is important to note that the sensitivity value in each case is a proportional rate of response. For example, a value of 0.1 indicates that for a unit change of value in that input variable, the output will vary on average by 0.1 of the input variation.

Model analysis of pathogen population change

Only two pathogen genotypes were available in sufficient numbers in the FAB outbreak data to facilitate analysis via machine learning: genotypes 13_A2 and 6_A1. The problems of small data are numerous but mainly revolve around overfitting. Overfitting occurs when a model adjusts excessively to the training data, seeing patterns that do not exist, and consequently performing poorly in predicting new data. In addition, outliers and noise become a real issue in small-data. The same suite of 24 machine learning classification techniques used to model early outbreaks were tested for their ability to predict distributions of 13_A2 and 6_A1. The goal was to develop an accurate model and use it to identify the most important variables for prediction, i.e., the principle driving factors associated with the occurrence of different genotypes. The same HADUK-Grid weather variables described above were used as model inputs, although values were averaged over 2006–2018 as genotype information was not available in the first 3 years of the FAB outbreak data. Two different modelling approaches were attempted that utilised the pathogen genotype data at different levels of spatial aggregation: models were developed to predict the genotype of individual outbreaks, and models were developed to predict the dominant genotype in each postcode district. The latter involved counting the number of outbreaks of each genotype in each postcode district. The models were trained, tuned, and tested as described above for the analysis of early outbreaks.

Results

Deliverable 1: occurrence of early outbreaks of late blight

To illustrate where the early outbreaks of late potato blight occurred the 5th, 10th and 20th percentiles of blight outbreaks (by reporting date) over the entire study period have been mapped in Figure 1.

Figure 2 is a map of the (median) day-of-year (i.e. days since January 1st) of first blight outbreak in each postcode district. In the plot, red postcodes are the earliest occurring outbreaks and yellow are the latest occurring outbreaks. A cut-off (colour bin) was used for every 5% of these days of first outbreak, meaning that some bin sizes cover more days than others; the very first outbreak occurred at day 46 (i.e. mid-Feb), the first 5% of the early outbreaks occurred between days 46 and 146 (mid-May) and so on. All 'first outbreaks' occurring after August 1st were grouped into one bin (days 213-365).

There is no clear pattern in this simple plot per postcode district, although it appears that most of the redder (earlier first outbreak) areas are along the coast and in the south (East Anglia and East Midlands) and more of the yellow (later first outbreak) areas are in inland and further north. This would indicate that first outbreaks tend to occur in parts of Great Britain that are warmer early in the year and would agree with the general understanding that temperature plays an important role in outbreak occurrence. There is further analysis of the controlling factors of outbreaks using GIS and spatiotemporal analysis tools, reported later in this report.

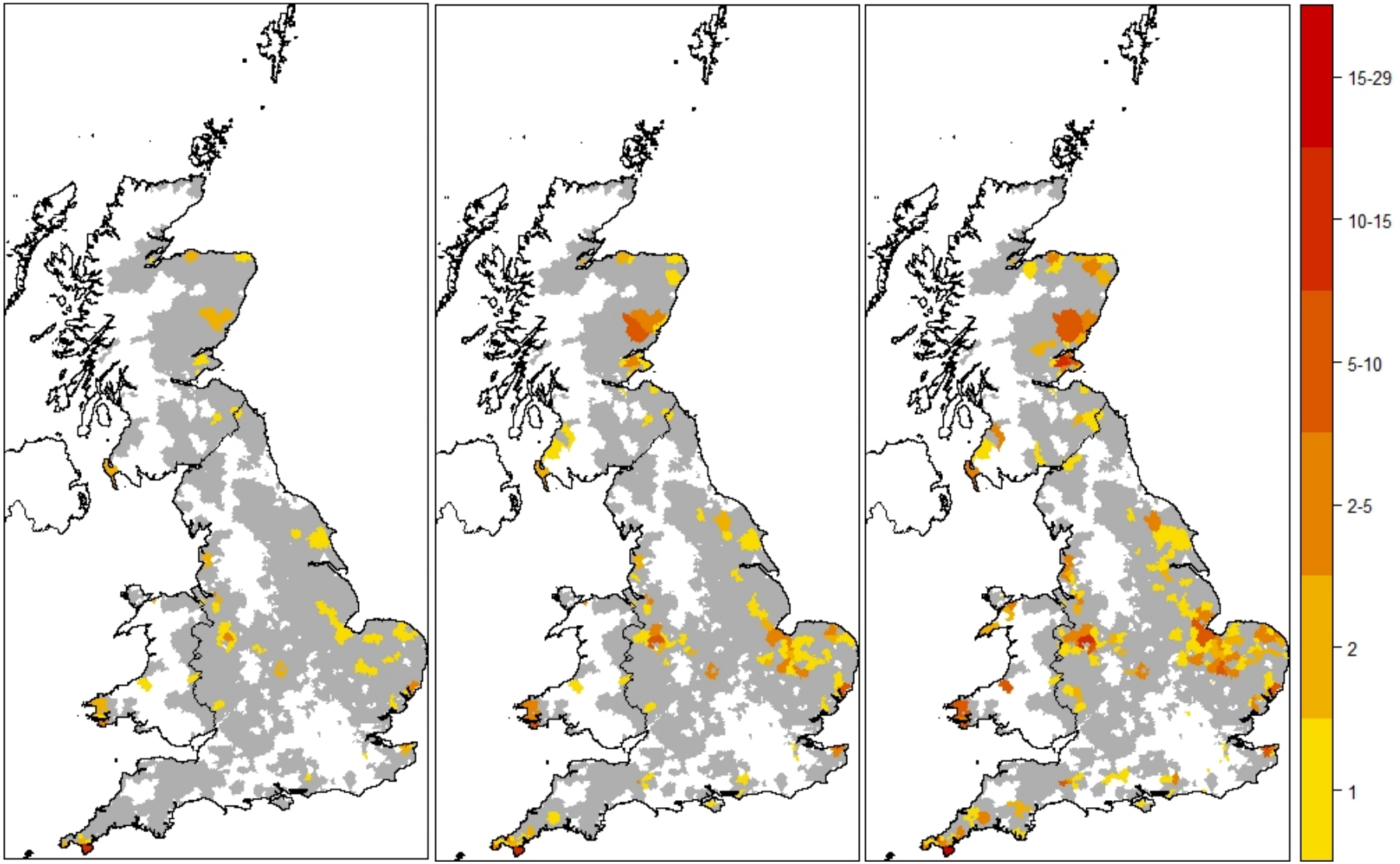


Figure 1. Choropleth maps of the 5th (left), 10th (middle) and 20th (right) percentiles of the blight outbreak dates of every year from 2005 to 2018. The colour scale indicates the number of sampled outbreaks per postcode district and grey zones are postcode districts in which >1 ha of potato was grown commercially.

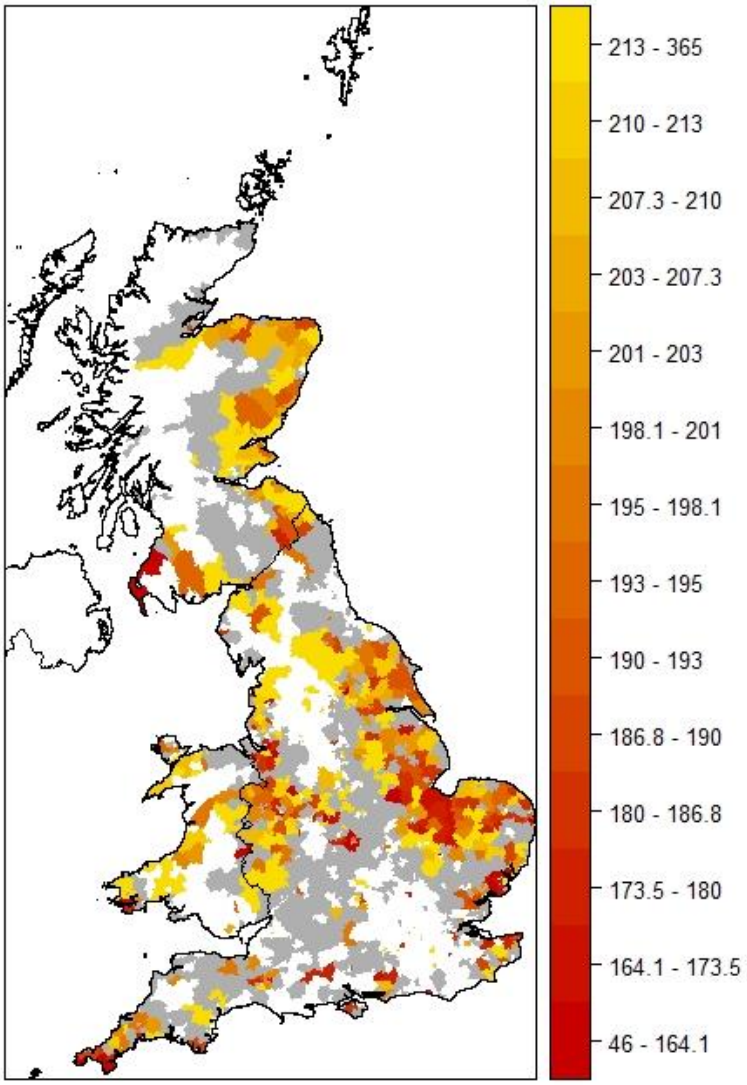


Figure 2. Median day of year of first late blight outbreak mapped by postcode district. Red indicates areas where outbreaks occur earlier in the year, on average. Yellow indicates areas where outbreaks tend to occur later.

Figures 3 and 4 show output from the Optimised Hot spot Analysis tool, showing statistically significant spatial clusters of high (hot spot) and low (cold spot) values of incidence for early outbreaks, 2003-2018. This shows the pattern and scales at which early outbreaks occur in GB. Hot spots of early outbreaks (10th and 20th percentile) were generally found in the south of England and Wales, particularly near the coast, whereas a large cold spot extends across the production regions of Scotland. The 10th percentile hot spots mainly relate to early potato growing regions in southwest England, Wales and to some extent, Scotland whereas the 20th percentile includes larger areas of potato production in the southwest and southeast of England. The local (Figure 3) or more extensive (Figure 4) cold spot in Scotland was somewhat unexpected given the results of the choropleth mapping (Fig. 1), which shows a few districts with high counts of early outbreaks in the Fife and Angus regions. However, these few regions are surrounded by neighbouring districts with no early outbreaks, resulting in the

observed cold spot. Overall, the results indicate a clear role of climate in occurrence of early outbreaks.

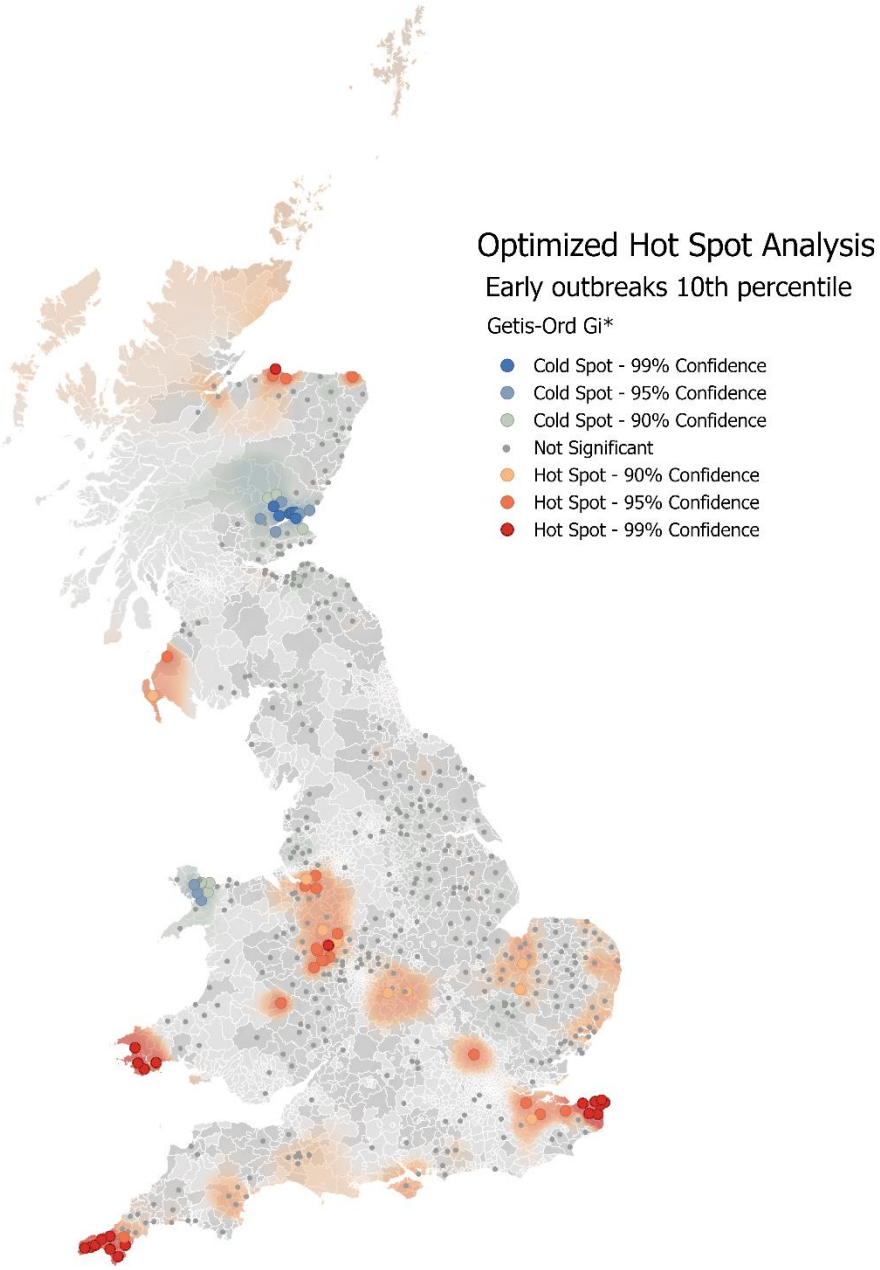


Figure 3. Statistically significant hot and cold spots derived by OHSA from the 10th percentile of late blight outbreaks by reporting date, 2003-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

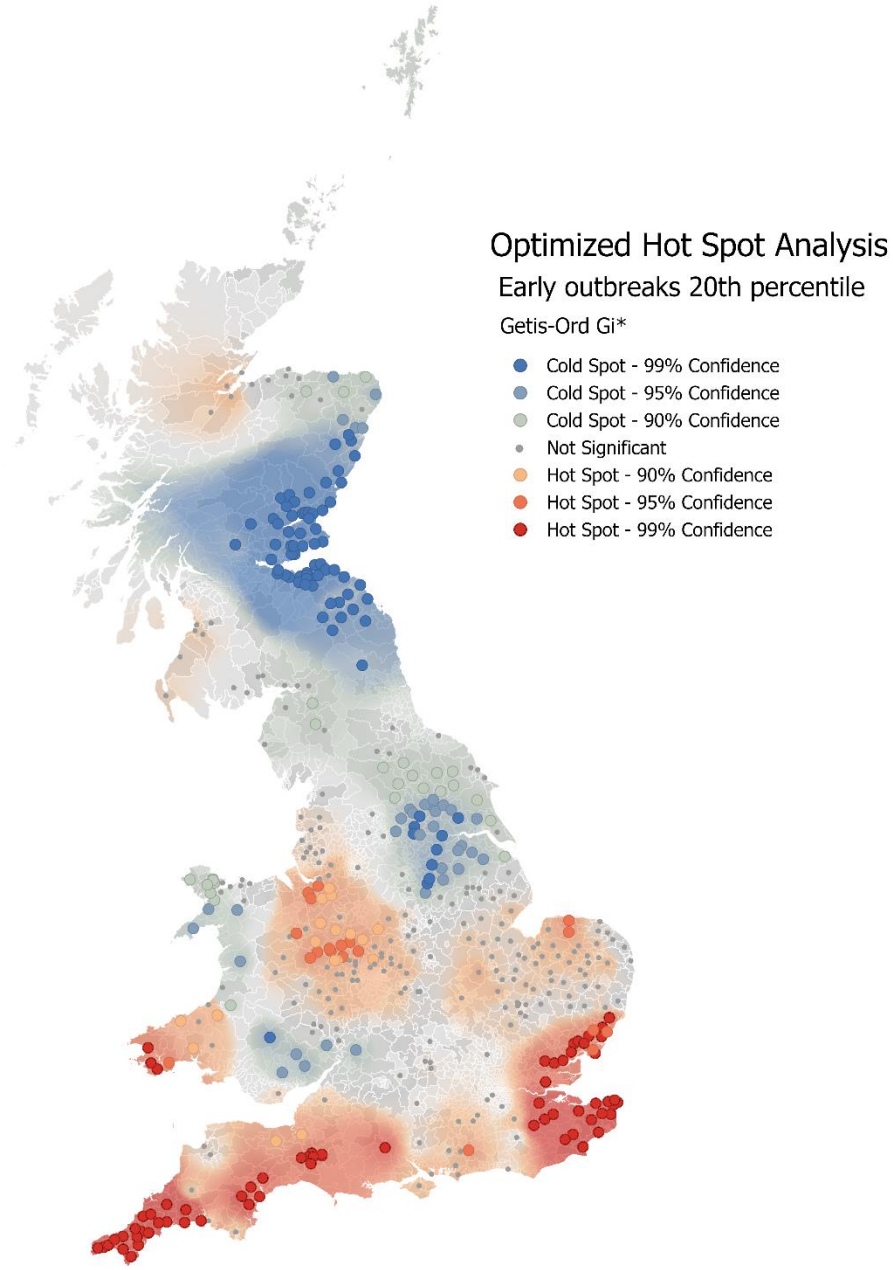


Figure 4. Statistically significant hot and cold spots derived by OHSA from the 20th percentile of late blight outbreaks by reporting date, 2003-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

Early outbreaks were further subdivided according to reporting year to analyse change in spatial distributions over time; however the data were insufficient (too few instances) for robust analysis and results were potentially misleading (i.e., a single outbreak could seem disproportionately significant within a small dataset). We therefore performed an emerging hot spot analysis (EHSA) to quantify change in early outbreak distributions over time. Similar problems occurred when analysing total incidence and pathogen genotype by year, and alternative techniques were again used.

Compared to the many hotspots defined over the whole data set (OHSA), the spacetime EHSA revealed fewer hot spots of early outbreaks that tended to be sporadic: locations that are on-again off-again hot spots in Kent, southwest England and Scotland (Fig. 5, Fig. 6). The cluster of sporadic and consecutive hot spots (uninterrupted runs of hot spots in the final time steps) of early outbreaks in the Angus / Aberdeenshire regions may be an indication that the warming climate could result in an increased frequency of early outbreaks in Scotland but could also reflect a bias due to an sampling regime that is more intensive than in other regions.

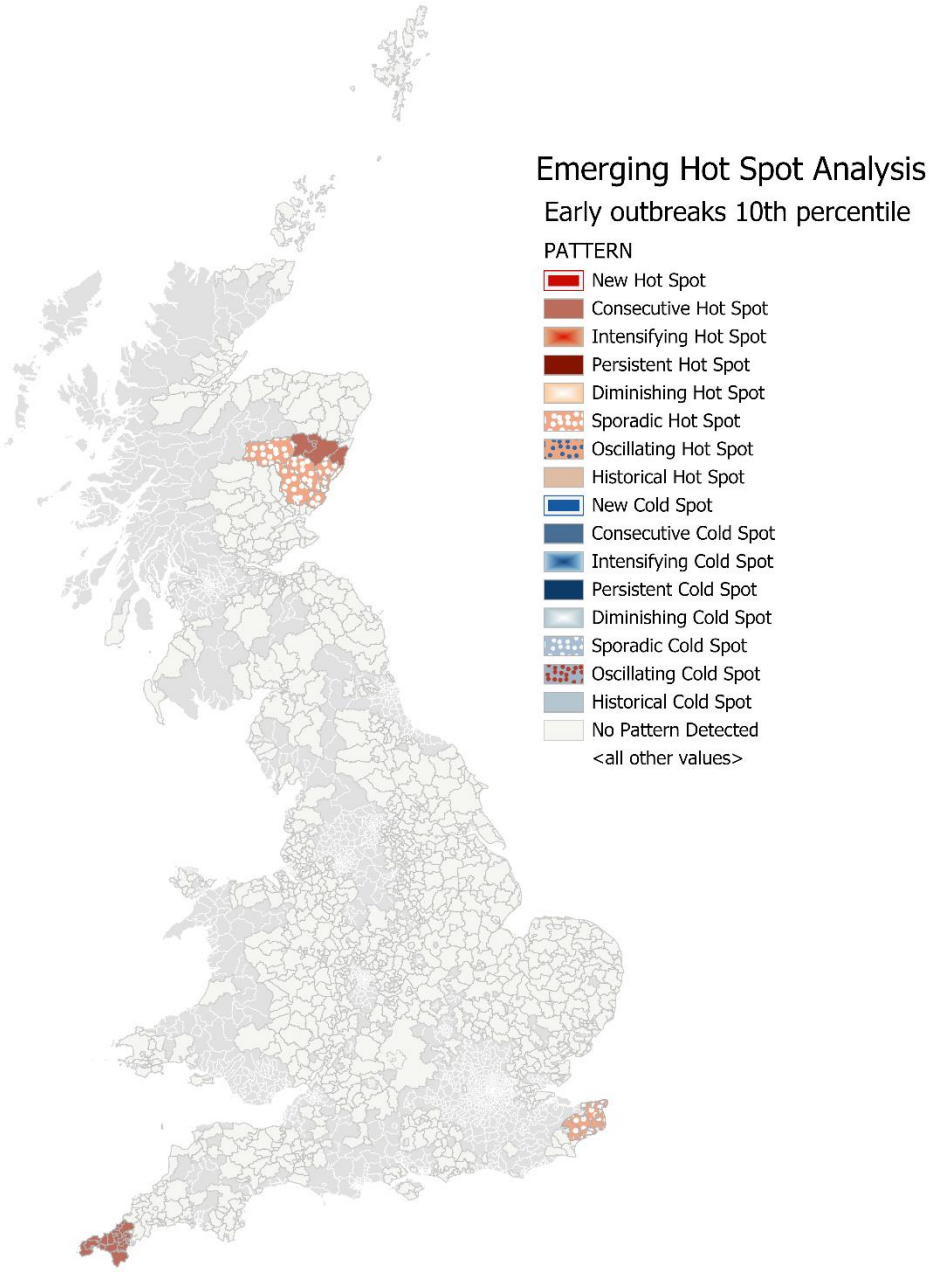


Figure 5. Space-time patterns of early outbreaks of late blight, derived by EHSA from the 10th percentile of late blight outbreaks (by reporting date), 2003-2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

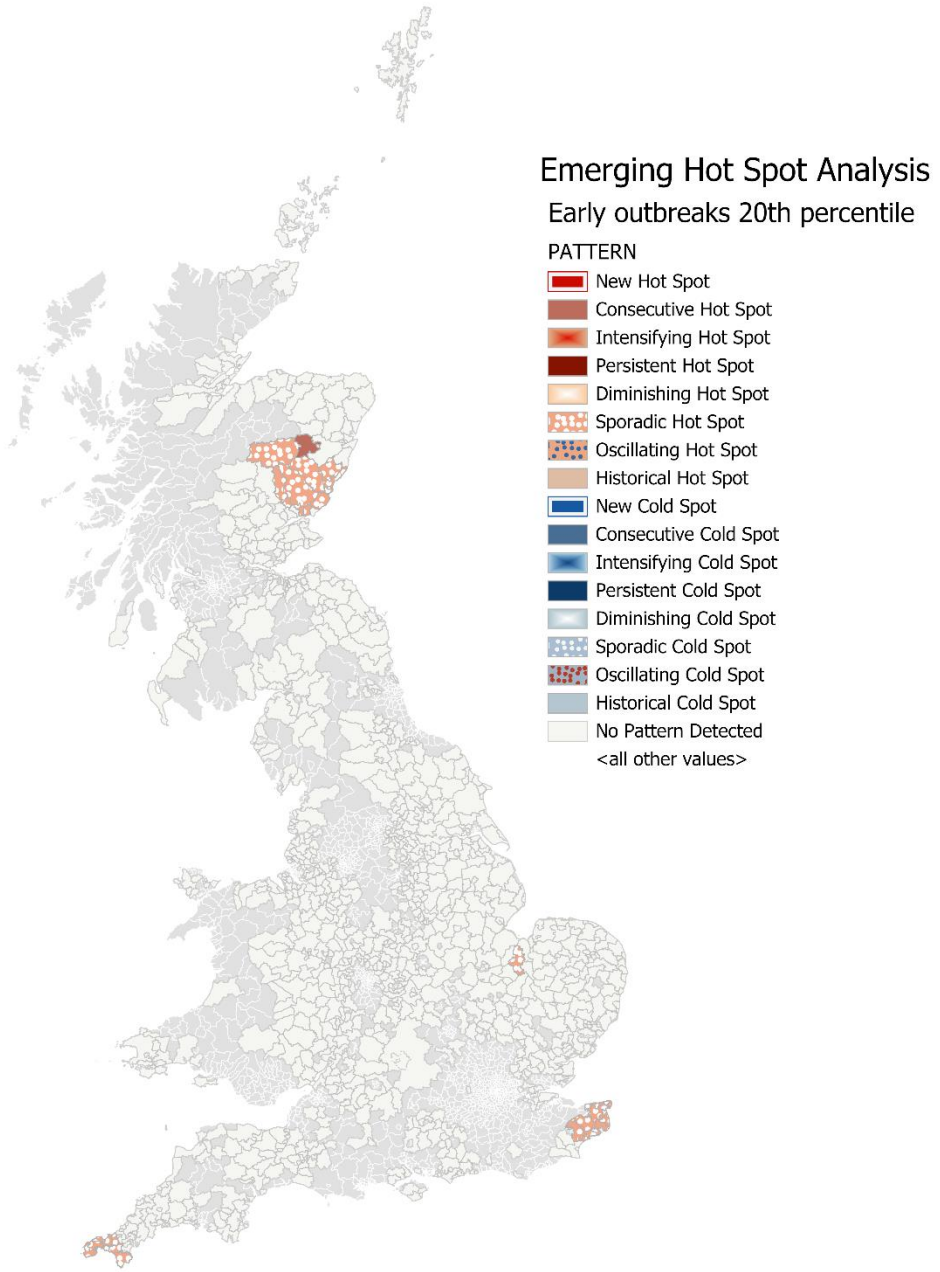


Figure 6. Space-time patterns of early outbreaks of late blight, derived by EHSA from the 20th percentile of late blight outbreaks (by reporting date), 2003-2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

Great Britain is well-known for the geographic variability of its weather, and this is reflected in the mapping of the long-term growing-season climate averages derived from the HADUK-Grid dataset (Fig. 7).

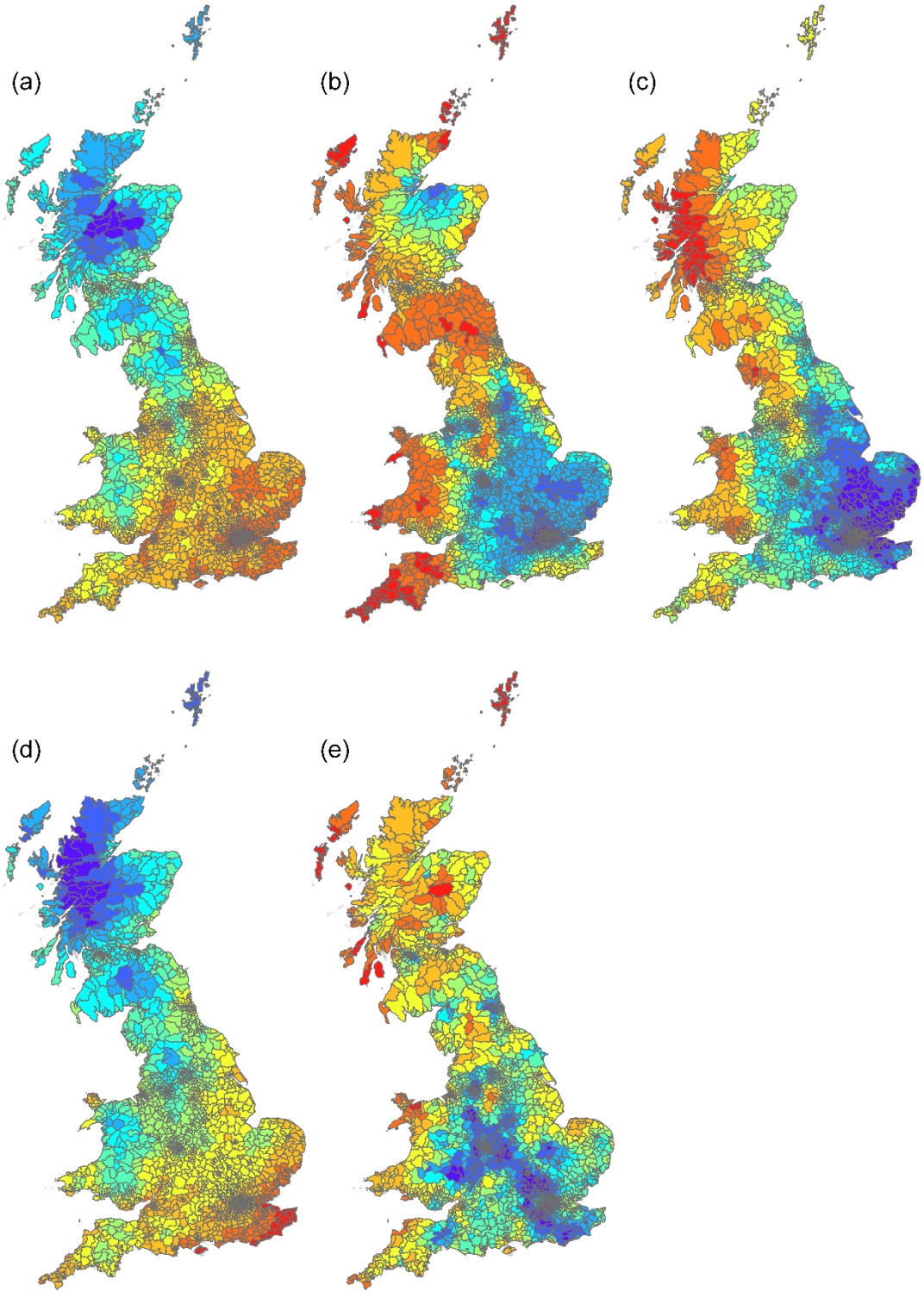


Figure 7. Weather data derived from the HADUK-Grid dataset, averaged over the potato growing season (May 1-October 31) and over the study period (2003-2018): (a) temperature, (b) relative humidity, (c) precipitation, (d) sunshine duration in hours, and (e) wind speed. Values range from blue (low) to red (high).

Of the suite of 24 machine learning algorithms tested, a 'Fine Decision Tree' proved to be the most accurate in predicting the classes of spatial clustering of early outbreaks (early20) derived by OHSA (Fig. 4). The model classified 2008 cases of the training set into the seven different types of spatial cluster with a training accuracy of 94.6%, and 501 held-out test cases with a testing accuracy of 91.2%. The similarity in training and test accuracies indicates that overfitting was not an issue. Note that nine cases were missing the required weather variables and were omitted from the analyses. A confusion matrix shows the spatial cluster classification from the OHSA (true class) versus the predicted class from the model for the held-out test data (Fig. 8). A confusion matrix is a cross-tabulation formed by the overall agreement-disagreement, where the row and column labels of the matrix represent observed and predicted classes, respectively, and each cell contains the corresponding number of test cases. Thus, the agreement values correspond to the diagonal cells, whereas the disagreement values correspond to the off-diagonal cells. The column summary displays the number of correctly and incorrectly classified observations for each predicted class as percentages of the number of observations of the corresponding predicted class, i.e., the positive predictive values and false discovery rates. The row summary displays the number of correctly and incorrectly classified observations for each true class as percentages of the number of observations of the corresponding true class, i.e., the true positive rates and false positive rates.

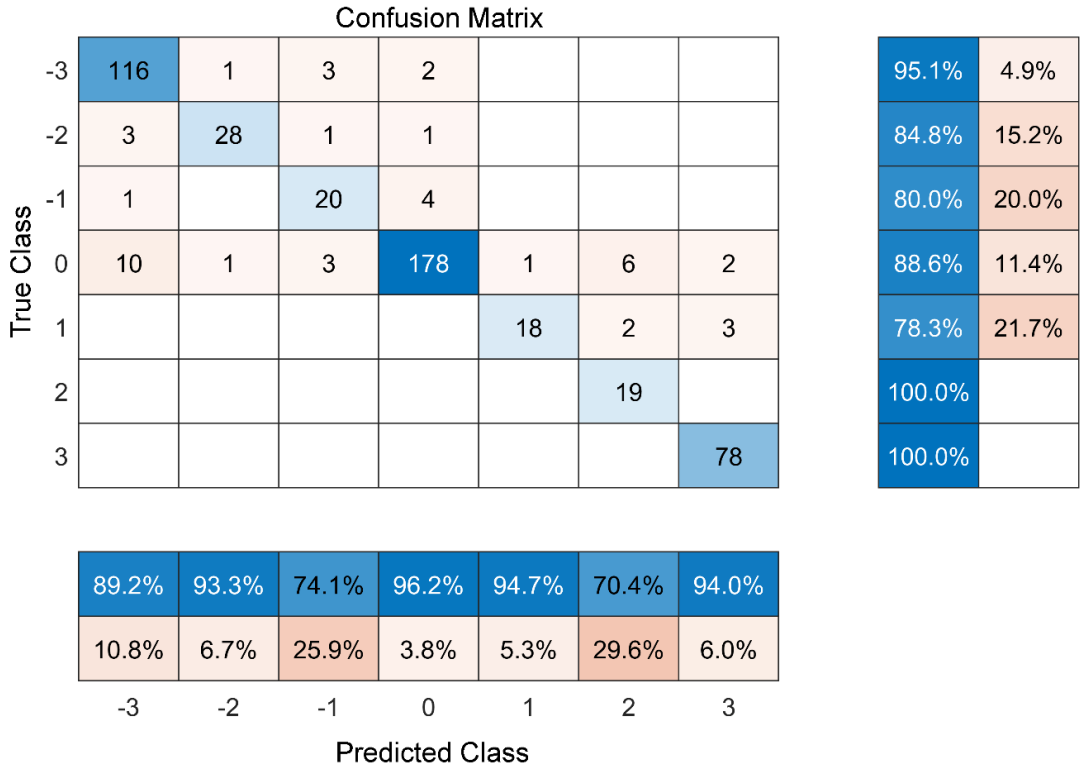


Figure 8. Confusion matrix and associated positive predictive values / false discovery rates (column summary), and true positive rates / false positive rates (row summary), for predictions of the decision tree on the test data.

The model was retrained using all the data to determine the importance of each weather variable in predicting the type of spatial cluster (Fig. 9). The importance of each variable was determined using the PREDICTORIMPORTANCE function of MATLAB. The results provide evidence that temperature is the principle driving factor for early outbreaks, followed by precipitation and sunlight. It is interesting to note that the climate maps for temperature (Fig. 7a) and sunshine duration (Fig. 7d) closely match the patterns of spatial clustering of early outbreaks (Fig. 4).

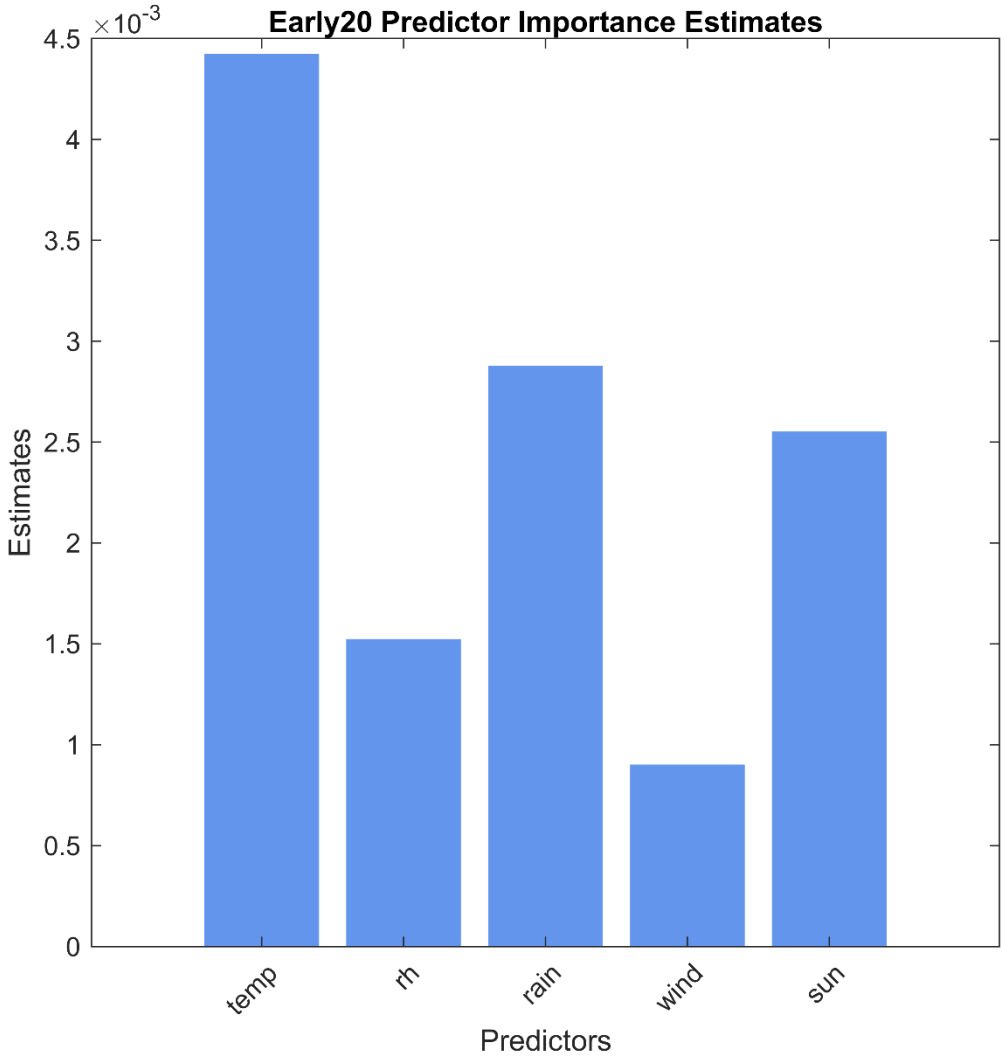


Figure 9. Importance of climate variables in predicting clusters of early outbreaks of potato late blight.

Deliverable 2: risk and rate of spatial spread of late blight

Figure 10 shows results of using Optimized Outlier Analysis within potato sectors, using default settings to identify the appropriate scale of analysis. This shows whether the level of similarity (clustering of either high or low values) or dissimilarity (outliers: high-low, low-high) in total incidence (2003-2018) is more pronounced than expected for a random distribution. Sectors in High-High or Low-High clusters are at risk of spread of disease from neighbouring (High) sectors, whereas those in Low-Low or High-Low are at less risk of spread of disease from neighbouring (Low) potato sectors.

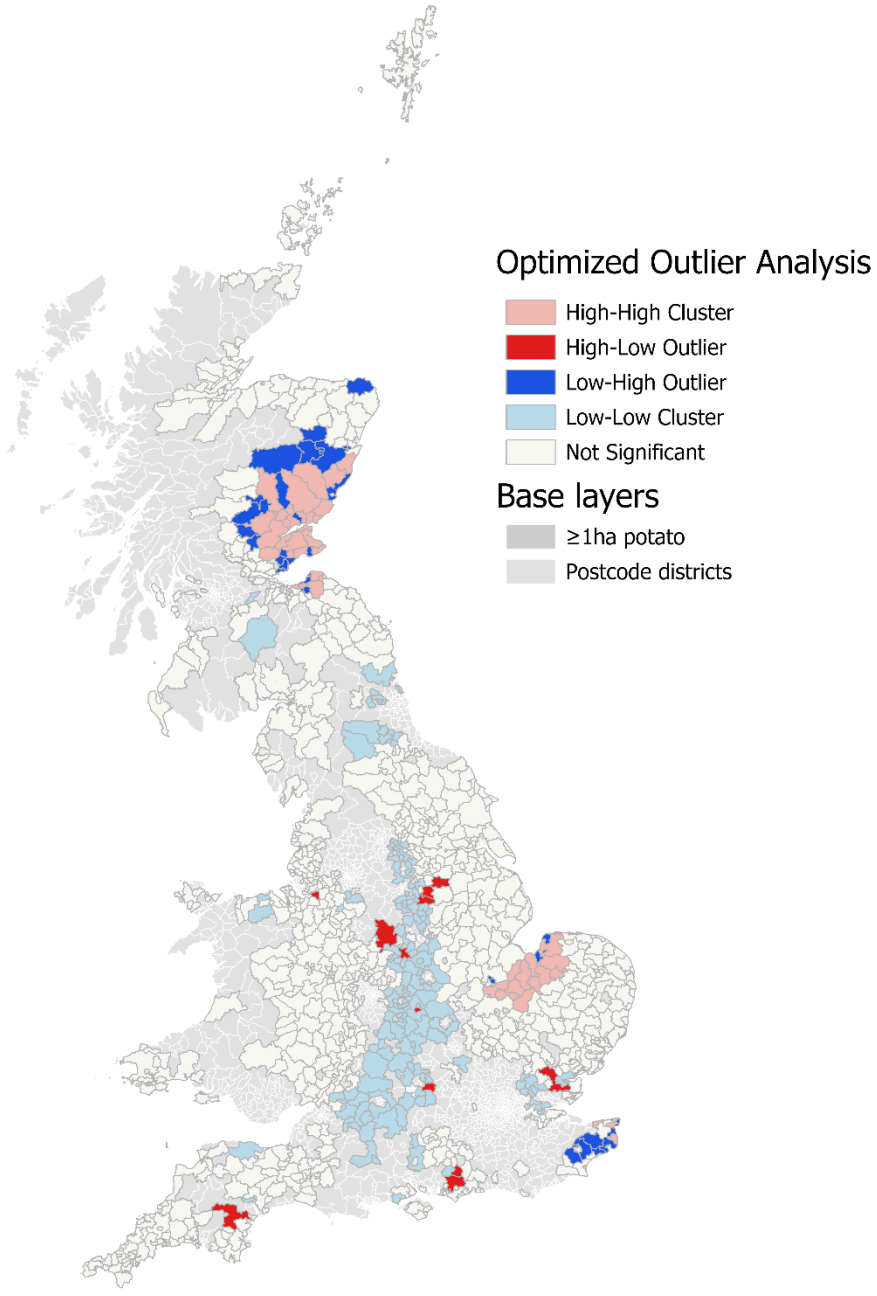


Figure 10. Output from the Optimised Outlier Analysis tool, showing statistically significant variation between observations and random variation. Postcode districts with no reported commercial potato crops were excluded from the analysis.

To determine a velocity of spread for 37_A2, data from only 2017 and 2018 were used as there were too few observations in 2016 to include in the analysis. We identified a 'western focus' of 37_A2 in the West Midlands where the first outbreak was reported (2016), and where the outbreaks were most concentrated. There were additional outbreaks recorded in East Anglia and the South East and one up in Auchincruive in subsequent years, but it is likely these originated from different sources of primary inoculum as opposed to dispersal of inoculum outwards from the West Midlands focus. The maps (Figs. 11 & 12) shows the Standard Distance for the first week in 2017 and 2018 where 37_A2 was recorded (inner light blue circle) and the Standard Distance for the final week when 37_A2 was recorded. The velocity of spatial spread for 37_A2 in 2017 was:

$$vel = \frac{SD_{\text{last}} - SD_{\text{first}}}{t_2 - t_1} = \frac{124{,}701\ \text{m} - 96{,}043\ \text{m}}{9\ \text{weeks}} = 3{,}184\ \text{m wk}^{-1}$$

This was repeated for 2018:

$$vel = \frac{SD_{\text{last}} - SD_{\text{first}}}{t_2 - t_1} = \frac{137{,}134\ \text{m} - 26{,}594\ \text{m}}{8\ \text{weeks}} = 13{,}817\ \text{m wk}^{-1}$$

This analysis was repeated for genotype 36_A2 for 2018 only (Fig. 13), as there were insufficient data for 2017:

$$vel = \frac{SD_{\text{last}} - SD_{\text{first}}}{t_2 - t_1} = \frac{298{,}033\ \text{m} - 20{,}884\ \text{m}}{16\ \text{weeks}} = 17{,}321\ \text{m wk}^{-1}$$

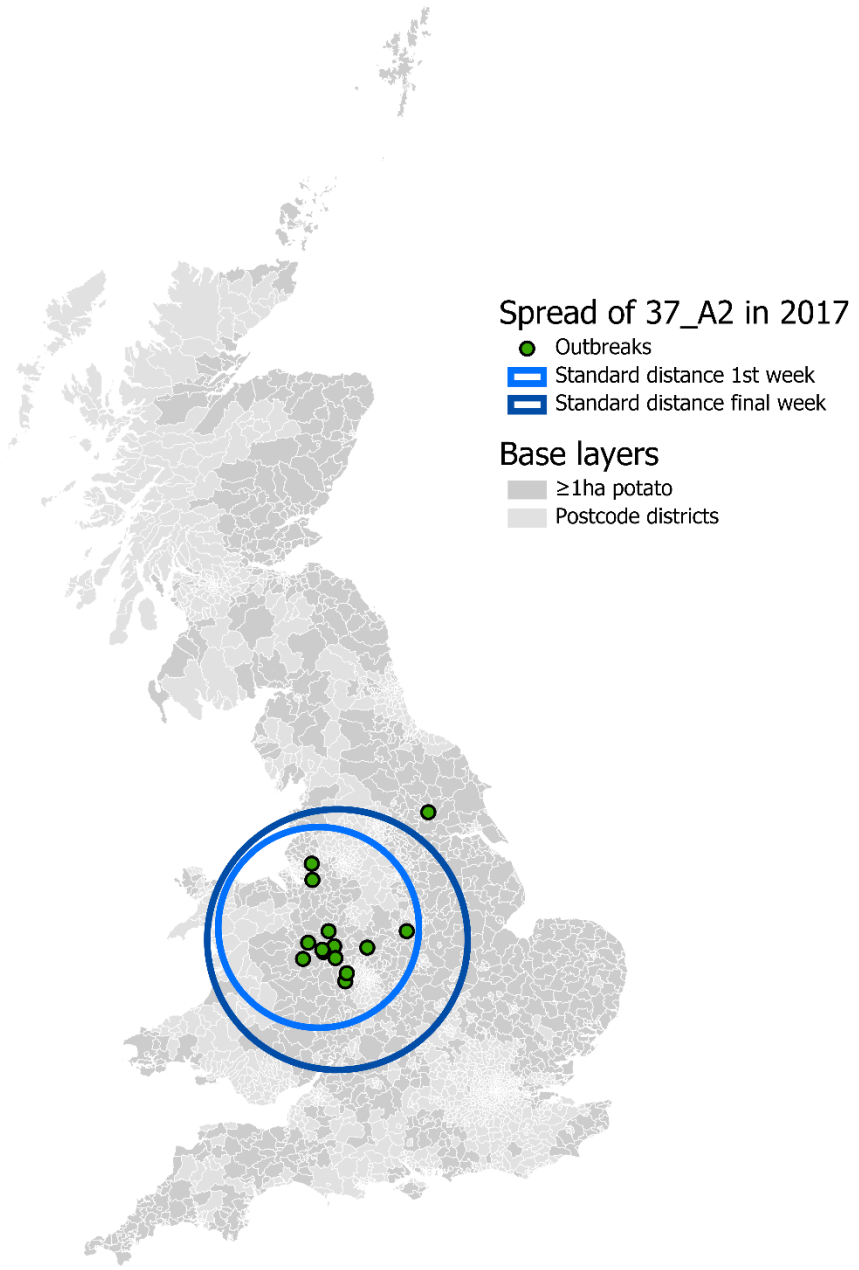


Figure 11. Standard Distance map representing the spread of genotype 37_A2 in the first and final weeks of the epidemic in 2017.

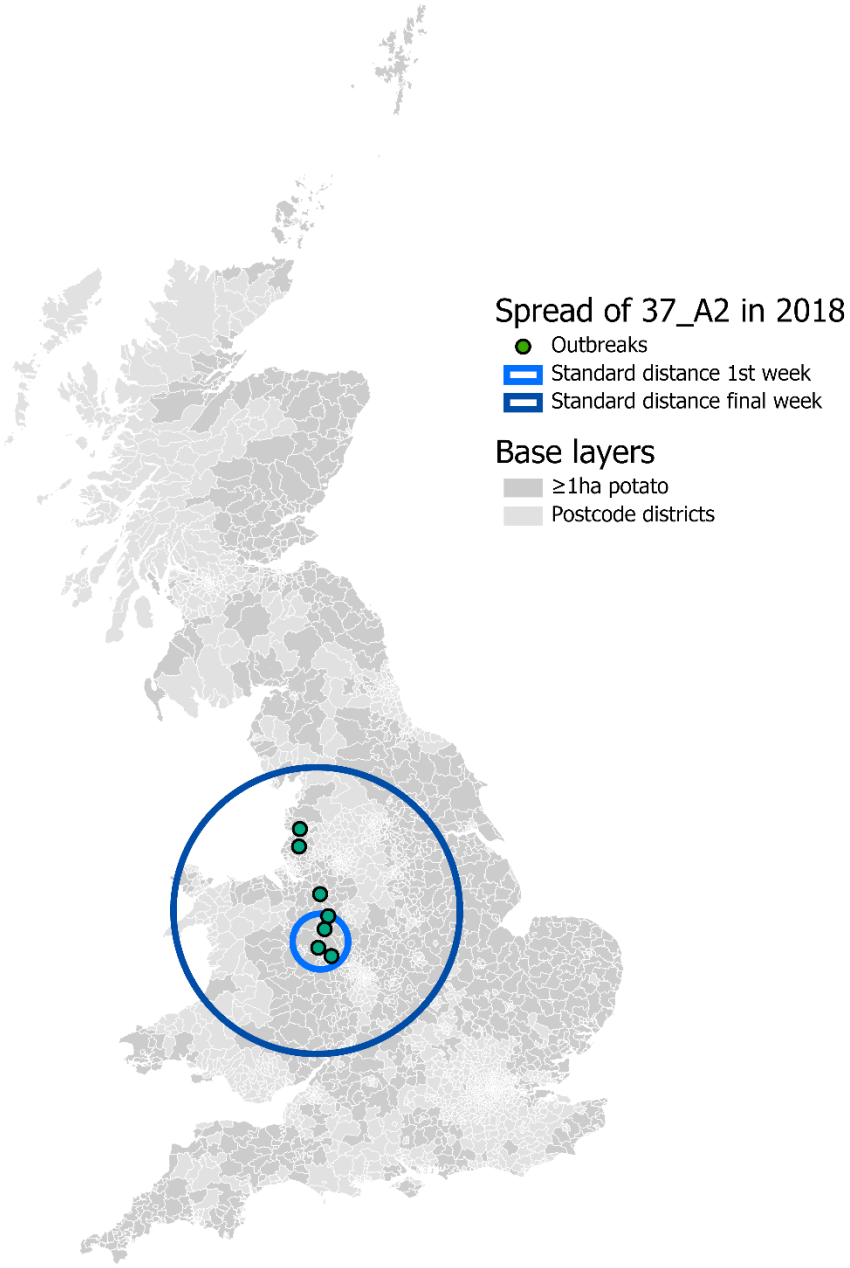


Figure 12. Standard Distance map representing the spread of genotype 37_A2 in the first and final weeks of the epidemic in 2018.

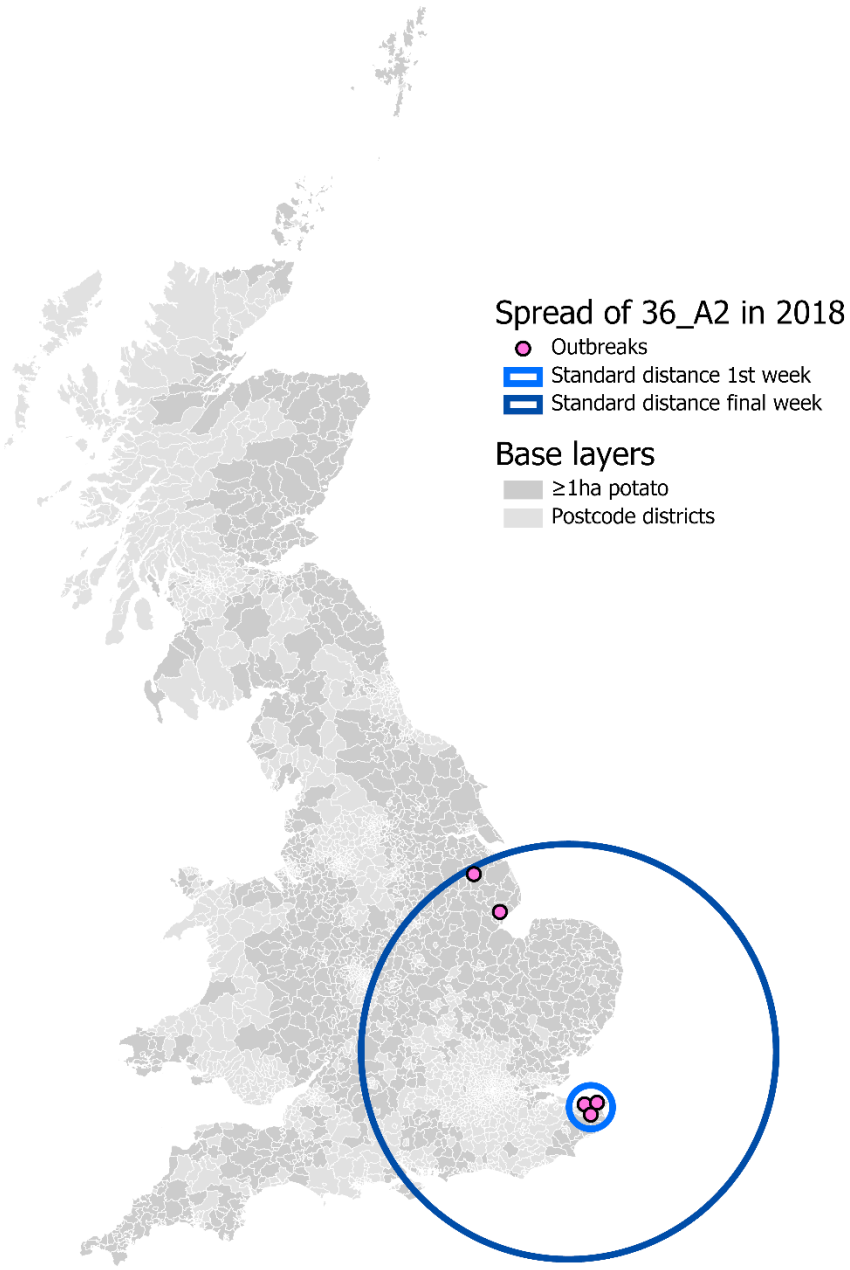


Figure 13. Standard Distance map representing the spread of genotype 36_A2 in the first and final weeks of the epidemic in 2018.

Deliverable 3: risk of late blight over time

In the following, historical maps of late blight occurrence across GB are presented to illustrate sampling intensity in relation to the density of potato cultivation and to identify areas at high or low disease risk.

Figure 14 (left) shows the total number of sampled late blight outbreaks within each postcode district from 2005-2018 and ranges from 1 to 60. There are many districts that have been sampled more intensively than others. Particularly notable are the east coast of Scotland (especially Fife, Angus and Aberdeenshire), parts of East Anglia, Kent the Midlands and the west coast of Wales.

Figure 14 (right) shows the mean area of commercially grown potatoes within each postcode district, normalised by the size of the postcode (i.e. the map shows the percentage of the postcode district area used for potato growing). The map shows where the potato growing intensity is highest and can be considered an indicator of areas more likely to be vulnerable to blight outbreaks. Since 2005, the potato growing intensity has been highest along the east coast of Scotland (especially in the regions of Fife and Tayside) and in East Anglia, East and West Midlands as well as in the early crops in southwest England, Wales and, to a lesser extent, Scotland.

Although sampling intensity varies according to postcode district there is a clear association between the potato crop distribution data and the number of sampled potato blight outbreaks. Although there are many unsampled postcode districts (grey in Fig. 14 left) these are mostly low potato density areas and the highest concentrations of blight sampling is recorded in areas with the highest potato growing density. Some deviations from this pattern are related to blight outbreaks sampled in 57 postcode districts where potatoes are not grown commercially. A total of 202 blight outbreaks were sampled in these 57 districts with some particularly clustering in some postcode districts in Wales (e.g. LL33, LL53, LL55, LL57 and SA48). Such outbreaks are predominantly from gardens, allotments and trials and some from tomato crops, but these outbreaks do not detract from the main findings as areas with <1 ha of commercial potato growing have been masked. A map of the total number of outbreaks from 2005 to 2018 normalised by potato density is also shown (Fig. 15). This normalisation in which the number of outbreaks was divided by the area of commercial potato planted (in square kilometres) within each postcode district indicates a more balanced sampling across the GB industry compared to the non-normalised data (Fig. 14).

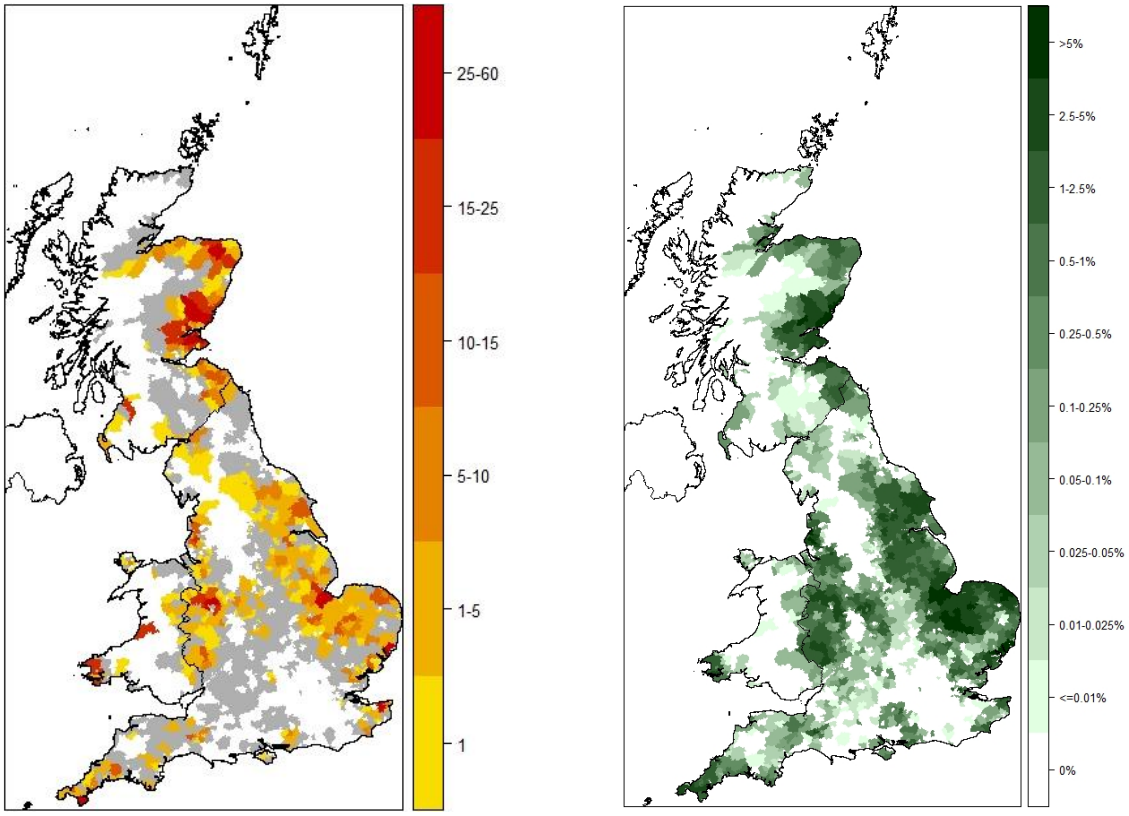


Figure 14. Left: Total number of sampled late potato blight outbreaks within each postcode district over the time period 2005-2018. Grey areas are postcodes where potatoes are grown, but where blight outbreaks were not reported. Right: Annual average area of potatoes grown within each postcode district over the time period 2005-2018, normalised by the size of the postcode (potato density).

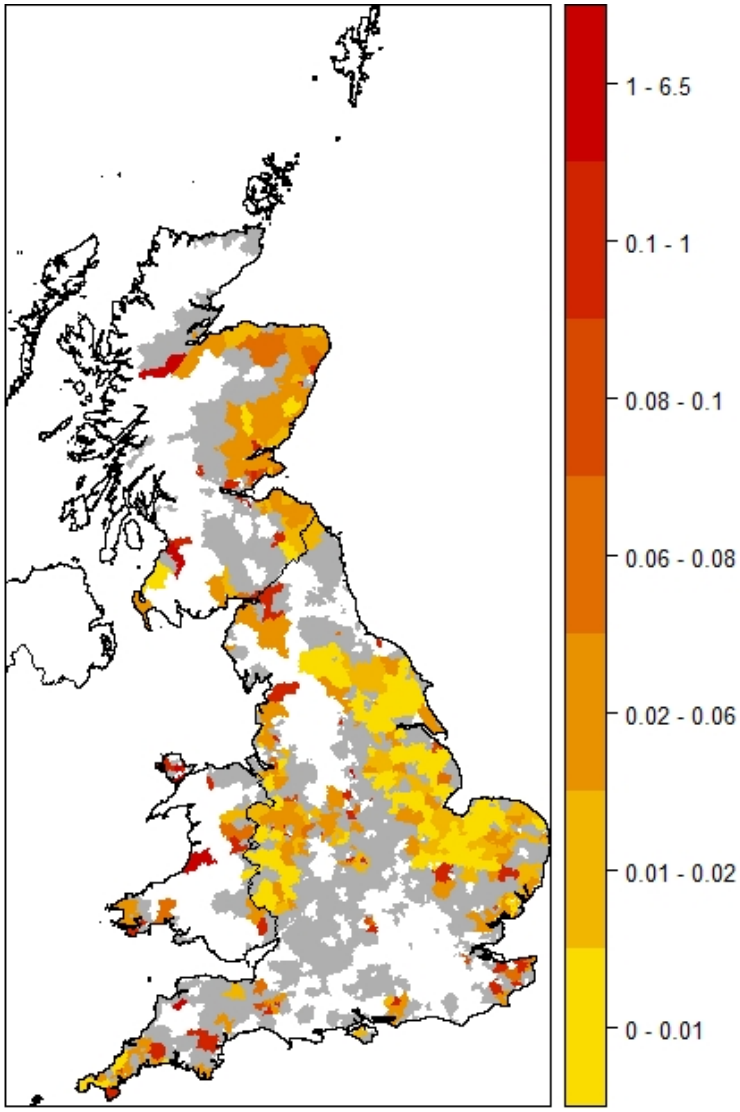


Figure 15. Total number of late potato blight outbreaks normalised by the annual average area of potatoes grown within each postcode district over the time period 2005-2018.

Analysis using the Optimised Hot Spot Analysis tool revealed statistically significant spatial clusters of high and low values of incidence for all outbreaks (total incidence), 2003-2018 (Fig. 16). This shows the pattern and scales at which late blight incidence was sampled across GB crops. Hot spots were mainly found in the Angus, Tayside, Fife and Aberdeenshire in Scotland, and in East Anglia and Kent in England, all areas of intense potato cultivation. No cold spots were identified.

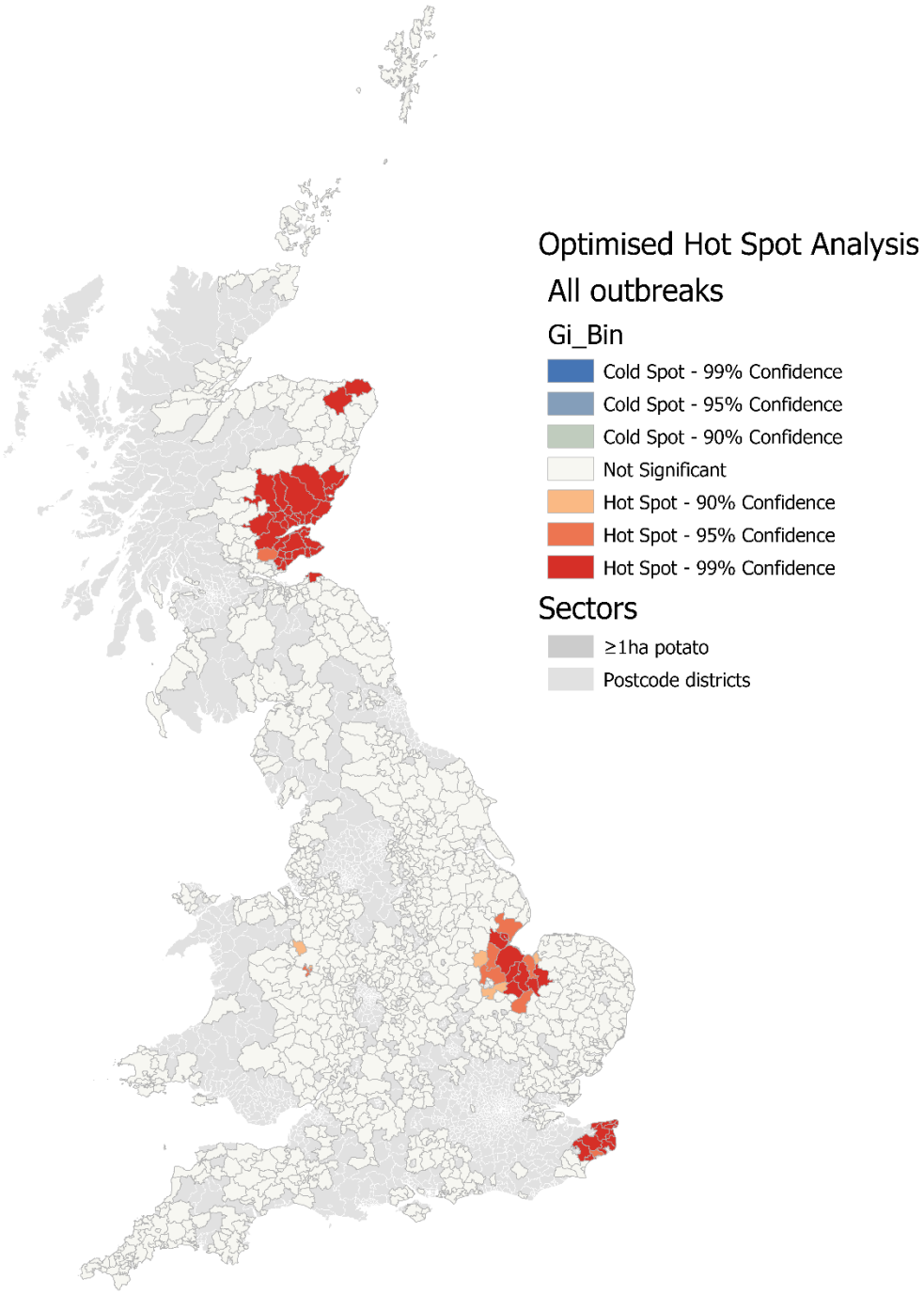


Figure 16. Statistically significant hot and cold spots derived by OHSA for all reported late blight outbreaks, 2003-2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

When analysed over space and time, the EHSA showed that the hot spots identified by the OHSA were of three types: sporadic hot spots that appear then disappear over time, consecutive hot spots where there is a run of statistically significant results in the final years, and new hot spots that are statistically significant in the final year analysed only (Fig. 17). The agreement between the OHSA and EHSA indicates that these regions are particularly problematic for late blight, either every other year or in consecutive years. This analysis over time emphasises the production area in the English midlands as a greater risk than the OHSA analysis alone.

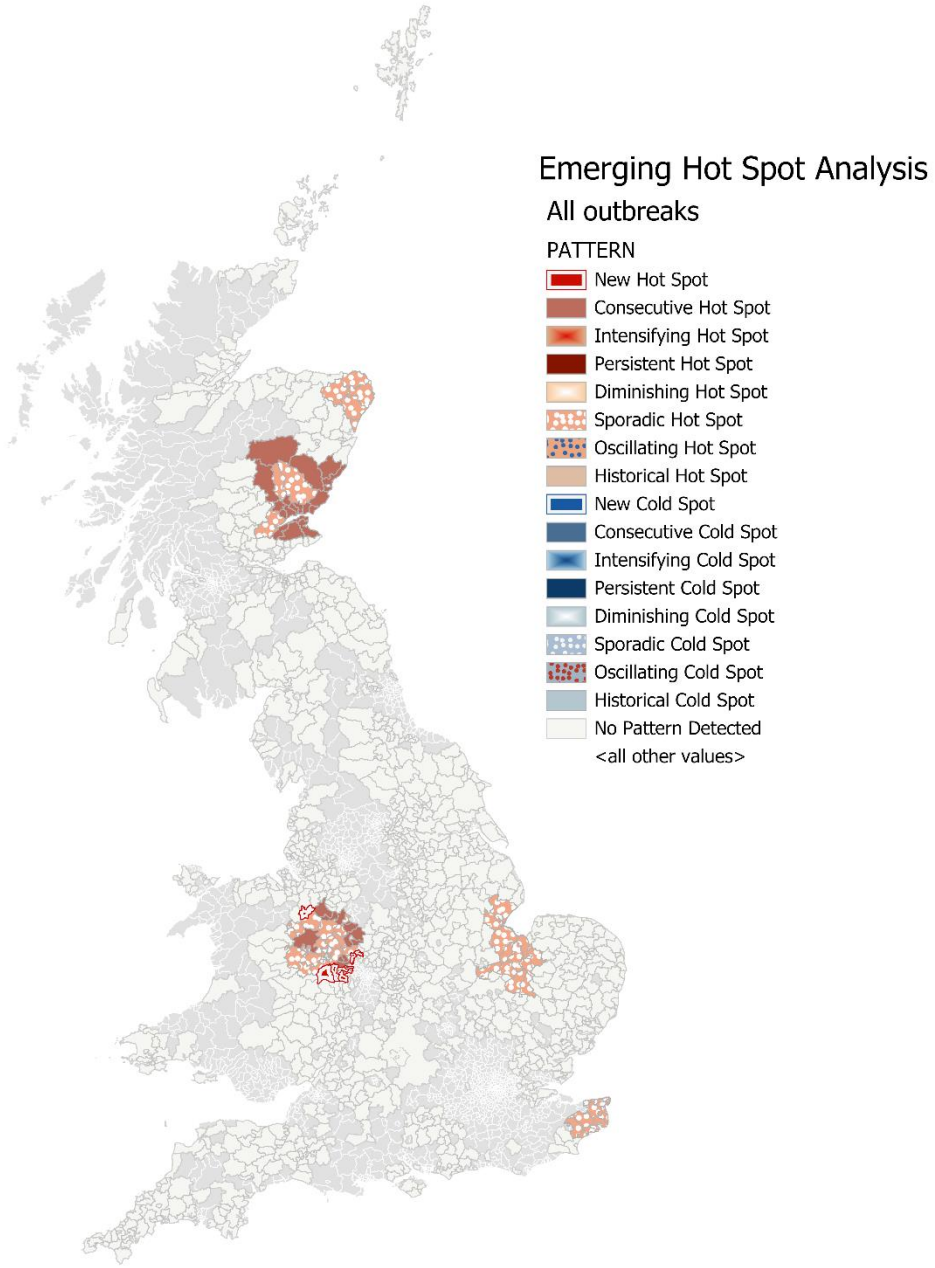


Figure 17. Space-time patterns of late blight incidence derived by EHSA, 2003-2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

Modelling

The neural network model developed to estimate risk of blight under different conditions generated a Receiver Operating Characteristic (ROC) curve (Fig. 18), which has an AUC (Area Under Curve) of 0.904. The highest accuracy (correct identification of true/false for late blight outbreak) is achieved with a threshold of 0.53 and is 82.4% with a True Positive rate of 79.7% and False Positive rate of 15.1%. The distribution of points shown in this graph demonstrate the True Positive and False Positive values obtained for different threshold values (in the range 0-1) above which the model output was taken to indicate an outbreak of Late Blight. An ROC that was a straight line from [0,0] to [1,1] would indicate a model no better than a random coin toss, while an ROC in which the curve reaches the top-left corner of the graph would indicate perfect prediction.

In situations where growers might prefer a higher True Positive rate (i.e. making sure that outbreaks are more often predicted accurately), this can be traded off against a higher False Positive rate. For example, to ensure that outbreaks are predicted 90% of the time, a model threshold of 0.33 could be used which would result in a False Positive rate of 34.0%. It is possible that growers would prefer this as the cost of 'wasted treatment' would be offset by reduced losses from outbreaks.

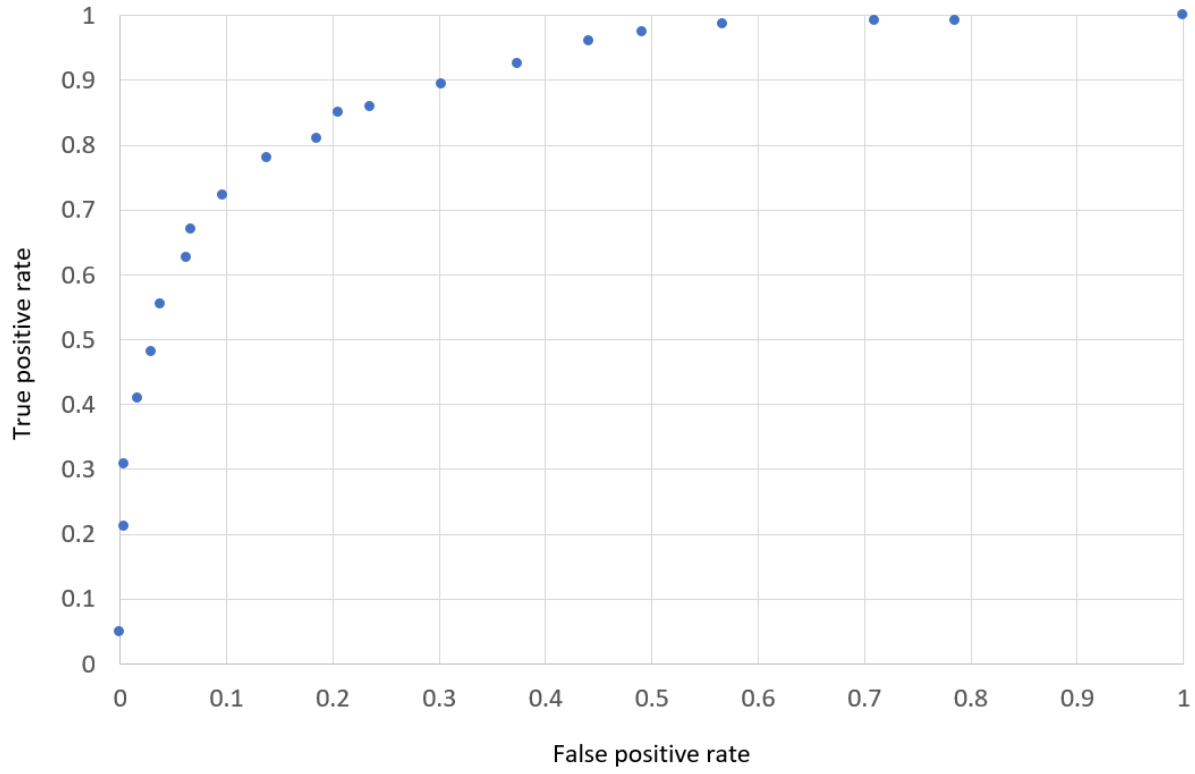


Figure 18. Receiver Operating Characteristic for model estimating presence/absence of late blight.

A sensitivity analysis of the trained model was carried out to determine the impact of individual input variables. These data are plotted in groupings by topography (Fig. 19), recent weather (Fig. 20), geology (Fig. 21), WRB soil class (Fig. 22) and WRB soil diagnostic properties (Fig. 23). The values given in each graph show the rate of change of model response (from 0 = no blight to 1 = blight) in proportion to the rate change in each input variable. Input variables are normalised within the range [0, 1].

Of the topographical factors elevation has a moderately negative effect on outbreak risk (i.e. lower elevations are more at risk), while slope has a strong impact (postcode districts with, on average, steeper slopes are more at risk). Curvature and aspect of the terrain do not show strong impacts on risk (Fig. 19). A plot of the weather-related factors (Fig. 20) shows that higher windspeed during the previous 28 days decreases the risk of outbreaks moderately. As expected, high rainfall, humidity and temperature in the 28 days prior to observation all give strong positive responses (i.e. increased risk). The role of different weather factors at different times prior to observation of late blight outbreaks is complicated by the fact that the date of outbreak initiation was not recorded in the field observations (for obvious reasons). Because of this, it is not possible to know if the weather conditions from up to 28 days prior to observation were before or after the start of the actual outbreak.

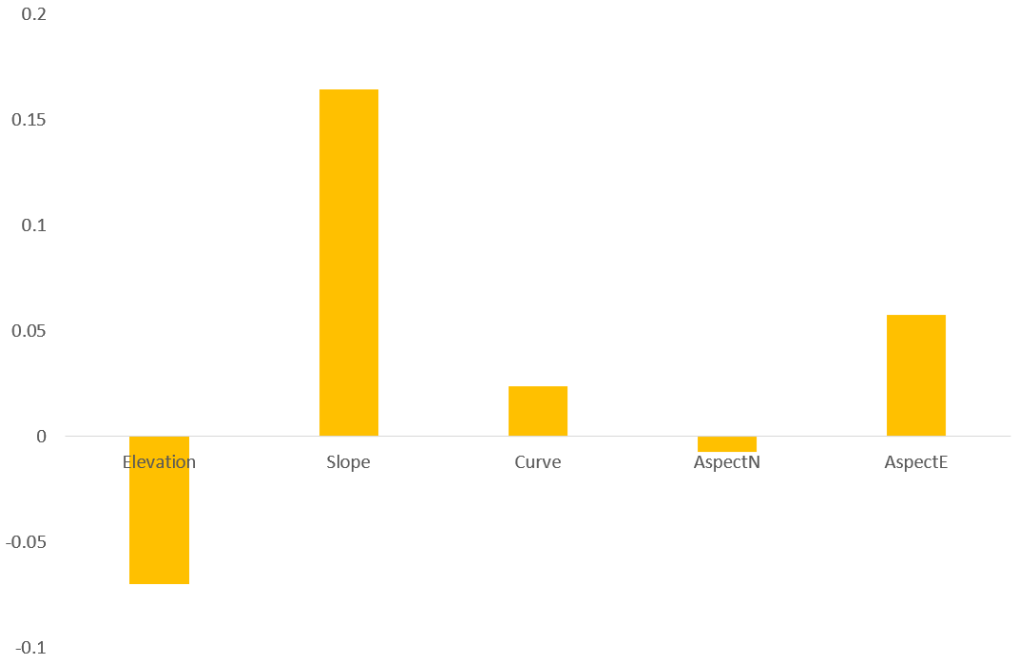


Figure 19. Sensitivity analysis of model by topography variables. AspectN is slope angle away from North (i.e. increasing towards the South), AspectE is slope angle from East (i.e. increasing towards the West) (it is necessary to have these to avoid a discontinuity in values from 359° degrees to 0°, which would cause issues with the modelling).

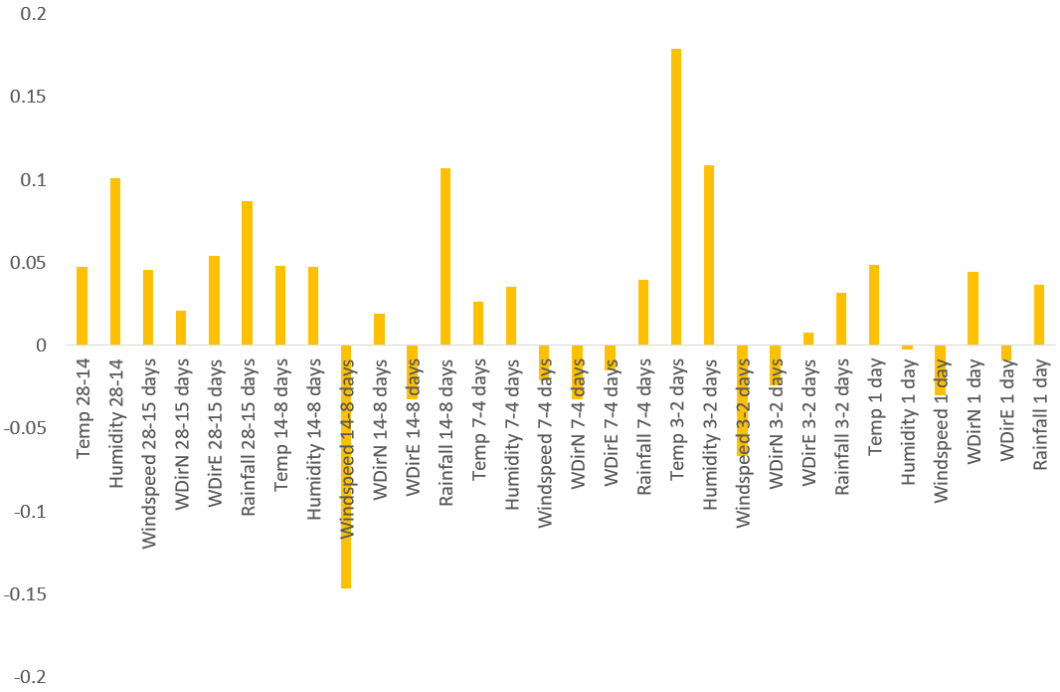


Figure 20. Sensitivity analysis of model by weather variables. Variables are averaged over the periods given before the field observation.

The underlying geological type can have positive or negative effects on late blight outbreak probability, but that none of these relationships are particularly strong (Fig. 21). Clay (increase in risk), sedimentary intermediate (e.g. sandstone, increased risk), igneous mafic (e.g. basalt, increased risk) and organic deposits (i.e. peat, decreased risk) are the strongest observations here.

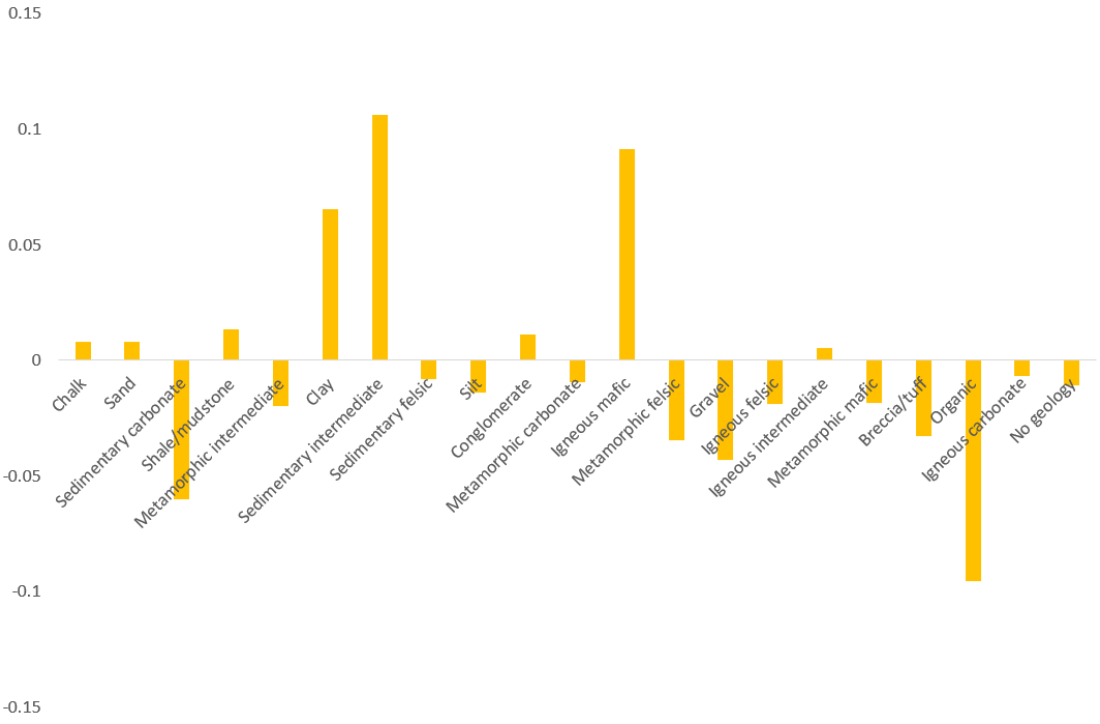


Figure 21. Sensitivity analysis of model by geological type.

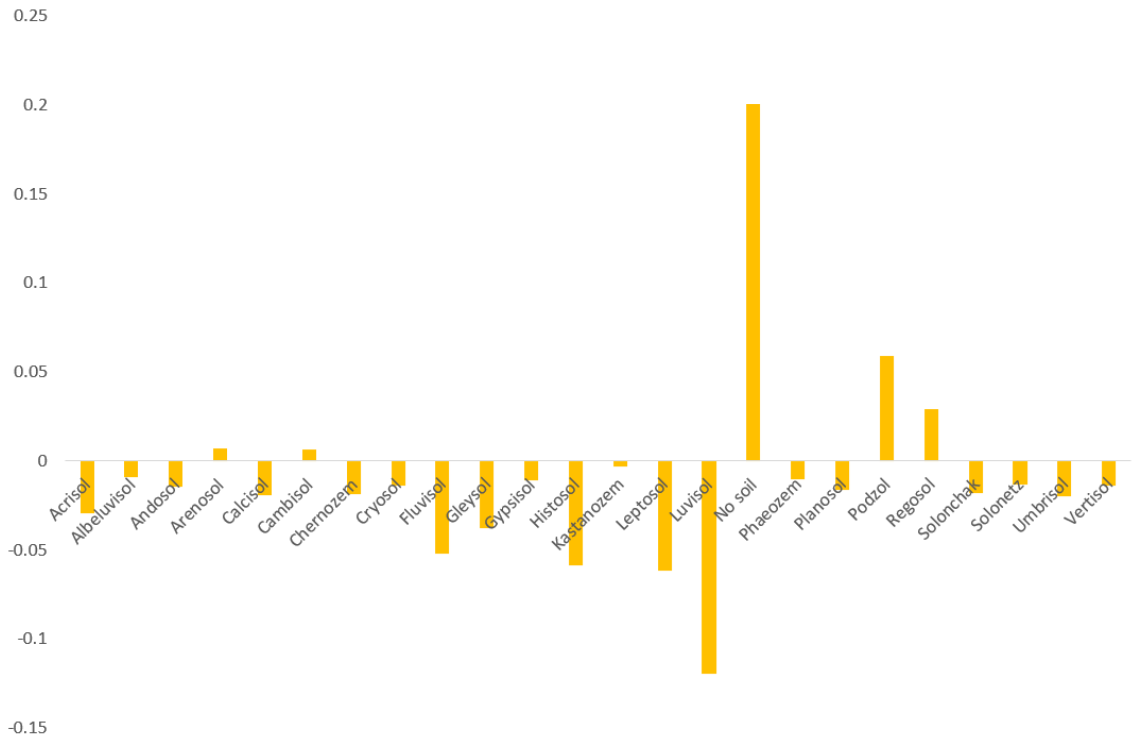


Figure 22. Sensitivity analysis of model by soil class.

While most soil types on which potatoes are grown have very little impact on outbreak risk, the 'no soil' class has a strong positive impact (Fig. 22). This soil class is seen on soil maps in urban areas, or in other places where the soil maps state that there is no soil (e.g. bare ground). For potato cropping to take place in these locations, soils must be artificially developed (e.g. allotments, gardens, raised beds). The Luvisol soil type, which is characterised by high clay content at depth, is seen as reducing the likelihood out late blight outbreaks. This soil type is common in Norfolk, Suffolk and along the north-east coast of England as well as throughout the Midlands.

The impacts of presence/absence of specific soil characteristics, which are determined through the observation of specific features or developmental conditions are plotted in Figure 23. The two most notable here are Calcaric (contains significant amounts of calcium) and Mollic (a feature of nutrient-rich soils with high organic matter and high biological activity), both of which cause reduced outbreak risk when present. No other specific features cause major differences to the risk of outbreak. Soils with Calcaric and Mollic characteristics generally have above-average pH, which may be significant.

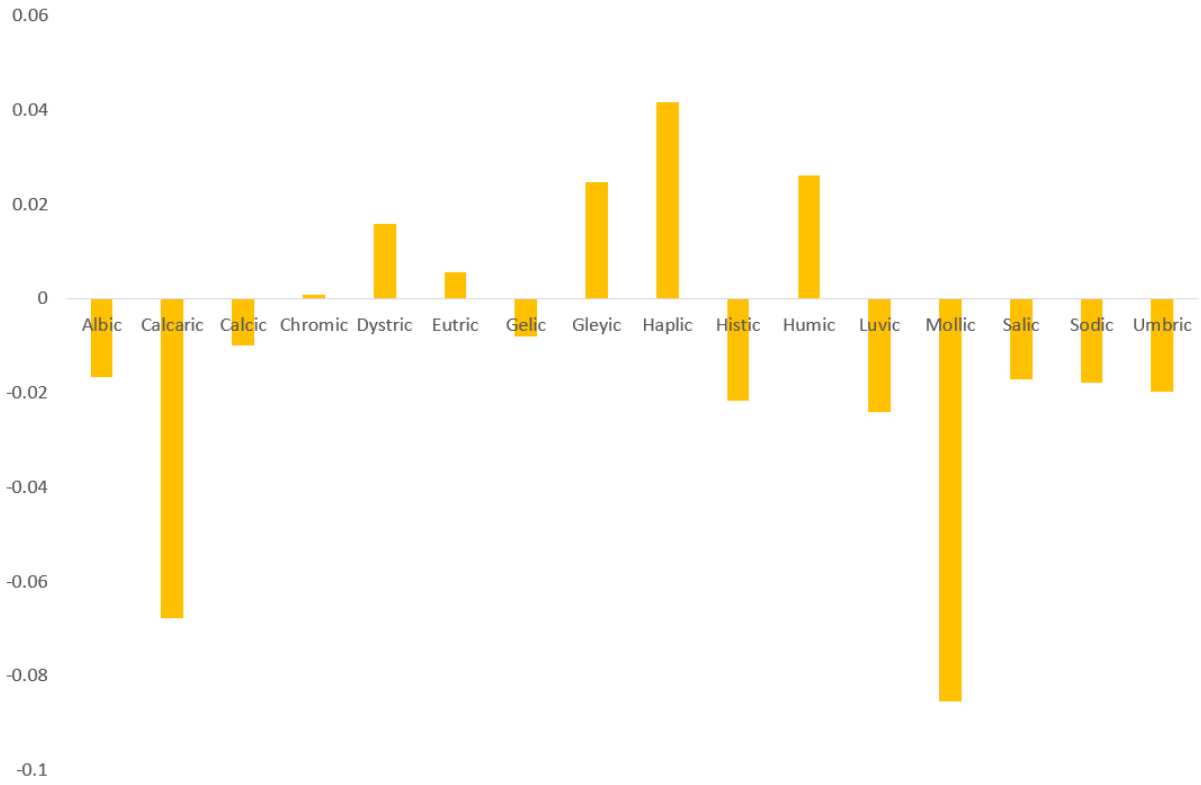


Figure 23. Sensitivity analysis of soil diagnostic properties.

The results of the modelling work and sensitivity analysis described above, as applied to various criteria are mapped to indicate the spatial distribution of risk (Figs. 24-29). The display of these maps is slightly different from previous maps in this section because of the underlying geographical projection of the data used. Additionally, the mapped data extent lies beyond GB, so areas of Ireland and northern France are included (these can be ignored). The sensitivity of the model to elevation has been used to produce a map of the impact of elevation differences across GB (Fig. 24). As the effect of increasing elevation is reduced outbreak risk, the values on the map follow a scale from near zero (red) to strongly negative (blue). So while the red areas are 'higher' risk than the blue areas, they are still below zero. Areas marked with dark blue are high elevation and lie outside existing regions of potato growing. Additionally, there are patches of small areas in the Fens that are blank due to the elevation being below sea level which prevented the model from running. The risks in this area are however, considered similar to the low-lying land surrounding it.

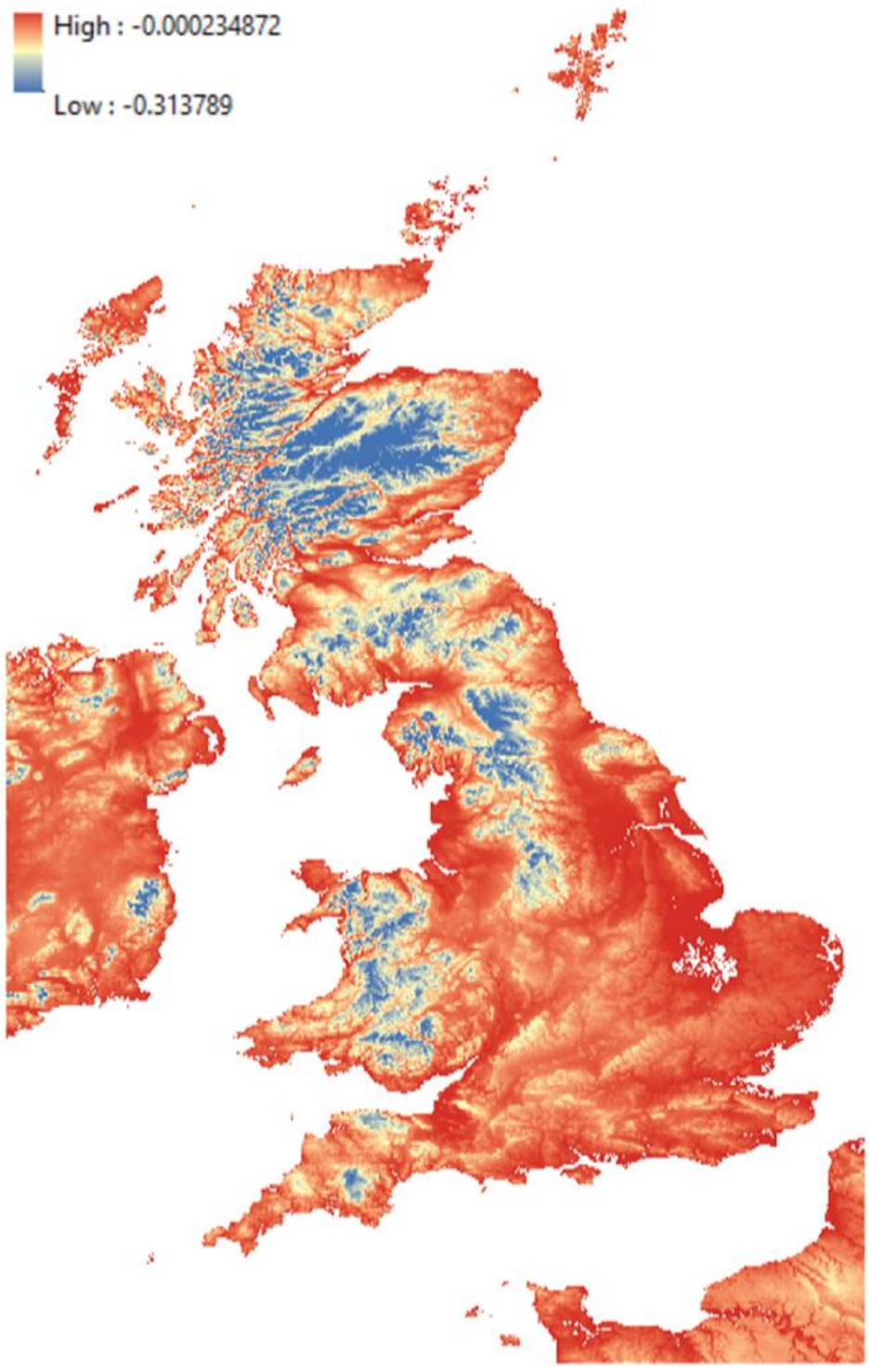


Figure 24. Map of contribution of elevation to outbreak risk.

The same type of outbreak risk mapping but using angle of slope instead of elevation is also shown (Fig. 25). As risk is positively associated with slope, areas in the map with steeper slopes are colour red for higher values of risk. Again, many of these areas are not in areas of potato growing. The three dropped lines in this map are caused by data processing errors within the GIS software used to derive slope values for the terrain.

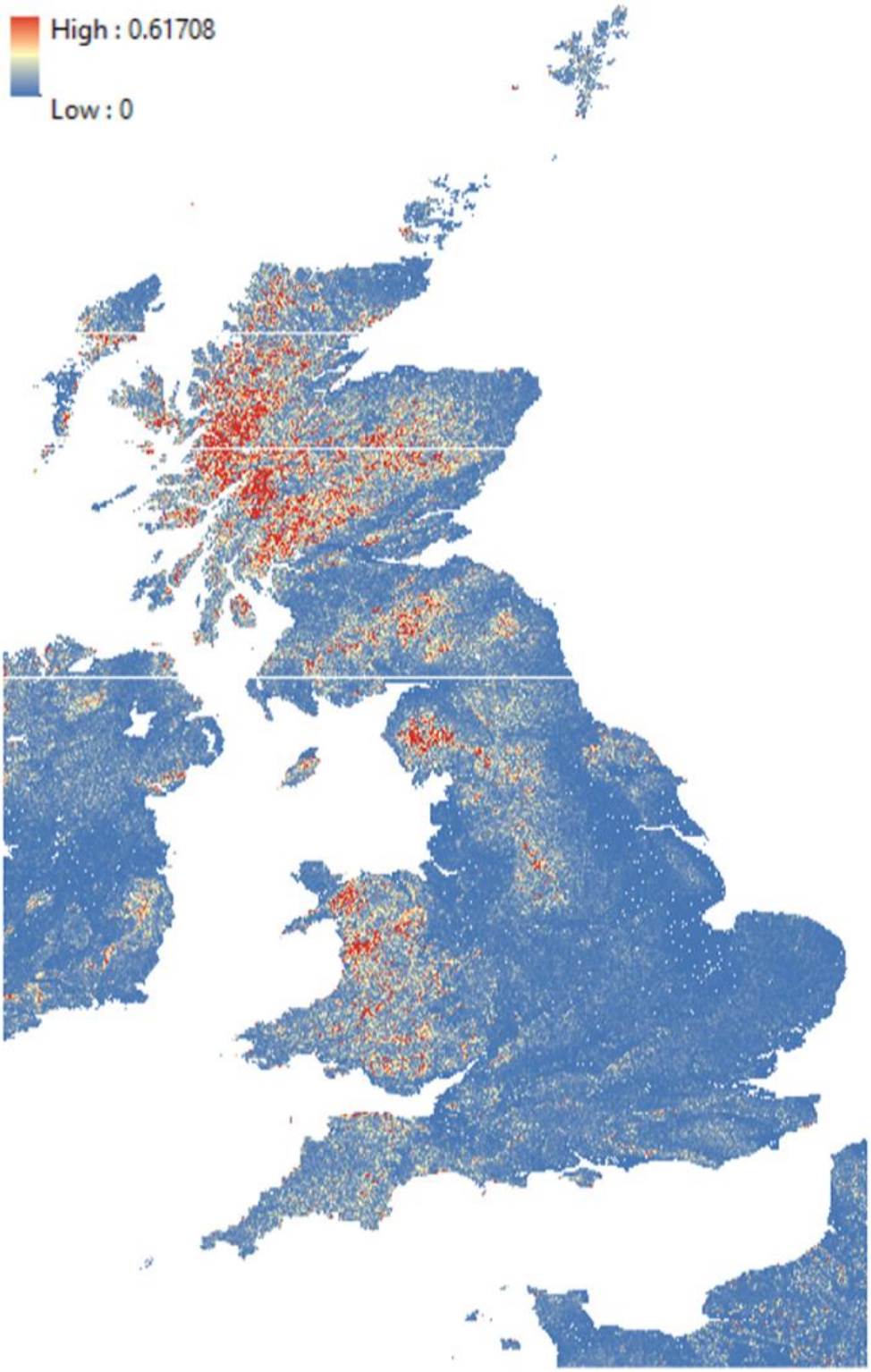


Figure 25. Map of contribution of slope to outbreak risk.

The map of the contribution of geological type to outbreak risk (Fig. 26) has values that lie within a narrower range than for Figures 24 and 25 and can be above or below zero depending on the geological type present. The north-south variation of values in this map is highly visible.

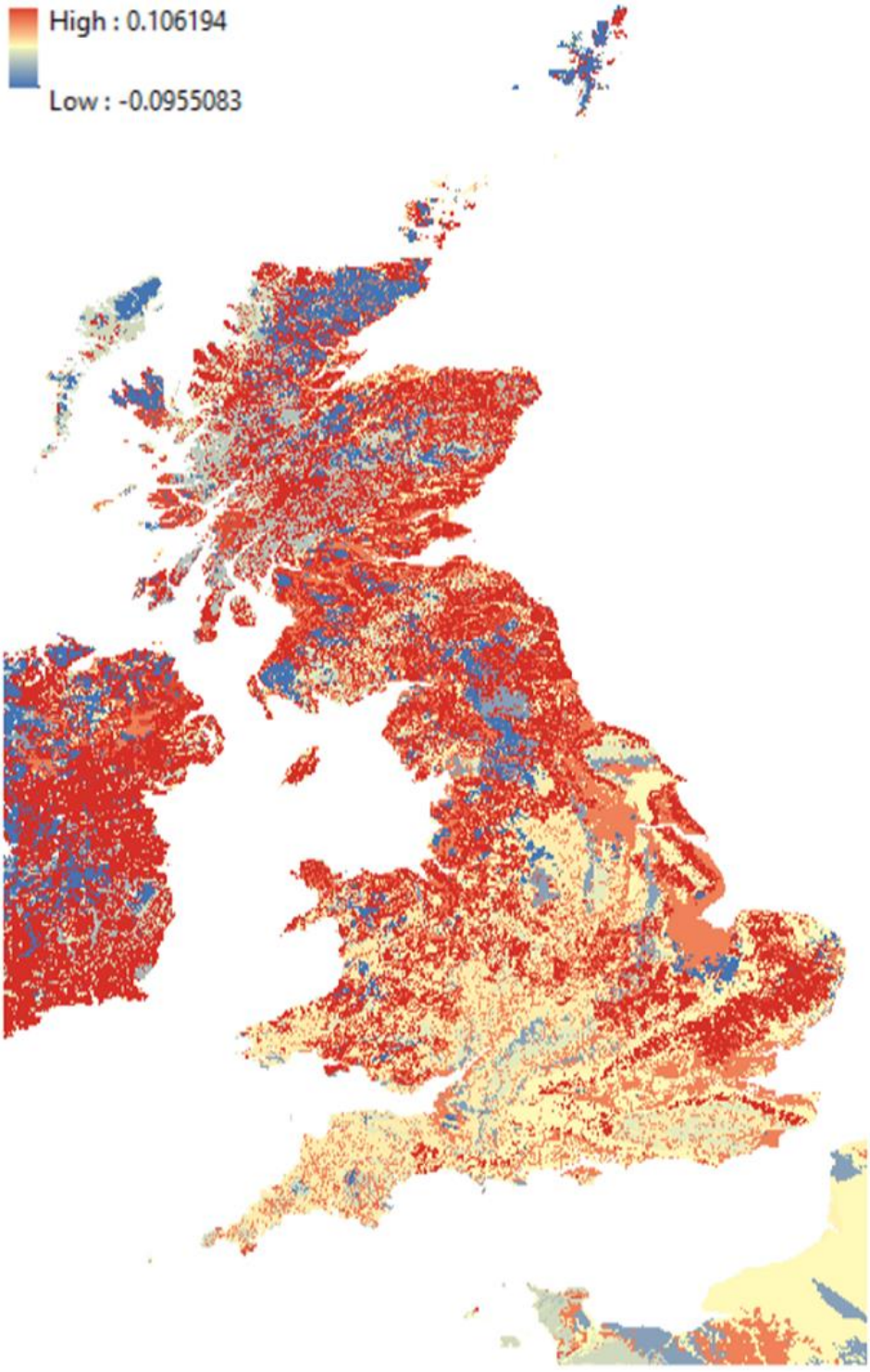


Figure 26. Map of contribution of geology to outbreak risk.

The map of the contribution of soil type to risk (Fig. 27) indicates distinct urban areas which follows on from the sensitivity analysis of no-soil types shown above (Fig. 22). Areas of water body are also highlighted and can be ignored. Areas of Luvisol with lower risk show up strongly in dark blue. We have not provided a map of the effects of soil diagnostic features on modelled risk as these values are generally smaller and lie within a narrow range with less meaning.

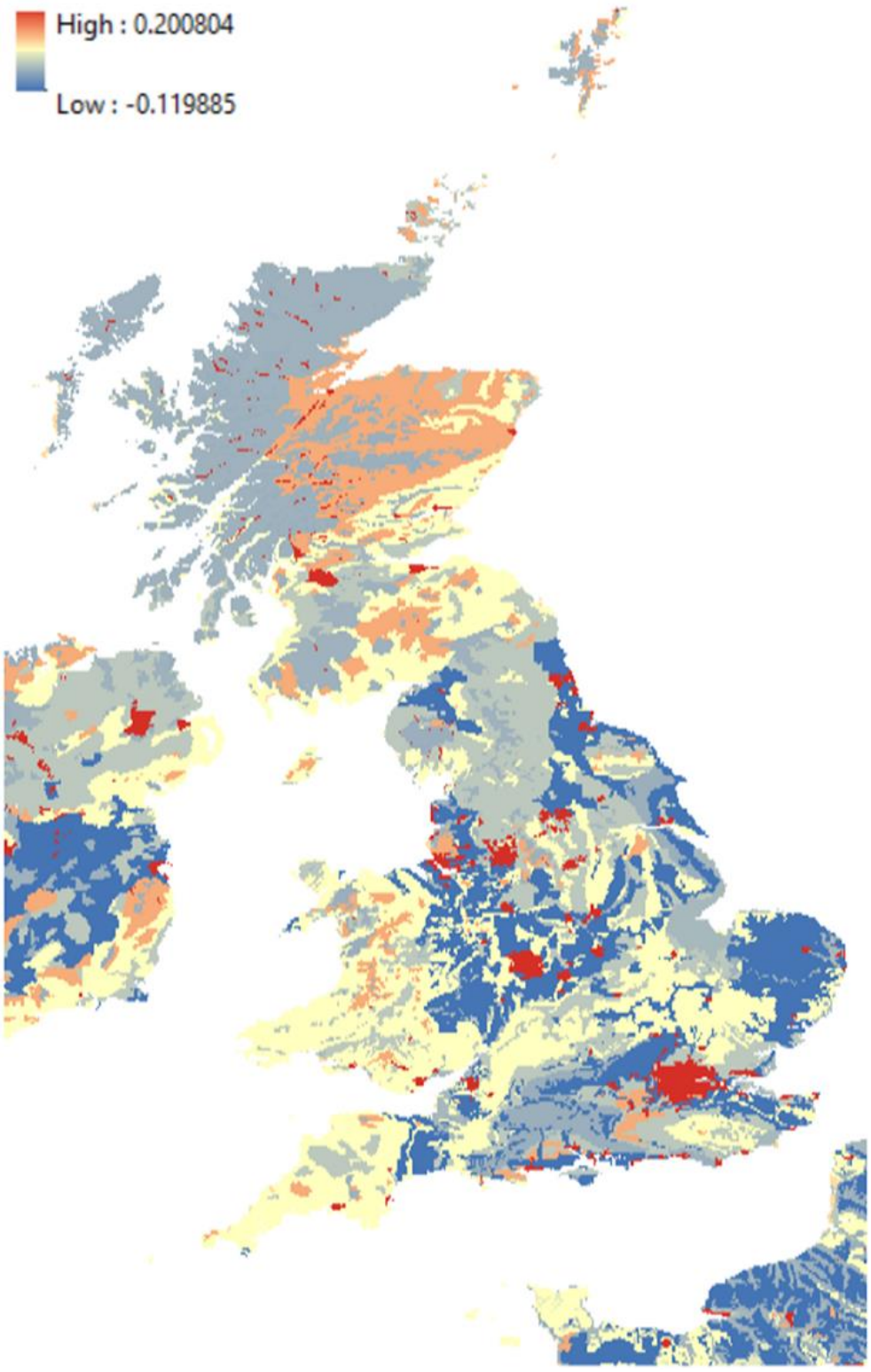


Figure 27. Map of contribution of soil class to outbreak risk.

The combined effects of topographic variables, geology and soil type/features are shown using a continuous scale from the minimum to maximum values (Fig. 28) and colour-coded by set ranges of values (Fig. 29). The issues with missing (below sea-level) data in the maps above is also seen in these summed maps. Missing values in East Anglia are, however, all in a narrow range on either side of zero.

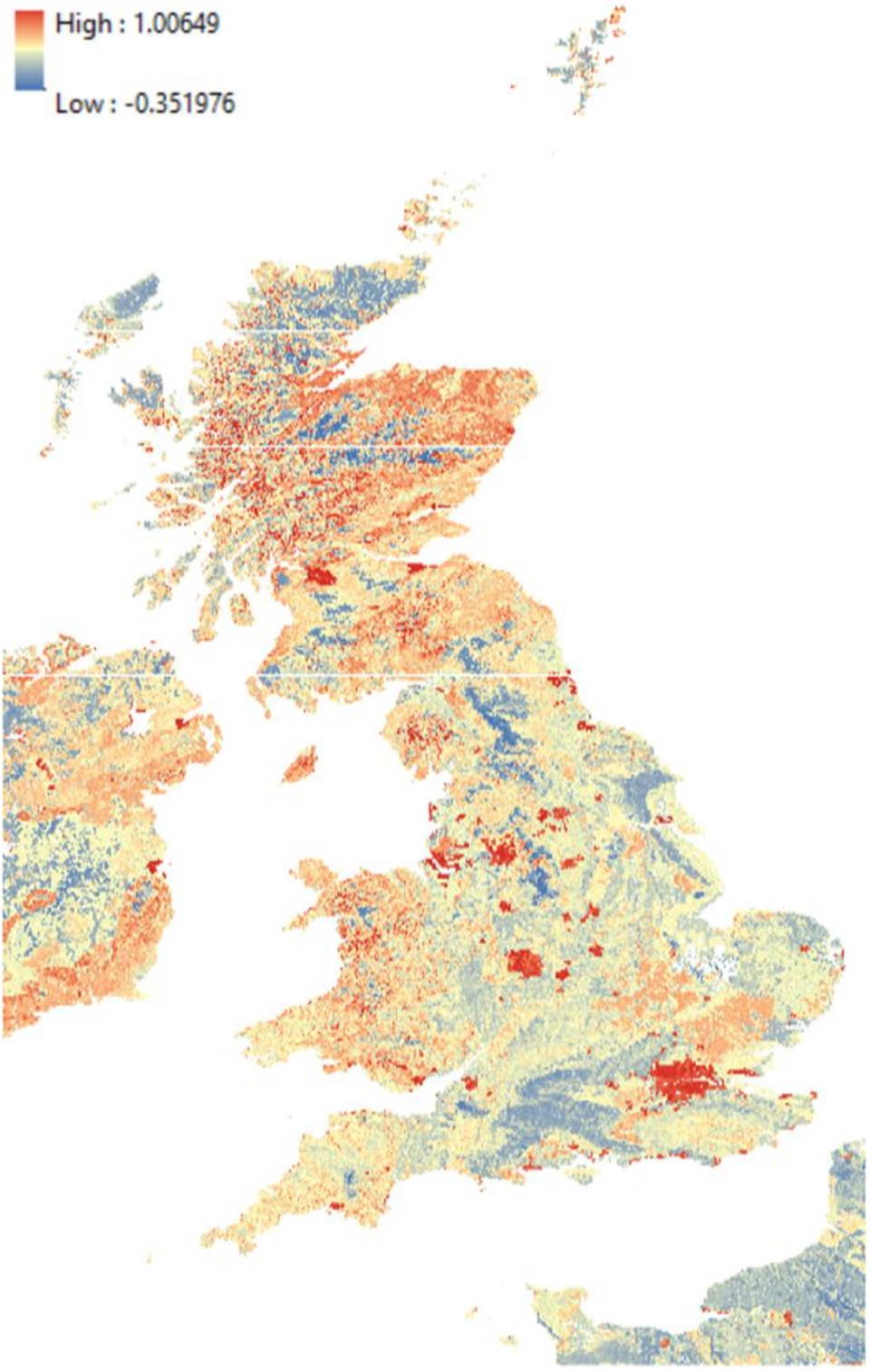


Figure 28. Map of combined contribution of topography, geology and soil to outbreak risk.

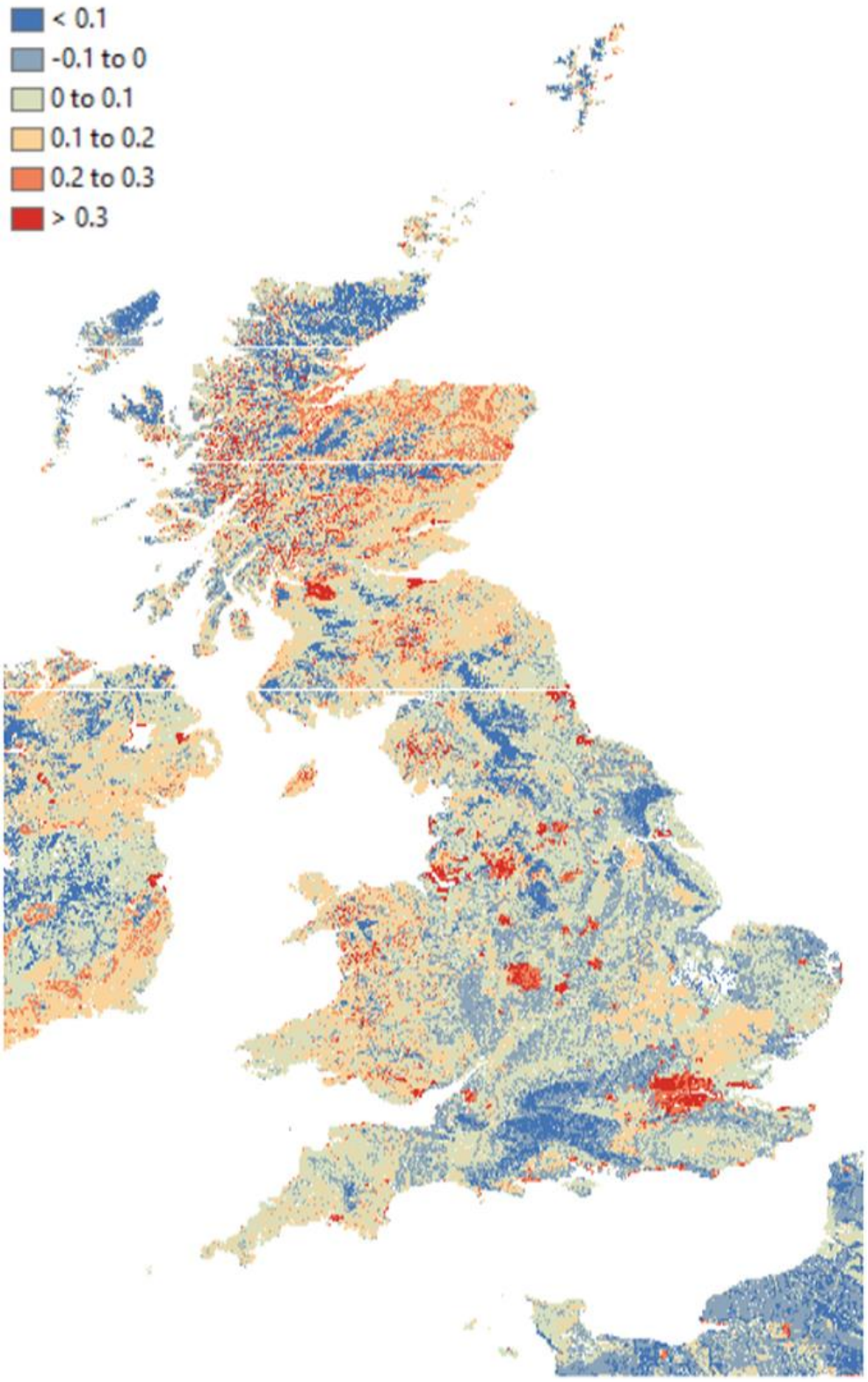


Figure 29. Category map of combined contribution of topography, geology and soil to outbreak risk. All low-lying areas in East Anglia not included in the map are in a narrow range of values around zero.

Deliverable 4: genetic makeup and spatial distribution of late blight pathogen genotypes

As shown on the table of sampled outbreaks the frequency of different *P. infestans* genotypes has changed over time. Genotypes 13_A2 and 6_A1 were by far the most common types amongst the FAB outbreak data (Table 1). This is followed by genotype 8_A1, which has persisted at a relatively low frequency. Some peak and then appear only sporadically (1_A1 and 2_A2) whereas others appeared late and are in the process of spreading more widely (37_A2 and 36_A2).

Table 1: Overview of the number of sampled blight outbreaks by year showing the records of the 10 most commonly observed genotypes at each outbreak.

		Genotypes									
Year	All	1_A1	2_A1	6_A1	7_A1	8_A1	23_A1	10_A2	13_A2	36_A2	37_A2
2006	162	20	18	18	5	18	0	11	78	0	0
2007	280	14	15	50	3	9	1	2	211	0	0
2008	203	1	3	31	2	8	0	0	181	0	0
2009	142	0	2	31	2	11	8	1	105	0	0
2010	82	0	4	20	0	6	4	0	48	0	0
2011	178	0	0	132	0	11	1	0	25	0	0
2012	344	4	0	228	3	16	2	0	104	0	0
2013	66	0	0	29	0	3	0	0	42	0	0
2014	257	0	1	171	0	8	0	0	84	0	0
2015	59	1	0	32	0	5	0	0	14	0	0
2016	171	0	0	111	1	3	5	0	41	0	3
2017	155	1	0	62	0	5	0	0	17	3	34
2018	68	1	0	29	0	0	0	0	8	12	17
Total	**2511**	**42**	**43**	**944**	**16**	**103**	**21**	**14**	**958**	**15**	**54**

The Directional Distribution (Standard Deviation Ellipse) tool was used to summarise the central tendency, dispersion, and directional trends of the main genotypes of interest across the study period (2006–2018) (Fig. 30).

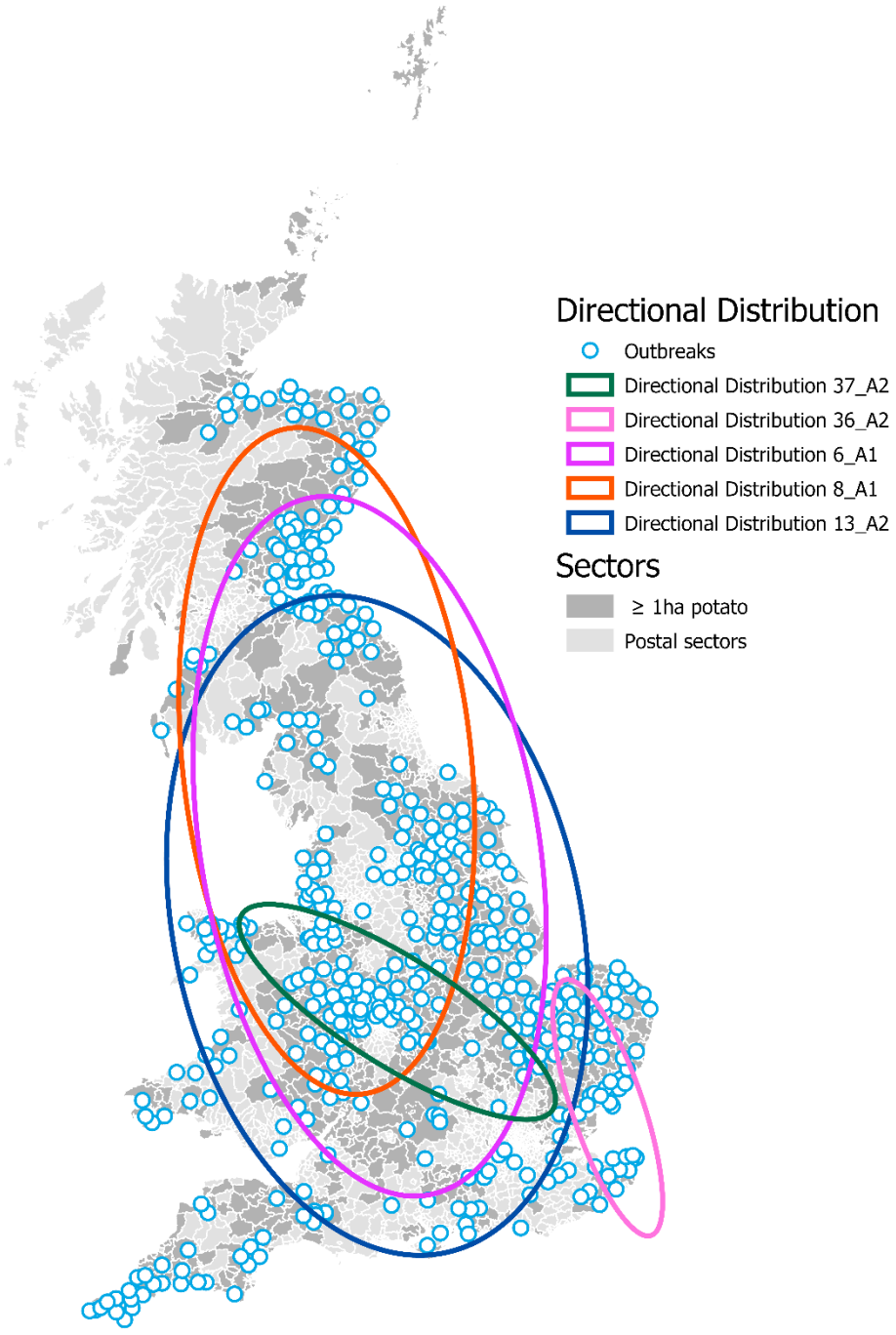


Figure 30. Spatial characteristics of the different distributions of genotypes: the central tendency, dispersion, and directional trends (2006–2018).

This analysis reveals that genotypes 13_A2, 6_A1 and 8_A1 share a similar orientation, but their average dispersions vary slightly in north-south orientation with 8_A1 having the most northerly and 13_A2 the most southerly means. This may indicate slightly different responses to climate or historical patterns of distribution (Appendix 2). Genotype 37_A2 appears to be spreading from the west to the east, whereas up until 2018, genotype 36_A2 was yet to become established beyond the southeast of GB. It has since spread northwards to Scotland, but the 2019 data is not included in this analysis.

Choropleth (colour-coded) maps were prepared to provide a simple summary of the total number of outbreaks in each postcode district for each of the genotypes studied (Figs. 31–35). They indicate the widespread distribution of 13_A2, 6_A1 and 8_A1 with the latter being at a significantly lower frequency. Genotypes 36_A2 and 37_A2, in comparison, remain more closely clustered to the postcodes in which they were first sampled.

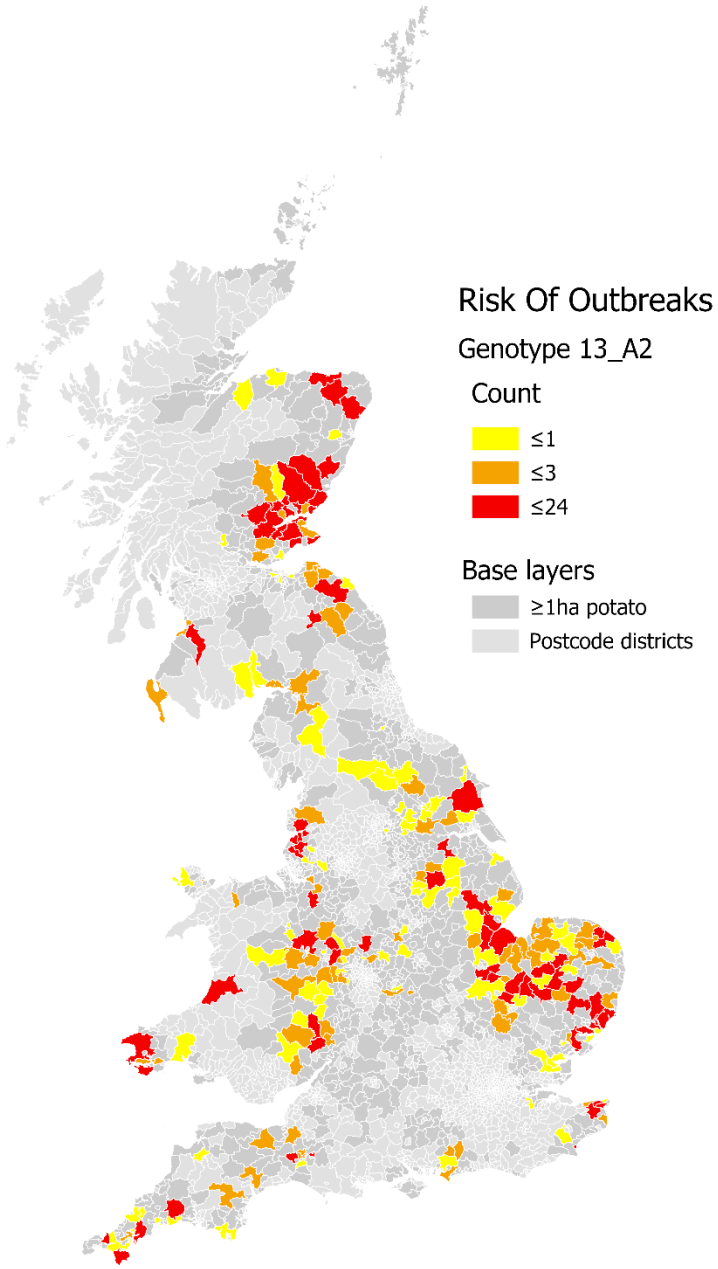


Figure 31. Choropleth map showing a count of all outbreaks of genotype 13_A2 within postcode districts containing >1ha potato (grown commercially), 2006–2018.

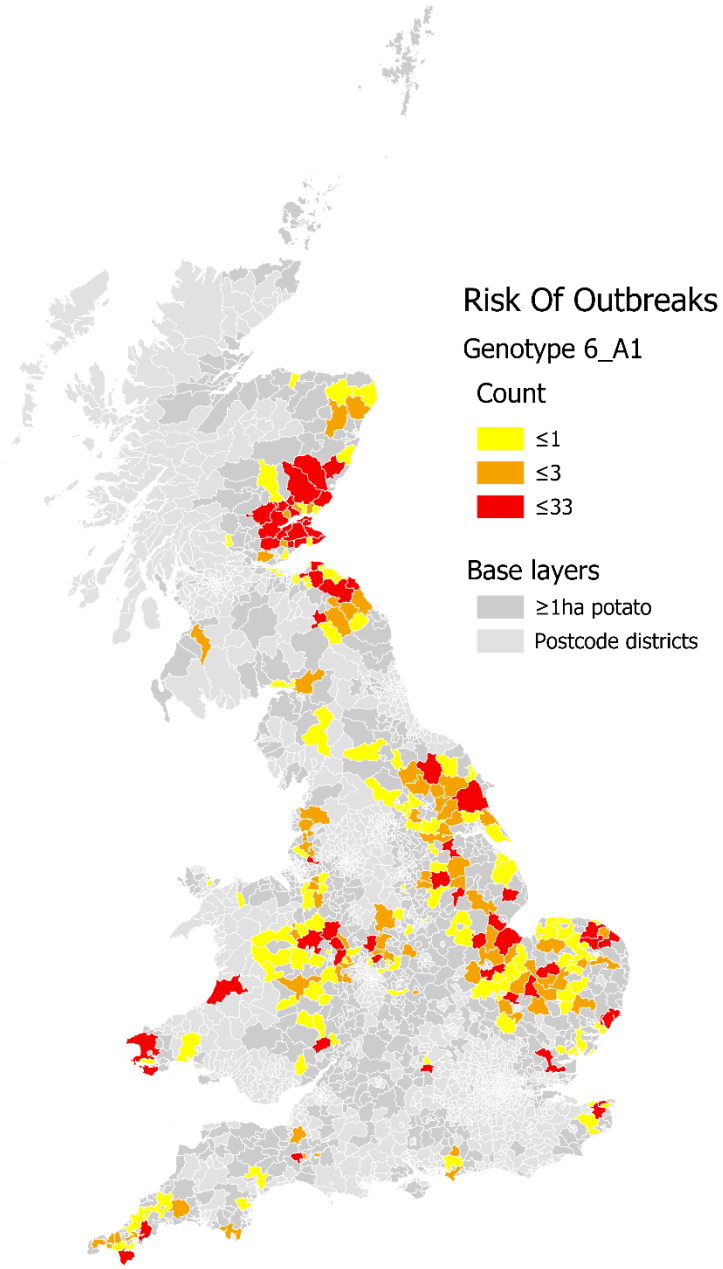


Figure 32. Choropleth map showing a count of all outbreaks of genotype 6_A1 within postcode districts containing >1ha potato (grown commercially), 2006–2018.

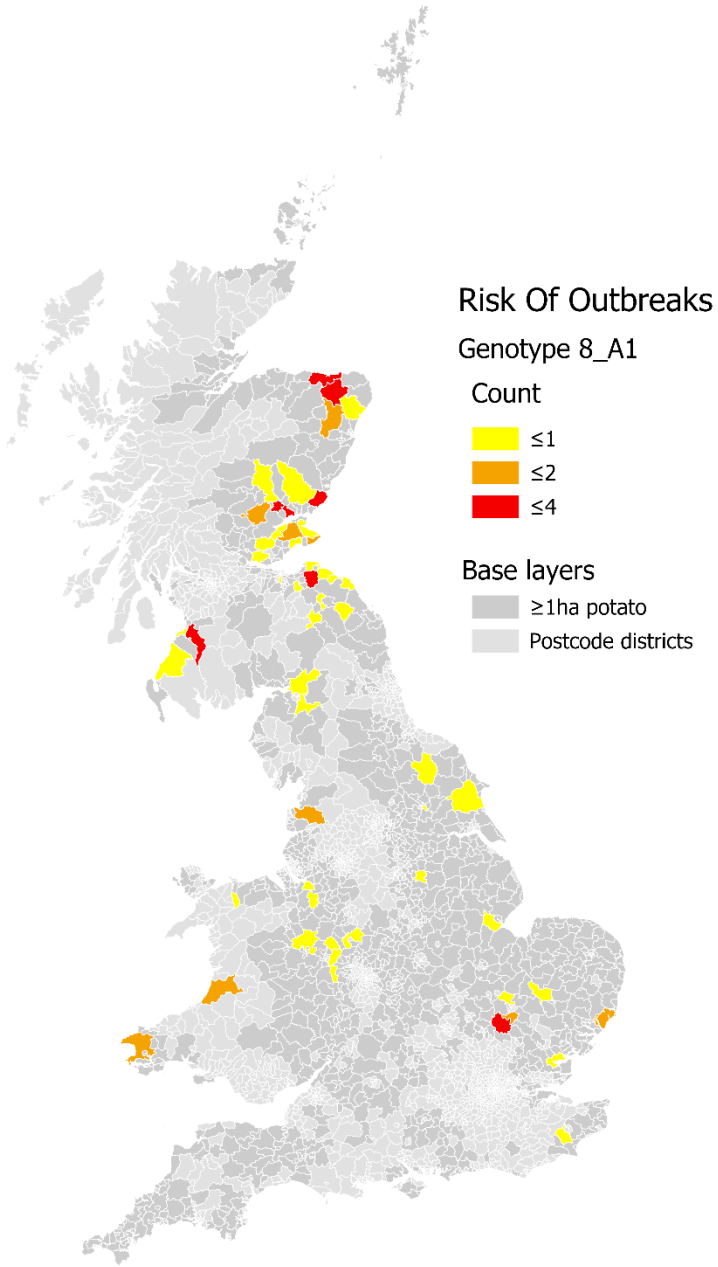


Figure 33. Choropleth map showing a count of all outbreaks of genotype 8_A1 within postcode districts containing >1ha potato (grown commercially), 2006–2018.

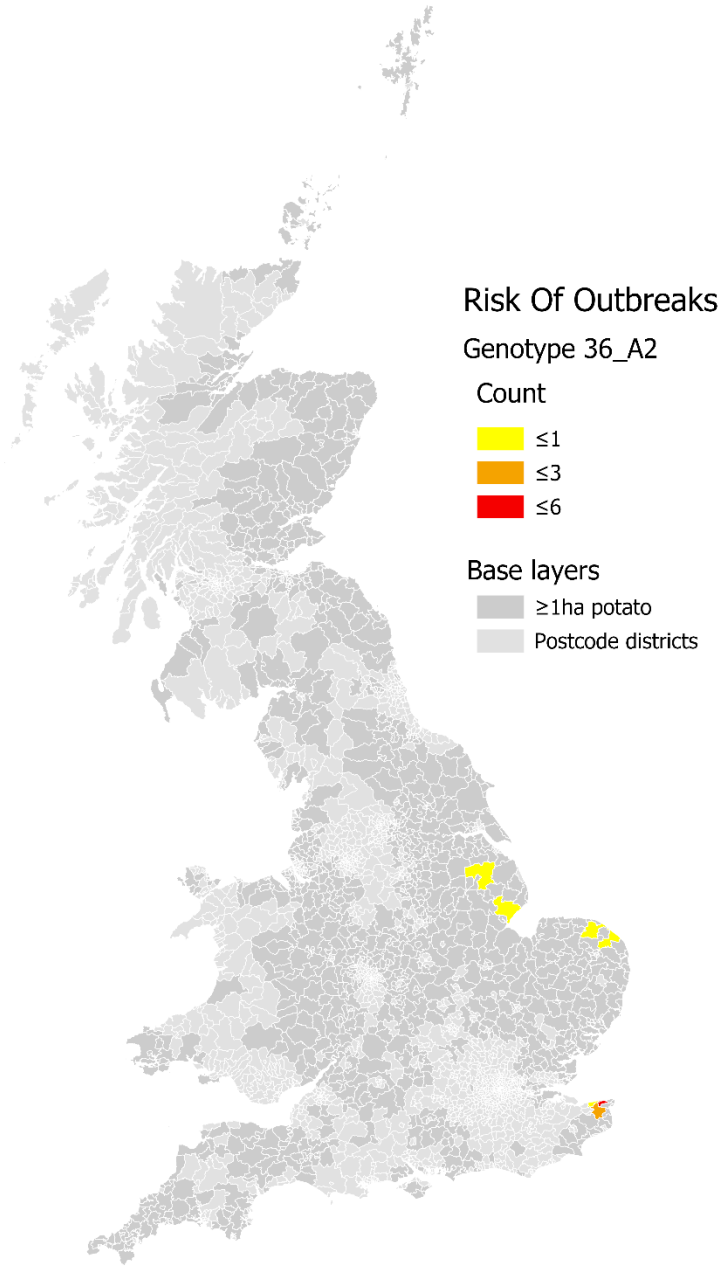


Figure 34. Choropleth map showing a count of all outbreaks of genotype 36_A2 within postcode districts containing >1ha potato (grown commercially), 2017–2018.

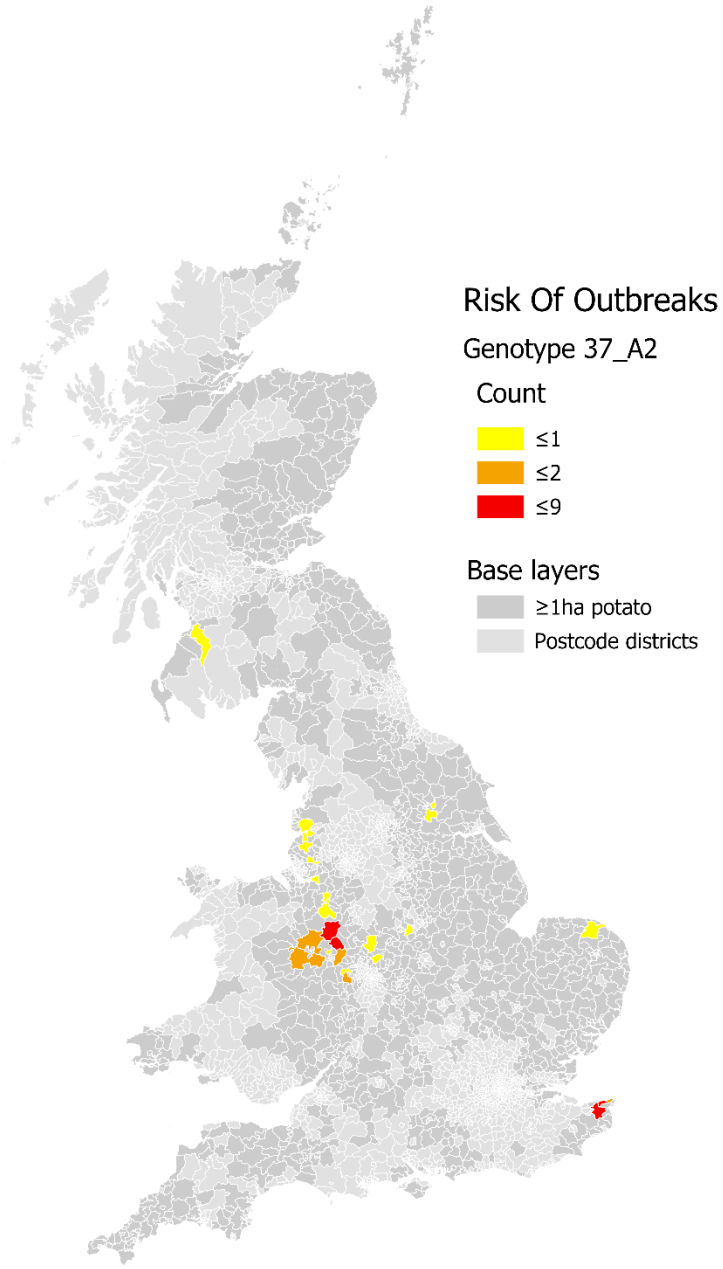


Figure 35. Choropleth map showing a count of all outbreaks of genotype 37_A2 within postcode districts containing >1ha potato (grown commercially), 2016–2018.

The OHSA provides information on the statistical significance of the clustering of high and low values of incidence identified in the above choropleth maps (Figs. 36–40). The results differed markedly for each of the genotypes studied. Although incidence of genotype 13_A2 was high in Scotland in the first half of the study period it declined from 2011 onwards with the consequence that the statistically significant hot spots occurred mainly in England and Wales (Fig. 36). The converse is true for genotype 6_A1 with hotspots in Scotland and northeast England. The drivers of this are not clear but may involve competition between these two genotypes or local expansion or collapse of populations (Fig. 37). Genotype 8_A1 was

sampled at a lower frequency but appears alongside 13_A2 in Wales and 6_A1 in Scotland, albeit in the more south-westerly and northern production regions of Scotland, whereas 6_A1 dominates along the eastern seaboard (Fig. 38). Genotype 36_A2 has been confined mostly to the south-eastern tip of GB since it was first reported in 2017 but is steadily moving northwards (Fig. 39). Genotype 37_A2 has spread from the Midlands across to the east and southeast of England and northwards into Scotland (in 2018). The incidences were however, not sufficiently to produce hot spots beyond the Midlands and Kent (Fig. 40).

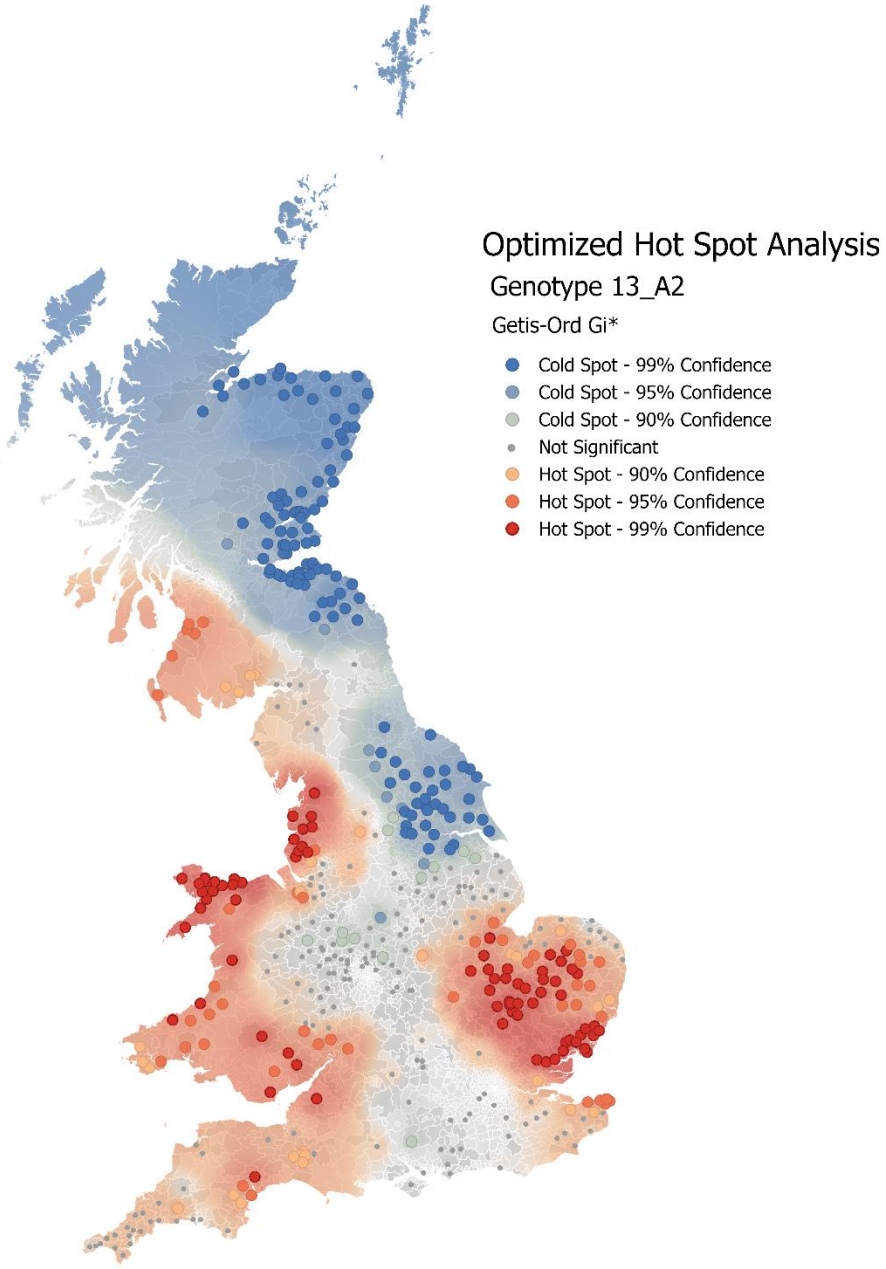


Figure 36. Statistically significant hot and cold spots derived by OHSA for genotype 13_A2, 2006-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

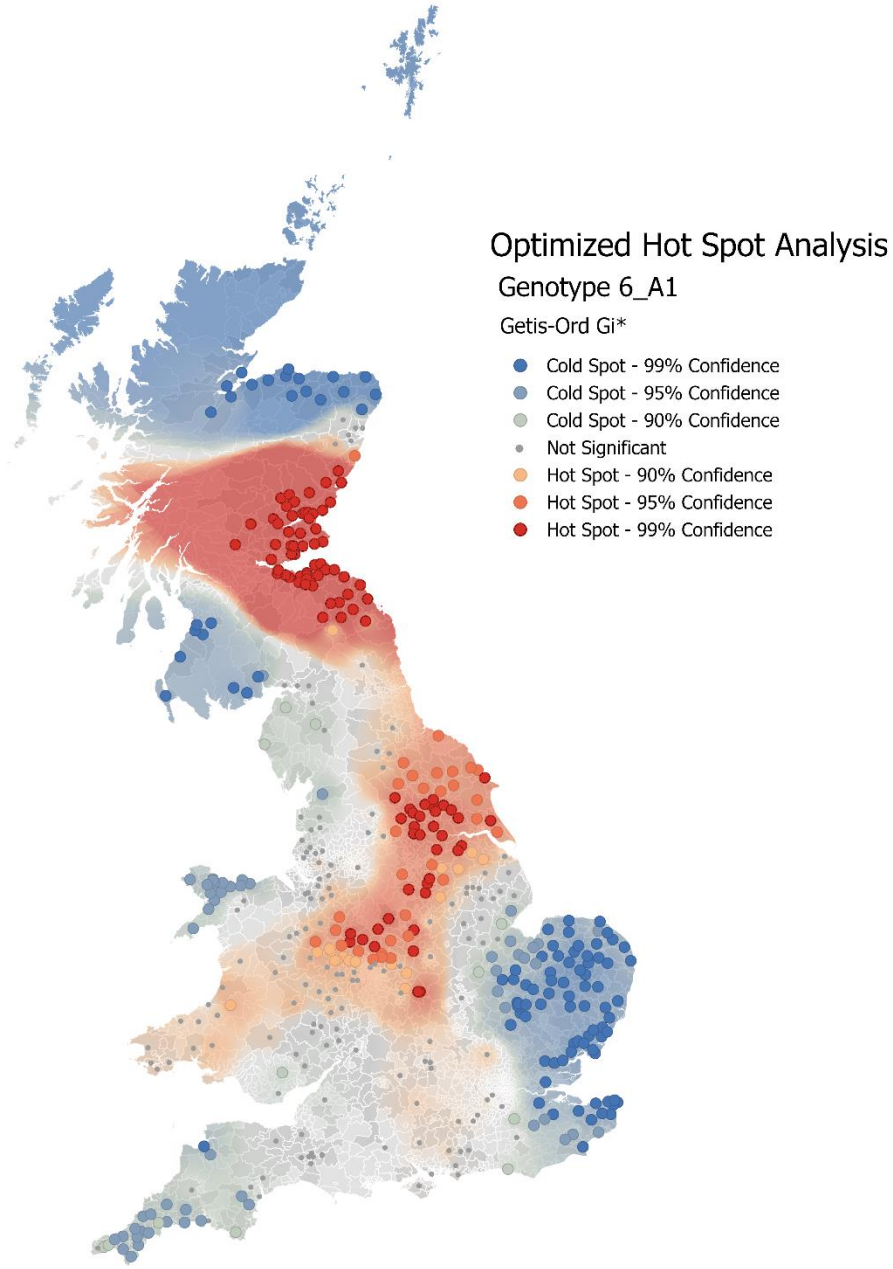


Figure 37. Statistically significant hot and cold spots derived by OHSA for genotype 6_A1, 2006-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

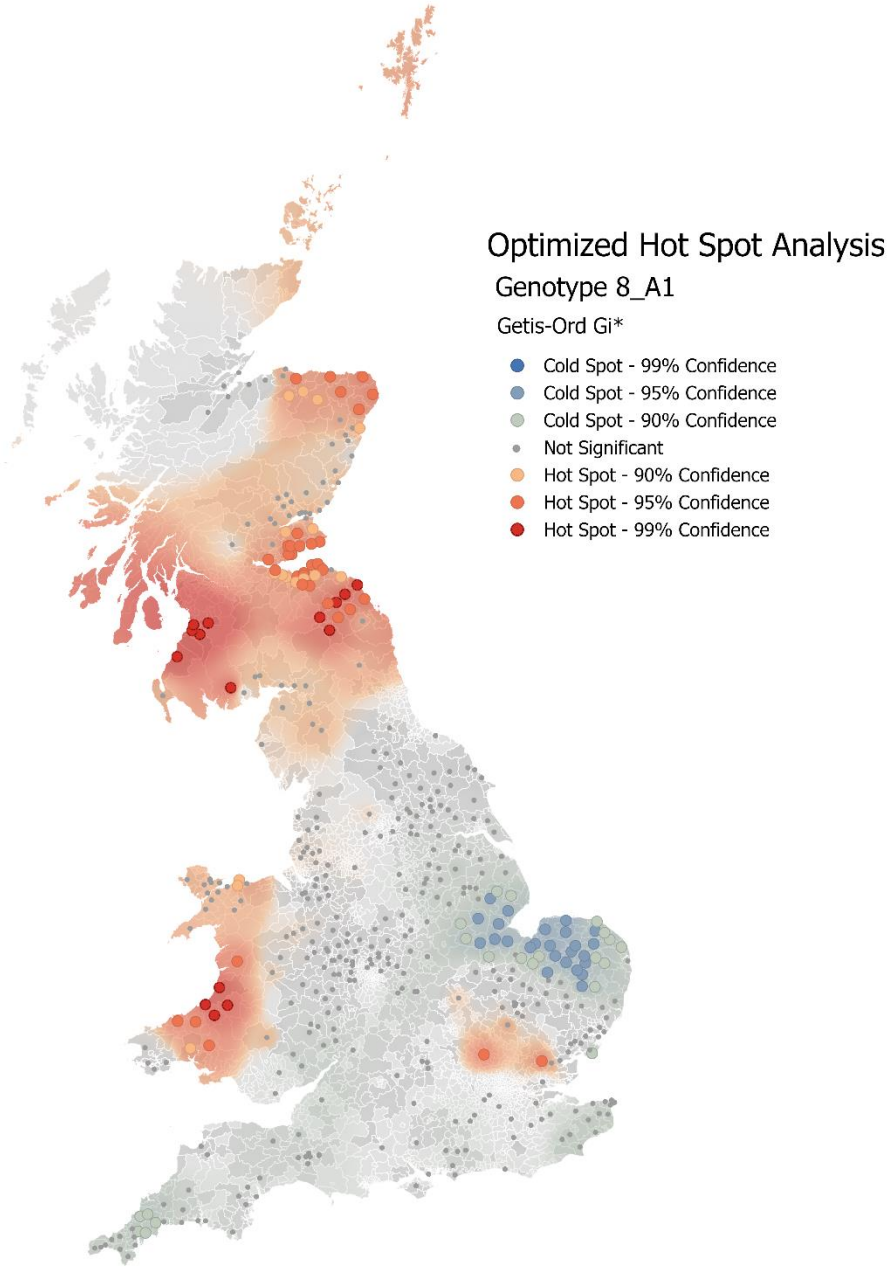


Figure 38. Statistically significant hot and cold spots derived by OHSA for genotype 8_A1, 2006-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

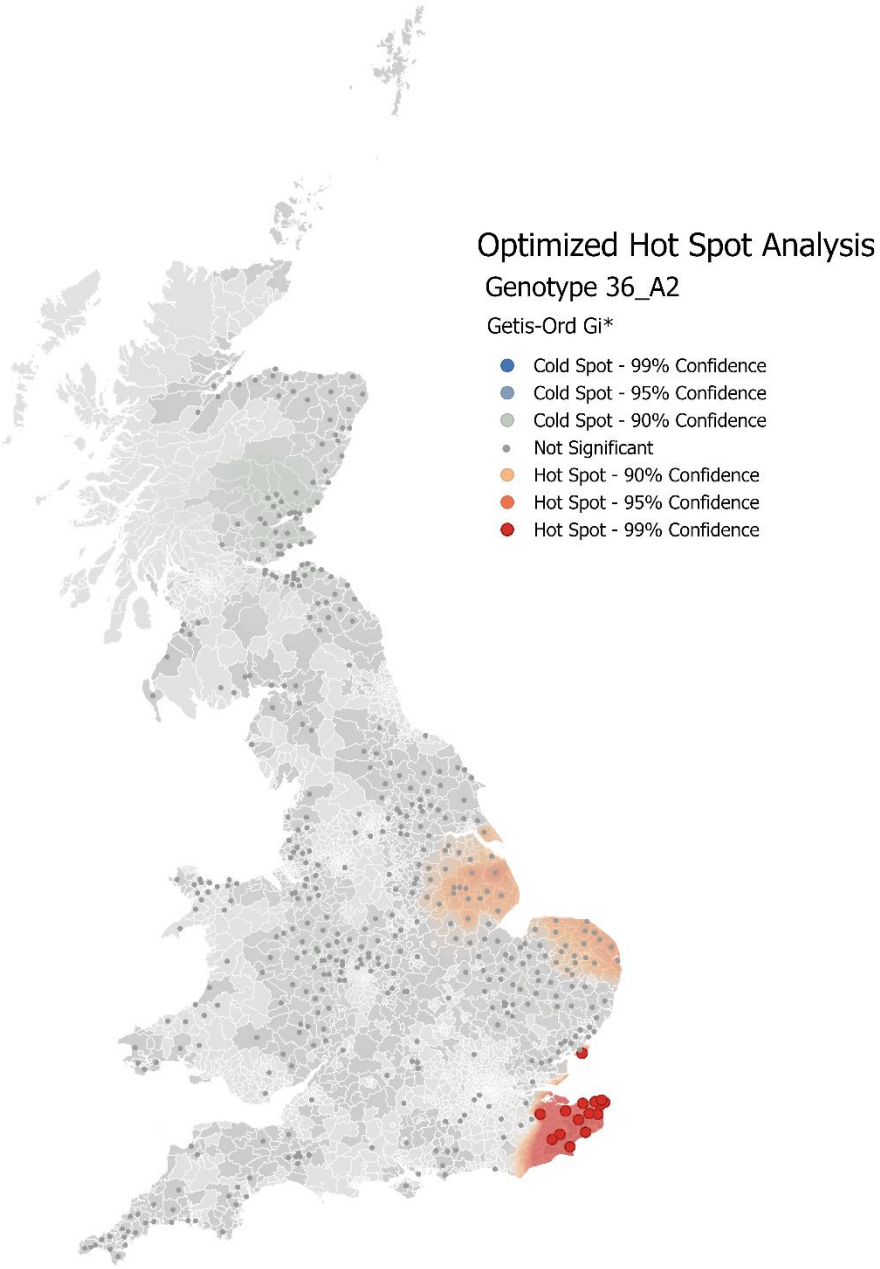


Figure 39. Statistically significant hot and cold spots derived by OHSA for genotype 36_A2, 2017-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

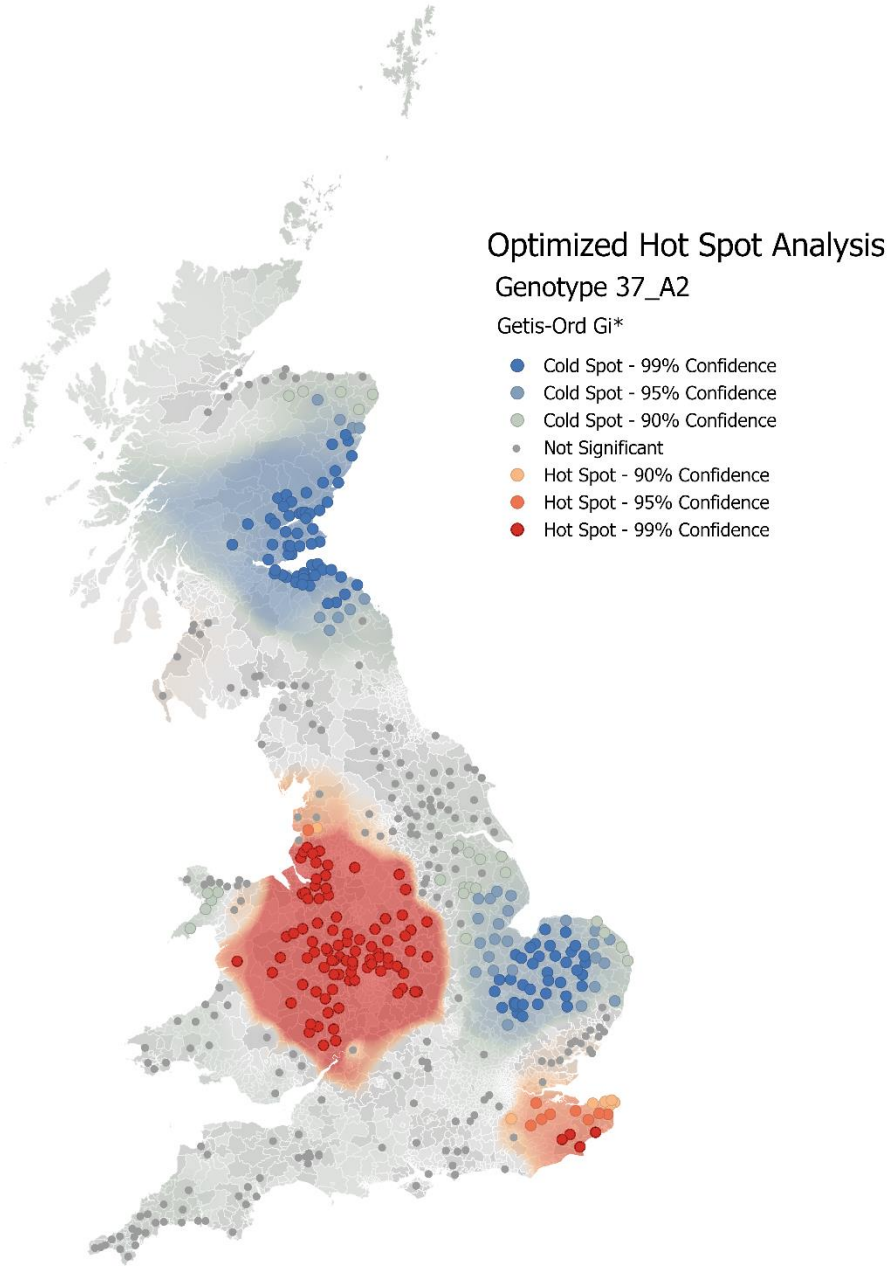


Figure 40. Statistically significant hot and cold spots derived by OHSA for genotype 37_A2, 2016-2018. Inverse distance weighting was used to interpolate a coloured raster surface from the OHSA points.

The EHSA corresponded well to the OHSA and facilitated an analysis of temporal trends in the patterns identified (Figs. 41–45). No cold spots were identified, and hot spots were of three types: consecutive, sporadic, and new (Table 2).

Table 2: Summary of Emerging Host Spot Analyses for late blight outbreaks, grouped by genotype, 2006–2018.

Genotype	Consecutive hot spots	Sporadic hot spots	New hot spots
13_A2	35	4	0
6_A1	50	32	0
8_A1	34	0	0
36_A2	16	24	1
37_A2	72	6	2

The incidence of genotype 13_A2 has been most frequent and consistent in Lancashire and East Anglia in recent years, leading to consecutive hot spots in those regions (Fig. 41). Whereas the OHSA did not identify a hot spot in the Shropshire region, the EHSA revealed four postcode districts in that area where sporadic hot spots occur. The results of the EHSA indicate that genotype 6_A1 has become established in the Fife/Tayside/Angus regions of Scotland and in Shropshire and East Anglia in England in recent years, leading to consecutive hot spots in those areas (Fig. 42). There were many sporadic hot spots in Anglia and Kent, indicating that the occurrence of this genotype is quite irregular and variable there. The EHSA flags up the broad spread of both these genotypes by around 2011 followed by local transitions with complex drivers. Genotype 8_A1 appears to be most prevalent in Scotland, albeit at a comparatively low frequency (Fig. 43) and may reflect its later displacement by 13_A2 and 6_A1. The newly introduced genotypes 36_A2 and 37_A2 appear to be firmly established in their centres of establishment in the most recent time steps analysed (Figs. 44 & 45). Of most interest to the potato industry are the new hot spots appearing towards the north for both genotypes, indicating spatial spread.

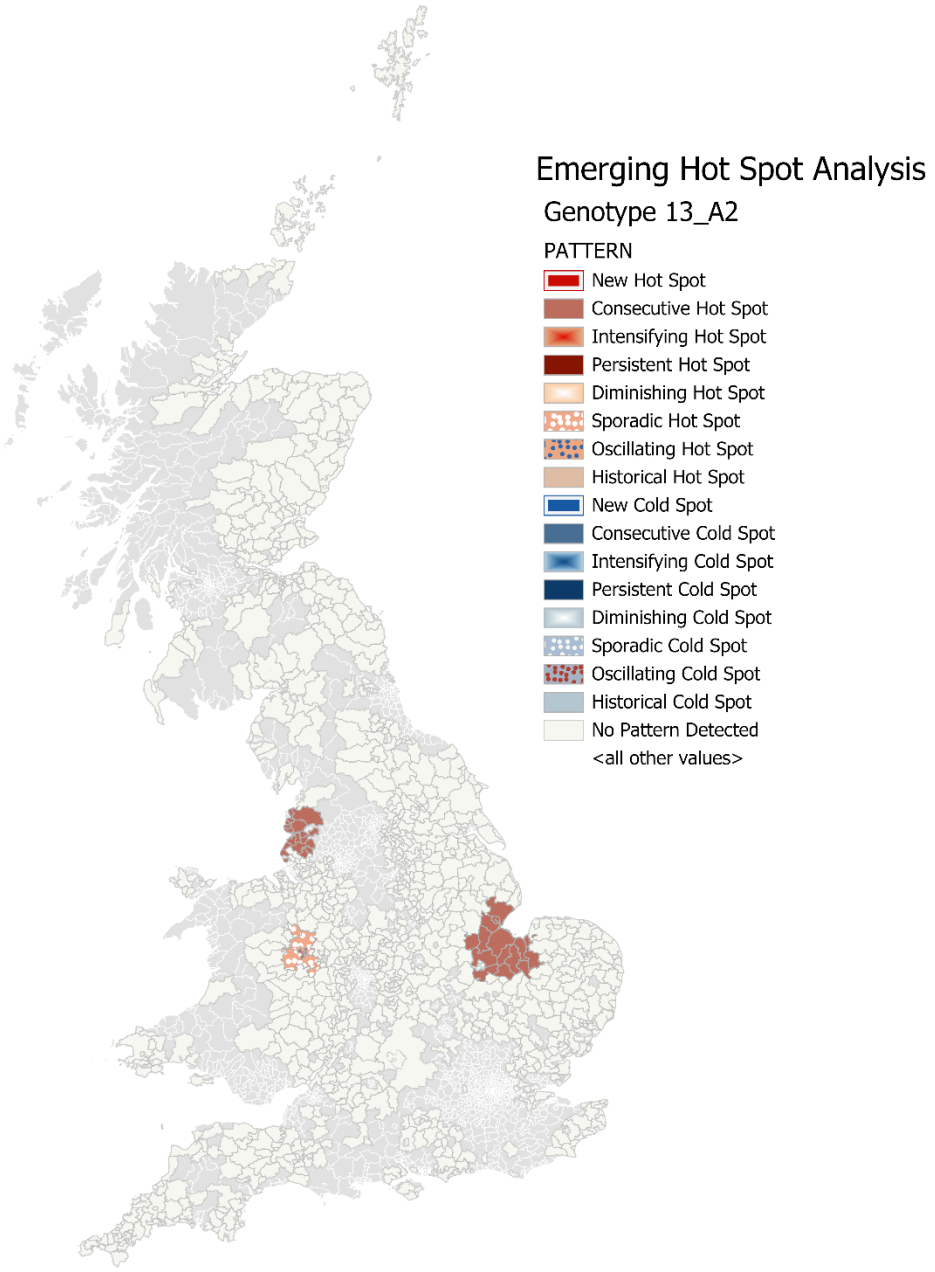


Figure 41. Space-time patterns of genotype 13_A2 incidence derived by EHSA, 2006–2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

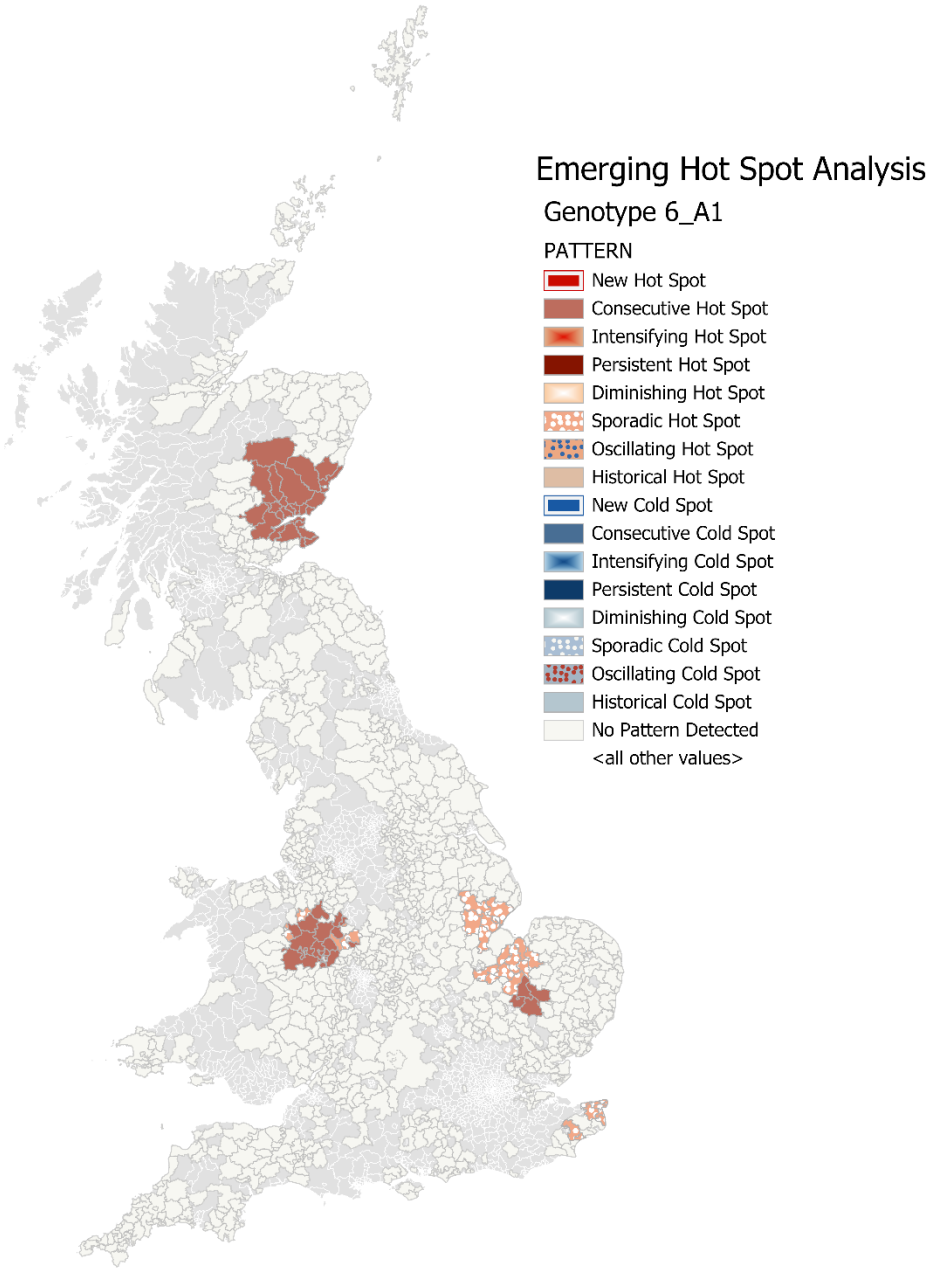


Figure 42. Space-time patterns of genotype 6_A1 incidence derived by EHSA, 2003–2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

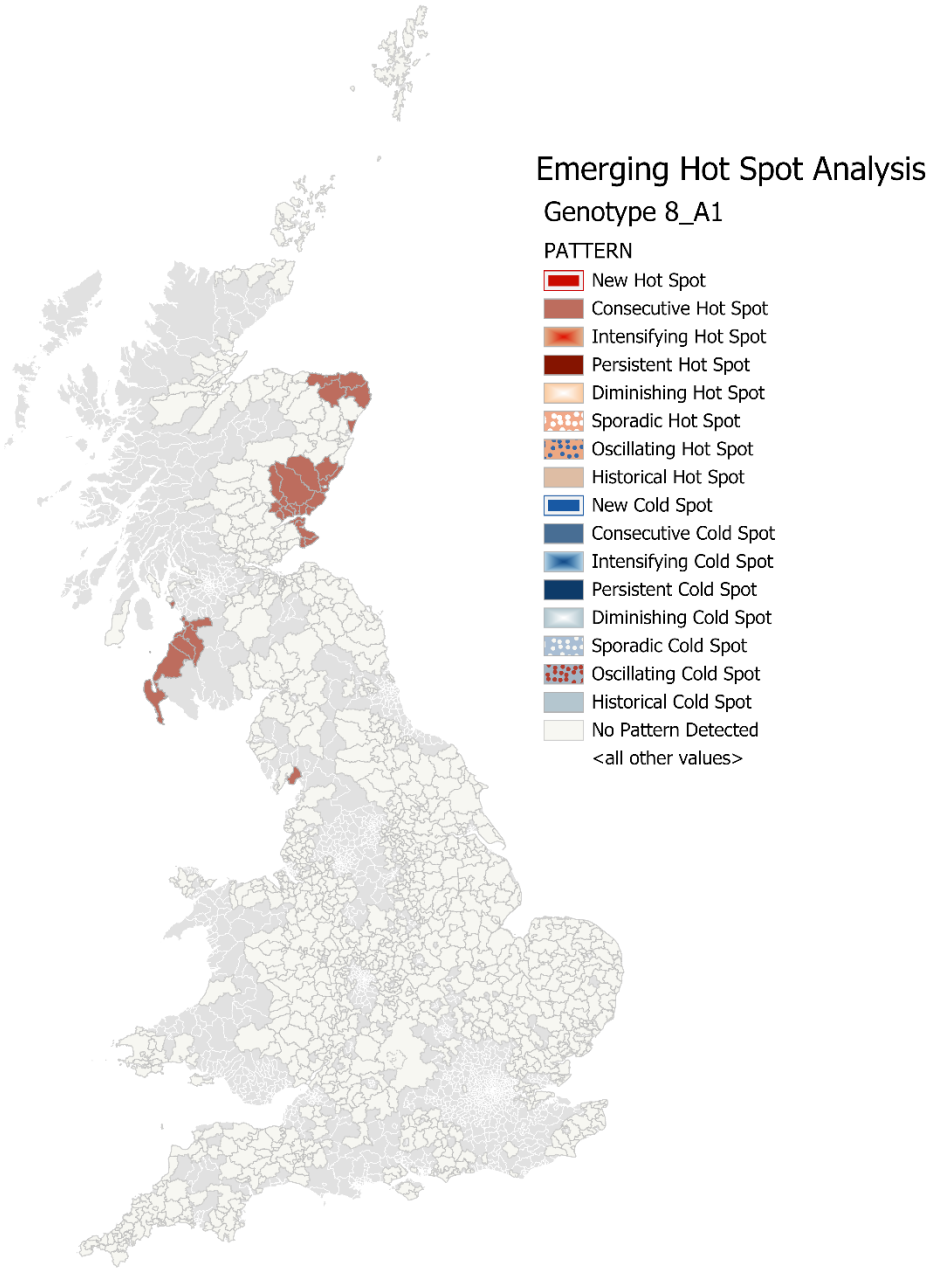


Figure 43. Space-time patterns of genotype 8_A1 incidence derived by EHSA, 2003–2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

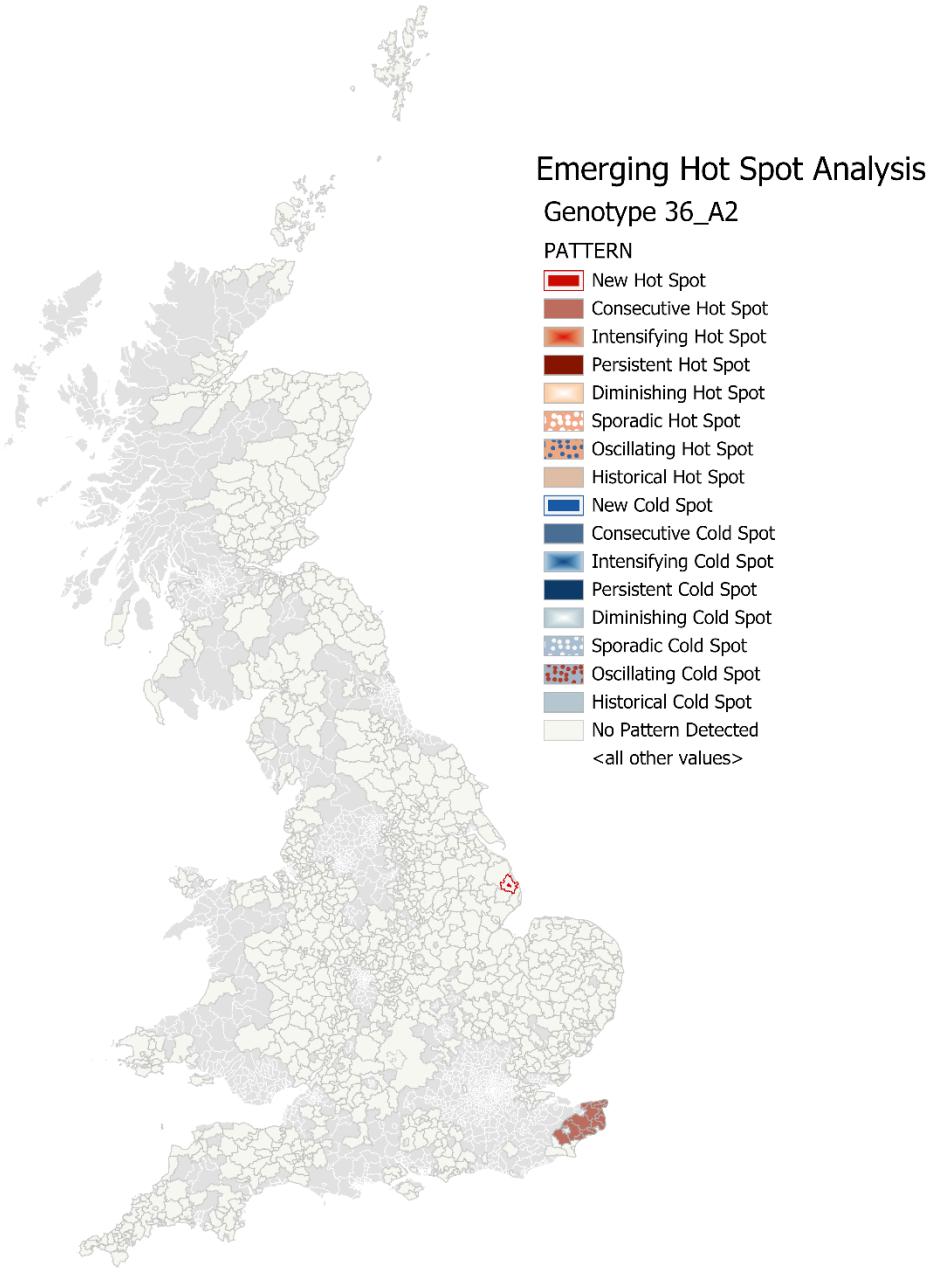


Figure 44. Space-time patterns of genotype 36_A2 incidence derived by EHSA, 2017–2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

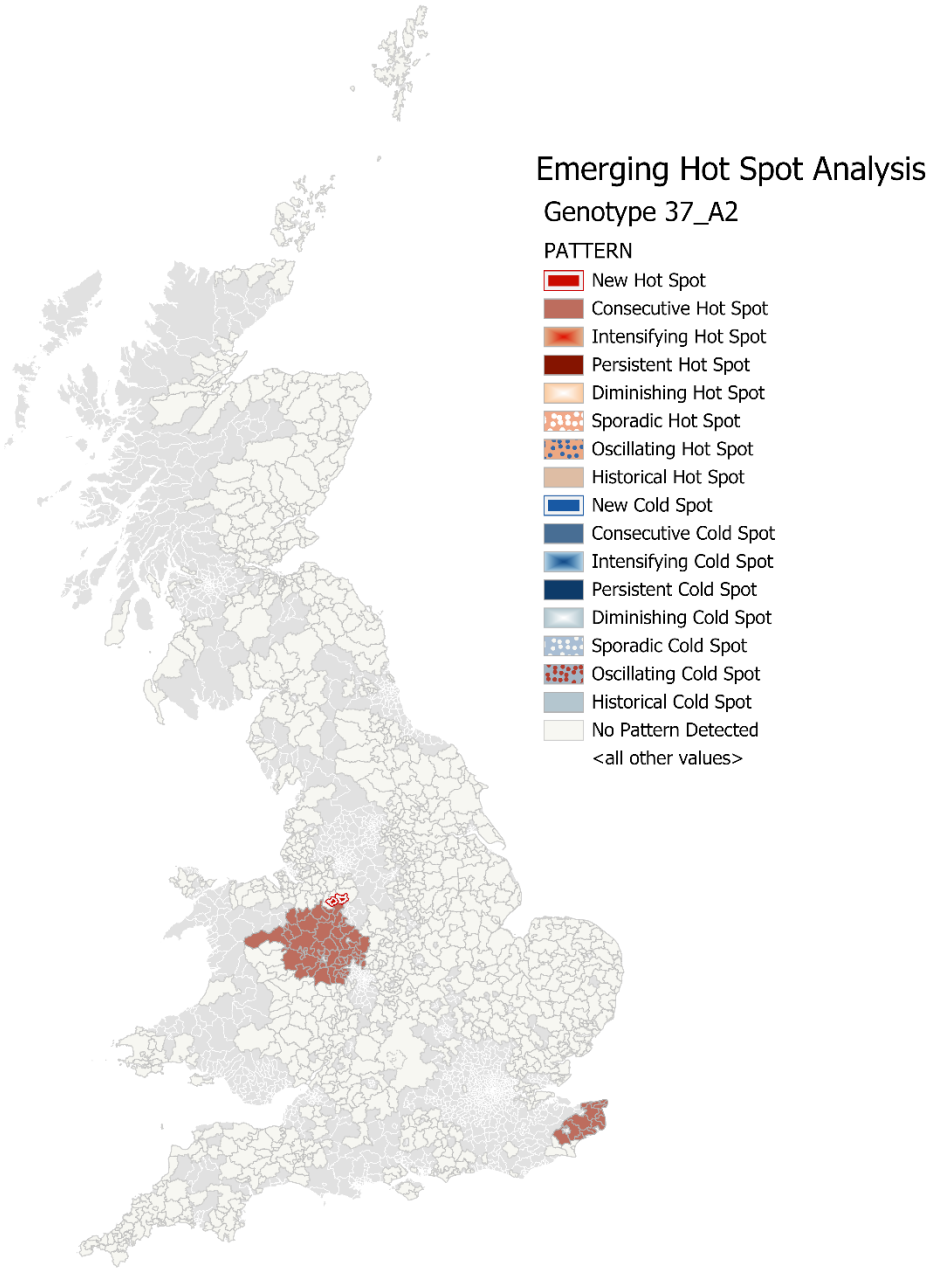


Figure 45. Space-time patterns of genotype 37_A2 incidence derived by EHSA, 2016–2018. Postcode districts with no reported commercial potato crops are shaded pale grey and were excluded from the analysis.

To visualise the temporal change in genotype distributions underlying the OHSA and EHSA analyses, we performed a KDE analysis to produce continuous surfaces of incidence intensity for 13_A2, 6_A1 and 8_A1 for 2006–2017 (Figs. 46-49). Note that 2010, 2013 and 2015 were lower blight intensity years with 82, 68 and 58 outbreaks reported, respectively. We did not plot 2018 data as the weather conditions were not favourable for blight and only 63 outbreaks were reported. Overall, the results agreed well with the OHSA.

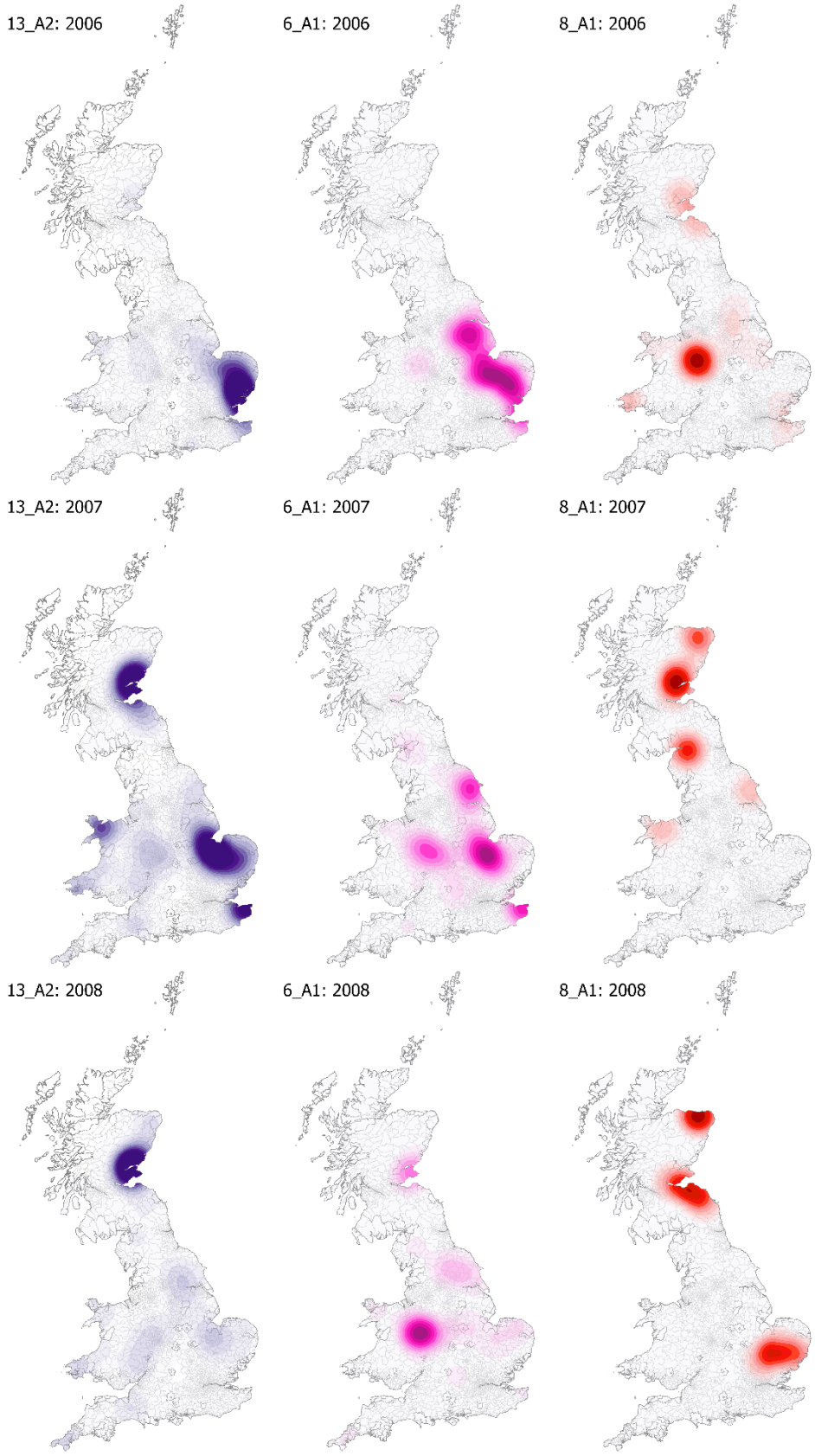


Figure 46. Kernel density distributions of late blight incidence showing inter-annual variation for genotypes 13_A2, 6_A1, and 8_A1, 2006–2008.

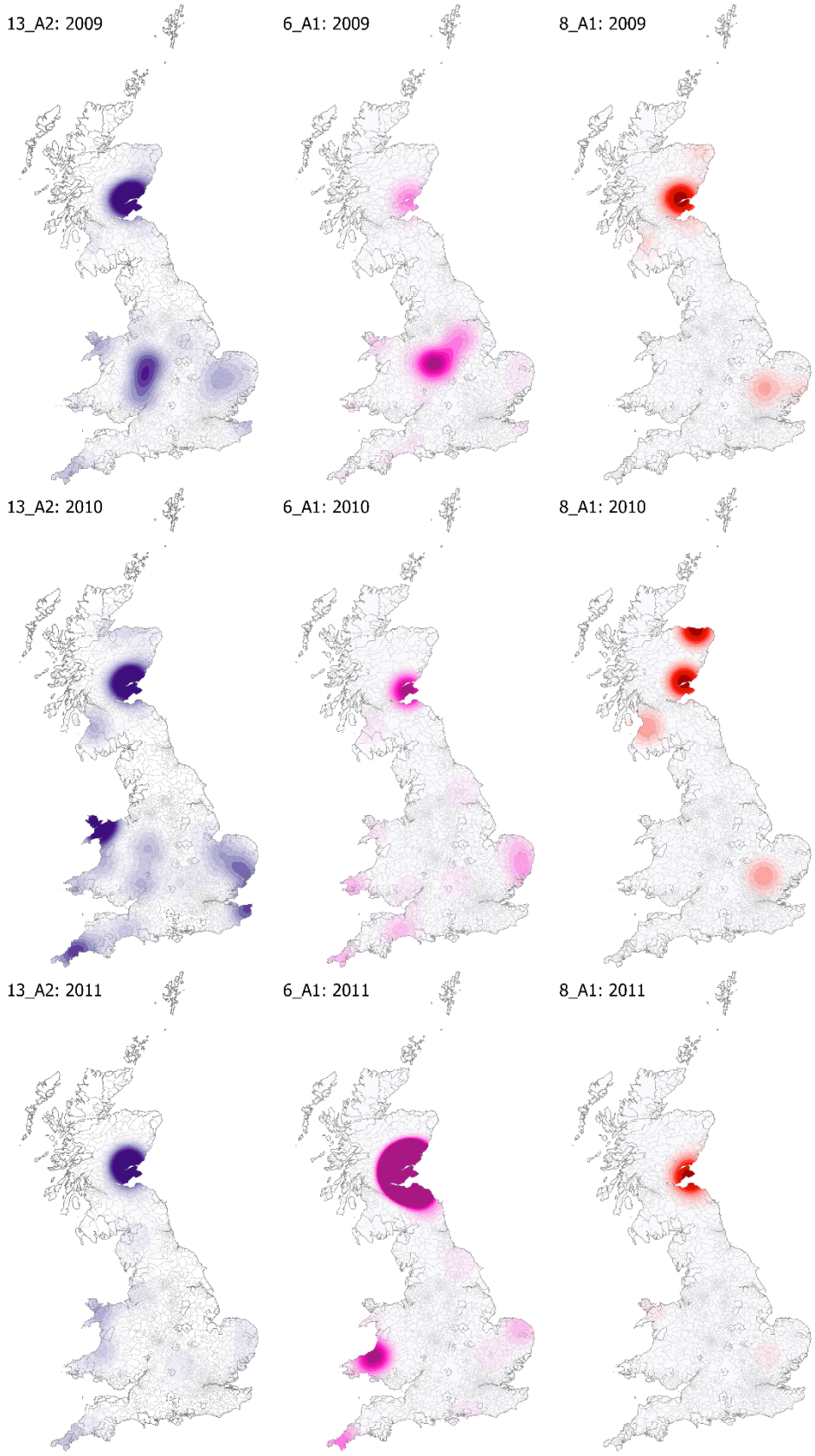


Figure 47. Kernel density distributions of late blight incidence showing inter-annual variation for genotypes 13_A2, 6_A1, and 8_A1, 2009–2011.

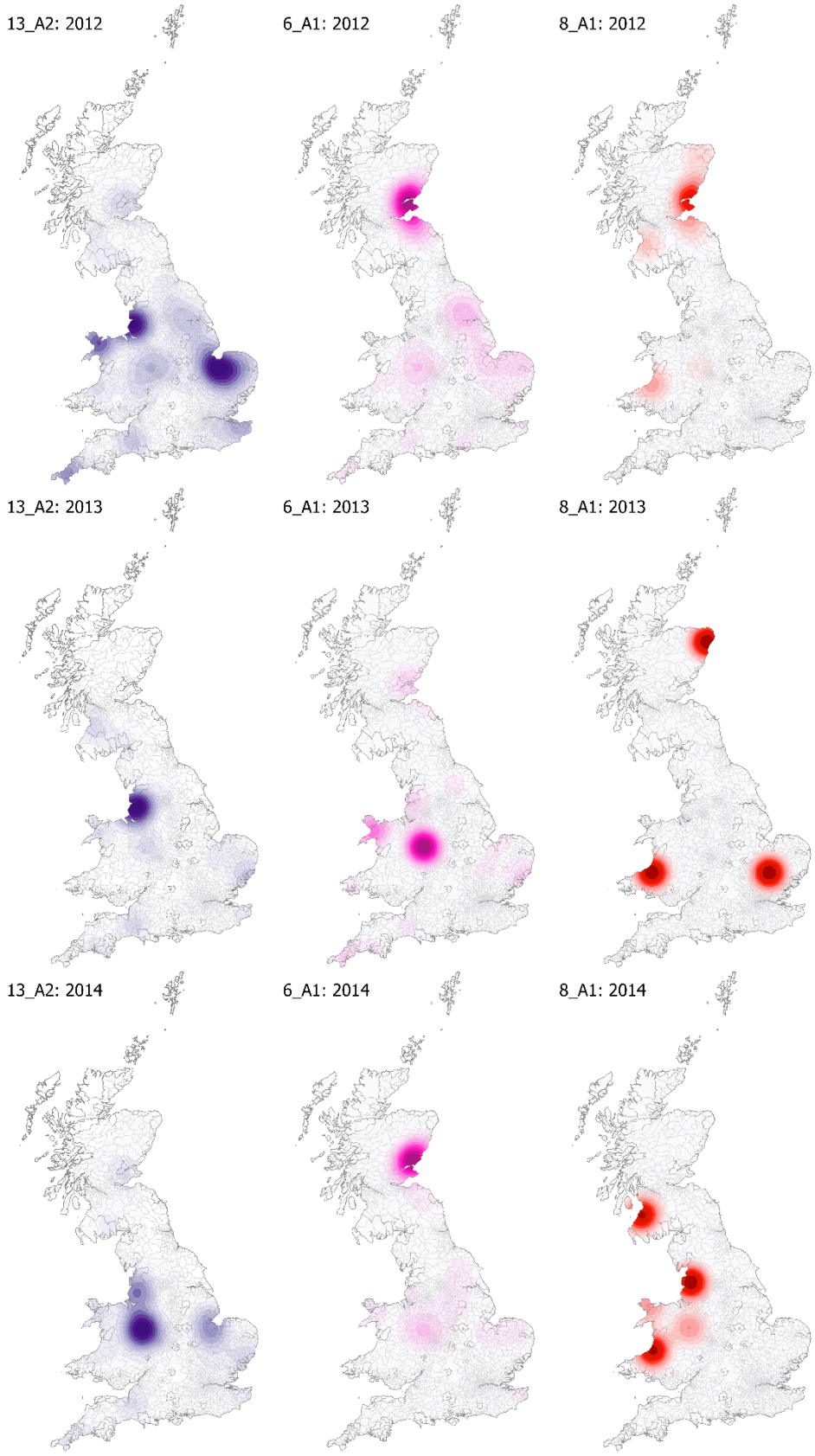


Figure 48. Kernel density distributions of late blight incidence showing inter-annual variation for genotypes 13_A2, 6_A1, and 8_A1, 2012–2014.

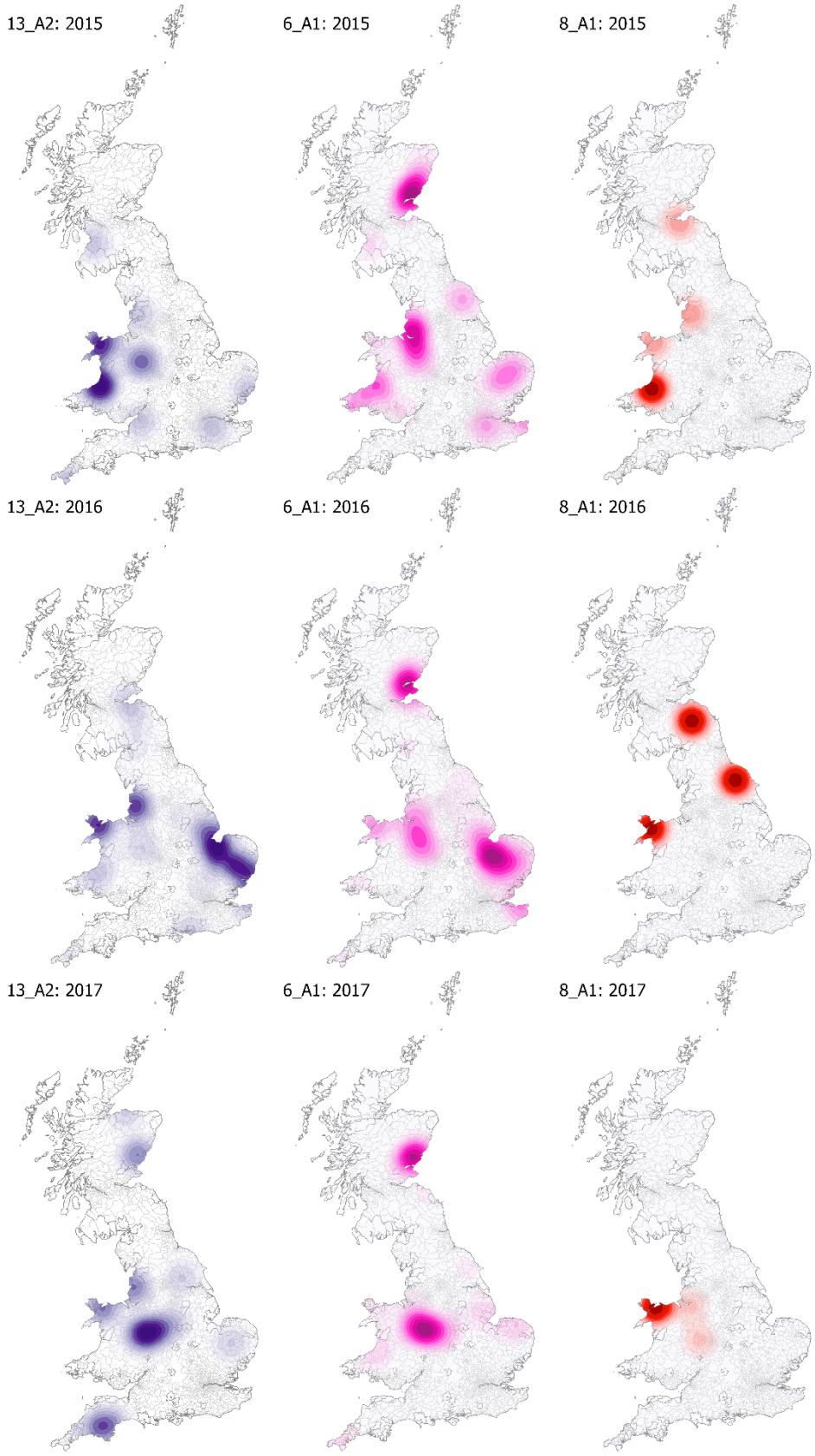


Figure 49. Kernel density distributions of late blight incidence showing inter-annual variation for genotypes 13_A2, 6_A1, and 8_A1, 2015–2017.

These analyses were useful in illustrating the large degree of variation in the distributions of genotypes within years, and within genotypes between years. A marked and rapid spread of the 13_A2 and 6_A1 genotypes was observed in the 2007 season from similar epicentres in eastern England 2006 (Fig. 38). Genotype 13_A2 established in Scotland early (Fig. 38) and was sampled at a high frequency there until after 2011 season when a late epidemic of 6_A1 dominated crops in Fife and Angus (Fig. 39). The epicentre of the genotype 6_A1 strain was slightly further north than 13_A2 which corresponds to findings of 6_A1 in Lincolnshire in 2004 and 2005 (data not shown). It spread north more slowly and was not established in Scotland until 2010. In the years since 2011, a displacement of 13_A2 by 6_A1 was apparent in Scotland (Figs. 48 and 49) but to a lesser extend in England and Wales.

In general, most areas of high intensity did not coincide for the three genotypes within years. Given this large variation both within and between years, and the relatively small numbers of outbreaks in some years, it was not possible to identify the principle environmental factors driving these population shifts.

Of the suite of 24 machine learning algorithms used for predicting the genotype of FAB outbreaks (using only cases of genotype 13_A2 or 6_A1), the highest training accuracy attained was 60.7% for a Fine Gaussian SVM (Fig. 50). Tuning the hyperparameters of this model to their optimal values only increased the accuracy by eight percent. The situation was slightly improved when attempting to predict the dominant genotype in each postcode district; a weighted KNN algorithm achieved a training accuracy of 69% (Fig. 51). Again, however, hyperparameter tuning did not improve model performance. We therefore selected the next best algorithm (Bagged Trees) and tuning increased the predictive accuracy by a few percent to 72.3%, with a testing accuracy of 66.9% and an area under the ROC curve of 0.72, which is considered a ‘fair’ result (Fig. 52).

1.1 ☆ Tree Last change: Fine Tree	Accuracy: 58.9% 5/5 features
1.2 ☆ Tree Last change: Medium Tree	Accuracy: 57.7% 5/5 features
1.3 ☆ Tree Last change: Coarse Tree	Accuracy: 56.8% 5/5 features
1.4 ☆ Linear Discriminant Last change: Linear Discriminant	Accuracy: 57.3% 5/5 features
1.5 ☆ Quadratic Discriminant Last change: Quadratic Discriminant	Accuracy: 54.3% 5/5 features
1.6 ☆ Naive Bayes Last change: Gaussian Naive Bayes	Accuracy: 54.0% 5/5 features
1.7 ☆ Naive Bayes Last change: Kernel Naive Bayes	Accuracy: 53.2% 5/5 features
1.8 ☆ SVM Last change: Linear SVM	Accuracy: 56.1% 5/5 features
1.9 ☆ SVM Last change: Quadratic SVM	Accuracy: 58.1% 5/5 features
1.10 ☆ SVM Last change: Cubic SVM	Accuracy: 45.8% 5/5 features
1.11 ☆ SVM Last change: Fine Gaussian SVM	Accuracy: **60.7%** 5/5 features
1.12 ☆ SVM Last change: Medium Gaussian SVM	Accuracy: 58.4% 5/5 features
1.13 ☆ SVM Last change: Coarse Gaussian SVM	Accuracy: 57.9% 5/5 features
1.14 ☆ KNN Last change: Fine KNN	Accuracy: 51.6% 5/5 features
1.15 ☆ KNN Last change: Medium KNN	Accuracy: 57.1% 5/5 features
1.16 ☆ KNN Last change: Coarse KNN	Accuracy: 55.7% 5/5 features
1.17 ☆ KNN Last change: Cosine KNN	Accuracy: 55.6% 5/5 features
1.18 ☆ KNN Last change: Cubic KNN	Accuracy: 57.2% 5/5 features
1.19 ☆ KNN Last change: Weighted KNN	Accuracy: 58.1% 5/5 features
1.20 ☆ Ensemble Last change: Boosted Trees	Accuracy: 59.1% 5/5 features
1.21 ☆ Ensemble Last change: Bagged Trees	Accuracy: 58.6% 5/5 features
1.22 ☆ Ensemble Last change: Subspace Discriminant	Accuracy: 57.1% 5/5 features
1.23 ☆ Ensemble Last change: Subspace KNN	Accuracy: 51.6% 5/5 features
1.24 ☆ Ensemble Last change: RUSBoosted Trees	Accuracy: 45.0% 5/5 features

Figure 50. Training accuracy of the suite of 24 classification algorithms used to predict the genotype of late blight outbreaks.

1.1 ☆ Tree Last change: Fine Tree	Accuracy: 62.8% 5/5 features
1.2 ☆ Tree Last change: Medium Tree	Accuracy: 59.4% 5/5 features
1.3 ☆ Tree Last change: Coarse Tree	Accuracy: 55.0% 5/5 features
1.4 ☆ Linear Discriminant Last change: Linear Discriminant	Accuracy: 62.5% 5/5 features
1.5 ☆ Quadratic Discriminant Last change: Quadratic Discriminant	Failed 5/5 features
1.6 ☆ Logistic Regression Last change: Logistic Regression	Accuracy: 61.8% 5/5 features
1.7 ☆ Naive Bayes Last change: Gaussian Naive Bayes	Accuracy: 58.7% 5/5 features
1.8 ☆ Naive Bayes Last change: Kernel Naive Bayes	Accuracy: 61.0% 5/5 features
1.9 ☆ SVM Last change: Linear SVM	Accuracy: 58.4% 5/5 features
1.10 ☆ SVM Last change: Quadratic SVM	Accuracy: 58.4% 5/5 features
1.11 ☆ SVM Last change: Cubic SVM	Accuracy: 59.2% 5/5 features
1.12 ☆ SVM Last change: Fine Gaussian SVM	Accuracy: 66.1% 5/5 features
1.13 ☆ SVM Last change: Medium Gaussian SVM	Accuracy: 59.2% 5/5 features
1.14 ☆ SVM Last change: Coarse Gaussian SVM	Accuracy: 57.6% 5/5 features
1.15 ☆ KNN Last change: Fine KNN	Accuracy: 66.1% 5/5 features
1.16 ☆ KNN Last change: Medium KNN	Accuracy: 55.3% 5/5 features
1.17 ☆ KNN Last change: Coarse KNN	Accuracy: 56.3% 5/5 features
1.18 ☆ KNN Last change: Cosine KNN	Accuracy: 59.2% 5/5 features
1.19 ☆ KNN Last change: Cubic KNN	Accuracy: 56.8% 5/5 features
1.20 ☆ KNN Last change: Weighted KNN	Accuracy: **69.0%** 5/5 features
1.21 ☆ Ensemble Last change: Boosted Trees	Accuracy: 66.4% 5/5 features
1.22 ☆ Ensemble Last change: Bagged Trees	Accuracy: 68.5% 5/5 features
1.23 ☆ Ensemble Last change: Subspace Discriminant	Accuracy: 61.8% 5/5 features
1.24 ☆ Ensemble Last change: Subspace KNN	Accuracy: 64.9% 5/5 features
1.25 ☆ Ensemble Last change: RUSBoosted Trees	Accuracy: 64.6% 5/5 features

Figure 51. Training accuracy of the suite of 24 classification algorithms used to predict the dominant late blight genotype in postcode districts.

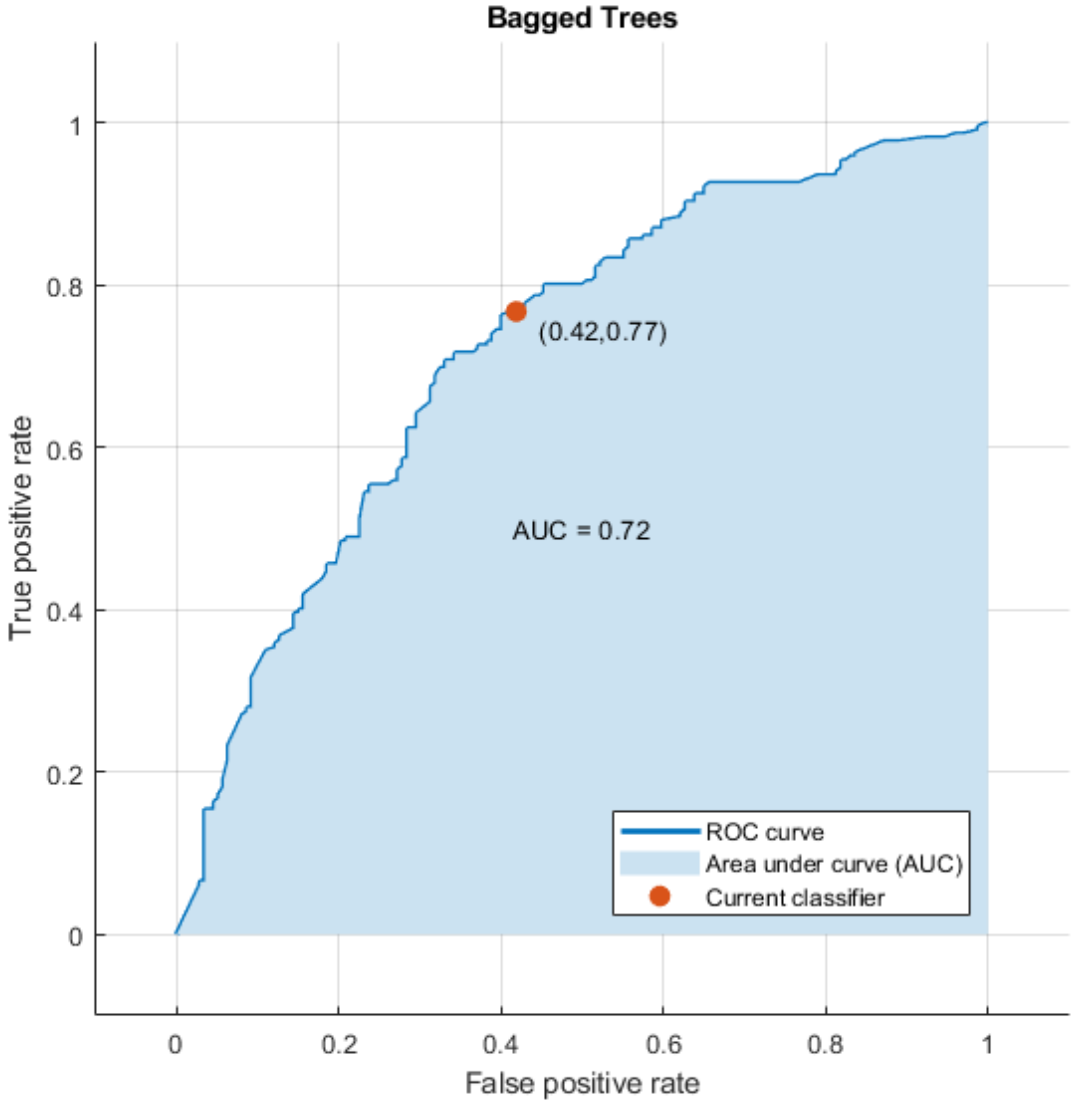


Figure 52. Receive Operating Characteristic curve of the model for predicting the dominant genotype in postcode districts.

Lack of predictive accuracy is not an unexpected result, as many studies have shown a large degree of phenotypic variation in environmental response, even within clonal lineages. It is possible that performance could be improved through the inclusion of other predictor variables. Nevertheless, it is of interest to note the importance of each weather variable in predicting the dominant genotype in each postcode district (Fig. 53). Figure 53 shows the importance of each predictor in the final model retained using all the data. The results are quite different to those obtained for early outbreaks (Fig. 9), with precipitation and humidity as the most important predictors. This provides tentative evidence that moisture is the principle driving factor for the competition between genotype 13_A2 and 6_A1. This finding is supported by the experimental analyses used to determine the new national warning system for late blight in the UK; the Hutton Criteria (Dancey et al. 2017). Controlled environment experiments with contemporary isolates of 13_A2 and 6_A1 showed significant levels of infection under drier conditions than previously observed.

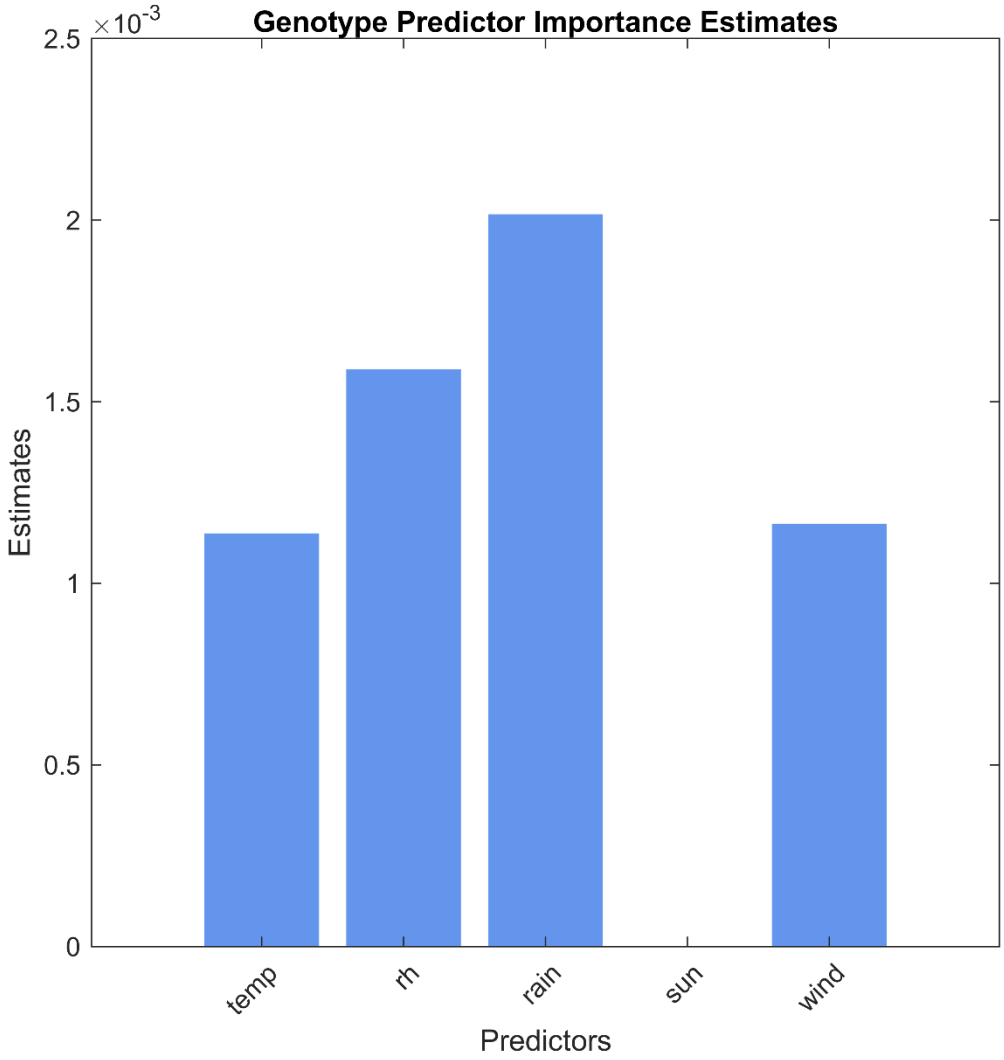


Figure 53. Importance of climate variables in predicting the dominant late blight genotype in postcode districts.

Discussion

In this project, a large and complex dataset of pathogen samples from late blight outbreaks sampled over 16 years from British potato crops was assembled and studied in relation to data on potato crop location and cropping density, hourly weather conditions and other factors such as soil and topography in order to evaluate the risks of early outbreaks and the rates of disease spread. Each of the pathogen samples had been genotyped and the patterns of genetic change in the pathogen population were also examined for evidence of the drivers of population change over time and in space.

The substantial dataset of sampled late blight outbreaks is an excellent resource, but a fundamental challenge in analysing data from crop disease surveillance programs as opposed to experimental data is observation bias. This bias can have several components, including spatial coverage bias (where not all fields are sampled) and detection bias (where some infected fields go undetected). This is often because the survey is focused on objectives other than a complete census of all outbreaks of disease. For example, in the case of the FAB potato late blight outbreak data used in this study, the principal objectives were to report on early outbreaks and sample the pathogen population in order to obtain information about population diversity, virulence, aggressiveness and fungicide sensitivity. The findings from this data affected management practices and was fed back to the potato industry. Sampling may thus be more intensive at the beginning of the growing season and decrease once blight is very active, or once a scout has already sampled the population in a specific geographic area. The FAB outbreak data could therefore be biased by 'imperfect' detection and it is possible that the resultant ArcGIS and modelling analyses is affected by patterns reflecting the difficulty or manner with which late blight was sampled rather than true patterns in occurrence and abundance. It can be seen, however, that the spatial distribution of reported outbreaks in the FAB data is reasonably uniform across the potato growing areas of GB, with some concentrated regions scattered along the eastern seaboard and throughout Wales and the South West (e.g., Fig. 14). These areas of higher incidence however, coincide with the areas where potato production is most intense, therefore the distribution of sampled outbreaks broadly matches the distribution of potato (Fig. 14) and particularly so when normalised to potato density (Fig. 15). The number of outbreaks reported in each country also mirrors the scale of potato production, which is greater in England, then Scotland and Wales, although incidence levels are relatively high for Scotland (Appendix 3). Pathogen sampling is also, dependant on suitable weather conditions for disease development and blight incidence varied from season to season and from one location to another. Such variation leads to localised disease epidemics which will have inevitably skewed the sampling intensity. Such marked changes in pathogen population size from one season to the next lead to 'founder effects' or 'genetic bottlenecks' which are known to influence the pathogen population structure over time (Goodwin, 1997). However, the long-term trend in outbreak reporting dates describes a typical epidemic curve for many crop diseases (Appendix 3), where the number of new outbreaks increases to a peak then declines with the proportion of uninfected crops and a shift to less favourable climatic conditions. Taken together, our evidence suggests that the quality and level of sampling activity is consistent and sufficient across most of GB and throughout the growing season and thus increases confidence in the outputs of this study. Nevertheless, there remains a potential for observation bias and the present findings should be interpreted with this caveat.

Occurrence of early outbreaks of late blight

Previous studies have shown that seasons in which late blight infection occurs earlier also result in greater infection pressure throughout the season with more late blight outbreaks sampled (Cooke, 2019). Because early infection increases both risk and the required expenditure on fungicides for blight management there is a clear advantage to the grower of prior knowledge of the regions prone to early infection. Analysis of all the outbreaks in this dataset indicated clearly that there is no single high-risk region from which early infection subsequently spreads (Figs 1-4). Disease occurred, on average, earlier in crops in the south of Britain than in the north but the main consistent hotspots of the first 10 and 20% of reported outbreaks across all seasons were in the early crops in southwest England, Wales and to a lesser extent in Scotland followed by the English Midlands and the Kent and Essex coast. Each season was different but, on average, the earliest 10% of outbreaks occurred in the last two weeks of May and the first week of June (Fig. 2). A more detailed study of the pattern of early disease hotspots over years showed fewer statistically significant hotspots overall and more hotspots that were defined as sporadic (Figs. 5 & 6). Modelling also showed that temperature was the best predictor of early outbreaks (Fig. 9). The onset and spread of late blight remains strongly dependent on the weather and a factor strongly influencing this analysis of early outbreaks was the variation in the disease pressure from season to season both at a national and a local level. The total number of sampled outbreaks ranged from 300 in the 2007 season to as few as 58 in 2015. The analysis applied is very sensitive at picking up different patterns of hotspots over time but relatively few passed these carefully defined thresholds. Colour-coded maps were produced to show the overall risk of early outbreaks by postcode district, and the week of the year these were most likely to occur (Appendix 1, Figs. A1 & A2). Using these we made general predictions that the earliest infections occur in regions with the highest density of potato crops and in warmer, low lying mainly coastal regions with earlier crops and more conducive blight weather. Climatic variation made it hard to predict more clearly or find other patterns in early outbreaks. In all potato growing regions, it is important that growers remain aware of primary infections at the start of the season and manage discard piles and volunteer potato plants (groundkeepers) and ensure high quality blight-free seed is used (Cooke et al., 2011). Such care will reduce early crop infection and aid management practices aimed at preventing rather than controlling disease.

Risk and rate of spread of late blight

An Optimized Outlier Analysis was used to examine all the sampled late blight outbreaks from 2003 to 2018 and identified clustering of disease incidence that provides an indication of future disease risk. The map generated from this analysis (Fig. 10 and Appendix 1, Fig. A3) indicated that the risk of spread of disease among neighbouring postcode districts was highest in the potato growing regions of Tayside, Fife, Lothian, East Anglia and parts of Kent. It was somewhat surprising that other potato growing regions such as the Midlands were not defined as high risk in this analysis, but this may relate to the year to year variation in local disease pressure.

The data available in this study allowed us to calculate the rate of spread of a new genotype across potato growing regions of GB. The velocity of spatial spread of newer genotypes 36_A2 and 37_A2 from early foci was calculated as between 3–17 km per week (Figs. 11 & 13). This highlights that even from a single point of infection the rate of crop to crop spread can have a severe impact on large areas of potato production within even one or two seasons.

The 37_A2 lineage was first sampled in the Midlands in late June 2016 and caused tuber blight outbreaks in the same region in addition to being sampled near Doncaster (Cooke,

2019). By the end of the 2017 season it had spread north to North Yorkshire as well as being sampled in eastern England and Kent (Figs. 11 & 12). Its spread appears to be consistent with crop to crop dispersal but new sources of infection via seed cannot be ruled out. This rapid spread had implications for the industry due to its insensitivity to fluazinam (Schepers *et al*., 2018) leading to weaknesses in some late-season spray programmes and resulting in an increased incidence of tuber blight problems in storage (Cooke, 2019). Genotype 36_A2 was first sampled in mid-April 2018 and despite it being a very low risk year it was also sampled north of The Wash later in the same season (Fig. 13). This lineage is known to be highly aggressive and its rapid spread is also causing management problems for growers.

Risk of late blight over time

An analysis of the accumulated total of sampled outbreaks showed some areas of above average late blight sampling and postcode districts with commercial potato production from which no outbreaks were sampled (Fig. 14). A comparison with the density of potato cultivation per district however, indicated the low blight sampling corresponded to low potato density and there were no major potato growing areas unsampled. When normalised according to potato cultivation density per postcode the distribution of sampling intensity was shown to be more uniform (Fig. 15). Not all potatoes are however, grown commercially and the FAB campaign has encouraged sampling from gardens, allotments and field trials as they provide another component of the pathogen population. Some of these regions appear as more intensively sampled patches with more samples for the relatively low potato cropping density (Fig. 15). Such samples were however not included in the hotspot data analysis and most maps thus indicate areas of commercial production in grey (e.g. Fig. 14) or show postcode districts with <1ha commercial potato that are masked out from the analysis (e.g. Fig. 16). A colour-coded map was produced to show the overall risk of late blight by postcode district (Appendix 1, Fig. A4).

A more detailed statistical analysis of hotspots in ArcGIS showed the most intense outbreak sampling from Kent, an area around The Wash in eastern England and Fife, Tayside and Angus in Scotland. Smaller spots were also recorded in the Midlands and Aberdeenshire (Fig. 16). No cold spots were recorded. These hot spot regions are areas with the greatest risk of blight over all the years and were also statistically significant hot spots in the space-time analysis (Fig. 17). There were three types of temporal trend identified in the hot spots of incidence: consecutive, sporadic, and new (appearing in the final year) with sporadic hot spot the most common. The prevalence of sporadic and lack of persistent hot spots reflects large inter-annual variation in the distribution of disease. It is perhaps striking that most potato growing districts are not statistically significant hot spots of blight sampling but, as the incidence data shows, this does not indicate an absence of blight but a reduced average risk.

A neural network model was developed to explore the relationship between outbreak risk and a series of detailed environmental factors. The overall accuracy of the model was high with the best compromise between true positive and false positive rate being of 80 and 15% respectively (Fig. 18). A sensitivity analysis to determine the impact of each variable on the model's ability to predict risk revealed some factors with a clear association but also showed the challenges of this type of analysis. As expected, weather had a strong impact on outbreak risk, particularly temperature, humidity, rainfall and windspeed (Fig. 20). Despite there being no known association between soil class or geological type and risk of late blight in the crop growing on such a substrate there was also evidence of an association with soil conditions and geological type. This analysis picked up above average pH Calcaric and Mollic soils as having a lower blight risk (Fig. 23). A difficulty faced in this analysis is that outbreaks were defined at postcode district level. Such districts vary in size and clearly comprise potato

growing land with a range of characteristics (soil, slope, aspect etc). Because of this, topographic factors were averaged across postcodes, instead of individual site conditions being applied to observations in the model. Despite this an association between topography (elevation, slope, aspect) and outbreak risk was also observed.

At the extremes there are clear differences with flat peaty soils on land at or below sea level in the Fens compared to land of greater elevation and mixed podzol and brown soils in parts of Aberdeenshire. But within regions or postcode districts it is unclear how representative an average slope or aspect measure is (the resolution is insufficient).

It is challenging to relate the mapped risks with the hot spot analysis. For example, the soil class risk map indicated lower risk in the Fens compared to a higher risk for most of the eastern Scottish growing area, yet the outbreak risk and hot spot analysis showed both areas with a high risk. It may be that the proportional contribution of soil type to blight risk is low and thus over-estimated compared to other factors.

All the modelling and sensitivity analysis performed with the neural network model was carried out using the full outbreak data. The original intention was to repeat this work using data from different genotypes and for early outbreaks. However, the number of data points available made it impossible to produce statistically robust models using this approach. For neural network modelling, the number of data points must be greater than the number of input variables, or the model becomes overfitted to the data and effectively meaningless for 'real' examples. Using fewer model inputs for individual genotypes would have produced significantly less accurate models. Future work to explore a resolution to this is recommended.

Genetic makeup and spatial distribution of late blight pathogen genotypes

Over the course of this study, the populations of P. infestans causing potato late blight on British crops have undergone major changes. New pathogen genotypes may have distinct advantageous traits enabling them to spread preferentially and displace others. This can happen very quickly under optimal conditions as the rate of inoculum production and its ability to spread can develop severe local to regional late blight epidemics. Examples of new traits may be an ability to overcome host resistance (Young et al., 2009, Montarry et al., 2008, Stellingwerf et al., 2018), a different temperature response (Cooke et al., 2012; Mizubuti & Fry 1998), increased aggressiveness (Young et al., 2018) or fungicide resistance (Schepers et al., 2018). Three clones, 8_A1, 13_A2 and 6_A1 have predominated with overall similar mean central tendency locations but with a slight skew to the south for 13_A2 and to the north in the case of 8_A1 (Fig. 30). Clone 8_A1 has been present in Europe since at least 1995 when it was reported widely in the UK and the Republic of Ireland (Day et al., 2004). It has been largely displaced by other clones but nonetheless persisted at a low frequency with hotspots in parts of Scotland and Wales (Fig. 38). Genotype 13_A2 had a significant impact on potato late blight management, particularly when it emerged in 2006-8, as it proved aggressive, insensitive to metalaxyl (Cooke et al., 2012) and overcame established sources of blight resistance (Lees et al., 2012). This clone spread rapidly from a serious outbreak on the Essex coast in 2006 to dominate crops across the UK in the following three seasons. Its spread north was also very rapid, found in 80% of Scottish outbreaks by 2007. Despite this initial wide distribution, this long-term analysis indicated that major hotspots of 13_A2 (Fig. 36) covered most of East Anglia, Lancashire and Wales with northern England and Scotland as cold spots. These cold spots reflect a later distinct transition to a population dominated by 6_A1 in Scotland in 2011 after which the frequency of 13_A2 in Scottish samples remained markedly lower than those from crops in England and Wales low (Appendix 2). This change in genotype explains why the emerging hotspot analysis indicated only sporadic hotspots in the Midlands and Shropshire

and persistent hotspots of 13_A2 in Lancashire and East Anglia (Fig. 41). In contrast, the long-term hotspot analysis of 6_A1 indicated a dominance from the north Midlands to northern England and most of the potato growing area of Scotland (Fig. 37). It has proved persistent with 40-80% of samples being of genotype 6_A1 every year since 2011 across much of Britain (Appendix 2, Figs. 38-41). The emerging hotspot analysis confirms its long-lasting presence in several regions of Britain as persistent or sporadic hotspots (Fig. 42). Sporadic hotspots are, in general, a sign of the variation in blight conducive conditions with 2010, 2013, 2015 and 2018 being drier warmer years with fewer outbreaks to sample. Genotype 6_A1 has no known insensitivity to any fungicide active ingredient but its persistence suggests an ability to survive well overwinter and that it is fit and aggressive in field epidemics (Cooke et al., 2012).

This analysis presents clear evidence of the recent local emergence and spread of two new threats. As discussed above, genotype 37_A2 emerged first in 2016 and spread rapidly to form a significant and spreading hotspot centred on parts of Shropshire (Fig. 40) and to a lesser extent in Kent. These are shown as persistent hot spots in the EHSA (Fig. 45) and probably reflect two distinct introductions. This lineage has subsequently spread to Scotland and Northern Ireland (www.euroblight.net). After widespread reporting of the fluazinam insensitivity of 37_A2 the number of hectares of potato in the UK treated with this product has fallen by over 80% (Garthwaite et al 2019). This drop has reduced the selection pressure on the 37_A2 population and the rate of spread and incidence of 37_A2 has subsequently declined to less than 10% of the sampled population in England and has prevented it fully establishing in Scotland (Appendix 2). This is a good example where population monitoring and risk assessment has generated timely new guidance on fluazinam use. Fluazinam producers and the Fungicide Resistance Action Group – UK advice has changed grower behaviour and is protecting this valuable active ingredient for future use. In parts of mainland Europe that are reported to be still relying on fluazinam more heavily, the proportion of 37_A2 remains close to 25% (www.euroblight.net). Although being first sampled one year later than 37_A2, genotype 36_A2 has formed hot spots in Kent and parts of East Anglia (Fig. 39 and 44)) and has spread further in 2019 (Appendix 2). With no reported fungicide insensitivity issues its ability to displace the existing genotypes appears to be related to aggressiveness as it formed larger lesions at low doses of all tested active ingredients to date (Lees et al., 2018).

Machine learning algorithms were used to model predictions of the dominant genotype in each postcode district. The model provided some evidence that precipitation and humidity are the most important predictors, suggesting moisture plays an important role in competition among genotypes (Fig. 53). This analysis was, however, challenging due to fluctuations in the number of outbreaks sampled per year due to between year climatic variation.

Finally, a kernel density analysis was used to display the smoothed 13_A2, 6_A1 and 8_A1 genotype distributions each year from 2006–2017 (Figs. 46-49). This supported the hotspot analysis and indicated a clear early spread of the 13_A2 and 6_A1 types over the 2006 and 2007 seasons followed by a patchy distribution in subsequent years. In 2011 a severe epidemic of predominantly 6_A1 late in the season in Scotland resulted in a skewed distribution in subsequent years. Genotype 8_A1 was established almost 10 years before 13_A2 and 6_A1 and its patchy broad distribution reflects this. A video was also produced to show the changing pattern of genotype distributions each year and is available online with this report or from the authors.

Future R&D

Further research is required to identify the principle drivers of change in the spatial distribution of genotypes. This will involve understanding any effects of the environmental variables identified by the modelling work as having an impact on late blight incidence. There is also a need to examine the role of primary inoculum on genotype distributions, i.e. the strength of association between genotype across consecutive growing seasons. There were insufficient data to perform space-time pattern mining or modelling of individual genotypes at a suitably fine temporal resolution. There is a need, therefore, to increase the number of outbreaks sampled each year under the AHDB Potatoes FAB campaign. This is important if we are to develop new tools to predict changes in the distribution of aggressive lineages in order to adapt short-term control strategies in a more timely fashion.

Appendix 1 - Visual aids

We have derived a series of visual aids from this work, and these are presented below arranged by project deliverable.

1. When and where do early outbreaks of late blight occur in different parts of GB?

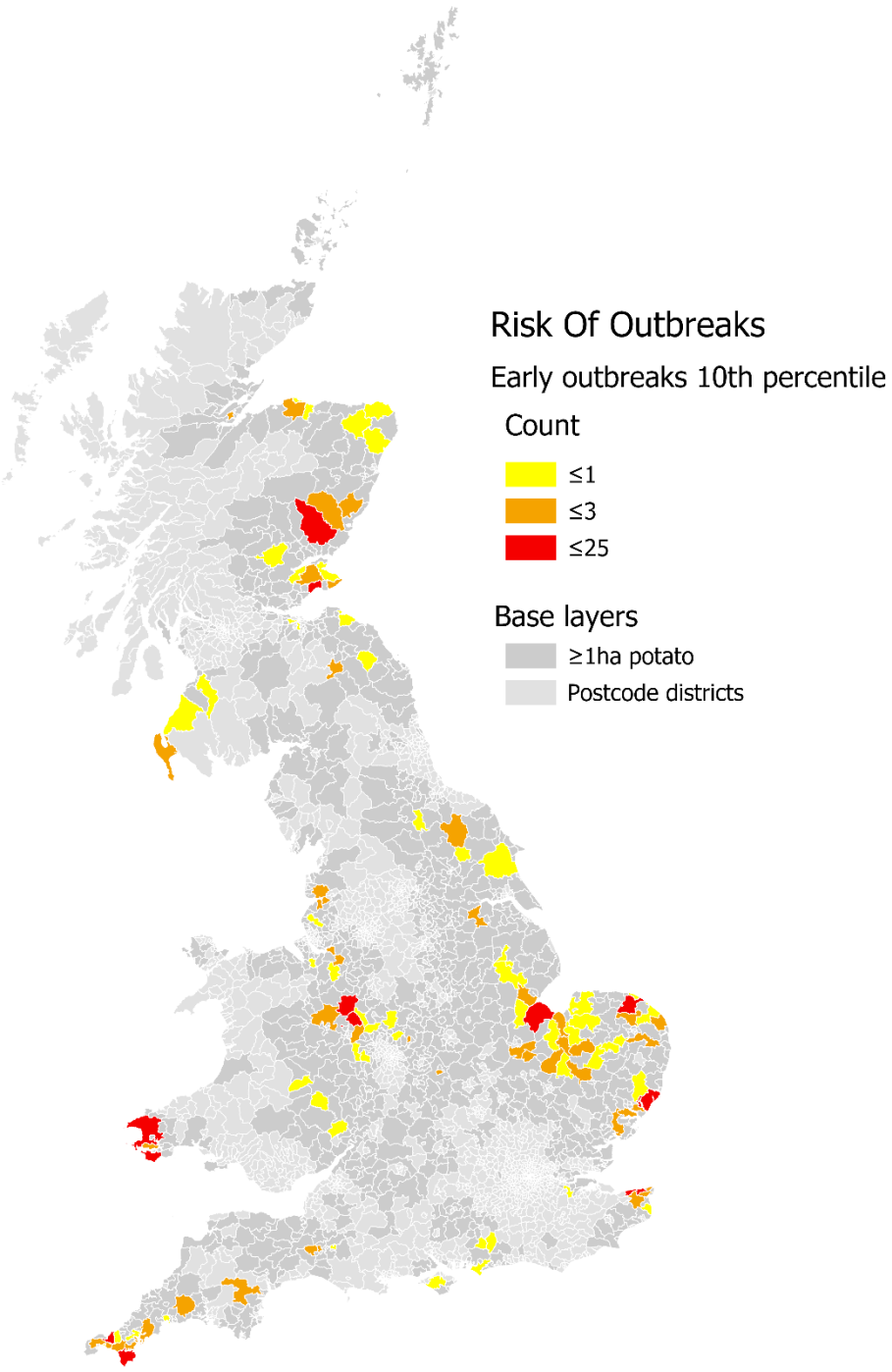


Figure A1. Choropleth map showing a count of the 10th percentile by date of outbreaks within postcode districts containing >1ha potato (grown commercially), 2003-2018.

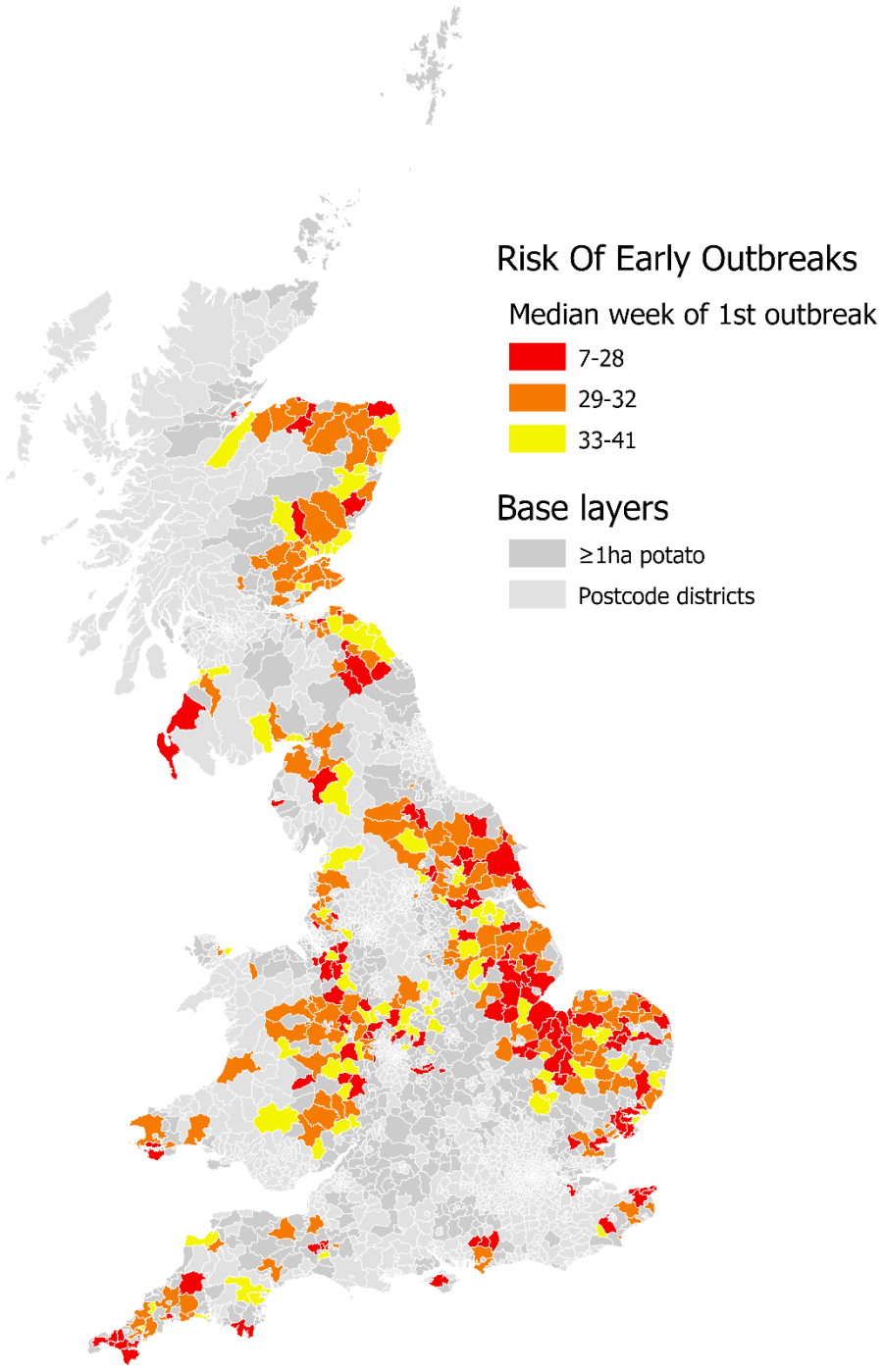


Figure A2. Choropleth map showing the median week of the year for the first late blight outbreaks within postcode districts containing >1ha potato (grown commercially), 2003–2018.

2. What is the risk of spatial spread of late blight in different parts of GB?

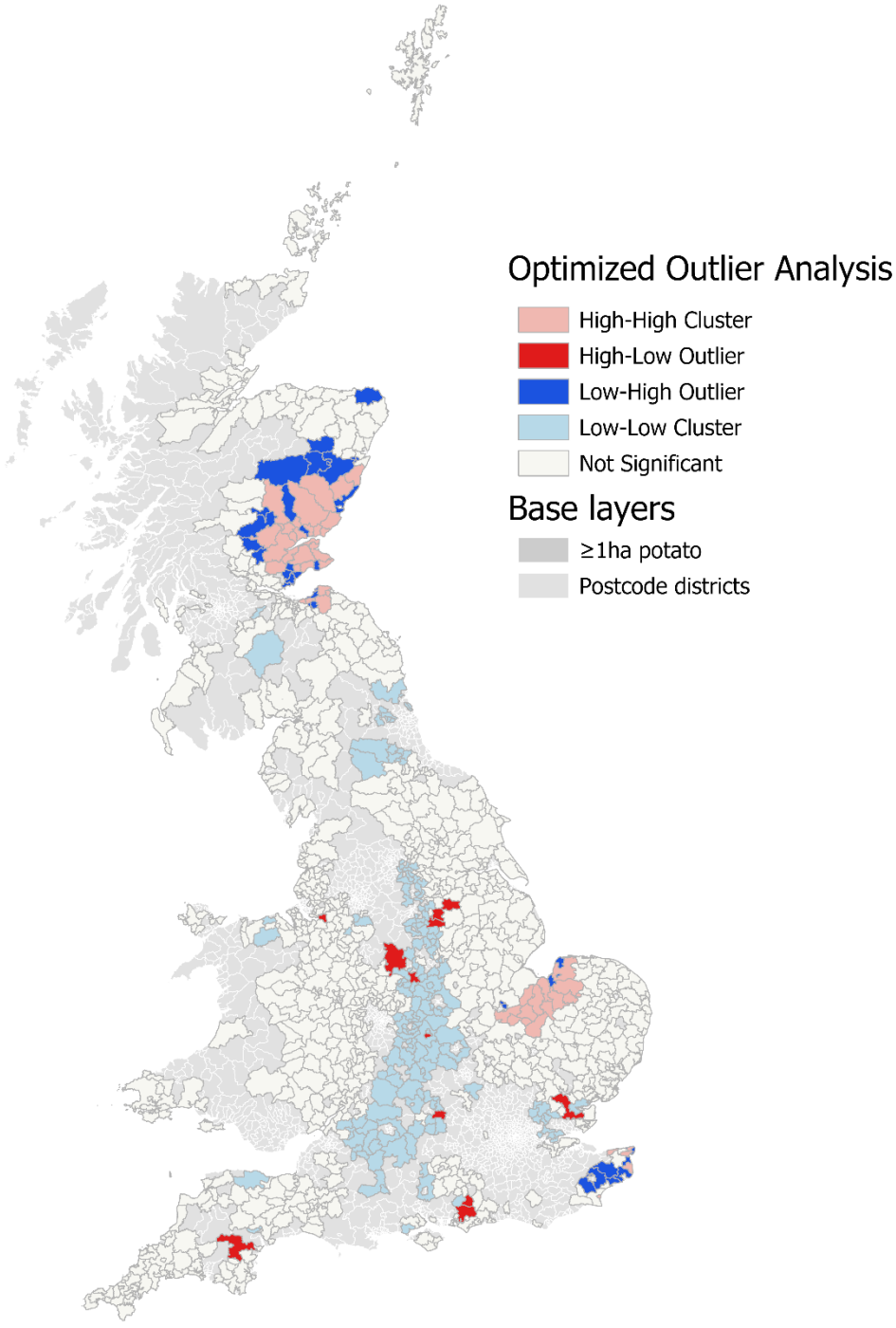


Figure A3. Risk of spread of late blight among postcode districts. Crops in High-High clusters or Low-High outliers are at risk of spread of disease from neighbouring (High) sectors.

3. What is the historical risk of late blight across different parts of GB?

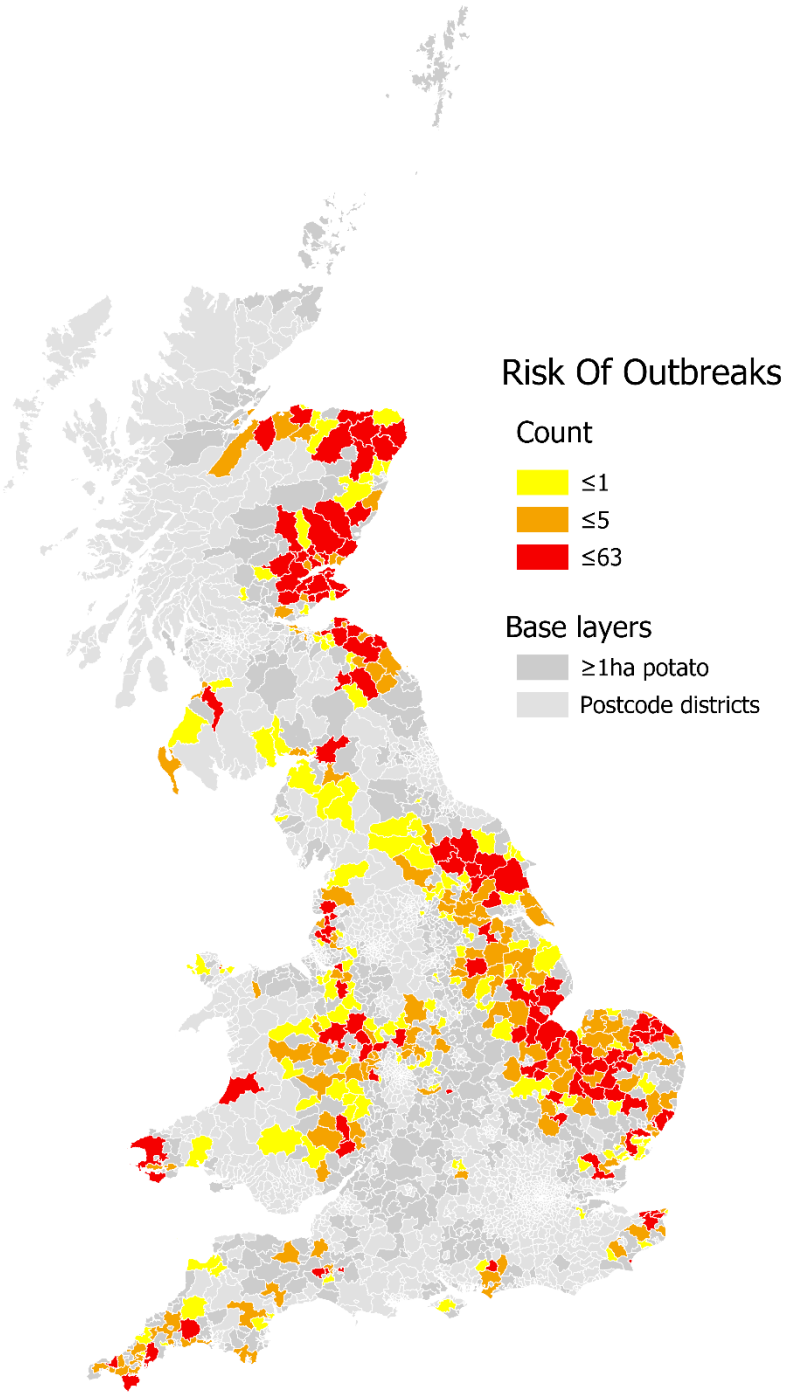


Figure A4. Choropleth map showing a count of all outbreaks within postcode districts containing >1ha potato (grown commercially), 2003-2018.

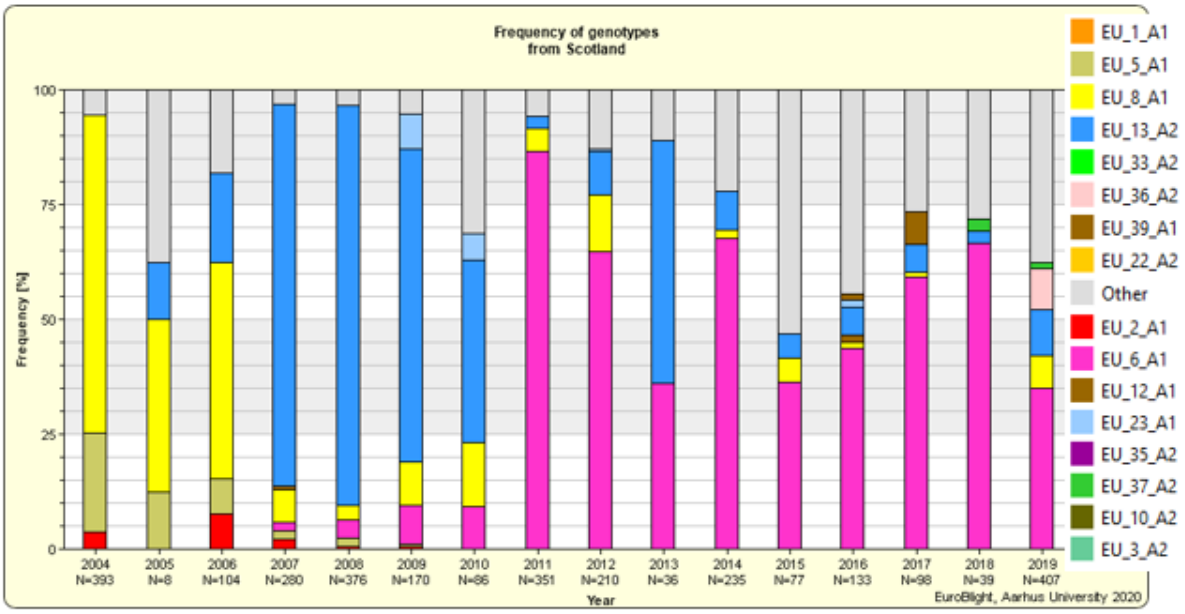
4. How has the genetic makeup and spatial distribution of the late blight pathogen changed over time, and what is driving this change?

In our analysis, we used Emerging Hot spot Analysis and KDE to explore this question. The figures given above show clear changes of genotypes in time and space but do not point to clear and obvious drivers based on the traits in the database. Climatic fluctuations from season to season impact the blight pressure and sampling intensity. The reduced population size and fewer samples over the 4 drier low blight pressure coupled with local high blight pressure in other seasons complicates such long-term analysis. It is easier to identify long-term presence of genotypes than it is to identify changes in distribution over the last 15 years. Eastern Scotland, Wales, the Midlands and Anglia all stand out as persistent areas in this. A visual analysis comparing the EHSA between 13_A2 and 6_A1 does show some spatial variation, but not enough to draw conclusions from. The Bagged Tree model developed to predict the dominant genotype in each postcode district did provide evidence that moisture plays a key role in driving pathogen population change, but the model was not accurate enough to draw any definitive conclusions. A video has been produced from the KDE maps (Figs. 38-41) that animates the changing spatial distribution of pathogen genotypes over time.

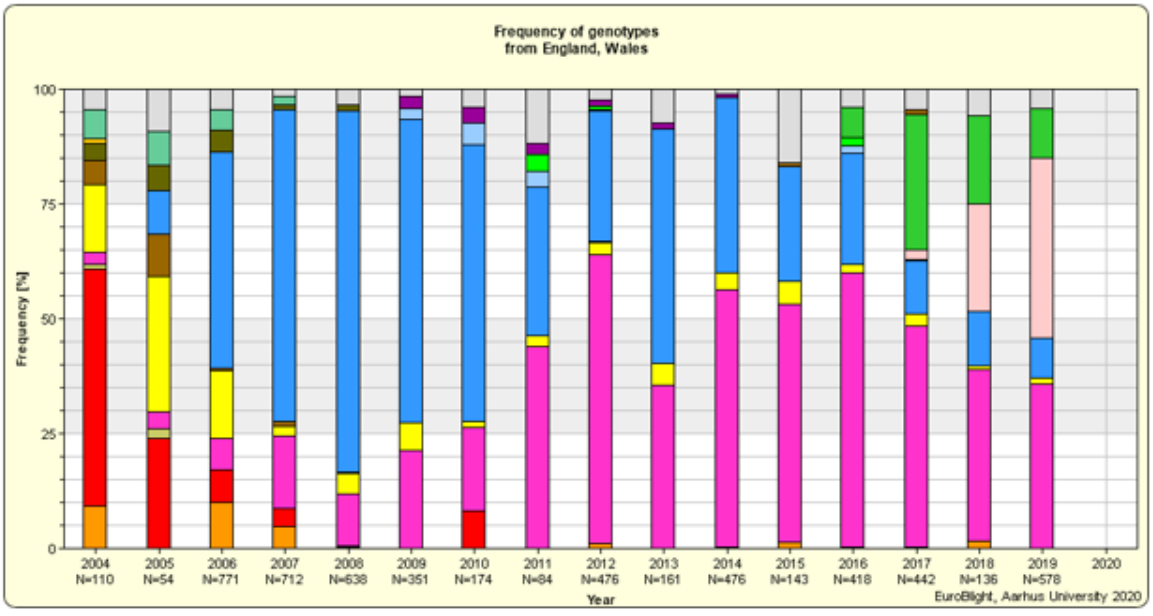
Appendix 2: *P. infestans* genotypes (2004-2019)

Proportion of different clonal genotypes of *P. infestans* collected from sampled FAB late blight outbreaks in a) Scotland and b) England and Wales from 2004-2019. Images generated from AHDB data submitted to EuroBlight database and reported via visualisation tool https://agro.au.dk/forskning/internationale-platforme/euroblight/

a)

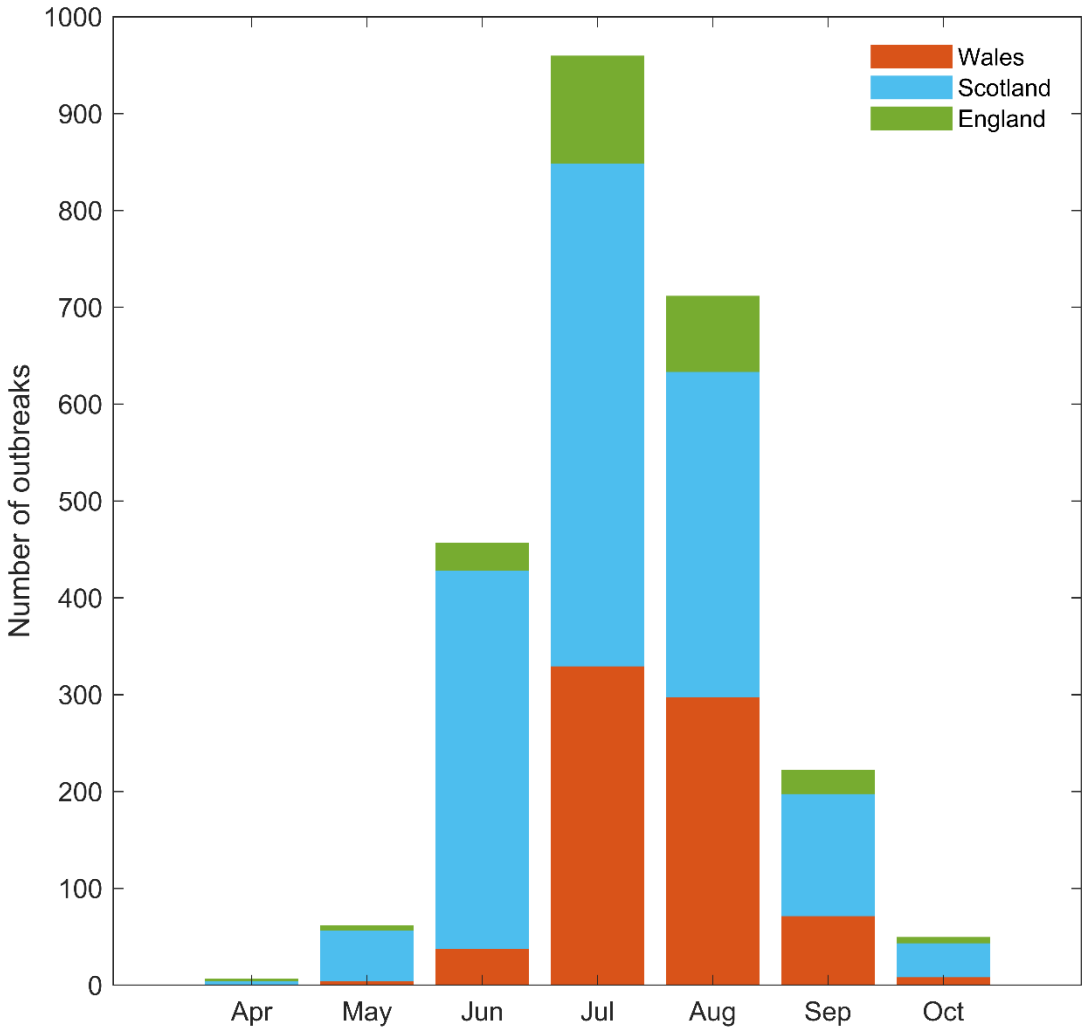


b)



Appendix 3: Blight outbreaks

Cumulative monthly totals of blight outbreaks sampled across England, Scotland and Wales of the period 2003-2018.



References

Aitkenhead, M.J., Coull, M.C., 2019. Mapping soil profile depth, bulk density and carbon stock in Scotland using remote sensing and spatial covariates. European Journal of Soil Science. 10.1111/ejss.12916.

Aitkenhead, M.J., Coull, M.C., 2016. Mapping soil carbon stocks across Scotland using a neural network model. Geoderma 262, 187-198. http://dx.doi.org/10.1016/j.geoderma.2015.08.034.

Aitkenhead, M.J., Rhind, S.M., Zhang, Z.L., Kyle, C.E., Coull, M., 2013. Neural network integration of field observations for soil endocrine disruptor characterisation. Science of the Total Environment 468-469, 240-248. 10.1016/j.scitotenv.2013.08.007.

Cooke DEL, Lees AK, Shaw DS, Bain RA, Ritchie F, Taylor MC, 2009. Survey of GB Blight Populations. In. *Final report of Potato Council project R274.* AHDB Potatoes web site http://potatoes.ahdb.org.uk/sites/default/files/publication_upload/20106%20Final%20Report%20Blight%20Populations%20R274.pdf.

Cooke DEL, Lees AK, Chapman AC, Cooke LR, Bain RA, 2013. GB Late Blight Populations: monitoring and implications of population changes. In. *Potato Council project R423 final report.* AHDB Potatoes web site http://potatoes.ahdb.org.uk/sites/default/files/publication_upload/Final%20Report_R423.pdf.

Cooke DEL, 2016. GB late blight population monitoring 2014 & 2015. In. *AHDB Potatoes Project report* *http://potatoes.ahdb.org.uk/sites/default/files/publication_upload/FAB%202014%20and%202015%20seasons.pdf.*

Cooke DEL. 2019. GB Late Blight Population Monitoring 2014 to 2018. Report for AHDB Potatoes project E000529. https://potatoes.ahdb.org.uk/sites/default/files/publication_upload/FAB_Final%20Report_2014to2018.pdf

Cooke DEL, Cano LM, Raffaele S, Bain RA, Cooke LR, Etherington GJ, Deahl KL, Farrer RA, Gilroy EM, Goss EM, Grunwald NJ, Hein I, Maclean D, Mcnicol JW, Randall E, Oliva RF, Pel MA, Shaw DS, Squires JN, Taylor MC, Vleeshouwers VG, Birch PR, Lees AK, Kamoun S, 2012. Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. PLoS pathogens 8, e1002940.

Cooke LR, Schepers HTAM, Hermansen A, Bain RA, Bradshaw NJ, Ritchie F, Shaw DS, Evenhuis A, Kessel GJT, Wander JGN, Andersson B, Hansen JG, Hannukkala A, Nærstad R, Nielsen BJ, 2011. Epidemiology and Integrated Control of Potato Late Blight in Europe. Potato Research 54, 183-222.

Dancey S, Cooke DEL, Skelsey P. 2017. Decision Support Systems in Great Britain; The Hutton Criteria. Proceedings of the Proceedings of the 16th EuroBlight Workshop, 2017. Aarhus, Denmark, 53-8.

Dancey S, 2018. Development and implementation of a new national warning system for potato late blight in Great Britain. PhD Thesis. University of Dundee.

Day JP, Wattier RAM, Shaw DS, Shattock RC, 2004. Phenotypic and genotypic diversity in *Phytophthora infestans* on potato in Great Britain, 1995–98. *Plant Pathology* **53**, 303-15.

Garthwaite D, Ridley L, Mace A, Parrish G, Barker I, Rainford J, Macarthur R, 2019. Arable crops in the United Kingdom 2018 In. Pesticide Usage Survey Report 284. Fera, York.

Goodwin SB, 1997. The population genetics of *Phytophthora*. Phytopathology 87, 462-73.

Hollis, D., McCarthy, M., Kendon, M., Legg, T., Simpson, I. 2019. HadUK- Grid—A new UK dataset of gridded climate observations. Geoscience Data Journal, 6(2): 151-159. https://doi.org/10.1002/gdj3.78

Kroner A, Mabon R, Corbiere R, Montarry J, Andrivon D, 2017. The coexistence of generalist and specialist clonal lineages in natural populations of the Irish Famine pathogen Phytophthora infestans explains local adaptation to potato and tomato. Mol Ecol 26, 1891-901.

Lees AK, Stewart JA, Lynott JS, Carnegie SF, Campbell H, Roberts AMI, 2012. The effect of a dominant *Phytophthora infestans* genotype (13_A2) in Great Britain on host resistance to foliar late blight in commercial potato cultivars. *Potato Research* **55**, 125-34.

Lees AK, 2018. Comparison of sensitivity to a range of fungicides in contemporary genotypes of Phytophthora infestans. AHDB Potatoes project report 2018/7. https://potatoes.ahdb.org.uk/sites/default/files/publication_upload/11120047%20Fungicide%20Sensitivity%20Testing%202018%20Report_Final.pdf

Mizubuti ES, Fry WE, 1998. Temperature effects on developmental stages of isolates from three clonal lineages of Phytophthora infestans. Phytopathology 88, 837-43.

Montarry J, Glais I, Corbiere R, Andrivon D, 2008. Adaptation to the most abundant host genotype in an agricultural plant-pathogen system--potato late blight. J Evol Biol 21, 1397-407.

Pettitt TR, Keane GJ, John SOL, Cooke DEL, Žerjav M, 2019. Atypical late blight symptoms following first recorded infections by Phytophthora infestans genotype EU_39_A1 in UK vine tomatoes. New Disease Reports 39, 16.

Schepers HTaM, Kessel GJT, Lucca F, Forch MG, Van Den Bosch GBM, Topper CG, Evenhuis A, 2018. Reduced efficacy of fluazinam against Phytophthora infestans in the Netherlands. European Journal of Plant Pathology 151, 947-60.

Stellingwerf JS, Phelan S, Doohan FM, Ortiz V, Griffin D, Bourke A, Hutten RCB, Cooke DEL, Kildea S, Mullins E, 2018. Evidence for selection pressure from resistant potato genotypes but not from fungicide application within a clonal Phytophthora infestans population. Plant Pathology 67, 1528-38.

Young GK, Cooke LR, Kirk WW, Tumbalam P, Perez FM, Deahl KL, 2009. Influence of competition and host plant resistance on selection in *Phytophthora infestans* populations in Michigan, USA and in Northern Ireland. Plant Pathology 58, 703-14.

Young GK, Cooke LR, Watson S, Kirk WW, Perez FM, Deahl KL, 2018. The role of aggressiveness and competition in the selection of *Phytophthora infestans* populations. Plant Pathology 67, 1539-51.