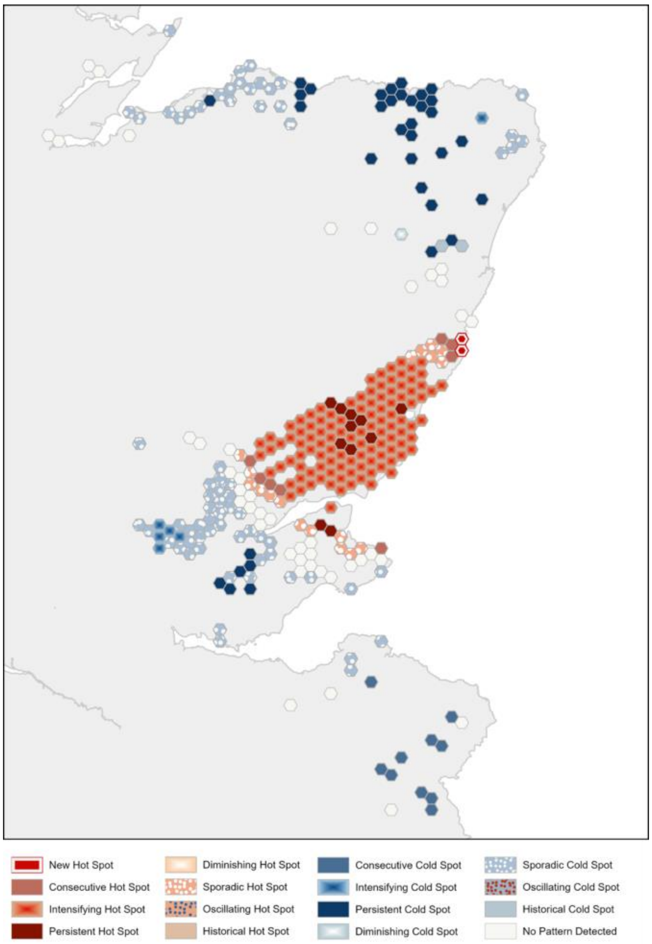
Modelling the spread of PCN in Scotland

Project Final Report



www.planthealthcentre.scot

This work was commissioned by Scotland's Centre of Expertise for Plant Health Funded by Scottish Government through the Rural & Environment Science and Analytical Services (RESAS) Division under grant agreement No PHC2020/08

Authors: Peter Skelsey[1]
[1] James Hutton Institute, Invergowrie, Dundee, DD2 5DA

Please cite this report as follows: P. Skelsey (2022). Modelling the spread of PCN in Scotland: Project Final Report. PHC2020/08. Scotland's Centre of Expertise for Plant Health (PHC). DOI: 10.5281/zenodo.7805729

Available online at: planthealthcentre.scot/publications

Dissemination status: Unrestricted

Content

1 Executive Summary

1.1 *Background*

Potato cyst nematodes (PCN) cost UK agriculture over $50 million/year and threaten food security in the developed and developing world. Improving our understanding of PCN epidemiology is a priority for the potato industry, with spatial and spatiotemporal modelling identified by the recent PHC PCN Working Group as essential components.

1.2 *Main objectives*

The aim of this project was to gain a better understanding of the spatial epidemiology of PCN in Scotland. The primary hypothesis was that PCN incidence is associated with a range of soil, crop, pest, agronomic, landscape, and climatic factors. The project sought to simultaneously analyse these factors over a range of spatial and temporal scales, by applying cutting-edge mapping, statistical, and artificial intelligence (machine learning) techniques to existing landscape-scale datasets, in order to: (i) fully characterise space and spacetime patterns of PCN incidence in Scotland, (ii) identify the principal drivers of PCN incidence in Scotland, (iii) produce a machine learning algorithm to predict future patterns of incidence, and (iv) perform a scenario analysis with the model to investigate the impact of various management scenarios. This represents a completely new avenue and scale of epidemiological exploration in the PCN literature.

1.3 *Key findings*

- Incidence of *G. pallida* and mixed populations are increasing over time, whereas incidence of *G. rostochiensis* is decreasing.
- There are statistically significant hotspots of *G. pallida* and mixed population incidence in the Dundee postcode area
- *G. rostochiensis* has a more widespread distribution than *G. pallida*, with statistically significant hotspots of incidence extending from the Dundee postcode area, up into Aberdeenshire and down into the Kirkcaldy postcode area.
- Changes in the spatiotemporal incidence of one species were partially mirrored by opposing changes in the other.
- Incidence of *G. pallida* is intensifying over space and time in in the north Kirkcaldy, east Perth, Dundee, and south Aberdeen postcode areas, with statistically significant hotspots of *G. pallida* and mixed population incidence in the Dundee postcode area.
- Incidence of *G. rostochiensis* has been persistently high over space and time in those areas, except for a large swathe of land across the Dundee postcode area where it is diminishing.
- The amount and proximity of surrounding seed and ware potato crops were important drivers of incidence, as were features specific to the infested site, such as the slope and size of the field and the number of years since it was last planted with potato.
- The principal drivers of PCN incidence can potentially be modified by growers or at a strategic national level to reduce risk of infestation.
- A machine learning algorithm was developed to predict PCN incidence to a high level of accuracy.

1.4 Recommendations and future work

- *National strategy*: The results of this study provide evidence that cross-contamination between infested fields and healthy potato crops is causing persistent and intensifying hotspots of incidence in densely cropped areas. There is currently little incentive or accountability for the ware sector to keep land PCN-free, therefore a national PCN strategy that addresses this issue, either through financial incentives or new legislation, could be key to preventing the ongoing loss of land for seed production.
- *Targeted control*: There is a potential role for targeted control methods within a national PCN strategy. The maps produced by this study provide powerful visual aids that can be used to tailor control options, restrictions, and financial incentives according to patterns of PCN incidence. Restrictions on the use of infested land for growing ware and on the disposal of crops, waste and soil could be tightened in hotspots of PCN incidence, and potentially eased or lifted in coldspots of incidence. Similarly, the standard sampling rates for official pre-crop soil testing for PCN could potentially be increased in areas where risk of pest spread is high and lowered elsewhere. In addition, financial incentives could be tailored to the species-level to encourage more widespread cultivation of varieties with moderate or high levels of resistance to *G. pallida* in regions where incidence of *G. pallida* is intensifying.
- *Mapping*: It is recommended that the ArcGIS analyses are repeated each growing season to provide an up-to-date overview of the evolving PCN situation in Scotland.
- *Recording*: Incorporation of additional agronomic variables in SPUDS (e.g., cropping history, soil type and management, irrigation practices) would serve to improve future modelling studies.
- *Modelling*: Improving our ability to forecast when and where PCN will spread in the future is one of the key recommendations of the Report of the Scottish PCN Working Group (2020). This research provides the foundation to meet that aim, with further work planned under the Scottish Government funded Strategic Research Programme 2022-2027 and through external funding.

2 Introduction

Potato cyst nematodes (PCN) *Globodera pallida* and *G. rostochiensis* are major pests of potato and other members of the Solanaceae family. They are among the most successful and highly specialized plant parasitic nematodes and among the most regulated quarantine pests globally. In Scotland, PCN has infested over 13% of land for growing potatoes and bulbs and is doubling every 7-8 years; an increase in spread of 5% per year. Current control of PCN in Scotland relies heavily on the Scottish Government programme of statutory PCN testing, which places restrictions on further production in land recorded as infested. This has a major impact on farm businesses and in particular the potato seed and flower bulb sectors, which require certified PCN-free land to grow their crop. The number of commercially grown varieties with dual resistance to both species of PCN that are suitable for the Scottish climate is limited, and although nematicides are routinely used against PCN, they are at high risk of withdrawal. As a result of the escalating PCN crisis, the Report of the Scottish PCN Working Group (Toth *et al.* 2020) recommends there is a need to "Determine the factors responsible for the spread of PCN using existing data sets and use outcomes to help inform the use of DSS and the development of a national strategy for PCN management."

The ability to forecast PCN outbreaks and spread first requires an understanding of the various factors that influence PCN population dynamics. The first step in understanding those dynamics is to identify patterns in PCN incidence, because when patterns of incidence are identified it is the outcome of underlying epidemiological processes and environmental heterogeneity that are being observed. In Scotland, Science & Advice for Scottish Agriculture (SASA) have been collecting georeferenced data on PCN incidence for many years as part of the statutory testing programme. These data represent a rich and valuable source of epidemiological information that have never been analysed to search for patterns of incidence. This would provide an unprecedented overview of the historical and evolving PCN situation in Scotland, and the incentive to search for the driving factors responsible for observed patterns of disease. Potential driving factors include weather, soil characteristics, cultivar choice, the distribution of potato crops, other landscape characteristics such as topography, and the PCN species. Remarkably, however, we have no information regarding the relative importance of these factors as drivers of PCN incidence. For example, soil temperature is widely accepted as the key environmental variable affecting the life cycle of PCN, but is weather a more important driver of incidence than soil characteristics, such as pH? Similarly, is resistance a more important driver of incidence than the density of potato production, or than soil or weather conditions? This study provides the first landscape-scale analysis of PCN incidence and its drivers and aims to improve our epidemiological understanding of the Scottish PCN situation.

3 Methodology

Source code for the PCN classifier described below is publicly available through the Zenodo archive at https://doi.org/10.5281/zenodo.6498423.

3.1 PCN patterns

ArcGIS Pro 2.8.3 (Esri UK Ltd, Aylesbury) and its geoprocessing tools were used to map and analyse space and spacetime patterns of PCN incidence (occurrence of either species), to give an unprecedented overview of the historical and current PCN situation in Scotland. Coordinates of PCN infested fields (hereafter referred to as points) from 2011 to 2020 were derived from SASA's Seed Potato Users Data System (SPUDS). The SPUDS software is used to store data from growing crop inspections and statutory testing, and incorporates details of the crop (cultivar, grade, origin, etc.), field history (location, area, soil testing data, etc.) and inspection findings (pests, diseases, trueness to type, etc.). All coordinates and map outlines are shown projected to the British National Grid, with measurement units in metres.

Spatial analyses included: Directional Distribution Analysis, Summarize Within, Hierarchical Density-Based Spatial Clustering of Applications with Noise, Optimized Hotspot Analysis, and Optimized Outlier Analysis. Spatiotemporal pattern analyses included: Summarize Percent Change, and Emerging Hotspot Analysis. These analyses are described in brief below; interested readers can directed to the ArcGIS Pro geoprocessing tool reference documentation for more information.

3.1.1 Directional Distribution Analysis

Given incident points, this tool creates standard deviational ellipses or ellipsoids to summarize the spatial characteristics of geographic features: central tendency, dispersion, and directional trends.

3.1.2 Summarize Within

This tool summarizes the number of points within polygons (here, postcode districts) to produce choropleth (colour-coded) maps of counts of points.

3.1.3 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

This analysis works by detecting areas where points are concentrated and where they are separated by areas that are empty or sparse. Points that are not part of a cluster are labelled as noise. The tool uses a range of distances to separate clusters of varying densities from sparser noise. The parameter 'Minimum Features per Cluster' that determines the minimum number of features required to consider a grouping of points a cluster, was set to 5% of the population size (42, 32, and 10 for *G. pallida*, *G. rostochiensis*, and mixed populations, respectively).

3.1.4 Optimized Hotspot Analysis

Given incident points, this tool creates a map of statistically significant hotspots (high point intensity) and coldspots (low point intensity) using the Getis-Ord Gi* statistic. It evaluates the characteristics of the input data and adjusts the parameters of the analysis to produce optimal results, i.e., it automatically aggregates incident data, identifies an appropriate scale of analysis, and corrects for both multiple testing and spatial dependence.

3.1.5 Optimized Outlier Analysis

This tool quantifies the spatial distribution of similar and dissimilar values of autocorrelation in patterns of points, using the Anselin Local Moran's I statistic. Like the Optimized Hotspot

Analysis, it creates a map of statistically significant hotspots and coldspots, but also determines the locations of spatial outliers. It also evaluates the characteristics of the input data to produce optimal results.

3.1.6 *Summarize Percent Change*

This tool calculates the percent change in the number of point features within polygons over two equal comparison time periods.

3.1.7 *Emerging Hotspot Analysis*

This analysis identifies temporal trends in the occurrence of spatial hot and coldspots, such as intensifying, diminishing, and sporadic hot and coldspots. A time step interval of 1 year and spatial settings determined as optimal by the OHSA tool were used to create the space-time cube.

3.2 *PCN drivers*

It is often the case in epidemiology that many different factors interact to produce patterns of incidence. Such combinatorial patterns are hard to detect using simple models, like logistic regression, which assume a linear relationship between predictor variables (hereafter referred to as features) and the response variable of interest. Non-parametric machine learning algorithms, on the other hand, can approximate the relationship between features and response variables with arbitrary functional forms that enable them to capture complex higher-order interaction effects and non-linear relationships. This helps to uncover the combinatorial interactions that are driving epidemiological patterns. In particular, ensembles of 'decision trees' are widely used in many fields of science to analyse the importance of underlying features. A decision tree is an explainable machine learning algorithm all by itself, but beyond its transparency, decision trees have a variety of inbuilt interpretation measures that help to understand the logic behind predictions and rank the importance of features.

Data defining the coordinates of PCN infested fields (positive samples), field area, and number of years since last cropped with potato were derived from SPUDS for the period spanning 2013-2020. These data were combined with features (variables) relating to soil characteristics (National Soil Map, National Soils Inventory for Scotland, FAO World Reference Base for Soils), topography (Ordnance Survey terrain-50 data), land cover (UKCEH Land Cover Maps, 25m resolution), potato crop distributions (SASA), and weather (UK Met. Office) to create an integrated PCN database containing 93 features (predictor variables). Weather and potato crop distribution data were matched to the year the sample was taken.

Feature engineering was used to generate meaningful potato crop distribution features (e.g., distance to nearest seed crop, distance to nearest ware crop) and a set of 28 bioclimatic features representing long-term averages of hourly temperature and precipitation variables over months and quarters of the year.

Data for individual years were gridded as 25 km^2 hexagonal cells using ArcGIS. This was the largest grid size determined as optimal by the OHSA tool for individual years. The Summarize Within tool was used to calculate the total number of positive samples, seed crops and ware crops, as well as the total area of seed and ware crops. All other features were averaged over each grid cell. Data from each annual grid was collated into a data table for machine learning.

Entries containing missing information were removed as were collinear variables ($r > 0.7$), leaving a total of 1036 entries with 60 features for modelling. The prediction task was framed as a binary classification problem, whereby grid cells containing positive samples were labelled as class 1 and grid cells with no positive samples were labelled as class 0. This produced a relatively balanced dataset for modelling, with 63% of cells labelled as class 1 and 37% as class 0.

A *k*-fold cross-validation procedure was used to train and tune ensembles of decision trees (classification ensemble) and obtain an unbiased estimate of the generalization accuracy of the entire model-building process to unseen data. This consisted of a 10-fold cross-validation for hyperparameter optimization using the Bayesian optimization approach, and a 10-fold cross-validation for model assessment. Predictive performance on the test folds was assessed using various binary classification metrics. The optimised model was then retrained using all the data to produce the finalised version. Ensemble learning was implemented using MATLAB's `fitcensemble` function.

To investigate the driving factors of patterns of PCN incidence, the importance of each feature for classification of incidence in the finalised model was estimated using MATLAB function `predictorImportance`. The function estimates predictor (feature) importance for each weak learner (tree) in the classification ensemble and returns the weighted average. The output contains an importance score for each feature in the dataset, which is essentially a measure of the number of times a feature is used to 'split a node' (e.g., ask a question or set a condition) and form a new branch or sub-tree, weighted by the number of samples it splits. A higher feature importance score therefore represents greater value for prediction. The feature importance scores were normalised to [0,1].

3.3 PCN prediction & scenario analysis

Although the inbuilt feature importance measures of decision trees are very useful for *explaining* complex relationships between features and the response, there are many other ML algorithms that may be more suited for *predicting* PCN incidence. To perform a scenario analysis of the impact of various potential management options on PCN incidence in Scotland, a suite of 32 machine learning algorithms were applied to the gridded incidence data to provide an initial exploration of the most accurate approach. This included Discriminant Analysis, Logistic Regression, Naïve Bayes, Support Vector Machines, k-Nearest Neighbour, and various types of Neural Network and Classification Ensemble techniques. A 5-fold cross-validation technique was used for performance testing.

Support Vector Machines (SVM) produced the lowest misclassification rates and were selected for further improvement. For SVM training, tuning of hyperparameters, and testing, the MATLAB function `fitcsvm` was used together with the *k*-fold cross-validation procedure described above. The finalised model was then used for scenario analyses to investigate the impact of changes to feature values on predicted PCN incidence. The top 2 most important features from each category of driver (crop distributions, field-specific, topography, soil characteristics) that were identified in the analysis of PCN drivers were adjusted in value by 25% (both an increase and then a decrease) with all other features held constant. The total number of grid cells classified as infested was compared to the baseline (unadjusted) predicted value to provide a factor change in PCN infestation levels. Weather features were excluded from this analysis as these cannot be modified to combat PCN. For a climate change risk assessment of blackleg see Skelsey et al. (2018).

4 Results

4.1 *PCN patterns*

4.1.1 *Temporal patterns*

To visualise how PCN incidence has changed over time, the total number of fields infested by *G. pallida*, *G. rostochiensis*, and mixed populations was plotted over time (Fig. 1). Whereas incidence of *G. pallida* is increasing over time, *G. rostochiensis* is in decline and incidence of mixed populations remains relatively static.

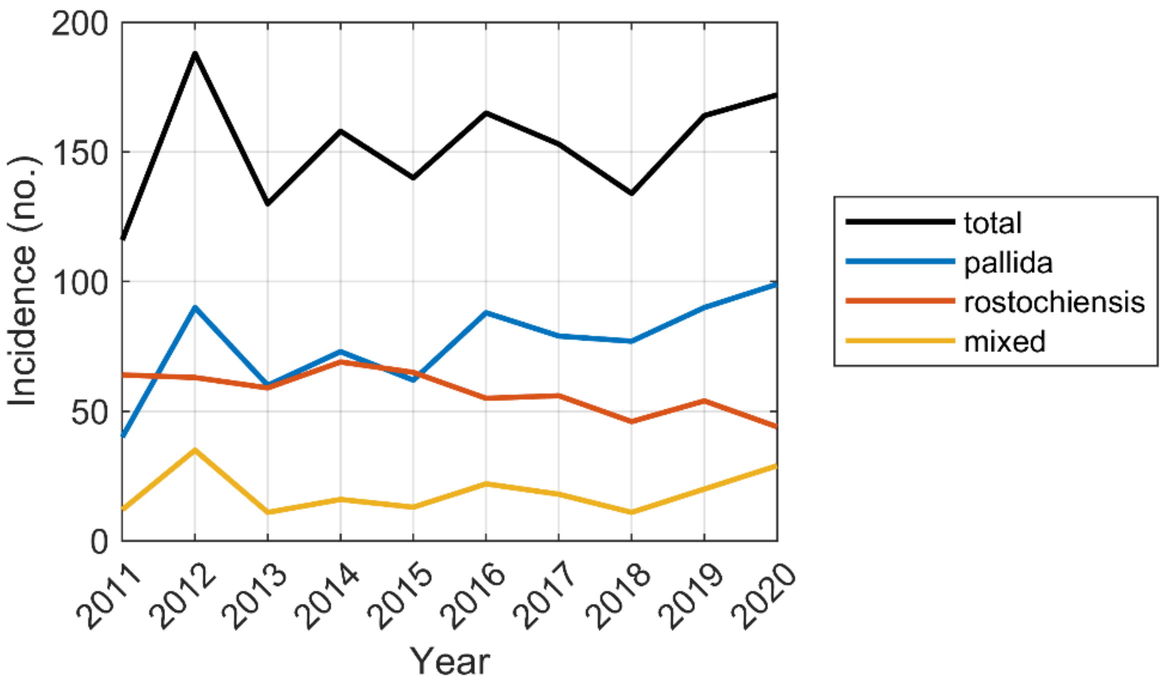


Fig. 1. Total number of fields testing positive for G. pallida, G. rostochiensis, and mixed populations in Scotland

4.1.2 Spatial patterns

The Directional Distribution Analysis revealed that the distribution of sites infested by *G. pallida, G. rostochiensis,* and mixed populations all share a similar central tendency and direction of distribution, but that *G. rostochiensis* is slightly more dispersed, reflecting a higher number of positive samples in northern and western locations.

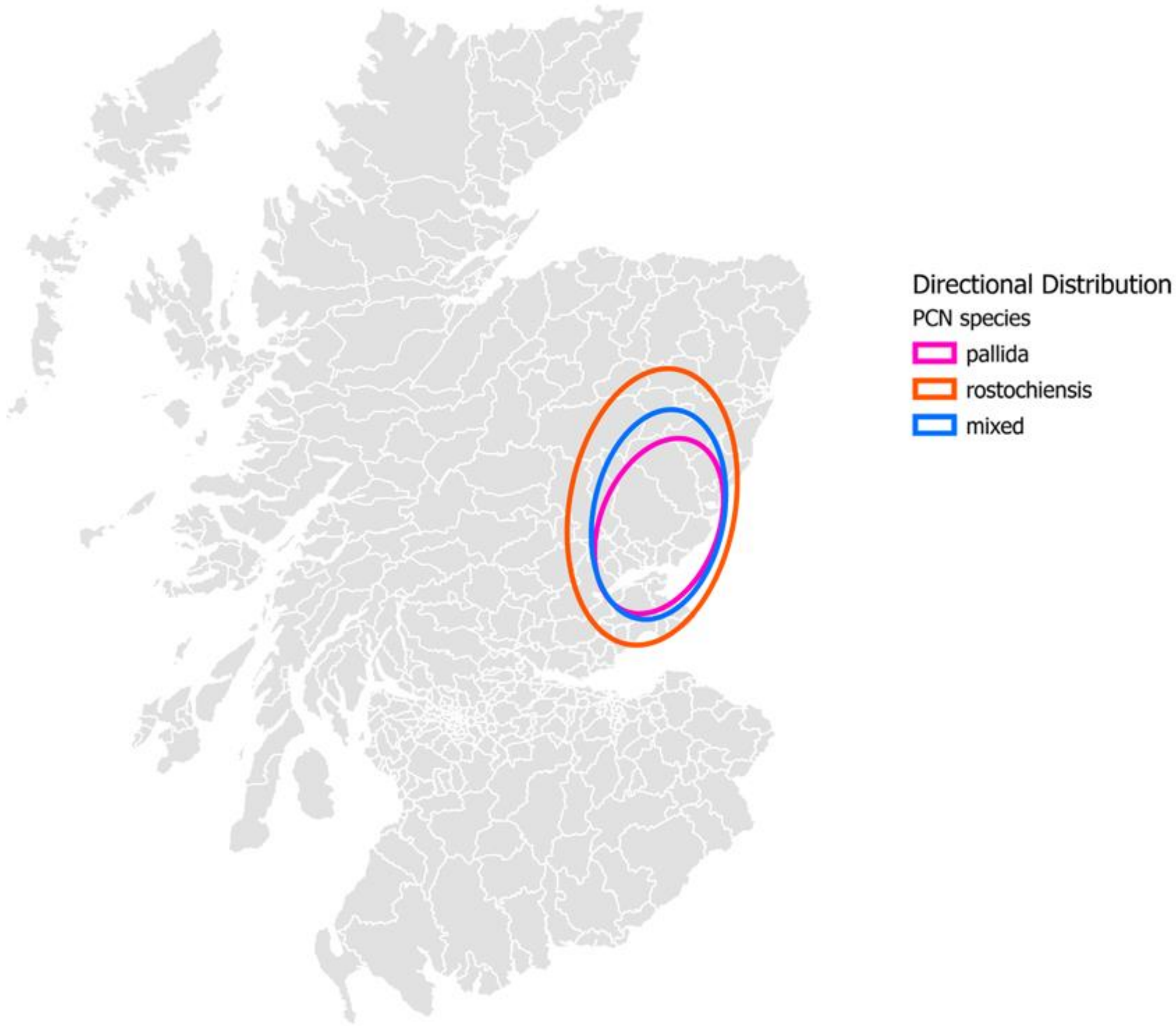


Fig. 2. Standard distance map representing the distribution of all positive sample locations of G. pallida, G. rostochiensis, and mixed populations in Scotland, 2011-2020.

Choropleth maps of PCN incidence confirmed the results of the Directional Distribution Analysis, and further revealed that incidence is most intense in the Dundee, Perth, and Kirkcaldy postcode areas for the three population types (Fig. 3).

The clustering analysis (HDBSCAN) identified 2 distinct, neighbouring clusters of PCN infestation in the Dundee postcode area for *G. pallida* and mixed populations, whereas *G. rostochiensis* is comprised of a large cluster spanning the Fife, Dundee, Perth, and south Aberdeen postcode areas, and a second cluster spread out along the NE coast of the country in the north Aberdeen and Inverness postcode areas (Fig. 4).

The Optimized Hotspot Analysis confirmed that the clusters identified by HDBSCAN were all statistically significant spatial clusters of high values of incidence (hotspots), except for the *G. rostochiensis* cluster along the NE coast, which was comprised of coldspots or spatial clustering of low values of incidence (Fig. 5). This is due to the relative isolation of these positive sample locations, which are few and scattered compared to those in the Dundee and south Aberdeen postcode areas.

The Optimized Outlier Analysis provided an assessment of incidence patterns and spatial outliers within the previously identified contiguous regions of hotspots across the Dundee and Aberdeen postcode areas (Fig. 6). For *G. pallida* there were scattered High-Low outliers in the Kirkcaldy and Perth postcode areas, and for *G. rostochiensis* these could be found in Kirkcaldy and further North in the Aberdeen and Inverness postcode areas. These show that incidence was high there relative to neighbouring locations, indicating a lower efficacy of PCN prevention/management compared to the surrounding area. Low-High outliers for all three population types were found scattered throughout the Dundee and Aberdeen postcode areas. These show that incidence was low there relative to neighbouring locations, indicating a higher efficacy of PCN prevention / management. The Dundee and Aberdeen postcode areas are characterised by dense potato production and high PCN incidence, therefore these Low-High outlier locations are worthy of further investigation to determine what, if any, differences in PCN management are being practiced.

Fig. 3. Choropleth maps showing a count of all positive sample locations of (a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020.

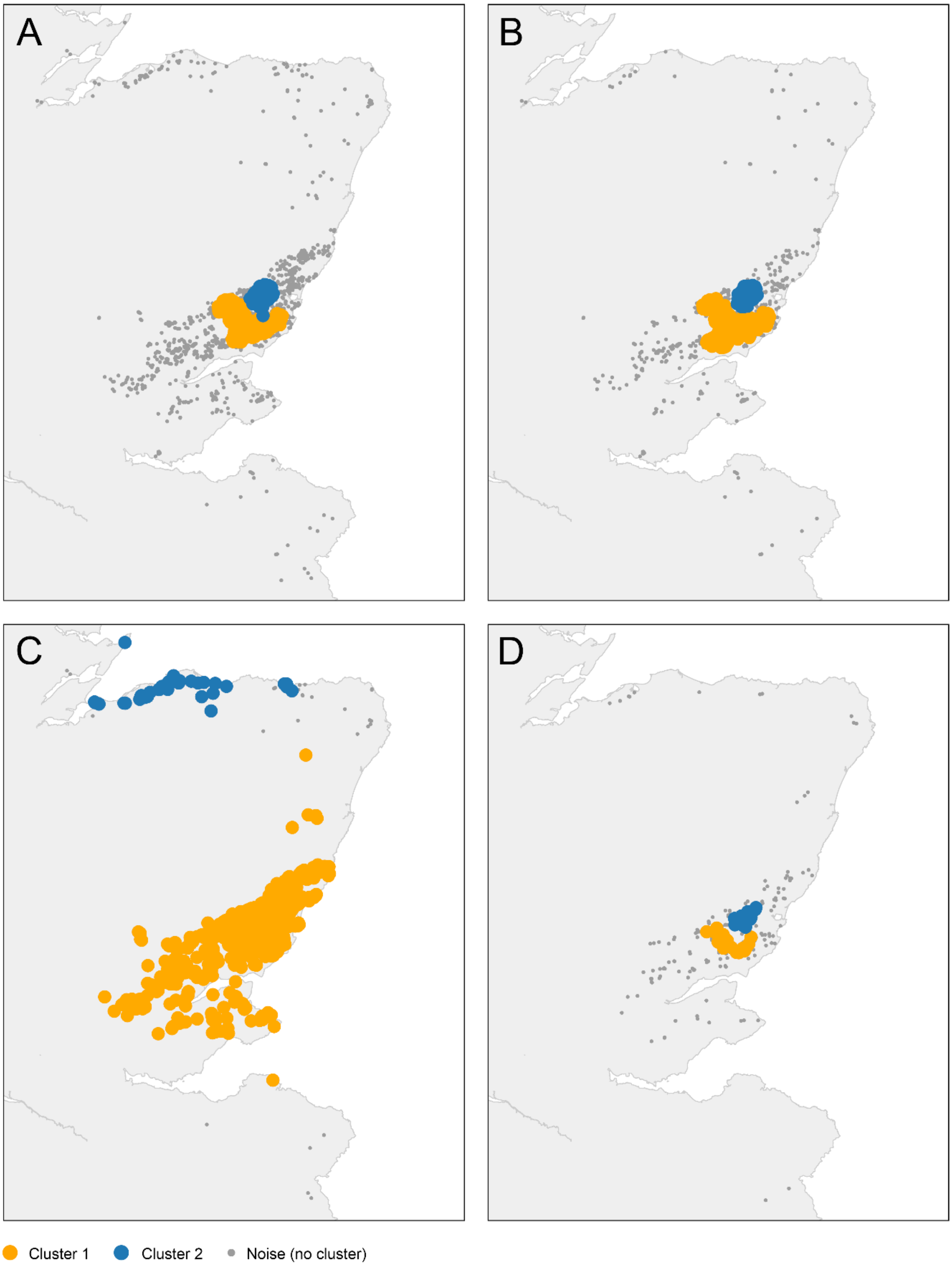


Fig. 4. HDBSCAN results showing clusters of positive sample locations of (a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020. Noise is data not assigned to any cluster.

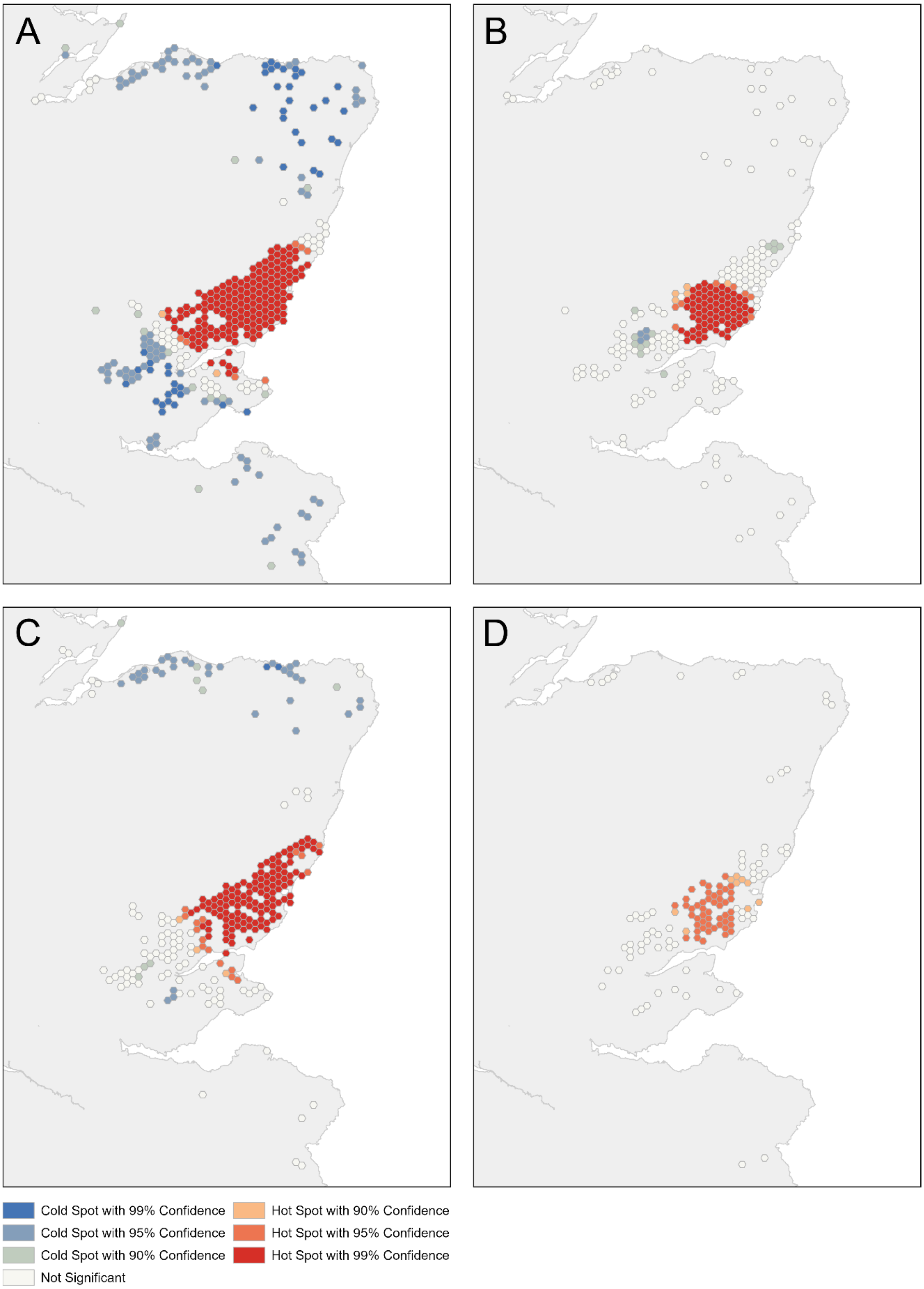


Fig. 5. Statistically significant hot and coldspots derived by Optimized Hotspot Analysis from positive samples locations for(a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020.

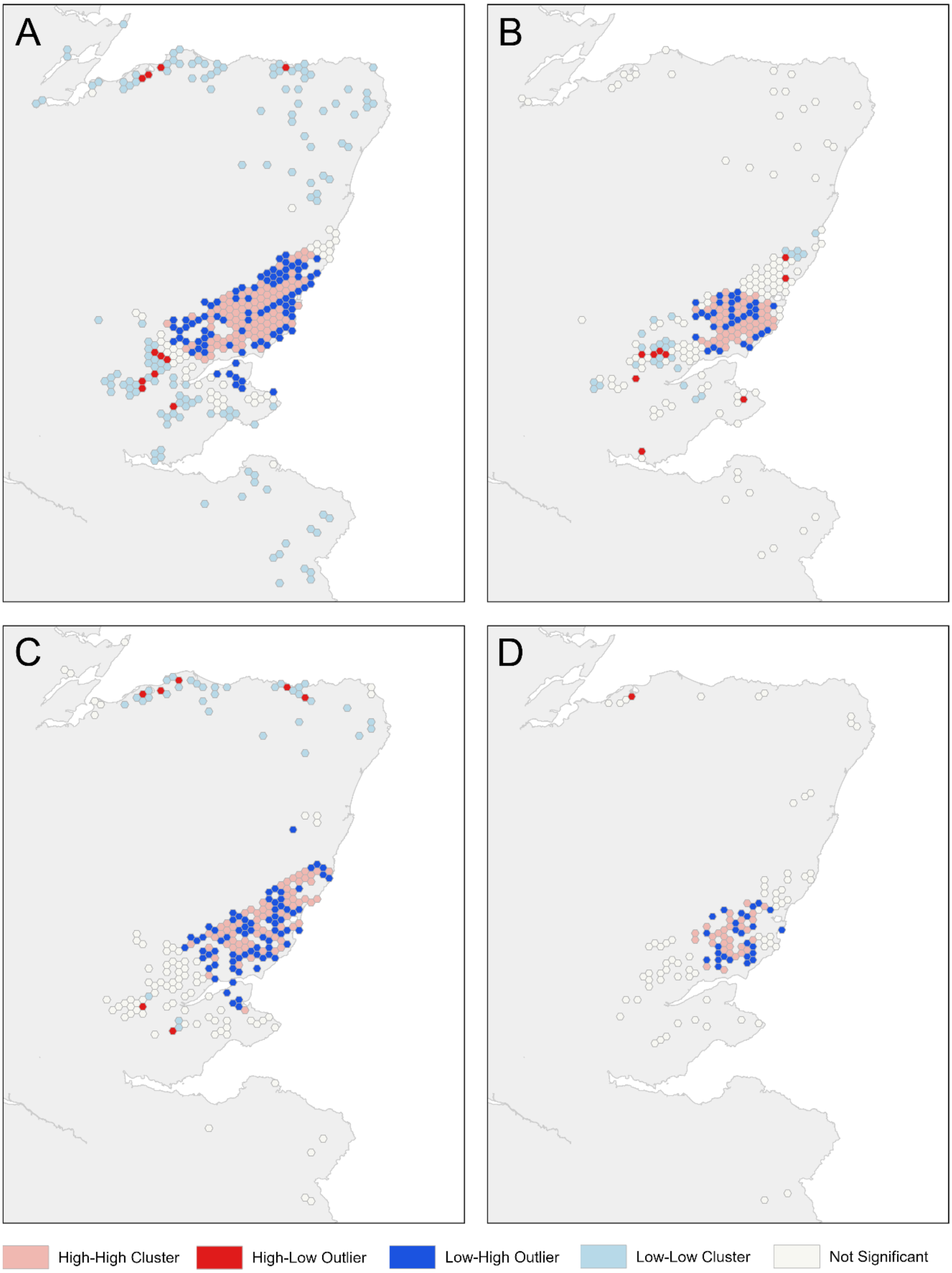


Fig. 6. Statistically significant hot and coldspots and spatial outliers derived by Optimized Outlier Analysis from positive samples locations for (a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020.

4.1.3 Spatiotemporal patterns

Choropleth maps of the percentage change in PCN incidence between the first and second halves of the study period suggested that changes in the spatial distribution of incidence of one species are partially mirrored by opposing changes in the other (Fig. 7). Of the 10 postcode districts where incidence of *G. pallida* had increased, *G. rostochiensis* was found in nine and had declined in six or declined/no change in seven. Of the 10 postcode districts where incidence of *G. rostochiensis* had increased, *G. pallida* was found in seven and had decreased in four. Of the 10 postcode districts where mixed populations had increased, *G. pallida* was found in nine and had increased in six, whereas *G. rostochiensis* had increased in five and decreased in five. Postcode districts DD5, DD8, and KY15 showed an increase in incidence of all three population types.

The Emerging Hotspot Analysis was more revealing with regards to the nature of temporal trends in patterns of incidence as it was performed at a finer spatial (5 km) and temporal (annual) resolution (Fig. 8). This analysis revealed that *G. pallida* was mostly intensifying across time and space in the north Kirkcaldy, east Perth, Dundee, and south Aberdeen postcode areas, whereas in the same areas incidence of *G. rostochiensis* was mostly persisting, except for a large swathe of land across the Dundee postcode area where incidence was diminishing. Both Figs. 7 and 8 are useful indicators of where resistance and other control options could best be deployed to better combat crop losses to PCN and loss of land availability for seed production.

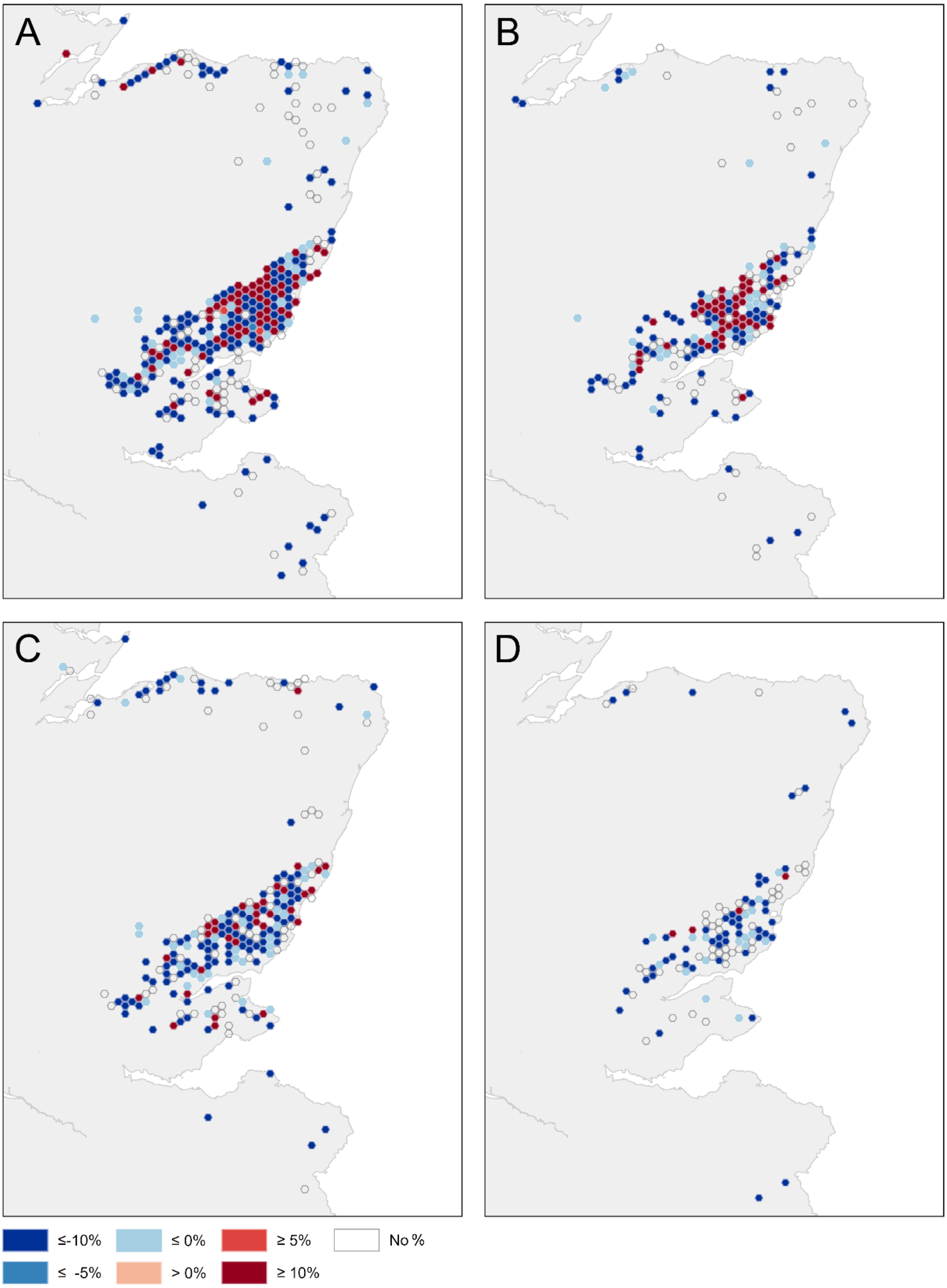


Fig. 7. Change in PCN incidence between the first (2011-2015) and second (2016-2020) halves of the study period, for (a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020.

Fig. 8. Space-time patterns of PCN incidence, derived by Emerging Hotspot Analysis from positive samples locations for(a) potato cyst nematode, (b) G. pallida, (c) G. rostochiensis, and (d) mixed populations in Scotland, 2011-2020.

4.2 PCN drivers

Performance metrics for the classification ensemble on the hold-out (unseen) test folds indicate that the model performed well in classifying PCN incidence in Scotland (Table 1). The optimised hyperparameters for the classification ensemble were: `Method = LogitBoost`, `NumLearningCycles` = 59, `LearnRate` = 0.89, `MinLeafSize` = 1, and `MaxNumSplits` = 571.

Table 1. Binary classification metrics for the PCN classification ensemble, averaged over 10 test folds

TPR	FPR	TNR	FNR	PPV	NPV	F1 score	Accuracy	AUROC
0.69	0.12	0.88	0.31	0.77	0.83	0.73	0.81	0.85

TPR = true positive rate, FPR = false positive rate, TNR = true negative rate, FNR = false negative rate, PPV = positive predictive value, NPV = negative predictive value, AUROC = area under the receiver operating characteristic curve

The results for the top 25 most important features ranked by importance score are given below (Fig. 9), and for the full set of features in the Appendix (not including the collinear features that were removed prior to analysis). Variables relating to the surrounding potato crop distributions were the most important, comprising four out of the ten most important features. It is of interest to note that soil characteristics were of low importance, with pH being the only soil feature in the top 10.

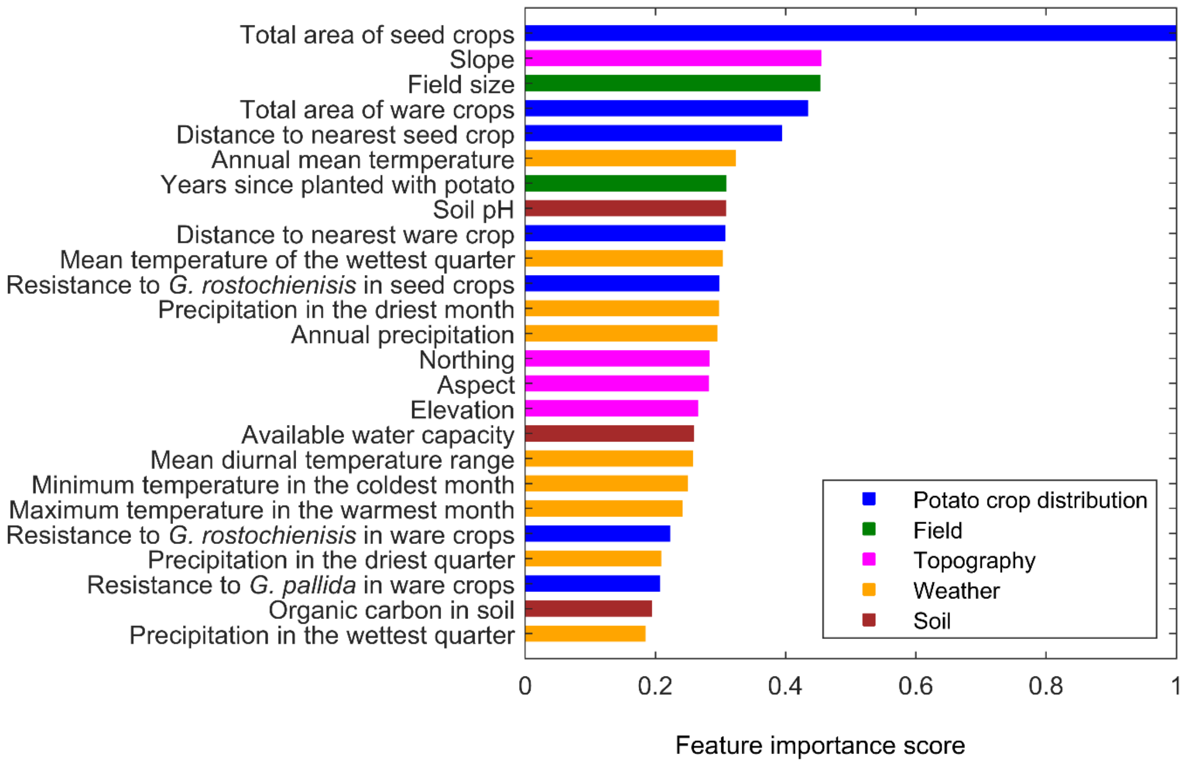


Fig. 9. Normalised feature importance ranking of the classification ensemble for PCN incidence, colour-coded by data source.

4.3 PCN prediction & scenario analysis

Performance metrics for the SVM on the hold-out (unseen) test folds indicate that the model was superior to the classification ensemble, and confidence can be placed in the results of the scenario analysis (Table 2). The optimised hyperparameters for the SVM were: `BoxContraint = 0.658`, `KernelScale` = 6.348, and `KernelFunction = gaussian`.

Table 2. Binary classification metrics for the PCN Support Vector Machine, averaged over 10 test folds

TPR	FPR	TNR	FNR	PPV	NPV	F1 score	Accuracy	AUROC
0.71	0.12	0.88	0.29	0.77	0.84	0.74	0.82	0.87

TPR = true positive rate, FPR = false positive rate, TNR = true negative rate, FNR = false negative rate, PPV = positive predictive value, NPV = negative predictive value, AUROC = area under the receiver operating characteristic curve

The scenario analysis revealed that changes to soil pH were by far the most effective in reducing incidence (Table 3). Both an increase and decrease in pH served to reduce incidence, with a shift to more acidic soils reducing incidence to zero. Similarly, moving fields southwards or northwards both served to decrease incidence. This was likely due to the relative sparsity of crop production in the north and south of the country. Interestingly, increasing the capacity of soils to hold water produced a decrease in incidence. This could be due to the relationship between soil texture, water content and nematode motility.

Table 3. Sensitivity of the PCN Support Vector Machine to changes in feature values

Feature[1]	Change in value	Factor change to incidence[2]	Change in value	Factor change to incidence
1	+25%	1.0350	-25%	0.9534
2	+25%	1.0204	-25%	0.9709
3	+25%	0.9825	-25%	1.0058
4	+25%	1.0204	-25%	0.9738
5	+25%	0.9825	-25%	1.0058
6	+25%	0.1691	-25%	0.0000
7	+25%	0.8921	-25%	0.8426
8	+25%	0.9563	-25%	1.0880

[1] 1 = Total area of seed crops, 2 = Slope, 3 = Field size, 4 = Total area of ware crops, 5 = Years since planted with potato, 6 = Soil pH, 7 = Northing, 8 = Available water capacity
[2] Ratio of the total number of infested grid cells to the baseline predicted value

4.3.1.1 Sub-sub-subtitle

Text under the subtitle of Heading level 4 (this level will not appear under contents)

5 Discussion

This study provides the first quantitative evidence of statistically significant spatial and spatiotemporal patterns of PCN incidence at the landscape-scale. It gives an overview of the distribution of *G. pallida* and *G. rostochiensis* in Scotland over a 10-year period (2011-2020), revealing where the hotspots of infestation are and the areas where the number of positive soil samples of both species are changing over time. It also provides the first quantitative analysis of the principal drivers of PCN incidence at the landscape-scale, and a model for predicting incidence with an accuracy of 82%.

The results of the ArcGIS analyses could be used to tailor control options, financial incentives, and restrictions according to the geographic distribution of PCN incidence and the corresponding risk of pest spread. For example, the Optimized Hotspot Analysis revealed significant spatial clusters of high PCN incidence across the Dundee, Perth, and Kirkcaldy postcode areas (Fig. 5). Tighter restrictions on the use of infested land for growing ware and/or a requirement to test ware land for PCN could be introduced in such areas to reduce pest spread, in addition to tighter restrictions on the disposal of crops, waste and soil (e.g., from grading, packing, and processing tubers). Conversely, in areas comprised of coldspots of incidence, such restrictions could potentially be eased or lifted. Similarly, the standard sampling rates for official pre-crop soil testing for PCN could potentially be adjusted according to the distribution of PCN incidence, with higher rates in areas where risk of pest spread is high and lower rates elsewhere.

Targeted control methods could also be developed for individual species. For example, *G. pallida* is intensifying over space and time in in the north Kirkcaldy, east Perth, Dundee, and south Aberdeen postcode areas, and that there is a corresponding decrease in *G. rostochiensis* incidence in the Dundee area (Figs. 7 & 8). Financial incentives could be offered in the regions where *G. pallida* is becoming dominant to encourage more widespread cultivation of varieties with moderate or high levels of resistance to *G. pallida*.

The feature importance analysis revealed that the total area of seed and ware crops and the distance to the nearest seed crop were important drivers of PCN incidence. This is not unexpected given that the main route by which PCN spreads is through the movement of infested material, primarily soil which may be transferred with tubers, plants, waste material or farm machinery. This result highlights a need for improved field hygiene and a strategy to prevent or minimise cross-contamination. There is currently little incentive or accountability for the ware sector to keep land PCN-free, therefore a national PCN strategy that addresses this issue, either through financial incentives or new legislation, could be key to preventing the ongoing loss of land for seed production. Development of a national strategy for PCN is one of the key aims of a new Scottish Government research project entitled "Delivering a sustainable potato industry for Scotland through management of Potato cyst nematode (PCN)". The results of this study suggest a potential role for targeted pest-geography-IPM solutions within that strategy. Nevertheless, it is recommended that the ArcGIS analyses are repeated each growing season to provide an up-to-date overview of the evolving PCN situation in Scotland.

The analysis of PCN driving factors provided further new epidemiological insights. The results showed that slope, the number of years between potato crops, and soil pH were all important drivers of incidence. These are all features that can potentially be modified by growers. Specifically, the scenario analyses revealed that modification of soil pH was a highly effective strategy for reducing incidence. This suggests there may be scope to reduce PCN levels through simple soil amendments. Controlled environment pot trials to better quantify the relationships between environmental variables, soil characteristics and PCN infestation/decline would be beneficial in future work. The soil features used in the current project were derived from the National Soil Map, National Soils Inventory for Scotland, and the FAO World Reference Base for Soils, and may not be fully representative of conditions in potato crop locations in the intervening years since these datasets were compiled. The results of such trials would provide useful empirical data on the drivers of PCN incidence/severity, and an improved set of soil features for PCN modelling.

Improving our ability to forecast when and where PCN will spread in the future is one of the key recommendations of the Report of the Scottish PCN Working Group (Toth *et al.* 2020). Within the confines of this limited study, it was possible to produce a machine learning model (the SVM classifier) that can predict gridded PCN incidence to an accuracy of 82%. Further work is planned to develop predictive tools for PCN for the potato industry under the Scottish Government funded Strategic Research Programme 2022-2027 and through various external funding sources.

6 References

1. Toth, I., Burnett, F., Burgess, P., Herron, C., Pickup, J. 2020. Potato Cyst Nematode (PCN) and the future of potato production in Scotland: Report of the Scottish PCN Working Group. Plant Health Centre. https://www.planthealthcentre.scot/publications/pcn-working-group-final-report.
2. Skelsey, P., Humphris, S., Campbell, E., Toth, I. Threat of establishment of non-indigenous potato blackleg and tuber soft rot pathogens in Great Britain under climate change. PloS one, 13(10): e0205711.

7 Appendix

Table 1. Normalised feature importance scores of the classification ensemble for PCN incidence.

Feature	Score
Total area of seed crops	1.0000
Slope	0.4551
Field size	0.4532
Total area of ware crops	0.4344
Distance to nearest seed crop	0.3946
Annual mean temperature	0.3236
Years since planted with potato	0.3092
Soil pH	0.3085
Distance to nearest ware crop	0.3076
Mean temperature of the wettest quarter	0.3039
Resistance to *G. rostochienisis* in seed crops	0.2984
Precipitation in the driest month	0.2976
Annual precipitation	0.2949
Northing	0.2831
Aspect	0.2824
Elevation	0.2656
Available water capacity	0.2594
Mean diurnal temperature range	0.2581
Minimum temperature of the coldest month	0.2501
Maximum temperature of the warmest month	0.2418
Resistance to *G. rostochienisis* in ware crops	0.2225
Precipitation in the driest quarter	0.2091
Resistance to *G. pallida* in ware crops	0.2073
Organic carbon in soil	0.1948
Precipitation in the warmest quarter	0.1848
Resistance to *G. pallida* in seed crops	0.1783
Temperature seasonality	0.1761
Isothermality	0.1749
Precipitation in the coldest quarter	0.1729
Na by ARD extraction	0.1595
Mean temperature of the driest quarter	0.1484
Ratio of carbon to nitrogen	0.1368
K by ARD extraction	0.1132
Mg by ARD extraction	0.1130
Exchangeable K	0.0977

Table 1 (continued)

Loss on ignition	0.0947
Percentage of clay	0.0918
Fe by ARD extraction	0.0883
Mn by EDTA extraction	0.0831
Cd by ARD extraction	0.0764
Exchangeable Mg	0.0755
Ca by ARD extraction	0.0734
Cr by EDTA extraction	0.0723
T by ARD extraction	0.0720
Ni by EDTA extraction	0.0712
Cu by ARD extraction	0.0685
Easting	0.0684
Zn by EDTA extraction	0.0668
Pb by ARD extraction	0.0665
Percentage of sand	0.0642
Mn by ARD extraction	0.0636
Exchangeable Na	0.0605
Exchangeable Ca	0.0599
Percentage of silt	0.0584
Zn by ARD extraction	0.0441
Cu by EDTA extraction	0.0376
St by ARD extraction	0.0345
Total P2O5	0.0276
Mo by ARD extraction	0.0262
Cd by EDTA extraction	0.0105

Plant Health Centre
c/o The James Hutton Institute
Invergowrie,
Dundee, DD2 5DA

Tel: +44 (0)1382 568905

Email: Info@PlantHealthCentre.scot
Website: www.planthealthcentre.scot
Twitter: @PlantHealthScot