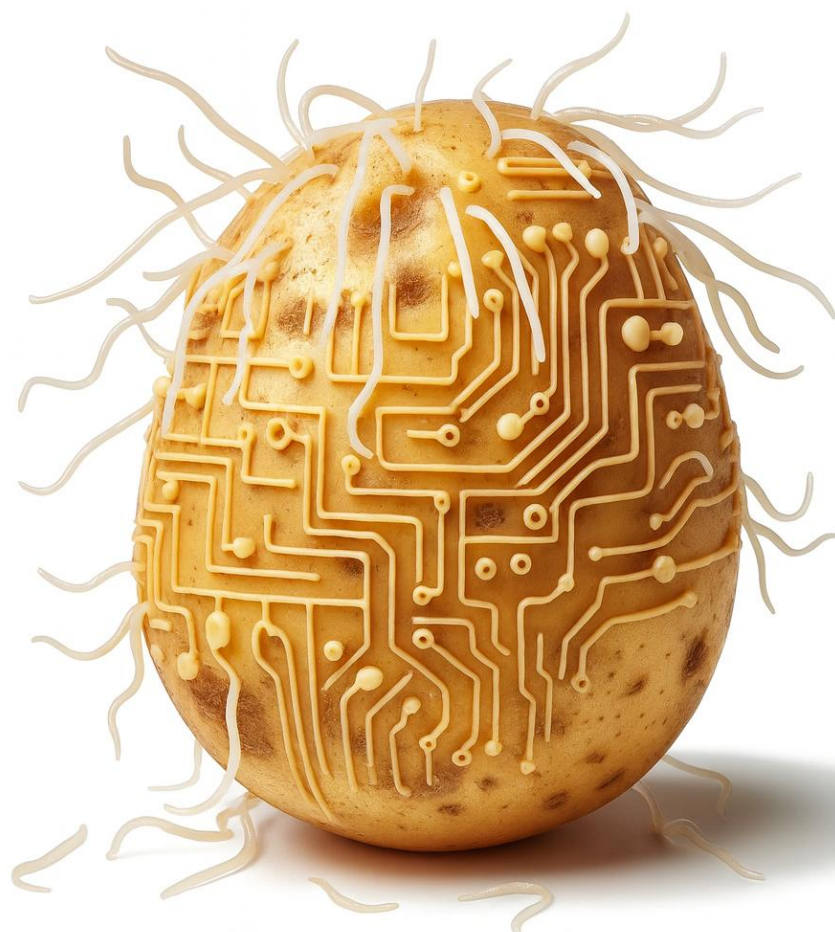


PCN-ML: Developing new models to predict PCN species presence / absence in Scottish soils

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 PHC2024/05

Authors: Peter Skelsey¹

¹Information and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA

Please cite this report as follows: P. Skelsey (2026). PCN-ML: Developing new models to predict PCN species presence / absence in Scottish soils. Project Final Report. PHC2024/05. Scotland's Centre of Expertise for Plant Health (PHC). DOI: <https://doi.org/10.5281/zenodo.21222963>

Available online at: planthealthcentre.scot/publications

Dissemination status: Unrestricted

Copyright: All rights reserved. No part of this publication may be reproduced, modified or stored in a retrieval system without the prior written permission of PHC management. While every effort is made to ensure that the information given here is accurate, no legal responsibility is accepted for any errors, omissions or misleading statements. All statements, views and opinions expressed in this paper are attributable to the author(s) who contribute to the activities of the PHC and do not necessarily represent those of the host institutions or funders.

Acknowledgements: I gratefully acknowledge Jon Pickup (SASA) for his assistance with data provision and reviews, and Claire Ghee (SASA) for her reviews.

Research Team: Peter Skelsey - computational epidemiologist

Details of Copyright Images: None.

Contents

1	Executive summary	3
2	Introduction and purpose of the PCN-Mapper app	6
3	System overview	7
3.1	Data inputs and structure	7
3.2	Analytical framework	7
3.3	Application workflow	8
3.4	User interface and outputs	8
4	Development of mapping components	9
4.1	Overview of the mapping approach	9
4.2	Interactive scatter charts linking sample data to mapped distributions	10
4.3	Geostatistical mapping methods	10
4.3.1	Kernel density estimation	10
4.3.2	Optimised hot spot analysis	10
4.3.2	Choropleth maps	12
4.3.3	Bubble charts	12
5	Development of machine learning models for nowcasting	14
5.1	Data sources and pre-processing	14
5.2	Exploration of modelling approaches	15
5.3	Cross-validation strategy and model selection	16
5.4	Final model architecture and deployment versions	17
5.5	Model performance and results	17
5.6	Interpretation of nowcast outputs	18
6	Discussion	19
7	Conclusions	21
8	References	22
	Appendix A — PCN-Mapper user guide	23

1 Executive summary

Background

Plant-parasitic nematodes (PPNs) are a major constraint on agricultural productivity worldwide, causing an estimated 8–15% of global crop losses annually, equivalent to approximately \$160 billion (Iqbal & Jones, 2017). Among these, potato cyst nematodes (PCN) are particularly damaging to potato production. Across Europe, PCN account for around 9% of yield losses (Turner & Subbotin, 2006), and in heavily infested soils yields may fall below the weight of seed planted (Whitehead & Turner, 1998).

In Scotland, two PCN species are present: *Globodera rostochiensis* and *G. pallida*. Female PCN form durable cysts containing hundreds of eggs that can remain viable in soil for up to 25 years. Cysts spread primarily via tubers, waste, soil movement, and contaminated machinery. Control therefore relies on strict biosecurity measures and regulatory oversight. Scottish legislation prohibits seed potato production on land confirmed as PCN-infested, while ware crops are permitted only under approved control programmes.

Species identification is central to effective management. *G. rostochiensis* can often be managed using resistant varieties combined with 6–7-year crop rotations, whereas *G. pallida* survives longer in soil, hatches over an extended period, and lacks strong varietal resistance. As a result, *G. pallida* typically requires longer rotations (10–15-years) and integrated control strategies, and poses the greater long-term threat to Scotland’s seed potato industry.

Statutory soil testing provides the regulatory foundation for PCN management and generates a substantial spatial dataset describing confirmed infestations across Scotland. However, while this system underpins certification and biosecurity, it also presents an opportunity: the accumulated testing data contain rich spatial and temporal structure that can be analysed to improve understanding of species-specific risk patterns at national scale.

Key Research Objectives

To address these challenges, this project aimed to develop an interactive PCN mapping and prediction application that brings together PCN testing data, spatial analysis, and machine learning. The app visualises the distribution and relative intensity of *G. pallida* and *G. rostochiensis* across Scotland, highlighting where infestations are most concentrated and where current risk is greatest.

Importantly, the application is designed to complement traditional soil testing. While statutory sampling remains the regulatory cornerstone for PCN management, all sampling approaches, including internationally adopted testing standards, are subject to detection limits. Spatial modelling provides an additional, independent assessment of risk that can support interpretation of negative or borderline results and help identify areas where underlying infestation may be present but below routine detection thresholds.

By combining observed and model-derived outputs, the app provides spatially explicit estimates of current infestation likelihood across Scotland. This includes areas that have not been directly sampled and locations where populations may exist at low or emerging densities below routine detection thresholds. These integrated analyses support evidence-based policy, guide surveillance and biosecurity prioritisation, and deliver species-specific risk information to growers and decision-makers.

Research undertaken

The research was organised around four linked objectives:

(O1) Collate, clean, and standardise statutory PCN testing data from multiple years and regions of Scotland, integrating these records with spatial reference data for analysis at multiple spatial scales.

(O2) Apply geostatistical methods to map the observed distribution of *G. pallida* and *G. rostochiensis*, identify areas of high infestation intensity, and derive spatial risk indicators.

(O3) Develop and train machine learning models to generate species-specific probabilistic estimates of current PCN presence and relative risk across Scotland, derived from spatial and temporal structure in the statutory testing dataset.

(O4) Integrate these analytical outputs into an interactive desktop application that enables SASA to visualise distributions, explore species-specific risks, and support targeted management advice and policy decisions.

MATLAB® (version R2025b) was used for all data processing, analysis and modelling.

Main findings

(O1) Statutory PCN testing data from 2010–2024 was provided by SASA and combined with postcode area, district, and sector identifiers to enable computation of regional summary statistics. The corresponding postcode boundary polygons were obtained from the UK Postcode Boundaries dataset published by EDINA at the University of Edinburgh (<https://datashare.ed.ac.uk/handle/10283/2597>).

(O2) Using statistical mapping techniques, the distribution of *G. pallida* and *G. rostochiensis* across Scotland was analysed to identify areas of higher infestation and significant spatial clustering. Key indicators were summarised at postcode district and postcode sector levels using interactive maps and charts that link individual soil test results to their geographic locations. Together, these tools transformed statutory testing data into clear, interpretable spatial intelligence to support targeted surveillance and species-specific management decisions.

(O3) Machine learning models performed well when evaluated using independent test data that were not used to train the model, demonstrating reliable discrimination between *G. pallida*, *G. rostochiensis*, and negative samples. Spatial and temporal predictors contributed most to model performance, reflecting consistent regional and year-to-year patterns in infestation. The resulting outputs provide a statistical “nowcast” of current PCN risk — a probabilistic estimate of present infestation likelihood derived from spatiotemporal patterns in statutory testing data. These nowcasts estimate the probability of species presence in unsampled areas and may also highlight locations where infestation could exist at low or sub-threshold densities not consistently detected by routine soil testing. This illustrates the potential role of machine learning as a complementary tool for national surveillance and risk-based testing strategies.

(O4) The analytical outputs were integrated into an interactive mapping application, PCN-Mapper, enabling users to explore species distributions and risk patterns at multiple spatial scales across Scotland. The tool combines observed sample results with model-based nowcasts to generate intuitive visualisations and summary statistics. This integration transforms complex testing data into accessible, decision-ready intelligence, supporting SASA in assessing species-specific risk, prioritising surveillance, and communicating trends in PCN distributions more effectively.

Recommendations

- Following training of relevant SASA staff and an initial testing phase, the PCN-Mapper app could be updated at the end of each growing season as new PCN testing data become

available, ensuring that maps, models, and risk indicators remain representative of current PCN populations.

- The following recommendations, while subject to SASA's resource availability and careful consideration of grower confidentiality, have the potential to enhance the delivery of actionable plant health information and thereby increase the overall impact of PCN-Mapper:
 - It is recommended that PCN-Mapper outputs are incorporated into SASA communications and advisory material accompanying positive PCN test notifications, providing tailored advice on species-specific risk and management.
 - Growers receiving negative test results could also be provided with contextual information on local PCN risk, drawn from PCN-Mapper, to encourage preventative management in high-risk areas.
 - PCN-Mapper could be used as a communication tool in grower and stakeholder meetings to illustrate regional and species-specific risk patterns, support discussion of management options, and raise awareness of emerging trends.
 - The app produces a wide range of high-level regional summary statistics, and these could be provided to agronomists, major seed houses, and policy teams to inform strategic decisions and long-term management policy.
 - Consider publishing high-level regional summary statistics from PCN-Mapper on the SASA website, e.g., the aggregated postcode-area level statistics that are automatically generated and saved to the user's device will maintain grower confidentiality. Making these summaries publicly available would increase transparency, help the industry understand regional PCN pressures, and promote collective action to manage risk.
- PCN-Mapper could be promoted alongside the PCN Scenario Planner developed as part of the Scottish Government-funded PCN Action Scotland project, with each tool serving distinct but connected purposes: PCN-Mapper for visualising and communicating risk at landscape scale, and the Scenario Planner for guiding management decisions on infested land.
- PCN-Mapper provides an evidence-based tool for prioritising soil testing in higher-risk areas, supporting efficient allocation of surveillance resources.
- Key priorities for future research include:
 - Maintain and archive records of historical field boundaries to enable accurate linkage of PCN test data to the same physical location across years. This would improve the ability to analyse long-term trends, distinguish genuine changes in infestation from mapping artefacts, and validate model predictions over time.
 - Develop and retain records at the sub-field (sampling-unit) scale, as this level of detail would provide valuable predictor variables for machine learning. In particular, information on the previous potato variety grown and other management practices would enhance model performance and improve understanding of factors influencing PCN spread, persistence and decline.
 - Investigate opportunities for real-time data sharing and automated updates between PCN-Mapper and SASA's PCN testing database to streamline future reporting and ensure timely model updates.

2 Introduction and purpose of the PCN-Mapper app

Statutory soil testing for potato cyst nematodes (PCN) is a cornerstone of Scotland's plant-health strategy, ensuring that seed potatoes are grown only on land tested and found free from infestation and providing an effective defence against the spread of these pests. While statutory pre-crop testing is an internationally recognised requirement for seed production, it also generates a rich dataset that remains under-utilised. These records contain valuable spatial and temporal information on where PCN occurs and how risk varies by species. This project set out to create a desktop application (PCN-Mapper) that transforms these data into accessible visual intelligence. The app combines ArcGIS mapping capabilities and machine-learning analysis to identify species-specific risk patterns and provide additional insight alongside statutory testing, including estimates of infestation likelihood across the wider landscape.

Effective visualisation of PCN species risk is critical for translating complex analytical data into clear, actionable insights that help non-modellers quickly identify where problems are concentrated and how risk varies spatially. This helps to bridge the gap between scientific analysis and practical decision-making in the seed industry. Static data tables and written reports are not effective tools for understanding and conveying the landscape-scale epidemiology of pests and diseases, making visual tools like PCN-Mapper of great benefit. Presentation of PCN data in a visual, spatially explicit format and in high-level regional summary statistics also enhances communication of risk and encourages coordinated action and collective responsibility across the potato sector.

PCN-Mapper is intended primarily for use by SASA staff as the national custodians of official PCN testing data. However, the app provides SASA with an enhanced analytical and communication tool to support surveillance, policy advice, and stakeholder engagement. From this central hub, information can flow to other key groups: policymakers gain an accessible evidence base for ministerial briefings and strategic decisions; growers can receive species-specific risk information and management guidance via their official PCN test notifications; and agronomists and seed houses could benefit from improved regional overviews and predictions that inform planning and management. At a broader scale, PCN-Mapper provides high-level regional statistics suitable for public reporting (e.g., a webpage) and stakeholder discussion, helping to raise awareness of regional PCN pressures, promote preventative management in high-risk areas, and safeguard the resilience and international reputation of Scotland's valuable seed potato industry.

PCN-Mapper complements the new "PCN Scenario Planner" developed through the PCN Action Scotland Project (<https://dss.pcnhub.ac.uk/>). PCN Action Scotland is a Scottish Government-funded initiative aimed at protecting Scotland's potato industry by improving PCN management and maintaining the country's high plant-health status. When a grower receives a positive soil test result, the Scenario Planner can be used to explore management options that accelerate population decline and assess their economic feasibility. In contrast, PCN-Mapper provides a broader, landscape-scale perspective by visualising and communicating patterns of species-specific risk across Scotland. Together, these tools offer an integrated approach to understanding, managing, and mitigating the threat of PCN. Future development may also incorporate additional farm-management information (e.g., crop rotations, varietal resistance deployment, and changes in field boundaries), allowing models to better capture within-field drivers of PCN risk.

3 System overview

This section described how PCN-Mapper integrates data, analysis and interface components into a single operational system. The app unites statutory PCN testing data with spatial analysis, mapping, and machine learning outputs to deliver interactive visualisations of observed PCN species distributions, quantitative risk indicators, and estimates of the likelihood of infestation in unsampled areas. The following subsections describe the data inputs, analytical methods, and system design that underpin the application.

3.1 Data inputs and structure

Statutory PCN soil testing data from 2010 to 2024 ($n = 26,296$ fields) were extracted from SASA's Seed Potato Unified Data System ([SPUDS](#)), which maintains regulatory testing records alongside a wider set of crop and field management information. Each field record includes a unique soil test identifier, ten-digit Ordnance Survey National Grid Reference (1 m resolution) coordinates, field area, the number of sampling units per field, and the total number of samples collected. For fields testing positive for PCN, additional attributes record the area and number of samples per sampling unit, together with the number of samples positive for *Globodera pallida*, *G. rostochiensis*, or both species (mixed samples). All coordinates were converted from OS National Grid References to British National Grid eastings and northings (metres) for analysis, and to geographic coordinates (latitude, longitude) for visualisation within PCN-Mapper.

Geospatial vector data (shapefiles) defining Scottish postcode areas, sectors and districts were used to calculate regional summary statistics and to visualise spatial results within the PCN-Mapper application (Pope, 2017).

Together, the statutory testing data and spatial boundary layers form the foundation for the analytical framework of PCN-Mapper, enabling both geostatistical mapping of observed infestations and machine-learning predictions of risk across unsampled areas.

3.2 Analytical framework

The analytical design of PCN-Mapper comprises two complementary strands. The first is a geospatial analysis, which maps the observed distribution and intensity of both PCN species across Scotland and derives a wide range of spatial risk indicators summarising infestation levels within each postcode area / district / sector. The second is a machine-learning analysis, developed to generate predictive "nowcasts" of PCN risk — probabilistic estimates of current infestation likelihood derived from spatial and temporal relationships identified in the statutory testing dataset. These nowcasts extend risk assessment beyond directly sampled fields and may help identify areas where underlying populations remain below routine detection thresholds.

Both analytical strands operate on the same cleaned and standardised dataset and use shared geographic reference layers to ensure spatial consistency. Their outputs are designed to be integrated within the PCN-Mapper interface, providing complementary insights, one empirical, based on observed data, and the other predictive, based on modelled relationships.

All geostatistical and machine-learning procedures were implemented in MATLAB® (version R2025b) using modular, version-controlled scripts to ensure transparency, reproducibility, and ease of future updates.

3.3 Application workflow

The PCN-Mapper application operates through a fully automated workflow from data input to visualisation, comprising the following steps:

1. Datasets, trained models, and geospatial shapefiles are verified and loaded into the system.
2. Geospatial analyses, regional summary statistics, and model predictions are generated.
3. Resulting output files are automatically saved to the user's machine.
4. Interactive maps and plots are rendered in the app interface.
5. Users can then interact with the outputs, for example, by selecting years to display, toggling species, choosing different plot types, or exporting maps as image files.

The system is designed for seamless updating: when new PCN testing data are added to the raw data files, PCN-Mapper automatically regenerates the machine-learning models, geospatial analyses, and visual outputs, ensuring that all results remain current and representative of the contemporary PCN situation.

3.4 User interface and outputs

The PCN-Mapper user interface was designed with accessibility, clarity, and reliability in mind (Fig. 1). The layout provides a clean, single-window workspace in which users can load data, run analyses, and view outputs without navigating through multiple menus. The maps are fully interactive, allowing users to pan, zoom, and explore different regions seamlessly, with dynamic updates to tooltips and overlays as the view changes. All interactive controls include tooltips (small pop-up text boxes that appear when the cursor hovers) explaining their function, and the visual elements—such as maps and plots—also display contextual tooltips that reveal underlying data values when hovered over. Informative alerts (pop-up messages) guide the user if required files are missing or an operation fails, explaining both the issue and the corrective action. This design ensures that the app remains intuitive for occasional users while providing sufficient feedback and transparency for expert users conducting analysis or model validation.

Each time PCN-Mapper is launched, it automatically regenerates all analytical outputs to ensure results remain current. Three main output files are produced. The `model_summary.csv` file provides a concise record of machine-learning model performance, including key test metrics used to evaluate predictive accuracy. The `model_predictions.csv` file contains spatially referenced probabilities of species presence and relative risk across Scotland, providing a “nowcast” of current PCN risk derived from spatial-temporal patterns in the statutory dataset. Finally, the `regional_statistics.csv` file summarises a wide range of derived statistics and risk indicators for every postcode area, district, and sector, supporting both local management decisions and high-level regional reporting. All output files contain metadata explaining the variables included, their units, and the methods used to generate them, ensuring transparency and reproducibility of results.



*Figure 1 – PCN-Mapper interface. (1) Main map display; (2) Start year selector; (3) End year selector; (4) Plot type selector; (5) Species selector (*G. pallida* / *G. rostochiensis*); (6) Context-sensitive input (postcode search or metric selector); (7) Information button; (8) Save map button; (9) Scale bar; (10) Basemap attribution.*

4 Development of mapping components

PCN-Mapper incorporates a suite of interactive mapping tools to visualise both observed and derived information on PCN species distribution and intensity across Scotland. These mapping components transform the statutory testing data into intuitive, spatially explicit, easily interpretable visualisations that support exploration of patterns and trends at multiple geographic scales. Interactive scatter charts link directly to the underlying survey data, allowing users to inspect individual test results and metadata at each sampled location. Choropleth maps and bubble charts summarise PCN metrics at broader spatial scales: choropleth maps show aggregated indicators by postcode district, while bubble charts display postcode-sector level summaries at sector centroids. Complementary geostatistical analyses, including kernel density estimation and optimised hot spot analysis, highlight areas of concentrated infestation and spatial clustering.

Note that all mapped figures in this report show a representative sample of the statutory PCN testing data, which have been further anonymised by random spatial displacement of field locations by up to 20 km to protect grower confidentiality. These figures are provided for illustrative purposes only and do not represent actual PCN test locations.

4.1 Overview of the mapping approach

The mapping components transform large volumes of georeferenced soil test data into interpretable spatial intelligence. Each map type serves a distinct user need, such as inspection of raw data, spatial summarisation, or hotspot detection. The app unifies these visual layers within a single interactive environment where users can toggle between map types, species distributions, and time periods to explore patterns dynamically. Mapping outputs are

available at multiple spatial scales, from individual test results up to postcode areas, districts, and sectors. This multi-scale design supports both detailed local inspection and regional overview. All map types share the same spatial reference data and coordinate system, ensuring alignment and comparability. The maps are fully interactive, allowing panning, zooming, hover-based tooltips, and export options. The following sections describe each mapping component in more detail, beginning with the interactive scatter charts that link directly to individual soil test records.

4.2 Interactive scatter charts linking sample data to mapped distributions

The scatter chart displays individual soil test locations as interactive points, colour-coded by species and test outcome. Hovering over a point reveals detailed metadata from the original survey record, linking the map of species distribution directly to the underlying data (Fig. 2).

4.3 Geostatistical mapping methods

In addition to the raw-data scatter charts, PCN-Mapper includes several spatial analytical layers derived through geostatistical processing. These layers quantify infestation intensity, identify statistically significant clusters, and summarise PCN risk at regional scales. The following subsections summarise the key approaches used.

4.3.1 Kernel density estimation

Kernel Density Estimation (KDE) was used to create continuous maps of PCN intensity from individual positive test results (Fig. 3). The method works by placing a smooth, bell-shaped (normal) weighting function (the kernel) around each positive sample location and summing the influence of all points across space. This produces a continuous surface where higher values indicate a greater probability of PCN species presence, effectively mapping the likelihood of infestation across Scotland.

Each record was weighted by the total area of infested sampling units, ensuring that larger infestations contributed proportionally more to the estimated density. The degree of smoothing, known as the bandwidth, controls how local or general the resulting patterns appear. Bandwidths were calculated automatically using a two-dimensional version of Silverman's rule, as implemented in ArcGIS Pro (Esri UK Ltd), to provide an objective and consistent measure of spatial spread (Silverman, 2018). The resulting probability surfaces were normalised to a 0–1 scale for visualisation and comparison between species.

4.3.2 Optimised hot spot analysis

The Optimized Hot Spot Analysis (OHSA) is a spatial analysis technique that identifies statistically significant clusters of high or low values in geographic data. This analysis exactly followed the method documented in the ArcGIS Pro Spatial Statistics toolset, which uses the Getis-Ord G_i^* statistic (Ord & Getis, 1995) to determine whether apparent clusters of detections represent genuine spatial patterns or random occurrences. Because the analysis focuses on clusters rather than individual sample points, nearby detections were first grouped into weighted points to represent local concentrations of PCN. A statistically significant hot spot represents an area where many positive detections occur close together, forming a cluster that is unlikely to have arisen by random chance. Conversely, a statistically significant cold spot indicates a cluster of locations with unusually low numbers of positive PCN detections, where the scarcity of cases is unlikely to be due to random variation.

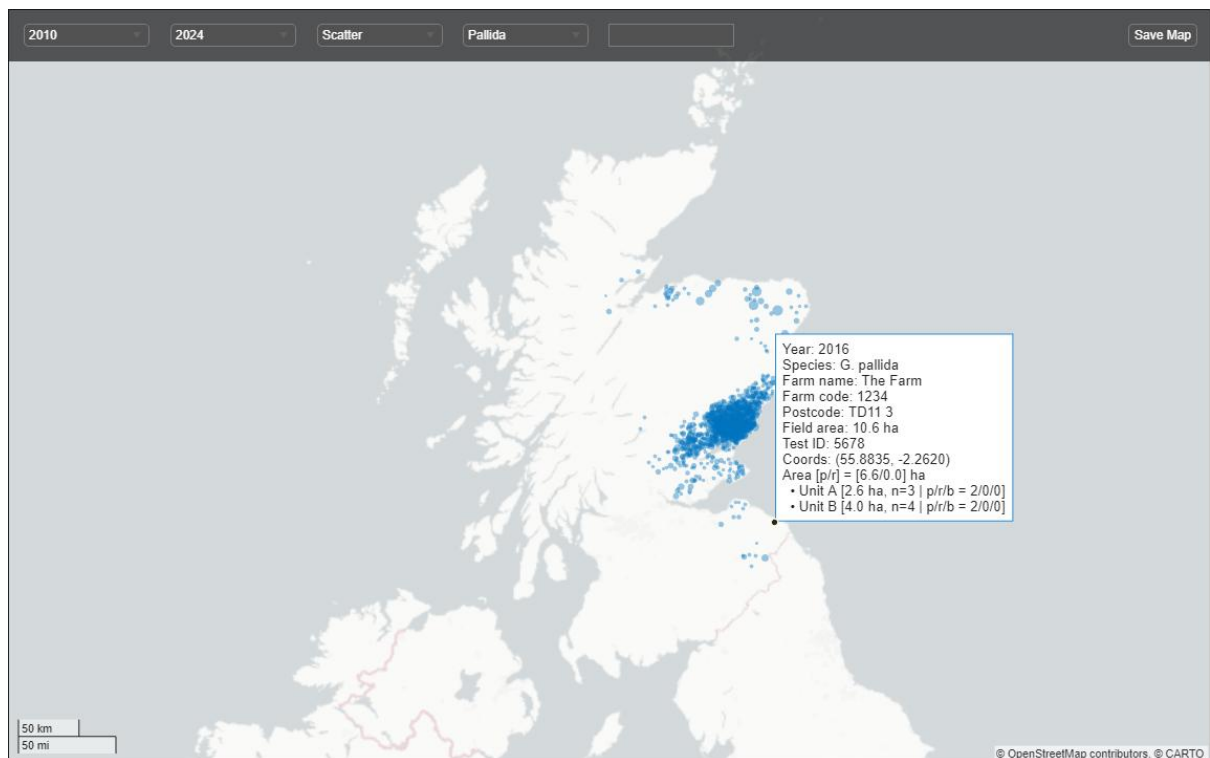


Figure 2 – Example interactive scatter chart showing statutory PCN test locations. Each point represents a test, colour-coded by species and scaled by the area of infested sampling units. Hovering reveals detailed test information, linking individual records to their spatial context. Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.

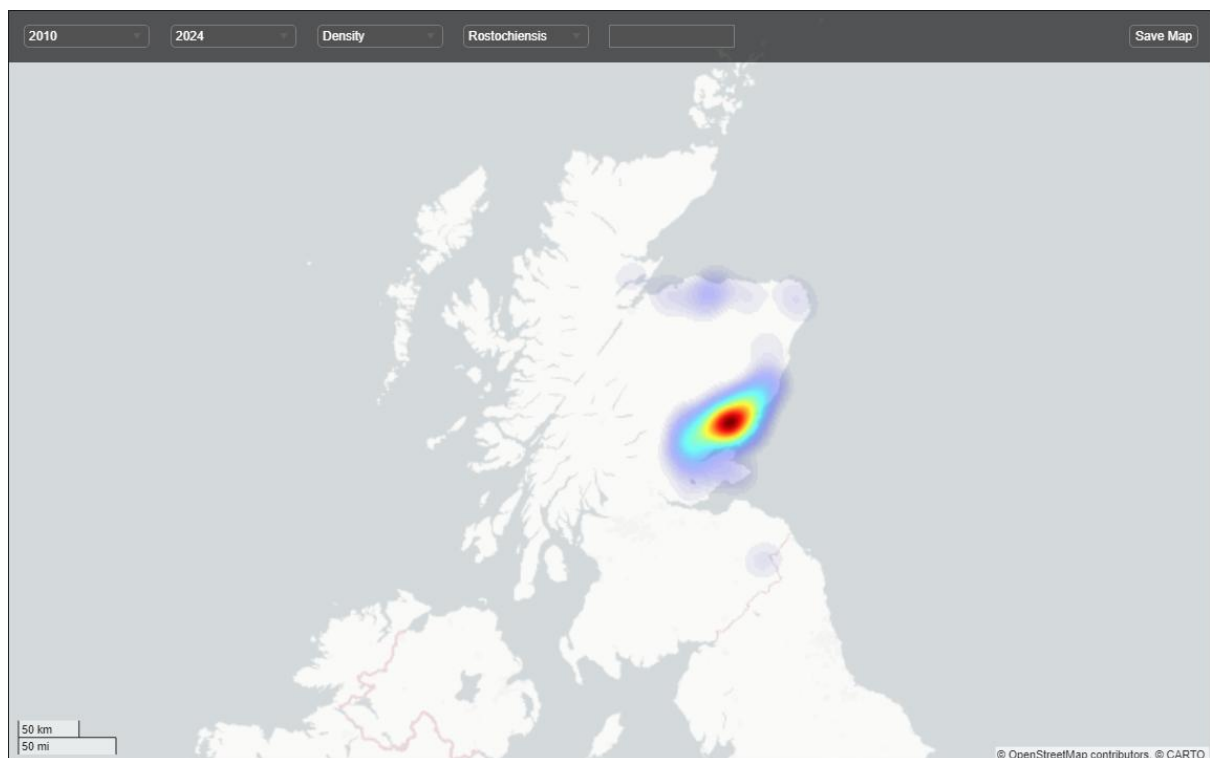
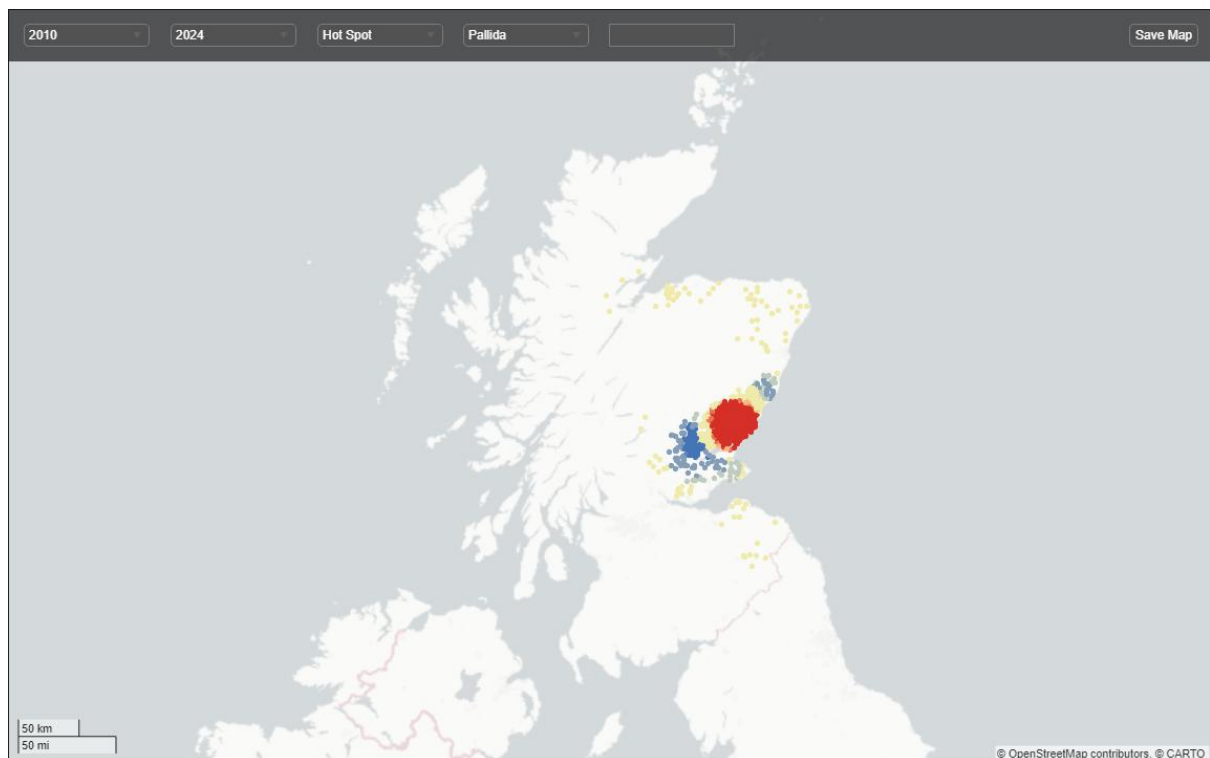


Figure 3 – Example kernel density estimation (KDE) surface illustrating the spatial intensity of *G. rostochiensis* across Scotland. Warmer colours indicate areas with a higher estimated probability of infestation. Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.



*Figure 4 – Example optimised hot spot analysis for *G. pallida* detections across Scotland. Red areas (hot spots) indicate zones of heightened risk where positive detections are concentrated; blue areas indicate zones of low activity with few or no detections. Darker shades represent increasing statistical confidence in the patterns. Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.*

4.3.2 Choropleth maps

Choropleth maps are a common form of thematic mapping in which variations in colour or shading represent differences in a variable across defined geographic areas. In PCN-Mapper, these maps display postcode districts shaded according to one of eight selectable species-specific metrics, allowing users to compare levels of infestation or risk across regions at a glance (Fig. 5). The colour scale automatically adjusts to the range of the selected metric, ensuring clear and consistent visualisation across different datasets. Districts with no positive detections are left unshaded, allowing infested areas to be highlighted more clearly.

4.3.3 Bubble charts

A bubble chart is a type of scatter plot that displays three dimensions of data using circular markers. The position of each bubble represents two variables on the x- and y-axes, while its size conveys a third. In PCN-Mapper, the bubbles are located at the centroid of postcode sectors, and bubble size can be varied according to one of eight selectable species-specific metrics (Fig. 6). Sectors with no positive detections are left blank, allowing areas of confirmed infestation to stand out clearly. This provides an intuitive way to compare relative PCN species intensity across regions at a finer spatial scale.

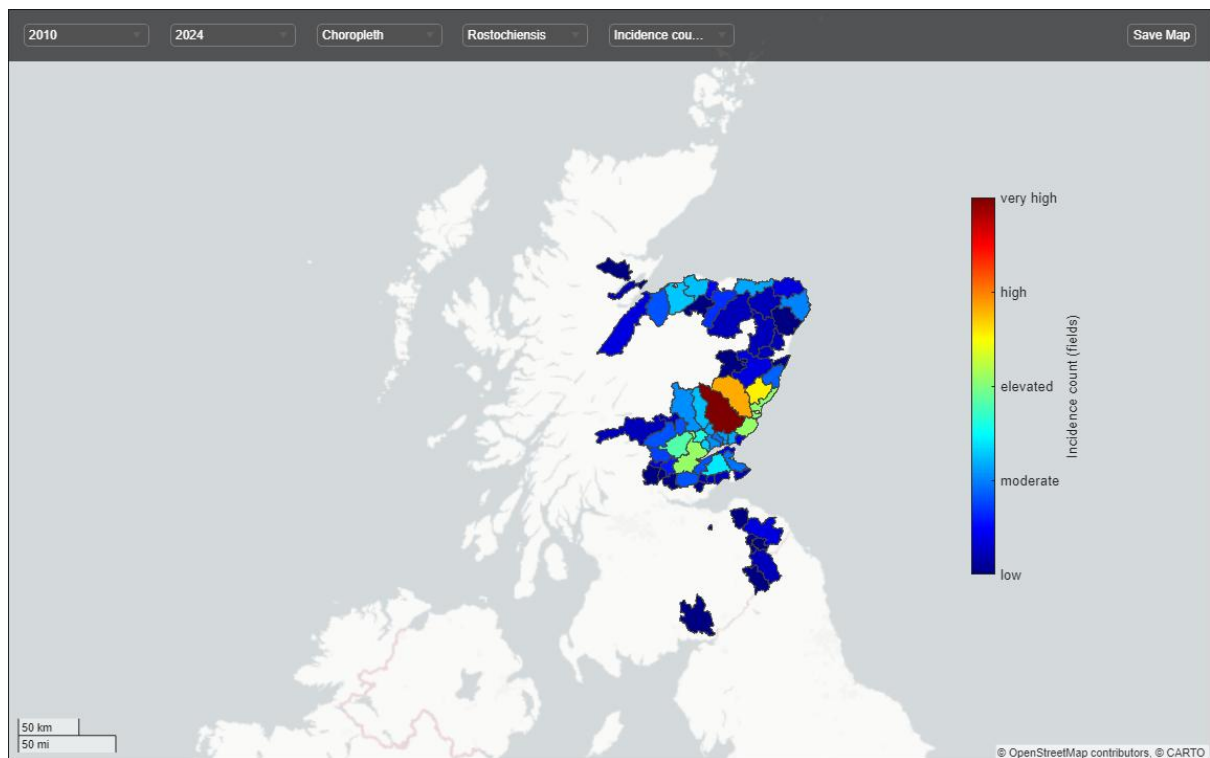


Figure 5 – Example of a choropleth map showing the total number of fields with positive detections of *G. rostochiensis* in each postcode district. Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.

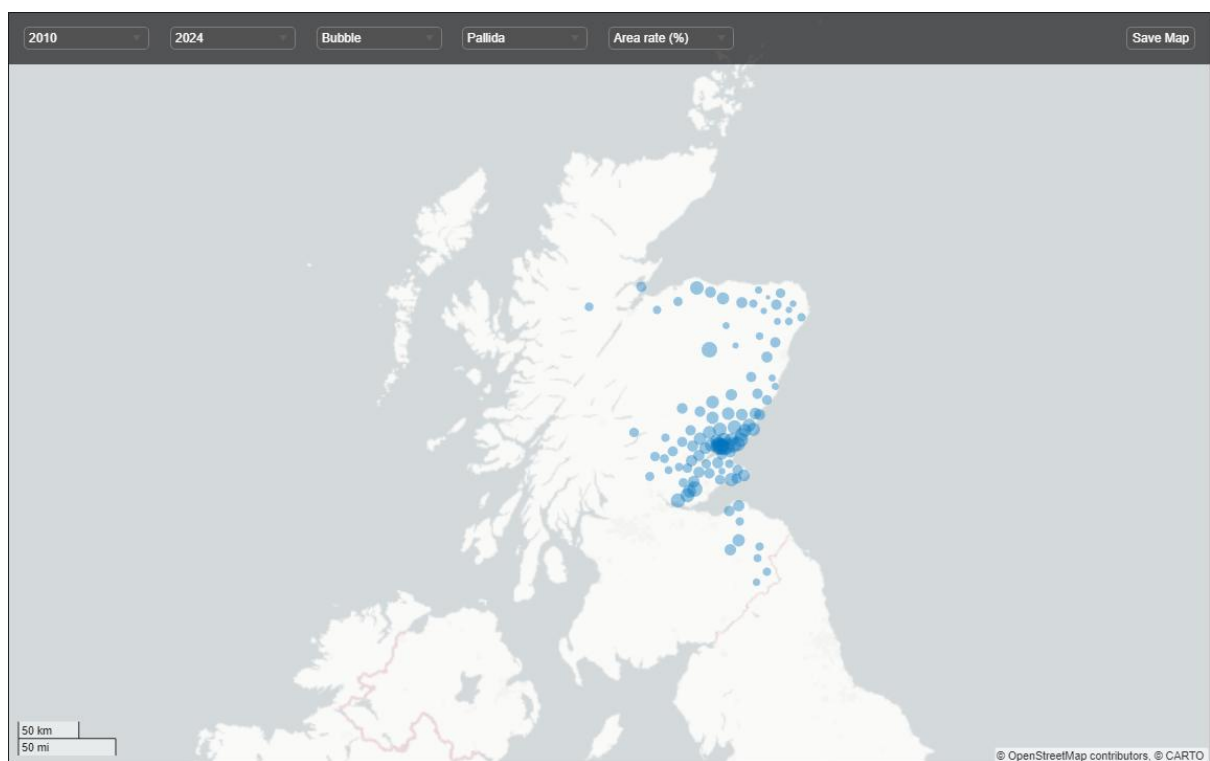


Figure 6 – Example of bubble chart map showing the percentage of sampled area that tested positive for *G. pallida*. Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.

5 Development of machine learning models for nowcasting

The goal of the machine learning component was to predict PCN species presence and relative risk in unsampled areas, complementing the geostatistical analyses based on observed distributions. This section outlines the process from data preparation to model deployment within the PCN-Mapper app, including the evaluation of multiple modelling approaches, algorithms, and validation procedures to ensure robust, generalisable predictions. An extended testing phase was carried out to determine which approaches performed best given the uneven distribution of positive and negative samples and the strong spatial patterns in the data. In this context, nowcasting refers to generating probabilistic estimates of the current spatial distribution of PCN species based on patterns learned from statutory testing data. Unlike forecasting, which projects future dynamics, nowcasting infers the most likely present distribution given historical and recent observations. Because statutory testing reports infestation only when detected through sampling, nowcasting can also provide a complementary probabilistic assessment of areas where infestation may exist at low or sub-detection densities, or where sampling has not occurred. The following subsections describe the data sources and pre-processing steps, the exploration of modelling approaches, the cross-validation and model selection framework, and the final model architecture, performance, and interpretation of the nowcast outputs.

5.1 Data sources and pre-processing

Machine-learning models were developed using the same statutory PCN testing dataset described in Section 3.1, after extensive cleaning and reformatting. Each record corresponded to a sampled field with confirmed spatial coordinates, year of sampling, and species-specific binary test outcomes (0 if all sampling units were negative, 1 if any sampling unit tested positive).

To standardise spatial sampling density and support prediction at regular intervals across Scotland, all cleaned observations were aggregated onto a hexagonal grid (250 km² per cell) using the centroid of each sampled field. Each observation therefore represents whether PCN was detected within a given grid cell during a particular sampling year; cells in which no PCN sampling occurred were excluded from the modelling dataset. The use of a regular hexagonal grid reduced spatial bias caused by uneven sampling, enabled reproducible spatial blocking for cross-validation, and provided a consistent spatial lattice for generating model predictions (nowcasts). The modelling framework therefore operates on grid cells containing at least one sampling record, meaning the dataset represents the spatial distribution of sampled locations rather than a complete national grid for every year. Model training is based on these sampled locations, while predictions are subsequently generated across the full grid covering Scotland.

For model training, the data were divided into two species-specific datasets: one for *G. pallida* and one for *G. rostochiensis*. Each dataset contained all negative samples together with those positive for the target species, including mixed samples where both species were detected. This structure allowed independent binary classifiers to be trained for each species, while preserving the full set of negative observations common to both.

After spatial aggregation, the grid-based dataset remained imbalanced between positive and negative training observations. Across all sampled grid cell-year observations, approximately 25% were positive for *G. pallida* and 75% negative, while for *G. rostochiensis* approximately 30% were positive and 70% negative. These proportions reflect the relative rarity of infested locations within the statutory testing dataset. Class weights were therefore incorporated during model training to prevent the algorithms from favouring the majority (negative) class and to ensure adequate sensitivity to positive detections.

An initial exploratory phase evaluated a wide range of potential predictor variables derived from spatial, temporal, and environmental data layers. These included: (i) spatiotemporal indicators describing proximity to previous detections and species distributions in surrounding years, (ii) bioclimatic variables calculated from UK Met. Office hourly weather observations, and (iii) soil and topographical descriptors from the National Soil Map of Scotland (<https://soils.environment.gov.scot/>). Principal component analysis (PCA) was tested as a dimensionality-reduction technique to minimise noise and redundancy among predictors.

Importantly, the epidemic signals captured by year and spatial coordinates alone were strong enough to support reliable predictive performance. The modelling was therefore simplified to use a compact, spatially explicit dataset defined only by year, X/Y coordinates (British National Grid), and class weights to correct for the class imbalance.

This processed dataset formed the common input for all subsequent modelling, cross-validation, and model-selection procedures described in Sections 5.2-5.6.

5.2 Exploration of modelling approaches

A wide range of machine-learning algorithms, resampling strategies, and model formulations were systematically evaluated to identify approaches capable of producing stable, generalisable, and spatially coherent predictions of PCN species presence. The aim was to develop models capable of discriminating *G. pallida* and *G. rostochiensis* occurrences across Scotland, accounting for severe class imbalance and overlapping spatial ranges. Model evaluation and selection were conducted within a nested cross-validation framework described in Section 5.3.

Early experiments used Error-Correcting Output Codes (ECOC) models to perform multi-class classification of the four possible states: *G. pallida* only, *G. rostochiensis* only, both species, and neither species. However, severe class imbalance and overlap in the spatial distributions of the two species limited the separability of classes in this formulation. The problem was therefore reframed as two independent binary classifiers, each operating on the species-specific datasets described in Section 5.1 and predicting the probability of presence for one species.

A set of representative algorithms spanning a range of modelling approaches was evaluated, including Naïve Bayes, discriminant classifiers, decision trees, *k*-nearest neighbours, logistic regression, support vector machines, neural networks, random forests, and boosted and bagged ensemble models with a wide range of base learners and ensemble aggregation methods. In addition, meta-learners were constructed by combining different subsets of these base models to test whether stacked or blended ensemble architectures could improve predictive performance and stability. Of these, decision trees, *k*-nearest neighbours, and ensemble methods consistently produced the most accurate and stable results under cross-validation. These three approaches were therefore implemented in the PCN-Mapper application and are the only algorithms available for developing new models within the tool and for generating the reported results.

When larger feature sets were tested, several feature-selection methods were evaluated to identify the most informative predictors and reduce overfitting. These included filter-based ranking approaches, redundancy-reduction techniques, and recursive elimination strategies, alongside embedded feature selection within regularised learners and ensemble models. Although these methods provided useful diagnostic insights, they produced only marginal or inconsistent improvements in predictive performance and were therefore not retained in the final modelling framework. This outcome suggests that the spatial and temporal structure

present in the statutory testing data already captures most of the predictive signal available for species occurrence.

To handle the approximately 3:1 negative-to-positive class imbalance, several resampling and weighting schemes were evaluated, including SMOTE-type oversampling, random undersampling, misclassification costs, and inverse-frequency class weighting. These approaches were implemented within the training folds of the cross-validation framework to avoid introducing information leakage between training and validation data. The latter proved most reliable and was adopted in all subsequent models.

Alternative probability thresholding strategies were also tested to optimise the conversion of predicted probabilities into binary presence/absence classifications. Instead of using a fixed 0.5 cutoff, thresholds were tuned during cross-validation to maximise balanced accuracy and better balance sensitivity and specificity across validation folds. However, this provided minimal improvement over the standard threshold and added complexity to model calibration, so the default 0.5 value was retained for the final deployment versions.

Hyperparameter tuning and model evaluation were conducted using Bayesian optimisation within a repeated nested cross-validation framework, providing an objective and reproducible basis for model selection.

5.3 Cross-validation strategy and model selection

A rigorous cross-validation (CV) was essential to assess model performance, particularly given the strong spatial dependence and class imbalance of the PCN data. A nested k -fold cross-validation (NCV) procedure was therefore used. In this framework, an outer loop estimates model performance on unseen data, while an inner loop performs hyperparameter tuning and model calibration. This structure limits overfitting, prevents information leakage, and provides unbiased estimates of predictive performance.

In k -fold CV, the dataset is divided into k subsets (folds). Each fold is used once as the test set while the remaining $k-1$ folds are used for model training. In the final modelling pipeline, blocked spatial k -fold CV was used, in which all samples within a single grid cell were assigned to the same fold (either training or testing). This ensured that spatial dependence among nearby observations did not lead to leakage of information between the training and testing sets, thereby providing more realistic estimates of model generalisability to new geographic locations.

The NCV procedure comprised two nested loops: an outer loop to assess predictive performance and an inner loop for hyperparameter tuning via Bayesian optimisation. Hyperparameters (configuration values affecting model complexity) were optimised by minimising the inner loop CV error (1-balanced accuracy). The best-performing settings were then used to retrain the model on the outer-loop training data before evaluation on the corresponding held-out test fold.

To reduce variance and improve the stability of performance estimates, the entire procedure was repeated (RNCV) using a $5 \times 5 \times 5$ repeated NCV design (five outer folds, each with five inner folds, repeated five times with different random partitions).

The performance metrics used to evaluate the models on the test data were: true negative rate (TNR, the proportion of negative instances correctly classified), false negative rate (FNR, the proportion of positive instances misclassified as negative), false positive rate (FPR, the proportion of negative instances misclassified as positive;), true positive rate (TPR, the proportion of positive instances correctly classified), positive predictive value (PPV, the proportion of predicted positives that are truly positive), negative predictive value (NPV, the proportion of predicted negatives that are truly negative), and balanced accuracy (the

arithmetic mean of the TPR and TNR). Additional performance metrics were also calculated, including the F1 score, Matthew's correlation coefficient (MCC), area under the receiver operating characteristic (AUROC), and overall accuracy, which are reported in the `model_summary.csv` output file described in Section 3.4 but not shown here for brevity.

5.4 Final model architecture and deployment versions

The results from the RNCV procedure were used to select the algorithms for *G. pallida* and *G. rostochiensis* that demonstrated the best ability to generalise to unseen data. All intermediate model versions generated during RNCV were then discarded, as they had served their purpose in model selection.

The final deployment versions were created using a single round of flat cross-validation on the full, species-specific datasets to optimise the hyperparameters of the selected algorithms. Each model was then retrained on all available data to maximise learning and provide the finalised classifiers used in PCN-Mapper. The trained models were saved as MATLAB.mat objects and automatically loaded by the application at startup, enabling rapid generation of spatial predictions and risk maps without the need for retraining.

Note: It is recommended that new models are trained each year as additional testing data become available. In addition, the models are designed to provide nowcasts of the current PCN situation, therefore they should not be used to generate retrospective predictions for previous years. Model retraining is initiated automatically within the app when new data are added to the required PCN testing dataset, ensuring that updated classifiers are derived and deployed without additional user input.

5.5 Model performance and results

The final model evaluation confirmed that both *G. pallida* and *G. rostochiensis* could be predicted with good accuracy using compact spatial-temporal feature sets. A classification ensemble provided the best performance for *G. pallida* (Table 1), while a *k*-nearest neighbour model performed best for *G. rostochiensis* (Table 2).

The tuned hyperparameter values for the final *G. pallida* ensemble were: `Method = AdaBoostM1`, `NumLearningCycles = 649`, `LearnRate = 0.267`, `MinLeafSize = 75`, `MaxNumSplits = 236`, `NumVariablesToSample = 3`, `PredictorSelection = curvature`. For *G. rostochiensis*, the optimal *k*-nearest neighbour configuration was: `NumNeighbors = 52`, `Distance = euclidean`, `DistanceWeight = inverse`.

Across all cross-validation folds and repeats, both models achieved high negative predictive values, indicating strong reliability in identifying PCN-free fields and hence good suitability for use in an operational setting for surveillance planning and risk-based screening. Balanced accuracy values were good and false-negative rates low, demonstrating consistent sensitivity to infested fields while maintaining a low rate of false alarms.

Overall, the models achieved a robust balance between sensitivity and specificity, providing dependable discrimination of infested and non-infested areas across Scotland. The resulting prediction probabilities underpin the nowcast layers within the PCN-Mapper application, where each value represents the modelled likelihood of PCN presence at the centroid of a grid cell. The interpretation and practical use of these nowcast outputs are described in Section 5.6.

Table 1 – Performance of models for classifying *Globodera pallida* presence/absence (micro-averaged across all test folds and repeats)

Models*	TNR†	FNR	FPR	TPR	PPV	NPV	Bal. Acc.
KNN	0.77 (0.15)	0.10 (0.13)	0.23 (0.15)	0.90 (0.13)	0.62 (0.19)	0.95 (0.07)	0.84 (0.07)
DT	0.88 (0.11)	0.26 (0.23)	0.12 (0.11)	0.74 (0.22)	0.72 (0.22)	0.90 (0.10)	0.81 (0.12)
ENS	0.86 (0.13)	0.20 (0.21)	0.14 (0.13)	0.80 (0.21)	0.71 (0.21)	0.92 (0.09)	0.84 (0.11)

* The machine learning classifiers are KNN = *k*-nearest neighbour, DT = decision tree, ENS = classification ensemble.

† The metrics are TNR = true negative rate, FNR = false negative rate, FPR = false positive rate, TPR = true positive rate, PPV = positive predictive value, NPV = negative predictive value, Bal. Acc. = balanced accuracy. Standard deviations across folds are given in parentheses.

Table 2 – Performance of models for classifying *Globodera rostochiensis* presence/absence (micro-averaged across all test folds and repeats)

Models*	TNR†	FNR	FPR	TPR	PPV	NPV	Bal. Acc.
KNN	0.70 (0.13)	0.18 (0.20)	0.30 (0.13)	0.82 (0.20)	0.58 (0.16)	0.88 (0.14)	0.76 (0.14)
DT	0.77 (0.14)	0.28 (0.25)	0.23 (0.14)	0.72 (0.25)	0.63 (0.19)	0.85 (0.12)	0.74 (0.12)
ENS	0.81 (0.13)	0.37 (0.28)	0.19 (0.13)	0.63 (0.28)	0.65 (0.24)	0.81 (0.14)	0.72 (0.15)

* The machine learning classifiers are KNN = *k*-nearest neighbour, DT = decision tree, ENS = classification ensemble.

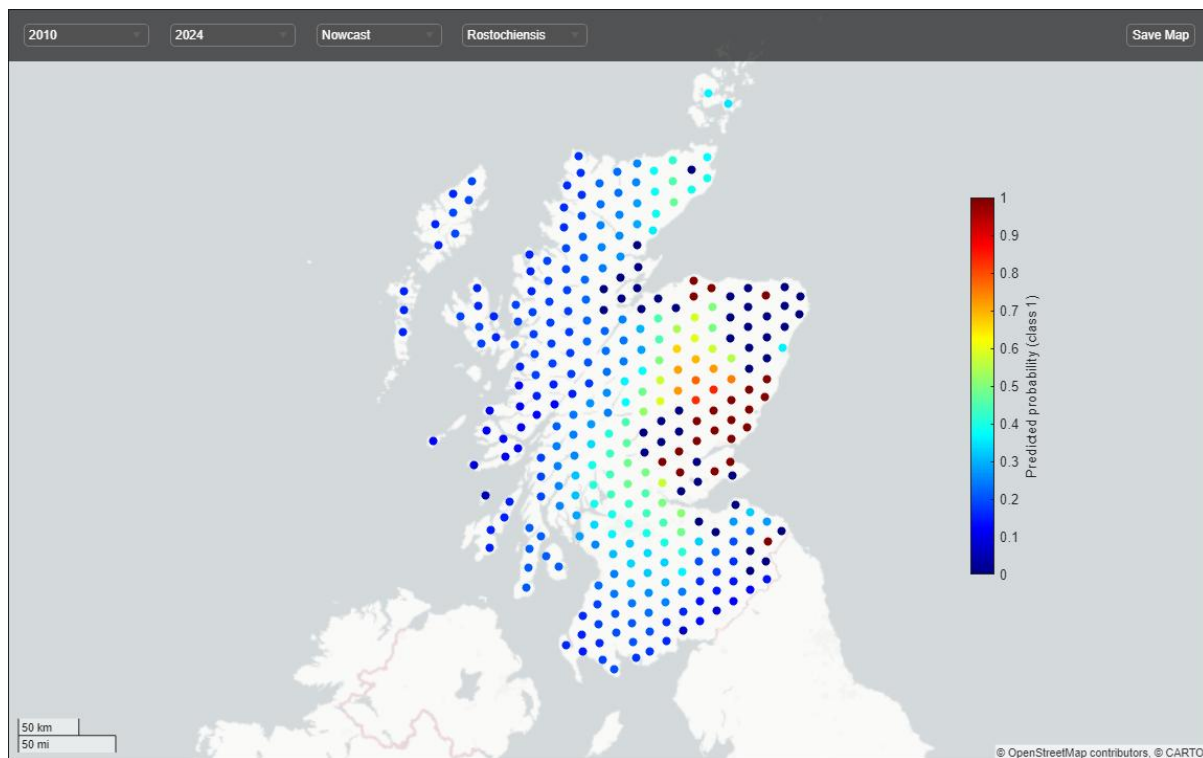
† The metrics are TNR = true negative rate, FNR = false negative rate, FPR = false positive rate, TPR = true positive rate, PPV = positive predictive value, NPV = negative predictive value, Bal. Acc. = balanced accuracy. Standard deviations across folds are given in parentheses.

5.6 Interpretation of nowcast outputs

The nowcast outputs represent the predicted probability of *G. pallida* or *G. rostochiensis* presence at the centroid of each grid cell (Fig. 7). Each value expresses the modelled likelihood of species presence at that location, inferred from spatial and temporal structure in the statutory dataset. Higher values therefore correspond to areas where infestation is more probable, while lower values indicate reduced, but not zero, risk.

The nowcast maps display continuous probabilities rather than binary classifications. These surfaces should therefore be interpreted as risk gradients rather than hard presence–absence boundaries. Higher probability values indicate a stronger modelled signal of infestation, while lower values correspond to areas where detections are less likely but not impossible. This probabilistic formulation allows gradual transitions between regions of differing risk and avoids the artificial discontinuities that thresholded (binary presence-absence) maps can introduce. Predicted binary labels are provided separately in the automatically generated `model_predictions.csv` output file described in Section 3.4.

The nowcasts are intended to provide a spatially explicit, data-driven representation of current PCN species risk rather than a forecast of future spread. They complement observed-data visualisations (such as KDE and hot-spot analyses) by identifying areas where the underlying probability of infestation is higher or lower given learned spatiotemporal structure. As such, they may assist in highlighting areas where infestation could be present but remain below routine detection thresholds.



*Figure 7 – Example nowcast map for *Globodera rostochiensis* using the finalised *k*-nearest neighbour model. Points represent grid cell centroids and are colour-coded from cool colours (low probability of infestation) to warm colours (high probability of infestation). Note: this figure is a representative, anonymised illustration and does not reflect actual PCN infestation patterns.*

It should be noted that the accuracy of the nowcasts is constrained by the distribution and representativeness of historical sampling. Probabilities in sparsely tested regions should therefore be interpreted with caution and used to inform surveillance and management prioritisation, rather than as definitive statements of pest absence or presence.

Collectively, the geostatistical analyses and machine-learning nowcasts provide complementary perspectives on the spatial distribution of PCN species across Scotland. The former describes observed clustering and infestation intensity based on confirmed detections, while the latter extends these insights to unsampled areas through probabilistic inference. Together, they establish a quantitative evidence base for the discussion that follows.

6 Discussion

This project developed PCN-Mapper, a data-driven decision-support tool designed to visualise PCN distributions across Scotland and generate spatial risk indicators based on SASA's statutory PCN testing data. By combining traditional geostatistical mapping with machine-learning nowcasting, PCN-Mapper provides clear visualisations of observed species-specific PCN distributions and predicts the probability of each species occurring in untested areas.

The scatter charts provide a clear representation of species distributions and, through the use of tooltips, a direct link to the underlying sample data, enabling a simultaneous visual overview and granular inspection of individual test records. Kernel density estimation and optimised hotspot analyses offer complementary perspectives on spatial risk, while postcode-district level choropleth and postcode-sector level bubble charts enable consistent and interpretable reporting at operationally relevant geographic units.

Machine-learning nowcasts extend this capability by generating species-specific probability surfaces across Scotland. Although based on a deliberately minimal feature set (year and spatial coordinates), the models produced stable and spatially coherent probability estimates. Their ability to consistently identify areas of lower modelled risk supports proactive disease-free certification and resource planning, while areas assigned higher probabilities may provide an early indication of elevated underlying risk and warrant prioritised surveillance and management attention.

SASA's statutory PCN testing is the internationally accepted method to confirm PCN presence and absence in Scotland, and it therefore provides a high-quality source of data to underpin all the mapping and modelling outputs in this project. However, as testing is targeted to support certification, it does not represent a uniform survey of all agricultural land. While geostatistical smoothing and spatially constrained model validation helps reduce bias from uneven sampling, this does not fully remove the influence of the underlying regulatory testing pattern. Thus, the outputs should be interpreted as risk surfaces shaped by historical sampling patterns rather than definitive infestation maps. They provide strategic guidance for targeting surveillance and may support more automated decision-making as datasets expand and model confidence increases.

Key strengths of PCN-Mapper include:

- Integration of statutory PCN testing data with modern spatial analytics.
- Combination of observed patterns and predicted risk surfaces.
- Robust evaluation and validation of machine learning models using spatially blocked repeated nested cross-validation.
- A user-friendly interface that uses tooltips to guide and inform the user.
- A large suite of charts and metrics allows users to explore PCN species distributions at multiple spatial resolutions (from individual test points to postcode-sector summaries), supporting both granular inspection and strategic regional assessment.
- Automatic generation and saving of a wide range of summary statistics at postcode area, district, and sector levels.
- Automatic generation and saving of model predictions.
- Reproducible analytical workflows with automatic generation and saving of model outputs and summary statistics.
- Automatic retraining of new machine learning models when new data are added to the required PCN testing dataset.

Limitations of PCN-Mapper include:

- Dependence on the availability and geographic distribution of statutory testing.
- Lack of PCN data from ware land.
- Lack of independent surveillance datasets for external validation of predictions.
- Statutory PCN testing reflects regulatory sampling priorities rather than a random survey of agricultural land, which may introduce spatial sampling bias.
- Predictive performance constrained by limited availability of sub-field scale predictors and detailed field management datasets (e.g., crop rotations and varietal resistance information).
- Nowcasts representing current estimated distribution rather than future forecast.

PCN-Mapper provides a unified, operationally focused system for visualising and assessing PCN species risk. While the system is constrained by the availability and structure of existing surveillance data, it demonstrates how statutory testing records can be transformed into actionable spatial intelligence to support targeted surveillance, certification, and management planning. As additional field data become available, the system can be retrained to maintain accuracy and incorporate emerging patterns.

Future development could focus on strengthening operational use and regulatory value. Priorities may include:

- Annual retraining on new statutory testing data, allowing the models to evolve with changing species distributions.
- Incorporation of sub-field (sampling-unit) scale data, where available, to further improve the machine learning models.
- Inclusion of land-use mapping and crop-rotation information to better distinguish arable and non-arable landscapes and improve interpretation of model results.
- Integration of field and sub-field management data (e.g., crop varieties, resistance deployment, and evolving field boundaries) to better capture within-field variation in PCN risk where such data can be integrated.
- Refinement of uncertainty communication, for example by displaying model confidence ranges alongside predicted risk.
- Development of spatial uncertainty surfaces to accompany model predictions, allowing users to identify regions where predictions are less certain and where additional surveillance would be informative.
- Operational automation, where new data are imported and models retrained seamlessly within the regulatory workflow.

7 Conclusions

- **PCN-Mapper integrates** statutory PCN testing data, geostatistical analysis, and machine-learning nowcasting into a single operational decision-support platform. By transforming routine sampling results into clear spatial intelligence and risk metrics, it provides clear evidence for understanding both observed and modelled PCN distributions across Scotland.
- **PCN-Mapper demonstrates** strong analytical performance and practical utility. Species-specific models generate stable probabilistic estimates of infestation risk, while interactive visualisations allow users to explore spatial patterns from individual sampling points to regional summaries. Together, these capabilities support surveillance planning, stakeholder engagement, and species-specific advisory work.
- **PCN-Mapper establishes** a scalable framework for data-driven PCN surveillance and management. As additional testing data become available, regular model retraining will allow the system to evolve with changing species distributions, strengthening its value for strategic decision-making and contributing to long-term efforts to manage PCN risk and safeguard Scotland's seed potato sector.

8 References

- Iqbal, S. & Jones, M. (2017) Nematodes. In: Thomas B., Murray B. & Murphy D. (Eds.) *Encyclopedia of Applied Plant Sciences*. Oxford, UK: Elsevier, pp. 113–119.
- Mesa-Valle, C.M., Garrido-Cardenas, J.A., Cebrian-Carmona, J., Talavera, M. & Manzano-Agugliaro, F. (2020) Global research on plant nematodes. *Agronomy* **10**, 1148.
- Ord, J.K. & Getis, A. (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis* **27**, 286–306.
- Pope, A. (2017) GB Postcode Area, Sector, District, [Dataset]. In: Edinburgh. U.O., ed. University of Edinburgh.
- Silverman, B.W. (2018) *Density estimation for statistics and data analysis*. Routledge.
- Turner, S.J. & Subbotin, S.A. (2006) Cyst nematodes. In: Perry R. & Moens M. (Eds.) *Plant nematology*. Wallingford, CT, USA: CABI, pp. 109–143.

Appendix A — PCN-Mapper user guide

Contents

A.1	Introduction	23
A.2	Copyright and license notice	23
A.3	System requirements.....	23
A.4	Quick start guide	24
A.5	Installing the PCN-Mapper application	25
A.6	Installing the assets package	25
A.7	First-run checks and expected files.....	25
A.8	PCN-Map per interface (user guide view)	27
A.9	Updating the PCN dataset.....	28
A.10	Maintenance and future updates	30

A.1 ***Error! Not a valid bookmark self-reference.*** Introduction

PCN-Mapper is designed as an applied decision-support tool for exploring and communicating the distribution and risk of potato cyst nematodes across Scotland. The application is intended for use by SASA staff as the national custodians of statutory PCN testing data, ensuring secure handling of sensitive information while enabling rapid analytical updates and visualisation. This section provides practical guidance on system requirements, data loading, map interaction, and ongoing maintenance to support routine operational use.

A.2 *Copyright and license notice*

PCN-Mapper © 2025 Dr Peter Skelsey & The James Hutton Institute.
All rights reserved.

PCN-Mapper, including all MATLAB® source code, analytical components, trained machine-learning models, processing logic, assets, and compiled applications, remains the intellectual property of the author and The James Hutton Institute.

Warranty and Liability

This software is provided “**as is**” without warranty of any kind.

The author and The James Hutton Institute accept no liability for any loss arising from its use.

A full licence document (“license.txt”) is included in the `assets` folder and forms part of this User Guide.

A.3 *System requirements*

This application is compiled for Windows 10/11 (64-bit) and is not compatible with Windows 7/8/8.1 or 32-bit systems. MATLAB standalone applications are not cross-platform, so this installer will not run on macOS or Linux. Platform-specific builds and MATLAB Runtime versions are required for non-Windows systems.

PCN-Mapper is computationally intensive and runs best on a reasonably powerful Windows machine; low-spec CPUs may result in slow performance.

A.4 Quick start guide

Follow these steps to install and launch PCN-Mapper.

Each step is described briefly here for convenience; the following sections (8.4 – 8.9) provide more detailed instructions and troubleshooting information.

1. Install the application

- Double-click `PCN-Mapper_v1.0.0_Installer.exe`.
- Allow the installer to download and install the required MATLAB Runtime (R2025b)

2. Install the assets package

- Extract `assets.zip` into a location where you have full read and write access
- Confirm the folder structure includes:
`assets/` (with `data/`, `models/`, `shapefiles/`), `LICENSE.txt` and `PCN-Mapper User Guide.pdf`

3. Launch the app

- Double-click `PCN_Mapper.exe`
- When PCN-Mapper is launched it will prompt you to locate the `assets` folder. Browse to the folder you extracted and select it. The app will repeat this process every time it is launched to ensure the correct `assets` folder is always utilised.
- The app will initialise and automatically generate a `results` folder alongside the `assets` folder where it will store various results files.
- This process takes ~15–30 seconds

4. Explore PCN patterns

- Select species (**G. pallida** or **G. rostochiensis**)
- Choose a visualisation (e.g., scatter, density, hotspots, choropleth, bubble, nowcast)
- Use the time-window selectors to explore trends (e.g., 2010–2024)

5. Save maps and outputs (optional)

- Click **Save** to export figures for reports and briefings

6. Updating the dataset

- Add new PCN records to `assets/data/pcn_data.xlsx` **before launching the app**
- If new data are detected, PCN-Mapper will ask whether to rebuild the models
- Select **Rebuild** after a full year of new results to keep predictions up-to-date

Tip: Do not rename or move any folders or files. If something appears missing or altered, the app will pause and show instructions to fix the issue.

A.5 Installing the PCN-Mapper application

To install the app, run the installation file `PCN-Mapper_v1.0.0_Installer.exe`. This is a web-based installer so you will need an internet connection to install the app. When you run the installer, it will automatically download a free piece of software called MATLAB Runtime to your computer. This is required to run the app, and it will install the same version of Runtime that the app was created with (version R2025b). You will require around 2GB of space for Runtime. If you already have the correct version of MATLAB Runtime then this step will be skipped. Runtime will be installed in `C:\Program Files\MATLAB\MATLAB Runtime`. You will not have to do anything with these files once installed.

During installation, you may choose where to install the application. The app itself does not require write access to its installation folder. However, you must have full read and write access to the location of the `assets` and `results` folders, as these are used to load data and store outputs during use.

A.6 Installing the assets package

The `assets` package is supplied separately (rather than bundled inside the installer) to protect the underlying PCN dataset and ensure it is not inadvertently redistributed outside authorised users.

Once installation is complete, extract the `assets.zip` file into a location where **you have full read and write access**. This creates the `assets` folder and the subfolders `data`, `models`, and `shapefiles`. These contain all the files required to build the datasets and run the app (Fig. A1). This includes two fully trained and finalised machine-learning models (one for *G. rostochiensis* and one for *G. pallida*), developed using data from 2010–2024 (Section 5) and stored in the `models` folder. The `assets` folder also contains the software license in `LICENSE.txt` and a .pdf copy of this user guide.

To ensure the app runs correctly, please leave all folder names and files unchanged, except for the PCN dataset located in `assets\data\pcn_data.xlsx`, which you may update to include new records (see Section A.9).

A.7 First-run checks and expected files

Once the application and assets package have been installed, launch PCN-Mapper to confirm that it is operating correctly. When opened, the app will prompt the user to locate the `assets` folder. Browse to the folder you extracted and select it. The app will repeat this process every time it is launched to ensure the correct `assets` folder is always utilised. Next, the app calculates all required internal datasets, geostatistical layers, regional summary statistics, and model predictions. A progress dialogue will display during this stage, and the process typically completes within 15–30 seconds. When initialisation is complete, the interface becomes fully interactive.

On first use (after initialisation), the application automatically creates a `results` folder in the same location as the `assets` folder. The app also generates three results files within that folder at start-up. These files are refreshed every time the app is opened so that users always begin with clean, up-to-date baseline outputs, preventing accidental reuse of older results and ensuring a consistent reporting workflow.

Verify that the `assets` directory structure matches the layout shown in Fig. A1. The folders and files are described in Table A1. If any required folders or files are renamed, moved, or in use by another application, the app will safely stop and display a pop-up alert explaining what is wrong and how to fix it. The issue must be addressed and the app restarted before

continuing.

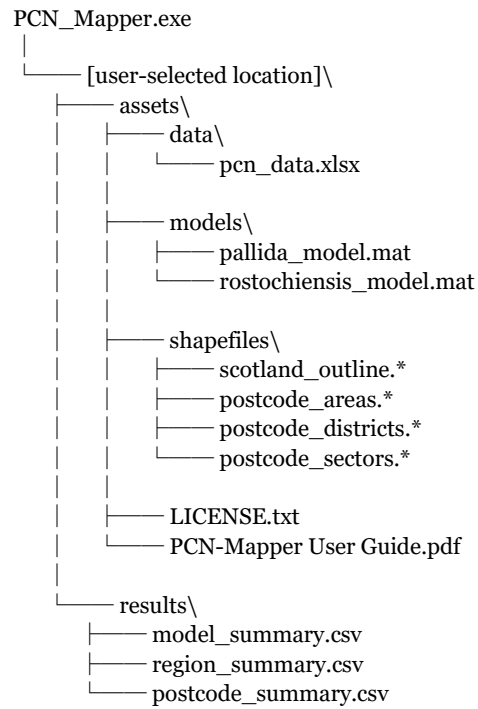


Figure A1 – Expected folder and file structure for PCN-Mapper after installation & first successful run.

Table A1 – PCN-Mapper directory contents and purpose

Item	Description
PCN_Mapper.exe	The PCN-Mapper application
assets (user-selected location)	Folder for the required datasets, models, & spatial files
data/pcn_data.xlsx	Confidential PCN records – user-updateable
models/pall_ml.mat	<i>G. pallida</i> (finalised) model file
models/rost_ml.mat	<i>G. rostochiensis</i> (finalised) model file
shapefiles/*	Postcode geography and Scotland grid files
LICENSE.txt	Software license terms
PCN_Mapper_User_Guide.pdf	Full user documentation
results (auto-generated)	Folder for app-generated outputs
model_predictions.csv	Predictions from the two nowcast models
model_summary.csv	Cross-validation performance metrics for both nowcast models
regional_results.csv	Regional summaries and risk outputs

A.8 PCN-Map per interface (user guide view)

The layout and controls are shown in Fig. A2. This section provides a concise overview for users; a comprehensive description of the interface and its underlying functionality is available in the main project report.



Figure A2 – Screenshot of PCN-Mapper interface.

Interface elements and controls:

1. **Map display.** Interactive map showing PCN testing results and model outputs across Scotland. Scroll to zoom, click-and-drag to pan.
2. **Start year selector.** Choose the first year of the time window for displaying PCN results (e.g., 2010–2024).
3. **End year selector.** Choose the final year of the time window. Must be equal to or later than the start year.
4. **Display mode.** Select the map display type: Scatter, Density, Hot Spot, Choropleth, Bubble, Nowcast.
5. **Species selector.** Select to display results for *G. pallida* or *G. rostochiensis*
6. **Context input.** Context-sensitive field that acts either as a postcode search (for highlighting and summarising areas) or as a metric selector when specific plot types are active.
7. **Information button.** Opens app details and licensing information.
8. **Save Map.** Export the current map to a file type and resolution suitable for reporting and presentations.
9. **Scale bar.** Shows scale in kilometres and miles for map interpretation.
10. **Attribution.** Basemap credits required by OpenStreetMap and CARTO licence terms.

Note that all interactive controls include tooltips (small pop-up text boxes that appear when the cursor hovers) explaining their function, and all visual elements, such as maps, also display contextual tooltips that reveal underlying data values when hovered over. Informative alerts (pop-up messages) guide the user if an operation fails, explaining both the issue and the corrective action.

A.9 Updating the PCN dataset

PCN-Mapper uses a structured Excel file (`pcn_data.xlsx`) to manage PCN testing records. This file contains two worksheets:

- **Soil Tests 2010-Present** — one row per laboratory test (includes both negative and positive field-level results)
- **PCN Positive Units** — one row per positive sampling unit at the sub-field scale (contains only positive units)

To update the dataset when new results become available:

1. Create a dated archive copy of `pcn_data.xlsx` before applying updates.
2. **Before launching the app**, open `assets/data/pcn_data.xlsx`.
3. In **Soil Tests 2010-Present**, append new field-scale records below existing entries.
4. In **PCN Positive Units**, append new rows **only** for positive sampling units.
5. Ensure all required columns are filled & **do not change** column names or order.
6. Save and close the file.
7. **Launch (or relaunch) PCN-Mapper** so the new data are fully incorporated into maps, summaries, and model outputs

Important: Every row added to *PCN Positive Units* must correspond to a valid test in *Soil Tests 2010-Present*. Do not delete, rename, or reorder columns.

Recommended update frequency

For PCN-Mapper to operate reliably, the dataset must only be updated when a full testing year has been completed. Do not add partial-year data. The predictive models require complete

annual data to ensure statistical validity and prevent bias from incomplete reporting periods. Only update the dataset once the entire year's results have been finalised and retrain the models at that point.

Model rebuild prompt

When PCN-Mapper detects that the dataset contains new records not included in the most recent nowcast models (Fig. A3), it will prompt you to either:

- **Rebuild** — regenerate the machine-learning models using the updated dataset
- **Keep old** — retain the existing models

Recommendation: Choose Rebuild when a full year of new results has been added to ensure the models remain up-to-date and accurate.

Rebuilding the models can take 20–60 minutes, depending on your machine's processing speed. During this time, the app will become temporarily disabled and a progress window will show the current training stage, model, and model performance metrics (Fig. A4). The app will become fully functional again once model training is completed, and results files will be automatically generated and saved as described in Section 8.5.

Data integrity checks

If changes are made to worksheet names, column names, or layout, the app will stop and show an alert explaining how to fix the issue. Correct the file and relaunch the app.

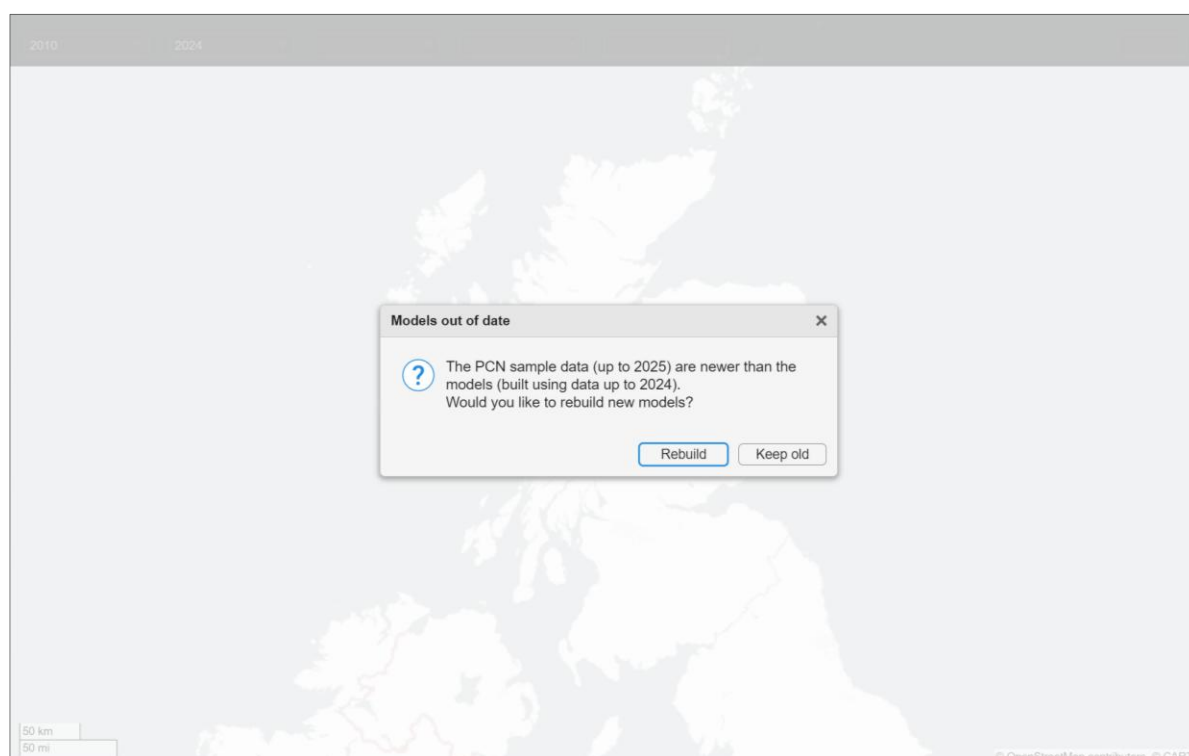


Figure A3 – Screenshot of PCN-Mapper detection of new data during initialisation.

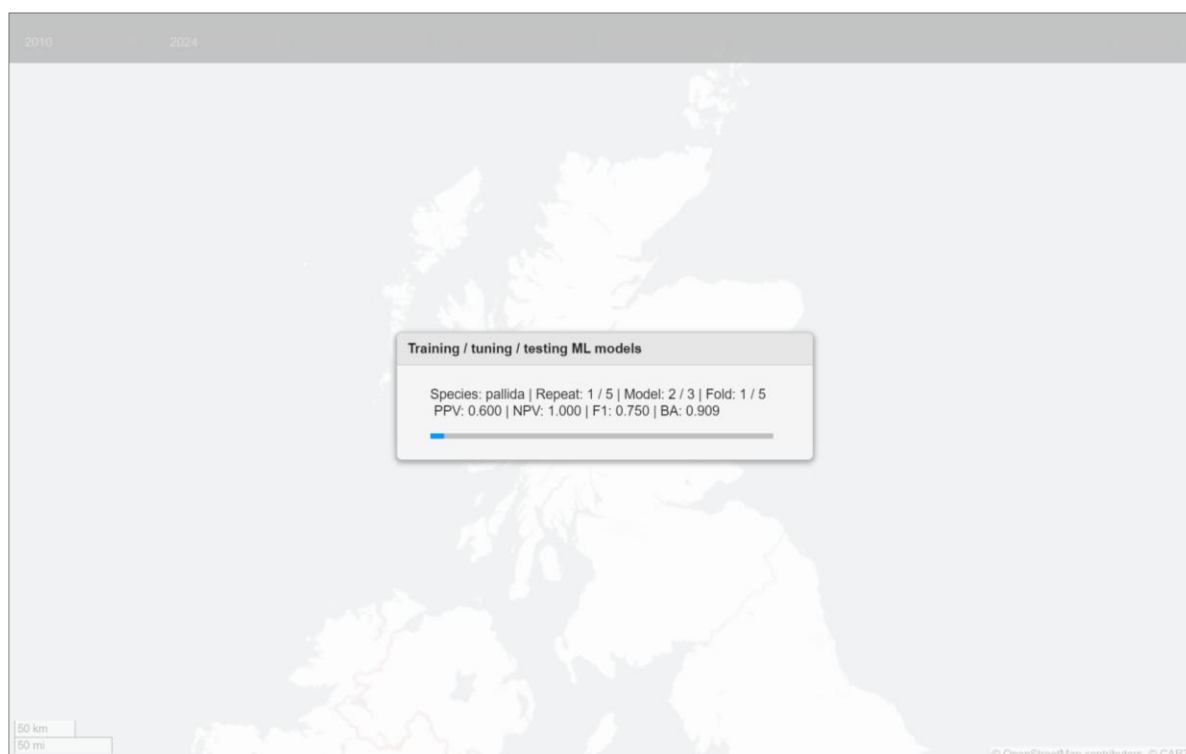


Figure A4 – Screenshot of PCN-Mapper rebuilding new up-to-date models. Progress is shown for each species, repeat run, model, and cross-validation fold. Key metrics are displayed during training: PPV = positive predictive value, NPV = negative predictive value, F1 = balance between precision and recall, BA = balanced accuracy, which is the overall accuracy adjusted for class imbalance.

A.10 Maintenance and future updates

PCN-Mapper has been designed to operate with minimal user maintenance. Routine care primarily involves keeping the input dataset up to date and rebuilding the predictive models when new annual results become available (see Section A.9). The following points outline recommended practice for long-term operation:

- **Software updates**

Future updates to the application may be issued to improve functionality, incorporate new data layers, or reflect changes to analytical methodology. When a new version is released, uninstall the previous version and install the latest PCN-Mapper package and assets folder.

- **MATLAB Runtime compatibility**

PCN-Mapper requires the specific MATLAB Runtime version installed during setup. Additional MATLAB Runtime versions may be present on your system (for other MATLAB-based applications), and this will not affect PCN-Mapper. However, if the required Runtime version is removed, the app will prompt reinstallation.

- **Data stewardship**

The file `pcn_data.xlsx` remains the authoritative local data source. Users should maintain secure storage and version control for this file, retaining dated archive copies before updates. This supports traceability and audit requirements.

- **Model refresh cycle**

For best performance, rebuild the nowcast models after updating the dataset with a **complete year of new results**. Rebuilding with partial-year data is not recommended, as it may reduce model stability and predictive accuracy.

- **System housekeeping**

The app automatically recreates required folders and baseline output files at start-up. Users should avoid renaming or modifying internal folders (assets, results, models, shapefiles, and data) to ensure correct operation. If files are altered, missing, or open in other applications, the app will provide guidance and request a restart.

- **Support and feedback**

If unexpected errors occur or enhancements are required, please contact the system maintainer. Feedback on usability, performance, and mapping outputs will inform future development and refinement of the tool.

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

LinkedIn: <https://uk.linkedin.com/company/plant-health-centre>



Royal
Botanic Garden
Edinburgh



RESAS
Rural & Environmental Science
and Analytical Services

