


Article

Spatio-Temporal Prediction of the Epidemic Spread of Dangerous Pathogens Using Machine Learning Methods

Wolfgang B. Hamer ^{1,*} , Tim Birr ², Joseph-Alexander Verreet ², Rainer Duttmann ¹ 
and Holger Klink ²

¹ Department of Geography, Physical Geography, Christian-Albrechts-Universität zu Kiel, Ludewig-Meyn-Str. 14, 24118 Kiel, Germany; duttmann@geographie.uni-kiel.de

² Department of Plant Diseases and Plant Protection, Institute of Phytopathology, Christian-Albrechts-Universität zu Kiel, Hermann-Rodewald-Str. 9, 24118 Kiel, Germany; t.birr@phytomed.uni-kiel.de (T.B.); javerreet@phytomed.uni-kiel.de (J.-A.V.); hklink@phytomed.uni-kiel.de (H.K.)

* Correspondence: hamer@geographie.uni-kiel.de; Tel.: +49-431-880-2955

Received: 8 November 2019; Accepted: 13 January 2020; Published: 15 January 2020



Abstract: Real-time identification of the occurrence of dangerous pathogens is of crucial importance for the rapid execution of countermeasures. For this purpose, spatial and temporal predictions of the spread of such pathogens are indispensable. The R package *papros* developed by the authors offers an environment in which both spatial and temporal predictions can be made, based on local data using various deterministic, geostatistical regionalisation, and machine learning methods. The approach is presented using the example of a crops infection by fungal pathogens, which can substantially reduce the yield if not treated in good time. The situation is made more difficult by the fact that it is particularly difficult to predict the behaviour of wind-dispersed pathogens, such as powdery mildew (*Blumeria graminis* f. sp. *tritici*). To forecast pathogen development and spatial dispersal, a modelling process scheme was developed using the aforementioned R package, which combines regionalisation and machine learning techniques. It enables the prediction of the probability of yield-relevant infestation events for an entire federal state in northern Germany at a daily time scale. To run the models, weather and climate information are required, as is knowledge of the pathogen biology. Once fitted to the pathogen, only weather and climate information are necessary to predict such events, with an overall accuracy of 68% in the case of powdery mildew at a regional scale. Thereby, 91% of the observed powdery mildew events are predicted.

Keywords: machine learning; random forest; infestation forecast; powdery mildew

1. Introduction

Modelling of the epidemic spread of diseases often requires the implementation of a geographical methodology independent of the disease studied. However, often only local and non-spatial studies of the behaviour of the pathogens have been conducted. This paper illustrates how such an implementation of geographical expertise can be achieved using the example of minimising fungicide treatments.

In the agricultural sector, there is always a risk of yield losses due to fungal infestations. To mitigate this risk, plant protection products have been used to a considerable extent. These are often applied regularly, in order to prevent possible infestation. In 2016, this resulted in 149,430 tonnes of fungicides and bactericides being applied in agricultural use in all states of the European Union [1]. In order to reduce the amount of fungicides applied, directive 2009/128/EC was passed by the European Parliament [2] to achieve a more sustainable use of pesticides.

One approach towards reaching this goal is to reduce the regular application of fungicides by making targeted forecasts of impending dangerous infestations. At present, there exist different models for different pathogens, but most models describe only single effects and do not represent the complex events in the field [3–8]. In addition to the individual effects determined under laboratory conditions, there have been numerous attempts in recent years to identify infection processes using machine learning methods. However, satellites and orthophotos have mainly been used to detect existing infestations [9–12].

The objective of this study is not to detect infestations but to generate a temporal and spatial prediction of the epidemic spread of infestations, which requires the interdisciplinary combination of phytopathological and geographical methods and knowledge. The procedure developed for this purpose will be illustrated using the example of powdery mildew, a fungus that infests wheat. It is more important for the target group that if the model does not predict a yield-relevant infestation, no infestation will occur, as opposed to the model incorrectly predicting an infestation. In order to achieve that objective, in addition to a corresponding data set consisting of several years of infestation and weather data, it was determined which regionalisation methods and machine learning procedures are necessary to accomplish it.

2. Materials and Methods

2.1. Study Area

Schleswig-Holstein was used as an exemplary study area. The northernmost federal state of Germany (Figure 1) is located between the North Sea in the west and the Baltic Sea in the east.

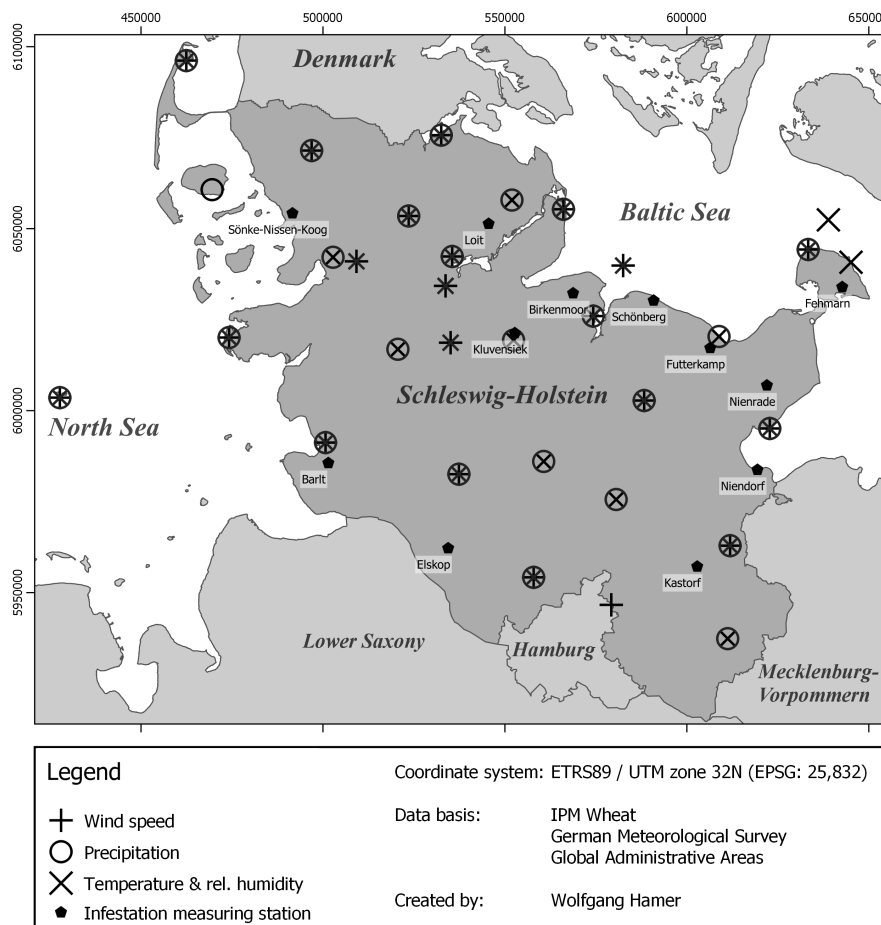


Figure 1. Map of the exemplary study area containing information on the weather and infestation measuring stations.

Following Fränzle [13], the research area was divided into marshes and mudflats in the west, residues of Saalian moraines east of the marshes, outwash plains in the centre, and uplands in the east, characterised by a series of ice advances. The highest elevation of the study area was 164 m, in the eastern uplands. Due to the soil composition in the study area, wheat is cultivated predominantly in the marshes and mudflats, and the eastern uplands. In the residues of Saalian moraines and the outwash plains, located in the centre of Schleswig-Holstein, less wheat is cultivated [14].

The climate of the region is determined by the westerlies and by its location in the transitional area between the North Atlantic and the European continental shelf [15]. The interplay of these influences is reflected in the seasonal course of the weather, with a cool and relatively dry spring, a rainy and moderately warm summer, and a sunny and mild autumn and winter [15]. The mean temperatures vary, on a long-term average, between 1.7 °C in winter, with higher temperatures at the coasts, and 16.5 °C in summer, with a stronger continental influence and higher temperatures in the south [16]. The long-term average precipitation is 809 mm/a, with higher values in the central and western part of the study area [16]. As a result of the westerlies, the highest wind speeds of the study area are reached on the west coast. On average, wind speeds around 4.3 m/s (at 10 m) have been recorded [16].

2.2. Sources of Data Used

The IPM (Integrated Pest Management) monitoring program of the Department of Phytopathology of Kiel University has studied the occurrence and dynamics of several pathogens in Schleswig-Holstein since 1993 (except for 2004) [17], and thereby, has created a unique data set of in-field observations of infestation events. Figure 1 shows the monitoring stations used since 1997. All observation sites are located absolutely in space using coordinates. Distributed in the wheat cultivation areas, they represent the disease situation in the study area selected for the exemplary application of the prediction approach. As described by Verreet et al. [17], the monitoring procedure took place weekly throughout the vegetation period, from growth stage (GS) 30 to 75 onwards at all leaf levels ($F - 7 =$ eight upper leaves to $F =$ flag leaf) [18]. In each studied plot, 30 plants were selected and disease incidence (percentage of infected plants or leaves per leaf position) and disease severity (percentage of leaf surface covered with pustules) of powdery mildew and other diseases were determined by counting pathogen-specific fungal structures. As the concept of the IPM monitoring program included the comparison of plant stocks treated with fungicides against untreated stocks, the control data can be used for the modelling approach. While most of the studied wheat varieties were available only for a restricted time, the variety Ritmo was used to allow for comparison throughout the experimental period. Powdery mildew was selected as an example of the application of the modelling approach. It has been well-described in the current literature, but is also difficult to predict because of its wind-borne behaviour. The spread of this pathogen is mainly influenced by weather conditions. The pathogenic spores need some wind speed to spread, but it cannot be too high, or the spores will not adhere to plants [19]. The infection caused by these spores is mainly influenced by the air temperature [20,21]: a humidity value close to 100% supports the pathogen and results in quick infection [21,22]. Precipitation can also influence the infection of the plant, but the impact of precipitation on the infection is typically negligible, except for heavy rain, which can wash the spores off of the plant surfaces [23]. In addition to the weather, the growth state of the plant is also important in the infection process. This can be illustrated, for example, by temperature values which are modified by a plant-specific function and summed (cumulative thermal unit (CTU)) [24]. For the scope of the current work, instead of predicting the actual disease incidence of powdery mildew, we focused on detecting whether a severe infestation event would occur, which was predicted by the exceedance of a pathogen-specific disease control threshold value. Klink [25] and Verreet et al. [17] used a threshold value of 70% disease incidence to trigger a fungicide treatment of the plant stock, to avoid yield-relevant damage by powdery mildew. Therefore, the exceedance of this threshold was used in our method to predict severe infestation events.

For the analysis of the climatic and weather situation in the study area, the data set of the German meteorological survey is available from the Open Data Server [16]. The Open Data Server provides regionalised climatic data summarised for 30 years, and for the years after 1995, hourly weather data for several locations in Schleswig-Holstein and the rest of Germany. The measuring stations of the German meteorological survey are shown in Figure 1. The number of available stations varies depending on the year and parameter considered. In 2019, temperature and humidity were measured at 26 stations, precipitation at 35 stations, and wind speed at 22 stations.

2.3. Methods Applied

In accordance with the objective of this paper, a modelling approach was developed as a combination of different interpolation and machine learning techniques. An overview of how these procedures work is given below.

Spatial interpolation methods are divided into deterministic and geostatistic methods [26]. In addition to the deterministic inverse distance weighting method, the geostatistical methods ordinary kriging, kriging with external drift, and random forest kriging were used in this study:

- **Inverse distance weighting (IDW)** is a common deterministic spatial interpolation method, which is defined as a weighted average of the data point values [27]. The weighting itself is a function of the distance:

$$\hat{z}(x_0) = \frac{\sum_{i=1}^n (d_i)^{-p} * z(x_i)}{\sum_{i=1}^n (d_i)^{-p}} \text{ if } d_i \neq 0, \quad (1)$$

where $\hat{z}(x_0)$ is the value to be estimated at the location x_0 , $z(x_i)$ is the known value at a specific location x_i , d_i is the distance between the estimated and the known data points, p is the inverse distance power, and n is the number of known data points closest to the estimated location. Although IDW does not model the spatial autocorrelation of the target variable (as the methods described below do), the procedure can still achieve good results. Wagner et al. [28] and Borges et al. [29], for example, found that the IDW interpolation achieved the best results for interpolating precipitation.

- **Ordinary kriging (OK)** was first established by Krige [30] and mathematically derived by Matheron [31]. Using this statistical interpolation method, the spatial dependence of the variable is not just assumed (as it would be with the IDW method), but analysed and integrated in the regionalisation. The analysis of the spatial dependence is based on variography: the similarity of point pairs (semivariance) is compared to their distance, and a function is adapted to the semivariance, which decreases with increasing distance Matheron [31]. By means of this variogram model, it is possible to calculate the weights (λ_i) by which the values of the surrounding points ($z(x_i)$) are multiplied, according to the kriging Equation (2) [31,32]:

$$\hat{z}(x_0) = \sum_{i=1}^n \lambda_i * z(x_i). \quad (2)$$

- **Kriging with external drift (KED)** uses the integration of independent variables in the process of Kriging interpolation. The principle of the procedure was established by Matheron [33] as a combination of OK and multiple linear regression of the dependent variable (the parameter of interest) with the independent variable (e.g., the coordinates or an elevation model). As described by Equation (3), a linear regression is fitted to the variables to predict the dependent variable by the (already regionalised) independent variables. The spatial prediction of the dependent variable starts with the application of the regression model ($\hat{m}(x_0)$) on the regionalised independent variables. In order to increase the accuracy of the prediction, the Kriging procedure ($\hat{z}(x_0)$), as in

Equation (2), is used to interpolate the residuals of the regression model. These two spatial data sets are then summed to calculate the aim variable ($\hat{e}(x_0)$):

$$\hat{e}(x_0) = \hat{m}(x_0) + \hat{z}(x_0). \quad (3)$$

The incorporation of covariates can lead to very good results, depending on the influence of such variables. Correspondingly, KED has proven to be especially appropriate for the interpolation of temperature [34] and wind speed [35]

- **Random forest kriging (RFK)** is similar to the KED procedure. However, instead of multiple linear regression, the relationship between the dependent and independent variables is represented by a random forest. The functionality of such a random forest is explained in more detail in the description of the machine learning methods.

In addition to the spatial interpolation methods, various machine learning methods were used. These are usually divided into unsupervised and supervised techniques. Unsupervised procedures search independently for patterns, while supervised procedures require a dependent variable (such as an infestation) for which the correlations with independent variables are examined. Three supervised machine learning methods were used in this work:

- **Decision trees (DT)** are based on the idea of recursive partitioning—splitting the data repeatedly into smaller subsets until they are homogeneous regarding the target variable. Of the different algorithmic methods created to find the ideal split, the C5.0 DT algorithm (developed by Ross Quinlan as an improvement to his C4.5 algorithm [36]), which uses the concept of entropy to create subsets with maximum purity, is the most common. The entropy for a specific subset is calculated as:

$$Entropy(S) = \sum_{i=1}^c -p_i * \log_2(p_i), \quad (4)$$

where S is the subset, c is the number of class levels, and p_i is the proportion of values in the specific class [37]. The quality of the splits, thus, is based on the summed entropy values of all subsets:

$$Entropy(T) = \sum_{i=1}^c w_i * Entropy(P_i), \quad (5)$$

with the total entropy (T) and the weighting (w_i) obtained by the proportion of examples in the subset [37]. The entropy values of the potential splits are weighted against each other, resulting in the information gain of potential splits, of which the one with the highest information gain is chosen.

- **Random forests (RF)**, developed by Breiman [38], combine the aforementioned DT with the bagging procedure, which uses bootstrap sampling—a random sampling with replacement method—to generate multiple predictions for a data set using the same prediction method [39]. The individual results are combined into one final prediction, using a plurality vote for classification and the average value for numerical outputs. The RF algorithm creates random subsets of the data set at each node of the developing tree [38]. The splits are evaluated using estimates, which test the ongoing grown trees at the cases not yet integrated into the model, in order to give an error estimate [38]. As a result, a forest of randomly grown DTs emerges. The combination of these trees, as an average for regression tasks and as the most frequent value for classification tasks, results in the final prediction of the RF model. RFs are assumed to be less prone to overfitting, as only parts of the data set are used to generate the individual trees [40]. Furthermore, they are assumed to be better at learning from larger data sets with a large number of features [37].
- **Boosted decision trees (BDT)** utilise the boosting procedure. Similar to the bagging method explained above, the boosting procedure creates a number of DTs [41]. In contrast to the decision trees of the RFs, however, the trees in BDT do not consist of random sub-data sets. At first,

only one DT is generated, based on the entire data set. The weight of misclassified instances in this first tree is, then, increased when the next tree is generated [42]. This process is repeated until the requested number of trees is reached. The final prognosis of this method is given by the majority decision of the generated DTs.

2.4. Modelling Approach

This section explains how the different spatial interpolation methods and machine learning techniques described in the previous section are combined into the process scheme of the proposed prediction approach (Figure 2). The application of the scheme requires that the data of independent parameters are available, which influence the dependent variable (i.e., the investigated infestation). These data should also reflect a temporal sequence, if this is assumed for the studied infestation. If these data are available locally, the first step of the scheme is to interpolate the points using deterministic or geostatistical regionalisation methods. The second step is to assign the respective values of the regionalised independent parameters to the infection events. This is necessary if the infestation is not triggered by the current conditions, but by the conditions of a foregone period. This approach allows for the aggregation of the hourly data of several days using the mean, the minimum, and the maximum values, assigning them to an observed infestation a few days later. This can vary depending on the infestation studied.

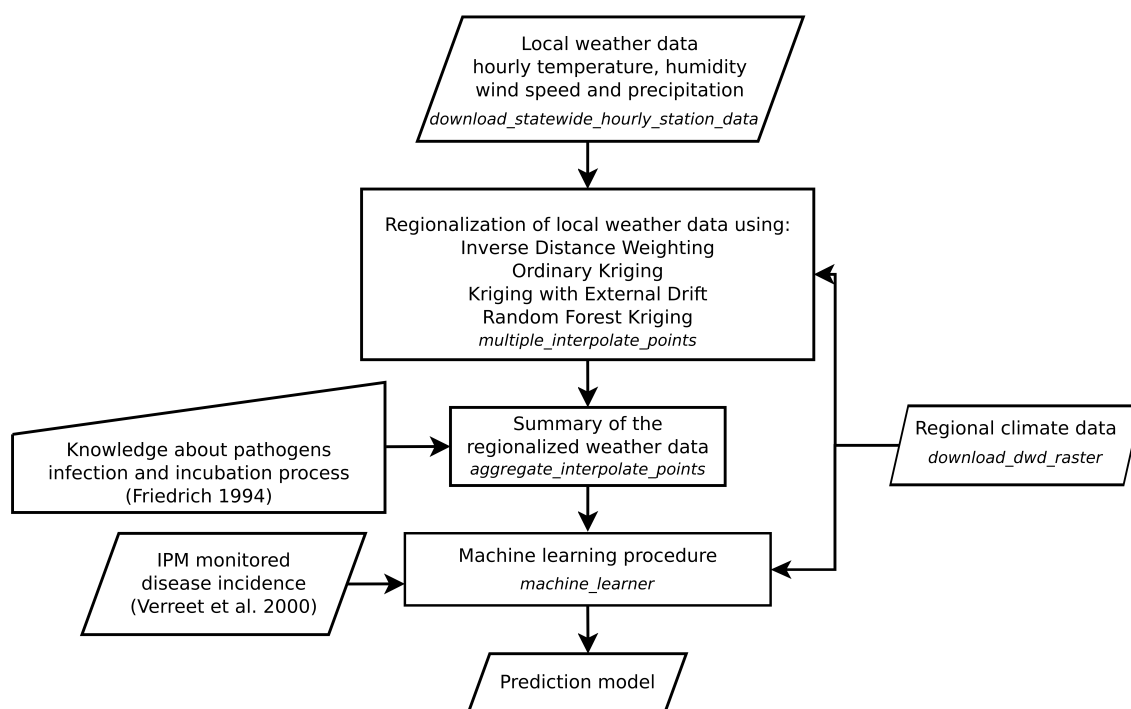


Figure 2. The proposed modelling approach combining machine learning and spatial interpolation methods (including the functions of the package papros [43]).

With the independent data modified in this way and the infestation data, a supervised machine learning algorithm was finally trained. The infestation data can be considered as either metric or as variable, classified by hazard. With the generated model, a spatial prediction of the infestation or the probability of the respective classes can then be made, based on daily updated independent data.

2.5. Use Case: Predict Powdery Mildew Infestation Events

The previous chapters described the methods applied and the combination of these procedures in the modelling approach. The modelling approach was, then, tested by applying it to the prediction of the pathogen powdery mildew (Section 2.2) for the study area of Schleswig-Holstein (Section 2.1).

The aim was to predict the probability of severe infestation events, which was defined by classifying the exceedance of the threshold value of 70% disease incidence (Section 2.2). The computer code, which was programmed to apply the process scheme (as described below) and to generate the results presented in this paper, is available in the Supplementary Data. The R programming language [44] was used, applying the functions of the package *papros* [43].

Firstly, following the description of the modelling approach (Section 2.4), the hourly local weather data (from the Open Data Server [16]) were interpolated. The weather parameters CTU, humidity, precipitation, temperature, and wind speed were considered, as these were determined to be relevant to the infection process (Section 2.2). In our study, no individual adaptation of the models to the respective autocorrelation was possible, as the interpolation of five parameters was executed for each hour of the years 1997–2019. Therefore, automated interpolation using the methods described in the previous section was tested. The covariables used for the KED and the RFK procedures were the longitude and latitude, and the regional climate data (1981–2010), which included the mean air temperature [45] from the Climate Data Center for the interpolation of the temperature; the cumulative thermal unit (CTU), the drought index calculated according to Pietzsch and Bissolli [46], for the interpolation of humidity; the mean precipitation for the interpolation of the precipitation; and the mean wind speed [47] for the interpolation of the wind speed. To determine which of the methods was best suited for this purpose, leave-one-out cross validations (LOOCV) were performed for each of the weather parameters at 500 randomly selected hourly values; this means that, for each of the 500 runs for each parameter, each weather measuring station was excluded and a spatial prediction was made with the other stations which had information on the parameter at the respective hour. For each hour and weather parameter, normalised root mean squared errors (NRMSE) were then calculated, resulting in 500 NRMSE values for each parameter at an average of 15 locations. After evaluating the optimal procedure for each parameter, this was used to interpolate temperature, CTU, humidity, wind speed, and precipitation at the locations of the infestation measuring stations, for each hour from 1997 to 2019.

In the second step, the regionalised weather data were summarised (Table 1) based on the knowledge of the pathogen infection and incubation processes (Section 2.2). The observed disease incidence of one day does not result from the weather of the same day but, instead, results much more from the previous weather conditions. As the main influence of the environment happens during the infection process, it is of interest to identify and aggregate the weather conditions during the infection and compare them with the observed occurrence of the pathogen several days later. This period between the end of the infection and the observation of the pathogen is referred to as incubation. The average infection and incubation durations for powdery mildew were calculated, depending on the average air temperature, using the equations of Friedrich [4], which resulted in a two-day infection time and seven-day incubation time. Therefore, the weather data of two days were summarised by the minimum, mean, and maximum values for each raster cell for each weather variable, and assigned to a date seven days later. Thus, the hourly resolution of the weather data was reduced to the daily resolution available by the IPM program. This was done for each in-field observation of pathogen behaviour in the IPM monitoring program. In addition to the summarised weather data, the daily and cumulative temperature units were calculated for the infection period, beginning on October 1 of the previous year. The equations used to do so were based on Soltani and Sinclair [24]. The results from the second step were summarised weather data. The daily and cumulative temperature unit for each of the IPM monitoring stations are presented in Figure 1, and that applies for each day of the last 20 years in Schleswig-Holstein. These data were, then, assigned to the observed disease incidence seven days after the summarised days. In the same way, the regional climate data and the elevation information [48] were assigned to the observed disease incidence.

The third step of the modelling process was the application of machine learning techniques. As with the spatial interpolation methods, the suitability of a machine learning method varies, depending on the target variable to be modelled and the covariates considered. A supervised learning technique with the ability to solve classification and regression tasks with a large number of independent variables

is required for the prediction of the infestation events. Classification and regression trees fulfil these requirements [37]. They also have the great advantage of showing how and on which variables the prediction of the model depends; however, according to Breiman [38], the accuracy of the predictions can be improved by using bagging procedures, such as that used in random forests. However, by using a bagging procedure, the insight that the DT procedure provides is reduced, as it is no longer just a tree which is the subject of a decision, but a large number of trees which are the subject of a majority decision. Therefore, methods such as the mean decrease accuracy have been used to gain insight into the importance of the model variables [49]. Accordingly, a comparison of the machine learning procedures listed in Section 2.3 should be used to determine whether the advantage mentioned by Breiman [38] can also be transferred to the prediction of powdery mildew cases, or whether this does not occur here. However, besides the machine learning method used, the aim of the learning process should not be disregarded in this comparison. Machine learning methods are usually trained with the aim of creating a model with the largest possible accuracy—such as the *C5.0* function [50] or the *random forest* function [51]. In the case of the prediction of infestation events, a non-detected exceedance of the disease control threshold value is considered graver than an under-run of the threshold detected as an exceedance. Therefore, we created a function which varied the number of trees in BDT and the number of variables randomly sampled as candidates for each split and the weighting of the classes for the RF algorithm using the area under the receiver operating characteristic (ROC) curve. The ROC plots the true positive rate of the model (the rate of correctly predicted exceedances) versus the false positive rate (the rate of wrongly predicted exceedances) [52]. Our function searches for a weighting which obtains the largest possible area under the ROC curve, in order to create a model with not only good accuracy but also with good sensitivity; which is the proportion of true positive values of the combined true positive and false negative values, not predicted exceedances [53]. This search within the function was done by holdout validation of the data entered to create the model. These were first divided in the ratio 70:30, according to the available years. With 70% of the data, models with varying weightings were generated iteratively. Based on the 30% remaining data, the receive operating characteristic area under the curve (ROC AUC) value of each model was checked, and finally, the weighting with the highest value was used to generate a model based on all data used in this step.

Analogous to the use of LOOCV for the spatial interpolation of the weather data, the prediction of the probability of infestations based on the machine learning methods was also tested using LOOCV. However, in contrast to the spatial interpolation LOOCV, where a single site was omitted while the forecast was made with the surrounding sites, the observed infestations of a whole year were excluded and compared with the predictions of a model adapted to the observations of the other years for the machine learning method LOOCV. Corresponding to the IPM stations, this led to an average of eight compared sites, each with an average of 12 observations for each year. The data set used to make a prediction for the year 2019 (91 observations), thus, consisted of 2047 entries. Each of these entries consisted of the parameters listed in Table 1. In addition to the target variable (the boolean classification of the infestation), these also included many independent variables. Most of these are hourly weather data, which were first regionalised using the procedures also listed in Table 1 (the choice of procedures is explained in the results) and then aggregated by minimum, average, and maximum values. In addition, the long-term climatic variables already regionalised by the DWD [16] make up a considerable proportion of the independent variables. The only variable that is not directly weather-related is the altitude model, which is also included [48]. The elevation model incorporates site characteristics that represent, for example, the natural structure of the study area (Section 2.1), as described by Fränzle [13]. For this purpose, the classification described by Fränzle [13] could also have been used as an independent variable, but since observed infestation values are not available for all natural units, an extrapolation of the predictions for these areas would not have been possible, which led to the use of the elevation model.

Table 1. Overview of the regionalisation methods applied (justification in the results) and the variables used.

Modell Feature	Selected Parameter
Air temperature interpolation method	Random Forest Kriging
Air humidity interpolation method	Inverse Distance Weighting
Precipitation interpolation method	Inverse Distance Weighting
Wind speed interpolation method	Kriging with External Drift
CTU interpolation method	Ordinary Kriging
Target variable	exceedance of 70 % disease incidence
Independent variables	minimum air temperature
	mean air temperature
	maximum air temperature
	minimum air humidity
	mean air humidity
	maximum air humidity
	minimum precipitation
	mean precipitation
	maximum precipitation
	minimum wind speed
	mean wind speed
	maximum wind speed
	CTU (based upon air temperature)
elevation information [48]	
air temperature (climatic) [16]	
precipitation (climatic) [16]	
drought index (climatic) [16]	
wind speed (climatic) [16]	

3. Results

The most suitable interpolation method depends on the spatial behaviour of the variable to be interpolated and the characteristics of the investigated area. In order to find out which method was most suitable for the interpolation of the hourly weather data in the study area, multiple leave-one-out cross validations were applied to each the parameters for 500 randomly selected hours. For each validation, the NRMSE was computed, which should be as low as possible (Section 2.5). Figure 3 shows box-plots of these NRMSE values for the four interpolation methods and five parameters of interest. The box-plots show that the results obtained by the methods were similar but that different interpolation methods achieved the smallest errors, depending on the parameters to be interpolated. In addition to the illustration (Figure 3), tests were performed to determine whether the interpolation procedures for the parameters differed significantly. The Wilcoxon rank-sum test [54] was used for this purpose. The lowest error in CTU interpolation was achieved, on average, when applying ordinary kriging and inverse distance weighting. As the OK NRMSE was slightly lower, this method was chosen for interpolation, even though the distribution of the errors was not significantly different from that of the IDW method. As with the other parameters, the option was chosen to use the IDW procedure if automated interpolation with the OK procedure was not be possible (e.g., because too few points were available). When interpolating the humidity, the lowest error was achieved with the IDW method, which differed significantly from all other methods except random forest kriging.

For the other interpolation methods, the procedure with the lowest error was chosen: in this case, the IDW procedure. The kriging with external drift procedure showed some high outliers, indicating considerably worse predictions in some cases for this parameter. This was also seen in the interpolation of precipitation. Again, the IDW interpolation achieved the lowest NRMSE and differed significantly from all other methods. A different picture emerged with the regionalisation of the temperature, where the RFK procedure produced the lowest error, but only differed significantly from the OK interpolation. RFK was selected for interpolation, according to the smallest error. Regionalisation of wind speed was

clearer and unambiguous, where the two methods using covariates had significantly lower errors than the other methods. KED was selected to interpolate this parameter.

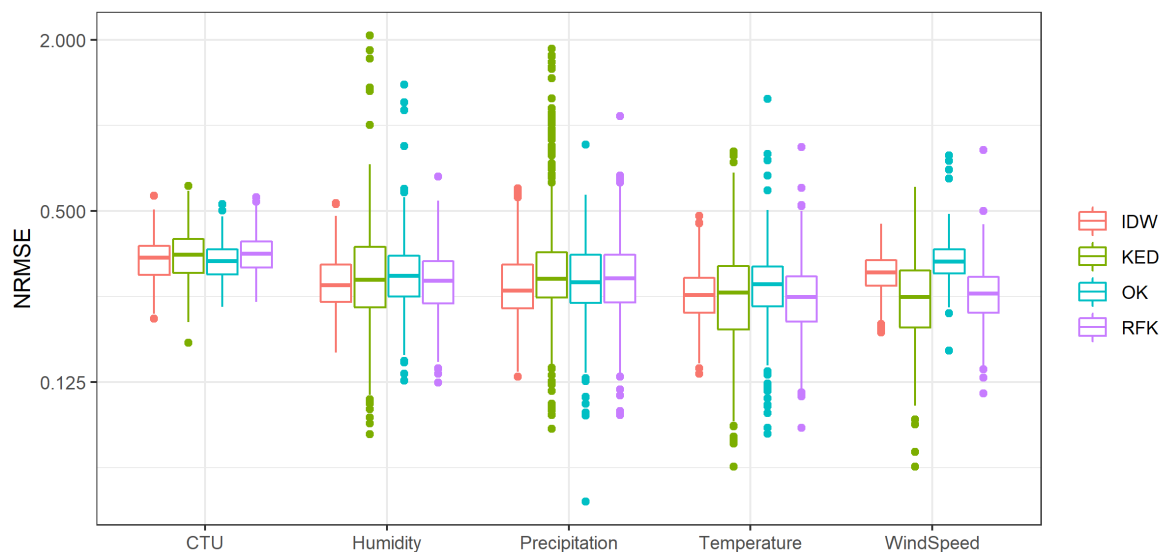


Figure 3. Normalised root mean squared error (NRMSE) of different deterministic and stochastic interpolation methods for the weather parameters cumulative thermal units (CTU), air humidity (%), temperature ($^{\circ}$ C), precipitation (mm/h), and wind speed (m/s).

Leave-one-out cross validation applied to the whole modelling approach (Section 2.5) allowed for estimation of the performance of the approach in predicting yield-endangering infestations with powdery mildew. Figure 4 illustrates the proportion of observed (bottom) and predicted severe events over different years using the true positive, true negative, false positive, and false negative values. The true positive value is the count of correctly predicted severe infestation events, while the count of unpredicted events is referred to as false negative. True negative are the correct predictions of non-occurring events, while the false positive value describes the number of predicted severe events which did not occur.

The proportion of endangering events varied over the examined years, with no infestations in 1999 and high infestation rates in 2003, 2009, and 2010. These years especially illustrate the differences between the various used and fitted machine learning methods. In 1999, the DT resulted in the highest true negative and the lowest false positive values, followed by the BDT and RF methods, which were not fitted to the ROC AUC but to the accuracy and the fitted BDT.

The fitted RF method showed the highest false positive rate, indicating that this method warned more often than the other procedures. The same was evident in the years of high infestation in 2009 and 2010. In these years, the BDT and the unfitted RF procedure did not predict about a third of the dangerous infestations as such. This proportion was lowest for the Random Forest approach. Over the years, it can be seen that the false negative rate (which is most dangerous for the farmer) was the lowest for the fitted RF. The fitted RF method differed considerably from the unfitted. In the BDT approach, this difference was less pronounced but can be seen in the years 2018 and 2019.

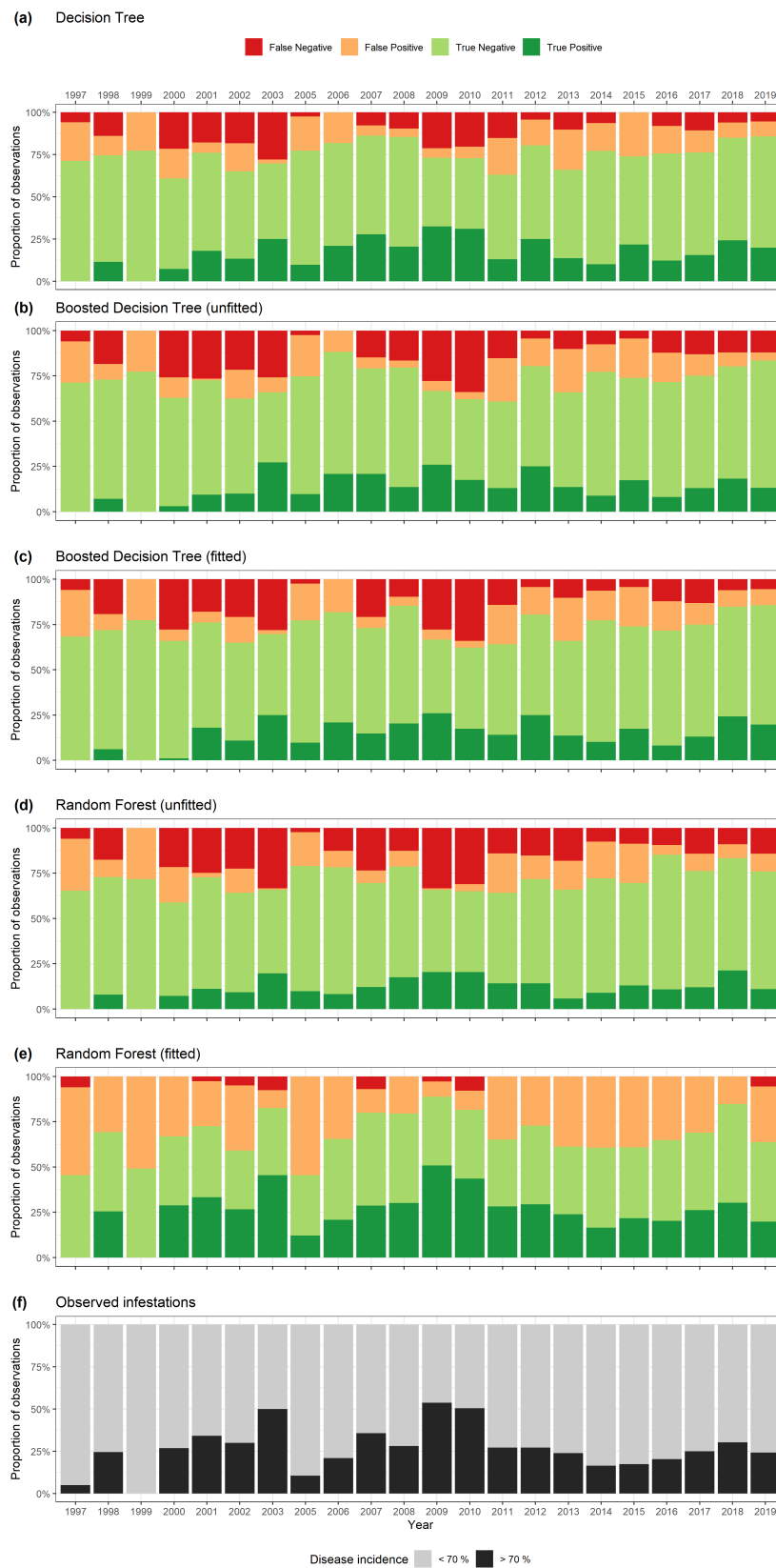


Figure 4. Observed disease incidence (f) and quality of the prediction of varying machine learning procedures (a–e) over the studied period from 1997 to 2019 in the study area. The procedures marked as “fitted” were adopted to reach a high receive operating characteristic area under the curve (ROC AUC) instead of a high accuracy.

Table 2 shows the statistical parameters of the iterative predictions resulting from the true positive, true negative, false positive, and false negative values summarised over all years. The accuracy describes the proportion of correct predictions, also visualised as true positive and true negative proportions in Figure 4. Specificity is the proportion of true negative values of the combined true negatives and false positives [53]; that is, it is the proportion of all underruns of the threshold value of all predicted underruns. Precision is the proportion of true positive values of the combined true positive and false positive values; therefore, it is the proportion of all occurred exceedances of all predicted exceedances [55]. The highest accuracy was achieved with Decision Trees, slightly followed by fitted and unfitted Boosted Decision Trees and unfitted Random Forests. The fitted Random Forests obtained a lower accuracy but showed, by far, the highest sensitivity. Almost 92% of dangerous incidents were predicted, as such. The second best method in this field, DT, only managed to do this by almost 60%. However, the good sensitivity was at the expense of the specificity, where the performance of the fitted Random Forests was clearly poorer than the unfitted ones. The Precision was similarly low for all procedures, with the highest values in Decision Trees. The ROC AUC value was highest for the adapted RF method, which was to be expected. The difference between the ROC AUC value and the value of the accuracy-fitted method was more pronounced than for the fitted and unfitted Boosted Decision Tree. These values were lower for both BDTs than for the DT method.

Table 2. Statistical measures of the performance of the different machine learning methods, separated by models either unfitted (uf) or fitted (f) following an optimal receiver operating characteristic (ROC) AUC.

	Decision Tree	Boosted DT (uf)	Random Forest (uf)	Boosted DT (f)	Random Forest (f)
Accuracy	0.751	0.727	0.711	0.733	0.679
Sensitivity	0.596	0.475	0.404	0.503	0.919
Specificity	0.812	0.826	0.832	0.824	0.584
Precision	0.556	0.520	0.488	0.531	0.466
ROC AUC	0.704	0.651	0.618	0.664	0.752

The spatial prediction presented in Figure 5 shows a clear differentiation of the probability of endangering infestations between the east and west coasts of the investigated area. In the observed disease incidences, this was also evident. This differentiation is most pronounced in the case of fitted random forests. In contrast, the prediction of the unfitted boosted decision trees show a stronger differentiation between north and south, indicating a stronger influence of parameters representing long-term continentality. In addition to these spatial dynamics, Figure 5 also shows the changes over time. The DT and the fitted BDT approaches show hardly any changes in the prediction of the probability of dangerous infestations. This indicates a strong inflow of the temporally non-dynamic climate variables and a limited influence of the hourly weather data on the model forecast. The (non-fitted) BDT and the RF approaches were more dynamic. Again, the RF approach demonstrates how training on ROC AUC can be used to predict the actual exceedances correctly, but also how false predictions are made at the end of the infestation period about non-occurring exceedances. This can also be seen in the false positives of the bar of 2019 in Figure 4. The lower probabilities of the unfitted RF approach, however, resulted in unpredicted dangerous infestations, and consequently, in a larger false negative fraction.

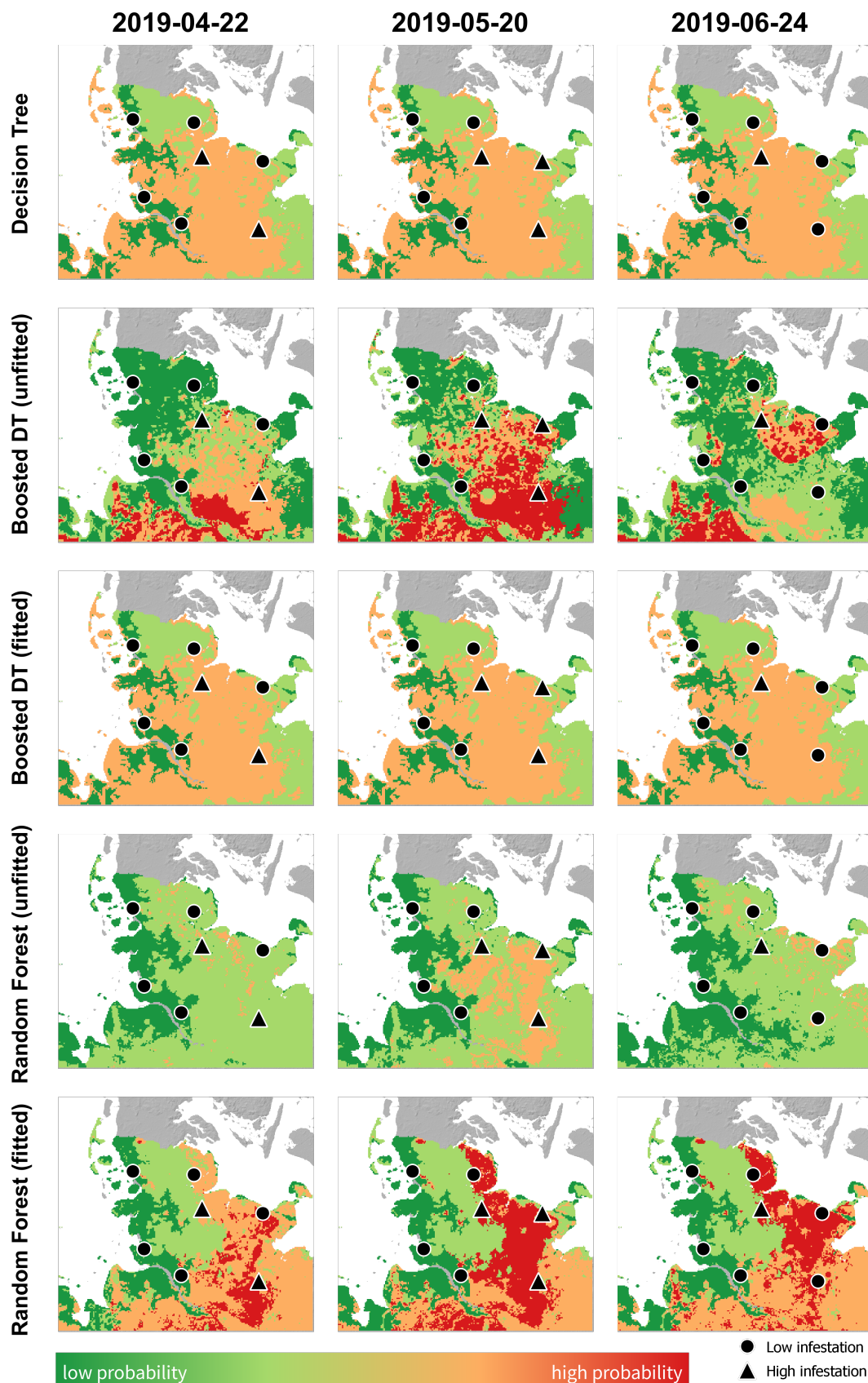


Figure 5. Predicted probabilities exceeding the powdery mildews’ threshold value (70%) on wheat in the study area, applying different machine learning procedures at selected dates in 2019. Observations of the IPM monitoring are included as spots and triangles.

4. Discussion

Previous work focusing on the use of machine learning in the context of phytopathological infestations has already produced satisfactory results. For example, Lu et al. [11] achieved an accuracy of 74% in the identification of anthracnose crown rot. Such accuracy is a result comparable with the overall accuracies reached in our study (Table 2). Similar to most studies published in this field, Lu et al. [11] used hyperspectral data, which were collected in-field using a mobile platform. Thus, no prediction was made based on the behaviour of the pathogen, but instead, an image classification of the plants leaves was performed. In the scenario we are looking at, it is the prediction and the transferability to the scale of Schleswig-Holstein that enables us to react promptly to dangerous infestations. Accordingly, a comparison of our approach with prediction systems that focus on the behaviour of the pathogen is more appropriate. At present, a number of modelling systems are already available in this field, such as MEVA-PLUS [6], WHEATPEST [8], InfoCrop [56], and WHEGROSIM [7]. These approaches have also achieved considerable accuracies, such as the WHEGROSIM model, which achieved a very high coefficient of determination of 0.89 [7]. Wen et al. [57] used machine learning methods, such as RF, to predict the behaviour of a pathogen. They predicted the spread of rust spores, which were influenced by the preceding weather. For Wen et al. [57], the RF method achieved the highest accuracy of 83% in predicting the spread. However, they used the variable of distance to the origin of the spores, which makes the transferability of the model difficult. However, similar to the other approaches, parameters not available at the regional and wider scales (such as an initial infestation value or the distance to the origin of the spores) are required to achieve these accuracies. Accordingly, these accuracies cannot be transferred to a spatial prediction on the scale of Schleswig-Holstein. Wheat, however, is not grown sporadically. It is an important part of most people's daily nutrition. As a result, wheat was grown over an area of 260,926 km² in 2017 in the European Union, according to the Food and Agriculture Organization (FAO) of the United Nations [58]. This accounts for 46.85% of the total area used for cereals in 2017 in the EU, which underlines the necessity of spatial predictions. An example prediction was presented in this paper. The results clearly show that spatial prediction is necessary.

Such spatial prediction was performed in our article, where the results showed that different interpolation methods achieve the best results according to the differing weather data used (Figure 3). However, these results must be taken with caution and should not be generalised for the parameters. Even if different methods were chosen based on the best result, only in the case of interpolation of the wind speed did the best prediction method differ significantly from the simplest deterministic IDW method. This effect relies on various causes. The empirical and theoretical variograms necessary for the stochastic methods were generated automatically, due to the large number of interpolations. During processing, control by expert intervention on the modelling and the integration of the autocorrelative behaviour of the target variable were not considered. For the KED and RFK methods, in which machine learning was implemented, the small number of climate stations in the early years may have led to the fact that these models were not representative of the entire study area. The choice of covariates must also be taken into account in these procedures. In the initial experiments, the elevation model was also integrated; however, this did not have a positive effect on the accuracy of the prediction. Although the eastern part of the study area was more relieved, the maximum elevation was only 164 m. In other areas, the consideration of this variable could lead to a higher relevance of covariable-considering methods.

Comparison of the machine learning methods also showed that there was no optimal method for every purpose. The correct procedure depended, rather, on the user's goal. In our example, the highest overall accuracy was achieved by the decision tree model (Table 2). In many cases, this may also correspond to the user's goal. For the example chosen here, however, a good negative prognosis by the highest possible sensitivity was deemed more important. The newly created enhancement of the machine learning approach using the area under the ROC curve has proven to optimise the sensitivity of the created model; however, this resulted in some loss in the model's overall accuracy. The overestimations of the risk situation—indicated by the low precision percentages—associated

with this optimisation, are acceptable for a good negative prognosis for the prediction of an epidemic spread of pathogens. This improvement had a stronger impact on random forests than on the boosted decision trees (Figure 4). This may be, in the case of boosted decision trees, because it only affected the number of trees generated, whereas in the case of random forests, it also included a weighting of the cases themselves. Figure 5 shows that the fitting to the ROC AUC value in the BDT prognosis was made by reducing the number of decision trees which generated the fitted BDTs results, as the spatial prediction of the fitted boosted decision trees was similar to that of the decision trees. However, the statistical measures in Table 2 and Figure 4 show that this was not always necessarily the case.

5. Conclusions

In conclusion, the use of geographical methodology has been shown to be suitable for predicting the epidemic spread of infestations. The combination of geostatistical regionalisation, machine learning methods, and long-term phytopathological data series (as depicted in Figure 2) has achieved verifiable predictions (e.g., Table 2) of the infestation events which endanger yields. In this context, the adapted random forest approach proved to be the most appropriate, as the improved prediction of actual cases improved by this approach outweighed the increased false positive rate. In future works, the results of transferring these models or the entire approach to other areas should be studied. These other areas may not only be spatial, but thematic as well. The *papros* package [43] (which was used to compute the results of our work) was designed to be modular, such that not only can weather data from other regions be easily obtained, but also completely different spatial data may be used. It should also be investigated to what extent the results achieved with other machine learning methods deviate from those presented here. Additionally, future works can show whether parameters that have not been taken into account up to now (e.g., those that have not been included on the basis of expert opinions) can influence the forecast, and thus lead to an improvement of the model and the knowledge of the pathogen.

Supplementary Materials: The R code used to generate the results of this study can be found at Zenodo (<http://doi.org/10.5281/zenodo.3532695>).

Author Contributions: Conceptualization, W.B.H., T.B., and H.K.; methodology, W.B.H.; software, W.B.H.; validation, W.B.H.; formal analysis, W.B.H.; resources, T.B., H.K., and J.-A.V.; writing—original draft preparation, W.B.H.; writing—review and editing, W.B.H., T.B., and H.K.; visualization, W.B.H.; funding acquisition, R.D. and J.-A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Stiftung Schleswig-Holsteinische Landschaft. We acknowledge financial support by Land Schleswig-Holstein within the funding programme Open Access Publikationsfonds.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BDT	boosted decision tree
DT	decision tree
f	fitted (by ROC AUC)
IDW	inverse distance weighting
IPM	integrated pest management
KED	kriging with external drift
OK	ordinary kriging
RF	random forest
RFK	random forest kriging
ROC	receive operating characteristic
ROC AUC	receive operating characteristic area under the curve
uf	unfitted (by ROC AUC)

References

1. FAOSTAT Pesticides Use. Available online: fenixservices.fao.org/faostat/static/bulkdownloads/Inputs_Pesticides_Use_E_All_Data.zip (accessed on 15 October 2019).
2. European Parliament. DIRECTIVE 2009/128/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 21 October 2009 Establishing a Framework for Community Action to Achieve the Sustainable Use of Pesticides. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02009L0128-20190726> (accessed on 15 October 2019).
3. Hau, B. *Epidemiologische Simulatoren als Instrumente der Systemanalyse mit Besonderer Berücksichtigung eines Modells des Gerstenmehltaus*; Acta Phytomedica; P. Parey: Hamburg, Germany, 1985.
4. Friedrich, S. *Prognose der Infektionswahrscheinlichkeit durch Echten Mehltau an Winterweizen (Erysiphe graminis DC. f. sp. tritici) anhand Meteorologischer Eingangsparameter*; Wissenschaftsverlag Mainz Aachen: Mainz, Germany, 1994.
5. Jensen, A.L.; Jensen, F.V. MIDAS—An Influence Diagram for Management of Mildew in Winter Wheat. In *UAI'96: Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1996; pp. 349–356.
6. Bruns, J.B. Untersuchungen zur wetterbasierten Befallssimulation und Verlustprognose von Echem Mehltau (*Erysiphe graminis* D.C. f. sp. tritici Marchal) an Winterweizen. Ph.D. Thesis, Georg-August-Universität Göttingen, Göttingen, Germany, 1996.
7. Rossi, V.; Giosuè, S. A dynamic simulation model for powdery mildew epidemics on winterwheat. *EPPO Bull.* **2003**, *33*, 389–396. [[CrossRef](#)]
8. Willocquet, L.; Aubertot, J.; Lebard, S.; Robert, C.; Lannou, C.; Savary, S. Simulating multiple pest damage in varying winter wheat production situations. *Field Crops Res.* **2008**, *107*, 12–28. [[CrossRef](#)]
9. Zhang, S.; Shang, Y.; Wang, L. Plant disease recognition based on plant leaf image. *J. Anim. Plant Sci.* **2015**, *25*, 42–45.
10. Delwiche, S.R.; Yang, I.C.; Graybosch, R.A. Multiple view image analysis of freefalling U.S. wheat grains for damage assessment. *Comput. Electron. Agric.* **2013**, *98*, 62–73, doi:10.1016/j.compag.2013.07.002. [[CrossRef](#)]
11. Lu, J.; Ehsani, R.; Shi, Y.; Abdulridha, J.; de Castro, A.I.; Xu, Y. Field detection of anthracnose crown rot in strawberry using spectroscopy technology. *Comput. Electron. Agric.* **2017**, *135*, 289–299, doi:10.1016/j.compag.2017.01.017. [[CrossRef](#)]
12. Mwebaze, E.; Biehl, M. *Prototype-Based Classification for Image Analysis and Its Application to Crop Disease Diagnosis*; Springer International Publishing: Cham, Switzerland, 2016; pp. 329–339; doi:10.1007/978-3-319-28518-4_29. [[CrossRef](#)]
13. Fränzle, O. *Streifzug durch 6000 Jahre Landnutzungs- und Landschaftswandel in Schleswig-Holstein*; Chapter Reliefentwicklung und Bodenbildung in Schleswig-Holstein, *EcoSys*; Verein zur Förderung der Ökosystemforschung zu Kiel e.V.: Kiel, Germany, 2004; Volume 41, pp. 11–35.
14. Meynen, E.; Schmithüsen, J. *Handbuch der Naturräumlichen Gliederung Deutschlands: 1953–1962*; Bundesanst. für Landeskunde u. Raumforschung: Bad Godesberg, Germany, 1962; Number Bd. 2.
15. Schlenger, H.; Paffen, K.; Stewig, R. *Schleswig-Holstein: Ein Geographisch-Landeskundlicher Exkursionführer*; Schriften des Geographischen Instituts der Universität Kiel, Hirt: Kiel, Germany, 1969.
16. DWD 2019 Open Data Server. Available online: <https://www.dwd.de/EN/ourservices/opendata/opendata.html> (accessed on 7 May 2019).
17. Verreet, J.; Klink, H.; Hoffmann, G. Regional monitoring for disease prediction and optimization of plant protection measures: The IPM wheat model. *Plant Dis.* **2000**, *84*, 816–826. [[CrossRef](#)]
18. Zadoks, J.C.; Chang, T.T.; Konzak, C.F. A decimal code for the growth stages of cereals. *Weed Res.* **1974**, *14*, 415–421, doi:10.1111/j.1365-3180.1974.tb01084.x. [[CrossRef](#)]
19. Cao, X.; Duan, X.; Zhou, Y.; Luo, Y. Dynamics in concentrations of *Blumeria graminis* f. sp. tritici conidia and its relationship to local weather conditions and disease index in wheat. *Eur. J. Plant Pathol.* **2012**, *132*, 525–535, doi:10.1007/s10658-011-9898-8. [[CrossRef](#)]
20. Eckhardt, H.; Steubing, L.; Kranz, J. Untersuchungen zur Infektionseffizienz, Inkubations- und Latenzzeit beim Gerstenmehltau *Erysiphe graminis* f. sp. hordei. *J. Plant Dis. Prot.* **1984**, *91*, 590–600.
21. Beest, D.E.T.; Paveley, N.D.; Shaw, M.W.; van den Bosch, F. Disease-weather relationships for powdery mildew and yellow rust on winter wheat. *Phytopathology* **2008**, *98*, 609–617. [[CrossRef](#)]

22. Hau, B.; de Vallavieille-Pope, C. Wind-dispersed diseases. In *The Epidemiology of Plant Diseases*; Cooke, B., Jones, D.G., Kaye, B., Eds.; Springer: Dordrecht, The Netherlands, 2006; pp. 387–416; doi:10.1007/1-4020-4581-6_15. [CrossRef]
23. Merchán, V.; Kranz, J. Studies on the effect of rain on the infection of wheat by *Erysiphe graminis* DC. f. sp. tritici Marchal. *J. Plant Dis. Prot.* **1986**, *93*, 255–261.
24. Soltani, A.; Sinclair, T. *Modeling Physiology of Crop Development, Growth and Yield*; CAB Books; CABI: Wallingford, UK, 2012.
25. Klink, H. Geoepidemiologische Erhebungen von Weizenpathogenen in Schleswig-Holstein unter Anwendung und Entwicklung des Integrierten Pflanzenschutzsystems (IPS-Modell Weizen) Für Einen Minimierten, Bedarfsgerechten Fungizideinsatz (1993–1996). Ph.D. Thesis, Christian-Albrechts-Universität zu Kiel, Kiel, Germany, 1997.
26. Hengl, T. *A Practical Guide to Geostatistical Mapping*; Amsterdam University Press: Amsterdam, The Netherlands, 2011.
27. Shepard, D. A Two-dimensional Interpolation Function for Irregularly-spaced Data. In *ACM'68: Proceedings of the 1968 23rd ACM National Conference*; ACM: New York, NY, USA, 1968; pp. 517–524; doi:10.1145/800186.810616. [CrossRef]
28. Wagner, P.D.; Fiener, P.; Wilken, F.; Kumar, S.; Schneider, K. Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. *J. Hydrol.* **2012**, *464–465*, 388–400, doi:10.1016/j.jhydrol.2012.07.026. [CrossRef]
29. Borges, P.d.A.; Franke, J.; da Anunciação, Y.M.T.; Weiss, H.; Bernhofer, C. Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. *Theor. Appl. Climatol.* **2016**, *123*, 335–348, doi:10.1007/s00704-014-1359-9. [CrossRef]
30. Krige, D.G. *A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand*; University of the Witwatersrand: Johannesburg, South Africa, 1951.
31. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
32. Cressie, N. The origins of kriging. *Math. Geol.* **1990**, *22*, 239–252, doi:10.1007/BF00889887. [CrossRef]
33. Matheron, G. *Le Krigeage Universel*; Cahiers du Centre de Morphologie Mathématique; Ecole des Mines de Paris: Fontainebleau, France, 1969; No. 1.
34. Benavides, R.; Montes, F.; Rubio, A.; Osoro, K. Geostatistical modelling of air temperature in a mountainous region of Northern Spain. *Agric. Forest Meteorol.* **2007**, *146*, 173–188. [CrossRef]
35. Eguía, P.; Granada, E.; Alonso, J.; Arce, E.; Saavedra, A. Weather datasets generated using kriging techniques to calibrate building thermal simulations with TRNSYS. *J. Build. Eng.* **2016**, *7*, 78–91. [CrossRef]
36. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
37. Lantz, B. *Machine Learning with R*, 2nd ed.; Packt Publishing Ltd.: Birmingham, UK, 2015.
38. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324. [CrossRef]
39. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140, doi:10.1007/BF00058655. [CrossRef]
40. Wyner, A.J.; Olson, M.; Bleich, J.; Mease, D. Explaining the success of adaboost and random forests as interpolating classifiers. *J. Mach. Learn. Res.* **2017**, *18*, 1–33.
41. Quinlan, J.R. *Bagging, Boosting, and C4. 5*; AAAI/IAAI: Palo Alto, CA, USA, 1996; Volume 1, pp. 725–730.
42. Schapire, R.; Freund, Y. *Boosting: Foundations and Algorithms*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2012.
43. Hamer, W. *Papros-Pathogen PROgnosis System*; 2019. Available online: <https://zenodo.org/record/2574046> (accessed on 7 May 2019). [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
45. DWD 2016 Multi-Annual Means of Grids of Air Temperature (2m) over Germany 1981–2010, Version v1.0. Available online: ftp://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/air_temperature_mean/DESCRIPTION_gridsgermany_multi_annual_air_temperature_mean_8110_en.pdf (accessed on 7 May 2019).
46. Pietzsch, S.; Bissolli, P. A modified drought index for WMO RA VI. *Adv. Sci. Res.* **2011**, *6*, 275–279, doi:10.5194/asr-6-275-2011. [CrossRef]

47. DWD 2014 Gridded Mean of Annual Wind Speeds from 10 m to 100 m (in 10 m Steps) above Ground and Weibull Parameters, for Germany, Version V0.1. Available online: ftp://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/wind_parameters/resol_1000x1000/DESCRIPTION_gridsgermany_resol_1000x1000_en.pdf (accessed on 7 May 2019).
48. EU Copernicus Programme. *European Digital Elevation Model (EU-DEM), Version 1.1*; European Environment Agency: 2016. Available online: <https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1/view> (accessed on 7 May 2019).
49. Nicodemus, K.K. Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. *Brief. Bioinform.* **2011**, *12*, 369, doi:10.1093/bib/bbr016. [[CrossRef](#)] [[PubMed](#)]
50. Kuhn, M.; Quinlan, R. *C50: C5.0 Decision Trees and Rule-Based Models*; R Package Version 0.1.2; 2018. Available online: <https://cran.r-project.org/web/packages/C50/index.html> (accessed on 7 May 2019).
51. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
52. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874, doi:10.1016/j.patrec.2005.10.010. [[CrossRef](#)]
53. Altman, D.G.; Bland, J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ Br. Med J.* **1994**, *308*, 1552. [[CrossRef](#)] [[PubMed](#)]
54. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [[CrossRef](#)]
55. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 2229–3981.
56. Aggarwal, P.; Kalra, N.; Chander, S.; Pathak, H. InfoCrop: A dynamic simulation model for the assessment of crop yields, losses due to pests, and environmental impact of agro-ecosystems in tropical environments. I. Model description. *Agric. Syst.* **2006**, *89*, 1–25. [[CrossRef](#)]
57. Wen, L.; Bowen, C.R.; Hartman, G.L. Prediction of Short-Distance Aerial Movement of *Phakopsora pachyrhizi* Urediniospores Using Machine Learning. *Phytopathology* **2017**, *107*, 1187–1198. [[CrossRef](#)]
58. FAOSTAT Crops Production. Available online: fenixservices.fao.org/faostat/static/bulkdownloads/Production_Crops_E_All_Data.zip (accessed on 15 October 2019).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).