Breeding Resistance Against Verticillium Wilt in Chrysanthemums

Statistical Consulting - Final Report

Khalajzadeh, Niloofar

s3420760@vuw.leidenuniv.nl

January, 2024

Abstract

This study investigates methods for measuring plant susceptibility to a specific disease and explores the factors influencing such susceptibility. Two approaches, statistical analysis and the use of the Area Under the disease progress Curve, are employed to assess plant susceptibility. The results show that data from Week 6, utilizing AUC as a measurement tool, and ANOVA with post-hoc analysis are effective parameters for evaluating plant susceptibility. Additionally, increasing the number of blocks while reducing the number of plants yields more reliable results. The study provides insights into classifying plants as resistant, tolerant, or susceptible based on AUC values and Week 6 data.

Keywords: Plant susceptibility, AUC, Statistical analysis, Disease progression

1 Introduction

1.1 Background

Chrysanthemums are a staple of autumn, bringing vibrant colors to porches and storefronts. However, lurking beneath their cheerful blooms is a potential threat: chrysanthemum verticillium wilt. This disease is caused by verticillium fungi that reside in the soil and can attack plants without early warning signs.

This group of fungi is known for its stubborn persistence in soil and its ability to interrupt a plant's water supply, which can lead to wilting and even death. These fungi reproduce through conidia, tiny spores that spread far and wide via air and water, quickly establishing new colonies.

When chrysanthemums are hit by verticillium wilt, the first signs are often a sudden drooping of the plant on a hot day and a change in leaf color to a pale yellow. This yellowing is a hallmark of plant susceptibility, a condition where the plant lacks the ability to fight off the disease, leading to symptoms like wilting and color changes.

As the disease takes hold, the leaves may start browning, a sign of increasing damage. Some chrysanthemums show tolerance, they may experience some yellowing and browning but manage to survive, maintaining

enough health to continue growing and blooming despite the disease's presence.

In severe cases, the plant may become overwhelmed, leading to a state of susceptibility where it loses all its leaves, significantly diminishing its blooms, or ultimately dying. On the other hand, resistant chrysanthemums are those that can withstand the fungus's effects without showing such drastic symptoms, maintaining their health and vibrant appearance.

Verticillium wilt affects more than just chrysanthemums; it's a widespread problem for various crops, causing symptoms from yellowing and browning to wilting and plant death, which can be economically damaging to agriculture.[1]

In attempting to unravel the complexities of this disease and it's effect on different cultivars, our study tracks various phenotypic traits. These include parameters like if and when the plant starts to show significant symptoms of the disease and the severity of the disease on each plant which is indicated as a percentage. By closely monitoring these traits, the aim is to develop a comprehensive understanding of the disease's impact on chrysanthemum plants and study the factors contributing to the susceptibility of the cultivars.

1.2 Research Questions

Guided by the aforementioned objectives, this study seeks to address the following research questions:

1. *Which parameters most accurately measure plant susceptibility or resistance?*
2. *How many plants should be tested in such a test?*
3. *How can plants be consistently classified into resistant, tolerant, or susceptible based on these parameters?*

2 Methods

2.1 Study Design & Measurements

The study conducted at *Wageningen University* was meticulously designed to assess the sensitivity of various chrysanthemum cultivars to different pathogens and soil types. Spanning a production cycle of approximately 7 weeks, the experiment took place from May to July 2023.

Data collection in the study was twofold: detailed *Disease Scoring* and *Plant Measurements*.

Disease Scoring involved recording the progression and severity of disease symptoms at specific time points, while Plant Measurements focused on evaluating the sensitivity of each cultivar as determined by different breeders.

In total, the study encompassed 1051 plants, 34 treatments, and 15 distinct cultivars. The treatments were organized in a way that the first six focused on the *Kennedy* cultivar. Treatments 7 through 34 were grouped into pairs, each featuring the same type of pathogen used: either a *Negative Control* or *Conidia* but differing in kind of cultivar. All these treatments utilized Potting Soil.

The *Kennedy* cultivar treatments were a bit more complex. Treatments 4 and 6 were similar to the latter groups, using *Potting Soil* with *Negative Control* and *Conidia* as pathogens, respectively. However, treatments 1, 2, 3,

and 5 had various unique settings, leading to their exclusion from further analysis to maintain consistency.

Each treatment was replicated across four blocks, with each block containing eight plants. The sole exception was the *80.087.000* cultivar, which was only planted in blocks 3 and 4.

2.2 Missingness

The dataset contained a few missing values across key variables such as Disease Scoring and Plant Length. Given their minimal presence, these missing entries were excluded from the analysis to ensure data integrity and accuracy. Moreover, it is assumed that they are of the Missing Completely At Random (MCAR) type.[2]

2.3 Data Exploration

The characteristics of the plants are illustrated in Table 1 and Table 2. The dataset primarily comprises plants categorized as *Very Sensitive* and *Sensitive*, showing a predominant tendency for rapid disease onset among the studied specimens.

In the analysis of the pathogen variable, only the *Kennedy* cultivar was subjected to the *Microsclerotia* pathogen. The *Negative Control* category, encompassing 487 plants, is excluded from the analysis due to the absence of affected plants, rendering its inclusion non-contributory to the study's primary focus on disease susceptibility and resistance.

In Figure 1, an analysis of the distribution of plants among different cultivars reveals an imbalance. Notably, the *Kennedy* cultivar is represented almost threefold compared to other cultivars in the experiment, indicating an outstanding focus on this specific variant. This disproportion highlights the *Kennedy* cultivar's central role in the study. However, beside *Kennedy* and *80.870.000* cultivars, it seems that the dataset is quite balanced.

Table 1: Summary of Continuous Variables

Variable	**Minimum**	**Mean**	**Maximum**
Plant Length in Week 5 (cm)	23.7	61.3	83.8
Relative Growth Rate (RGR) in Week 5 (cm)	6.5	49.1	73.6
Plant Length in Week 9 (cm)	12.4	72.0	98.3
RGR in Week 9 (cm)	0.4	59.8	87.1

Table 2: Summary of Categorical Variables

Variable	Category	Number of Plants
Sensitivity	Very Sensitive	191
	Sensitive	128
	Reasonably Sensitive	64
	Intermediate with Damage	63
	Relatively Tolerant	32
	Tolerant	32
	Other	45
Soil Type	Regular Soil	63
	Potting Soil	492
Pathogen	Microsclerotia	63
	Conidia	492

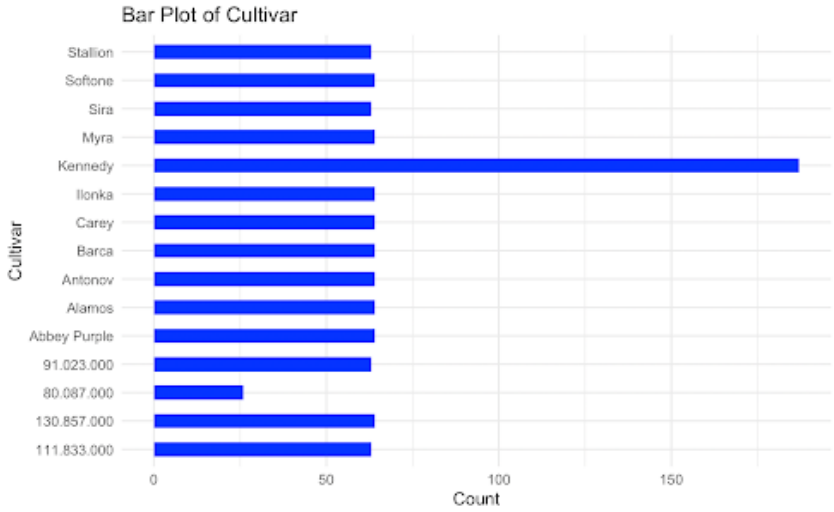

Figure 1: Distribution of Cultivars in the Dataset.

Figure 2 displays the average number of diseased leaves for each cultivar. The *Kennedy* cultivar, which constitutes the largest portion of the study's sample, shows a lower average of affected leaves compared to some other cultivars. This outcome is somewhat unexpected. Given the higher quantity of *Kennedy* plants, one might anticipate a correspondingly higher incidence of leaf disease. However, the data indicates a relatively lower occurrence of leaf affliction in the *Kennedy* cultivar. This might be a potential resistance to leaf disease in these plants.

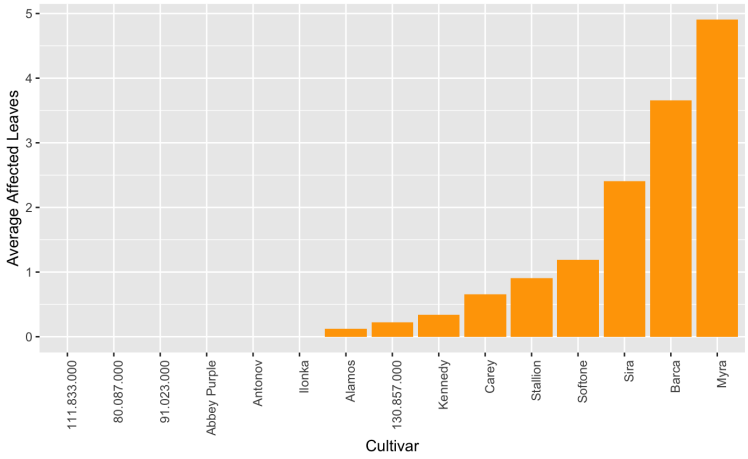

Figure 2: Average Number of Affected Leaves per Cultivar in Week 4

2.4 Statistical Analysis

In addressing the research questions, the initial step involves visualizing the mean affected rates for each cultivar across all blocks at various time points.[3] This visualization aims to assess any disparities between the blocks. For instance, the plots of *Kennedy* and *Myra* cultivars are illustrated in Figure 3 and Figure 4. While these plots are presented, the corresponding plots for other cultivars have been included in the Appendix.

Preliminary observations suggested minimal differences in disease severity between the blocks. However, to substantiate this observation, a more precise statistical approach is required.

2.4.1 First Approach: Using Methods that Measure the Plant's Susceptibility

Area Under the Curve (AUC) is employed as a key analytical tool. AUC is a method used in statistics to measure the extent of a variable's deviation over a period.[4] In this context, it quantifies the progression of disease severity across the time span of the study. The objective is not solely to identify the most sensitive or resistant cultivars but to understand the broader parameters influencing plant susceptibility or resistance. By calculating AUC for each cultivar within each block, it is possible to gain a comprehensive view of the disease's impact over time. However, this should be taken into account that the study has limitations in the available data. For instance, the variation in factors like *Soil Type* and *Pathogen*, which is explored for the *Kennedy* cultivar, is less comprehensive for other cultivars. Because of this reason, they are excluded from further analysis in order to have solid results.

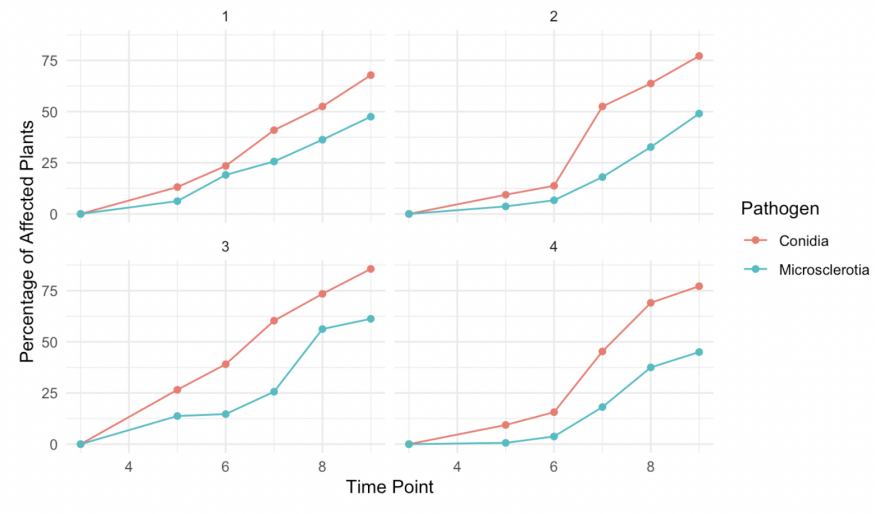

Figure 3: Mean Affected Rate of Plant over Time by Block and Pathogen for Kennedy

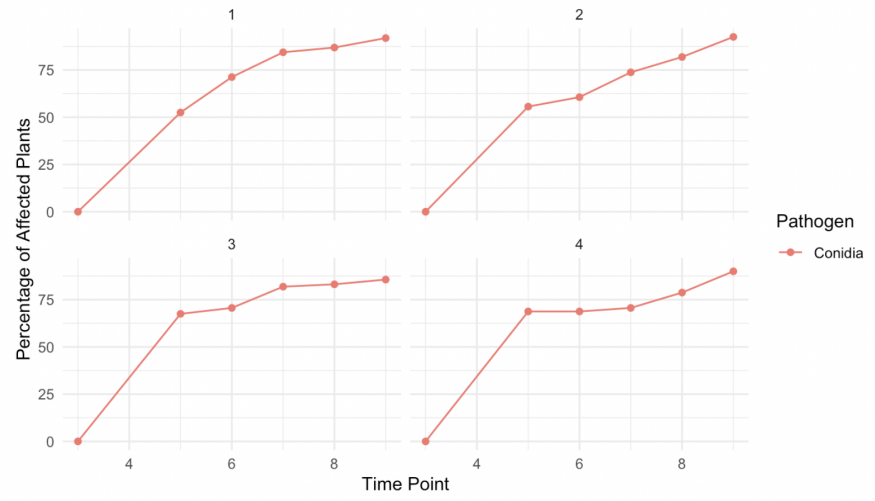

Figure 4: Mean Affected Rate of Plant over Time by Block and Pathogen for Myra

To statistically validate the observed patterns and assess the significance of block and cultivar effects on disease severity, an *Analysis of Variance* (ANOVA) is conducted using the AUC values. This statistical test is crucial in determining whether variations in disease severity are due to differences in cultivars or blocks.

ANOVA is a robust statistical method used to compare means across multiple groups and is particularly effective in experiments where multiple factors are at play, as in this study. The results of ANOVA would provide insights into whether specific cultivars or growing conditions (as represented by different blocks) have a statistically significant impact on the progression and severity of the disease.

Subsequent to the ANOVA, the study utilizes the `emmeans` and `cld` functions in R, commonly applied in post-hoc analysis following ANOVA.

The `emmeans` function is employed to ascertain the *estimated marginal means* (EMMs or least-squares means) for each cultivar.[5]

The critical elements of this result include:

- ***emmean***: The estimated mean for AUC for each cultivar, providing a measure of disease impact over time.
- ***SE* (Standard Error)**: A measure of the accuracy of the emmean, indicating the variability of the estimate.
- ***df* (Degrees of Freedom)**: Reflecting the number of independent data points used in calculating the mean.
- ***lower.CL*** and ***upper.CL***: The lower and upper bounds of the 95% Confidence Interval, showing a range within which the true mean is expected to lie with 95% certainty.

The *Cumulative Link Model*(cld) function's output categorizes cultivars into distinct groups, based on statistical significance comparisons of their means:[5]

Every element in this function is defined exactly the same as `emmeans` function.

This approach allows for a nuanced understanding of how cultivars compare with each other in terms of disease severity.

The study further incorporated the `contrast` function in R, focusing on pairwise comparisons of the estimated marginal means for each cultivar against the reference cultivar *Kennedy*. This specific post-hoc analysis method allows for a detailed examination of how each cultivar's mean AUC compares with that of *Kennedy*.

Contrast Analysis: Each line in the contrast output represents a comparison between the mean AUC of *Kennedy* and another cultivar. For instance, a contrast like *111.833.000 - 130.857.000* means the EMM of cultivar *111.833.000* is being compared with *130.857.000*.

- ***Estimate***: This value represents the estimated difference in mean AUCs between the two cultivars. A negative estimate indicates that the mean AUC of the first cultivar is lower than that of the second.

- ***t.ratio***: This is the test statistic associated with each pairwise comparison, providing a basis for testing the null hypothesis of no difference between the cultivars' means.
- ***p.value***: This is the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true. A small p-value (commonly less than 0.05) suggests a statistically significant difference between the AUCs of the two cultivars compared.

The application of the *Tukey* method for adjusting p-values in these multiple comparisons helps control the risk of *Type I errors* (false positives).

2.4.2 Second Approach: Using Parameters in the Data to Capture Plant's Susceptibility

In addressing the challenge of assessing plant susceptibility, the study pursues a predictive approach focusing on the final stage of the disease progression. This entails forecasting the mean of affected plants rate at week 9 based on earlier stages of the disease and additional plant growth metrics.

The methodology involves dividing the data into training and testing sets. The aim is to develop a predictive model where the response variable is the mean of affected plants rate at *Week 9*. As predictors, the model incorporates mean rates from earlier weeks (*Week 5* to *Week 8*) and other variables, namely the mean *Plant Length* and *Relative Growth Rate* at *Week 5*.

Linear regression is employed as the statistical tool for prediction.[6] This choice is substantiated by several factors:

1. Linear regression offers an easily interpretable model structure, where the relationship between predictors and the response variable is straightforward.
2. The presumption in this scenario is that the relationship between the earlier week rates and the final week's rate is linear.
3. Linear regression is particularly effective for prediction when the relationship between variables is not overly complex and the data adheres reasonably well to the assumptions of linear modeling.

3 Results

3.1 First Approach: Using Methods that Measure the Plant's Susceptibility

In this section, the research questions posed at the outset of the study will be answered. It begins by examining the results of ANOVA, which assesses the significance of cultivar and block variations on the AUC. The details of this analysis are presented in Table 3.

The F-value for *Cultivar* is 48.6, with a p-value of less than $2e^{-16}$ indicates a statistically significant difference among cultivars in terms of their AUC values. In contrast, the F-value for *Block* is 2.3 with a p-value of 0.08. This result shows that the differences in AUC values across various blocks are not statistically significant, as the p-value exceeds the standard significance threshold of 0.05. The lack of a significant block effect implies that environmental or locational factors represented by different blocks do not substantially influence the plants' disease susceptibility as measured by AUC.

Table 3: Summary of ANOVA before Simulating Blocks

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cultivar	14	992782	70913	48.6	$2e^{-16}$ ***
Block	3	10218	3406	2.3	0.08
Residuals	40	58303	1458		

Table 4: Cultivar's classification based on AUC

Cultivar	Amount of AUC
Sira	406.3
Myra	364.7
Barca	283.4
Stallion	257.7
Kennedy	236.6
130.857.000	228.2
Softone	200.5
Carey	128.7
Ilonka	109.4
Antonov	39.8
Alamos	33.7
Abbey Purple	27.1
111.833.000	16.3
91.023.000	5.5
80.087.000	1.25

Addressing the First and Third Research Questions with AUC Analysis:

Upon establishing that block effects do not significantly influence AUC, a table reflecting the average AUC on blocks for each cultivar is compiled. The analysis uses AUC as a key indicator, enabling a systematic classification of chrysanthemum cultivars into categories of resistance, tolerance, and susceptibility. These categories are determined based on their average AUC values, which correspond to the extent of disease impact observed in each cultivar. This method presents a clear approach for consistently categorizing cultivars, facilitating focused efforts in breeding and research for enhancing disease resistance in chrysanthemum plants. The cultivars are organized in descending order in Table 4. A higher AUC signifies greater disease impact, reflecting a higher susceptibility, while a lower AUC indicates relative resistance. Cultivars such as *Sira* and *Myra* are at the top of the list with AUCs of 406.3 and 364.7, respectively, categorizing them as highly susceptible. Conversely, *80.087.000*, with the lowest AUC of 1.25, is identified as highly resistant.

Post-hoc Analyses on Cultivar Susceptibility:

Beside AUC method, the post-hoc tests following ANOVA, the `emmeans` and `cld` functions, propose valuable insights into the susceptibility of different chrysanthemum cultivars. Detailed results of these analyses are provided in the Appendix due to space constraints.

The `emmeans` analysis indicates the estimated average AUC for each cultivar. For example, the cultivar *Sira* exhibits the highest average AUC estimate, showing a higher susceptibility compared to other cultivars. The confidence intervals associated with these estimates mention their precision, with narrower intervals denoting more precise estimates.

The `cld` analysis facilitates direct comparisons between cultivars. Cultivars grouped under the same letter, such as *91.023.000* and *111.833.000*, do not show significant differences in their disease impact, as indicated by their grouping under *1*. These groupings are instrumental in identifying cultivars with statistically similar disease impacts. Cultivars assigned unique grouping letters significantly differ from others within their respective groups. In summary, the `emmeans` results offer an in-depth view of each cultivar's resistance or susceptibility to disease, while the `cld` results provide a more straightforward method for comparing cultivars against each other.

Moreover, in the analysis, the `contrast` function is employed to conduct pairwise comparisons of the estimated marginal means for each cultivar against the *Kennedy* cultivar, which is selected as the reference due to its central role in the study. This approach also provides insights into the relative disease impact of each cultivar in comparison to *Kennedy*. Statistically significant differences, indicated by low p-values, reveal that the AUC for certain cultivars markedly differs from that of *Kennedy*. These differences highlight the variance in disease susceptibility or resistance among the cultivars. For instance, instances where *Kennedy* appears in the comparison with significant p-values (e.g., *Kennedy - Myra*, *Kennedy - Sira*) suggest a notable divergence in disease impact for these cultivars compared to *Kennedy*.

Addressing the Second Research Question with Analyzing Different Number of Blocks:

In the continuation of the study, simulations with varying numbers of blocks are conducted to assess how changes in the number of blocks might affect the detection of significant effects in the analysis which is done

so far.[7]

Initially, the analysis is replicated with a small number of blocks, such as 2 or 3, but the outcomes remained unchanged, with no significant effects observed. However, a marked shift is noted when the number of blocks is increased to 10 or 11 as it shows in Table 5. In these scenarios, the block effect becomes significant, suggesting that environmental or location-specific factors represented by different blocks exert a notable influence on the AUC. This finding indicates that with a lower number of blocks, such as 4, certain environmental effects that could be significant may not be adequately captured.

To corroborate these findings, further analyses, including estimated marginal means, pairwise comparisons, and cld, are performed for the different block scenarios. These detailed analyses have been included in the Appendix.

Table 5: Summary of ANOVA after Simulating Blocks

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cultivar	14	992782	70913	48.6	$2e^{-16}$ ***
Block	10	25948	2595	2.0	0.04 *
Residuals	33	42574	1290		

3.2 Second Approach: Using Parameters in the Data to Capture Plant's Susceptibility

The study's second approach focuses on predicting the susceptibility of plants by forecasting the mean of final disease severity rate at *Week 9*. This prediction is based on the disease severity rates observed at earlier timepoints (*Week 5* to *Week 8*) and additional plant-related variables measured up to *Week 9*, such as *Plant Length* and *Relative Growth Rate* at *Week 5*.[8] The results of this predictive analysis are as follows:

In the pursuit of identifying the most effective predictors for predicting plant susceptibility in *Week 9*, the study first considers all available variables. It is observed that data from *Week 6* and *Week 8* shows significant predictive power. Notably, *Week 8*'s significance is anticipated as it is closest to *Week 9*. However, to optimize both cost and precision, a **backward selection** approach is employed, leading to a focus on earlier timepoints. The result is peresented in Table 6.

Table 6: Summary of Model's Precision

Model	MSE(Mean Square Error)	RMSE(Root Mean Square Error)
Week 5	294.5	17.1
Week 6	267.7	16.3

Upon further analysis, the model utilizing *Week 6* data demonstrates superior performance compared to the *Week 5* model, as evidenced by lower MSE and RMSE values. This shows that utilizing *Week 6* data for

predicting *Week 9* plant statuses yields more accurate results than relying on *Week 5* data. This finding is in line with the expectation that predictions closer to the target week are generally more precise.

For the last part, the **Welch Two Sample t-tests** are conducted to compare the mean of affected plants rate at *Week 5* and *Week 6* between two groups of cultivars: *Resistant* and *Susceptible*.[9] These groups are defined based on their Mean AUC values, which are indicators of the cultivar's sensitivity to the disease over time. The t-test is a statistical method used to determine if there is a significant difference between the means of two groups, which is particularly useful in this context to understand if the disease progression differs significantly between resistant and susceptible cultivars. The result is provided in Table 7.

Table 7: Welch Two Sample t-Test

.	**Mean in Group Resistant**	**Mean in Group Susceptible**	**p-value**
Week 5	2.0	36.5	$5.373e^{-09}$
Week 6	3.9	42.6	$8.345e^{-11}$

The clear distinction in the mean of affected plants rate between resistant and susceptible cultivars at these early stages indicates that early intervention based on these data points could be effective in managing the disease and making informed decisions about cultivar resistance in future planting cycles. Cultivars showing higher affected plants rate at these stages are more likely to end up being classified as susceptible by *Week 9*. Also, the significant differences found in the t-tests validate the approach of using mean of affected plants rate as a criterion for classifying cultivars into resistant or susceptible categories.

4 Conclusion

This study aimed to assess the susceptibility of different plant cultivars to chrysanthemum wilt by analyzing the disease progression over time. The study employed a multi-faceted approach, including statistical analysis and modeling, to address the research questions.

The **first research question** addressed the identification of effective parameters for measuring plant susceptibility to a specific disease. Two distinct approaches were employed to answer this question.

The first approach involved statistical analyses, including ANOVA and subsequent post-hoc tests such as *emmeans* and *pairwise comparisons*. These analyses were utilized to assess the significance of different factors, particularly the impact of blocks, in determining plant susceptibility. Additionally, the *Area Under the Curve* method was employed to create graphical representations of the mean affected plant rates for each cultivar within each block. Subsequently, mean AUC values were calculated, considering the non-significant effect of blocks.

The second approach focused on predictive modeling, aiming to forecast mean affected plant rates at later time points based on data from earlier stages. This approach aimed to ascertain whether it is feasible to predict disease progression and plant sustainability using information gathered from earlier time points.

In conclusion, the study suggests that both the utilization of data from *Week 6*, with AUC as a measurement tool, and the application of ANOVA and post-hoc analyses to evaluate plant sustainability are parameters and tools for assessing plant susceptibility to the disease.

The **second research question** addressed the optimal number of plants to be included in the test. The findings indicate that increasing the number of blocks in the study has an impact on the reliability of the results, with the effect of blocks becoming statistically significant. It is important to mention that employing more blocks while reducing the number of plants in the study may be a preferable approach.

The **third research question** focused on the consistent classification of plants into resistant, tolerant, or susceptible categories based on the identified parameters. One approach is to utilize the AUC results and create a table that arranges cultivars in ascending or descending order based on their AUC values, allowing for the identification of those with the highest susceptibility and resistance. Additionally, classification can be performed using data from *Week 6* as a predictor for later weeks, where cultivars with mean affection rates ranging from 0-20% are classified as resistant, those with mean rates between 20-50% as tolerant, and those above 50% as susceptible.

5 Discussions & Implications

The findings of this study bear significance for agricultural practices and research endeavors. These results offer valuable intuition that can aid both farmers and researchers in the selection of cultivars with greater resistance to the disease, potentially leading to a reduction in the necessity for extensive disease control measures. This is also good to mention that based on the fact, it is concluded that the results from earlier time points can be used to predict plants susceptibility in the future, this can help farmers to save costs of maintenance on the affected plants.

However, it's worth noting that the reliability of these results could be further enhanced with the inclusion of additional data from earlier time points, which would contribute to the development of more robust predictive models with a broader range of predictors.

A substantial limitation of this study pertains to its sample size. The inclusion of a larger dataset comprising a greater variety of cultivars, replicates, and variations in soil types and pathogens would enhance the generalizability of the findings, proposing a more comprehensive understanding of susceptibility factors.

Additionally, future research efforts could explore additional predictors, such as genetic markers or environmental factors, to further refine disease progression models and improve their accuracy.

References

[1] T. BARNETT, *Chrysanthemum verticillium wilt: Learn about mum verticillium control*, 2021.

[2] S. van Buuren, *Flexible Imputation of Missing Data*. Second Edition, 2018.

[3] A. A. G. Kwanchai A. Gomez, *Statistical Procedures for Agricultural Research*. John Wiley Sons, 1984.

[4] C. Bento, "Roc analysis and the auc — area under the curve," 2022.

[5] a. L.-S. M. R. p. Lenth R emmeans: Estimated Marginal Means, 2020. [Online]. Available: `https://CRAN.R-project.org/package=emmeans`.

[6] J. Brownlee, "Linear regression for machine learning," *Machine Learning Algorithms*, 2023.

[7] T. M. D. et al., "Simulation-assisted machine learning," *Bioinformatics*, 2019. [Online]. Available: `https://doi.org/10.1093/bioinformatics/btz199`.

[8] R. T. et al., *An Introduction to Statistical Learning*. 2013.

[9] [Online]. Available: `https://en.wikipedia.org/wiki/Welch%27s_t-test`.

Appendix

GitHub link for code and plots:

https://github.com/Khalajzadehn/Statistical-Consulting/blob/main/Statistical-Consulting.pdf

Cumulative Link Models

A cumulative link model is a model for ordinal-scale observations, i.e., observations that fall in an ordered finite set of categories. Ordinal observations can be represented by a random variable Y_i that takes a value *j* if the *i*th ordinal observations falls in the *j*'th category where $j = 1, ..., J$ and $J \geq 2$. A basic cumulative link model is

$$\gamma_{ij} = F(\eta_{ij}) \quad \eta_{ij} = \theta_j - {x_i}^T \beta \quad i = 1, ..., n \quad j = 1, ..., J-1 \tag{1}$$

Where

$$\gamma_{ij} = P(Y_i \leq j) = \pi i1 + ... + \pi_{ij} \quad with \quad \sum_{j=1}^{J} \pi ij = 1 \tag{2}$$

are cumulative probabilities. π_{ij} is the probability that the *i*th observation falls in the *j*th category, η_{ij} is the linear predictor and ${x_i}^T$ is a p-vector of regression variables for the parameters, β without a leading column for an intercept and *F* is the inverse link function. The thresholds (also known as cut-points or intercepts) are strictly ordered:

$$-\infty \equiv \theta_0 \leq ... \leq \theta_{J-1} \leq \theta_J \equiv \infty \tag{3}$$

Linear Regression

Consider the model function

$$y = \alpha + \beta x \tag{4}$$

which describes a line with slope β and y-intercept α. In general such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the errors. Suppose we observe n data pairs and call them $(x_i, y_i), i = 1, ..., n$. We can describe the underlying relationship between y_i and x_i involving this error term ϵ_i by

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{5}$$

This relationship between the true (but unobserved) underlying parameters α and β and the data points is called a linear regression model.

Area Under the Curve (AUC)

The area under the curve can be calculated through three simple steps. First, we need to know the equation of the curve(y = f(x)), the limits across which the area is to be calculated, and the axis enclosing the area. Secondly, we have to find the integration (antiderivative) of the curve. Finally, we need to apply the upper limit and lower limit to the integral answer and take the difference to obtain the area under the curve.

$$Area = \int_a^b y.dx = \int_a^b f(x).dx = g(b) - g(a) \tag{6}$$