

UNIVERSITY OF NOTTINGHAM



The University of
Nottingham

UNITED KINGDOM • CHINA • MALAYSIA

SCHOOL OF MATHEMATICAL SCIENCES

Statistical Shape Analysis of Wheat Root Systems

Andrew Hyde

A thesis submitted to the University of Nottingham for the
degree of
PHD

DECEMBER 2017

ABSTRACT

The roots of a plant play a vital role in its growth and development, but due to practical difficulties of observing underground roots, the study of their shape has long been neglected. Recent advances in CT imaging technology have allowed for accurate non-destructive imaging of root systems in soil. This technique has formed the basis of the FutureRoots project.

The main challenge with analysing the shape of a plant root system is that they have varying topological structure, so traditional shape analysis methods cannot be applied.

In this thesis, we develop three approaches for analysing wheat root systems.

The first approach involves measuring a set of pre-chosen root traits, and analysing this set using conventional statistical methods. This approach is effective but may miss potentially important shape information and the large number of measurable traits reduces the potential power of statistical tests.

The second approach is to perform pairwise comparisons based on the Hausdorff Metric and use Multidimensional scaling to reduce a large set of pairwise comparisons to a dataset which can be analysed with conventional statistical methods. This approach can detect and test for overall shape differences but can fail to detect subtle differences.

The third approach is to apply the Persistent Homology technique from Topological Data Analysis, which is designed to find underlying topological differences between two shapes. This method successfully finds differences but it is difficult to interpret the results.

We will apply these three techniques to simulated data and a real life dataset.

In addition, because of experimental considerations, the wheat roots had to be unnaturally constrained to a small area so we have developed a method to estimate how they would have grown unconstrained.

ACKNOWLEDGEMENTS

I'd like to thank my supervisors Professor Ian Dryden and Professor Andrew Wood for supervising my PhD and giving invaluable advice and constant encouragement. I'd like to thank the FutureRoots team for providing their expertise and assistance on Biological and Computer Imaging, as well as the experience of working with a fine team of researchers. I must also thank the European Research Council (ERC) for provided funding for both the CPIB FutureRoots program and my PhD research.

On a personal note, I'd like to thank my mum, dad and brother for their amazing love and support over the last four years, I could never have done this without them.

CONTENTS

1	INTRODUCTION	9
1.1	Root Shapes	9
1.2	CPIB Future Roots Project	10
1.2.1	Growth Paper Method	12
1.3	Thesis Overview	13
2	LITERATURE REVIEW	16
2.1	Biological Background	16
2.1.1	Terminology	16
2.2	Shape Analysis	17
2.2.1	Procrustes Analysis	17
2.3	Multidimensional Scaling	19
2.4	Linear Discriminant Analysis	20
2.5	Generalized Linear Model	21
2.6	Angular Statistics	22
2.6.1	Simple Statistics	22
2.6.2	Tests for Uniformity	24
2.7	Combining the results of hypothesis tests	25
2.8	Hausdorff Distance Metric	28
2.9	Topological Data Analysis	30
2.9.1	Introduction	30
2.9.2	A Simplified Explanation of Persistent Homology	31
2.9.3	A technical overview of Persistent Homology	37
2.9.4	Types of Filtration	42
2.9.5	Distance between Persistence Diagrams	45
3	ROOT SYSTEMS	48

Contents

3.1	Introduction	48
3.2	Approach	49
3.3	General structure of wheat root systems	50
3.4	Root System Shape	50
3.5	RSML	51
3.6	Compensating for Plant Overgrowth	51
3.6.1	Unravelling method	52
3.6.2	Handling Laterals and Sub-Roots	53
3.7	Instability	60
3.8	Appropriateness	61
3.9	Conclusion	61
4	ANALYSIS OF MEASURED TRAITS	62
4.1	Introduction	62
4.2	Simple Summary Statistics	62
4.3	Statistics of Roots	63
4.3.1	Length	63
4.3.2	Extreme Statistics	64
4.3.3	Convex Hull	64
4.3.4	Angle Statistics	65
4.3.5	Topology Statistics	66
4.3.6	Miscellaneous Statistics	67
4.4	Analysis of Traits of Root Systems	67
4.4.1	Tukey's Honest Significance Differences Test	68
4.5	Analysis of Traits of Individual Root Strands	70
4.5.1	Weighting	72
4.5.2	Conclusion	73
5	PAIRWISE DISTANCE MEASUREMENT	74
5.1	Pairwise Distance Measurement	74
5.1.1	Root System Shape Space	76
5.1.2	Adapting metrics to ignore nuisance differences	76

Contents

5.1.3	Finding the minimal shape distance	77
5.2	Distances between Line segments	79
5.2.1	Segment Reduction	79
5.2.2	Minimum distance	79
5.2.3	Total Minimum Distance	80
5.3	Hausdorff Distance	81
5.3.1	Alternative	82
5.4	Multidimensional Scaling	83
5.4.1	Choice of MDS k value	83
5.5	Analysis Approach	84
5.6	Conclusion	85
6	TOPOLOGICAL DATA ANALYSIS	87
6.1	Persistent Homology	87
6.2	Conventional Persistent Homology	89
6.2.1	Converting root systems to point clouds	89
6.2.2	Delaunay Triangulation	89
6.3	Line-based Filtration	91
6.4	Alternative Point Cloud Representations	91
6.5	Comparing Persistence Diagrams	92
6.5.1	Alternative approaches	92
6.6	Limitations of TDA	93
6.7	Interpretation of Persistent Homology results	93
6.7.1	Visualizing Homology Classes	94
6.8	Visualizing Locations of Holes in the Vietoris-Rips	94
6.9	Conclusion	98
7	SIMULATION STUDY	99
7.1	Simulated Dataset	99
7.2	Common Cases	99
7.3	Pairwise Distance Approach	100
7.3.1	Elimination of Angle Differences	100

Contents

7.3.2	Hausdorff Distance	101
7.3.3	Pseudo Hausdorff Distance	105
7.4	Topological Data Analysis	111
7.5	Choice of k for MDS	113
7.6	Conclusion	114
8	PRACTICAL DATASET	115
8.1	Dataset	115
8.2	Relationships with Summary Statistics	117
8.2.1	Root Count	118
8.2.2	Multi-Analysis of Various Statistics	118
8.3	Comparison of Factor Means	122
8.4	Analysis of individual root strand properties	137
8.4.1	Root Strand Length	137
8.4.2	Multi-Analysis of Various Statistics	138
8.5	Results of Trait Analysis	143
8.6	Other Properties	146
8.6.1	Horizontal Growth	146
8.7	Pairwise Distance Approach	146
8.8	Topological Data Analysis	156
8.9	Conclusion	161
9	CONCLUSIONS AND FURTHER WORK	162
9.1	Results of our Investigations	162
9.2	Further work	165
9.2.1	Time Series Analysis	165
9.2.2	Soil Analysis	166
9.2.3	QTL Analysis	167

INTRODUCTION

1.1 ROOT SHAPES

In this thesis we will be exploring methods for analysing the shape and structure of root systems belonging to wheat plants.

Wheat has been a staple food for much of the world throughout history, and is vital to supporting the ever growing population of the world. In the middle of the 20th Century, in response to a predicted food crisis, Norman Borlaug and others worked to develop new varieties of wheat and rice that had a higher yield and were resistant to disease. These initiatives, later termed the "Green Revolution", have allowed food production to keep pace with population (Lynch, 2007).

Despite their success, there were drawbacks to the crop varieties produced by the Green Revolution. They were designed to be grown in soil that was highly fertilised and well irrigated, which is unfortunately not the case in many developing countries. Low fertility and drought have had detrimental effects on the yield of these plants, that have offset the gains made by the Green Revolution. There is a need to develop new varieties of crops that are drought resistant and can produce a high yield even in low fertility soil. (Den Herder et al., 2010).

The Green Revolution primarily focused on improving "Shoot biomass" and "Seed Yield", which are both "above ground" traits of plants that are easily observed. Traits that were "below ground" such as the roots and their

1.2 CPIB FUTURE ROOTS PROJECT

ability to draw nutrients and water from the soil were much more difficult to measure and had to be ignored.

Measuring traits of the roots within the soil is a difficult problem. Soil is typically dense, opaque and the deeper sections are compacted by the weight of the soil above. Additionally, the roots of wheat plants are very thin and delicate, so even a slight disturbance of the soil can break them, potentially killing the plant. Analysing roots non-destructively has not been possible until recently.

1.2 CPIB FUTURE ROOTS PROJECT

The Future Roots project is an ERC-funded project based at the "Centre for Plant Integrative Biology" (CPIB) in the University of Nottingham. It was created to analyse the root systems of plants using non-invasive x-ray techniques (Mairhofer et al., 2013). Its goal is to analyse the root systems of wheat and other crops in order to understand and improve their performance in drought conditions and other hostile environmental conditions.

The major issue in analysing the root systems of plants is that it is very difficult to observe them in a way that is not destructive. Plant roots are delicate and removing them from soil will disrupt their structure and likely kill the plant.

The Future Roots project aims to non-invasively image plant root systems using a CT scanner in order to detect the location of plant roots in soil. This produces cross-sectional views of the soil, an example of which can be seen in Figure 1. The images of the soil are then analysed to track the growth of the roots from the plant stem, as seen in Figure 2.

In addition, the Future Roots project has developed a state of the art automated greenhouse that can grow, monitor and scan plants on a day-to-day basis. This is being used to run experiments to identify phenotypes and eventually help improve growth in plant strains. The project will be

1.2 CPIB FUTURE ROOTS PROJECT

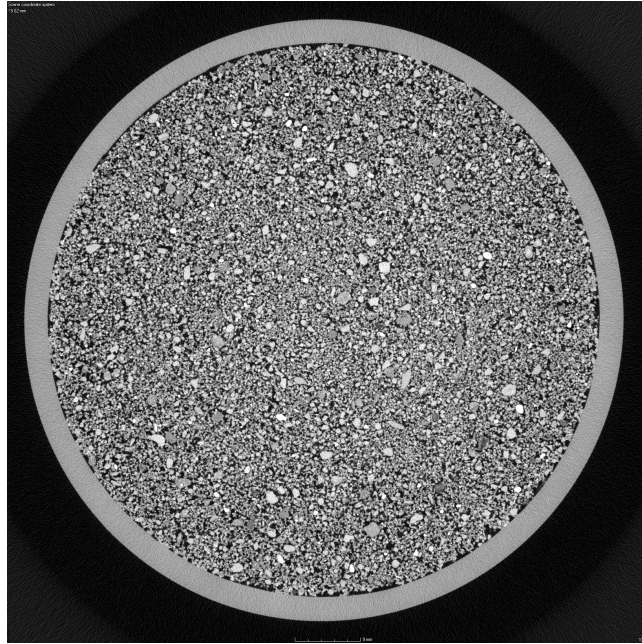


Figure 1: Cross-section CT scan of a soil column.

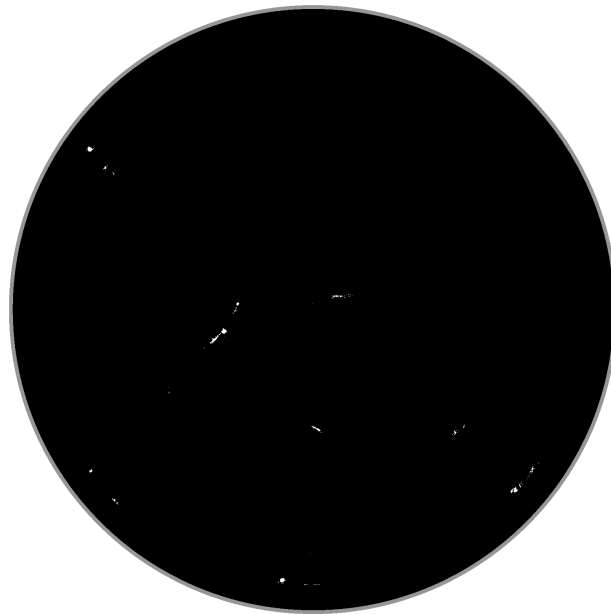


Figure 2: The same soil column in Figure 1, with pixels that likely contain roots marked in white. The roots are small and difficult to see in the noise of the soil.

1.2 CPIB FUTURE ROOTS PROJECT

primarily concerned with wheat. For a visual example of a wheat plant's root system see Figures 3 and 4.

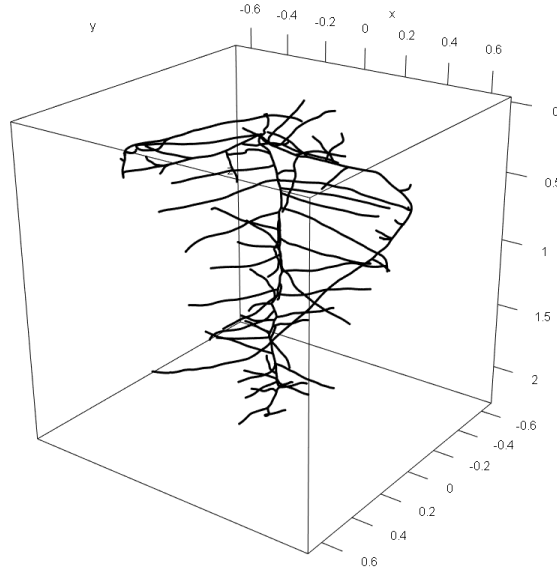


Figure 3: Plot of a root system from the Savannah line grown in high nitrate soil.

1.2.1 *Growth Paper Method*

In addition to the 3D datasets created by scanned soil columns, we have a collection of 2D scans of plants that have been created using a "growth paper" method. In this method, plants are grown on a paper that contains water and nutrients and is designed to simulate a soil medium. The plant roots are constrained to only grow along the paper in a single flat plane. Once grown, the plants can be easily photographed and their root structure extracted using image analysis. While the focus of this thesis will be on analysing the 3D root systems, many of the methods used can be applied to 2D root systems as well, with some possible modification.

1.3 THESIS OVERVIEW

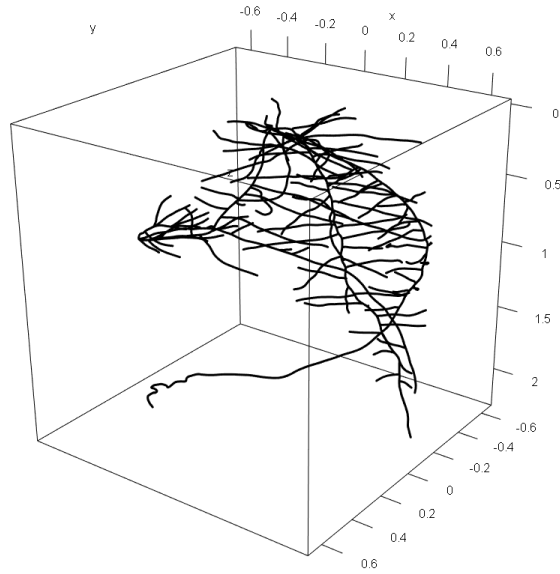


Figure 4: Plot of a root system from the Rialto line grown in low nitrate soil.

1.3 THESIS OVERVIEW

We'll now give a brief overview of the thesis and the contents of each chapter.

- Literature Review

This gives background information on the biological and statistical techniques that we will be using in this Thesis. The sections on the Hausdorff distance (Section 2.8) and Topological Data Analysis (Section 2.9) will both be used extensively during our thesis.

- Root Systems

Here we describe the root systems that we have scanned and how they are structured. We will outline the approach that we will be using to analyse these root systems.

We also describe an algorithm for 'unravelling' roots that have been unnaturally constrained by the columns in which they were grown.

- Analysis of Measured Traits

Here we will use our first approach which is to measure 'traits' from our root system. This approach is the conventional approach that biologists will use to analyse roots.

- Pairwise Distance Analysis

Here we will describe our second approach, which is to use metrics to identify pair-wise shape differences between pairs of root systems. This approach is much more flexible than using simple traits in that it can handle differences in structure much more easily.

- Topological Data Analysis

Here we will describe our third approach, which is based on the 'Persistent Homology' technique from Topological Data Analysis. This method transforms the data to a new form that emphasises the topological features in an object and de-emphasises geometric features.

- Simulated Study

In this chapter we apply our methods to a large simulated dataset, to see how our approaches are able to identify differences between different sets of root systems and how well they fail to identify differences between root systems if no underlying differences exist. We will also use this data to calibrate parameters of the Multi-dimensional Scaling (MDS) method that we need for two of these approaches.

- Practical Dataset

In this chapter we will apply our approaches to a real world dataset provided by the Future Roots program, testing for significant relationships with Irrigation and Genotype, and attempting to interpret the results.

- Conclusion and Further Work

1.3 THESIS OVERVIEW

Here we summarise our work and present potential avenues for future research.

LITERATURE REVIEW

2.1 BIOLOGICAL BACKGROUND

2.1.1 *Terminology*

When referring to root system, it is important to realize that there is a technical distinction between an "individual root" and a "root system". A "root system" refers to all of the roots that belong to an individual plant, whereas a "root" is just a single unbroken part of that.

"Primary roots" are large roots that are first grown by the plant in order to anchor itself into the soil. In wheat these will be larger than other roots and do not split into multiple primary roots (although they will grow small offshoots).

"Lateral roots" are the offshoot roots that grow off the primary ones. In wheat these will be much smaller than the main root.

"Line" is the technical way of referring to plants with the same genetics. Two plants belonging to the same line are genetic clones of each other.

2.2 SHAPE ANALYSIS

Shape Analysis, a relatively new topic in mathematics, is the study and comparison of geometric shapes from various populations (e.g. the skull shapes of males and females from a gorilla species). Using shape analysis one can determine whether any differences found are significant and outside the variation expected by random chance (Dryden and Mardia, 1998).

Shape Analysis for the most part is focused on situations where the general structure or topology of the shape is assumed to be fixed, for example facial recognition where each person is expected to have two eyes, two ears, one nose, etc. with variations in size, location and orientation of the various features. This assumption is not true when comparing the shapes of root systems, since typically these will have different numbers of roots arranged in different structures. This means that a lot of conventional methods and results will not be applicable to our analysis.

2.2.1 *Procrustes Analysis*

Procrustes analysis is a simple shape analysis method that uses 'landmarks' to quantify and analyse a shape. First we determine a set of k landmarks that correspond to common, unambiguous, identifiable points on the objects. We can then use the sum of squares of the differences between common shape points on the two objects as a squared distance measure. However this is not enough on its own as the distances are affected by rotation, scale and translation.

The idea of Procrustes is to rotate/scale/translate one shape to superimpose onto the other to create the best match possible (i.e. minimise the geometric distances between them). Once that is done, the distance between the two superimposed shapes is called the Procrustes distance and that is used for calculation.

Depending on the scenario you may consider 'scale' to be a significant difference. This is certainly true for analysing roots, as the size of the roots will affect the amount of energy needed to grow them as well as the amount of water and nutrients the roots can find. In this case, you simply do not correct for scale when matching the two object shapes.

For 2D shapes, the Procrustes mathematical model is each of the two shapes has a 'configuration' of landmarks, $y = (y_1, \dots, y_k)'$ and $w = (w_1, \dots, w_k)'$, where each is a vector in the k -dimensional complex space, $\mathbb{C}^k$. These vectors are assumed to be centred, i.e. $y^*1_k = w^*1_k = 0$.

We wish to match w onto y . This can be done using 'complex linear regression' to find the transformation that best matches the two together.

The complex regression model is expressed as

$$y = (a + ib)1_k + [\beta \exp i\theta] w + \epsilon, \quad (1)$$

where $a + ib$ is a constant term, θ and β are rotation/scale parameters, and ϵ is the error term. This can be simplified to

$$y = X_D A + \epsilon. \quad (2)$$

Our goal is to minimise the sum of squares of the error terms

$$D^2(y, w) = \epsilon^* \epsilon, \quad (3)$$

where $A = (a + ib, \beta \exp i\theta)$ is the similarity transformation parameter, and is designed to eliminate rotation and scale. However, we can choose to fix $\beta = 1$ to only eliminate rotation.

We will not be able use Procrustes Analysis here (or any shape method that uses landmarks) as there is not a fixed set of landmarks that can be made for root systems due to the varying number of roots.

2.3 MULTIDIMENSIONAL SCALING

We have several options for evaluating pairwise distances between two objects, allowing us to build a distance matrix representation of a dataset. However this information is not in a form that will be practical to analyse. The underlying space is complex and non-Euclidean while most statistical techniques assume the data is in a nice Euclidean space.

Multidimensional Scaling (MDS) allows us to take a set of objects with known distances, and convert them to points within a vector space that have distances that closely match the original objects (Mardia et al., 1979). More formally, we have k objects on which a distance function $\delta_{i,j}$ between the i th and j th objects is defined. We can produce a dissimilarity matrix Δ with each of these distances,

$$\Delta = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,k} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,k} \\ \vdots & \vdots & & \vdots \\ \delta_{k,1} & \delta_{k,2} & \cdots & \delta_{k,k} \end{pmatrix}. \quad (4)$$

Our goal therefore is to create a vector $\{x_1, \dots, x_k\}$ in $\mathbb{R}^k$ that minimises,

$$\min_{x_1, \dots, x_k} \sum_{i < j} (\|x_i - x_j\| - \delta_{i,j})^2. \quad (5)$$

where $\|\cdot\|$ is a vector norm.

For this report, we will be only considering the classical MDS methods, that define $\|\cdot\|$ as the standard Euclidean metric. The classical MDS is a form of Principal Component Analysis, in that if the distances, $\delta_{i,j}$, are Euclidean then MDS is equivalent to PCA.

To perform classical MDS on an n by n distance matrix D , with k dimensions,

1. Calculate the squared distance matrix $D^{(2)} = [d_{i,j}^2]$.

2. Apply 'Double centring', $B = -\frac{1}{2}JD^{(2)}J$, where J is the centring matrix, $J = I - \frac{1}{n}11'$.
3. Our matrix of MDS components is then, $X = E_k\Lambda_k^{1/2}$, where E_k is the matrix of eigenvectors and Λ_k is the diagonal matrix of k eigenvalues of B .

2.4 LINEAR DISCRIMINANT ANALYSIS

For our methods we will be using several categorical variables, and we will wish to examine how the numerical variables relate to them, and try to predict what effects changing those numerical variables will have.

Linear Discriminant Analysis (LDA) is a method for classifying data into categories (Friedman et al., 2001). For the classic version, it uses a linear combination of variables to create a discriminator function. The sample is classified into one of two classes depending on whether the classifier is greater than a given threshold.

This method can be further extended into multiple variables, where each class has a linear discriminator function associated with it and a sample can be assigned to the class whose function gives the highest value.

Our primary goal is to ensure a high classification rate (i.e. samples are correctly assigned to their actual class).

One issue with LDA is that it is prone to 'over-fitting'. If a dataset is used to 'train' the classifier, then using the same dataset to 'test' how well the classifier performs will give misleading results. It is advisable to split the dataset into a separate "training"/"testing" sets, or alternatively, use cross-validation to leave out the sample being tested from the calculation, and repeat for all samples to get an overall rate.

2.5 GENERALIZED LINEAR MODEL

The generalized linear model (GLM) is a generalization of the concept of linear models to support a wider range of distributions on observed values (Madsen and Thyregod, 2011).

We will primarily be using it to model binary response variables, however it can support any distribution from the exponential dispersion family, whose probability density takes the form,

$$f_Y(y; \theta) = c(y, \lambda) \exp \left(\frac{\theta^T T(y) - A(\theta)}{d(\lambda)} \right), \quad (6)$$

where λ is a known precision parameter, θ is a parameter related to the mean of the distribution. This family contains many common distributions including the 'normal distribution' and 'binomial distribution'.

The model states that $Y_1, Y_2, \dots, Y_n$ are mutually independent observations whose density is an exponential dispersion model with the same variance function, $V(\mu)$. In addition, we have a data matrix of n variables X and coefficients β , which create a linear predictor $\eta = X\beta$.

We can relate our linear predictor to our canonical parameter θ using a link function g ,

$$\eta = g(\mu). \quad (7)$$

This link function is different for each exponential distribution, and depends on $T(y)$, $d(\lambda)$ and $A(\theta)$.

For the binomial model, the link function is $g(\mu) = \ln \left(\frac{\mu}{1-\mu} \right)$ and the variance function is $V(\mu) = \mu(1 - \mu)$. The fitted value, $\hat{p}_i = \hat{\mu}_i = g^{-1}(X_i \hat{\beta})$, will be the estimated 'priori probability that the entry i is 1 (and not 0)'.

The model supports the Akaike Information Criterion (AIC), which we will use to measure the 'quality' of the fit and remove redundant variables. (Fox, 2000). The AIC is calculated using the maximum value of the likeli-

hood function L for a particular fit and the number of parameters k . We wish to minimize,

$$\text{AIC} = 2k - 2\ln(L). \quad (8)$$

This criterion provides a trade-off between a model's 'goodness-of-fit' and its complexity.

2.6 ANGULAR STATISTICS

One question that is of interest to the CPIB project is to see if individual roots have a preferred angle that they wish to grow in, or if the roots grow in a random uniform pattern. We will be using angular statistics to test this.

2.6.1 Simple Statistics

Mean Direction

Let $X_i, i = 1 \dots n$ be a sample of unit vectors from a $(p - 1)$ -dimensional hyper sphere and define,

$$\hat{S} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (9)$$

Let $|\cdot|$ denote the vector norm, $|v| = \sqrt{v^T v}$. If $|\hat{S}| > 0$, the mean direction of the sample is represented as

$$\bar{X} = \frac{\hat{S}}{|\hat{S}|}. \quad (10)$$

If $|\hat{S}| = 0$, then the mean direction of the sample is undefined.

An important quantity is the resultant length, $\bar{R}$ where,

$$\bar{R} = |\hat{S}|, \quad (11)$$

which can be used as a simple measure of concentration.

Von Mises-Fisher Distribution

The Von-Mises-Fisher distribution is a simple uni modal distribution for use on the $p - 1$ hypersphere (Mardia and Jupp, 2009).

It has a mean direction parameter $\boldsymbol{\mu}$, concentration parameter $\kappa \geq 0$ and density,

$$f(\mathbf{x}, \boldsymbol{\mu}, \kappa) = C_p(\kappa) \exp(\kappa \boldsymbol{\mu}^T \mathbf{x}), \quad (12)$$

where,

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2} I_{p/2-1}(\kappa)}, \quad (13)$$

and I_v is the modified Bessel function of the first kind at order v .

If the concentration parameter, $\kappa = 0$, then $f(\mathbf{x}, \boldsymbol{\mu}, \kappa) = 1/2\pi$, and the data is uniformly distributed (regardless of the value of $\boldsymbol{\mu}$).

Estimating parameters

Define,

$$A_p(\kappa) = \frac{I_{p/2}(\kappa)}{I_{p/2-1}(\kappa)}. \quad (14)$$

Then given a sample $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ from a vMF distribution with parameters $\boldsymbol{\mu}$ and κ , then the MLE for the two parameters are given by,

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n \mathbf{x}_i}{|\sum_{i=1}^n \mathbf{x}_i|}, \quad (15)$$

and,

$$\hat{\kappa} = A_p^{-1}(\bar{R}). \quad (16)$$

$\hat{\kappa}$ is well-defined because $A_p(\cdot)$ is known to be monotonic.

2.6.2 Tests for Uniformity

Rayleigh Test

The Rayleigh Test assumes that the data is distributed with a von-Mises-Fisher distribution, and tests for the κ parameter being zero. Formally we consider the hypothesis,

$$H_0 : \kappa = 0 \quad (17)$$

$$H_1 : \kappa > 0 \quad (18)$$

When $\kappa = 0$ the expected value of S defined in Equation 9 is equal to $E(S) = \mathbf{0}$. This is equivalent to $E(\bar{R}) = 0$. The test statistic for this is,

$$T = np\bar{R}^2, \quad (19)$$

where the distribution of T is,

$$T \sim \chi^2(p) \quad (20)$$

The test does have a problem however when the assumption that the data is von-Mises-Fisher distributed is violated. A non-uniform multi-modal distribution can have $\bar{R} = 0$ naturally, which this test will fail to reject.

Kuiper's Test

Given a simple two-dimensional angular data, we can use the Kuiper's test to test for uniformity since this test is invariant under circular distributions.

Given a sample of n angles from an angular distribution $(\theta_1, \theta_2, \dots, \theta_n)$

$$z_i = F(\theta_i), \quad (21)$$

$$D^+ = \max_i [i/n - z_i] D^- = \max_i [(i-1)/n - z_i] \quad (22)$$

$$V = D^+ + D^- \quad (23)$$

This test is non-parametric which means it is inherently less powerful than the Rayleigh test will be when its assumptions are met. However given that the Rayleigh's assumption of uni-modality may not be reliably met in all settings it is a useful tool to have.

Multiple subsets with different means

We will be dealing with plants that have been scanned at different angles to each other. We plan on getting a set of angle data from each one, and we want to test if these angles are uniformly distributed overall.

However, each plant has a different general orientation that is randomly distributed and is convoluted with our data. This means that we will need to compensate for this effect.

If we assume that each of the angles from each root comes from a von Mises-Fisher distribution (vMF) with a common concentration parameter, κ , but different location parameters, μ , our model is

$$x_{i,j} \sim \text{vMF}(\mu_i, \kappa), \quad (24)$$

where $x_{i,j}$ is the angle j from subset i , $i = 1 \dots N$, N is the number of subsets, $j = 1 \dots n_i$, n_i is the number of angles in subset i .

2.7 COMBINING THE RESULTS OF HYPOTHESIS TESTS

When performing a hypothesis test, a rejection threshold is chosen for the p-value that will minimize the probability of a false negative/positive (and

trade off between them). However this calculation is based on the assumption that only a single hypothesis test is performed.

When multiple hypothesis tests are performed, the chances that at least one of these tests will be below the given threshold increases with the number of tests performed. Given n independent tests with the same probability of Type-1 error, α , then when each of the null hypotheses are true,

$$P(\text{At least one Type-1 error}) = 1 - (1 - P(\text{Type-1 error}))^n = 1 - (1 - \alpha)^n, \quad (25)$$

which means that the chances of a Type-1 error will increase quickly with the number of tests, n . When doing a large amount of comparisons, we will need to compensate for this effect.

A simple way is to reduce the p-value threshold to reduce the chance of a Type-1 error to be reasonable. We can do this by inverting equation 25. This can be used to create a new acceptance value for p-values, α' ,

$$\alpha' = 1 - (1 - \alpha)^{\frac{1}{n}}. \quad (26)$$

As n increases, α' will asymptotically converge to,

$$\alpha' = -\frac{\ln(1 - \alpha)}{n}. \quad (27)$$

Fisher's Method

Fisher's method combines multiple hypothesis test p-values into a large meta hypothesis test. If we combine a large amount of tests and it fails to reject the null hypothesis, then we rule out any individual hypothesis rejections that occur.

For k hypothesis tests, where the p-value of the i th hypothesis is p_i , the test statistic T is calculated as

$$T = -2 \sum_{i=1}^k \ln(p_i), \quad (28)$$

where T is χ^2 distributed as

$$T \sim \chi^2(2k). \quad (29)$$

Holm-Bonferroni method

One disadvantage to Fisher's method is that it is an 'all or nothing' approach. It is very possible that of the tests that are classified significant, only some of them are 'false positives', and thus a more nuanced method is required.

There are multiple methods for dealing with this problem, depending on the assumptions you can make about the results and how strict or lenient you wish to be. For this project we are using the Holm-Bonferroni method, which does not require the results to be independent and ensures the probability of a type-1 error is at most α .

Let $H_1, \dots, H_n$ be a family of n hypothesis tests and $p_1, \dots, p_n$ their corresponding p values. We order the hypothesis tests by their p-value from lowest to highest, signified by $H_{(1)}, \dots, H_{(n)}$ and $p_{(1)}, \dots, p_{(n)}$.

For a given significance level α , we then find the first index k such that,

$$p_{(k)} > \frac{\alpha}{n + 1 - k}, \quad (30)$$

we then reject any hypothesis test H_i where $i < k$. If none of the hypothesis tests satisfy the requirement then we reject all the given hypotheses.

We can alternatively create an adjusted p-value for the hypothesis tests,

$$p_{(i)}^* = \min \left[\max_{j \leq i} \{ (n + 1 - j)p_{(j)} \}, 1 \right], \quad (31)$$

then we can reject any hypothesis test $H_{(i)}$ where $p_{(i)}^* < \alpha$ and this result will be the same as the previous method.

2.8 HAUSDORFF DISTANCE METRIC

The Hausdorff distance is a standard way of defining a metric on subsets of a metric space (Henrikson, 1999). Given a space S with a metric d , we can create a new metric d_H that acts on subsets of S .

A metric is a generalization of the notion of distance, and is defined to be a function $d : S \times S \rightarrow \mathfrak{R}$ that satisfies the following properties (Burago et al., 2001).

For any $x, y, z \in S$,

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \implies x = y$
3. $d(x, y) = d(y, x)$
4. $d(x, z) \leq d(x, y) + d(y, z)$

Given two non-empty subsets X and Y , we can define d_H as,

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} m(x, Y), \sup_{y \in Y} m(y, X) \right\}, \quad (32)$$

where $m(x, Y)$ is given by,

$$m(x, Y) = \inf_{y \in Y} d(x, y). \quad (33)$$

The Hausdorff metric is the "largest distance between an element of one set to its closest neighbour in the other set", or in simpler terms, the worst-'best distance' between our two subsets.

The Hausdorff metric provides a very natural way of extending a metric that measures distance between objects to measure distances between sets

of those objects. It has been widely used in Image Analysis to match object shapes (see Dubuisson and Jain (1994) and Charpiat et al. (2005) for examples).

There are several advantages to the Hausdorff metric.

- The Hausdorff metric satisfies the triangle inequality, which is useful property for many algorithms such as clustering.
- If the underlying metric space has the properties of being complete (every Cauchy sequence has a limit) or totally bounded (the space can be covered using a finite collection of open spheres of arbitrary size ϵ), then the induced Hausdorff metric space will also have these properties.
- Since the Hausdorff distance measures the most extreme difference and discards smaller ones, it is resilient to noise, which makes it ideal for doing image matching on photos which commonly have some amount of noise.
- The construction of the metric does not rely on arbitrary assumptions about the underlying space. There are infinitely many ways of measuring distances between sets that we could use and, simply by random chance, some will give false positives. Since the Hausdorff distance

There is a major disadvantage to this approach however, while the Hausdorff metric's focus on the largest difference makes it resilient to noise, it is still discarding information. Smaller differences are ignored by the Hausdorff metric, as long as there is a larger difference, and these small differences may be important factors.

We will be using the Hausdorff distance metric, as well as a modified version, in Chapter 5.

2.9 TOPOLOGICAL DATA ANALYSIS

2.9.1 *Introduction*

Topological Data Analysis (TDA) is a statistical shape analysis method that focuses more on topological features of the shape, rather than geometric features. This relatively new area of statistics uses a variety of mathematical methods that have been developed in recent years (Carlsson, 2009).

Topological Data Analysis has been used to analyse connections and shapes in a variety of different areas. Here are a few examples of how the TDA methods have been applied.

- Neurons in the brains neural network. The process was able to detect previously unknown structure features, specifically that in response to stimuli the brain forms a large number of 'cliques of neurons bound within cavities' (Reimann et al., 2017).
- DNA of various forms of breast cancer. Connections between different cancer types were formed using cluster analysis, and these connections were analysed for their relationship to survivability and other factors. The analysis identified a previously unknown sub-group of cancers with a common DNA mutation, that gave patients a 100% survival rate (Nicolau et al., 2011).
- Performance of NBA basketball players. TDA was able to classify the behaviour of players into 13 distinct classifications, rather than the traditional 5 positions that were previously thought of in basketball (Lum et al., 2013).
- Movements of swarms of insects. Each 'frame' of movement was analysed to examine how the topology of connections between 'close' insects changed over time. This approach was able to identify several features that were difficult to detect by eye. (Topaz et al., 2015).

Normally in most statistical methods, the topology of the sample space is assumed to be known and fixed, or at the very least any differences are eliminated or ignored. In Biology, analysis usually relies on statistics and metrics related to a subject (e.g. height, weight, etc.) to allow comparisons between them.

In Topology, the notion of metrics and 'distance' is replaced with a more generalized concept of 'connectedness', where each point has a local neighbourhood set that it is connected to. This means that a lot of information and distinctiveness is lost from the data (e.g. famously, a tea cup and a doughnut have the same topology since they both are 'solid objects with a single hole'), however this does allow analysis between vastly different structures.

Most importantly, it allows for the detection of structure differences at all scales. This will be important in our application of root shape analysis, since there will often be many smaller lateral roots that might otherwise be ignored.

The TDA methods we will be using are based around the concept of "Persistent Homology". Since persistent homology draws upon some very advanced methods in topology, we will first give a simplified version that skips over a lot of detail but gives an outline of the method.

2.9.2 *A Simplified Explanation of Persistent Homology*

Persistent Homology is a statistical analytical method designed to identify 'topological features' within a random dataset. 'Topological features' refers to any feature that would make a structure or shape topologically distinct, including connected structures and holes.

There are many different types of data for which the method can apply but usually the data will be a point cloud of discrete points, along with some metric for measuring the distances between them. The problem with analysing such a dataset is that discrete points have a trivial topology (noth-

ing is 'connected'), so we assume the point cloud is a partial random observation of an underlying connected structure with a non-trivial topology. It is this underlying structure that we wish to analyse.

So we are assuming that there are connections between the points in our point cloud, but they are not observable. So we have to infer what these connections may have been using some rule, and create new connections in our dataset.

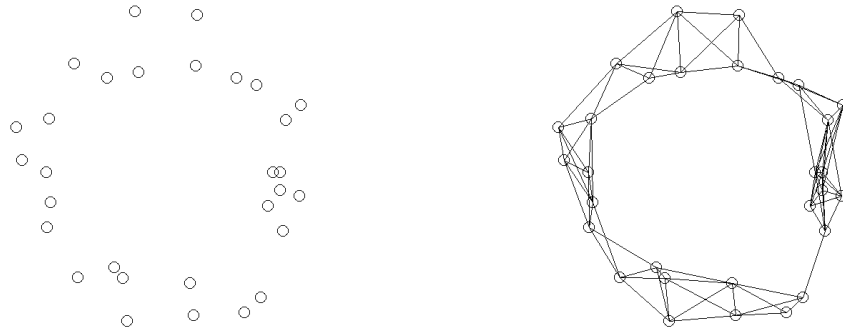


Figure 5: An example point cloud, with connections filled in on the right. The connections are made if the distance between two points is less than d .

A simple example of such a rule would be to join two points if the distance between them is less than a given distance d . An example is shown in Figure 5, where a point cloud appears to form in a circle, and we join any two points that are closer than d .

We can now use homology to find all the connected components and holes within our new connected point cloud. Looking at Fig 5 again, we can see a large central 'hole' is present (likely representing the hole of the underlying circle structure), but there are a lot of smaller holes in the circle boundary. These smaller holes are caused by the points not being in perfectly spaced or in straight lines, and most importantly, they are not features within the underlying circle structure and are 'noise'. So we need to adjust our initial rule so that they can be filled in.

We adjust our rule so that now if any set of k points are all pairwise closer than our distance threshold d , then they are filled in with a k -simplex (which means that all of the area between the k points is connected). So if a hole is small then it will be filled in. Using this, a lot of the small holes in Figure 6 are all filled in, while the larger central hole is left untouched. This method of joining together subsets, when all points are pairwise within a given distance, is the basis for the Veitoris-Rips complex.

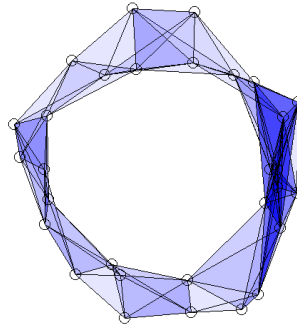


Figure 6: The Veitoris-Rips complex of the point cloud in Figure 5, when $d = 1$. Holes are filled in if their longest side is less than d . Filled holes have been colored with a transparent shade of blue, more intense shades of blue simply signifies there are multiple filled triangles overlaid.

This method worked well for the example in Fig 5, but there is a problem, that is the choice of value for d . If d is too low, a lot of points will be disconnected that should not be, and some insignificant holes will not be filled in (see Figure 7). If d is too large, then all points within the original point clouds convex hull will be connected into a single object and the topology is trivial (see Figure 8). The natural question we might ask is "what value of d should we use?", but unfortunately one single value of d is not enough.

The problem is if the underlying structure has multiple holes of different sizes or multiple disconnected components at different distances, the optimal d required to properly detect any one of these given features may be different depending on the scale of the feature. It may also be that no single value of d will identify all relevant topological features.

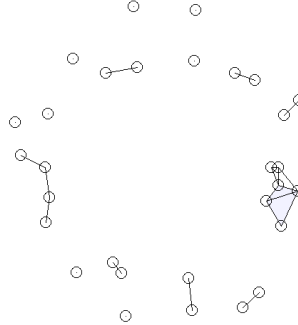


Figure 7: The Vietoris-Rips complex of the point cloud in Figure 5, when $d = 0.3$. The main central 'hole' is missing, and there are multiple disconnected components.

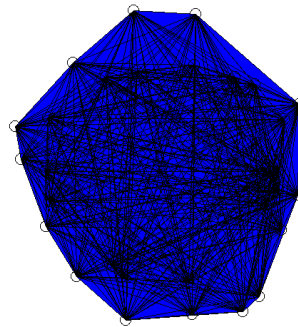


Figure 8: The Vietoris-Rips complex of the point cloud in Figure 5, when d is arbitrarily large. The structure has become the convex hull of the original dataset, with every point within connected.

FILTRATIONS Instead of choosing a single value for d , we instead vary the value of d from 0 to ∞ . This creates a sequence of connected data sets, which we call a 'Filtration'. The filtration forms the basis for persistent homology and can be created in numerous ways. We can see how the filtration for our example builds in Figure 9.

The filtration is a function that maps a real valued parameter (called the 'time parameter') to a connected version of our original dataset. A key requirement is that as the time parameter increases, the set will only become more connected (or stay the same). As the time parameter tends to ∞ , all points and components will become connected.

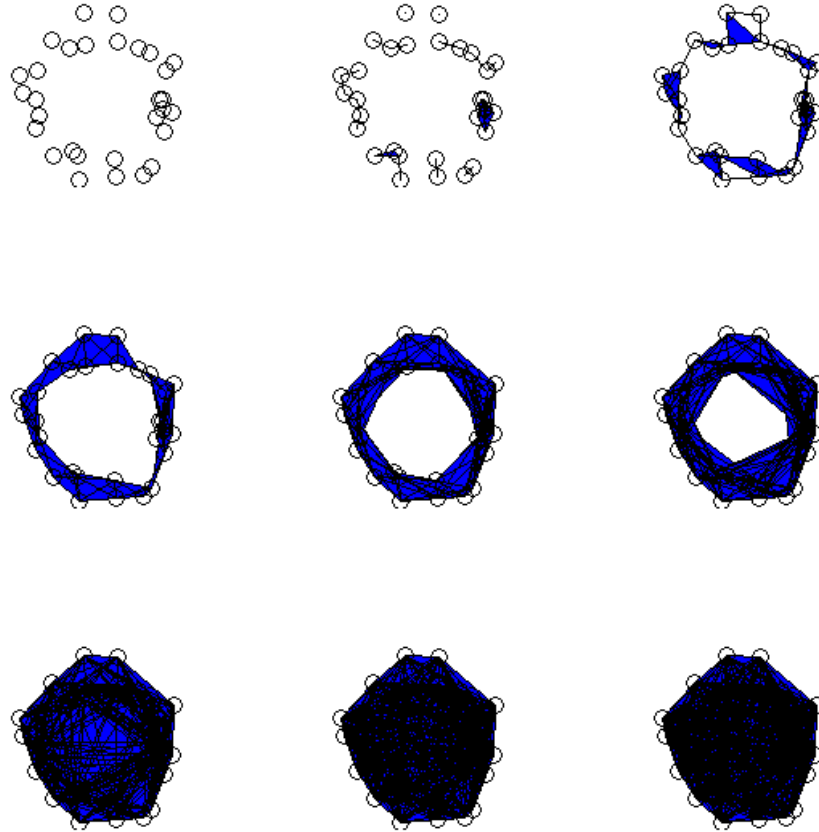


Figure 9: A sample from the filtration of the point cloud in Figure 5, with d varying from 0 to 2.5.

PERSISTENT HOMOLOGY So as filtrations 'time parameter' grows, topological holes will appear at various points, but then disappear at later points as the point cloud becomes more connected. We can track these holes as they appear and disappear, to get a 'birth time' and 'death time' for each of these features. This collection of 'birth/death times' is the transformed version of the data that TDA analyses, and is called a Barcode diagram (or alternatively a 'Persistence Diagram').

The plot of the persistence diagram for our example is shown in Figure 10. Each feature is plotted as a horizontal line that extends from the 'birth time' to its 'death time' (which are measured on the horizontal axis). The vertical order of the lines is arbitrary, in this thesis they will be ordered in ascending 'total time'. The total time of the feature is how long it 'persists' for, and is considered to be a measure of how 'significant' a given feature is. In our example barcode, there is one 'large' hole that persists for a long time and corresponds to the hole in the underlying circle, and two 'small' holes that do not persist for very long and are 'noise'.

The most important property of the barcodes is that they are stable (Cohen-Steiner et al., 2007). In other words, small changes in the original dataset will only result in small changes in the resulting barcode.

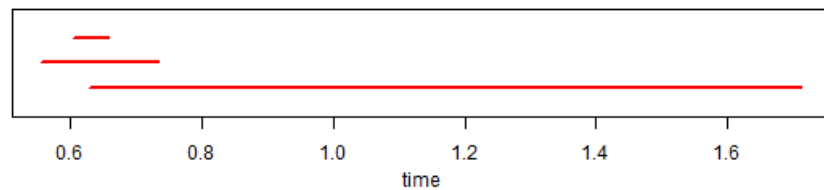


Figure 10: The barcode representation of the 'holes' found by Persistent Homology of the point cloud in Figure 5

In our example we only looked at 2D 'holes' but we can also include the 'connected components' in our barchart (see Figure 11). Each of these corresponds to a discrete point in our point cloud and all start at the 0 time, since in our initial point cloud, each point is a 'connected component' on its

own. The 'death time' of connected components is when it becomes merged with another 'larger' connected component. The longest bar at the bottom corresponds to a discrete point that is present at the 0 time and, as it grows, it merges with all the other points and become the single largest connected component. This bar is infinite in length, since the component has no death time, but by convention is drawn truncated.

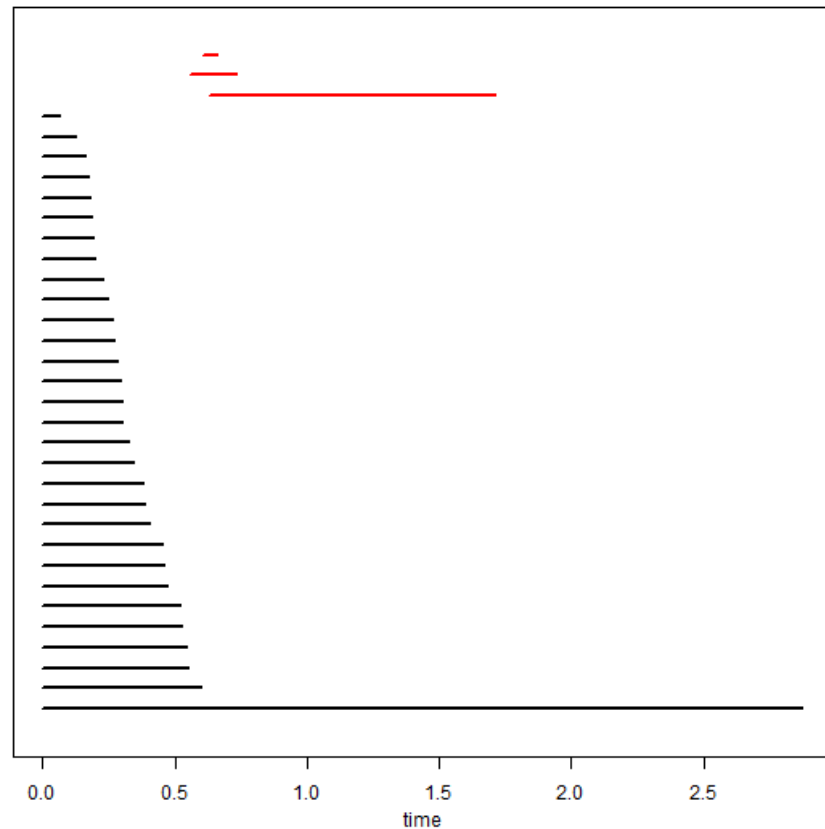


Figure 11: The barcode representation of all the 'connected components' and 'holes' found by performing Persistent Homology on the point cloud in Figure 5.

2.9.3 A technical overview of Persistent Homology

For our simplified explanation we have glossed over several important concepts.

Simplicial Complexes

The first concept is that of Simplicial Complexes. These are what we use to represent our point cloud and the connections between the points.

SIMPLEXES A k -simplex is the mathematical extension of the notion of a triangle to arbitrary dimensions. A k -simplex comprising of $k+1$ points, and these points are 'affinely independent' (which ensures the object is not 'flat' from some perspective). The interpretation of the low-dimension simplices is very simple. A 0-simplex is a single point, a 1-simplex is a line-segment between two points, a 2-simplex is a triangle and a 3-simplex is a tetrahedron. An $(n+1)$ -dimensional simplex is represented as $[v_0, v_1, \dots, v_n]$, where the v_i 's are the points that the simplex connects to.

A k -simplex has $k + 1$ 'faces' which are $(k - 1)$ -simplices, which are the original simplex but with a different point removed. So a tetrahedron has four triangles as its faces, each triangle has three line segments as its faces and each line segment has two points as its faces.

ORIENTATION Normally simplexes do not have an orientation, but they can be given one. The orientation of a simplex is given by the ordering of its elements, with two simplexes having the same orientation if and only if their orderings are an even permutation of each other (an even permutation can be generated with an even number of 'swaps'). The negative sign is used to denote an equivalent simplex of the opposite orientation.

Since 0-simplexes have only one point (with only one permutation), they have no orientation.

SIMPLICIAL COMPLEXES A simplicial complex K is a set of simplices such that

1. Any face of a simplex from K is also in K
2. The intersection of any two simplices is a face of both simplices.

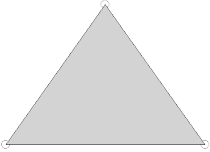
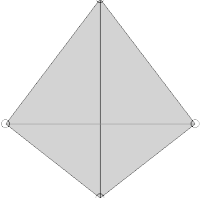
0-Simplex	Point	
1-Simplex	Line	
2-Simplex	Triangle	
3-Simplex	Tetrahedron	

Table 1: Table of the low dimensional simplices.

A simplicial complex can be considered to be a group using the symmetric difference of two sub-complexes as an operator (denoted with $+$) i.e. the set of all simplices that are in one set but not in both. The empty set is the identity element for this group.

In addition, if we are using oriented simplexes then the combination of two simplexes that are the same but with opposite orientation, gives an empty set. i.e. for a oriented n -simplex A , $A + (-A) = \emptyset$.

The set of n -simplices within a sub-complex, will form a basis for the sub-complex.

HOMOLOGY Homology is a topological method that is designed to systematically identify topological features within simplicial complexes. Its purpose is to classify spaces using their loops (or closed cycles). Two loops are considered to be 'homologous' if they can be continuously deformed to

match each other, and they are not 'homologous' if there is some topological feature preventing this. For example, two loops can fail to be homologous if they are on different non-connected objects. They can also fail if one encloses a 'hole' that the other loop does not.

We can use this classification of loops to identify topological structures that are present within the space and measure how many of each type of 'hole' are within the space.

BOUNDARY OPERATOR The boundary operator is used to find the 'non-shared faces' of a set of oriented n -simplices (called an n -chain). If the boundary of a set is empty (i.e. every n -simplex 'shares' its faces with others), then we know that it is a cycle or loop.

The boundary of an oriented n -simplex, $A^n = [v_0, \dots, v_n]$, is defined as

$$\partial_n(A^n) = \sum_{i=0}^n (-1)^i [v_0, \dots, (v_i), \dots, v_n], \quad (34)$$

where $[v_0, \dots, (v_i), \dots, v_n]$ is the original n -simplex with the i th vertex removed, and $(-1)^i$ signifies that odd i 's will have opposite orientation.

We can extend this definition linearly to n -chains. For a n -chain, $K^n = \sum_1^k A_i^n$, we define

$$\partial_n(K^n) = \sum_1^k \partial_n A_i^n. \quad (35)$$

The boundary operator maps the set of n -chains to the set of $(n-1)$ -chains, and in fact, acts as a group homomorphism (Edelsbrunner and Harer, 2010).

An important property of the boundary operator is that if you apply it twice, i.e. take the boundary of a boundary, then you will always get the empty set. This means that every boundary is also a cycle.

$$\partial(\partial([v_0, \dots, v_n])) = \emptyset \quad (36)$$

HOMOLOGY GROUPS We will use 'Homology groups' to classify the loops or cycles within our complex. First, we choose a dimension k to analyse.

We note that the elements of the subgroup $Z_k = \ker(\partial_k)$ are all cycles, since they map to the empty set. We also note that the elements of the subgroup $B_k = \text{im}(\partial_{k+1})$ are also cycles (since all boundaries are cycles). In addition, since the operation of Z_k is commutative (i.e. it is an Abelian group), then its subgroup B_k is a "normal subgroup" of Z_k . Therefore we can form a "quotient group" using the set of sets, $\{z + B_k, z \in Z_k\}$ where $z + B_k = \{z + b : b \in B_k\}$. This quotient group has the operation $\times$, which is defined as $(a + B_k) \times (b + B_k) = (a + b + B_k)$.

The quotient group, $H_k(K)$ which is formed from Z_k and B_k , is called the k -th homology group,

$$H_k(K) = Z_k / B_k = \{zB_k : z \in Z_k\}. \quad (37)$$

The k -th homology groups are topologically invariant, two complexes with the same topology will give isomorphic homology groups.

The rank of this group is called the k -th Betti number, and is a measure of the k -dimensional 'holes' within the complex. The rank is defined to be the size of the smallest 'generating set' (a set of elements, $X \subseteq Z_k$, where every element in Z_k as a combination of the elements in X and their inverses.)

Filtration

Consider a function $f : K \rightarrow \mathbb{R}$, where K is a family of simplicial complexes, that is non-decreasing for increasing sequences of faces. In other words if $\gamma \in K$ is a face of $\sigma \in K$ then $f(\gamma) \leq f(\sigma)$. The Vietoris-Rips method, for example, uses 'the total boundary length of all the simplices with the complex' as the filtration function.

Then for every $a \in \mathbb{R}$, we can create a sub-complex, $K(a) = f^{-1}((-\infty, a])$. Then if $a < b$, then $K(a)$ is a sub-complex of $K(b)$.

In practice we will use only a finite set of $n - 1$ values for a to work with (denoted by $a_i, i = 1 \dots (n - 1)$). We will also denote $K_i = K(a_i)$.

A filtration is a collection of $(n + 1)$ complexes such that,

$$\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K \quad (38)$$

Persistent Homology

When $0 \leq i \leq j \leq n$, the inclusion $K_i \rightarrow K_j$ is used to induce a homomorphism on the homology groups $H_k(K_i) \rightarrow H_k(K_j)$. This allows us to track the creation or destruction of holes in the filtration.

At each stage, some holes may be 'created' (i.e. are not linked to, from the previous filtration homology group) and some may be 'destroyed' (i.e. linked to the identity element of the next filtration's homology group). We can track how long these holes persist for, from the time they are created to the time they are destroyed.

If a 'hole' persists for a long time, then it is far more likely to be a significant or 'real' feature and not just an artefact.

We collect this information as a set of tuples, (n, b, d) , where n is the dimension of the hole, b is the 'birth time' and d is the 'death' time. By definition, $b < d$. This set is known as a 'Persistence diagram', or as a 'Barcode'

2.9.4 *Types of Filtration*

VIETORIS-RIPS COMPLEX One way of creating a filtration from a point cloud is using the Vietoris-Rips complex. Let the set of points act as our vertices. Consider the set of all simplexes of dimension ≤ 3 or less that can be formed using these vertices, this will act as our K .

Given a set of n vertices, $\{v_1, v_2, \dots, v_n\}$ the Vietoris Rips complex (K) is defined as,

$$\begin{aligned} K = & \left\{ [v_i] : i = 1 \dots n \right\} \\ & \cup \left\{ [v_i, v_j] : i, j \in 1 \dots n \wedge (i < j) \right\} \\ & \cup \left\{ [v_i, v_j, v_k] : i, j, k \in 1 \dots n \wedge (i < j < k) \right\}, \end{aligned} \quad (39)$$

where $\{[v_i]\}$ are the point simplices formed from the input vertices, $\{[v_i, v_j]\}$ are the line simplices between each pair of vertices, and $\{[v_i, v_j, v_k]\}$ are all possible triangle simplices formed from three different vertices.

The function f used to make our filtration is , for arbitrary $v_0, \dots, v_n$,

$$f([v_0]) = 0, \quad (40)$$

$$f([v_0, v_1]) = \|v_0 - v_1\|_d, \quad (41)$$

$$f([v_0, \dots, v_n]) = \sum_{i=1}^n f([v_0, \dots, (v_i), \dots, v_n]), \quad (42)$$

where $\|v_0 - v_1\|_d$ is the distance between vertices v_0 and v_1 .

A way of visualising this in two dimensions is, circles are placed over the point cloud centred at each point and their sizes start at 0 and are slowly increased at the same rate. As the size of these circles increases we draw lines between every pair of points whose circles overlap. When a set of 3 points are all connected and their total boundary length is less than the size of the circles, we fill in the face formed by the 3 points. This continues until every possible line and face has been formed, at which point the point cloud is a single connected object with no holes.

This method of filtration has the advantage that there is a 'fast' algebraic method for computing in that is $O(n^3)$ in worst case (where n is the number of points in the point cloud).

GRID COMPLEX A grid complex is an alternate method of building a complex from a point cloud. A triangulated grid of points is created in the space surrounding the point cloud. This forms the complex on which we build our filtration.

If we then choose a threshold value, r , a sub-complex is formed by choosing only simplexes when all the vertices are within r distance of one of the points in our point cloud. We can see this method in action in Figure 12.

This method is typically quick to calculate, and can be easily extended to more general scenarios. However it typically results in more complex structures.

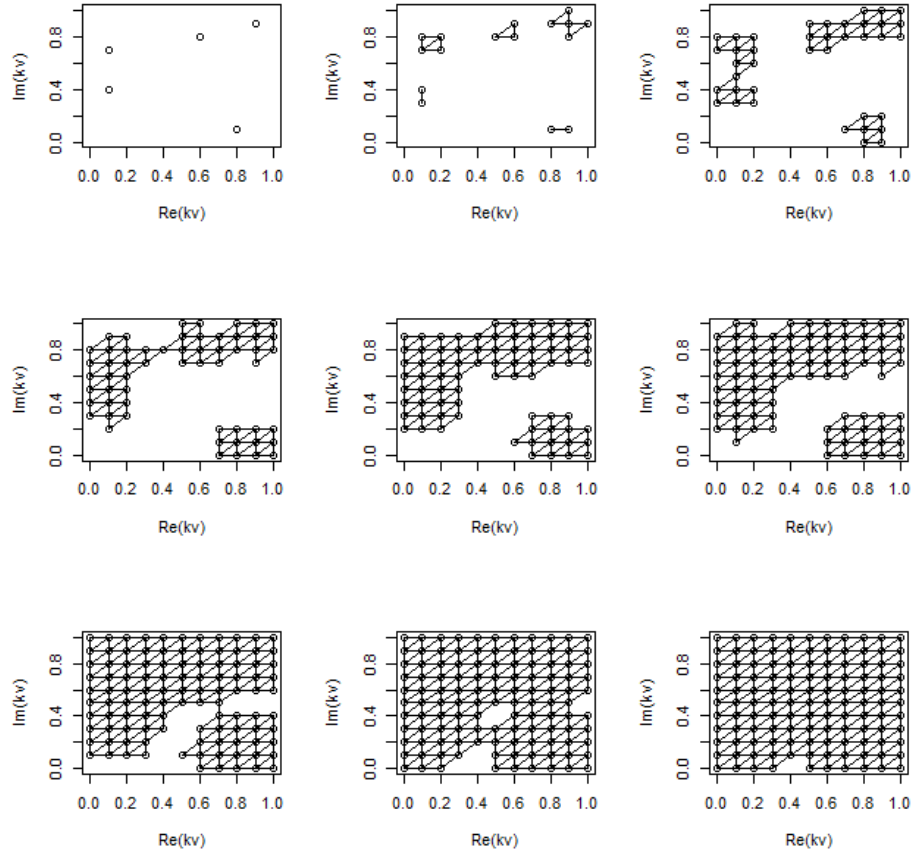


Figure 12: Filtration of an example grid complex.

FILTRATIONS BASED ON DIRECTIONS We can build our filtration by using the points position. As we did in the Vietoris-Rips complex, we build

our underlying complex K using the points and all simplexes formed between them of less than dimension 3.

We choose a function for evaluating the individual points in the point cloud, f , that assigns a positive real number to each point. Then, for a given threshold, r we included every simplex $[v_0, \dots, v_n]$ if for every point in the simplex, $f(v_i) < r$.

Simple examples for f are the vertical distance from the lowest point, distance from the origin, .

2.9.5 Distance between Persistence Diagrams

Persistence diagrams capture the topology of the underlying shape, and now that we can calculate them, we naturally will wish to compare them.

The Persistence diagram is a set of persistent homology classes (which capture a single topological feature). These classes have a 'start' and 'end' time, and are represented as $\mathbf{x} = [x_L, x_U]$, where $x_L, x_U \in \mathfrak{R}$ and $x_L \leq x_U$. We can interpret these as 'points' on a 2D diagram. The simple distance between two persistent homology classes, ($\mathbf{a} = [a_L, a_U]$ and $\mathbf{b} = [b_L, b_U]$), is calculated as $||\mathbf{a}, \mathbf{b}|| = \max \{|b_L - a_L|, |a_U - b_U|\}$.

PAIRING POINTS BETWEEN PERSISTENCE DIAGRAMS In order to determine distance between diagrams, we wish to pair up each point in one diagram with a point in the other diagram, without any overlap. So effectively, we wish to find a bijection between the two sets of points that minimises the total distance.

However there is no guarantee that there will be equal numbers of points in each diagram, which makes such a bijection impossible. So in order to pair up any extra points, we shall assume that there are an infinite number of points on the diagonal of each diagram (in the form (x, x)). These can be interpreted as being 'imaginary' trivial classes that are born and die at the same level of filtration (and thus are completely insignificant).

So our pairing method will have two types,

- Pairs between 'real' diagram points: $[a_i, b_j]$
- Pairs between a 'real' diagram point and an 'imaginary' diagram point on the opposing diagrams diagonal: $[a_k, (t, t)]$ or $[(t, t), b_l]$

We require that each diagram point, belongs to exactly one pairing.

It should be clarified, that even when there are equal numbers of points, we still allow points to be paired with a point on the diagonal. We can see an example of such a pairing in Figure 13.

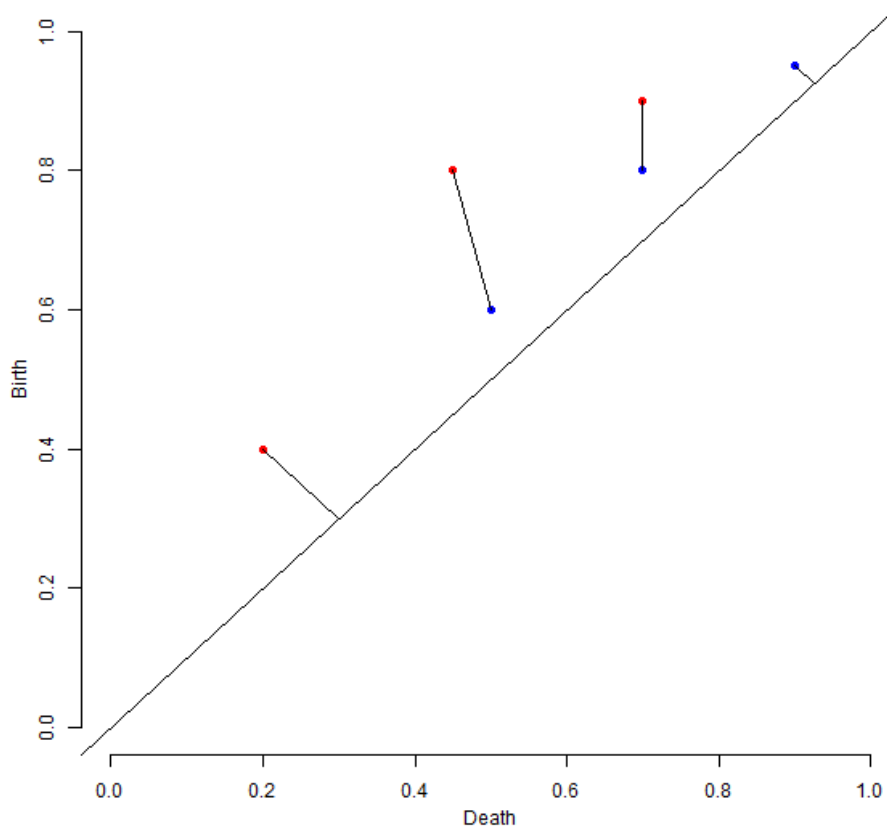


Figure 13: An example of a pairing between the red points of one persistence diagram to the blue points of another.

WASSERSTEIN DISTANCE The Wasserstein distance between two persistent homologies is defined as being the minimum largest distance between paired points amongst all possible pairings,

$$d_p(f, g) = \left(\inf_{\eta} \sup_{[a, b] \in \eta} ||a - b||^p \right)^{1/p}, \quad (43)$$

where η is any bijection pairing between the points of the persistence diagrams.

BOTTLENECK DISTANCE This is a special case of the Wasserstein distance, when $p \rightarrow \infty$.

The bottleneck distance is the metric used to justify the stability of the persistent homology method in (Cohen-Steiner et al., 2007).

The downside of these two methods is that they can be computationally expensive, especially for large datasets.

ROOT SYSTEMS

3.1 INTRODUCTION

The FutureRoots project aims to run a series of scientific experiments that grow various types of crops under different environmental conditions. They will scan the roots of these crops in order to get a scan of the hidden root system. From this point, we will need to consider what questions we need to ask of the shapes of these root systems, and how they can be answered.

These questions may test a-priori assumptions. So for example, we may suspect that one type of crop grows deeper roots than another. In this case, we must use statistics that are designed to specifically measure the believed trait, and then we can test these statistics to see if they match the assumptions.

We may also not have specific assumptions of how the root shape is different, just that there may be differences under various treatments, and we want to know what those differences may be. This is a much more open ended question, and a lot more difficult to answer, since there is a lot of ways to measure roots. The natural approach would be to measure as many traits as possible and to see if any of those have significant differences under the various treatments. However this approach means we must be careful to avoid potential 'p-farming', since we are looking at so many traits.

A simpler question we may ask, is whether there are significant shape differences, but not to worry about asking what those specific differences

3.2 APPROACH

are. This approach is obviously less insightful if we get a positive result, but if we fail to get a positive result, then this tells us exactly the same as before. This is useful if we suspect a-priori that there are no shape differences but still wish to test for them.

In addition, since we are not concerned with identifying a shape difference, our method can be more general and less rigid. However if we should get a positive result, then we must be prepared to ask the natural follow up question of what those differences are.

The final question is that we may wish to experiment with a new method for measuring traits, to see if this can gain us any insight into the structure of root systems that previous methods may have missed.

3.2 APPROACH

To recap, in the wheat species we will be looking at, root systems do have some common structure. First the seed grows a long primary root in order to gain nutrients for growth, typically vertically straight downwards. It then grows multiple seminal roots to the side, that anchor the plant in the soil and provide more nutrients. Each of the roots then start to grow smaller offshoot lateral roots. Occasionally lateral roots may grow their own offshoot sub-lateral roots.

The key problem we have to deal with is that the number of seminal roots, and the number of laterals is random. This means that the structure or topology of the root system will be different for every plant.

Most statistical methods are built upon the assumption that the structure of the sample data is fixed, or at least that any differences in structure should be eliminated using summary statistics.

However the number of roots, their types and their connections to each other, are important factors that should not be discarded.

3.3 GENERAL STRUCTURE OF WHEAT ROOT SYSTEMS

We will be dealing with root systems that were grown on 2D growth paper and in a 3D column of soil. From a mathematical perspective, these roots are embedded in Euclidean space ($\mathbb{R}^2$ and $\mathbb{R}^3$, respectively).

Each root strand is modelled as a single connected line in Euclidean space. This line has a direction and start/end points implied by the direction of growth.

We will be ignoring the diameter of the root strand since the resolution of the CT scanner is not sufficient to reliably estimate the diameter of the root strand. For smaller roots, they may occupy only one or two pixels in total which is sufficient to determine their approximate path. It should also be noted that this means there will likely be very small segments of roots that were not possible to detect and are therefore missing. This also means that the end point or 'tip' of a root strand may not be accurate.

3.4 ROOT SYSTEM SHAPE

We need to design what we consider to be important aspects for the 'shape' of a root system. In traditional shape analysis, two objects are defined to have the same shape if one shape can be translated and rotated to match the other (we can also optionally include operations like scaling and reflection if the application requires it).

This definition can extend to roots but there is a problem, roots have a definite vertical orientation (i.e. the direction of gravity). If a plant grew its roots upwards out of the soil, we would definitely consider it to be a different shape to a normal plant whose roots grew into the soil.

For horizontal orientation (e.g. north, south, east and west), there is no empirical evidence to suggest that wheat roots have a bias to one direction or another, and during the experiment there was no attempt to keep their

horizontal orientation consistent. For these reasons, we will assume that horizontal orientation is a nuisance parameter.

Likewise, horizontal translation is considered a nuisance parameter (i.e. we do not care where in the soil it was planted), but vertical translation is considered important (a seed that was planted five centimetres below the surface is different from a seed that was planted one metre below the surface).

To summarize, two root systems are considered to have the same shape if we can horizontally translate and rotate one of the roots (keeping its vertical position and orientation constant) to match the other.

3.5 RSML

The raw data format for storing plant root systems is the Root System Markup Language or "RSML". This format is based on the XML standard and can store 2D and 3D image metadata, plant root properties and geometries, continuous functions along individual root paths as well (such as root diameter or water uptake). This is the format the FutureRoots project uses to store its data.

The RSML data stores the path of an individual root strand as a sequence of points, called a polyline.

3.6 COMPENSATING FOR PLANT OVERGROWTH

An issue that arose with the experiments conducted at the Hounsfield facility is that the root systems spread out much farther in the soil than anticipated, and the soil columns were too small to contain them. This meant that the roots would often outgrow the soil available, hit the edge of the column and be forced to follow the curve of the edge or bounce back inwards.

This happen in the majority of wheat plants grown and was unfortunately unavoidable. We can see this effect happening in Figure ??.

This is a problem for shape analysis since each of the roots is constrained into an unnatural cylinder and this becomes the default shape of any root system. This makes the plants more similar than they would be otherwise. We need to find a way to compensate for this effect and try to estimate how the root would have grown in an unrestricted soil environment.

Fortunately, we have observed empirically that root strands will tend to grow in straight lines (subject to occasional random perturbations), we will therefore make the assumption that a root strand that hit the wall would have grown in a straight line. In addition, we will assume that the total length of the root, and the distances between lateral roots along a seminal root, are the not affected by the collision with the column wall. Using these assumptions, we can 'unravel' the root system to estimate how it might have grown if it was unrestricted in the soil.

Before we continue, it is very important to stress that this method is not intended to replace the original data. The assumption that roots travel in straight lines with random perturbations, is a strong assumption that could easily be inaccurate. The purpose of this method, is to supplement the original dataset, by emphasising shape differences between root systems that were unnaturally constrained in a way that feels natural and realistic.

3.6.1 *Unravelling method*

First, we must decide what are the circumstances the determine that a root has hit the edge of the column, or is so close to the edge that its growth is constrained.

We begin by deciding on a 'near to the edge' region. Within this region we are considering roots to be close enough to the edge that path in which they growth is constrained. There is measurement error in both the location of roots and the edge of the column, so we will make this region fairly gen-

erous. This region will be defined as anything not within a central cylinder centered within the column. The border of this region is shown in Figure 14.

Let's first consider a seminal root that starts at the shoot origin and travels into the edge region (and possibly leaves). An example of this is shown in Figure 15. The root bounces back away from the column wall, and it is probable that if the wall had not been there it would have kept going straight.

For each pair of adjacent points in our line, there is a change in the horizontal direction. We are assuming that the changes that occur within the edge region are unreliable and should be eliminated. Therefore for each pair of points where at least one point is within the edge region, we set the change in horizontal direction to 0. We then recreate the new path of the root.

We can see an example of the new path in Figure 16. The root follows a much straighter path than before.

3.6.2 *Handling Laterals and Sub-Roots*

Dealing with laterals is a bit more complex, this is because they are dependent on their parent roots. If the parent root changes, then the laterals will have to shift with them. We wish to preserve the laterals relationship with its parent, specifically the location and angle at which it emerges. We should therefore shift and rotate the roots to match their parents new location/orientation. An example is shown in Figure 19.

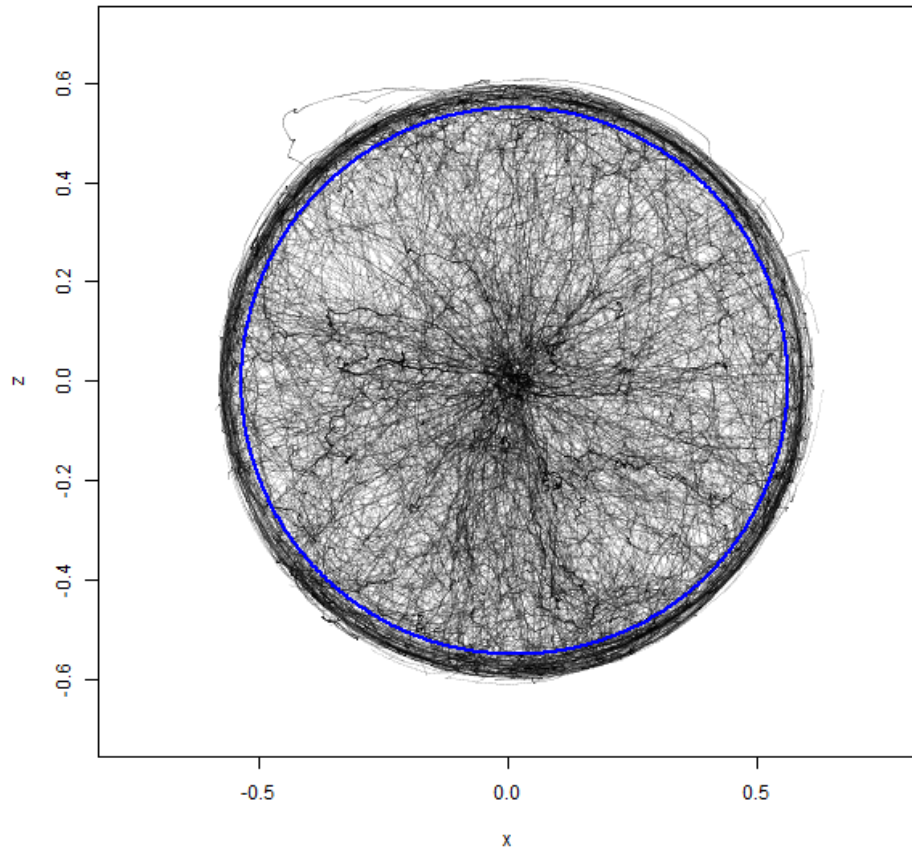


Figure 14: A circle is used to separate the region close to the edge of the column. The circle is 0.55 in radius and centered at (0.14,0).

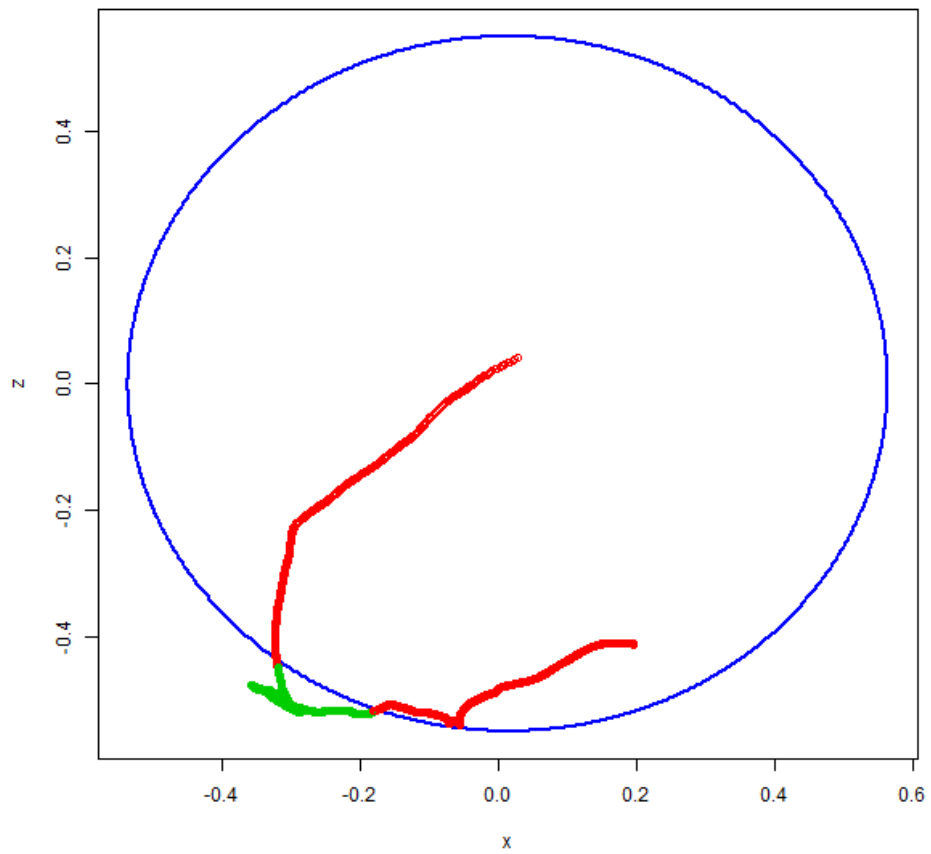


Figure 15: A single root strand from a wheat plant is shown with a circle region. The green region of the root strand, outside the circle, is considered to be 'unreliable'.

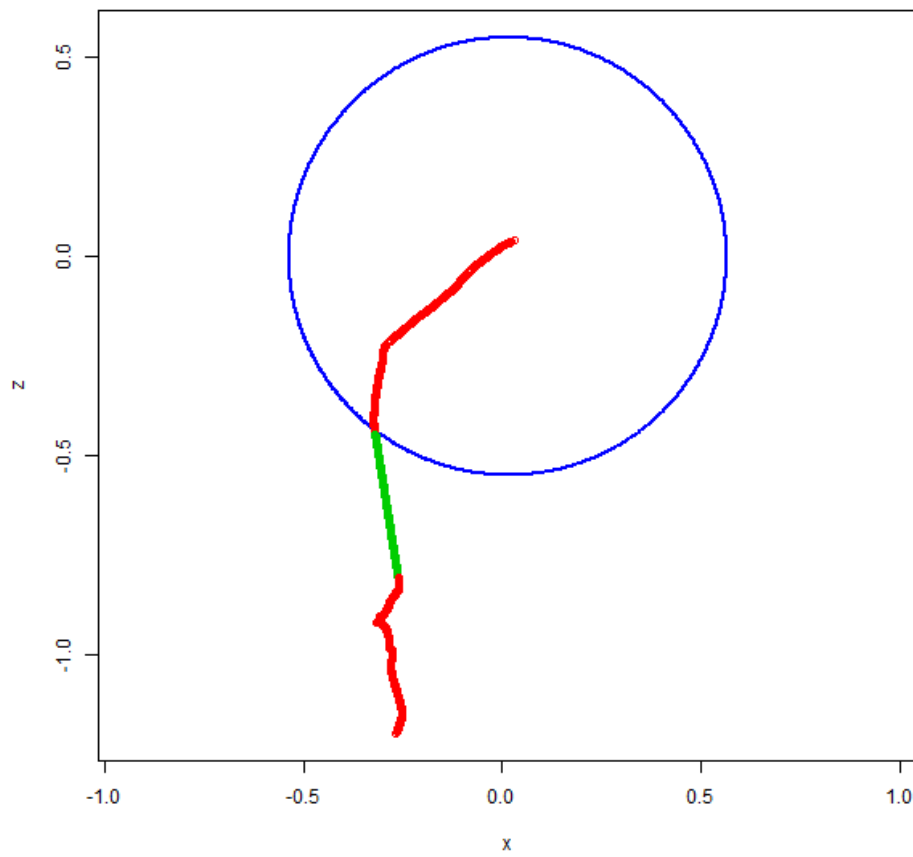


Figure 16: The single root strand from Figure 15, after the green section has been 'straightened'.

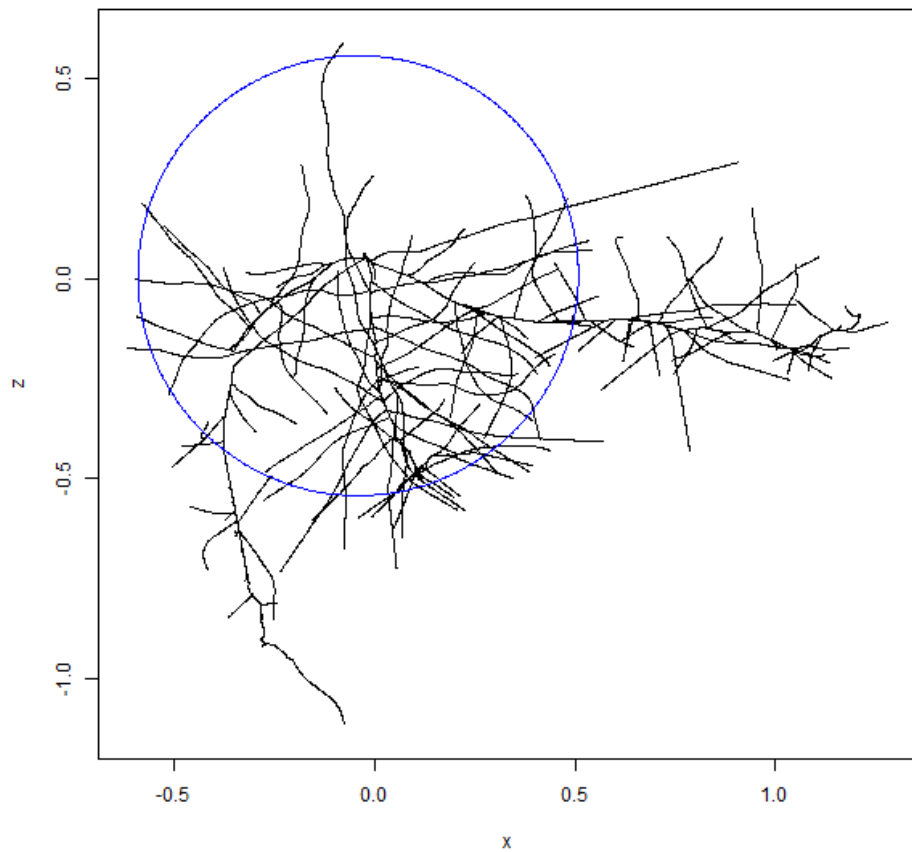


Figure 17: A root system where each of the root strands has been 'straightened' and each of the laterals adjusted to their corresponding location.

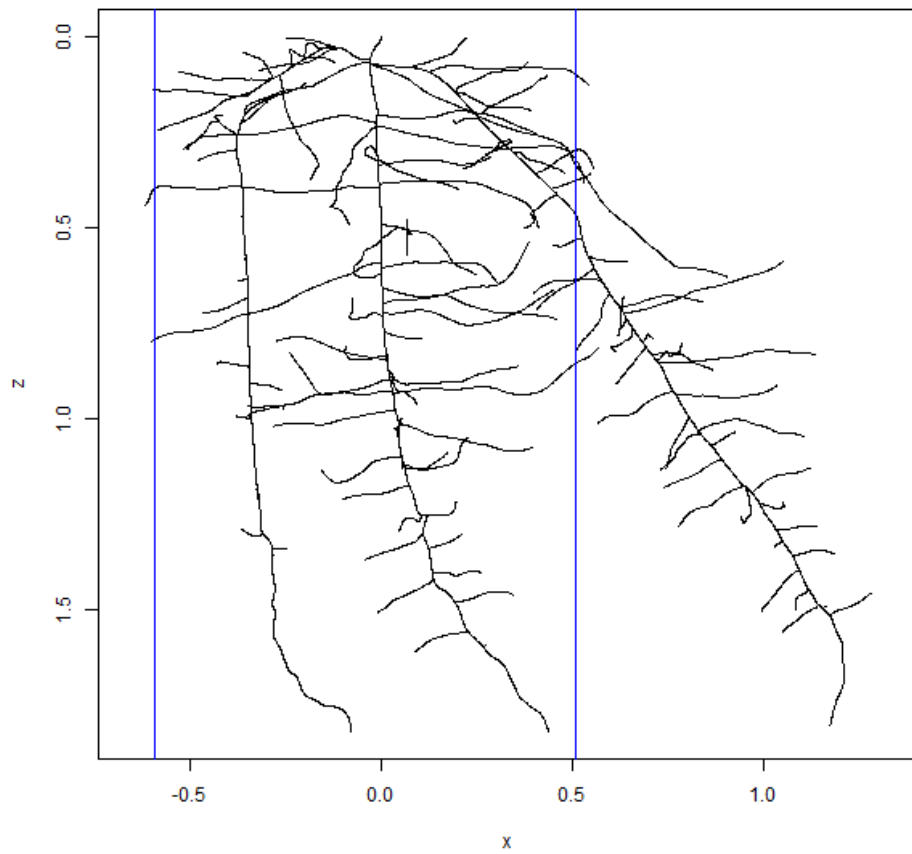


Figure 18: A root system where each of the root strands has been 'straightened' and each of the laterals adjusted to their corresponding location.

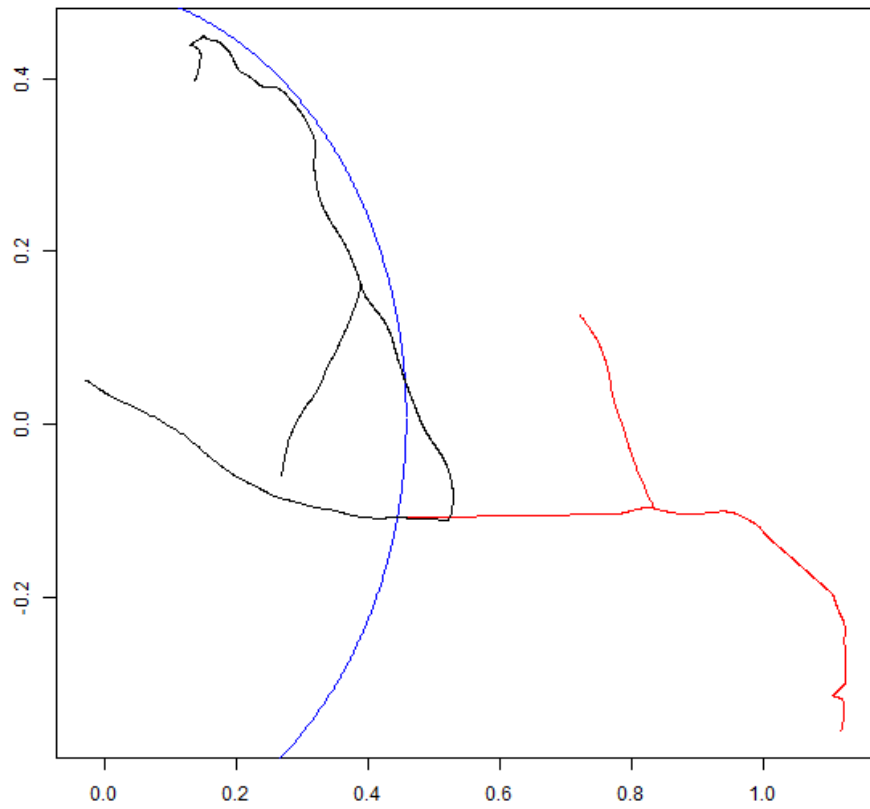


Figure 19: The original seminal root and lateral root are drawn in black, with the compensated roots shown in red. Each lateral has to be moved and rotated to ensure that the interior angles of its intersection with the parent root are preserved.

3.7 INSTABILITY

The process is very dependent on the choice of boundary. If a root is in the process of making a slight turn when it crosses the boundary, then only the first part of the turn is reflected in the estimated trajectory. Since a small change in angle can have a large effect over a long distance this means that the root strand will take a very different path than it would have if the barrier started before the slight turn or after it.

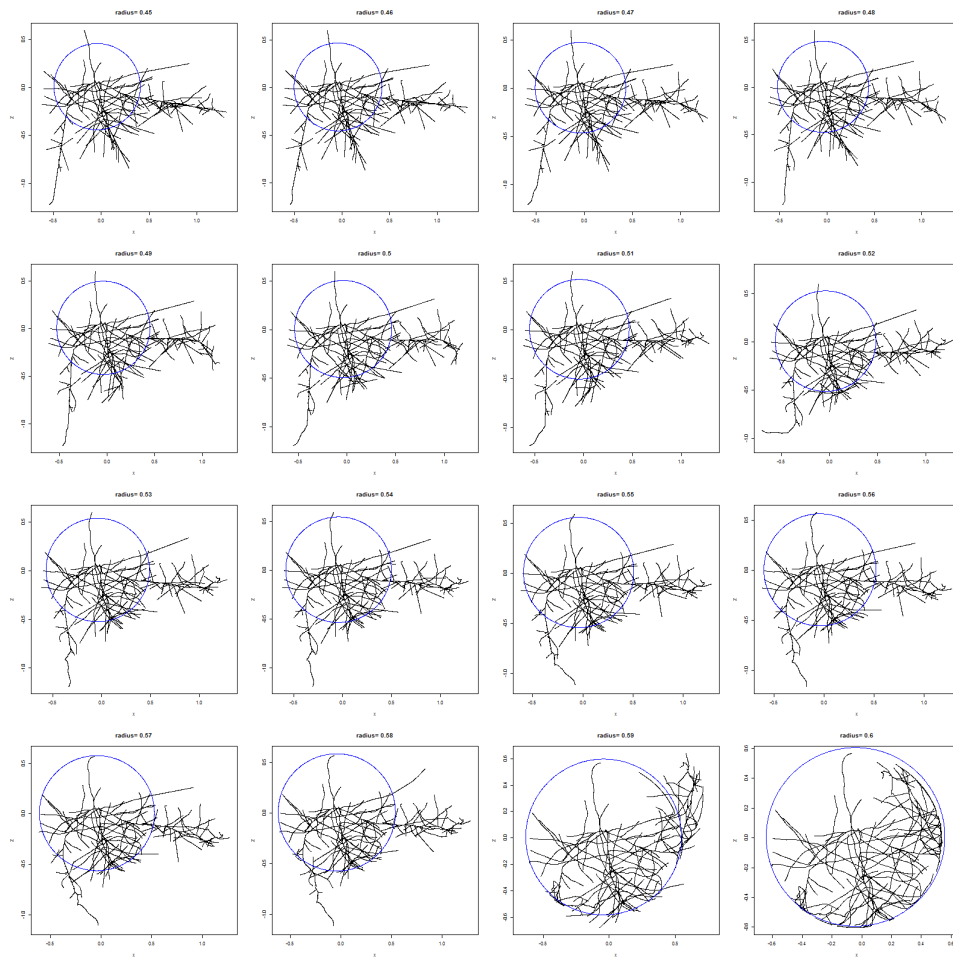


Figure 20: The compensated root system is dependent on where the barrier is located.

3.8 APPROPRIATENESS

An immediate objection that might be raised to this method is that the unravelled data may not be realistic, and more problematically, we cannot determine how unrealistic it is. We do not know how the plant would have grown without the column restricting it, and our unravelled method is simply assuming that they would have grown straight, which could easily be wrong.

For this reason, it is important that we do use this data as a replacement for the original scans, but as supplementary data that provides an artificial perspective.

3.9 CONCLUSION

Root systems provide a challenging type of data to analyse, with both important geometric and topological traits that may need to be considered. We also need to be able to eliminate the nuisance factors of 'horizontal translation' and 'horizontal orientation' (while ensuring that 'vertical translation' and 'vertical orientation' are still measured factors in our analysis).

The plants in our experiments frequently grow large enough to hit the sides of their columns which unnaturally restricts their movement, so by developing a method to approximate what their growth might have been we may find important insights.

In the next chapter we will begin to form our analysis framework by analysing root systems using measured traits.

ANALYSIS OF MEASURED TRAITS

4.1 INTRODUCTION

The first approach we will be looking at is the conventional method used by biologists to analyse root systems. That is to measure a set of given traits and see how these are affected by any dependent variables, such as the plant's genotype or growth conditions.

This approach is simple, straightforward and can capture a lot of information about the root system shape. However as we will show, there is inevitably going to be some information that is lost.

The reason some shape information is lost, is because roots have a randomised structure or topology, in terms of the number of roots and how they are joined together. Features that are present in one root system, are not guaranteed to be present in another.

4.2 SIMPLE SUMMARY STATISTICS

By 'trait', we will be referring to 'simple summary statistics'. We will use the term 'simple statistic' to refer to any statistic whose output has a fixed dimension.

For example, the 'Set of Root Lengths' is not a simple statistic since its size will vary with the number of roots within the root system. However,

'Average Root Length' will always be a single positive real number, and therefore is a simple statistic.

So for this chapter we will be focusing on these kinds of statistics, that we can measure and know what the topology of their output will be. This restriction makes analysis much simpler and we can use plenty of conventional univariate and multivariate methods.

A key problem with this approach is that it downplays a lot of the problematic topological and structural information present in the root system, since this is very difficult to represent in a fixed-dimensional space. Hence we will later be expanding our approach to include more unconventional methods of analysis in Chapters 5 and 6.

So analysing and comparing simple statistics is fairly straightforward since we can use most conventional statistical methods. The only downside is that there may be more complex shape or topology traits that cannot be easily expressed this way.

4.3 STATISTICS OF ROOTS

Here we will provide a quick list of traits that we will be measuring for root systems.

4.3.1 *Length*

The root strand is measured from its start to its end, following the path of the strand. For primary roots, the start is the base of the plant (which will be the origin point) and for laterals, the start is the point when it branches off from its parent root. For all types of roots, the end is the tip of the root.

There can be some ambiguity as to where we should measure the 'path' of a root since roots are technically not lines but have a width. In principle, we're measuring the shortest path between the start and end points that

is contained entirely within the root. In practice however, the width of roots are so thin compared to their length, any measurement differences is negligible. In fact in 3D scans, the width of roots is often smaller than the resolution of the CT scanner, making any measurement of such impossible.

When a root branches in two, there can be some ambiguity as to which is the original and which is the offshoot lateral root. In most cases it is fairly clear as the lateral root will be smaller than it's parent root and travel in a very different trajectory. However in cases where there is ambiguity, the shorter root will be treated as the offshoot.

4.3.2 *Extreme Statistics*

Extreme statistics capture information about the outlying regions of root systems. Of interest is how far down the roots grow, which we will call the 'Maximum Depth', and how far from the roots' origin do they grow horizontally, which we will call 'Maximum Width'.

Both of these statistics are influenced by the general 'size' of the plant, so we also measure the 'Max. Width / Max. Depth Ratio'. This eliminates the size factor and gives us a very simple shape measurement.

These statistics do have the inherent flaw that since they are measuring outliers, the results are susceptible to having large random outliers.

4.3.3 *Convex Hull*

The 'Convex Hull' is the smallest filled shape/set that encloses the entire root system and is also 'convex'. A set is convex if a straight line can be drawn between any two points that lies completely within the set. Informally, in two dimensions, the Convex Hull can be visualised as the area enclosed by stretching a rubber band around the shape. The convex hull is unique and can be calculated easily by statistical software (Eddy, 1977).

This gives us a measure of the area that the root system encloses, and is less prone to outlier influence than extreme statistics.

From the convex hull we are interested in measuring the 'Surface Area' and 'Volume' (or, equivalently, the 'Perimeter' and 'Area' in 2D).

We can also measure the 'Isoperimetric Ratio', which is defined as $B^d / V^{(d-1)}$, where ' d ' is the dimension of a shape, ' B ' is the 'boundary' measurement ('Surface Area'/'Perimeter') and V is the 'volume' measurement ('Volume'/'Area'). This measures how circular or compact a shape is, and may provide a useful shape statistic.

4.3.4 *Angle Statistics*

In this section, we are considering 'Angles' to refer to the direction of a path as a vector, rather than relative angles between vectors. This choice is to line up with standard biological terminology.

For each root strand, there are two angles that we are interested in. The 'Emergence Angle', which is the initial direction of growth that the root begins in, and the 'Tip Angle', which is the final direction of growth that the root ends in.

In addition for lateral roots, which split off from a parent root, we want to know the 'Parent Intersection Angle', which is the direction of the parent root at the point where the lateral root splits off.

Working with angles in a root system is complicated by the fact that there are subtle differences between angles in 2D and 3D. The rotation group in 3D is not commutative. This means that when dealing with 3D rotations, the order in which the rotations are applied is important.

In addition, we are assuming that the horizontal 'orientation' of the plant is a nuisance variable that should be eliminated. This will not be an issue when comparing angles within the same root system, but since different plants do not have the same 'horizontal orientation', when comparing angles from two different plants it will need to be taken into account.

This issue means that for 3D plants, if you rotated the scan horizontally around the vertical line that passes through the root system's origin, then the results of our analysis should not change. Equivalently for 2D plants, if you reflect the scan around the vertical line passing through the roots origin, then the now flipped root system should not affect our analysis.

However the biggest issue with measuring angles, is where we should measure an angle from. The basic idea is to take the point at which we want to start measuring the angle from, then take a second point that is further along the root, and measure the direction of the straight line from the first point to the second. However, from observation, we have seen that often a lateral root will start growing out from its parent root at an approximately 90° angle, but then suddenly change directions after a few millimetres.

This means that our choice of where the second point is along the root, will affect our result. Unfortunately there isn't a single 'correct' choice for this. Different locations are effectively measuring different things, and in fact, we will be including measurements of angles along multiple distances.

We can choose to make it a fixed distance along the root such as half a centimetre; if the root length is shorter than our fixed distance we simply use the farthest point possible. Alternatively, we can make it a relative distance along the root, such as 10% of the root strands total length.

4.3.5 *Topology Statistics*

These statistics aim to capture the 'structure' of the root system independent of any distances involved. It is difficult to summarise a graph of connections into a simple statistic, but there are some basic measures that we can use.

The simplest measure is to count the number of roots of each type. Giving us a 'Primary Root Count' and a 'Lateral Root Count'. We can combine these to create the "Total Root Count" for the root system, which will be significant later in our analysis.

4.4 ANALYSIS OF TRAITS OF ROOT SYSTEMS

We can also measure how the root system branches and we use the 'Average Lateral Root Count per Parent Root'. This measures how often a root branches out to create a new lateral root. A minor clarification is that we count all primary roots as 'parent roots' even if they do not have any lateral roots yet, since it is very likely that they will grow laterals at a later point. Lateral roots on the other hand, seldom branch off to form 'sub-lateral' roots, so we only include them as 'Parent roots' if they have branched.

4.3.6 *Miscellaneous Statistics*

Another interesting factor is a root strands' 'Tortuosity'. Tortuosity measures how straight or curvy a path is, and is defined as the distance between the start and end points divided by the total distance along the root's path. The result always lies between 0 and 1, with a straight line having a tortuosity of 1. This gives us a simple shape measurement for the root strand.

In addition, we can include metadata. These are not measured traits of the skeletonised root system, but measurements of the plant itself. These include traits such as the 'Plant Height', 'Total Biomass', 'Column Weight' and 'Water Uptake'. Metadata contains important statistics of how well the plant is performing and we will need to analyse both their response to experimental variables and any relationships they might have with traits of the root system.

4.4 ANALYSIS OF TRAITS OF ROOT SYSTEMS

Once we have assembled our measured traits, we now have to investigate how they respond to experimental treatments. The natural approach is to perform linear regression on each of our traits.

If we wish to model the relationship between a measured trait, y and our set of explanatory variables X , we can model this relationship as,

$$y_i = X_i^T \beta + \epsilon_i, \quad (44)$$

where:

- i is the index of the plant.
- y_i is the measured trait y of plant i .
- X_i is the vector of explanatory variables corresponding to plant i .
- β is the vector of unknown parameters.
- $\epsilon_i \sim N(0, \sigma^2)$ is the random error term.

Using this model we can test for significant relationships with our explanatory variables. There is a problem though, we have to be careful of over-fitting. We have a large number of traits to analyse and as we perform each test on each trait, we increase the chance of a Type 1 error (a false positive). This means that we must lower our threshold of what we consider to be acceptable p-values. There are numerous methods for combining hypothesis tests detailed in Section 2.7. Using such methods, we can find a set of traits that we can be reasonably sure are significantly related to our explanatory variables.

4.4.1 *Tukey's Honest Significance Differences Test*

If our experimental design involves division into groups, we can use 'Tukey's Honest Significance Differences Test', or 'Tukey's HSD Test' to analyse group means. Tukey's HSD test analyses the pairwise differences between all possible pairs of means. This has been shown to be useful since often when a significant difference is found, it leads us to question which of the group

means are responsible. This test provides us with a nice way of testing this relationship.

If we specify a significance threshold, then we can think of the group means as vertices in a graph, and if Tukey's test considers them to not be significantly different, then there is an edge connecting them together. Now we can use this graph to create a visualisation, so that we can see the connections between groups. The obvious way of visualising it would be by drawing all the edges, however that may not be the best way. Consider if we had five groups, none of whose means are significantly different, to display this we need to draw $(5 \times 4)/2 = 10$ edges. However if we instead had ten groups, again none of whose means are significant, then we need to draw $(10 \times 9)/2 = 45$ edges, which is disproportionately more edges.

So instead of visualising the graph using edges, we will try to visualise the graph using 'maximal cliques' (Godsil and Royle, 2013). A 'clique' in a graph is a set of vertices where every vertex is connected by an edge. So in our case, a clique would be a set of the group means that are all not significantly different from each other. A 'maximal clique' is simply a clique that cannot be extended by adding in another neighbouring vertex.

The maximal cliques are able to convey information about which vertices are connected without giving undue emphasis to larger maximal cliques. In addition we can use properties of our data set to predict how the cliques will be arranged.

If our dataset is balanced, so that each group has the same number of samples, and we assume homoscedasticity, then the observed differences between group means will all have the same variance. Therefore, when comparing any two group means, $\bar{x}$ and $\bar{y}$, we will fail to reject if $|\bar{x} - \bar{y}| < K$, where K is a threshold dependent on the datasets' size, error term variance, and the desired significance level. Most importantly, if the data is balanced, then the threshold K will be the same for all pairs of $\bar{x}$ and $\bar{y}$.

This means that if we have two group means, $\bar{x}$ and $\bar{y}$, where $\bar{x} < \bar{y}$, that are not significantly different (i.e. $|\bar{x} - \bar{y}| < K$). Then for any group

mean, $\bar{z}$, that lies between them, $\bar{x} < \bar{z} < \bar{y}$, we know that $|\bar{z} - \bar{x}| < K$ and $|\bar{y} - \bar{z}| < K$, and therefore $\bar{z}$ is not significantly different from $\bar{x}$ or $\bar{y}$.

If the dataset was not balanced, then we cannot make this conclusion since the test statistic threshold would be dependant on relative sample sizes of each group being tested. It would be possible for two small groups to not have enough data points to be considered significantly different from each other, but for there to be a third large group between them that does have enough data points to be considered significantly different to the two smaller groups. This means that if we ordered our group means from smallest to largest, then each maximal clique will span a contiguous range without any gaps.

4.5 ANALYSIS OF TRAITS OF INDIVIDUAL ROOT STRANDS

In our analysis of root systems, we run into the problem that our data sets are fairly small, reflecting the fact that growing plants is time consuming and requires a lot of space and resources. This means that there may be subtle differences that we cannot detect because we do not have the statistical power. However with a slight change in perspective, we can increase the effective size of our dataset but with a caveat.

The idea is to analyse traits related to individual root strands. A full-grown plant will likely have at least twenty root strands that can be detected by the LCT scanner. Thus the set of individual root strands is numerically larger than the set of plants. So we can massively increase our available data. So we can transform a dataset of 40 plants into a dataset of 1000 root strands.

However there is a problem. In standard analysis we assume that all of our observed values are independent of each other, outside of their relationship with the explanatory variables. This no longer holds, since there are now large sets of root strands that come from the same plant.

It is entirely possible that during the experiment, one plant may have received slightly more sunlight or been attacked by a very minor infection, which affect its growth positively or negatively. Obviously we try to eliminate such variables but that is not always possible. Normally we assume that any random benefit to an individual in a group may be offset by a random negative to another individual in that group. However since these effects now also apply to a large set of individual root strands, any effect would be amplified.

This means we need to include a set of generic 'plant effect' terms in our explanatory variables for each plant grown. However the exact values of these effect are not reproducible, since if we repeated the experiment we would need to grow entirely new plants with new 'plant effect' terms. In addition, our new 'plant effect' makes any effects related to genotype or the plant's growth conditions redundant since these are contained within the plant effect.

We can solve all these problems by making our 'plant effect' term a 'random effect', and thus make our model a 'mixed effects model'. This means we do not assume any fixed values for our plant effects but that for our experiment they were randomly drawn from a normal distribution with zero mean and unknown variance. This allows our dataset to be reproducible, at least in principle, since the variance of this random effect should not change when the experiment is repeated. In addition, the plant effects will no longer mask any fixed effects related to plants, since the random effects must have a zero mean. If we wish to model the relationship between a root strand trait, y and a set of explanatory variables x , our new mixed effects model is

$$y_i = X_i^T \beta + u_{p_i} + \epsilon_i, \quad (45)$$

where:

- y_i is the measured trait of the i th root strand.

- X_i is the vector of explanatory terms corresponding to the i th root strand.
- β is a vector of the unknown fixed effects.
- P_i is the index of the plant that the i th root strand belongs to.
- $u_{P_i} \sim N(0, \sigma_U^2)$ is the unknown random plant effect for plant i .
- $\epsilon_i \sim N(0, \sigma^2)$ is the unknown random error for the i th observation.

This new model allows us to find relationships involving individual root strands without being affected by plant effects.

4.5.1 *Weighting*

There may be an underlying relationship between the environmental conditions that we wish to analyse and the number of root strands, i.e. a plant may grow more roots when in irrigated conditions than in drought. This is perfectly fine for the normal analysis where we are analysing individual plants, but when analysing individual root strands there will be far more 'irrigated' root strands than 'drought' strands, effectively creating an unbalanced dataset.

In order to get around this, we will weight the individual root strands so that the total weight for the root strands of each plant is equal to 1, i.e. each root strand's weight is inversely proportional to the number of root strands in the plant that it belongs to.

In practice the dataset is still unbalanced, but weighting should help to ensure that relationships with root count do not cause some plants to be over-represented.

4.5.2 *Conclusion*

Our analysis of conventional simple statistics gives us a neat framework to ask questions of a root system dataset and potentially get answers. Our framework is based on well-known classical methods, whose advantages and limitations have been well studied and documented. A full analysis of the wheat root data will be carried out in Chapter 8.

However while this approach will provide invaluable data, as we stated in our introduction, there are some important traits that we cannot reasonably analyse this way. This is why we must expand our approach to less conventional methods, which we will do in the next few chapters.

PAIRWISE DISTANCE MEASUREMENT

Using simple statistics is a very useful way of extracting geometric information from our root shapes, but it is limited.

If plant root systems have different numbers of roots or the way these roots are structured is different, then we cannot continuously deform one root system into another. If we require continuity along individual root strands to be preserved, adding or removing a root will break that continuity, and therefore each root system will have a different topology. Since each plant root system can have a different topology, and the simple statistic's output has a fixed dimension, this means that any continuous simple statistic must discard some information that is related to the plants topology. This does not mean that simple statistics are invalid or not useful, but it does mean that we may miss important features.

Rather than map each root to a statistic in order to analyse differences with a group of roots, we can consider measuring the pairwise distances between roots. This allows us to map the topology of root systems into a metric space. This allows for much more flexibility as the resulting metric space does not have to be Euclidean.

5.1 PAIRWISE DISTANCE MEASUREMENT

In our set of root systems, which we will call R , we wish to measure the 'shape differences' between two root systems, using a function $D : R \times R \rightarrow$

$[0, \infty)$. By 'shape differences', we expect that for two root systems, $\mathbf{r}_1$ and $\mathbf{r}_2$, then $D(\mathbf{r}_1, \mathbf{r}_2) = 0$ if and only if $\mathbf{r}_1$ and $\mathbf{r}_2$ have the same 'shape'. This way we can ensure that no shape information is being lost.

In section 3.4, we discussed what we define to be the 'shape' of a root, and what we consider to be nuisance information. As a quick recap, we consider that 'horizontal translation' and 'horizontal orientation' are considered to be nuisance information that should be ignored. Note that some other factors such as 'size', 'vertical translation' and any 'rotation outside the horizontal plane' are considered to be part of a root's shape, even though in many applications of shape analysis they would also be ignored.

So if we use ' $\mathbf{r} + \mu$ ' to denote the root system $\mathbf{r}$ translated by a horizontal vector μ , then we require that for any horizontal vectors μ_1 and μ_2 ,

$$D(\mathbf{r}_1, \mathbf{r}_2) = D(\mathbf{r}_1 + \mu_1, \mathbf{r}_2 + \mu_2), \quad (46)$$

and if we use $\Theta(\mathbf{r}, \theta)$ to denote the root system $\mathbf{r}$ rotated about the central vertical axis by the angle θ , then we require for any two angles θ_1 and θ_2 ,

$$D(\mathbf{r}_1, \mathbf{r}_2) = D[\Theta(\mathbf{r}_1, \theta_1), \Theta(\mathbf{r}_2, \theta_2)]. \quad (47)$$

Since D is zero if and only if the two root systems being measured have the same shape, then we can be sure that no shape information is being discarded. This is not a perfect approach since it is still possible that some subtle shape differences may be masked by larger, more obvious shape differences. We can reduce the impact of this by our choice of D , and by having a large sample size such that there is more likely to be pairs of plants with only the subtle shape differences.

We ideally wish for our function D to be a metric but currently that is not possible since $D(\cdot, \cdot)$ may be zero for two different root systems, which contradicts a necessary requirement for metrics. However this will only happen for two root systems with the same shape, so with a small change in perspective we can make it a metric.

5.1.1 *Root System Shape Space*

Using the notion of 'shape', we can create an equivalence relation ' $\sim$ ', such that $\mathbf{r}_1 \sim \mathbf{r}_2$ if and only if $\mathbf{r}_1$ and $\mathbf{r}_2$ have the same shape. Using this we can partition root systems into equivalence classes and build a quotient space $S = R/\sim$. This new space will represent the shape of root systems.

The elements of our shape space S are sets of root systems that have been rotated and translated horizontally. We will denote these sets as $\{\mathbf{r}\}$, where r is a representative element within the set, and the set consists of all root systems that can be rotated and translated horizontally to be equal to r .

We can naturally extend D to this new space shape by defining $D' : S \rightarrow [0, \infty)$ to be,

$$D'(\{\mathbf{r}_1\}, \{\mathbf{r}_2\}) = D(\mathbf{r}_1, \mathbf{r}_2) \quad (48)$$

which will always be unambiguously defined regardless of the choice of $\mathbf{r}_1$ and $\mathbf{r}_2$, because equations 46 and 47 ensure we ignore horizontal rotation and translation. Our goal is therefore to choose D such that D' will be a metric on S .

 5.1.2 *Adapting metrics to ignore nuisance differences*

If we have a metric, $F : R \rightarrow [0, \infty)$, that measures differences between root systems and does not ignore the nuisance parameters of horizontal translation and rotation, then naturally we will wish to adapt it to create a shape metric as above.

In other words for a metric on the root systems $F : R \times R \rightarrow [0, \infty)$, we wish to create an equivalent metric $F' : S \times S \rightarrow [0, \infty)$ on the root system shape space.

Let S_1 and S_2 be shape equivalence classes of root systems in S , then we can define our F' as being the smallest possible distance between any two pairs of root systems chosen from S_1 and S_2 ,

$$F'(S_1, S_2) = \inf_{\mathbf{r}_1 \in S_1, \mathbf{r}_2 \in S_2} F(\mathbf{r}_1, \mathbf{r}_2). \quad (49)$$

If F satisfies the conditions of being a metric, then F' will satisfy the following three properties,

1. $F'(S_1, S_2) \geq 0$
2. $F'(S_1, S_2) = 0$, if and only if S_1 S_2 are equal or can be rotated to match each other.
3. $F'(S_1, S_2) = F'(S_2, S_1)$

However the triangle inequality does not automatically transfer from F to F' , since it is possible for a path through a third equivalence class S_3 to cover less total distance, if it has two elements close enough to S_1 and S_2 . This property would need to be proven separately for a given choice of F .

5.1.3 Finding the minimal shape distance

Horizontal Translation

We begin by eliminating the horizontal translation. This is where, for practical considerations, we will deviate slightly from our theoretical approach outlined above in order to simplify calculations.

For each root system we have a single common point, the origin point, which is defined as the point where the plants stem breaks through the soil. This is defined for every root system, and we can easily translate a root system so that their origin points match. We will consider this to be enough to eliminate any horizontal translation differences.

For simplicity, we will simply translate all root systems so that the origin points are at the 'zero' point in 2D/3D space. This is not mathematically the same as our approach above but we feel that it is justified since, from observation, root systems appear to be 'built' around the origin point, and any shape differences start appearing immediately from that point.

Horizontal Rotation

If we have translated two root systems horizontally to have a common origin point, all that is left is to find rotations that minimise our metric distance.

So for two root systems that have been translated, $\mathbf{r}_1$ and $\mathbf{r}_2$, we wish to minimise $F [\Theta(\mathbf{r}_1, \theta_1), \Theta(\mathbf{r}_2, \theta_2)]$ over two angles θ_1 and θ_2 .

We can simplify this problem significantly if our metric is already rotationally invariant, so that for an angle θ ,

$$F [\Theta(\mathbf{r}_1, \theta), \Theta(\mathbf{r}_2, \theta)] = F(\mathbf{r}_1, \mathbf{r}_2), \quad (50)$$

then the distance will only be dependent on the difference between the two angles,

$$F [\Theta(\mathbf{r}_1, \theta_1), \Theta(\mathbf{r}_2, \theta_2)] = F [\mathbf{r}_1, \Theta(\mathbf{r}_2, \theta_1 - \theta_2)]. \quad (51)$$

So we only need to minimise $f(\Delta\theta) = F [\mathbf{r}_1, \Theta(\mathbf{r}_2, \Delta\theta)]$ over a single angle, $\Delta\theta$, which is a much simpler proposition than minimising over two angles.

From observation we have found that in general, $f(\Delta\theta)$ is not a simple function and will have multiple local minima. So our approach will be to evaluate f at 1° increments across the 360° range to find a good starting point for minimising towards a local minimum.

5.2 DISTANCES BETWEEN LINE SEGMENTS

If we are representing our roots as a sequence of straight line segments, we need a method of efficiently measuring the distance measures between line segments and combining them into a distance measure for the root overall.

5.2.1 *Segment Reduction*

Our roots contain a large number of roots which have very detailed paths that use a large number of line segments to accurately represent them. However there are circumstances (such as when finding the rotation that minimises a distance between two roots) where we only need an approximate distance. Thus if multiple line segments are nearly parallel, they can be merged together and will not lose much accuracy.

5.2.2 *Minimum distance*

Finding the minimum distance between two lines is a fairly simple case. It is a simple calculus task to find points on two (infinite length) lines that minimize the distance. We can perform this on the extended lines from the line segments and then find the closest points that are actually on the line segments. Since the distance function is uni-modal, these points on each line segment are the closest pair of points.

We can improve performance by checking if the two line segments are almost parallel or if one of the segments is small enough to be considered to a point, and making a special case if so. We can test this by calculating $D(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{u})(\mathbf{v} \cdot \mathbf{v}) - (\mathbf{u} \cdot \mathbf{v})^2$, where $\mathbf{u}$ and $\mathbf{v}$ are vectors of the two line segments and $(\cdot)$ is the dot product, and checking if $D(\mathbf{u}, \mathbf{v})$ is less than an arbitrary small threshold, ϵ .

5.2.3 Total Minimum Distance

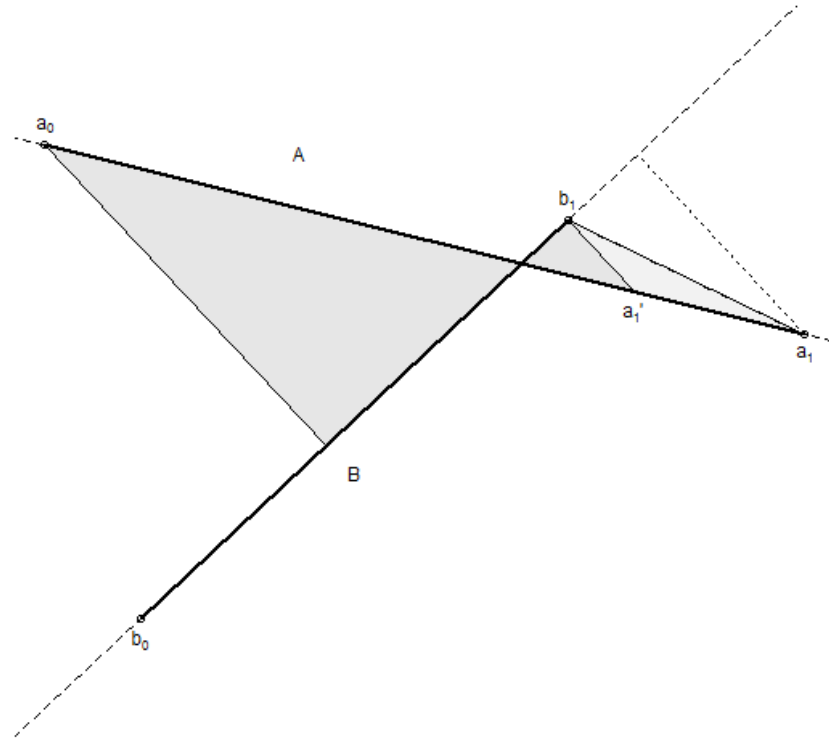


Figure 21: Diagram outlining the points that need to be handled for calculating total minimum distance from line segment **A** to line segment **B**.

This is a more complex scenario that requires a more sophisticated algorithm. Here we wish to find the integral of the 'minimum distance from a point of line segment **A** to any point of line segment **B**' across the range of the line segment **A**. Unlike before, this is not a symmetric situation, and the answer will be different if **A** and **B** were reversed.

An illustrative diagram is shown in Figure 21 that shows the points that need to be calculated. Bear in mind that in Figure 21, the two lines are effectively in the same 2D plane but this will not be the case in most 3D cases.

5.3 HAUSDORFF DISTANCE

If $\mathbf{a}$ is a point of line segment $\mathbf{A}$, $p(\mathbf{a}, \mathbf{B})$ is the closest point on line segment $\mathbf{B}$ to point $\mathbf{a}$ and $|\cdot|$ is the Euclidean distance metric, we wish to compute,

$$\int |\mathbf{a} - p(\mathbf{a}, \mathbf{B})| d\mathbf{a} \quad (52)$$

Let $\mathbf{A}'$ and $\mathbf{B}'$ be the extended infinite lines that contain $\mathbf{A}$ and $\mathbf{B}$ respectively. For each end point of line segment $\mathbf{A}$, denoted $\mathbf{a}_0$ and $\mathbf{a}_1$, we can find the corresponding closest points on line $\mathbf{B}'$. Since $p(\mathbf{a}, \mathbf{B})$ is a linear function of $\mathbf{a}$ this means that for $\mathbf{a} = \mathbf{a}_0 + t[\mathbf{a}_1 - \mathbf{a}_0]$,

$$p(\mathbf{a}_0 + t[\mathbf{a}_1 - \mathbf{a}_0], \mathbf{B}') = p(\mathbf{a}_0, \mathbf{B}') + t[p(\mathbf{a}_1, \mathbf{B}') - p(\mathbf{a}_0, \mathbf{B}')]. \quad (53)$$

Thus we can integrate t from 0 to 1 to get our answer, and this is a fairly simple task. It is equivalent to finding the area of the quadrilateral formed by points $\mathbf{a}_0$, $\mathbf{a}_1$, $p(\mathbf{a}_0, \mathbf{B}')$ and $p(\mathbf{a}_1, \mathbf{B}')$. However, it may be that $p(\mathbf{a}_0, \mathbf{B}')$ or $p(\mathbf{a}_1, \mathbf{B}')$ are not points within line segment $\mathbf{B}$. In which case, we can find $\mathbf{a}'_0$ and $\mathbf{a}'_1$ for which $p(\mathbf{a}'_0, \mathbf{B}) = \mathbf{b}_0$ and $p(\mathbf{a}'_1, \mathbf{B}) = \mathbf{b}_1$. Between $\mathbf{a}'_0$ and $\mathbf{a}'_1$ the procedure is the same as before, and we only have to consider the sections outside this range.

We can assume without loss of generality that $\mathbf{a}_1 \neq \mathbf{a}'_1$. For any point between them, the closest point on $\mathbf{B}$ will be $\mathbf{b}_1$. Unfortunately integrating the distances across this small subsection is not equivalent to finding the area of the triangle formed by $\mathbf{a}_1$, $\mathbf{a}'_1$ and $\mathbf{b}_1$, otherwise we could join it with the main quadrilateral.

5.3 HAUSDORFF DISTANCE

The Hausdorff distance is the main distance metric we will be using and was the motivating force behind this method.

The Hausdorff distance is discussed and defined in Section 2.8, but as a brief reminder, the Hausdorff metric measures the difference between two

sets as the “largest distance between an element of one set to its closest neighbour in the other set”, i.e. the worst-‘best distance’.

This metric has several very useful properties that will be useful in our analysis of root systems. The metric is continuous, so a ‘small’ change in either root system (including the addition or removal of a ‘small’ root strand), will, at worst, result in a similarly ‘small’ change in the Hausdorff distance. This means the method is resilient to small errors in measurement. The metric is also invariant to a common translation and rotation, and responds linearly to a common scaling factor.

One disadvantage of using the Hausdorff distance is that it does not always respond to changes. The Hausdorff distance is defined using the two points that have the worst ‘best distance’, and if there is a change in the root systems that do not affect these two points or change which two points have the worst ‘best distance’ then there will be no change in the Hausdorff distance. This means that in a local space around the two root systems being considered, a lot of points are ignored when determining the Hausdorff distance.

We would like it so that any change in either root system would be represented by some change in the distance between them. This ensures that potentially useful information is not being needlessly discarded.

5.3.1 *Alternative*

The Hausdorff distance is effective and simple to calculate but has the major disadvantage that changes to the input can fail to result in any change in the output (since the calculation focuses on the distance between the two most extreme points). We must also consider a variation of the Hausdorff distance that is always dependent on all points.

We will be considering what we will call the ‘Pseudo-Hausdorff distance’. Where the Hausdorff distance is defined as the ‘worst best distance’, the Pseudo-Hausdorff is defined as the ‘average best distance’. So for two fi-

nite sets A and B , with an underlying distance metric $d(\cdot, \cdot)$, we define the Pseudo-Hausdorff distance as,

$$H_p(A, B) = \frac{1}{2} \left(\frac{1}{|A|} \sum_{a \in A} d'(a, B) + \frac{1}{|B|} \sum_{b \in B} d'(b, A) \right), \quad (54)$$

where, d' is the partial Hausdorff distance for a point a to a set of points B ,

$$d'(a, B) = \min_{b \in B} d(a, b). \quad (55)$$

5.4 MULTIDIMENSIONAL SCALING

We can use MDS analysis (as discussed in 2.3), to find the set of n points in k -dimensional Euclidean space that closest matches a given $n \times n$ distance matrix.

As we increase the number of dimensions k , the resulting MDS distance matrix will get closer to a minimal-distortion representation of the sample distance matrix.

We can use this Euclidean set as a stand in to analyse our set of roots for statistical differences, using traditional linear methods such as analysis of variance.

5.4.1 *Choice of MDS k value*

An important factor in creating a set of MDS variables, is the choice of k . This is a balancing act, as if k is too small, then our MDS variables may omit important differences. However as k gets larger, the larger MDS variables become numerically unstable and sensitive to small changes, to the point that they become effectively meaningless in terms of the underlying population.

Deciding on an optimal value for k is not a simple task, as it is a result of the underlying population distribution that we cannot directly observe. In addition, the choice of k will be dependent on the how many samples we have and we will need to determine a correct value of k for each value of n .

Since each MDS variable is independent we can analyse how much variance each variable contributes to the total. The variance of the k th MDS variable, is non-increasing with k . In order to properly choose k , we need to ensure that each MDS variable up to k contributes enough to the total variance to ensure they are significant, and the sum of those contributions represents enough of the total variance. We will denote the variance of the i th MDS variable as σ_i^2 and the total variance as $T = \sum_{i=1}^n \sigma_i^2$. So we wish to find the largest value of k such that, for a specified 'contribution parameter' $\delta \in (0, 1)$,

$$\sum_{i=1}^k \sigma_i^2 \geq \delta T, \quad (56)$$

and also for a 'minimal variance parameter', $\epsilon \in (0, \delta)$,

$$\forall_{i \in 1 \dots k} : (\sigma_i^2 \geq \epsilon T). \quad (57)$$

Deciding on an appropriate value for k is a difficult problem, and there are potential dangers in having the value too high or too low. If we wish to minimise the possibility of a false significance, it is better to err on the side of a smaller k .

5.5 ANALYSIS APPROACH

Once our data has been transformed into a conventional set of k MDS variables, we need to be able to ask questions of it. We do this by treating our MDS variables as our representatives of each plant root system.

So, for example, if we wanted to see if two groups of plants had different shaped roots, then we can pool the two groups together and calculate a

5.6 CONCLUSION

pooled distance matrix, and corresponding MDS variables. We can then test our MDS variables for a change in distribution between the two plant groups.

Testing the MDS variables can be accomplished using a multivariate analysis method, such as Hotelling's t -squared statistic. This can test to see if the two groups have different means for their MDS variables, which would be enough to show that there are significant shape differences. Of course, there may be the case of the two groups having equal means but other distribution differences (such as different variances) that you can test for, however that would require a much larger sample size to work with, so that may not be practical.

This approach does a good job of testing for significant shape differences, but it is quite lacking when it comes to interpretation. It is difficult to identify what differences in the MDS variables signify in practical terms for our dataset, although plotting individual roots along each MDS axes can help to give some explanation.

5.6 CONCLUSION

By focusing on pairwise distances between root systems, we can bypass the restrictions that traditional simple shape statistics have with regards to topological or structural differences. The Hausdorff distance is a perfect fit for this approach and has been used in image analysis before (see Section 2.8).

Given a set of pairwise distances we can convert them to a conventional dataset using the Multidimensional Scaling method. We still haven't determined the proper value of k to be used in the MDS method but that is a question we will try to answer using a simulation study in Chapter 7.

The pairwise distance approach has so far focused on shape differences. In the next chapter we will be looking at Topological Data Analysis, which

5.6 CONCLUSION

focuses on the topology of our shape and we will use this to create a new pairwise distance measure that emphasises topological differences.

TOPOLOGICAL DATA ANALYSIS

Topological Data Analysis (TDA) is a new approach in statistics and a very active area of research. We hope that we can apply its techniques to our plant root system shape data.

Plant roots are very difficult to analyse since we need to account for the differing topology of each root system. Up until now, we have taken approaches that downplay topological differences and choose to focus on the pure geometry of the roots. However with Topological Data Analysis, we have a method that will put an emphasis on topological trait differences.

It is important to understand that despite the name, Topological Data Analysis is not a purely topological method and does depend on geometric information. This means TDA could, for example, consider a doughnut and a coffee cup to be different objects despite having the same topology. This is not a problem however, just a simple misunderstanding of what the approach is trying to accomplish. The goal is to use topology methods to analyse data and not to solely analyse the topology of the data on its own.

6.1 PERSISTENT HOMOLOGY

The chief method we will be using is Persistent Homology, which is described in Section 2.9. This takes a point cloud and a filtration method, and uses them to find a set of 'homology classes'.

The ‘filtration method’ is used to progressively ‘blur’ the point cloud. So as the filtration parameter increases, the points in our cloud connect together and form lines, loops, holes, faces, voids, etc. until, eventually, every point in the cloud is connected together in one big solid object that has the topology of a sphere. The goal of Persistent Homology is to analyse how this ‘blurred’ representation of the point cloud changes as the filtration parameter increases.

The Persistent Homology technique results in a set of ‘intervals’. Each interval has a ‘start time’ and an ‘end time’, and is associated with a ‘homology class’. These sets of intervals are sometimes called ‘barcodes’ or ‘persistence diagrams’, which is motivated by how they are plotted. For consistency, we will refer to them as ‘persistence diagrams’.

In theory, if two point clouds are randomly sampled from some underlying shape, then they will have similar persistence diagrams, and as more points are added they will converge to a common diagram.

Understanding and interpreting what a ‘homology class’ is and what it represents in the point cloud is difficult, and this is a major drawback to TDA. In loose terms, a ‘homology class’ represents either a ‘feature’ or a ‘hole’ in the shape. The corresponding ‘interval’ of a homology class then represents when the class first appears during the filtration, and when it disappears. The idea is that if a class has a very short interval, it is not an important feature and likely be noise, but if the class has a very long interval then it is likely to be important to the structure of the point cloud.

One complication of using this approach is that our input data is not a point-cloud, but a collection of lines or curves. This means that some adjustments will be needed to ensure the technique works properly.

6.2 CONVENTIONAL PERSISTENT HOMOLOGY

6.2.1 *Converting root systems to point clouds*

Since conventional persistent homology is performed on a point-cloud, a simple solution is to represent our set of lines using a point cloud.

The method of how we convert to a point cloud is important, because the choice of method can have an effect on the result. It is important that our approach is as consistent as possible, so that comparison is fair. So if two root systems have similar shapes they should have similar point clouds.

The number of points in our cloud, n , is also important as it can affect the result, a large number is typically better for accuracy. However Persistent Homology calculations can get computationally expensive for large point-clouds, so a large n may not be feasible.

We wish to have our points to be as spread out evenly as possible across the root system, and to have the distance between successive points be as consistent as possible. If we choose an average distance d , then we can get the corresponding n by dividing the total root lengths in the system by d .

For points in a line this is a fairly simple task as we can choose k evenly spaced points easily. A more difficult case is when multiple lines intersect we must ensure that these points are not over-represented by assigning points to each line. If we are not concerned with having 'exactly' n points and do not mind an approximation, we can simply merge points in the intersection if we feel the intersection is over-represented.

6.2.2 *Delaunay Triangulation*

The Cech complex and VietorisRips complex are standard methods for creating persistent homology filtrations, but they are difficult to compute, especially given a large number of points in our set. The Delaunay triangulation

is a useful way to reduce the filtration to be much simpler. An example of applying the Delaunay triangulation to a root system is shown in Figure 22.

Let n be the number of points and d the number of dimensions. The number of simplexes in the Delaunay triangulation is $O\left(n^{\lceil d/2 \rceil}\right)$, (Seidel, 1995), which is much less than the total simplex combinations formed in the Vietoris-Rips complex, $O(n^{d+1})$.

Using the Delaunay triangulation as the base complex and building a filtration on it, is much simpler and faster than the Vietoris-Rips or Cech methods.

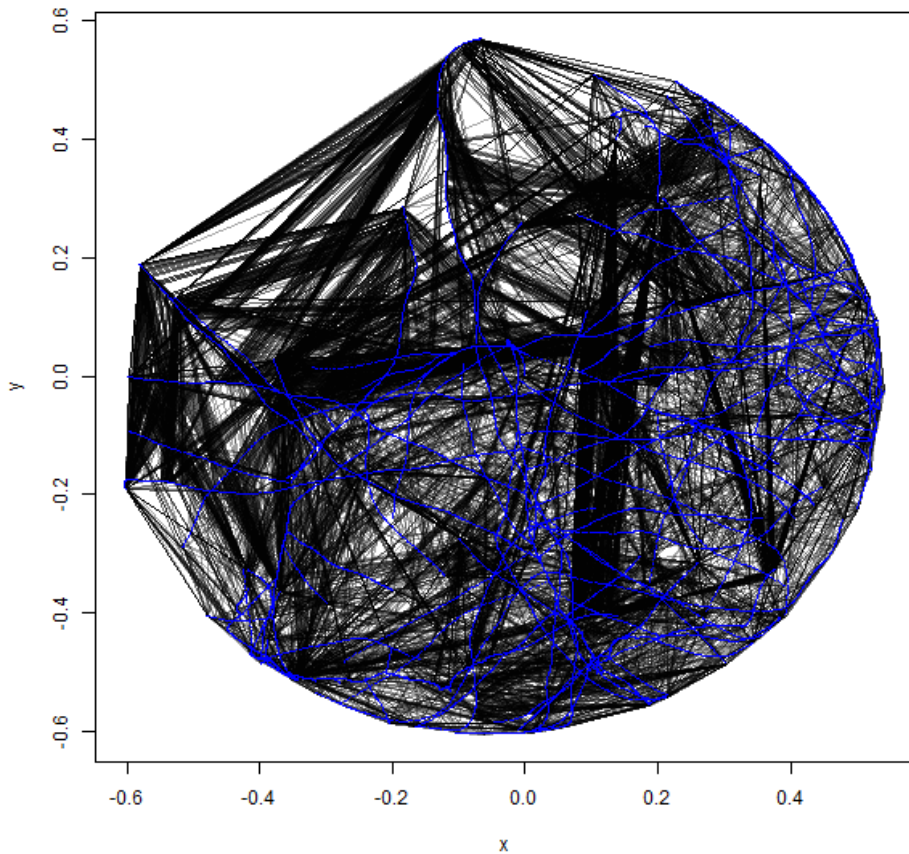


Figure 22: A topdown view of the Delaunay Triangulation of a point-cloud version of a root system. The original root system is shown in blue. Tetrahedrons with negligible area have been omitted.

6.3 LINE-BASED FILTRATION

An alternative approach when working on root systems is to treat the subset of lines in Euclidean space as the space on which the homology is performed. What this means in practice, is that a filtration is created by assigning every point a number, and then looking at the homology of the growing subspace formed by the filtration.

This allows us to create a 'persistence diagram' for that root system that does not depend on any choice of representation (which converting to a point-cloud will have).

Since root systems will not have any loops or voids in them, then neither will the filtration subspaces. Thus we only need to calculate the homology for the '1-dimensional holes'.

The main disadvantage with this approach is that the subspace may not capture all information related to the topology of the root system. It also does not have the guarantee of stability that the point-cloud approach has.

6.4 ALTERNATIVE POINT CLOUD REPRESENTATIONS

Our approach includes every point in the root system as being a significant point. Since most of the root system is considered to be a straight line, this means there is no distance between points and it is much harder for large holes to form. While this approach is fine, we can consider an alternate approaches that only include select 'important' points in our point cloud representation to allow for greater distance between points.

One simple approach is to consider using only the start and end points of each root. We may assume that this would give more emphasis to smaller lateral roots, since there are more of them and therefore more points; however the fact that the end points of the primary roots are much further away from nearby points means they will have a much larger persistence class.

6.5 COMPARING PERSISTENCE DIAGRAMS

A persistence diagram is a multiset of points, $\{(u, v) \in \mathbb{R}^2 : u, v \geq 0, u \leq v\}$. The assumption is that the larger the difference between u and v , then the more 'important' the feature is. We can imagine an infinite number of zero-length points on the diagonal, (v, v) , that do not correspond to any feature.

Comparing two diagrams and working out a distance between them is an open area of research and standard methods may not perform well in all cases. Since our data is very unusual we will be considering some alternative methods to see if they work better.

6.5.1 *Alternative approaches**k most significant classes*

The key issue that makes comparing persistence diagrams difficult is that different persistence diagrams have different number of 'points' within them. We can bypass this problem by choosing k points from each diagram to act as representatives for their diagram. This does mean discarding information, but it does simplify analysis.

We select the k points that have the greatest persistence ($\|x_U - x_L\|$), since these are most likely to represent significant topological features. We order these points from greatest persistence to least persistence, $\{x_{(1)}, \dots, x_{(k)}\}$ (where $\|x_{U_i} - x_{L_i}\|$ is decreasing in i).

We can then define a distance between two diagrams as,

$$D_p(X, Y) = \left[\sum_{i=1}^k |x_{(i)} - y_{(i)}|^p \right]^{\frac{1}{p}}. \quad (58)$$

Distribution analysis

Here we treat the internal sizes of the points in the multiset as a random sample from a distribution. Then we can use the estimated deviance between these distributions as a distance measure.

We can also easily test for differences in mean and variance as a simple hypothesis test. This test can be extended to allow us to test multiple persistence diagrams simultaneously.

6.6 LIMITATIONS OF TDA

Topological Data Analysis is a new and active area of research and unfortunately there are still some limitations to its effectiveness. The biggest issue is that it is difficult to interpret and draw conclusions from results. This is not surprising since Homology groups are difficult to visualise and interpret on their own, so differences between them are very unlikely to be easily interpreted either.

This leads to the common impression that TDA is a 'black box', where you insert some data, it outputs a result, but it is impossible to understand what that result means for the dataset or how that result came about.

We can help improve this by attempting to visualize the homology generators. This is not simple to do with the filtrations built on Euclidean space (such as Chech and Vietoris-Rips filtrations) but we can find an approach with the line filtrations that are built on the root system itself.

6.7 INTERPRETATION OF PERSISTENT HOMOLOGY RESULTS

The biggest problem with TDA as it stands, is that it has a 'black box' nature. You put data into the method and you may get a significant result but it is very difficult to understand how this result may relate back to the original data.

6.7.1 *Visualizing Homology Classes*

In Figure ??, we see the homology generators of a line filtration based on the vertical distance.

The first row pictures the root with each point coloured from blue to red to show the ascending values of the filtration at that point.

The second row shows the homology generators that are found using Persistent Homology method. They are plotted as a Persistence diagram and a Bar chart. Each generator is coloured with a randomly chosen colour.

The third row shows the root but each point is associated with a Homology generator. The idea is that a subset of points can be associated with each Homology generator. Each of the subsets is a 'group generator element' for each of the persistent homology groups that the particular generator spans (and only those groups). These subsets can be easily found by tracking when the classes of the filtration merge together.

The segments of root that change vertically are associated with large homology generators. This shows that they are very significant features in the context of the vertical distance filtration. The central vertical stem is the largest feature, which is as expected.

The segments of root that are mostly horizontal (and vary only a little vertically) however, tend to be broken up into small classes. This means that these are not significant features in the context of the vertical filtration and can be ignored.

6.8 VISUALIZING LOCATIONS OF HOLES IN THE VIETORIS-RIPS

We can use a more complex version of this process to visualize the locations of holes in the Vietoris Rips complex. This is done by choosing a representative cycle for each homology class that best encloses a hole at its peak size.

Unlike before, these holes are not constrained to be contained on the root system.

The question is how do we interpret holes in the root system. A hole is formed when roots are close at one point, become distant, and then get close to each other again. This can be interpreted as an inefficiency in the system as the roots should ideally be spreading out further and further.

In Figures 23 and 24, we can see two root systems with very different distribution of 'holes'. The first has more holes and the holes are larger in size. Conversely, the second has fewer holes and they are smaller. The key difference is that while both systems have three large primary roots; the second system has one primary growing down and two of its roots spread outwards in different directions, and the first system has three roots that initially spread out at evenly spaced angles, but then grow downwards creating a void in the centre.

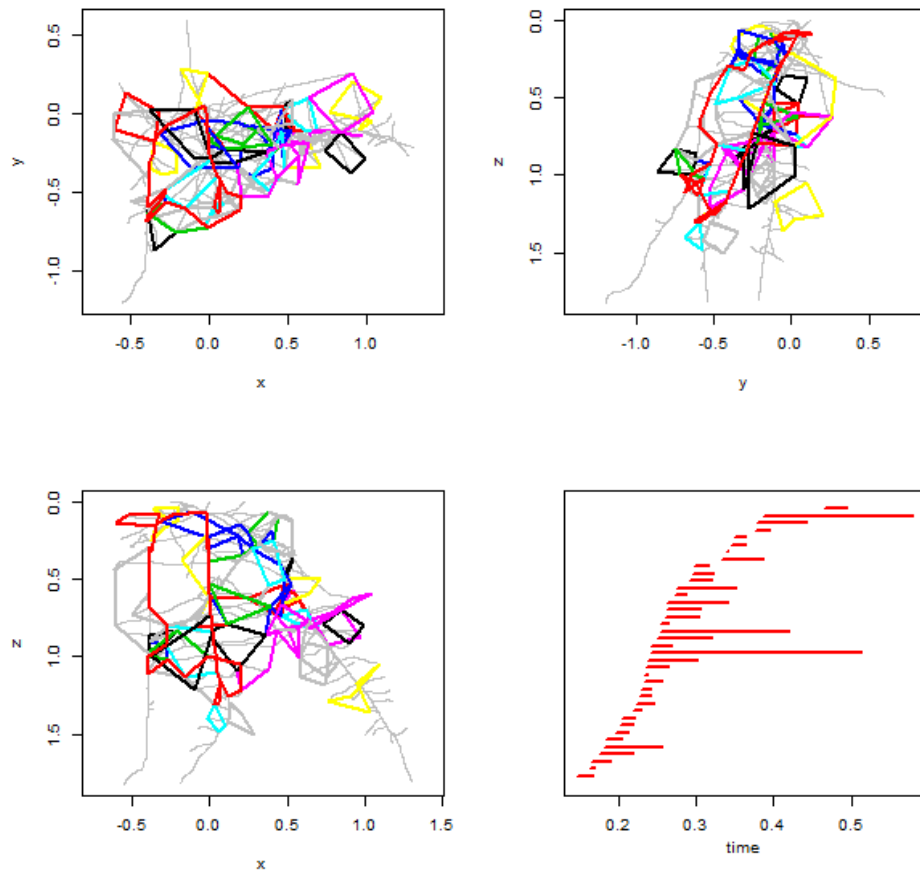


Figure 23: Visualization of the holes found by the Veitoris-Rips persistent homology method. There are some very large holes that span the root system.

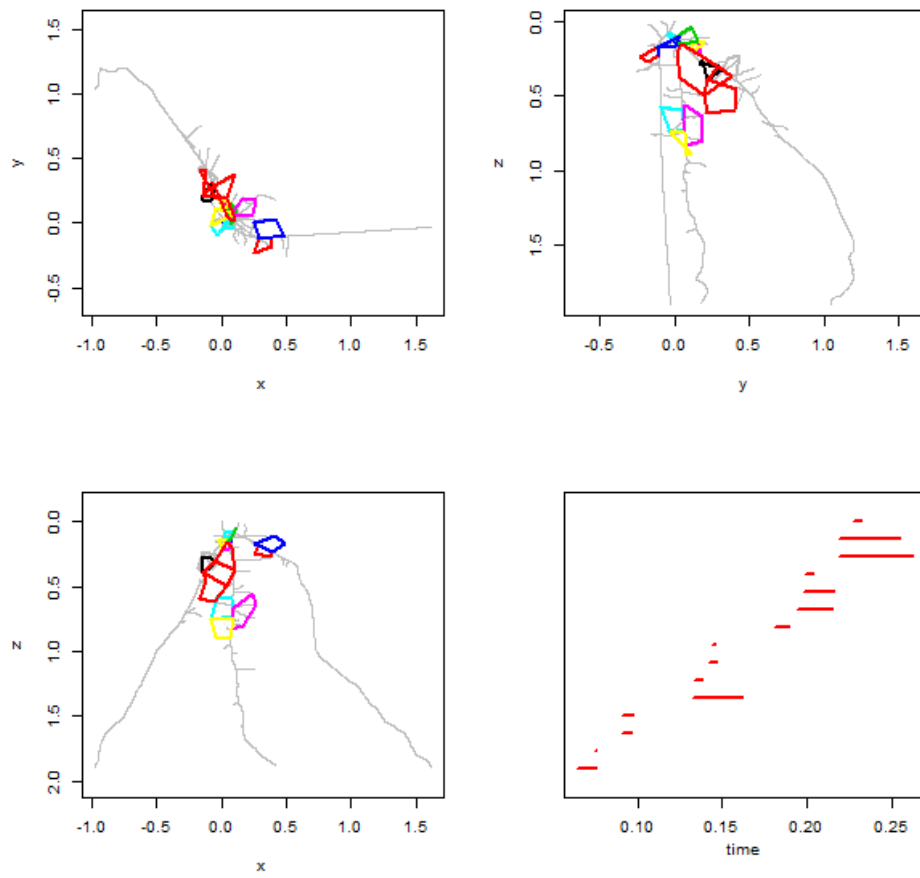


Figure 24: Visualization of the holes found by the Vietoris-Rips persistent homology method. The holes are much smaller.

6.9 CONCLUSION

Topological Data Analysis presents a new perspective to analyse data, that emphasises topological differences between structures rather than geometric. It is our hope that this new topological perspective will assist us in identifying new phenotypes that have not been identified before.

In the next chapter we will conduct a simulation study to see how the Pairwise Hausdorff and TDA methods perform on simulated root system data.

SIMULATION STUDY

7.1 SIMULATED DATASET

We will be using a standard simulated dataset from Lobet et al. (2016). This is a dataset of 10,000 simulated RSML roots, using a variety of input parameters. The original intention of this dataset was to provide a standardised dataset for analysing the performance of image analysis methods.

The data was generated using two different algorithms that simulate 'monocot' plants and 'dicot' plants respectively. There are many subtle biological differences between these two types of plants, however the main difference that we are interested in, is that a 'monocot' root systems are fibrous and have multiple large roots branching from the seed, whereas 'dicot' root systems have a single main root called the 'tap root'. The wheat plants we are studying are 'monocot' plants.

There are 5,000 monocot and 5,000 dicot plants in total. Each are divided into 250 groups of 20 plants, where each group has different randomly generated parameters.

7.2 COMMON CASES

It is unlikely that we will be analysing a dataset of 10,000 3D roots, due to the expense of acquiring 3D scans. We expect any datasets to be comprised

7.3 PAIRWISE DISTANCE APPROACH

of around 20-50 plants. In addition we will be focusing on wheat roots that are 'monocot' plants.

In order to better simulate the datasets we will likely be encountering, we will only be looking at small subsets of the 5,000 simulated monocot roots.

There are three scenarios that we wish to consider.

1. Two randomly selected 'simulation groups'.
2. Two groups of randomly selected plants, sampled from any the whole dataset.
3. Two groups of randomly selected plants, sampled from a single randomly chosen 'simulation group'

Each of these scenarios gives two groups, consisting of 10 plants each. Now, simply simulating a given scenario once is insufficient, so we simulate each scenario hundreds of times, and see what variety of results we get.

We will use linear discriminant analysis to see how well our pairwise methods perform in these three scenarios. The goal is to use the first k MDS variables to predict which group the plant came from. We will use cross validation to determine how well the LDA performs. We can then use the performance of these LDA to choose a value for k .

Note that we have not concerned ourselves with verifying the assumptions of LDA. The purpose of this analysis is not to create a simple measure of fit quality for each method, and not to suggest the model is accurate.

7.3 PAIRWISE DISTANCE APPROACH

7.3.1 *Elimination of Angle Differences*

In order to eliminate the factor of horizontal rotation, we fix one root system and rotate the second root system about the vertical axis by the angle θ to find the minimum pairwise distance.

There are several questions we need to answer. The first is how many values of θ to we need to evaluate. We need to find the global minimum of the function $\theta \rightarrow \mathfrak{R}$. This function is continuous but may not be smooth.

7.3.2 Hausdorff Distance

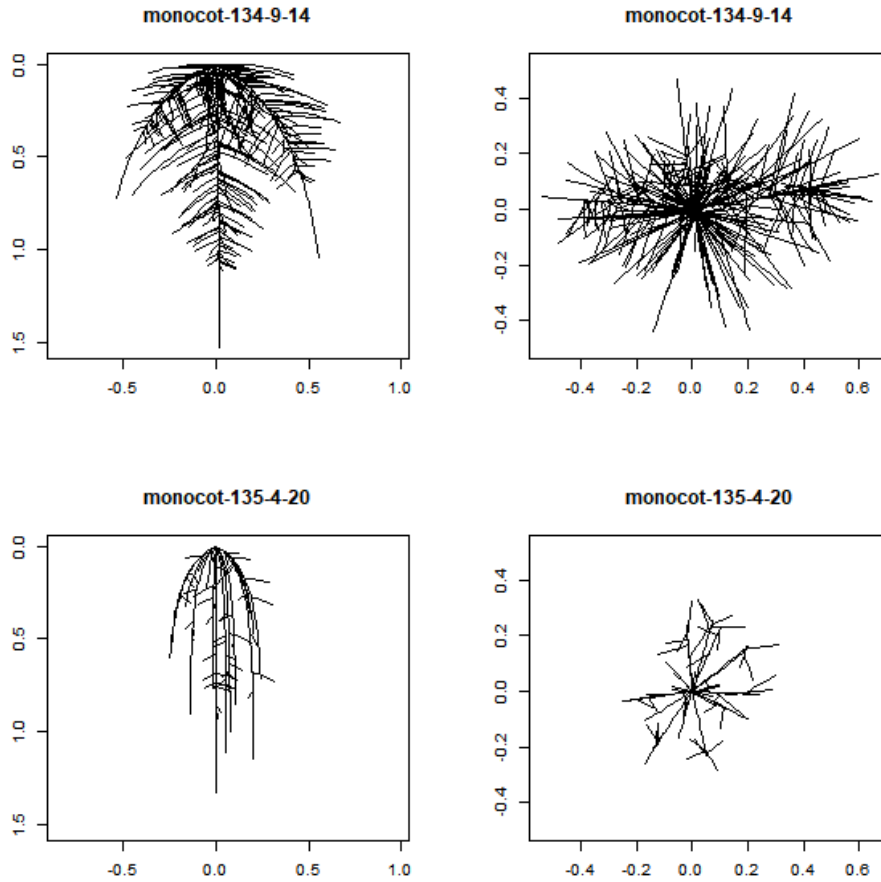


Figure 25: Plot of two different root systems, “monocot-134-9-14” and “monocot-135-4-20”. Side views on the left, top-down views on the right.

We can see an example of this rotation function in figure 26, and it’s clear that the function is very unpredictable. There are multiple local minima across the range of θ and two of them are close to being the global minimum.

If we can start at a point close to the global minimum, then we can use a steepest descent method to optimize across this function. Finding a point

7.3 PAIRWISE DISTANCE APPROACH

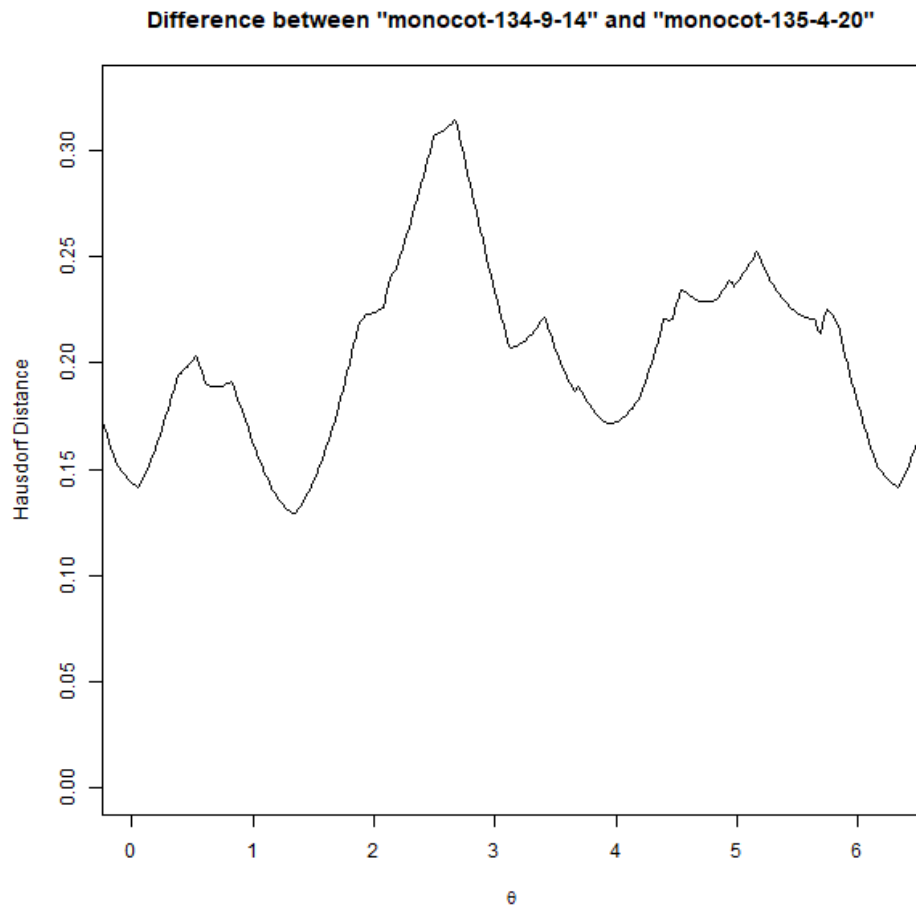


Figure 26: Effect of rotation on the pairwise Hausdorff distance between two different plant root systems (see Figure 25). In this case, rotation has a large effect on the distance measure.

7.3 PAIRWISE DISTANCE APPROACH

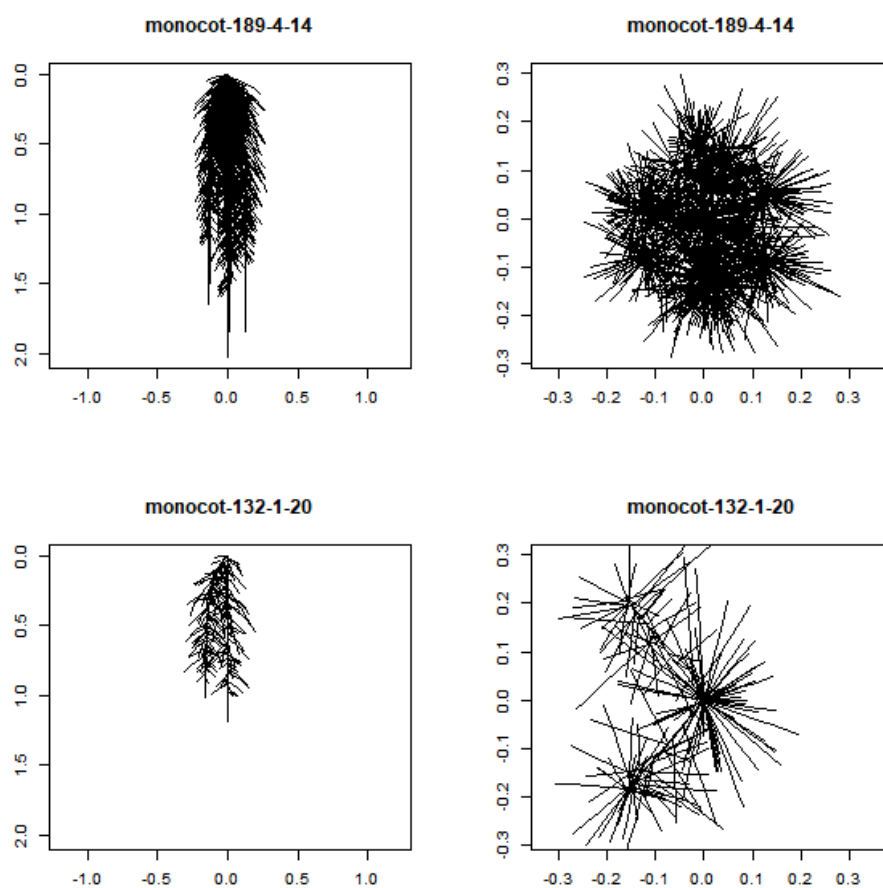


Figure 27: Plot of two different root systems, "monocot-189-4-14" and "monocot-132-1-20". Side views on the left, top-down views on the right.

7.3 PAIRWISE DISTANCE APPROACH

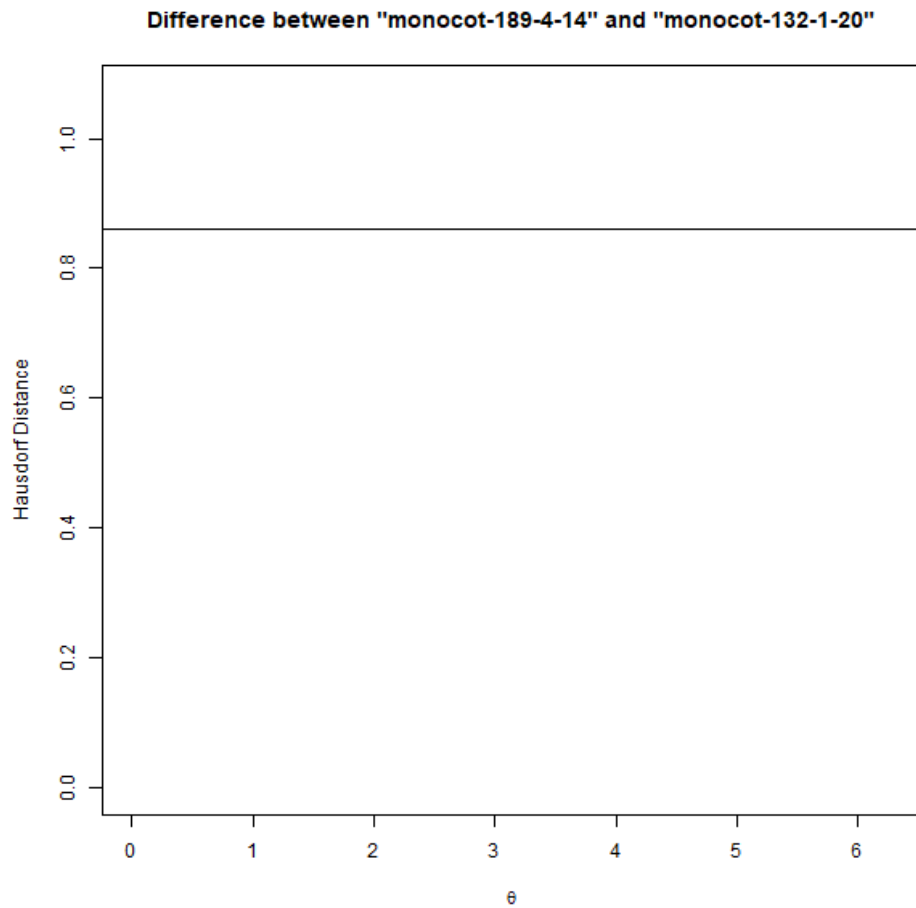


Figure 28: Effect of rotation on the pairwise Hausdorff distance between two different plant root systems (see Figure 27). In this case, rotation has little effect.

close to the global minimum can be found by evaluating the function at angles θ degrees apart along (where θ is small) and using the smallest angle as a starting point.

However with the Hausdorff function there is a special case to this, where rotation does not have a large effect. For an example, see figure 28, where the function appears to be a straight line (it is not a straight line if we looked closer). What has happened here is that both roots have a long straight central stem that goes down from the origin point, and the tips of these stems are their most extreme points. One of these plants stems is much longer than the other, and the Hausdorff method identifies these two points as being the most important extreme differences. Since the two points are so close to the vertical axis, horizontal rotation has little effect on them and so the rotation function appears to be flat.

In terms of performance under LDA, we can see in Figure 29, that the method is quite successful at assigning values in case 1, even when $k = 1$. As k increases, the probability of correct classification falls a bit, but this is expected as later MDS variables are more noisy and less likely to add useful information. There seems to be a peak around $k = 3$, but given the large confidence intervals this is not significant.

In cases 2 and 3, the LDA achieves a roughly 50% success rate. This is what we expected to see, since the MDS variables were not related to the arbitrary group assignments, so this should perform about the same as a coin toss.

7.3.3 *Pseudo Hausdorff Distance*

We can apply the same rotation analysis to the Pseudo-Hausdorff distance metric. We can see equivalent examples in Figures 30 and 31. As before the rotation functions are very unpredictable with several local minima, however the rotation function in Figure 31 is not an almost-straight line as it

7.3 PAIRWISE DISTANCE APPROACH

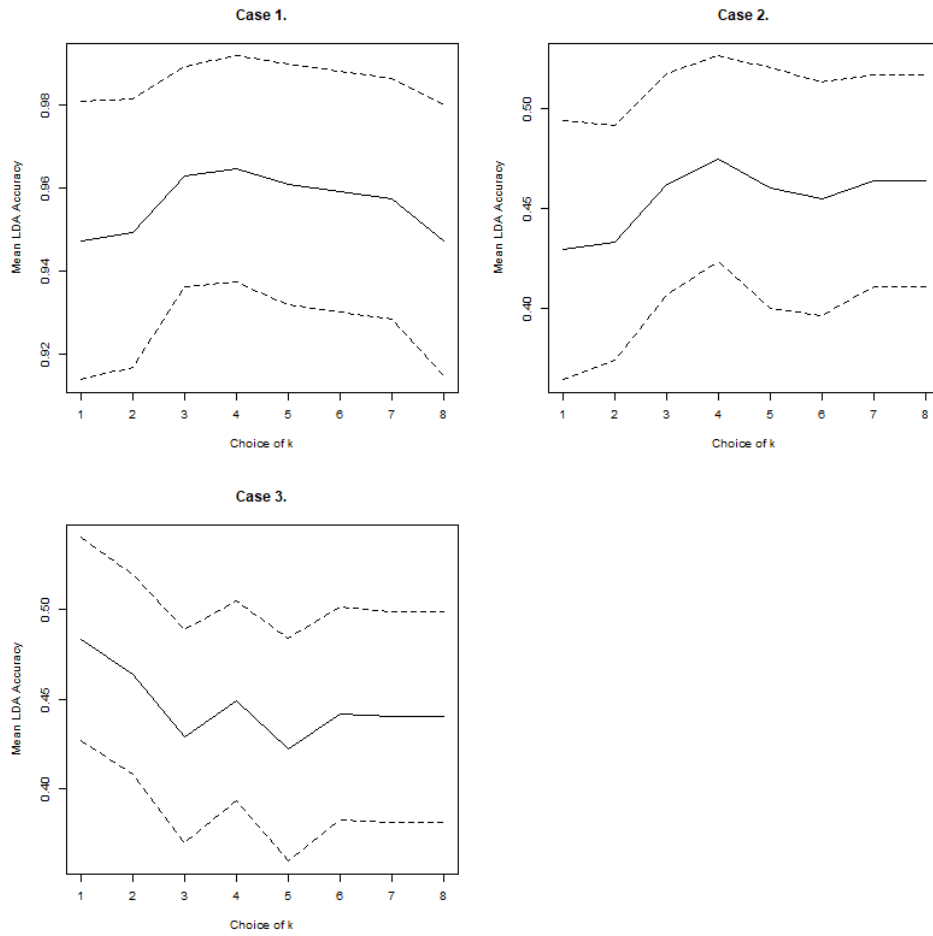


Figure 29: LDA correct classification probability for various choices of k . The dotted lines represent the upper and lower 95% confidence intervals.

was in Figure 28. This is because the Pseudo-Hausdorff is always affected by changes in the underlying points.

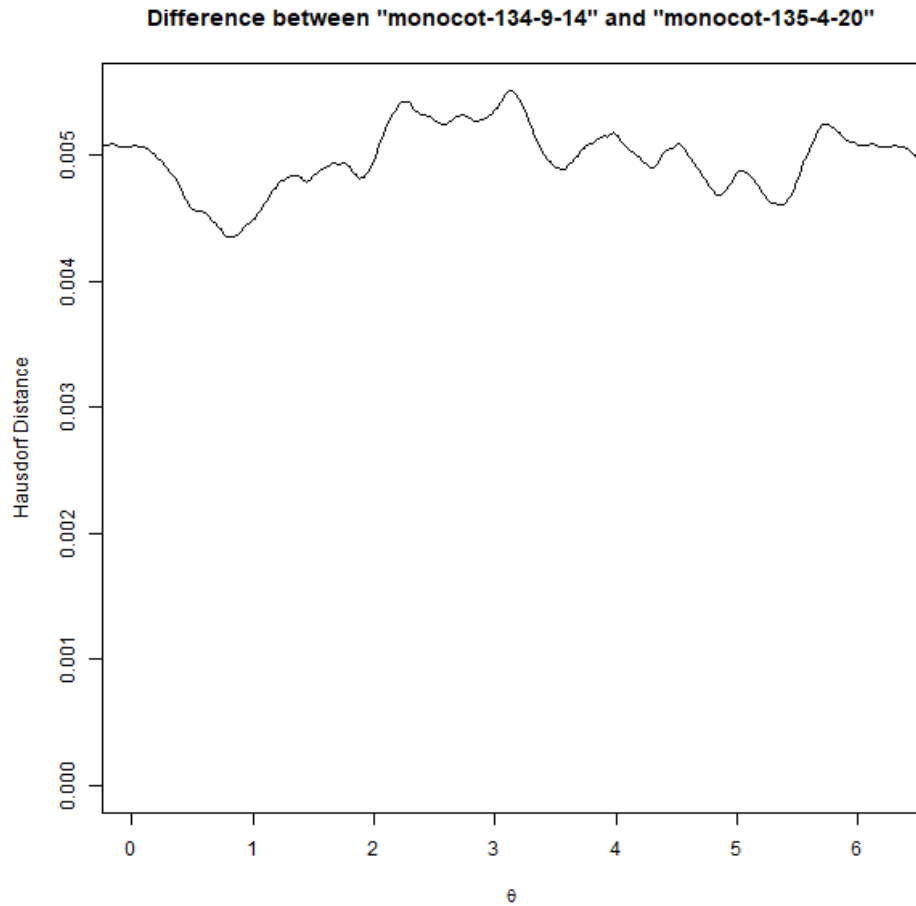


Figure 30: Effect of rotation on the pairwise Pseudo-Hausdorff distance between two different plant root systems. (see Figure 25)

Looking at the effect that different number of approximation points has on the Pseudo-Hausdorff metric, in Figure 32, we can see the rotation functions at various levels of approximation. Unlike the regular Hausdorff metric, this has a strong effect on the result, with each of the values appearing to be inversely proportional to the number of levels. This is likely because of a large number of 'small distances' between adjacent points affecting the average.

This is an unfortunate effect and definite downside to this approach. We will still include the Pseudo-Hausdorff distance in our analysis, since we can

7.3 PAIRWISE DISTANCE APPROACH

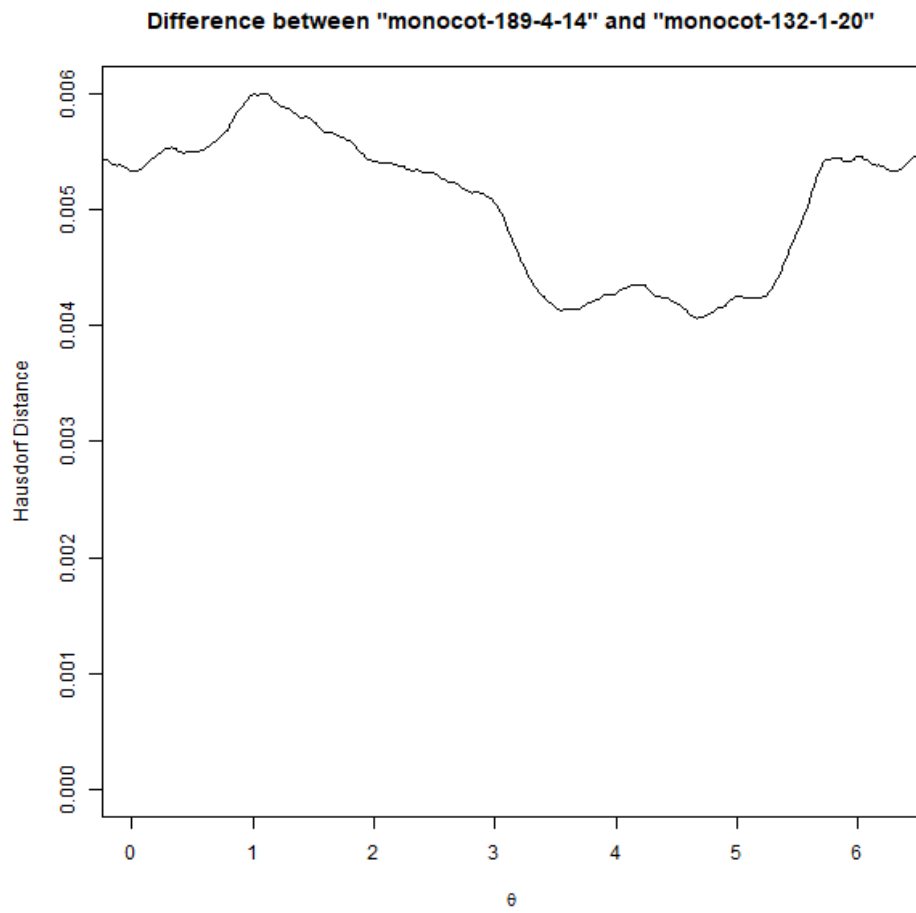


Figure 31: Effect of rotation on the pairwise Pseudo-Hausdorff distance between two different plant root systems. (see Figure 27)

partially control this effect by keeping the approximation levels consistent throughout all comparisons.

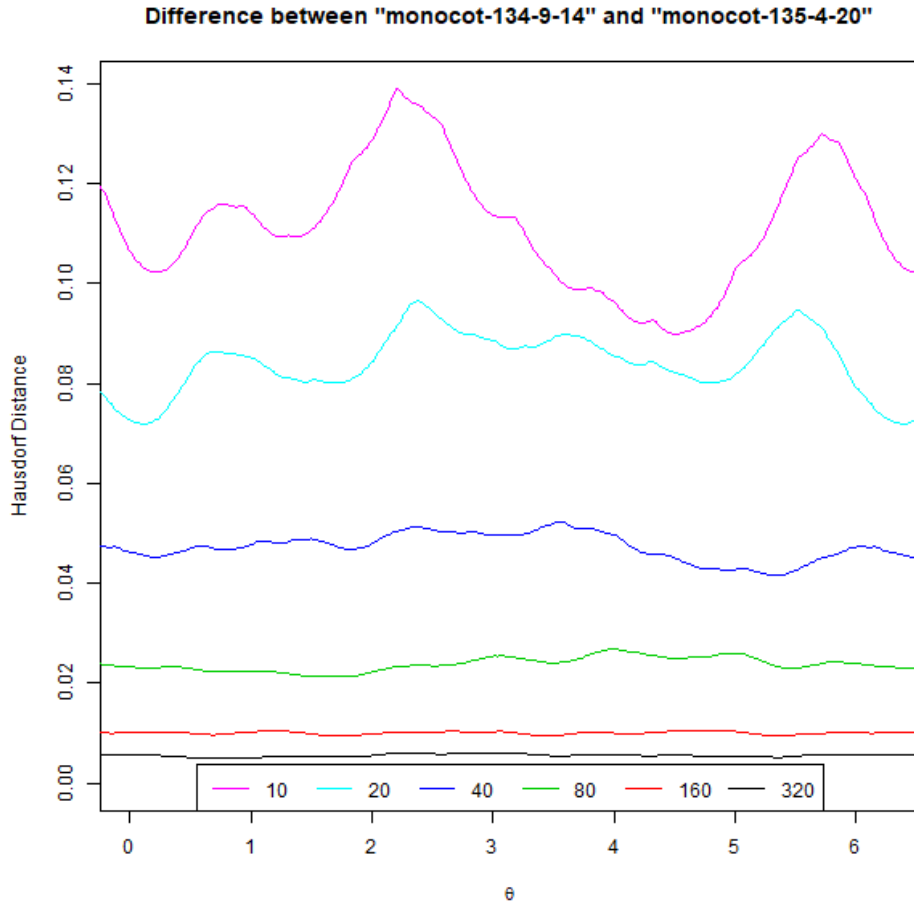


Figure 32: Effect of rotation and the number of approximation points on the pairwise Pseudo-Hausdorff distance between two different plant root systems.

For performance under LDA, we can see in Figure 33, that, like the regular Hausdorff method, we achieve good accuracy in case 1. There does not seem to be a peak, but the accuracy falls for higher k .

Cases 2 and 3 have the same expected 50% accuracy as discussed in the Hausdorff case.

7.3 PAIRWISE DISTANCE APPROACH

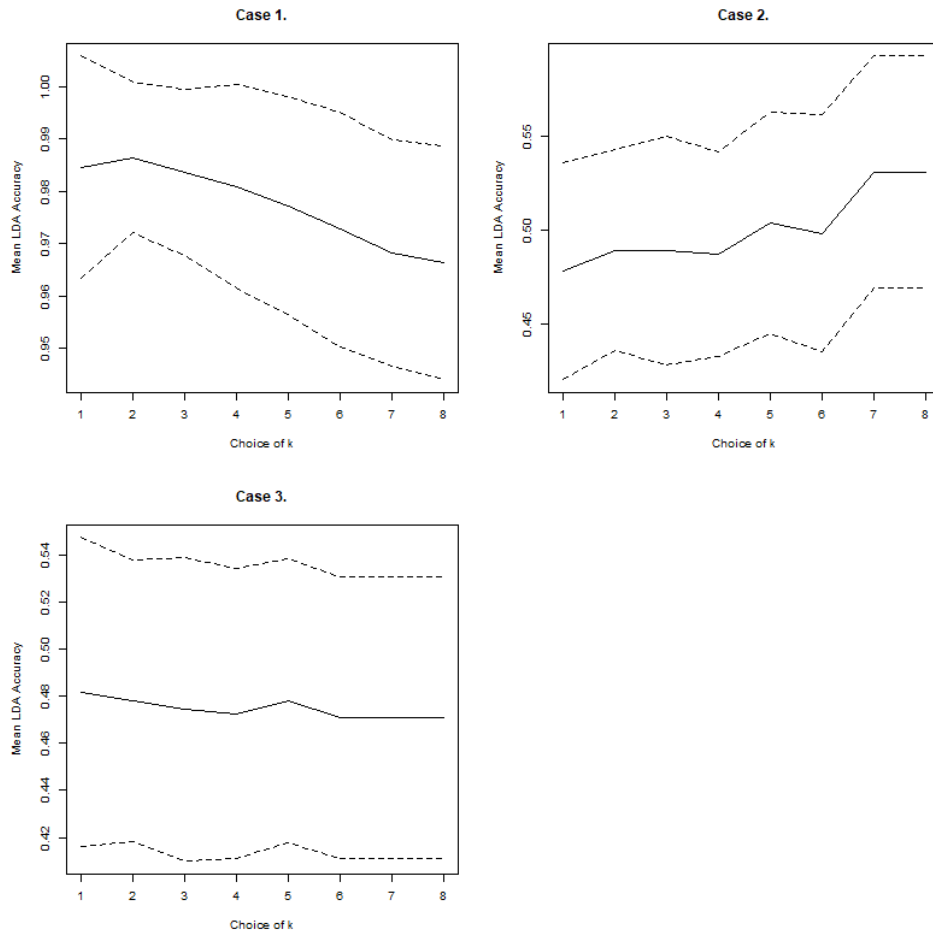


Figure 33: LDA correct classification probability for various choices of k . The dotted lines represent the upper and lower 95% confidence intervals.

7.4 TOPOLOGICAL DATA ANALYSIS

We do not need to account for rotation with the TDA approach so the only question is the level of approximation required. We will be using the Vietoris-Rips complex and bottleneck distance metric, since these have readily available computationally efficient methods (Fasy et al., 2017) and they work well on our plant data.

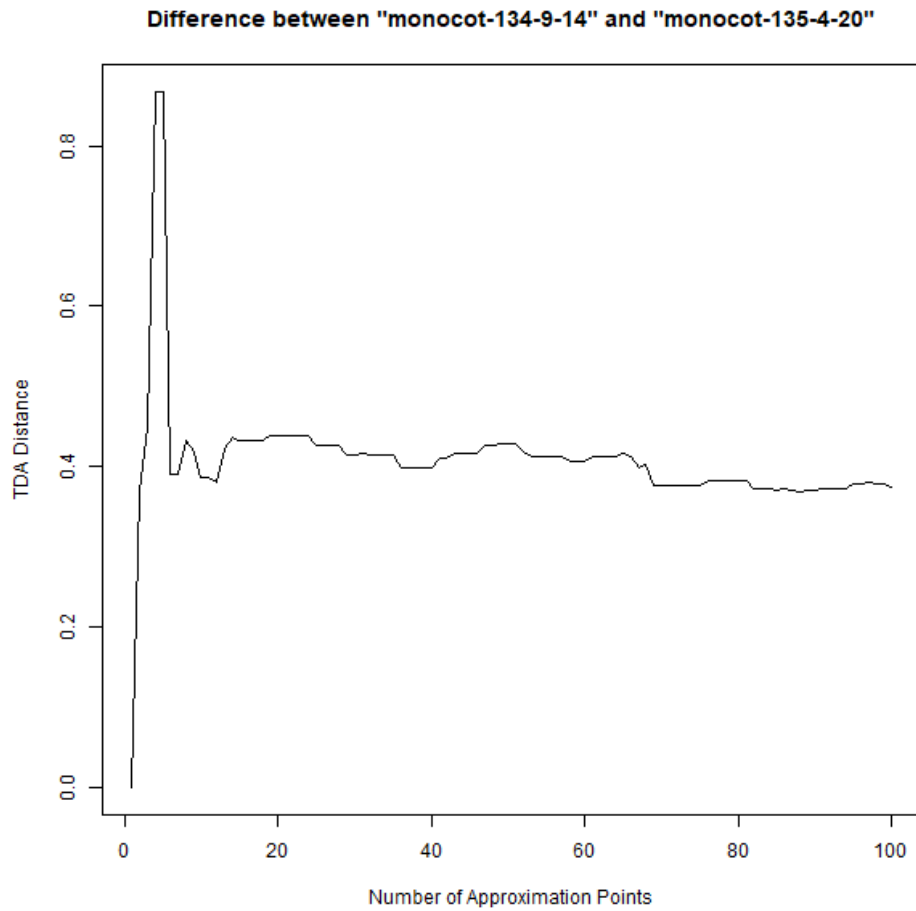


Figure 34: Effect of the number of approximation points on the value of a TDA comparison, using the Vietoris-Rips complex and bottleneck metrics.

From Figures 34 and 35, we can see that the bottleneck distance seems to approach to being around a single value, albeit with significant variation. Unfortunately it is not computationally reasonable to calculate the

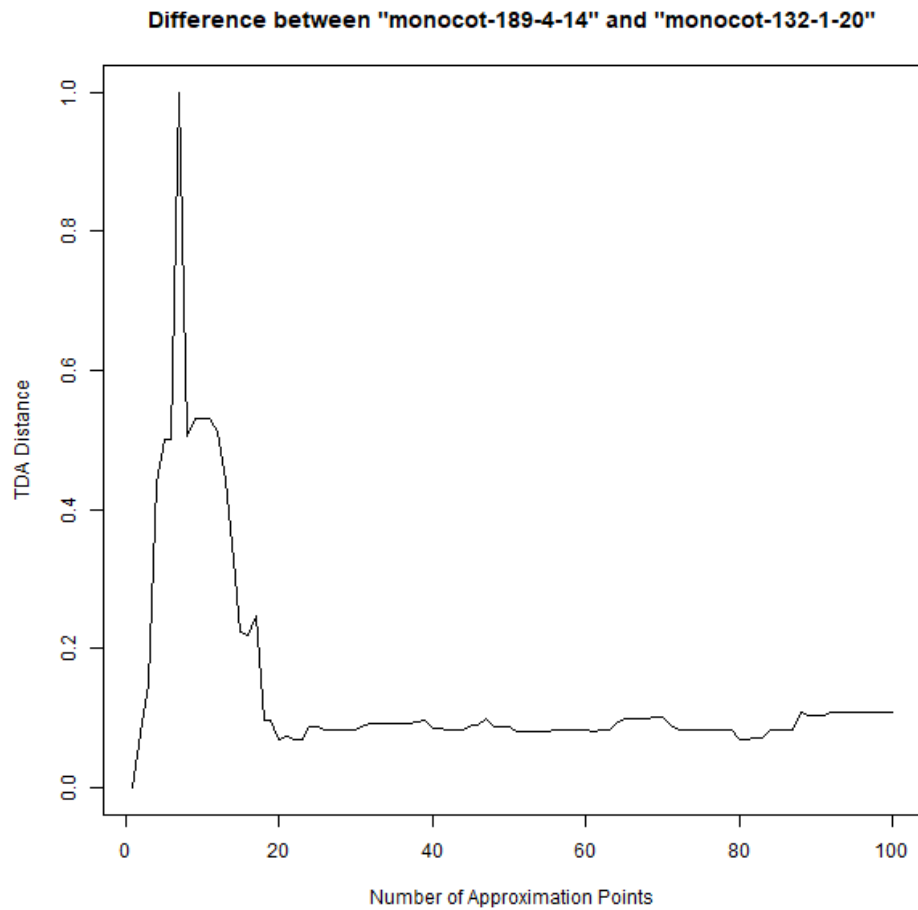


Figure 35: Effect of the number of approximation points on the value of a TDA comparison, using the Vietoris-Rips complex and bottleneck metrics.

7.5 CHOICE OF K FOR MDS

Rips complex for large numbers of points, so we will have to settle for a compromise.

For performance under LDA, we can see in Figure 36, that, like the regular Hausdorff method, we achieve good accuracy in case 1. There does not seem to be a peak, but the accuracy falls for higher k .

Cases 2 and 3 have the same expected 50% accuracy as discussed in the Hausdorff case.

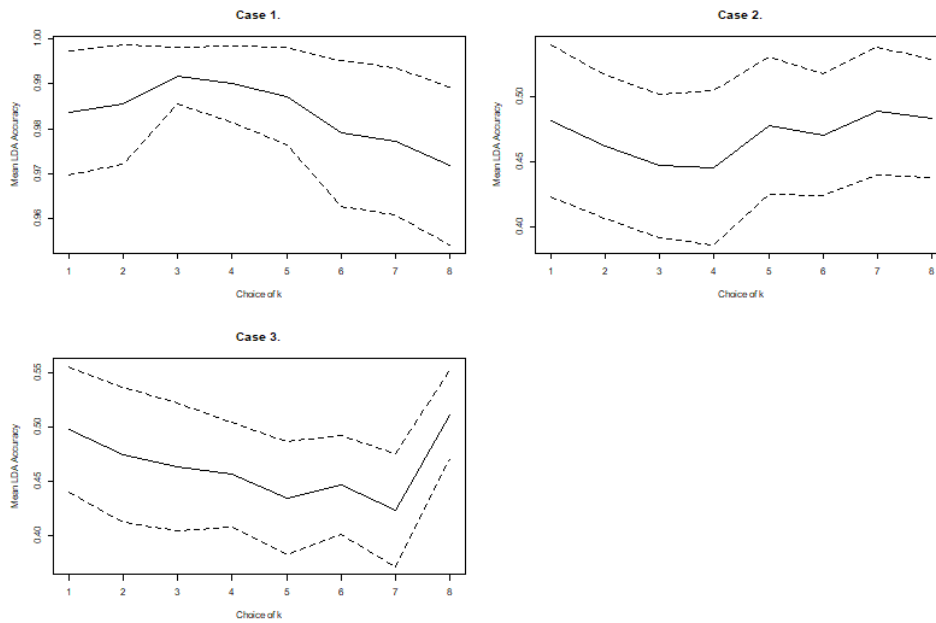


Figure 36: LDA correct classification probability for various choices of k . The dotted lines represent the upper and lower 95% confidence intervals.

7.5 CHOICE OF K FOR MDS

Looking at the effect of k on classification accuracy in Figures 29, 33 and 36 that a large value of k does not seem to make a large difference on the classification rate. Each of them seems to have a significant boost at either $k = 2$ or $k = 3$ and therefore we will choose $k = 3$ as our choice for comparisons in the analysis of practical data.

7.6 CONCLUSION

The Hausdorff and TDA approaches have both performed well on our simulated dataset, able to find significant differences where they occur and failing to find significant differences where there are none. Our simulated dataset has a large amount of shape variation though, and it is possible that real world datasets may have more subtle differences.

In the next chapter we will apply our results to a real world dataset.

PRACTICAL DATASET

8.1 DATASET

The data consists of 32 CT scans of the root systems of individual wheat plants. This provides us with a 'skeletonised' view of all the roots. Examples of these skeletonised plants can be seen in Figure 37.

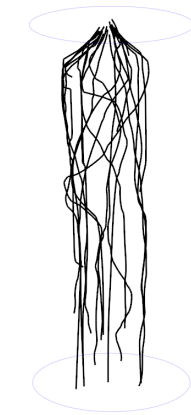
The wheat plants are grown from 4 different genotypes, denoted as '44', '72', 'Rialto' and 'Savannah'. The '44' and '72' plants are cross-breeds of the 'Rialto' and 'Savannah' genotypes. 8 plants were chosen from each genotype, 4 were grown in a 'drought' environment and 4 were grown in an 'well-watered' environment.

So, we have a balanced experiment, with a 'Genotype' factor (with 4 levels), an 'Irrigation' factor (with 2 levels), and 4 replicates in each group, for a total of 32 data points.

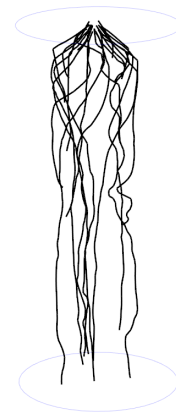
We are interested in seeing if there is interaction between the two factors. We have enough replicates that this is feasible without much loss of experimental power.

In addition to the CT scans of the wheat plants, we have several additional variables that measure various properties of the wheat plant at harvesting. These are,

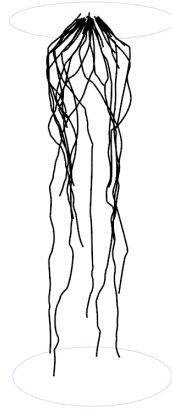
- Plant height (cm)
- Total above-ground biomass (g)



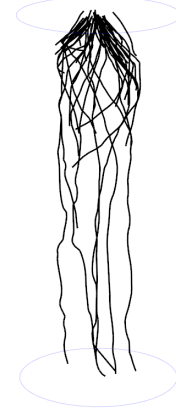
(a) 44/Drought



(b) Savannah/Drought



(c) 72/Drought



(d) 72/Well-Watered



(e) Rialto/Well-Watered



(f) Savannah/Well-Watered

Figure 37: A selection of wheat Roots from our dataset, rendered in 3D.

- Fertile shoots
- Infertile shoots
- WinRHIZO total length (m)
- Initial column weight
- Final column weight
- Water uptake (L/plant)

Most of these are self-explanatory, with the exception being WinRHIZO Total Length, which measures the total length of the root system using WinRHIZO, a different method to the CT scans. We wish to see if our method matches that of the WinRHIZO measurements.

8.2 RELATIONSHIPS WITH SUMMARY STATISTICS

There are a wide variety of summary statistics that can be measured from our root system and we wish to see if these are affected by the genotype and irrigation factors, and their interaction.

Since there is such a large number of statistics that are used to measure the plant roots, we have to be careful to ensure that we are not 'farming for significant p-values'. In order to prevent this, we will reduce our rejection threshold. It is difficult to choose a reduced p-value threshold since the choice of threshold is always essentially arbitrary. We have chosen a threshold of 0.001, and our reasoning is simply that we have roughly 50 traits that we can measure from a plant and so we divide the standard p-value threshold of 0.05 by 50 to get 0.001, which a simple round number. We could use a more sophisticated method but we believe that 0.001 is a satisfactory choice.

Let Y_i be a statistic of the i th plant, we can model its relationship with the Genotype and Irrigation factors as,

$$Y_i = \mu + \text{Genotype}_i\beta_1 + \text{Irrigation}_i\beta_2 + \text{Interaction}_i + \epsilon_i, \quad (59)$$

where Genotype_i , Irrigation_i and Interaction_i are rows of the contrast matrices that represent the genotype and irrigation factors, and their interaction factor respectively. μ , β_1 , β_2 and β_3 are the parameters to be estimated and $\epsilon_i \sim N(0, \sigma^2)$, is the error term.

We can reduce the number terms using AIC reduction methods, if necessary. We use the stepwise reduction algorithm from the 'mass' r package (Ripley et al., 2013) to choose the combination of parameters that give the lowest AIC.

8.2.1 *Root Count*

As an example of the results, we will use the statistic of Root Count.

Root count is a simple measure of the plants root system, that counts the number of offshoots that are made, regardless of length. If the plant has more roots, it will have an improved ability to take up water but will require more water due to the increased biomass. So it is very possible that the plant may respond when it is in drought conditions, either growing more roots to try to find more water, or growing less roots to try to conserve what little water it finds.

As we can see in Figure 38, there is a substantial increase in root count when the environment is well watered. There does not seem to be a relationship with the genotype. We can confirm this analytically using ANOVA, as shown in Table 2.

8.2.2 *Multi-Analysis of Various Statistics*

We perform our analysis on multiple properties. The results of this can be see in Table 3. The table contains the results of partial tests for the significance of each factor, the results of an overall test for significance, and

8.2 RELATIONSHIPS WITH SUMMARY STATISTICS

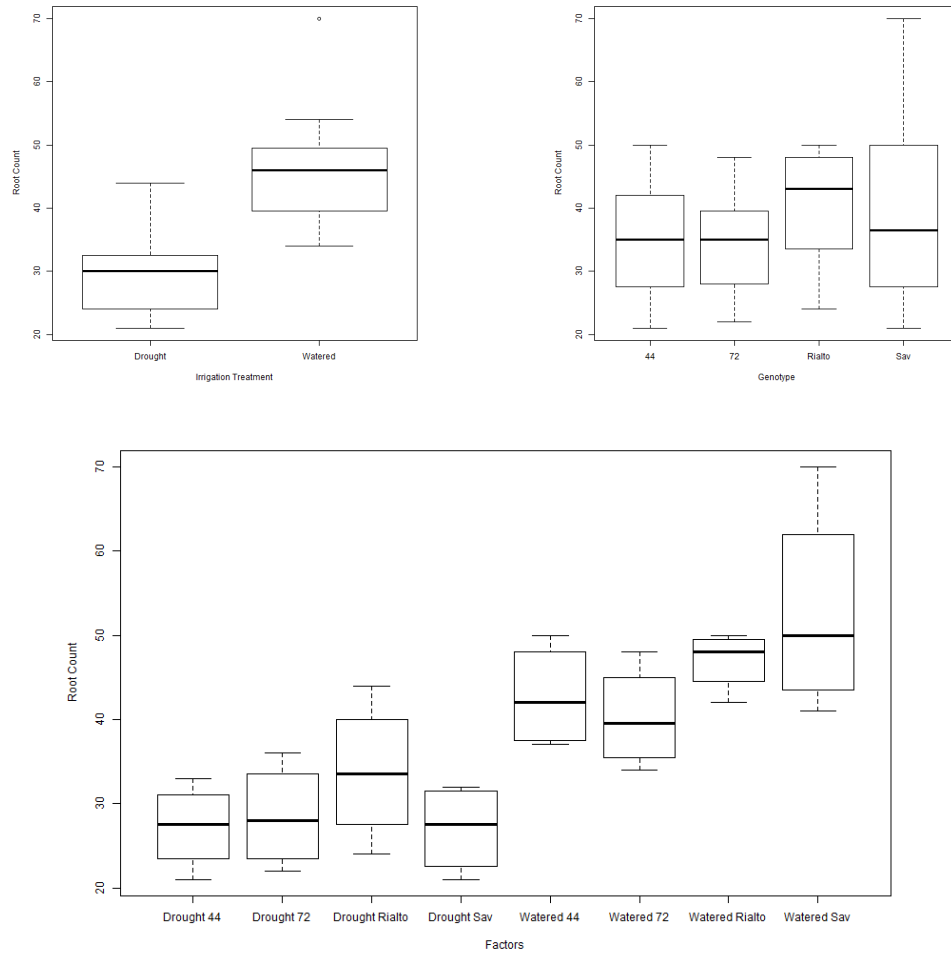


Figure 38: Plots of root count with genotype and irrigation treatment.

Variable	Df	Sum Sq	Mean Sq	F val	P-Value
Genotype	3	239	80	1.537	0.23
Irrigation	1	2195	2195	42.313	1×10^{-06} (***)
Interaction	3	239	80	1.538	0.23
Residuals	24	1245	52		
Total	31	3918	2406		

Note: (·) $p < 0.1$; (*) $p < 0.05$; (**) $p < 0.01$; (***) $p < 0.001$

Table 2: Anova test performed on the 'Root Count' measurement. There is a strong relationship with the Irrigation factor.

the result of an overall test after irrelevant factors have been removed by AIC reduction.

We have also included several statistics related to the ‘unravelled’ version of the root system (see Section 3.6).

We can see that quite a few of them have a relationship with our factors, mostly with the Irrigation factor.

	Genotype	Irrigation	Interaction	P.Value	Reduced
Root Length Total	0.16	0.00034*	0.15	0.0041	0.0041
Root Length Average	0.42	0.00038*	0.86	0.022	0.00016*
Width	0.92	5e-04*	0.28	0.022	0.00027*
Width (Unravelled)	0.0064*	0.55	0.027*	0.0061	0.0061
Height	0.15	0.0075*	0.88	0.077	0.012
Width/Height Ratio	0.95	0.0024*	0.33	0.07	0.0014
Width/Height Ratio (Unravelled)	0.0084*	0.39	0.025*	0.0065	0.0065
Horizontal Angle	0.2	0.31	0.89	0.49	-
Vertical Angle	0.001*	0.026*	0.15	0.0016	0.0016
Avg Tortuosity	0.0072*	0.98	0.37	0.035	0.0054
Avg Tortuosity (Unravelled)	0.00034*	0.38	0.12	0.0014	0.0014
Root Count	0.23	1e-06*	0.23	9.4e-05*	8.4e-07*
Convex Hull Surface Area	0.5	0.00095*	0.18	0.017	0.017
Convex Hull Surface Area (Unravelled)	0.0084*	0.73	0.3	0.034	0.0072
Centre of Mass (Horizontal)	0.68	0.025*	0.42	0.24	0.02
Centre of Mass (vertical)	0.27	0.0028*	0.84	0.061	0.002
Plant Height	1.3e-05*	1.2e-11*	0.016*	3.5e-10*	3.5e-10*
Total above-ground biomass	1e-07*	<2.2e-16*	0.84	1.1e-14*	<2.2e-16*
Fertile shoots	0.31	3.4e-12*	0.029*	1.3e-09*	1.3e-09*
Infertile shoots	0.16	7.9e-11*	0.0036*	1.1e-08*	1.1e-08*
WinRHIZO total length	0.4	<2.2e-16*	3.9e-07*	<2.2e-16*	<2.2e-16*
Initial column weight	0.71	0.5	0.08	0.27	0.27
Final column weight	0.71	0.095	0.95	0.69	0.07
Water uptake	4.2e-06*	3.8e-16*	0.0071*	6.8e-14*	6.8e-14*
RLD 0-20cm	6.3e-09*	<2.2e-16*	4.9e-06*	<2.2e-16*	<2.2e-16*
RLD 20-40cm	0.00085*	<2.2e-16*	0.0014*	<2.2e-16*	<2.2e-16*
RLD 40-60cm	5.3e-11*	1	3.8e-05*	2.7e-10*	2.7e-10*
RLD 60-80cm	0.00015*	1.2e-05*	5.5e-07*	4.1e-08*	4.1e-08*

Table 3: Results of Anova tests performed on a large number of properties. The Genotype, Irrigation and Interaction columns contain the p-values of partial tests for relationships with those individual terms. The p-value column has the overall test for any relationship. The ‘reduced’ column has the overall test when unnecessary terms have been removed with AIC reduction methods.

8.2 RELATIONSHIPS WITH SUMMARY STATISTICS

	Mean	72	Rl	Sv	W	W:72	W:Rl	W:Sv
Root Length Total	68*	-6.8	3	-9	11	-2.6	0.059	18
Root Length Average	2.6*	-0.37	-0.41	-0.37	-0.73*	0.27	0.28	0.19
Width	0.61*	0.017	-0.019	-0.0099	0.037	-0.013	0.058	0.048
Width (Unravelled)	2.1*	-0.36	-0.42*	-0.12	-0.51*	0.41	0.56*	0.85*
Height	4.4*	0.0046	0.036	0.062	0.071	-0.026	0.0069	-0.031
Width/Height Ratio	0.14*	0.0038	-0.0054	-0.004	0.0062	-0.0021	0.012	0.011
Width/Height Ratio (Unravelled)	0.47*	-0.081	-0.098*	-0.033	-0.12*	0.094	0.12*	0.19*
Horizontal Angle	-0.57	0.37	0.43	2.1	0.17	-0.87	-1.2	-1.2
Vertical Angle	0.97*	-0.068*	-0.044	-0.11*	-0.093*	0.069	0.1*	0.053
Avg Tortuosity	0.92*	0.0054	-0.0043	-0.025	0.0073	-0.024	0.0068	-0.012
Avg Tortuosity (Unravelled)	0.95*	0.006	-0.0043	-0.021*	0.0041	-0.015	0.0083	0.0015
Root Count	27*	1.2	6.5	-0.25	16*	-3.8	-2.2	10
Convex Hull Surface Area	13*	-0.24	-0.46	-0.021	0.23	0.16	1*	0.46
Convex Hull Surface Area (Unravelled)	34*	-2.9	-5	1.2	-3.3	2	5	8.5
Centre of Mass (Horizontal)	0.1*	0.0025	0.008	-0.0062	0.04	0.019	-0.043	0.045
Centre of Mass (vertical)	1.7*	-0.078	-0.16	-0.0012	-0.14	-0.047	-0.0055	-0.13
Plant Height	55*	0.65	-4	1.6	7.4*	8.2*	3.2	9.1*
Total above-ground biomass	24*	1.9*	-0.65	3.2*	7.7*	0.50	-0.087	0.76
Fertile shoots	4.8*	-0.5	-0.25	-0.75	2.8*	0.75	1.2	2.8*
Infertile shoots	2.5*	0	0.25	1.8*	-2.2*	0.75	-0.25	-2*
WinRHIZO total length	26*	-1.4*	0.26	-1.6*	4.9*	2*	-0.75	2.9*
Initial column weight	46*	0.54	0.63	0.82	1.6*	-1.4	-1.3	-2.6*
Final column weight	46*	0.032	0.43	-0.0077	-0.28	0.015	-0.3	-0.15
Water uptake	17*	-1.9*	-1.6*	-2.2*	6.7*	0.69	-2.6*	-0.50
RLD 0-20cm	1.1*	0.092*	0.032	-0.052	0.66*	0.095	-0.28*	-0.11
RLD 20-40cm	0.91*	-0.052	-0.11*	-0.11*	0.46*	-0.065	0.042	0.13*
RLD 40-60cm	0.80*	-0.1*	0.07*	0.12*	-0.0075	0.092*	0.0025	-0.065*
RLD 60-80cm	1.3*	-0.16*	0.052	-0.22*	-0.33*	0.18*	0.12	0.5*

Table 4: Coefficients of fitted linear models on properties in Table 3. The asterisk * next to a coefficient denotes that the coefficient is significant at the 0.05 threshold.

W=Well-watered, Rl=Rialto, Sv=Savannah

8.3 COMPARISON OF FACTOR MEANS

	Mean	72	Rl	Sv	W	W:72	W:Rl	W:Sv
Root Length Average	2.3*	-	-	-	-0.55*	-	-	-
Width	0.6*	-	-	-	0.06*	-	-	-
Root Count	29*	-	-	-	17*	-	-	-
Plant Height	55*	0.65	-4	1.6	7.4*	8.2*	3.2	9.1*
Total above-ground biomass	24*	2.1*	-0.7	3.6*	8*	-	-	-
Fertile shoots	4.8*	-0.5	-0.25	-0.75	2.8*	0.75	1.2	2.8*
Infertile shoots	2.5*	0	0.25	1.8*	-2.2*	0.75	-0.25	-2*
WinRHIZO total length	26*	-1.4*	0.26	-1.6*	4.9*	2*	-0.75	2.9*
Water uptake	17*	-1.9*	-1.6*	-2.2*	6.7*	0.69	-2.6*	-0.50
RLD 0-20cm	1.1*	0.092*	0.032	-0.052	0.66*	0.095	-0.28*	-0.11
RLD 20-40cm	0.91*	-0.052	-0.11*	-0.11*	0.46*	-0.065	0.042	0.13*
RLD 40-60cm	0.80*	-0.1*	0.07*	0.12*	-0.0075	0.092*	0.0025	-0.065*
RLD 60-80cm	1.3*	-0.16*	0.052	-0.22*	-0.33*	0.18*	0.12	0.5*

Table 5: Coefficients of fitted linear models on properties in Table 3, that were reduced using AIC reduction and are significant at the 0.001 threshold. The asterisk * next to a coefficient denotes that the coefficient is significant at the 0.05 threshold. A blank entry means that term was removed during AIC reduction.
W=Well-watered, Rl=Rialto, Sv=Savannah

8.3 COMPARISON OF FACTOR MEANS

The linear models are very useful for inferences from our data but are not ideal for interpreting differences between the effects of different factors. In order to better interpret how our data varies in practice, we can use Tukey's HSD (honest significance difference) test. This test when performed on data with multiple groups will calculate the significance of differences between all possible pairs of means.

This allows us to tell if, for example, the mean Rialto plant height is significantly different from the mean Savannah plant height, or if the mean well-watered root count is significantly different from the mean drought root count.

In Figures 39 - 51, we have taken the significant properties we found in Section 8.2.2 and plotted boxplots of each property in terms of genotype individually, irrigation individually, and the combination of the two. Each factor's boxplots are plotted in ascending order of their mean value (so the

order is different for each plot). Blue rectangles are then added that join together any levels that are not significantly different according to a Tukey HSD test.

For example, in Figure 42 on page 127, we see three boxplots of Plant Height.

The first plot in the upper left is the plant height versus genotype, and all four boxplots are within the same blue rectangle. This means that none of the means of the genotypes are considered significantly different from each other (note that this ignores the irrigation and interaction factors).

The second plot in the upper right is plant height versus the irrigation factors, and both bars are in separate blue rectangles. This means that the drought and well-watered means are considered significantly different from each other.

The third plot in the bottom is plant height versus the combined genotype and irrigation factor. This is the most complex plot, since there are a lot more levels. Three of the blue rectangles overlap, which occurs when two groups means are significantly different to each other but are not significantly different to a third group mean. This is possible since 'not significantly different' is not the same thing as saying that the means are equal.

We can see that '72/Watered' and 'Savannah/Watered' are separated from the rest of the group with their own blue rectangle. This could mean that these two genotypes grow significantly taller, but only in watered conditions.

For the other groups, it's a little more ambiguous but we can see some differences. For example, 'Rialto/Drought' and 'Rialto/Watered' are significantly different, as well as '44/Drought' and '44/Watered', so we can tell that being watered does have a positive effect on all the genotype group means.

We can also see that the four 'Genotype/Drought' groups are all not significantly different to each other, which is very different from the four

8.3 COMPARISON OF FACTOR MEANS

'Genotype/Watered' groups, where '72/Watered' and 'Savannah/Watered' are significantly different from 'Rialto/Watered' and '44/Watered'.

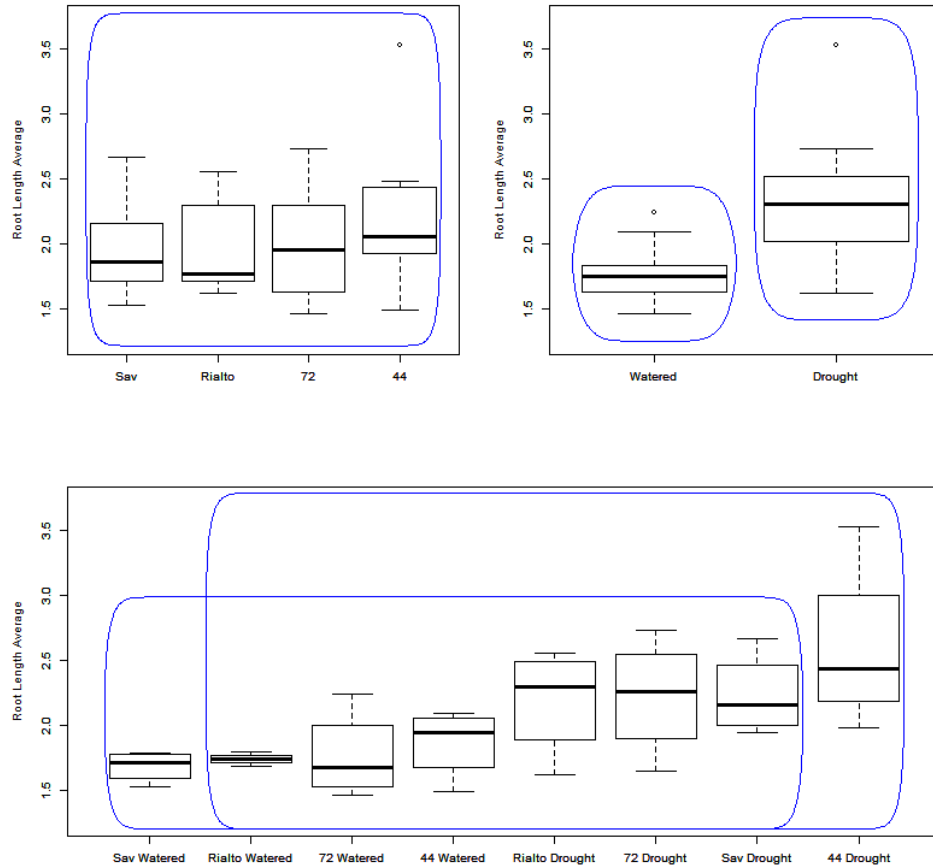


Figure 39: Boxplot of "Root Length Average" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

In Figure 39, "Root Length Average" is clearly dependent on the irrigation factor, with being 'Well-watered' reducing the average. We can see that '44 Drought' is considered significantly different from 'Sav Watered', however AIC reduction removed the interaction term as unnecessary, this is likely just the Irrigation factor being expressed.

In Figure 40, 'Width' has a relationship with the Irrigation factor but nothing else is significant. This is an interesting case because there isn't enough evidence to see a strong relationship with any of the Genotype/Irrigation

8.3 COMPARISON OF FACTOR MEANS

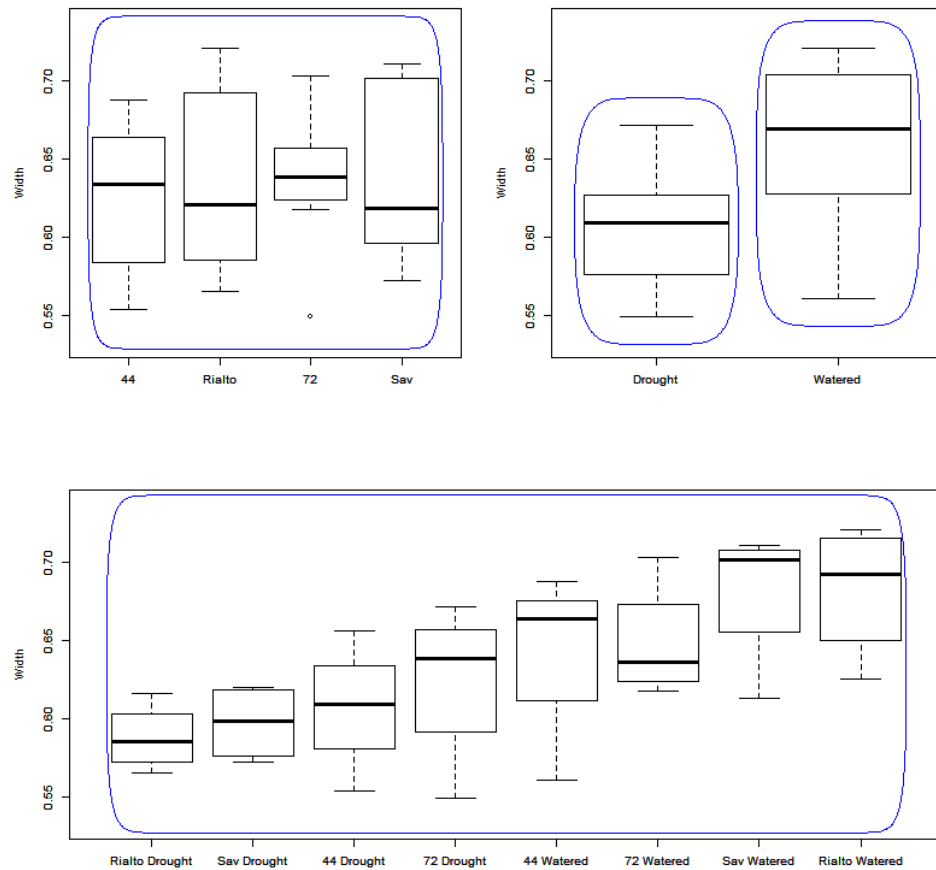


Figure 40: Boxplot of "Width" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

pairs individually, and it's only when combined that we see the relationship with Irrigation. This occurs because when the Irrigation data is pooled into their two sets there is sufficient evidence to find a significant difference, but individually none have enough evidence..

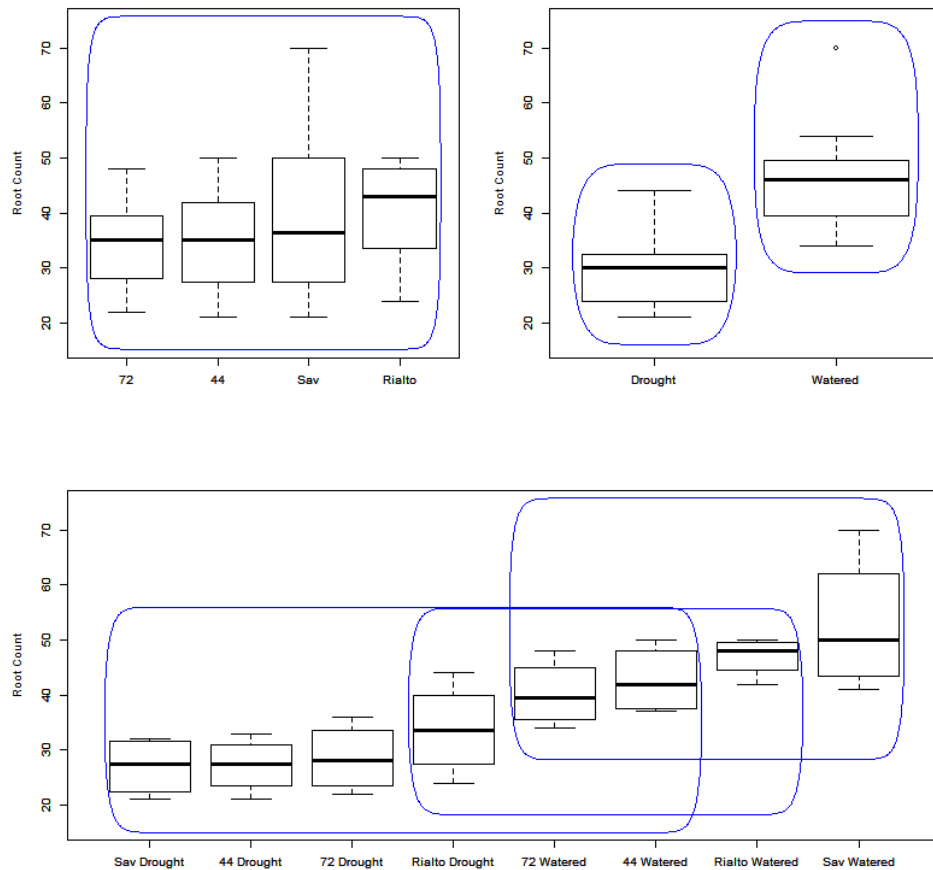


Figure 41: Boxplot of "Root Count" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

In Figure 41, we see "Root Count" is affected by the Irrigation factor. Rialto/Drought is considered significantly different from the other Genotype/Drought pairs, but since the Interaction factor was removed by AIC reduction, we can conclude this is not evidence of a deeper relationship.

In Figure 43, there is a relationship with Irrigation and Genotype, however the Genotype effects are masked by the large Irrigation effect when

8.3 COMPARISON OF FACTOR MEANS

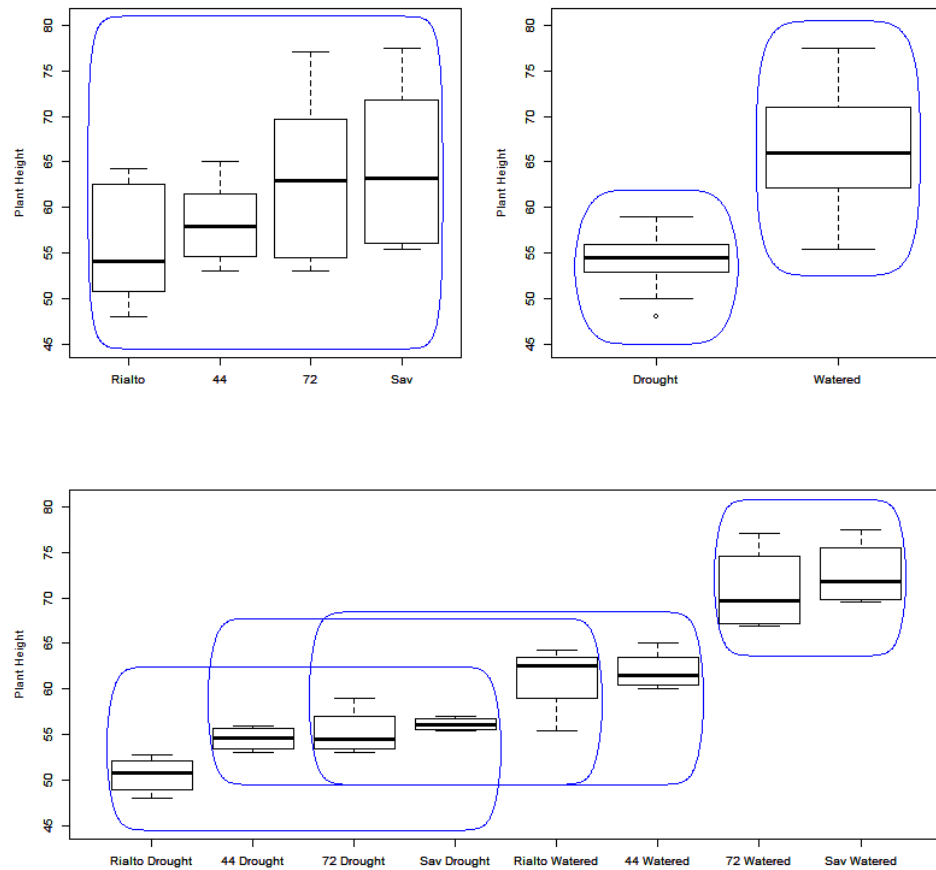


Figure 42: Boxplot of "Plant Height" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

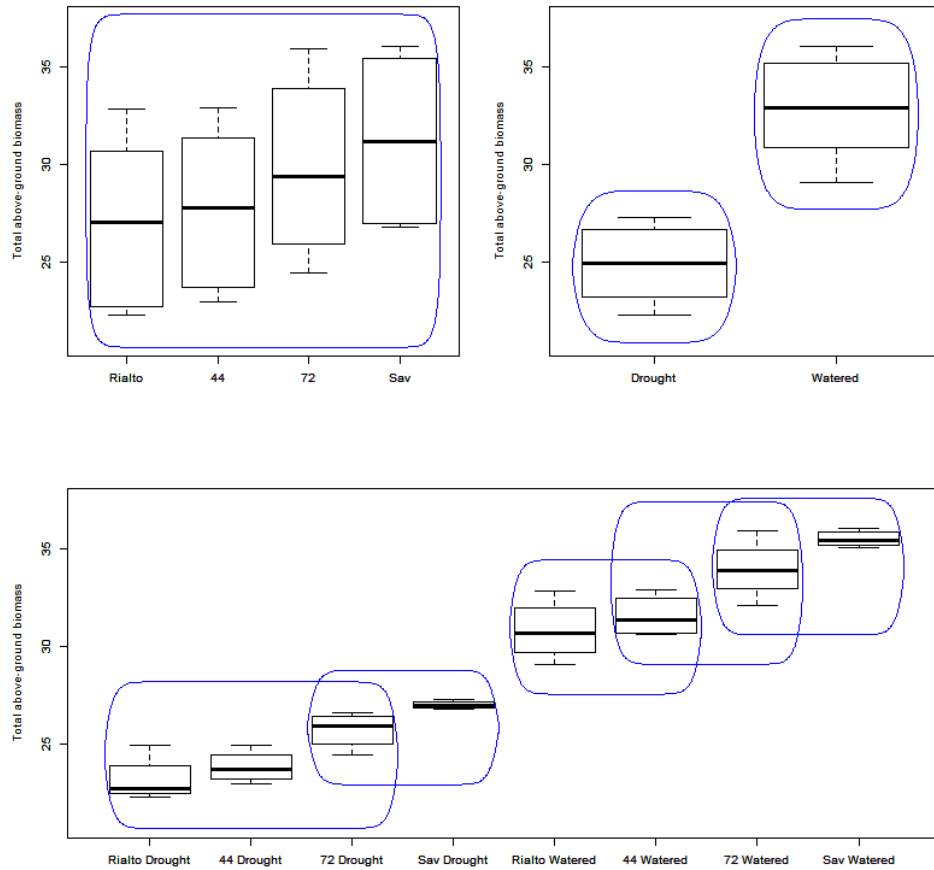


Figure 43: Boxplot of "Total above-ground biomass" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

looking at Genotype alone. We can see that there is a similar difference between "Rialto" and "Savannah" factors in both drought and watered conditions.

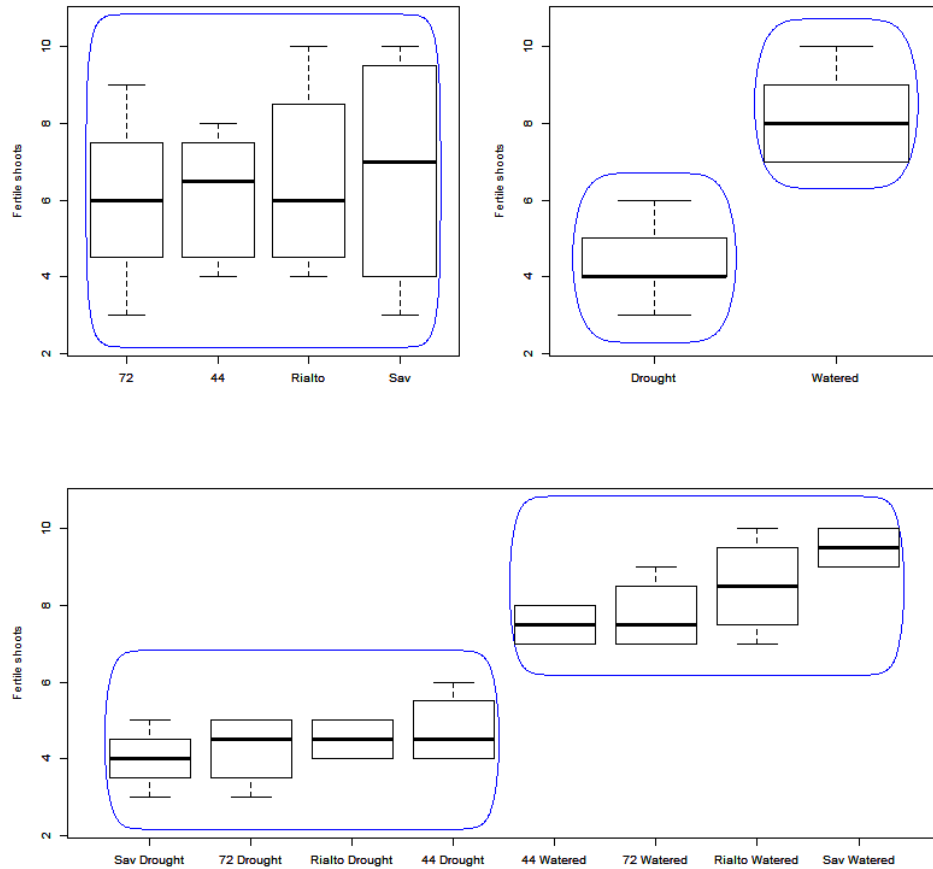


Figure 44: Boxplot of "Fertile shoots" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

In Figure 44, there is a very clear evidence of a relationship with Irrigation and no evidence for Genotype or their interaction.

For the remaining Figures 45-51 the various measures exhibit a strong relationship with Genotype, Irrigation and their Interaction. Each of the Genotypes responds slightly differently to drought and watered conditions.

8.3 COMPARISON OF FACTOR MEANS

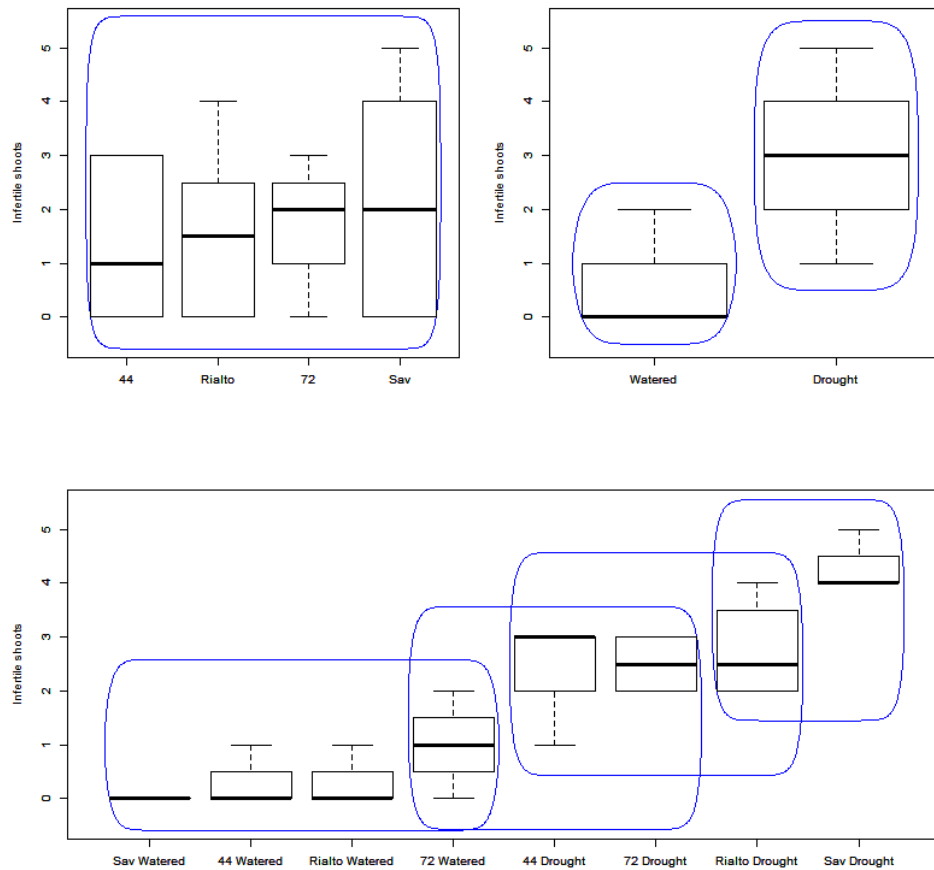


Figure 45: Boxplot of "Infertile shoots" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

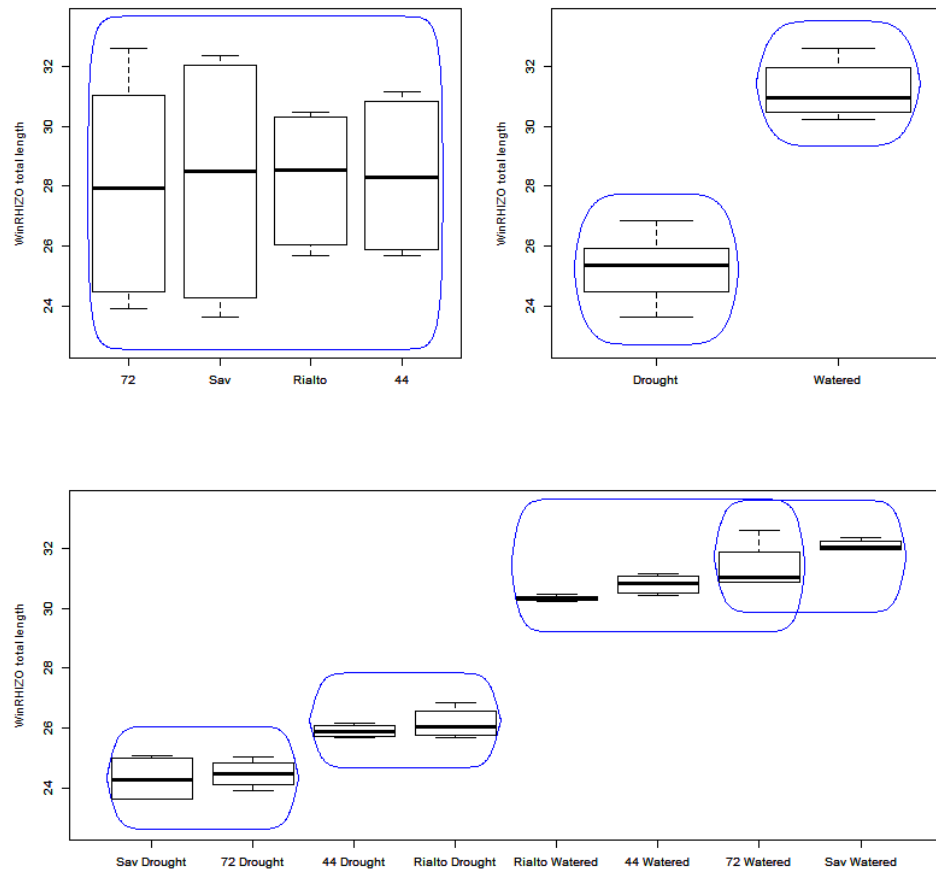


Figure 46: Boxplot of "WinRHIZO total length" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

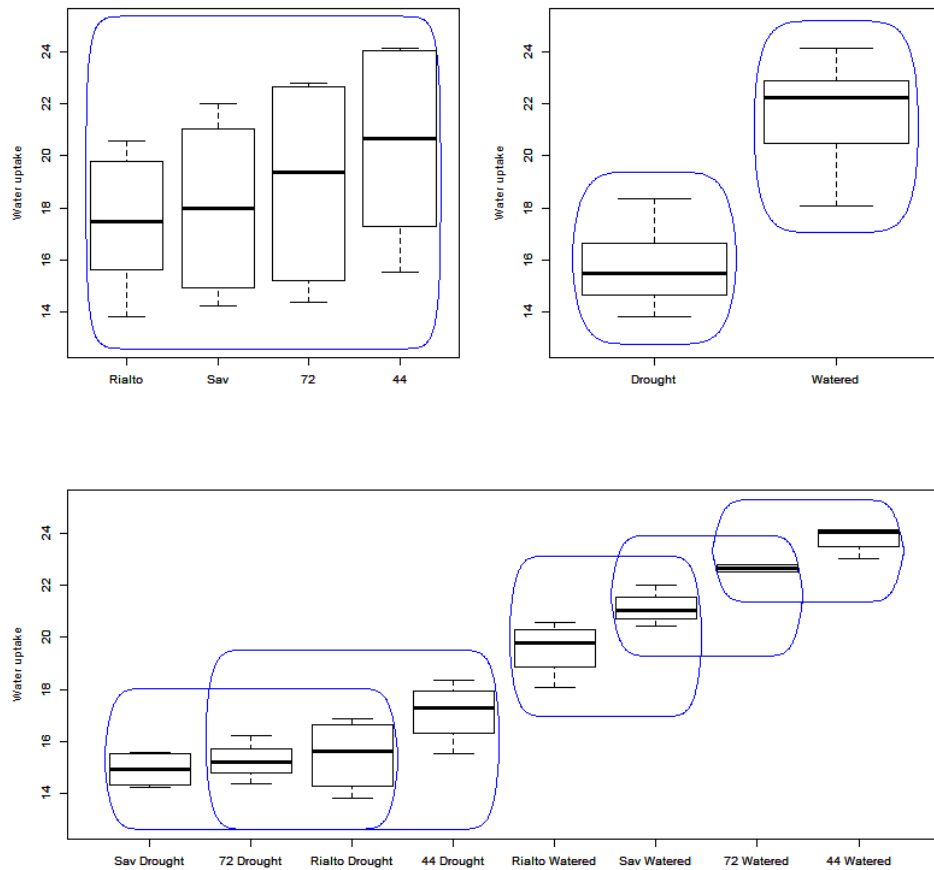


Figure 47: Boxplot of "Water uptake" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

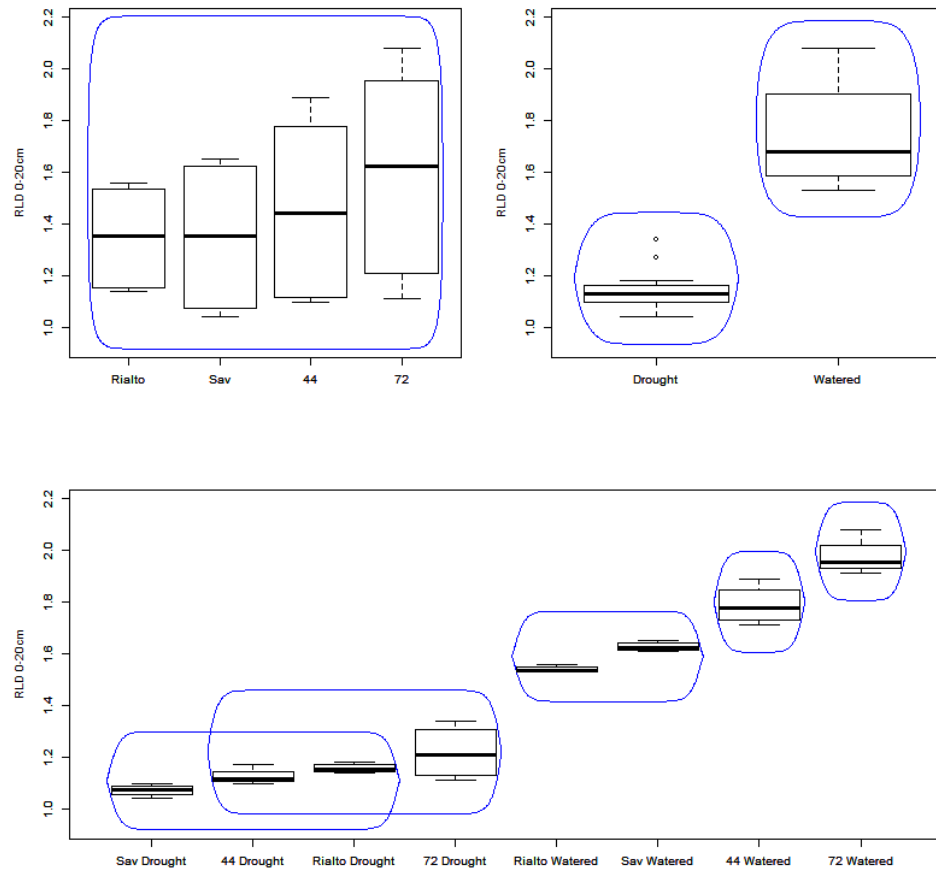


Figure 48: Boxplot of "RLD 0-20cm" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

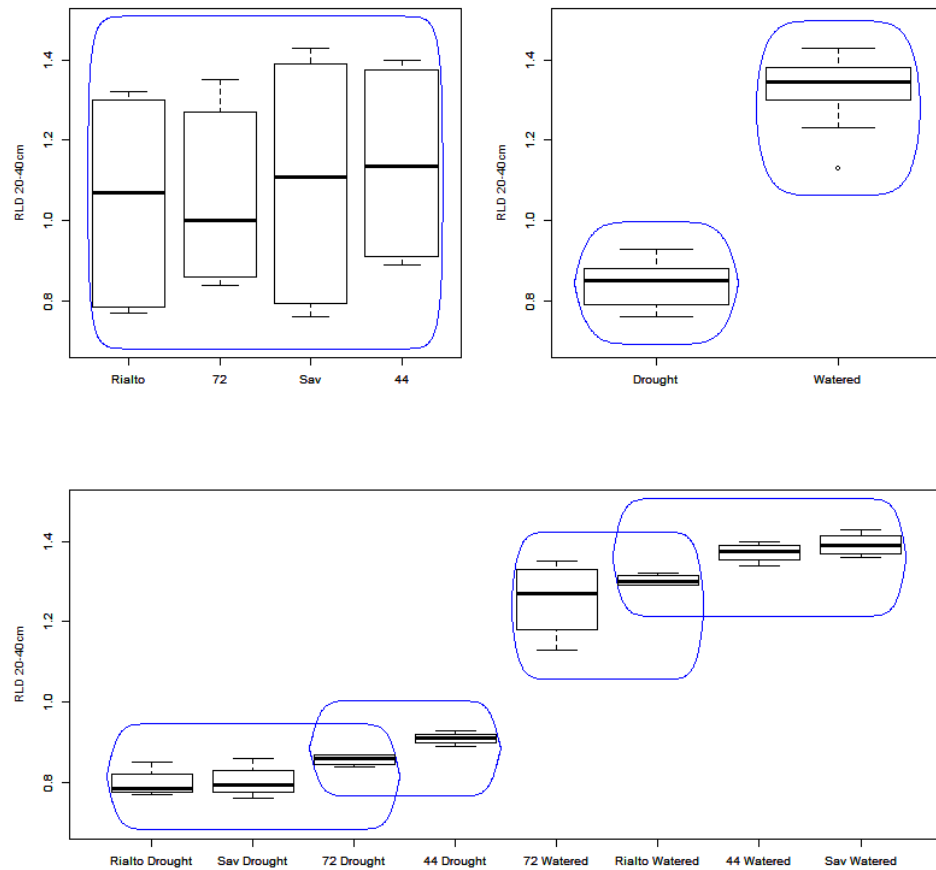


Figure 49: Boxplot of "RLD 20-40cm" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

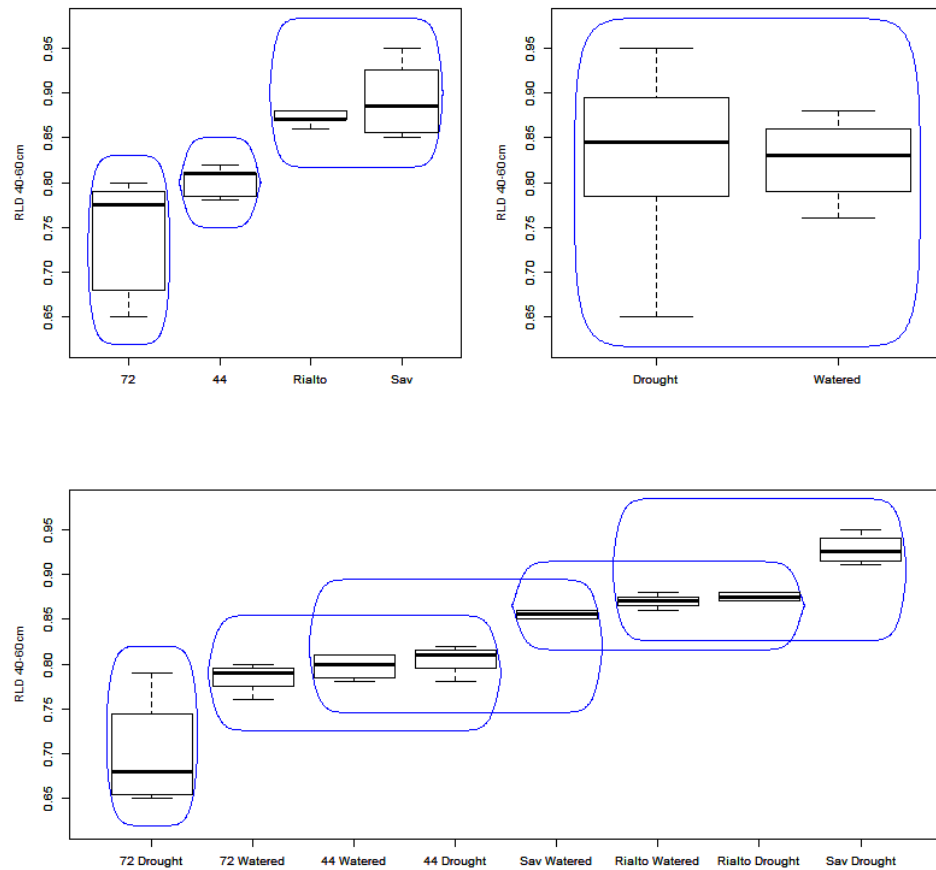


Figure 50: Boxplot of "RLD 40-60cm" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.3 COMPARISON OF FACTOR MEANS

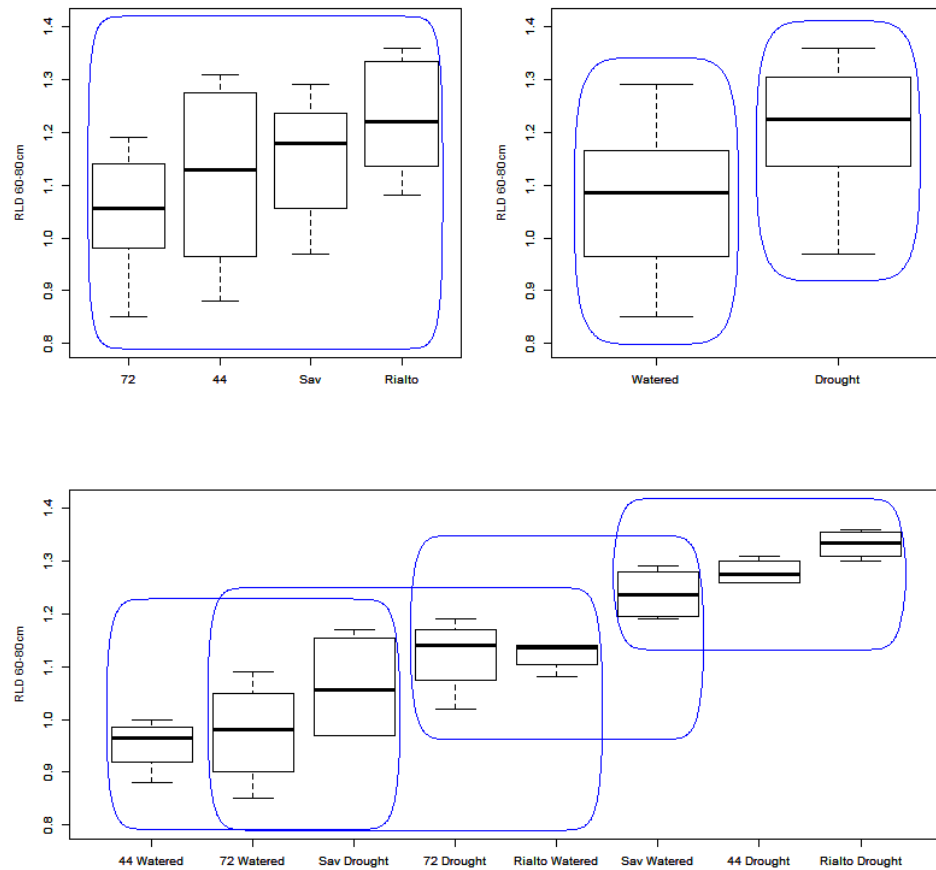


Figure 51: Boxplot of "RLD 60-80cm" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

Our dataset has a relatively small number of plants, but each individual plant has a large number of root strands, and naturally we might ask if we could analyse properties of these strands to increase the number of points in our dataset. The problem with this approach is that individual root strands are not independent of each other. If a plant had its growth stunted, then all the roots of the plant would be smaller, and if that reason was not related to its genotype or irrigation conditions then we have a hidden factor that will bias our results.

So each individual plant may have a bias affecting its root's properties, that is not related to genotype or irrigation. We can model this as a 'random effect', to create a mixed effects model. Let $Y_{i,j}$ be a property of the j th root strand from the i th plant, then we model it as,

$$Y_{i,j} = \mu + \beta_1 \text{Genotype}_i + \beta_2 \text{Irrigation}_i + \text{Interaction}_i + Z_i + \epsilon_{i,j}, \quad (60)$$

where, $Z_i \sim N(0, \gamma^2)$, is the i th plants random effect. The values of Z_i are random and would be different every time we repeated the experiment with new plants. Each Z_i is distributed as a normal variable with mean 0 and unknown variance γ^2 .

Using this model we can analyse individual root strands, which effectively increases the number of data points from 32 to 1197, with the trade off of requiring a more complex model which has unobservable effects (the individual plant effects).

8.4.1 *Root Strand Length*

We wish to see if each the length of each root strand is related to our environment factors. We model our mixed effects model, and the results are found in Table 6 and Figure 52. We can see that 'Watered' has a negative

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

effect on root length (-0.678), which is reasonably consistent with the effect that 'Watered' has on 'Root Length Average' (-0.73). Note: We wouldn't expect them to line up perfectly since we are using different methods on slightly different datasets. The 'plant effect's variance is quite small compared to residuals (and we will later see that it is not significant), which would imply that individual plant conditions (that are unrelated to Irrigation/Genotype), do not seem to have an effect on root length average.

	Estimate	Std. Error	t value
(Intercept)	2.53	0.181	13.9
72	-0.353	0.254	-1.39
Rialto	-0.392	0.247	-1.59
Sav	-0.325	0.2575	-1.27
Watered	-0.678	0.238	-2.84
72:Watered	0.24	0.337	0.711
Rialto:Watered	0.283	0.328	0.862
Sav:Watered	0.150	0.333	0.452

(a) Fixed Effects

	Variance	Std.Dev
Plant Effect	0.03432	0.1852566
Residuals	2.63633	1.623678

(b) Random Effects

Table 6: Results of a mixed model analysis of Root Lengths.

8.4.2 Multi-Analysis of Various Statistics

We apply our Mixed model to a variety of traits, and summaries of the results of those analyses are shown in Tables 7 and 8. Four traits were found to be significant at the 0.001 threshold, "Root Length", "Height", "Vertical Angle" and "Avg. Tortuosity (unravalled)".

Interestingly, "Vertical Angle" is significantly lesser for "Savannah" plants and has a significant random plant effect. This is because, the Wheat plants spread some of their roots directly downwards and others travel outwards first before travelling down. However, several Savannah plants don't have

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

many roots growing directly down, and this is not correlated with Irrigation.

	Genotype	Irrigation	Interaction	P.Value	Reduced
Root Length	0.24	2.7e-05*	0.73	0.0025	3.3e-05*
Width	0.92	0.055	0.75	0.62	-
Width (Unravalled)	0.3	0.043*	0.83	0.34	0.34
Height	0.078	1.2e-07*	0.72	8e-04*	8e-04*
Width/Height Ratio	0.62	0.19	0.53	0.47	-
Width/Height Ratio (Unravalled)	0.61	0.2	0.54	0.47	-
Horizontal Angle	0.97	0.52	0.53	0.9	-
Vertical Angle	5.2e-05*	0.0098*	0.054	0.00015*	0.00015*
Avg Tortuosity	0.0017*	0.96	0.34	0.013	0.013
Avg Tortuosity (Unravalled)	1.4e-05*	0.32	0.071	0.00017*	0.00017*
Centre of Mass (Horizontal)	0.86	0.021*	0.98	0.53	0.53
Centre of Mass (vertical)	0.19	5.7e-05*	0.96	0.0046	0.0046
Start Pos (Vertical)	0.81	0.96	0.39	0.8	-
End Pos (Vertical)	0.14	2.5e-05*	0.87	0.0017	0.0017

Table 7: Results of Anova tests performed on a large number of properties. The Genotype, Irrigation and Interaction columns contain the p-values of partial tests for relationships with those individual terms. The p-value column has the overall test for any relationship. The 'reduced' column has the overall test when unnecessary terms have been removed with AIC reduction methods.

	Mean	72	Rl	Sv	W	W:72	W:Rl	W:Sv	$\sigma(\text{Plant})$
Root Length	2.3*				-0.5*				0.15
Height	2.2*	-0.31	-0.37	-0.38	-0.62*	0.17	0.28	0.12	0
Vertical Angle	0.98*	-0.073*	-0.046	-0.12*	-0.096*	0.075	0.11*	0.058	0.029*
Avg Tortuosity (Unravalled)	0.95*	0.0061	-0.0025	-0.021*	0.004	-0.014	0.0064	0.0029	0.0036

Table 8: Coefficients of fitted linear models on properties in Table 7, that are significant at the 0.001 threshold. The asterisk * next to a coefficient denotes that the coefficient is significant at the 0.05 threshold. A blank entry means that term was removed during AIC reduction. W=Well-watered, Rl=Rialto, Sv=Savannah

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

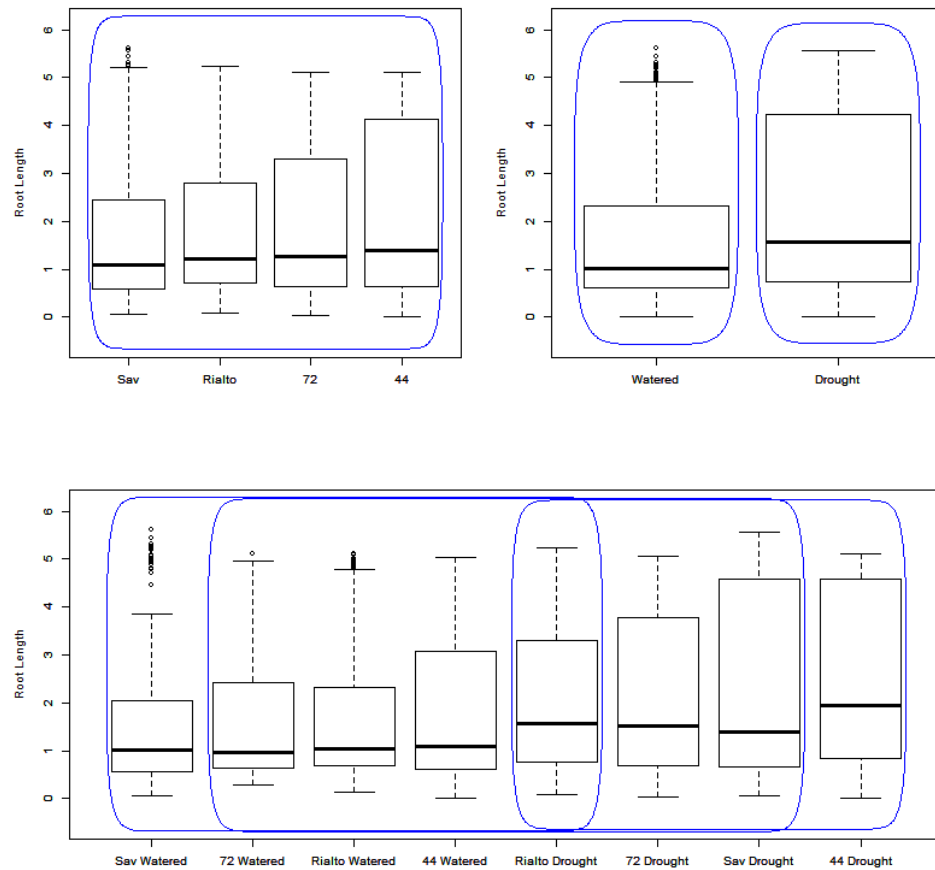


Figure 52: Boxplot of "Root Length" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

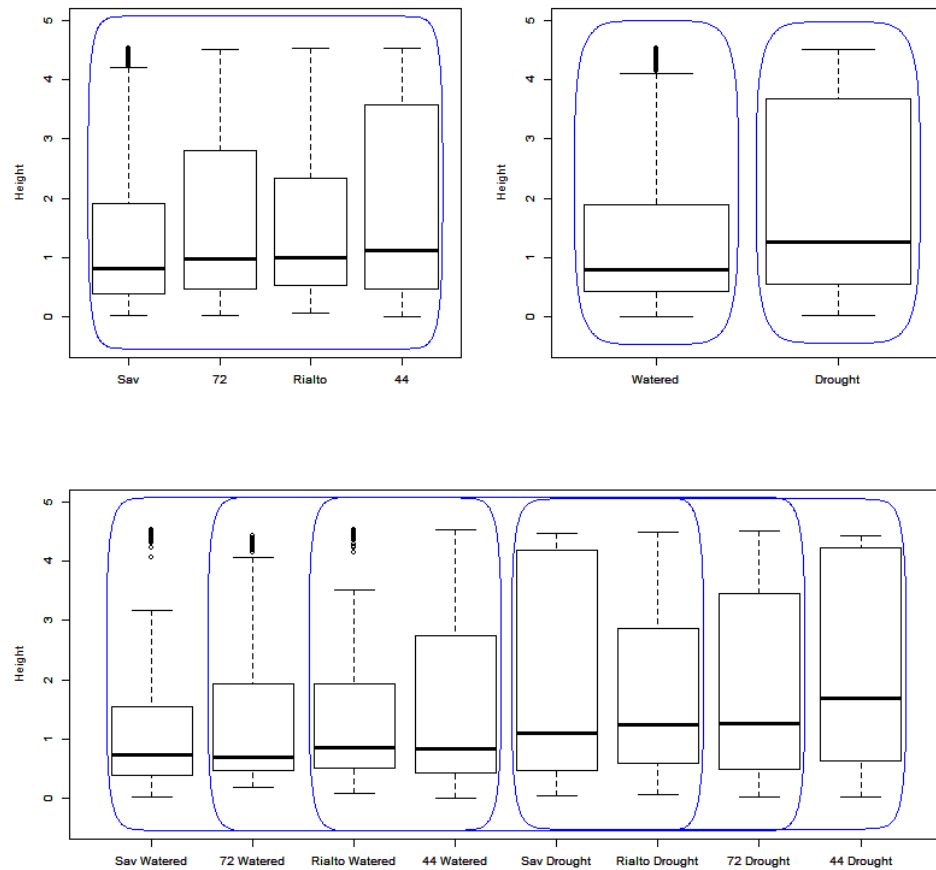


Figure 53: Boxplot of "Height" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.4 ANALYSIS OF INDIVIDUAL ROOT STRAND PROPERTIES

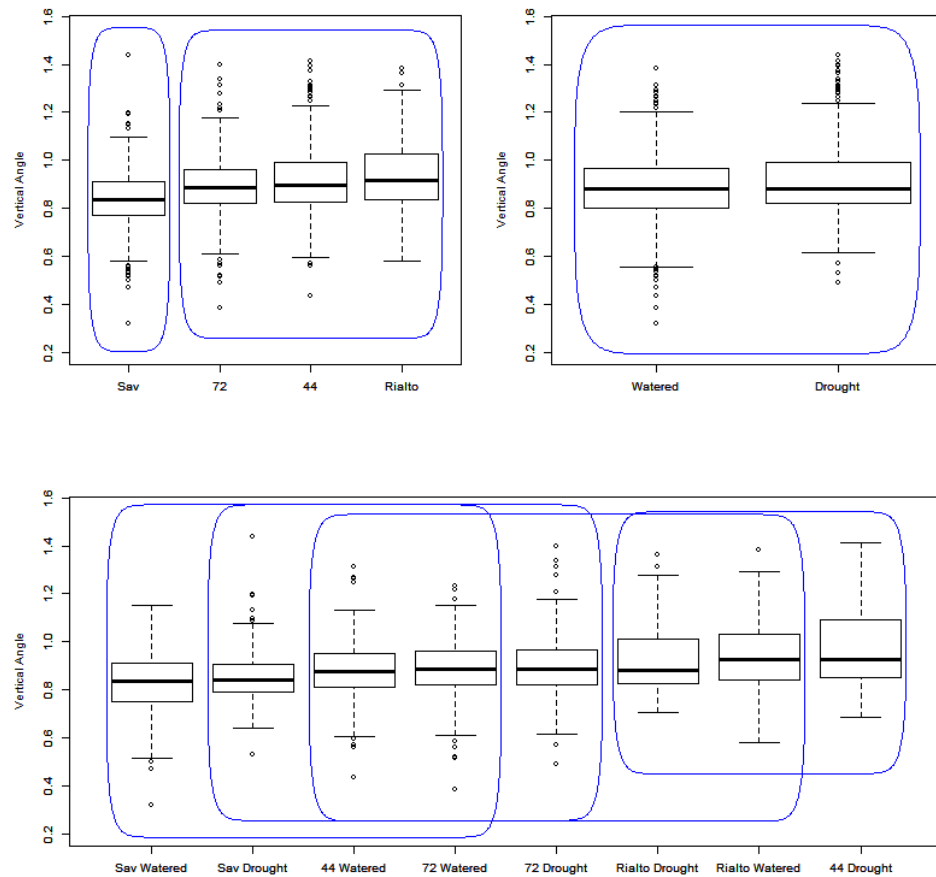


Figure 54: Boxplot of "Vertical Angle" versus the genotype, irrigation factors and their combination. The factors are plotted in ascending order in terms of mean. The blue boxes denote that any two factors within the same box are not significantly different.

8.5 RESULTS OF TRAIT ANALYSIS

There are numerous traits and properties that are strongly associated with the species and drought factors and their interactions. It is difficult to interpret all of the complex interactions between these factors though.

No trait appeared to have a significant relationship with the genotype factor alone, which is surprising since Rialto and Savannah are known to have different root lengths. This may be because the wheat was scanned before being fully grown.

In Table 9, we can see the effects that water has on measured traits (that we can identify as being significant). We have also included a measure of the "relative effect", this is the effect as a percentage of the standard deviation of the trait. The "Root Count", "Root Length Total" and "Plant Height" are the traits most affected by Irrigation, with more roots being grown in a well-watered environment. The average root length is slightly reduced but not by a very large margin.

However we can break these effects down even further in Tables 10, 11, 12 and 13 and look at individual effects. Each has a slightly different response, for example Plant Height is only increased by being watered in "72" and "Savannah". We can conclude that the different genotypes respond to drought in different ways, which is an important result since that is not an assumed fact and we can use it to determine which is more resilient.

Most of the significant results we have inferred are related to metadata about the plant, and not the scanned roots. Root Count and Root Length Total were the most significant factors and, even then, only in the pooled effect of irrigation. This seems to show there is not a lot of shape variation within the scanned root data, however while we have tested a large number of traits, there is no reason to believe that we have not missed an important trait. This approach has provided a lot of insight but there is no way to tell if it is complete.

8.5 RESULTS OF TRAIT ANALYSIS

	Effect	Std. Error	Relative Effect
Root Length Total	7.27	2.68	56.464%
Root Length Average	-0.273	0.0896	-2.119%
Width	0.03	0.0103	0.233%
Root Count	8.28	1.89	64.358%
Convex Hull Surface Area	0.322	0.125	2.506%
Plant Height	6.25	1.21	48.596%
Total above-ground biomass	4.01	0.502	31.139%
Fertile shoots	1.97	0.247	15.300%
Infertile shoots	-1.31	0.213	-10.200%
WinRHIZO total length	2.98	0.224	23.188%
Water uptake	3.05	0.386	23.740%
RLD 0-20cm	0.295	0.0341	2.290%
RLD 20-40cm	0.244	0.0162	1.899%

Table 9: Significant effects of the 'Watered' factor on our set of measured traits. (The 'Drought' factor has the negative of these effects.)

	Effect	Std. Error	Relative Effect
Fertile shoots	1.16	0.436	8.986%
Infertile shoots	-1.44	0.339	-11.172%
WinRHIZO total length	2.64	0.253	20.497%
Water uptake	5.08	0.446	39.511%
RLD 0-20cm	0.35	0.028	2.718%
RLD 20-40cm	0.286	0.0214	2.225%
RLD 40-60cm	-0.0294	0.0129	-0.228%
RLD 60-80cm	-0.183	0.0328	-1.423%

Table 10: Significant partial effects of the combination of 44 and 'Watered' factors on our set of measured traits. (the 'Drought' equivalent has the negative of these effects)

	Effect	Std. Error	Relative Effect
Plant Height	10.4	1.47	80.945%
Fertile shoots	1.41	0.436	10.929%
WinRHIZO total length	3.22	0.253	24.985%
Water uptake	3.92	0.446	30.457%
RLD 0-20cm	0.537	0.028	4.175%
RLD 20-40cm	0.169	0.0214	1.311%
RLD 40-60cm	-0.0419	0.0129	-0.325%
RLD 60-80cm	-0.161	0.0328	-1.248%

Table 11: Significant partial effects of the combination of 72 and 'Watered' factors on our set of measured traits. (the 'Drought' equivalent has the negative of these effects)

8.5 RESULTS OF TRAIT ANALYSIS

	Effect	Std. Error	Relative Effect
Fertile shoots	2.16	0.436	16.757%
Infertile shoots	-1.44	0.339	-11.172%
WinRHIZO total length	2.15	0.253	16.728%
RLD 0-20cm	0.102	0.028	0.794%
RLD 20-40cm	0.216	0.0214	1.681%
RLD 40-60cm	0.0431	0.0129	0.335%

Table 12: Significant partial effects of the combination of Rialto and 'Watered' factors on our set of measured traits. (the 'Drought' equivalent has the negative of these effects)

	Effect	Std. Error	Relative Effect
Plant Height	12.2	1.47	94.934%
Fertile shoots	3.16	0.436	24.529%
Infertile shoots	-1.69	0.339	-13.114%
WinRHIZO total length	3.93	0.253	30.542%
Water uptake	2.39	0.446	18.567%
RLD 0-20cm	0.19	0.028	1.474%
RLD 20-40cm	0.306	0.0214	2.380%
RLD 40-60cm	0.0281	0.0129	0.219%
RLD 60-80cm	0.102	0.0328	0.792%

Table 13: Significant partial effects of the combination of Savannah and 'Watered' factors on our set of measured traits. (the 'Drought' equivalent has the negative of these effects)

8.6 OTHER PROPERTIES

8.6.1 *Horizontal Growth*

We wish to see if the plants growth has a bias to one side or another (potentially caused by orientation to the sun or some other external environmental factor), or if the roots spread out randomly in all directions in the horizontal plane.

We can do this by performing the Rayleigh Test for Uniformity on each plants set of horizontal growth angles. This will give us a p-value that indicates when the growth angles have a bias to one particular direction. We have to perform this test separately on each root system, in order to eliminate the 'orientation of the column in the CT scanner' which is a nuisance variable. Once we have performed these tests for each plant, we can combine them into a single overall test, using Fisher's method.

Performing this combined test for our data gives a p-value of 0.998, which is not significant. From this we conclude there's no evidence in our data of a bias to a given direction.

8.7 PAIRWISE DISTANCE APPROACH

In this section we will analyse the dataset using our two Hausdorff metrics. We will be running this analysis of the original and unravelled versions of our dataset.

We begin by performing the large sets of comparisons, to produce four 32x32 matrices that are plotted in Figures 55, 56, 57 and 58. We use hierarchical clustering to try to group similar plants together. It is difficult to analyse the plants in this format, and there does not appear to be any obvious clustering.

8.7 PAIRWISE DISTANCE APPROACH

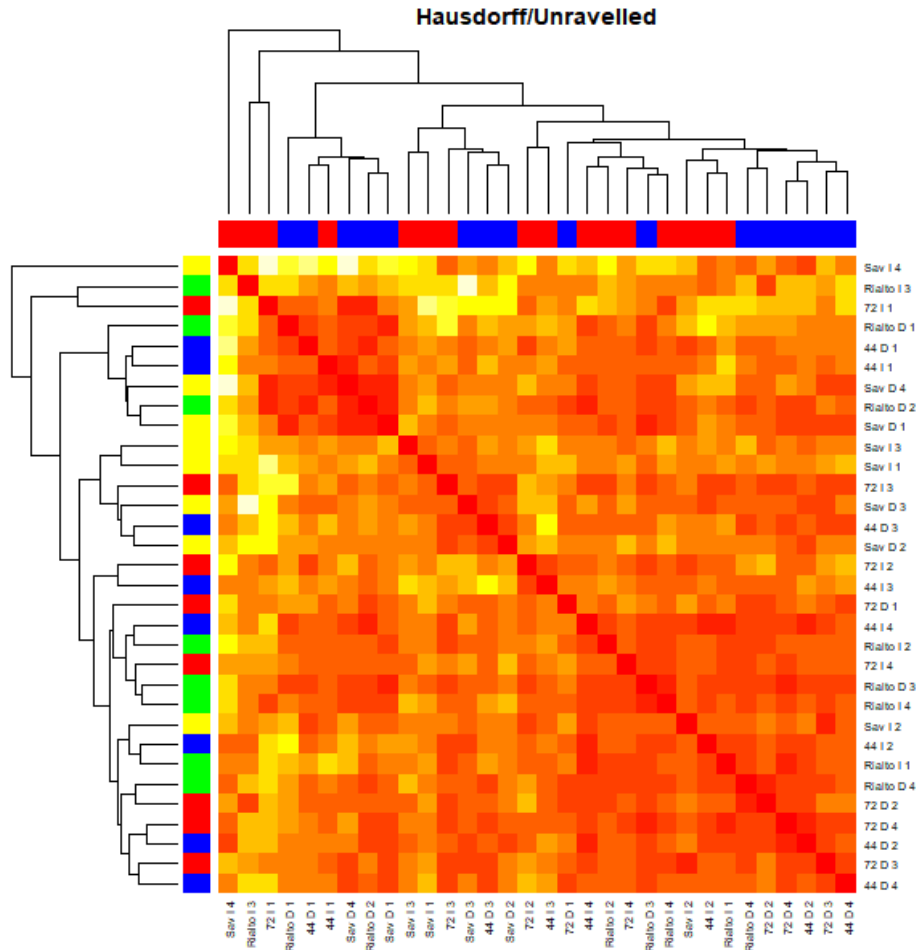


Figure 55: Heatmap view with Hierarchical clustering of the pairwise distances calculated by using the Hausdorff distances on the original data. The coloured sections on the top and left axis represent the Irrigation and Genotype factors on that entry.

8.7 PAIRWISE DISTANCE APPROACH

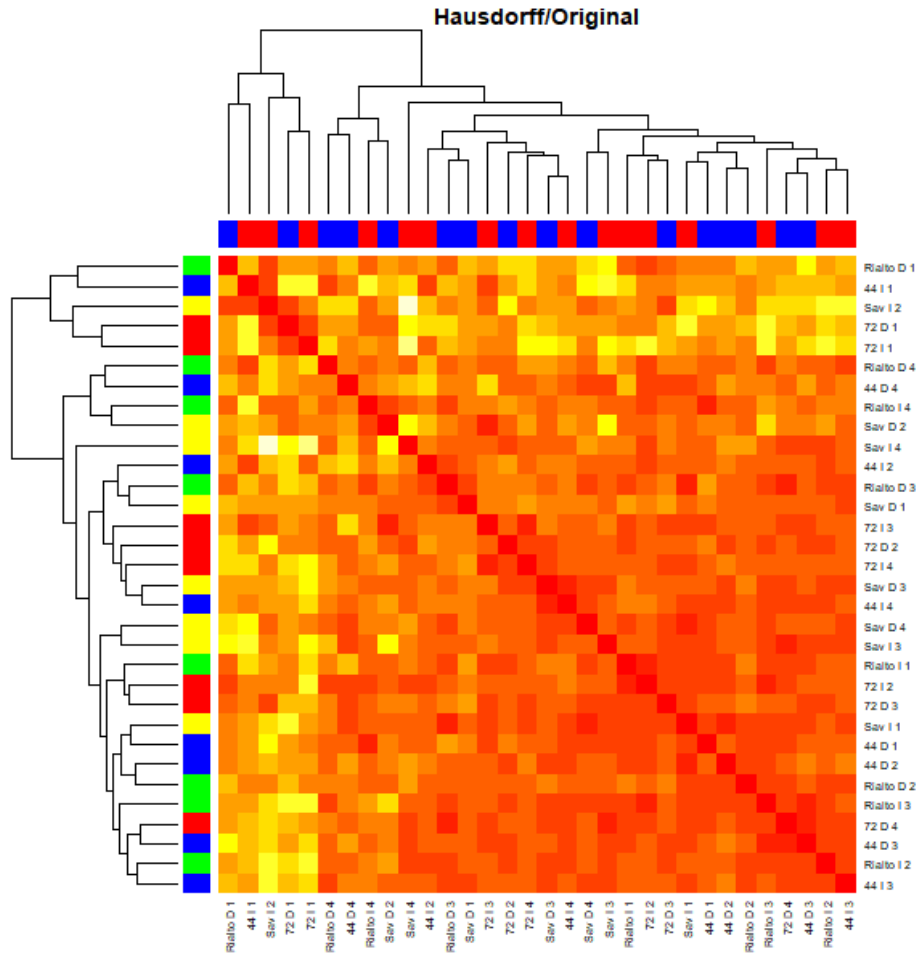


Figure 56: Heatmap view with Hierarchical clustering of the pairwise distances calculated by using the Hausdorff distances on the unrav-elled data. The coloured sections on the top and left axis represent the Irrigation and Genotype factors on that entry.

8.7 PAIRWISE DISTANCE APPROACH

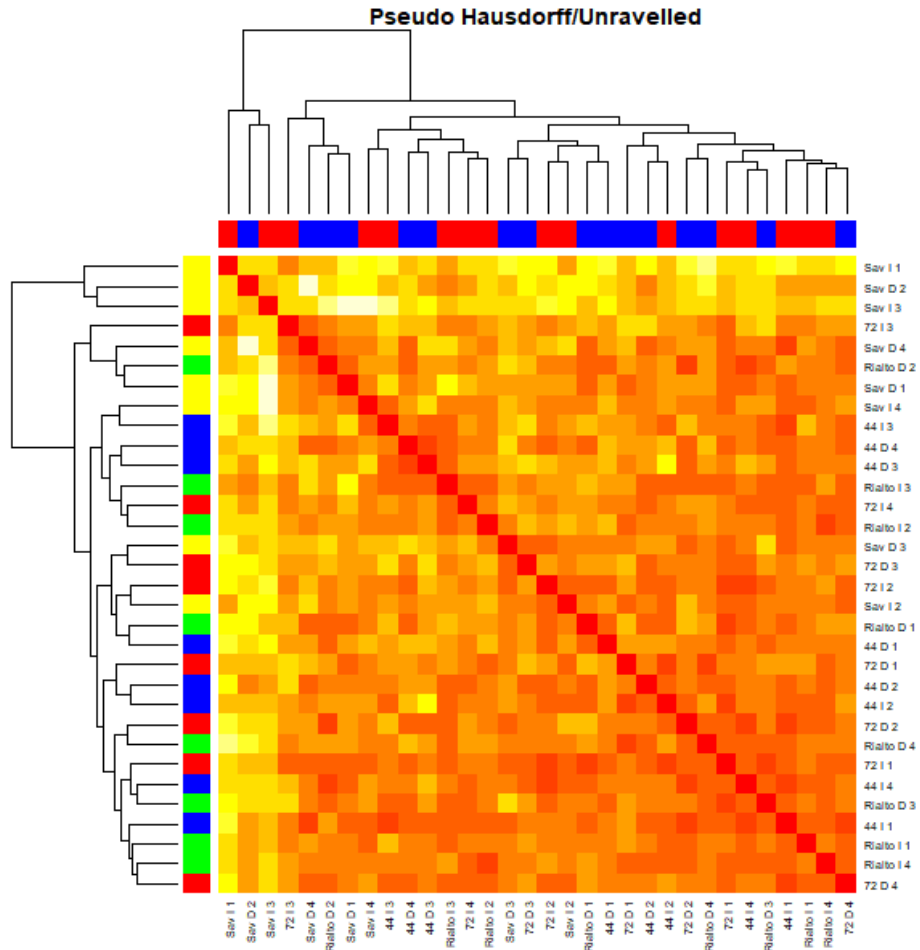


Figure 57: Heatmap view with Hierarchical clustering of the pairwise distances calculated by using the Pseudo Hausdorff distances on the original data. The coloured sections on the top and left axis represent the Irrigation and Genotype factors on that entry.

8.7 PAIRWISE DISTANCE APPROACH

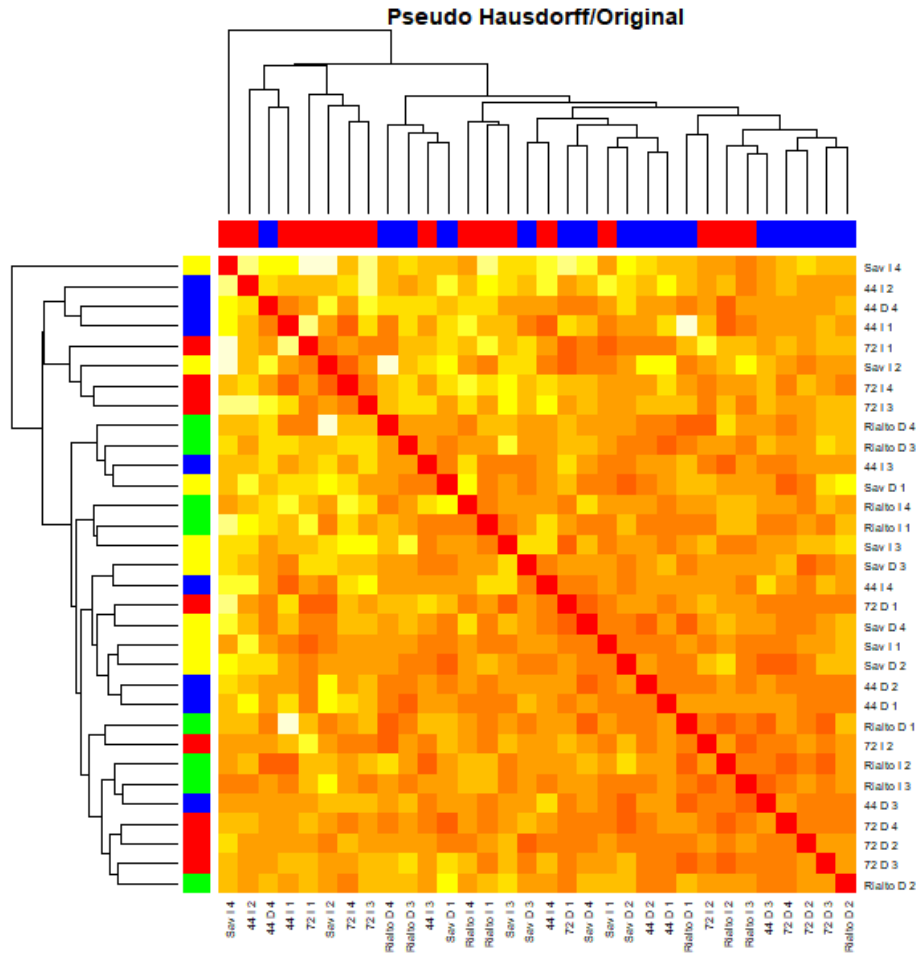


Figure 58: Heatmap view with Hierarchical clustering of the pairwise distances calculated by using the Pseudo Hausdorff distances on the unravelled data. The coloured sections on the top and left axis represent the Irrigation and Genotype factors on that entry.

We now transform this data using Multi-dimensional scaling to a set of 3 MDS variables. The choice of 3 MDS variables was discussed and justified in Section 7.5. We can see plots of the MDS variables for the various combinations in Figures 59, 60, 61 and 62. Visually, it is difficult to find any form of pattern within the results.

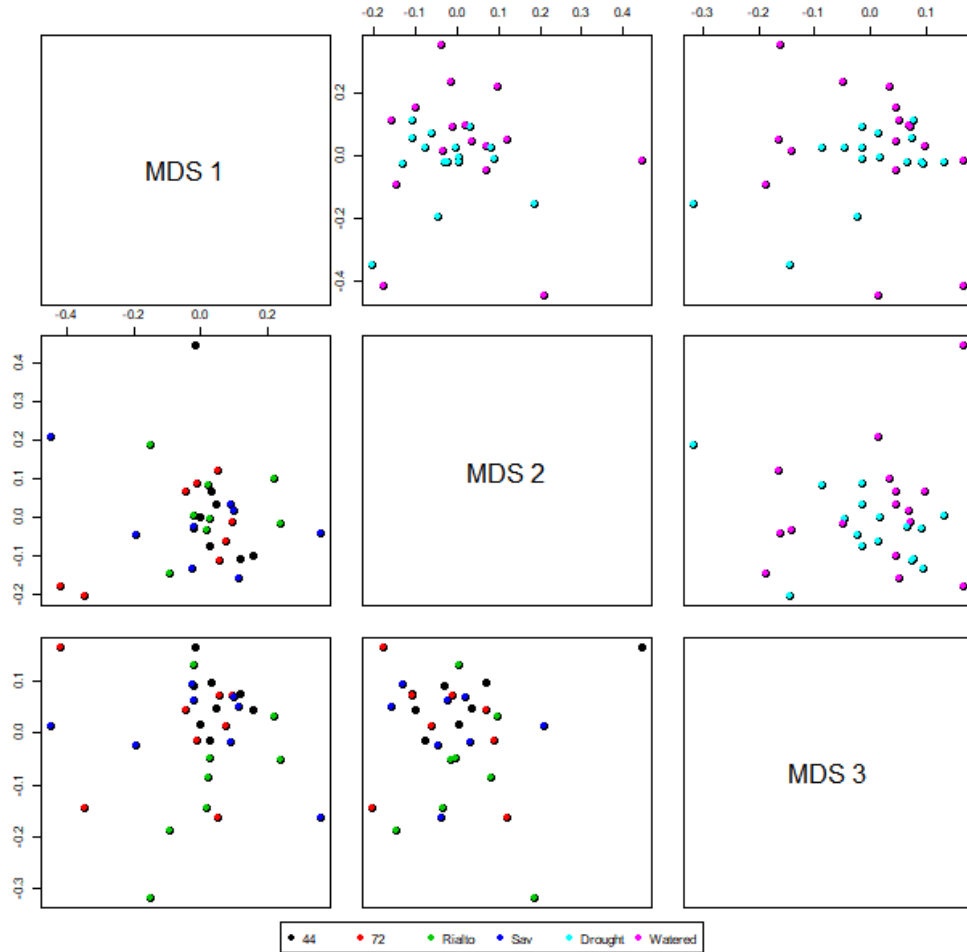


Figure 59: Scatterplot of the MDS variables derived from the Hausdorff metric over the original dataset.

We can perform Multivariate Analysis-of-variance (MANOVA) tests to see if any of the results have significant relationships with our factors (Anderson et al., 1958). The results of these tests are seen in Table 14. Since we are performing 12 tests we will set our p-value significance threshold to be at the $\frac{0.05}{12} = 0.00416$ threshold. Unfortunately, while the Hausdorff/Original seems to be fairly low, ultimately none of these tests appear to be significant,

8.7 PAIRWISE DISTANCE APPROACH

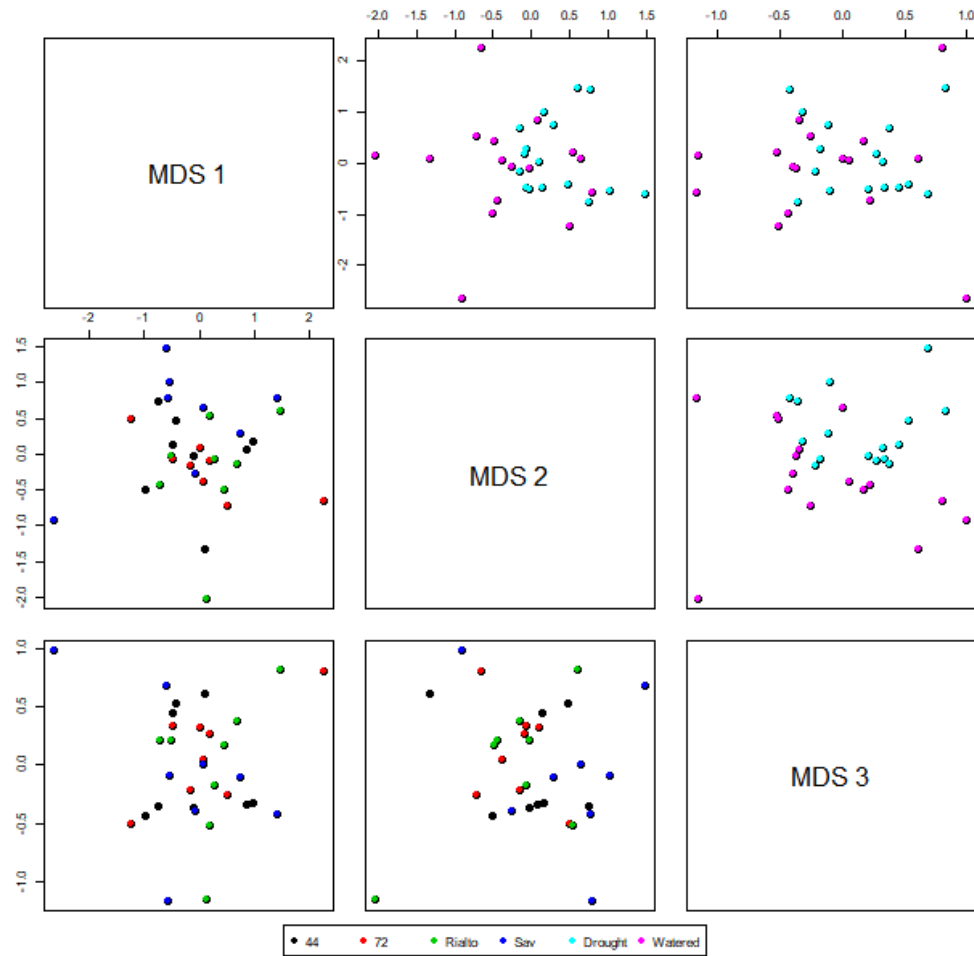


Figure 60: Scatterplot of the MDS variables derived from the Hausdorff metric over the unravelled dataset.

8.7 PAIRWISE DISTANCE APPROACH

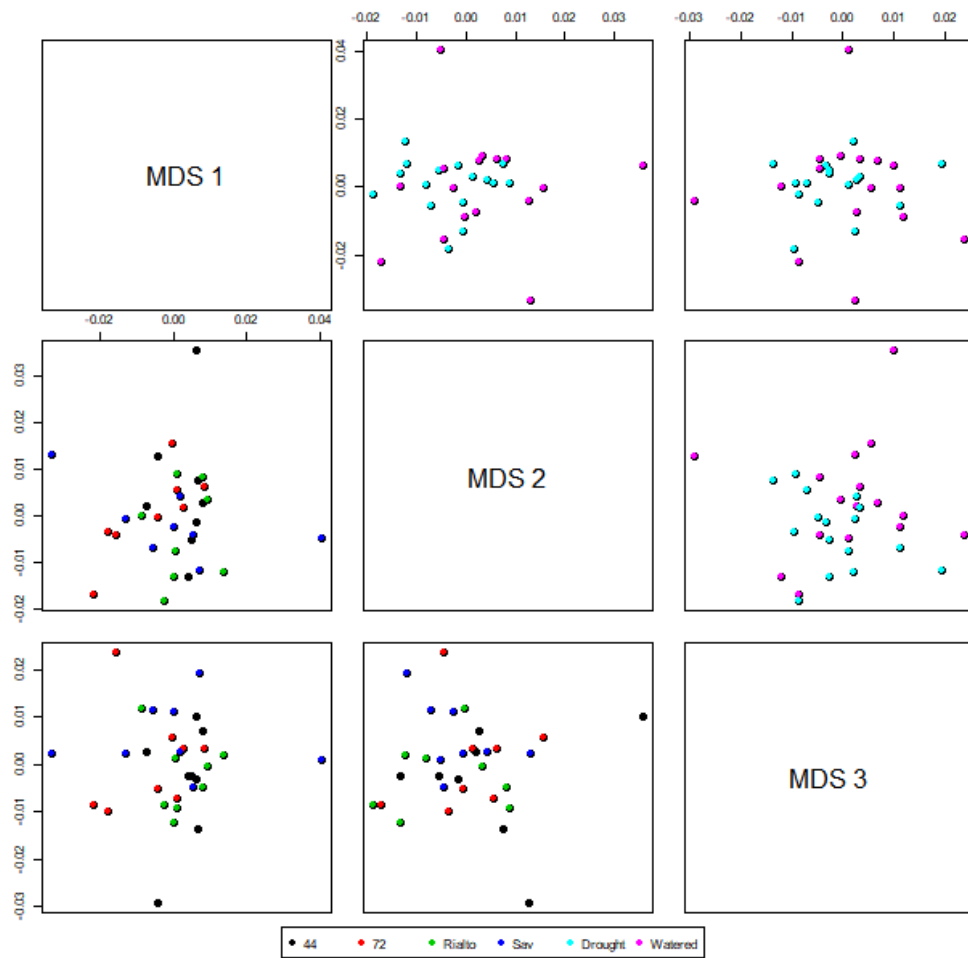


Figure 61: Scatterplot of the MDS variables derived from the Pseudo Hausdorff metric over the original dataset.

8.7 PAIRWISE DISTANCE APPROACH

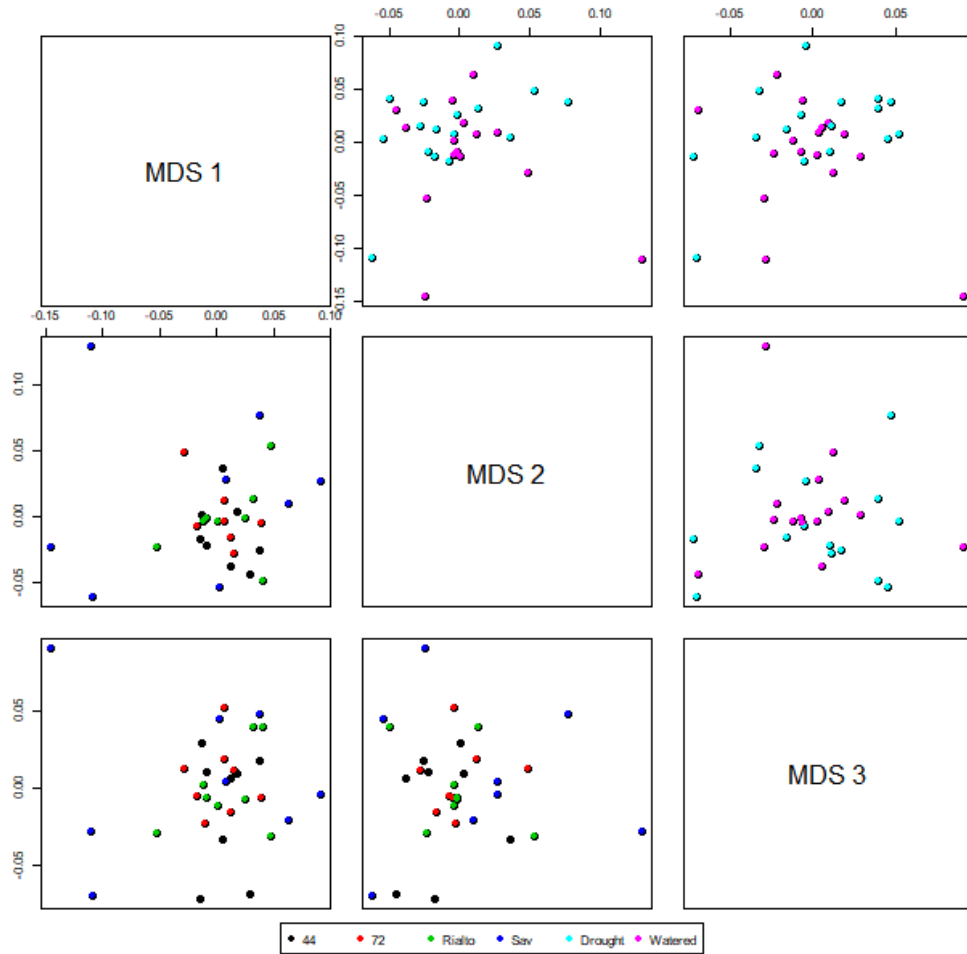


Figure 62: Scatterplot of the MDS variables derived from the Pseudo Hausdorff metric over the unravelled dataset.

which is consistent with the conclusions of our traits analysis, in Section 8.5, since very few traits had a significant relationships with our factors.

The MANOVA tests have several assumptions, such as normality and homogeneity of variances and covariances, and these appear to be reasonable for our data, although it is difficult to tell given the small sample size. There are also some minor outliers in the data which the MANOVA test can be sensitive to.

The one measured trait that had a very large relationship with Irrigation was the "Root count", and it seems the Hausdorff style metrics are having difficulty picking that up. This is not entirely unexpected since two roots that follow each other closely can effectively 'mask' each others effect on the results, so it may be that root count cannot be effectively estimated using this approach.

This is a disappointing result since it does not find any significant relationships, but it does the advantage of being a little more definitive than our traits results (since a non-significant trait result could simply be that there a missing trait that we missed). We can say with confidence that there are no relationships with shape at least in terms of 'broad strokes' shape distribution. Any relationships that exist must be subtler and not obvious from a visual glance.

This is in contrast to the traits results where we had to add the proviso that "we may have missed an obvious significant trait." With this approach we can at least say "if we have missed any significant relationships then they are probably not very obvious."

	Irrigation	Genotype	Interaction
Hausdorff/Original	0.013	0.61	0.79
Hausdorff/Unravelled	0.61	0.28	0.81
Pseudo Hausdorff/Original	0.44	0.77	0.71
Pseudo Hausdorff/Unravelled	0.33	0.37	0.7

Table 14: P-Values for various tests on different combinations of metric and dataset for a relationship with different factors.

8.8 TOPOLOGICAL DATA ANALYSIS

Now we wish to perform a TDA analysis on our original data and the estimated unravelled version.

We perform the Persistent Homology method based on the Vietoris-Rips complex on each of the plants and its unravelled version. The results of this can be seen in Figures 63 and 64. It is difficult to see any form of pattern from the barcodes alone. We may notice that the 1-dimensional homology classes in the unravelled dataset's barcodes are more spread out than those in the original dataset's barcodes.

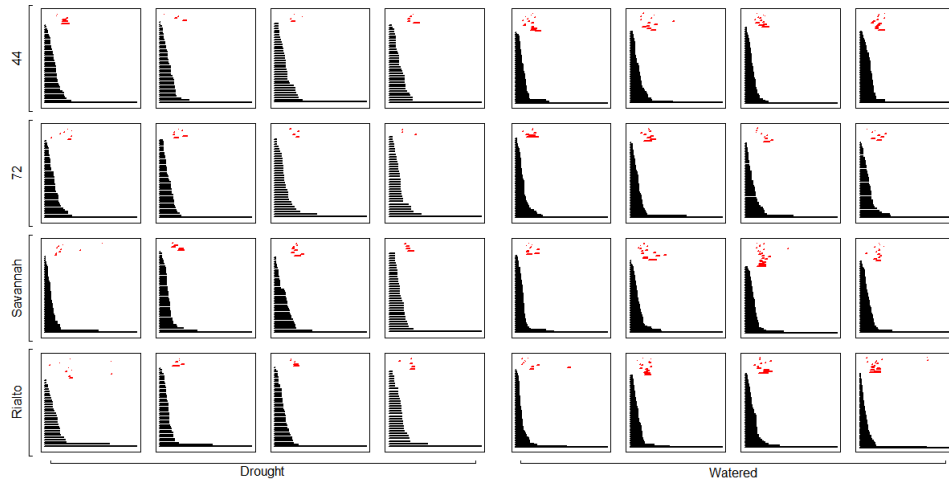


Figure 63: Barcodes of the Vietoris-Rips of the plants in the original dataset. 0-dimensional homology classes are pictured in black, and 1-dimensional homology classes are pictured in red.

We can perform the bottleneck distance analysis using both the 0-dimensional and 1-dimensional classes, and then perform MDS to get an overview. We can see the results in Figures 65 and 66, and it looks like MDS 3 of the original data is affected by the Irrigation. We can test for relationships using manova tests (at the $\frac{0.05}{6} = 0.0833$ threshold since we are doing 6 tests), the results of which are seen in Table 15, and it does seem that there is a relationship between Irrigation and the original data MDS variables.

In addition there seems to be a "horse-shoe effect" (as labelled by (Kendall, 1971)) in the plot of the 'Original data - MDS 1 vs MDS 2'. This occurs when

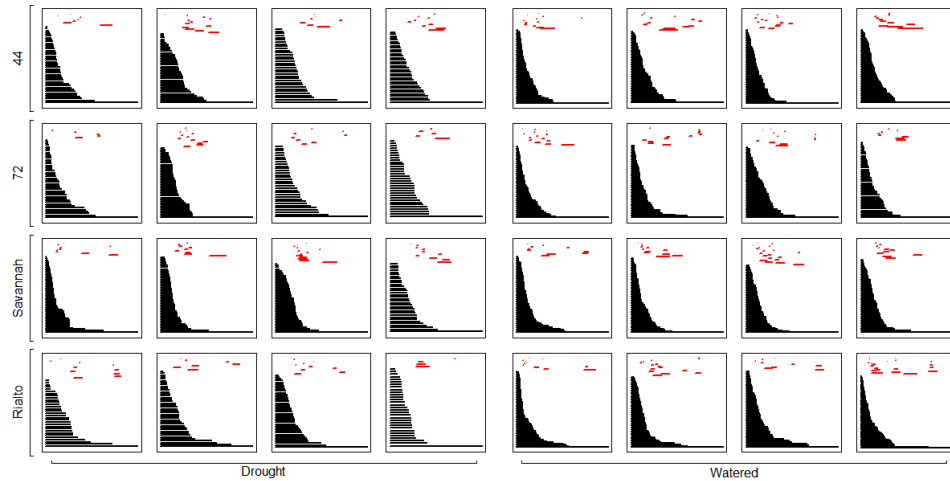


Figure 64: Barcodes of the Vietoris-Rips of the plants in the unravelled dataset. 0-dimensional homology classes are pictured in black, and 1-dimensional homology classes are pictured in red.

large distances are difficult to distinguish from moderate distances, and the points form a horseshoe shape (Mardia et al., 1979, p412)

	Irrigation	Genotype	Interaction
Original	0.002*	0.27	0.95
Unravalled	0.9	0.036	0.075

Table 15: P-values for various MANOVA tests on different datasets for a relationship with different factors.

After some exploration it appears that MDS 3 is strongly correlated with the Root Count, see Figure 67, which we know from before is affected by Irrigation.

A problem with this result, is that while we could interpret that MDS 3 is a measurement of "Root Count", it is also possible to interpret that MDS 3 is measuring something else that is correlated with root count or measuring a combination of things that includes Root Count. This is an unfortunate limitation of TDA that it is difficult to interpret the results and their relationship with our data.

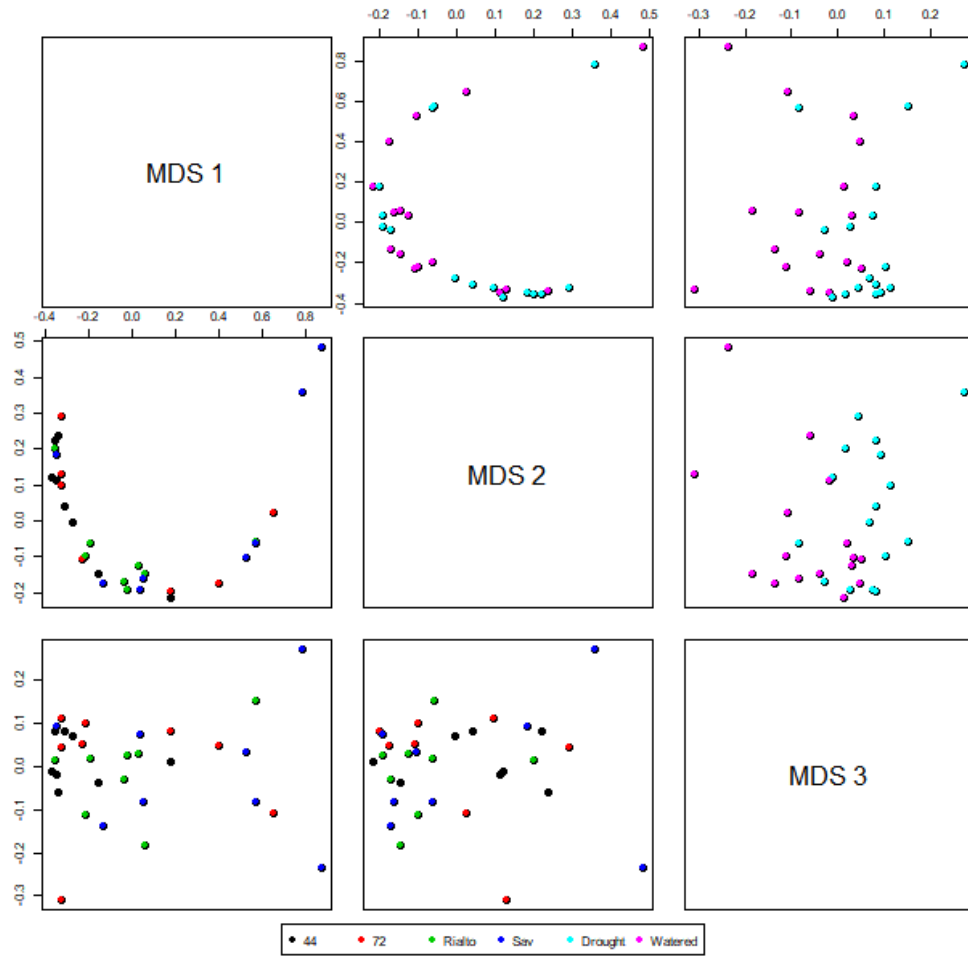


Figure 65: MDS variables from the TDA analysis of the original datasets

8.8 TOPOLOGICAL DATA ANALYSIS

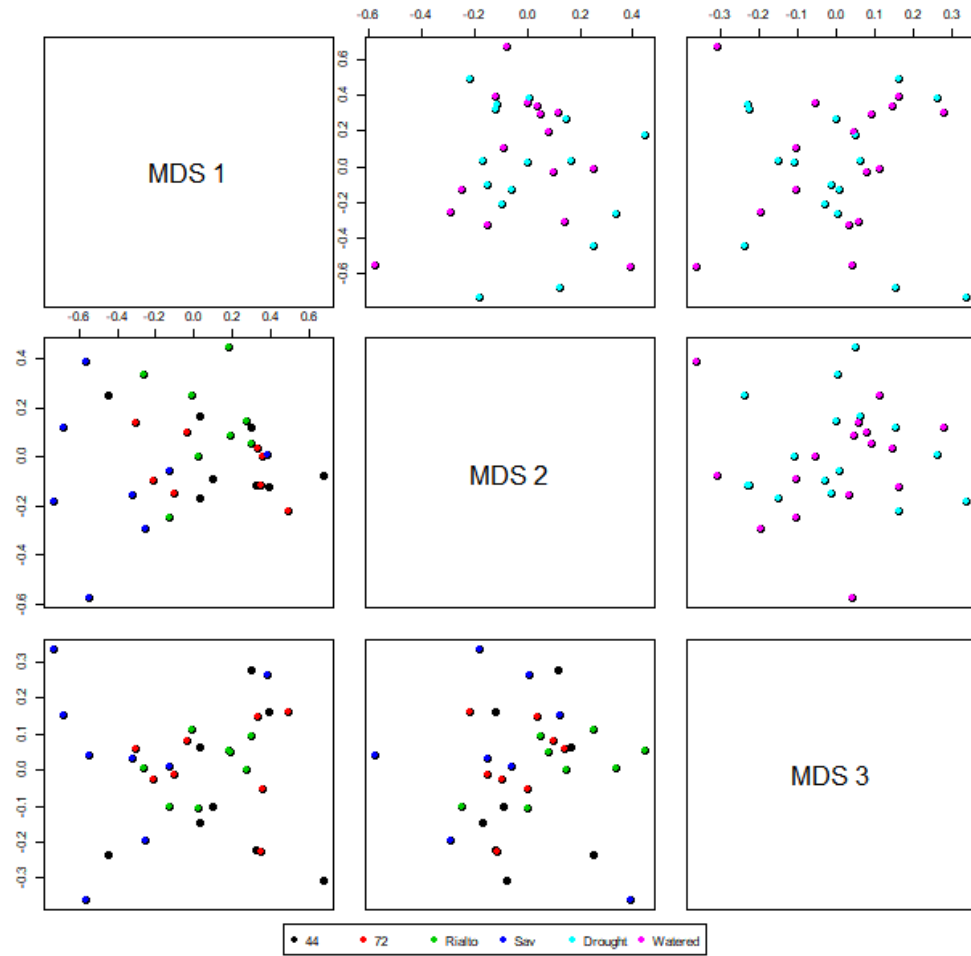


Figure 66: MDS variables from the TDA analysis of the unravelled datasets

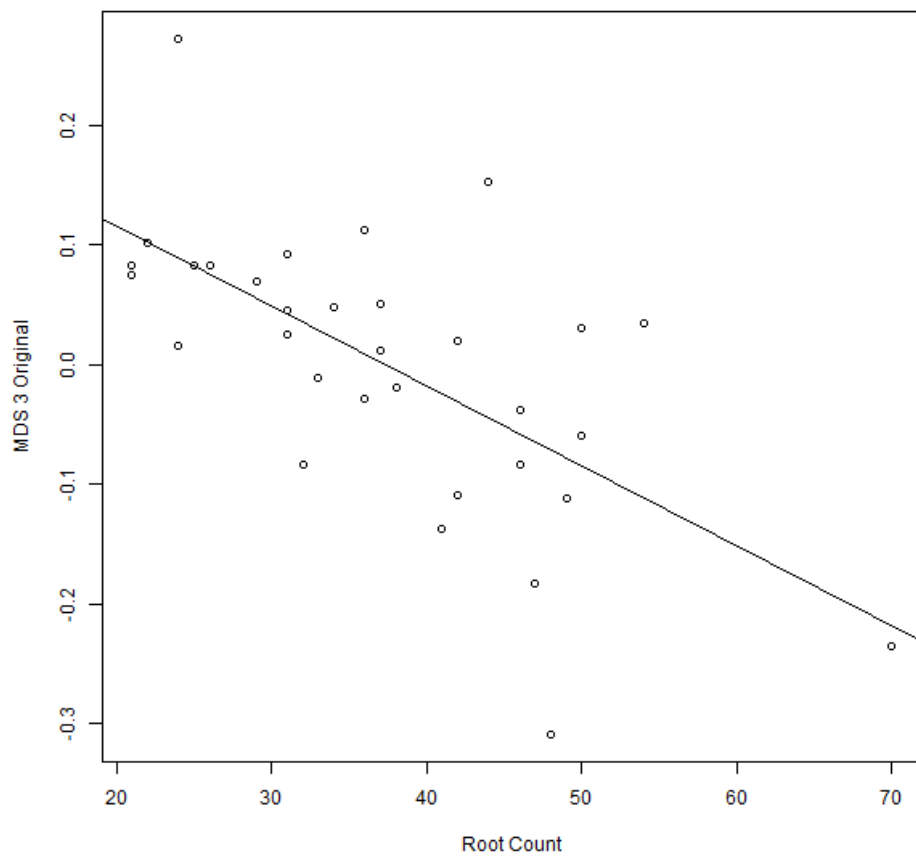


Figure 67: Plot of MDS 3 (original data) Root Count.

8.9 CONCLUSION

Our three approaches have allowed us to examine our datasets from three different perspectives and we have found similar results from all three. Our traits analysis provided us with a lot of significant relationships but mostly with 'metadata' traits and not with traits from our scanned root systems, with "Root Count" being the only trait that was related to Irrigation. The Hausdorff metrics failed to find any significant relationships, but the TDA approach was also able to detect a significant relationship with a variable that was strongly correlated with Root Count.

From these results we can safely conclude that there is a significant relationship between Root Count and Irrigation.

CONCLUSIONS AND FURTHER WORK

In this chapter we will summarise the findings of our thesis and consider possible future avenues of research.

9.1 RESULTS OF OUR INVESTIGATIONS

The CPIB Future Roots project has developed a method for scanning the roots of plants while they are in soil, and has been running experiments on wheat plants to understand responses to drought conditions. The aim of this thesis has been to develop a framework in which these scanned root systems can be analysed. This framework comprises of three different approaches.

In Chapters 1 and 2, we introduce the background of the Future Roots project, the biology of root systems, and the existing analysis methods that we will be basing ours upon.

In Chapter 3, we look at root systems and how they are structured and identify the approaches that we will be needing to take to fully understand them. We also describe nuisance factors of "horizontal rotation and translation" that will need to be eliminated from our analysis. We then described a problem with plant roots growing too large and hitting the boundary of their container, disrupting their path, and propose an 'unravelling' method to estimate the path the root would have taken if the boundary were not there.

In Chapter 4, we describe the approach based on measuring “traits” or simple statistics of a plant root system and testing for significant relationships. This is the conventional approach that is used to analyse root shapes. In our simulated study and our practical data, it proves itself to be an effective method for shape analysis. Its advantages are that it is simple, based on well-established ideas and results can be easily interpreted, but its key disadvantages are that there may be traits that are missed, and that there are some complex properties of root systems that cannot be properly described by fixed-dimensional statistics.

In Chapter 5, we develop methods for pairwise comparison of root systems. This approach, which uses metrics to find a distance between two roots, is a lot more flexible than traits since we can ensure that any difference between the two root systems means that our statistics will identify them as being different. We describe the Hausdorff metric and, after identifying weaknesses of the Hausdorff metric, we describe a slight variant which we call the Pseudo-Hausdorff metric that compensates for them. We also discuss how the nuisance effects of ‘horizontal rotation’ and ‘horizontal translation’ can be eliminated from comparisons between root systems.

The pairwise comparison approach has the advantage that it can capture any form of shape differences between two roots systems and test for significant relationships, however it does have the disadvantage that if there are multiple shape differences it is difficult to identify them from the output and one strong shape difference may mask a smaller one.

In Chapter 6, we apply Topological Data Analysis methods to root systems, specifically the Persistent Homology method. This builds a ‘barcode’ of homology classes or ‘holes’ that represent important features in our root system. In our simulated study and our practical data, this method performed effectively and was able to identify significant relationships. The advantage of this approach is that it exposes an aspect of a dataset that was previously hidden.

The major disadvantage is that it is very difficult to interpret what this new ‘aspect’ actually represents. Barcodes are very abstract and it is difficult to understand how they relate to the original point cloud.

In Chapter 7, we performed a simulation study using an existing dataset of simulated images. Our approaches were able to effectively and consistently identify shape differences within the simulated sets, and would also fail to identify shape differences where none existed.

In Chapter 8, we applied our approaches to an existing practical dataset from the FutureRoots program. The dataset consisted of plants chosen from four genotypes (“44”, “72”, “Savannah” and “Rialto”) grown under “Drought” and “Watered” irrigation conditions. The results of analysing this dataset using each of our three approaches were consistent with each other and we were able to show that under drought conditions the number of roots grown is significantly reduced.

The ‘measured traits’ approach was also able to make use of additional metadata that showed complex relationships between Irrigation and Genotype, for example, ‘Plant Height’ was significantly increased under well-watered conditions for two of the genotypes (72 and Savannah) but had no change for the other two (44 and Rialto).

The pairwise distance using Hausdorff distance metrics failed to find a significant relationship which is a pity but from that we can at least assume that there is not a ‘large’ amount of shape differences between the various populations.

The TDA approach provided us with a large amount of new data, which is difficult to interpret its meaning. We were able to show that there is a significant relationship with the irrigation condition, likely due to a strong correlation with root count.

Our dataset is not huge and does not appear to have any large shape differences from what we can tell, so this was a challenging dataset to analyse. We believe that we have demonstrated the effectiveness of all three approaches as well as the limitations that each one possesses.

9.2 FURTHER WORK

9.2.1 *Time Series Analysis*

One experiment that we wished to analyse was that of 'time series', where a single plant is repeatedly scanned over the length of its growth cycle. The goal would be to create a sequence of daily scans that follow the plants root system from the beginning of its germination until it is ready to be harvested. We could then analyse its growth and try to answer questions about how predictable its future structure was and how existing structures might change.

Unfortunately we were unable to perform these experiments due to logistical issues. Scanning a wheat column takes time and there was not enough room on the schedule to perform daily scans of more than one or two plants. In addition, wheat plants have a chance of failure, especially when growing in drought conditions, and if the plant failed too quickly, we could not be certain the data would be useful.

Despite this, some time series scans of growing wheat plants and barley plants were successfully performed, but then there were further issues with the 'skeletonization' software having difficulty consistently lining up the various scans to a common frame of reference. It would have been difficult to tell if small changes in structure were related to actual changes or measurement error.

Due to the low amount of samples available and the various measurement issues, we decided to focus on the traditional experiments that simply scanned the plants at the end of their growth cycle.

Despite this, we feel that examining a plants root system as it grows is potentially an important and insightful experiment to perform, if the issues could be solved.

9.2.2 *Soil Analysis*

One avenue that was considered but not pursued, was analysis of the raw CT images. The CT scans capture a lot of detail about the state of the soil as well as the plant's root system. It is possible to analyse this data to get information about the location of soft soil, hard soil, water and air pockets.

The first difficulty to performing such an analysis was the size of the data, with even a single column scan containing several gigabytes of information. This makes developing a full analysis method difficult since it requires a lot of processing time to perform even a basic analysis. Approximation of the soils contents is not feasible either, since even though the data is very high resolution, the small soil particles and the plants roots are still at the limit of what that high resolution can capture.

This difficulty could be overcome, but there is also the issue that the CT scan has a built-in error that is not simple to account for. CT scanners work by projecting a stream of x-rays through the object at a wide range of angles, and recording what passes through. The idea being that dense objects will absorb the rays slightly while air and less dense objects will let them pass. From this you can reconstruct the contents of the object you were scanning. The problem is that dense objects can leave a 'shadow' on the nearby area and the scan will detect slightly denser areas around a dense object, and it is difficult to know how far this effect extends.

The 'shadow' issue could potentially be fixed with a higher resolution scan or a more sophisticated scanning technique, but that is not feasible with the current technology available.

Despite the difficulties, we believe that an analysis of the soil and its relation to the wheat plants growth would be an useful analysis, especially if it were combined with the time series approach outlined above.

9.2.3 *QTL Analysis*

In addition to metadata about the plants root system and growth, we also have genetic information about its DNA. The roots used in our experiments are cross-breeds that have been created to use in a method called QTL Analysis. Given a single continuous phenotype from a large variety of different plant genotypes, QTL analysis can track the changes in DNA to look for positions in a DNA strand that are correlated with changes in the phenotype.

QTL analysis is applicable to the results obtained from our methods, since the MDS variables generated from pair-wise analysis are continuous phenotype data. It is possible that we may find a DNA region that is correlated with the MDS variable, that was not known to correlate with more conventional traits.

REFERENCES

- Anderson, T. W., Anderson, T. W., Anderson, T. W., Anderson, T. W., and Mathématicien, E.-U. (1958). *An introduction to multivariate statistical analysis*, volume 2. Wiley New York.
- Burago, D., Burago, Y., and Ivanov, S. (2001). *A course in metric geometry*, volume 33. American Mathematical Society Providence.
- Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.
- Charpiat, G., Faugeras, O., and Keriven, R. (2005). Approximations of shape metrics and application to shape warping and empirical shape statistics. *Foundations of Computational Mathematics*, 5(1):1–58.
- Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120.
- Den Herder, G., Van Isterdael, G., Beeckman, T., and De Smet, I. (2010). The roots of a new green revolution. *Trends in plant science*, 15(11):600–607.
- Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*, volume 4. John Wiley & Sons New York.
- Dubuisson, M.-P. and Jain, A. K. (1994). A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE.
- Eddy, W. F. (1977). A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, 3(4):398–403.

References

- Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Soc.
- Fasy, B. T., Kim, J., Lecci, F., Maria, C., included GUDHI is authored by Clement Maria, V. R. T., by Dmitriy Morozov, D., by Ulrich Bauer, P., Kerber, M., and Reininghaus, J. (2017). *TDA: Statistical Tools for Topological Data Analysis*. R package version 1.5.1.
- Fox, J. (2000). *Multiple and generalized nonparametric regression*. Number 131. Sage.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.
- Godsil, C. and Royle, G. F. (2013). *Algebraic graph theory*, volume 207. Springer Science & Business Media.
- Henrikson, J. (1999). Completeness and total boundedness of the hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1:69–80.
- Kendall, D. (1971). Seriation from abundance matrices. *Mathematics in the archaeological and historical sciences*, pages 215–232.
- Lobet, G., Koevoets, I., Tocquin, P., Périlleux, C., et al. (2016). Using a structural root system model for an in-depth assessment of root image analysis pipeline. *bioRxiv*.
- Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific reports*, 3:srep01236.
- Lynch, J. P. (2007). Roots of the second green revolution. *Australian Journal of Botany*, 55(5):493–512.
- Madsen, H. and Thyregod, P. (2011). *Introduction to General and Generalized Linear Models*. Chapman and Hall CRC.

References

- Mairhofer, S., Zappala, S., Tracy, S., Sturrock, C., Bennett, M. J., Mooney, S. J., and Pridmore, T. P. (2013). Recovering complete plant root system architectures from soil via x-ray mu-computed tomography. *Plant Methods*, 9(8):1–7.
- Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.
- Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.
- Reimann, M. W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., Dłotko, P., Levi, R., Hess, K., and Markram, H. (2017). Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in computational neuroscience*, 11:48.
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M. B. (2013). Package mass. *Cran R*.
- Seidel, R. (1995). The upper bound theorem for polytopes: an easy proof of its asymptotic version. *Computational Geometry*, 5(2):115–116.
- Topaz, C. M., Ziegelmeier, L., and Halverson, T. (2015). Topological data analysis of biological aggregation models. *PloS one*, 10(5):e0126383.