Algorithms for Predictive Plant Monitoring

Umut Çakır umut.cakir@boun.edu.tr

Molecular Biology and Genetics, 2014, Boğaziçi University

Muhammed Orhun Gale morhun@sabanciuniv.edu

Computer Science and Engineering, 2018, Sabancı University

Ozan Bicen
Electronics Engineering - Faculty of Engineering and Natural Sciences

Stuart J. Lucas
SU Nanotechnology Research and Application Center

Nihal Öztolan Erol
SU Nanotechnology Research and Application Center

Abstract

Pressure on food production increases day by day because of population growth and climate change. To overcome this pressure, increasing the yield in every arable land is essential. Since osmotic stress limits crop yield, in this project, selecting *T. Durum* wheats which have osmotic stress resistance by various statistical methods and Machine Learning algorithms is aimed. A data set which includes some morphological data of *T. Durum* wheats as root length, surface area, root volume, average root diameter and number of tips are examined. After these processes, we selected AS37, AS74 and AS131 as osmotic stress resistant samples.

Keywords: Osmotic stress tolerance, *T. durum* wheat relatives, Machine Learning, Cluster Analysis

1. Introduction

Population growth and climate change are some of the factors which increase the pressure on food production. The United Nations Food and Agricultural Organization (Fao & Dwfi, n.d.) (FAO) argues that if estimated population growth takes place, in 2050, humanity must increase the overall food production by 70%. Besides the required increase, climate change decreases crop yields and extends the problem. Wheat is one of the most-produced cereals. It is the biggest contributor of vegetable protein to human diet and after the rice, it is the most produced human food crop. (CORDIS | European Commission', n.d.) (EU-Genetic Markers) Therefore, sustainability of wheat production and increasing the its yield play a crucial role to compensating the pressure.

According to FAO, there are two main crop yield definition which one of them is aggressive whereas another one is less competitive. Aggressive one focuses on the ratio of number of seeds sown and crops harvested while passive one wants more yield per unit area. Also, there are limiting factors that affect the crop yield like water limitation. (FAO) Since limiting factors determine the bounds of the amount of the product, it can be argued that crop yield and effects of limiting factors are negatively correlated. Water limitation restrains the production because of the dehydration, which causes osmotic stress, it brings along. Thus, osmotic stress resistivity can be an important feature to raise the crop yield in areas which have limited water availability or getting dry because of the climate change. According to Öktem et al. (2006), osmotic stress tolerance is about the genes and the tolerance can be increased by genetic modification. (pp.194-203) Therefore, selecting plants whose genotypes provide higher osmotic stress tolerance and cultivating them can increase the crop yield in water limited areas. (Dhanda, Sethi, & Behl, 2004)

Agriculture 4.0 refers to a technological toolbox which consists of a set of advanced technologies such as Big Data, Internet of Things (IoT) and Artificial Intelligence and aims to deal with the challenges in agricultural production. (Lezoche, Panetto, Kacprzyk, Hernandez, & Alemany Díaz, 2020, pp. 3-8) Also, Machine Learning (ML) algorithms are used in different aspects of agri-food industry. (Miranda, Ponce, Molina, & Wright, 2019, pp.22-27) One of these applications is yield estimation. Chlingaryan (2018), successful yield predictions have been made recently by ML techniques such as Artificial Neural Networks, Support Vector Regression, M5-Prime Regression Trees and k-nearest neighbour. (Chlingaryan et al., 2018, pp.63-64) Another application is identification and classification of plants for stress phenotyping. Stress phenotyping aims to detect plants whose genotypes provide stress resistance and reducing the yield losses that stem from stress factors. For this purpose, unsupervised ML algorithms like k-Means, Hidden Markov Model (HMM), Boltzman Machine (BM) and Gaussian Mixture Model (GMM) can be used. (Singh, Ganapathysubramanian, Singh, & Sarkar, 2016, pp.110-114) In this context, these algorithms also might be used for detection of plants that have osmotic stress tolerance and after the detection of proper plants, their yields can be estimated.

In this project, our aim is to detect the *T. Durum* samples which demonstrate growing in osmotic stress conditions as if they are in well-watered conditions by using various statistical methods and unsupervised machine learning algorithms in order to indicate the accordance between these techniques and form an opinion about why these samples become distinct from other samples.

2. Method

1. 12 different Turkish *T. durum* cultivars were crossed with 19 tetraploid relatives, 14 of them were *T. diccoccoides*, 4 of them were *T. dicoccon* and 1 of them was *T. araraticum*. All samples were collected in Turkey and obtained from gene bank of the Field Crops Central Research Institute, Ankara.
2. Successful hybrids were back crossed with *T. durum* parent.
3. Samples were self-fertilized for 5 generations and then from last generation (F_5 generation), 500 samples were selected and planted in Ankara province.
4. From each F_6 generation, 10 seeds were washed with 0.5% sodium hypochlorite solution and sown on to a plate that contains 5% PEG-6000 for germination for 24 hours. 100 different F_6 hybrids

were selected. Hybrids, also called sample, were replicated several times (usually 4 or 5) and hybrids were identified with a unique identifier, which is "Sample ID".

5. 5 seeds were transferred to plate that contains 40% PEG-6000, which is used for osmotic stress analysis and referred as "osmotic stress (OS)". Another 5 seed were transferred to plate that contains only MS/MES medium, which is used for control group and referred as "well-watered (WW)". This step was called day 0.
6. Both groups of plates (OS and WW plates) were incubated at room temperature in the dark and different morphology parameters of the roots were recorded 3 and 5 days after previous step. These parameters are length of root in cm (L), surface area of root in cm^2 (SA), average diameter of root in mm (AvD), root volume in cm^3 (RV), and number of roots (Tips). Parameters and parentage information data of F_6 hybrids are available in (Budak, 2017). In the data, OS3 and OS5 columns stand for measurement in osmotic stress condition at 3rd and 5th day, respectively. Likewise, WW3 and WW5 columns stand for measurement in well-watered condition at 3rd and 5th day, respectively. Up to here, above steps were done by Budak (2017). Following steps were done by us.

   Root morphology analysis was done as follows:
7. If any replicate does not have appropriate measurement in any column (either OS3, OS5, WW3, or WW5), that replicate is discarded for further analysis.
8. Instead of dealing with multiple replicates for each sample, for each sample average value of each parameter were calculated, and average value was used for each sample. For instance, let's say, X sample has 3 replicates and length measurements for OS3 column are 8.00, 9.00, and 13.00. Instead of using each measurement, average value ((8.00+9.00+13.00)/2 = 10.00) is used for length measurement in OS3 column for sample X.
9. Basic statistics were conducted, and correlation matrices were prepared between each parameter. Significance of each parameter between either OS3/WW3 or OS5/WW5 was tested.
10. K-means and GMM clustering graphs were created in order to find OS samples that have resistance against osmotic stress, simply called candidates. For each day (3rd and 5th day) graphs were created and samples in OS were marked if these samples fall into WW group. Meanwhile, for each day 95% confidence intervals were calculated for each parameter in OS. For each parameter except tips candidates were determined if samples have higher value than upper bounds of confidence interval. For tips candidate samples were determined if tips has lower value than lower bounds of confidence interval.
11. Because some determined samples did not have appropriate values in some parameters according to WW even though they had proper value according to OS, in order to eliminate those samples and obtain clear result 95% confidence intervals also were calculated for each parameter in WW. For all parameters except tips lower bound of confidence interval in WW was used as a cut-off limit for samples in OS that were described in previous step. For tips higher bound of confidence interval in WW was used for cut-off limit. That is, for all parameters except tips samples in OS were marked only if higher value than lower bound of WW and for tips samples in OS were marked only if lower value than upper bound of WW.
12. Then, because lower tips samples did not have appropriate root parameters, in other words, lower tips samples were not good candidate according to root parameters, we eliminated lower tips samples and determined samples which have higher in length, surface area, and root volume as candidates.

3. Results

a) Data organization and cleaning

Because raw data were unclean and not suitable for most analysis, data were properly organized such that they can imported to many software/libraries such as pandas data frame or SPSS. Then, arithmetic averages of each sample for every parameter were calculated. Each sample has several replicates and instead of using each replicate for analysis averages were calculated for each replicate group and considered as a single measurement for one sample. For instance, Table 1 shows that AS165 has 3 replicates and instead of using each replicate value for AS165, average of these values was calculated, and average value which shown in last row in the table was used for further analysis.

Table 1: Root morphology data for AS165 at 3rd day

		Well-watered, 3 DAS				
Sample ID	**Rep ID**	**L (cm)**	**SA (cm2)**	**AvD (mm)**	**RV (cm3)**	**Tips**
AS165	a	19.670	3.928	0.636	0.062	21
AS165	b	11.238	2.817	0.798	0.056	7
AS165	c	15.022	3.301	0.699	0.058	30
Average		15.310	3.348	0.711	0.059	19.333

Besides that, if a replicate had no value (NaN) in any treatment (OS 3, OS5, WW3, WW5), that replicate was not considered and deleted in order to obtain clear conclusion. For this reason, 19 samples out of 100 were lost.

b) Correlation between root parameters

After organization and cleaning of the data, correlation matrices were created to check whether there is any correlation between any two parameters. Figure 2 shows correlation matrices between parameters in WW3, WW5, OS3, and OS5 separately. Almost all matrices are identical: In general, there is a strong positive correlation between length & surface area, reasonable positive correlation between length & root volume and length & tips, on the other hand, reasonable negative correlation between average diameter & length was observed.

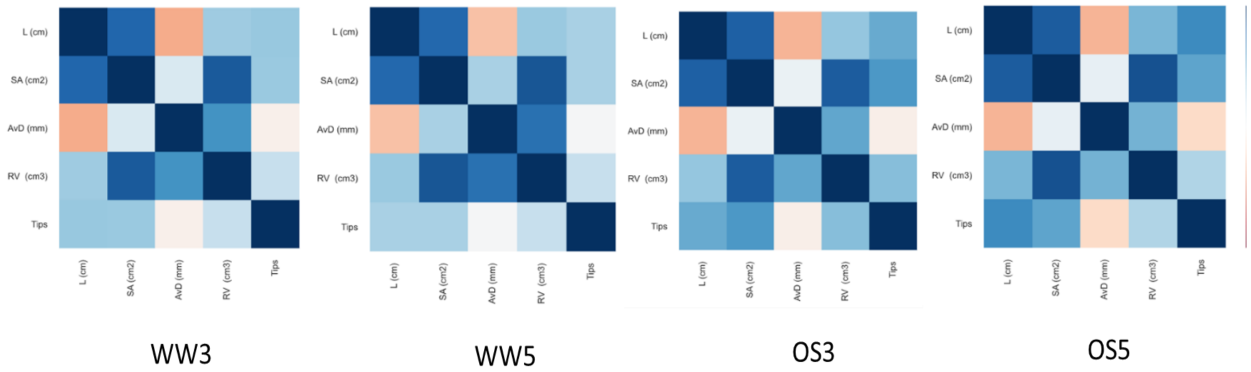


Figure 2: Correlation matrices between parameters in WW3, WW5, OS3, and OS5, respectively.

Individual Pearson correlation values are shown in appendix 1. Because all correlation matrices are almost identical, we can combine all four to one matrix to conclude which two parameters are significant or not. Table 2 shows correlations regardless of treatments and days. According to the table, there is a significant positive correlation between length & surface area, length & root volume, and length & tips, but a significant negative correlation between length & average diameter. That is, length is either positively or negatively correlated between any other four parameters. Likewise, surface area and root volume are positively correlated between any other four parameters. However, there is no significant correlation between average diameter & tips.

Table 2: Correlations Regardless of Treatment and Day

		Lcm	SAcm2	AvDmm	RVcm3	Tips
Lcm	Pearson Correlation	1	,907**	-,217**	,668**	,347**
	Sig. (2-tailed)		,000	,000	,000	,000
SAcm2	Pearson Correlation	,907**	1	,152**	,916**	,325**
	Sig. (2-tailed)	,000		,006	,000	,000
AvDmm	Pearson Correlation	-,217**	,152**	1	,485**	-,067
	Sig. (2-tailed)	,000	,006		,000	,231
RVcm3	Pearson Correlation	,668**	,916**	,485**	1	,251**
	Sig. (2-tailed)	,000	,000	,000		,000
Tips	Pearson Correlation	,347**	,325**	-,067	,251**	1
	Sig. (2-tailed)	,000	,000	,231	,000	

**. Correlation is significant at the 0.01 level (2-tailed).

c) Comparing Days in the Same Treatment

Treatments on different days in respect to each parameter were compared. Table 3 and Table 4 shows median values of length, surface area, average diameter, root volume, and tips at OS3 & OS5 and WW3 & WW5, respectively. Significant parameters at different days were marked with "***". According to Table 3 and Table 4, length, surface area, and root volume are significant on both treatments at different days. Average diameter is not significant on both treatments at different days (p=0.75 for OS and p=0.37 for WW). On the other hand, tips is significant for OS, but not significant for WW (p=0.38).

Table 3: Median value of parameters at OS3 and OS5; *** : $p < 0.001$

Median	**Lcm**	**SAcm2**	**AvDmm**	**RVcm3**	**Tips**
OS3	10.34 ***	2.02 ***	0.65	0.03 ***	12.20 ***
OS5	16.82 ***	3.57 ***	0.66	0.06 ***	21.25 ***

Table 4: Median value of parameters at WW3 and WW5; *** : $p < 0.001$

Median	**Lcm**	**SAcm2**	**AvDmm**	**RVcm3**	**Tips**
WW3	15.67 ***	3.53 ***	0.68	0.06 ***	15.00
WW5	21.84 ***	4.95 ***	0.71	0.09 ***	13.67

d) Comparing Treatments in the Same Day

Treatments in the same days in respect to each parameter were compared. Table 5 and Table 6 shows median values of length, surface area, average diameter, root volume, and tips at OS3 & WW3 and OS5 & WW5, respectively. Significant parameters in the same day were marked with "***". According to Table 5 and Table 6, at both days there is a significant difference between treatments according to length, surface area, and root volume parameters, but not average diameter. On the other hand, at 3rd day tips is not significant between treatments (p=0.06), but at 5th day tips is significant. Because tips has some extreme values whose may come from wrong measurements, all tips which are bigger than 50 equalize to 50 and again at 3rd day tips is not significant (p = 0.29). Moreover, average diameter is not informative because there is no significant difference between either treatments in the same day or days in the same treatment (see "Comparing Days in the Same Treatment" section), we will not use average diameter in the next steps.

Table 5: Median value of parameters at OS3 and OS5; *** : $p < 0.001$

Median	Lcm	SAcm2	AvDmm	RVcm3	Tips
OS3	10.34 ***	2.02 ***	0.65	0.03 ***	12.20
WW3	15.67 ***	3.53 ***	0.68	0.06 ***	15.00

Table 6: Median value of parameters at OS3 and OS5; *** : $p < 0.001$

Median	Lcm	SAcm2	AvDmm	RVcm3	Tips
OS5	16.82 ***	3.57 ***	0.66	0.06 ***	21.25 ***
WW5	21.84 ***	4.95 ***	0.71	0.09 ***	13.67 ***

e) Selecting Samples which Have Osmotic Stress Resistance

95% Confidence intervals (CI) for each parameter were calculated for OS3, OS5, WW3, and WW5 for each day separately in order to determine which samples have resistance against osmotic stress. Then, OS samples whose any parameter did not fall into OS CI and fall into WW CI were determined. Because length, surface area, and root volume are higher in WW than OS condition and tips is lower in WW than OS condition, (see Table 3 to Table 6), while WW lower bound limit was used for length, surface area, and root, WW higher bound limit was used for tips. Average diameter was not used because it was not informative. We assumed that weight of length, surface area, root volume, and tips were equal to determine osmotic stress resistance samples.

Table 7: Upper/lower bound limits of parameter at 3rd and 5th day for selecting osmotic stress resistance sample

Upper/Lower Limit	Lcm	SAcm2	RVcm3	Tips
3rd Day	15	3.24	0.05	12.8
5th Day	21.2	4.7	0.08	17.9

According to Table 7 at 3rd day, we did not consider tips because at 3rd day tips was not significant between OS and WW condition (see Table 5). If we want to determine samples according to each parameter separately, we came to the following result shown in Table 8. There is no sample that is present in all three parameters. Only AS189 and AS195 (bold) are present in two parameters out of three.

Table 8: Candidates which have resistance according to each parameter separately at 3rd day

L	SA	RV
AS91	**AS189**	AS7
AS123	**AS195**	AS53
AS165		AS57
AS182		AS64
AS188		**AS189**
AS191		AS498
AS195		
AS205		
AS493		

According to Table 7 at 5th day, in that case we used tips because at 5th day tips is significant between OS and WW condition (see Table 6). If we want to determine samples according to each parameter separately, we came to the following result shown in Table 9. There is no sample that is present in all four parameters. Only AS37 is present in three parameters out of four. There are many candidates according to tips such that we could not fit all of them in one column.

Table 9: Candidates which have resistance according to each parameter separately at 5th day

L	SA	RV	Tips	Tips (continue)	Tips (continue)
AS36	**AS37**	**AS37**	AS12	AS142	AS252
AS37	AS74	AS57	AS53	AS145	AS455
AS123	AS131	AS74	AS91	AS147	AS466
AS124		AS122	AS101	AS152	AS467
AS131		AS498	AS122	AS154	AS472
AS188			AS128	AS159	AS478
AS191			AS134	AS169	AS479
AS195			AS136	AS175	AS482
AS261			AS137	AS186	AS494
			AS138	AS197	AS495
			AS139	AS206	

f) Tips effect on root development

In order to see tips effect on root development only a range of tips was selected at 5th day. We determine samples that have lower number of tips, that is, samples whose tips are smaller than 17.9 (WW higher bound limit) to see effect on length, surface area, and root volume. If plant has small number of tips, root length and surface area always become smaller. In other words, if we want to determine samples whose length and surface area is bigger (close to WW condition), tips should be higher. Likewise, it is almost true for root volume; if plant has small number of tips, root volume always become smaller except AS122 sample. (AS122 tips and root volume are 15.33 and 0.105667, respectively.) On the other hand, if plant has big number of tips, we could not come to a general conclusion like small number of tips. However, if a plant has big number of tips, there are some samples that has bigger length, surface area, and root volume or any combination of two out of three parameter. AS37 sample has high number of tips, so all three parameters are higher. AS74 and AS131 samples has also high number of tips, thus for AS74 surface area and root volume are bigger and for AS131 length and surface area are bigger. Their values under osmotic stress at 5th day are shown in Table 10.

Table 10: Candidates whose tips are higher

Sample ID	L (cm)	SA (cm2)	AvD (mm)	RV (cm3)	Tips
AS37	22.53612	5.1372	0.71842	0.0948	35.6
AS74	20.96398	4.92894	0.72938	0.1002	25.4
AS131	25.59553	4.907775	0.61125	0.07675	28.25

g) Cluster Analysis

Gaussian Mixture Model (GMM) was used as a ML technique is in order to detect candidate samples that were selected as osmotic stress resistant by statistical methods. From the data set, significantly correlated trait doubles were selected. Data of these doubles were fitted to the GMM algorithm in MATLAB. To be able to say a sample can be a stress resistant candidate, we set the criteria as being marked as WW sample while the sample is an OS sample in the model which was created by GMM. Initially, we used an algorithm (see Appendix 2.1) to find the models of different covariance structure options. By using this algorithm 4 different possible models were made. Then, a different algorithm (see Appendix 2.2) was used to model posterior probabilities of each selected double. While 5 out of 6 posterior probability models look like unshared covariances and full sigma, Surface Area vs. Number of Tips model look like shared covariances and full sigma (see Appendix 2.8). In these models, AS37 was marked as WW in 4 out of 6 models, AS74 was marked as WW in 3 out of 6 models and AS131 was marked as WW in 3 out of 6 models (see Appendix 2.9-26). These samples were marked as WW in Length vs. Surface Area, Length vs. Number of Tips, Root Volume vs. Number of Tips models mutually. In Surface Area vs. Number of Tips model, AS37 was marked as WW while AS74 and AS131 were marked as OS however, their possibilities to be a member of WW group were given as around 40-50% by the algorithm (See Appendix 2.14-20-26). Models that these samples were not marked as WW, Length vs. Root Volume and Surface Area vs. Root Volume, were classified as two totally differentiated groups because of the extreme Root Volume values (See Appendix 2.9-13-15-19-21-25).

4. Discussion

In this experiment, our aim was by using samples, which are F_6 generation of multiple crosses between *T.durum* parent with wild tetraploid parent (*T. diccoccoides*, *T. dicoccon* or *T.araraticum*), to determine osmotic stress resistance samples according to length, surface area, average diameter, root volume, and tips parameter.

100 samples with root parameters under osmotic stress and well water condition at 3rd and 5th day were used to determine samples which have resistance against osmotic stress. These root parameters are length, surface area, average diameter, root volume, and tips. In order to reduce variability and increase significance of the result the confidence levels of each sample was replicated generally 4 to 5 times and arithmetic averages of the parameters were used for the analysis. Some replicates did not have value in any treatment/day, so these replicates were dropped for the analysis. For instance, if a replicate has values in WW3, WW5, and OS3 but not OS5, that replicate was dropped, so 19 samples were dropped, and 81 samples were remained.

According to correlations regardless of treatment and day (Table 2), while positive correlation is observed in length & surface area, length & root volume, length & tips, surface area & average diameter, surface area & root volume, surface area & tips, average diameter & root volume, and root volume & tips, negative correlation is observed in length & average diameter. There is no significant correlation in average diameter & tips. Mentioned significances are also present for WW3 and WW5. However, surface area & average diameter is not significant for OS3. Besides, length & root volume, surface area & average diameter, and root volume & tips are not significant for OS5. Even though correlation matrices are almost identical, there is significant different between treatments (WW or OS) even at different days (3rd or 5th day). In WW condition, if length increases, root volume increases; however, in OS condition, if length increases root volume does not increase. Similarly, in OS condition when surface area and root volume increases, average diameter and tips do not change, respectively. The reason may be that water is limited and a plant directs resource to one parameter to survive. For instance, the plant does not want to increase root length and root volume at the same time if water is limited in the environment. Besides that, order of parameter here is not important, that is, we do not assume that the first parameter affects second parameter or vice versa. For example, we do not assume that length increases root volume or not. We are just dealing with positive or negative correlations between any two parameters.

When different days in the same treatment were compared (OS3 & OS5 and WW3 & WW5), average diameter is not significant in both treatments. That is, average diameter does not increase in time, but it does not always mean that total diameter is changing because average diameter is sum of diameters and divided by tips. In WW condition, total root diameter does not change significantly in time, but in OS condition, total root diameter increases significantly in time. Then, in WW condition tips is not significant at different days, but in OS condition tips is significant at different days, probably because in WW condition plants have already access to water, thus they do not need to increase number of root to access water, however, in OS condition plant have access to limited water (faced to osmotic stress), so they increase number of root to access water in time.

When treatments in the same day were compared (OS3 & WW3 and OS5 & WW5), average diameter is again not significant. At 5th day tips is significant, but at 3rd day tips is not significant. Why significance is observed at 5th day not 3rd day is that probably plants growth its root "normally" like in WW condition in the beginning and if they had faced to osmotic stress, they redirect metabolic effort to root development but it takes some time. 3 day are not enough, but 5 days are enough to redirect metabolic effort to root development to increase number of roots (tips).

Samples in OS condition were marked if parameter of interest is out of confidence interval in OS and in confidence interval in WW. In other words, if a sample in OS condition is similar to WW condition according to any parameter, these samples were selected as a candidate that has resistance against osmotic stress. We believe that 5th day is more informative in determining candidate samples because at 3rd day tips is not significant between OS3 & WW3. In addition, if tips is lower, other parameters (length, surface area, and root volume) become lower: Samples with lower tips typically do not have resistance against osmotic stress. Because of this finding, parameters at 5th day may be more informative to determine candidate samples. According to 5th day, samples with high tips in OS condition have a chance to resistance against osmotic stress if other parameters are similar to WW condition. These candidate samples are AS37, AS74, and AS131; their tips are high; thus, their length, surface area, and root volume are similar to WW condition.

In our analysis, we used arithmetic means of replicates for each sample. If any parameter has an extreme value, may be outlier, that value may interfere with our result and cause to wrong conclusion. We did not eliminate these types of extreme values, so our conclusion may be wrong. To select candidate samples, firstly samples with high tips were determined and then if other parameters were close to WW condition, these samples were selected for candidate samples. Because our first criterion was tips and if tips have an extreme value, we may come to wrong or incomplete list of candidates. Table 11 shows parameters of replications for AS37, AS74, and AS131. For instance, replication a in AS37 has 64 tips, which may be outlier and because of 64 average of tips may be a misleading arithmetic average. Very smaller tips such as 14 in replication b in AS37 also may cause to misleading arithmetic average. Shortly, outlier at one end (very high or very low) skews the arithmetic average and cause to misleading result. Instead of using directly arithmetic means, we should either eliminate these outliers before analysis or use other method such as median instead of arithmetic means.

Table 11: Parameters of replications for candidate samples in OS condition

	Replication	**L (cm)**	**SA (cm2)**	**AvD (mm)**	**RV (cm3)**	**Tips**
AS37	a	27.0676	6.6261	0.7792	0.129	64
AS37	b	18.5438	3.1047	0.5329	0.041	14
AS37	c	22.8499	5.7807	0.8053	0.116	34
AS37	d	19.7783	4.8724	0.7842	0.096	44
AS37	e	24.441	5.3021	0.6905	0.092	22
Average		22.53612	5.1372	0.71842	0.0948	35.6
AS74	a	23.6741	8.5303	1.1469	0.245	22
AS74	b	17.8843	3.5306	0.6284	0.055	14
AS74	c	21.4914	4.343	0.6432	0.07	40
AS74	d	25.2991	5.4003	0.6795	0.092	32
AS74	e	16.471	2.8405	0.5489	0.039	19
Average		20.96398	4.92894	0.72938	0.1002	25.4
AS131	a	30.7651	4.7037	0.4867	0.057	38
AS131	b	30.9692	7.0977	0.7295	0.129	45
AS131	c	20.8763	3.7259	0.5681	0.053	16
AS131	d	19.7715	4.1038	0.6607	0.068	14
Average		25.59553	4.907775	0.61125	0.07675	28.25

Additionally, 19 samples were eliminated in data cleaning because these 19 samples do not have any measurement in at least one treatment/day. For instance, if a sample does not have measurement only in OS3, that sample was eliminated before analysis. To recover these 19 samples as much as possible we can use a different strategy. At least if any of these 19 samples have measurement in OS5 and WW5, we can take them into consideration because we found that 5th day is more informative than 3rd day for determining candidate samples.

We determined that AS37, AS74, and AS131 have osmotic stress resistance. In order to find biological reason(s) why these samples have osmotic stress resistance, transcriptome, proteome, single nucleotide polymorphism or mutation analysis can be conducted. By these techniques which genes/proteins have a role in osmotic stress resistance can be found.

Besides the issues about sample set, it was observed that GMM algorithm marked the selected samples as statistical methods remark in the cases which were properly modelled. Adjustments in the data set may fix problematic models too and algorithm may mark samples completely as expected. For future works with this data set, biologically irrational samples must be removed in accordance with the biological information. It may make ML algorithms work more efficient. After the selection stage is completed, selected samples' yield can be predicted by using ML algorithms in OS and WW conditions.

5. References

Budak, H. (2017). High-throughput SNP genotyping of modern and wild emmer wheat for yield and root morphology using a combined association and linkage analysis. *Functional & Integrative Genomics*, *17*. https://doi.org/10.1007/s10142-017-0563-y

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018, August 1). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*. Elsevier B.V. https://doi.org/10.1016/j.compag.2018.05.012

Dhanda, S. S., Sethi, G. S., & Behl, R. K. (2004). Indices of Drought Tolerance in Wheat Genotypes at Early Stages of Plant Growth. *Journal of Agronomy and Crop Science*, *190*(1), 6–12. https://doi.org/10.1111/j.1439-037X.2004.00592.x

Fao, & Dwfi. (n.d.). *Yield gap analysis of field crops: Methods and case studies*. Retrieved from www.fao.org/publications

Genetic markers signal increased crop productivity potential | News | CORDIS | European Commission. (n.d.). Retrieved 25 August 2020, from https://cordis.europa.eu/article/id/118823-genetic-markers-signal-increased-crop-productivity-potential

Lezoche, M., Panetto, H., Kacprzyk, J., Hernandez, J. E., & Alemany Díaz, M. M. E. (2020, May 1). Agri-food 4.0: A survey of the Supply Chains and Technologies for the Future Agriculture. *Computers in Industry*. Elsevier B.V. https://doi.org/10.1016/j.compind.2020.103187

Miranda, J., Ponce, P., Molina, A., & Wright, P. (2019). Sensing, smart and sustainable technologies for Agri-Food 4.0. *Computers in Industry*, *108*, 21–36. https://doi.org/10.1016/j.compind.2019.02.002

Oktem, H., Eyidoøan, F., Selçuk, F., da Silva, J., & Yücel, M. (2006). Osmotic Stress Tolerance in Plants: Transgenic Strategies (pp. 194–208).

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016, February 1). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*. Elsevier Ltd. https://doi.org/10.1016/j.tplants.2015.10.015

6. Appendices

1. Correlation matrices between the parameters:

Table 1.1: Correlations for WW3

		Lcm	SAcm2	AvDmm	RVcm3	Tips
Lcm	Pearson Correlation	1	,784**	-,252*	,404**	,364**
	Sig. (2-tailed)		,000	,023	,000	,001
SAcm2	Pearson Correlation	,784**	1	,367**	,881**	,480**
	Sig. (2-tailed)	,000		,001	,000	,000
AvDmm	Pearson Correlation	-,252*	,367**	1	,737**	,143
	Sig. (2-tailed)	,023	,001		,000	,202
RVcm3	Pearson Correlation	,404**	,881**	,737**	1	,412**
	Sig. (2-tailed)	,000	,000	,000		,000
Tips	Pearson Correlation	,364**	,480**	,143	,412**	1
	Sig. (2-tailed)	,001	,000	,202	,000	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 1.2: Correlations for WW5

		Lcm	SAcm2	AvDmm	RVcm3	Tips
Lcm	Pearson Correlation	1	,774**	-,377**	,323**	,457**
	Sig. (2-tailed)		,000	,001	,003	,000
SAcm2	Pearson Correlation	,774**	1	,268*	,842**	,466**
	Sig. (2-tailed)	,000		,016	,000	,000
AvDmm	Pearson Correlation	-,377**	,268*	1	,729**	-,019
	Sig. (2-tailed)	,001	,016		,000	,867
RVcm3	Pearson Correlation	,323**	,842**	,729**	1	,329**
	Sig. (2-tailed)	,003	,000	,000		,003
Tips	Pearson Correlation	,457**	,466**	-,019	,329**	1
	Sig. (2-tailed)	,000	,000	,867	,003	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 1.3: Correlations for OS3

		Lcm	SAcm2	AvDmm	RVcm3	Tips
Lcm	Pearson Correlation	1	,800**	-,492**	,246*	,365**
	Sig. (2-tailed)		,000	,000	,027	,001
SAcm2	Pearson Correlation	,800**	1	,056	,767**	,491**
	Sig. (2-tailed)	,000		,622	,000	,000
AvDmm	Pearson Correlation	-,492**	,056	1	,637**	-,017
	Sig. (2-tailed)	,000	,622		,000	,880
RVcm3	Pearson Correlation	,246*	,767**	,637**	1	,370**
	Sig. (2-tailed)	,027	,000	,000		,001
Tips	Pearson Correlation	,365**	,491**	-,017	,370**	1
	Sig. (2-tailed)	,001	,000	,880	,001	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

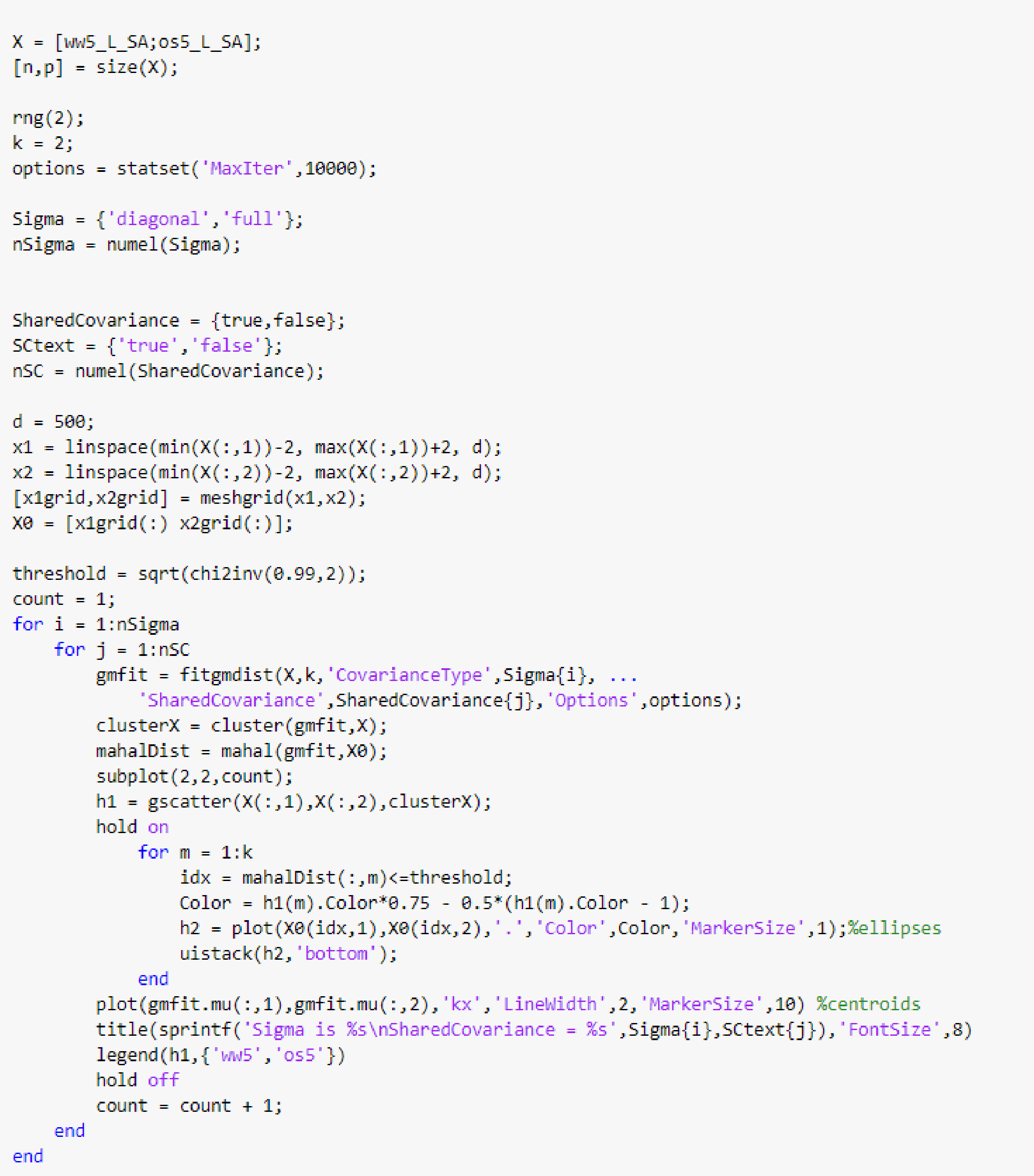
Table 1.4: Correlations for OS5

		Lcm	SAcm2	AvDmm	RVcm3	Tips
Lcm	Pearson Correlation	1	,754**	-,590**	,152	,625**
	Sig. (2-tailed)		,000	,000	,176	,000
SAcm2	Pearson Correlation	,754**	1	-,092	,757**	,549**
	Sig. (2-tailed)	,000		,413	,000	,000
AvDmm	Pearson Correlation	-,590**	-,092	1	,496**	-,307**
	Sig. (2-tailed)	,000	,413		,000	,005
RVcm3	Pearson Correlation	,152	,757**	,496**	1	,199
	Sig. (2-tailed)	,176	,000	,000		,074
Tips	Pearson Correlation	,625**	,549**	-,307**	,199	1
	Sig. (2-tailed)	,000	,000	,005	,074	

**. Correlation is significant at the 0.01 level (2-tailed).

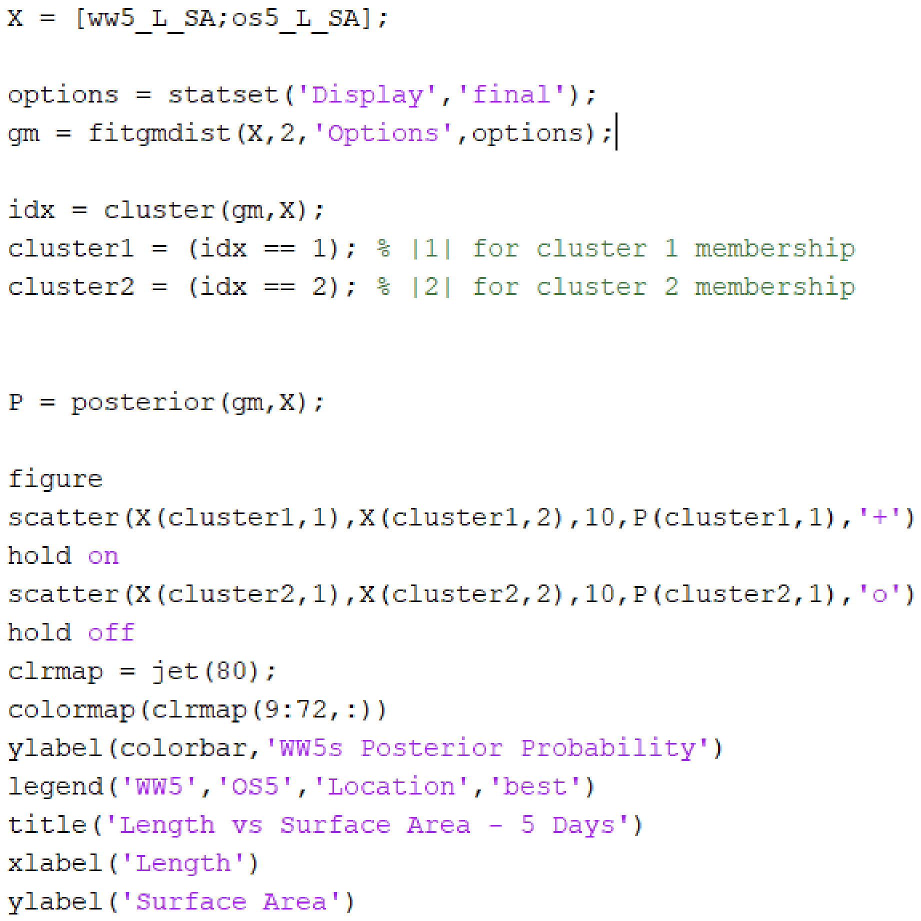
2. Cluster related documents:

Appendix 2.1



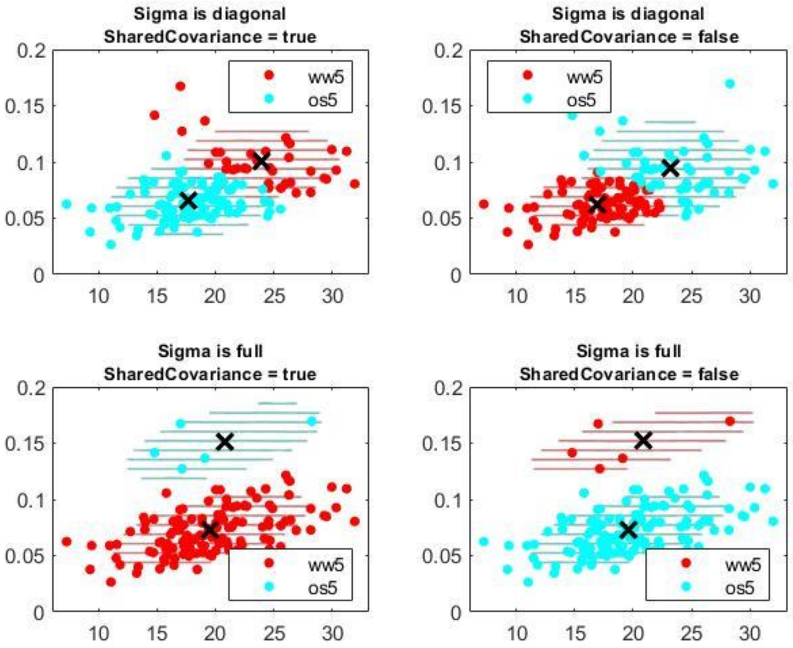
```
X = [ww5_L_SA;os5_L_SA];
[n,p] = size(X);

rng(2);
k = 2;
options = statset('MaxIter',10000);

Sigma = {'diagonal','full'};
nSigma = numel(Sigma);


SharedCovariance = {true,false};
SCtext = {'true','false'};
nSC = numel(SharedCovariance);

d = 500;
x1 = linspace(min(X(:,1))-2, max(X(:,1))+2, d);
x2 = linspace(min(X(:,2))-2, max(X(:,2))+2, d);
[x1grid,x2grid] = meshgrid(x1,x2);
X0 = [x1grid(:) x2grid(:)];

threshold = sqrt(chi2inv(0.99,2));
count = 1;
for i = 1:nSigma
    for j = 1:nSC
        gmfit = fitgmdist(X,k,'CovarianceType',Sigma{i}, ...
            'SharedCovariance',SharedCovariance{j},'Options',options);
        clusterX = cluster(gmfit,X);
        mahalDist = mahal(gmfit,X0);
        subplot(2,2,count);
        h1 = gscatter(X(:,1),X(:,2),clusterX);
        hold on
            for m = 1:k
                idx = mahalDist(:,m)<=threshold;
                Color = h1(m).Color*0.75 - 0.5*(h1(m).Color - 1);
                h2 = plot(X0(idx,1),X0(idx,2),'.','Color',Color,'MarkerSize',1);%ellipses
                uistack(h2,'bottom');
            end
        plot(gmfit.mu(:,1),gmfit.mu(:,2),'kx','LineWidth',2,'MarkerSize',10) %centroids
        title(sprintf('Sigma is %s\nSharedCovariance = %s',Sigma{i},SCtext{j}),'FontSize',8)
        legend(h1,{'ww5','os5'})
        hold off
        count = count + 1;
    end
end
```

Appendix 2.2

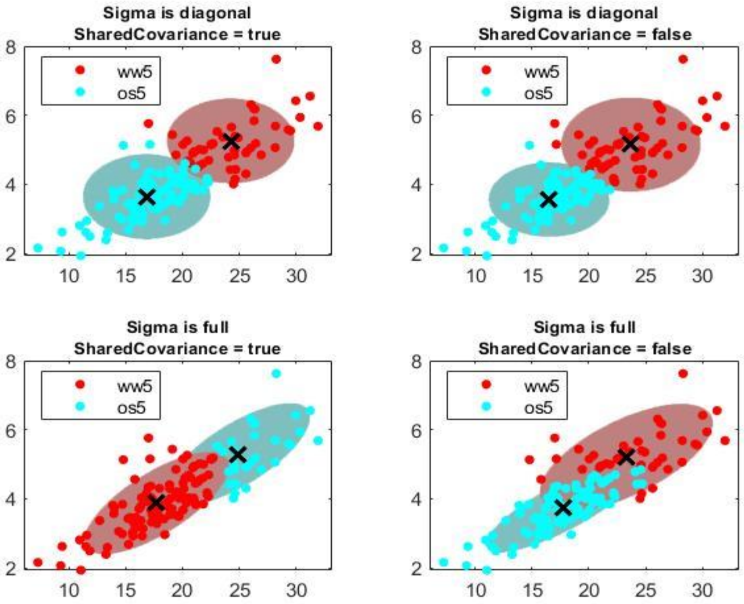


```
X = [ww5_L_SA;os5_L_SA];

options = statset('Display','final');
gm = fitgmdist(X,2,'Options',options);

idx = cluster(gm,X);
cluster1 = (idx == 1); % |1| for cluster 1 membership
cluster2 = (idx == 2); % |2| for cluster 2 membership


P = posterior(gm,X);

figure
scatter(X(cluster1,1),X(cluster1,2),10,P(cluster1,1),'+')
hold on
scatter(X(cluster2,1),X(cluster2,2),10,P(cluster2,1),'o')
hold off
clrmap = jet(80);
colormap(clrmap(9:72,:))
ylabel(colorbar,'WW5s Posterior Probability')
legend('WW5','OS5','Location','best')
title('Length vs Surface Area - 5 Days')
xlabel('Length')
ylabel('Surface Area')
```

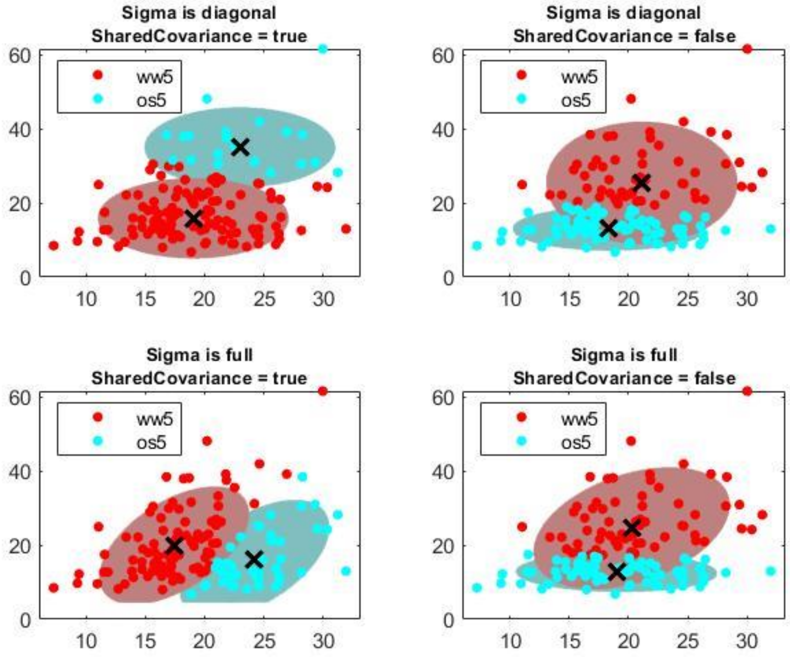
Appendix 2.3: Length vs. Root Volume



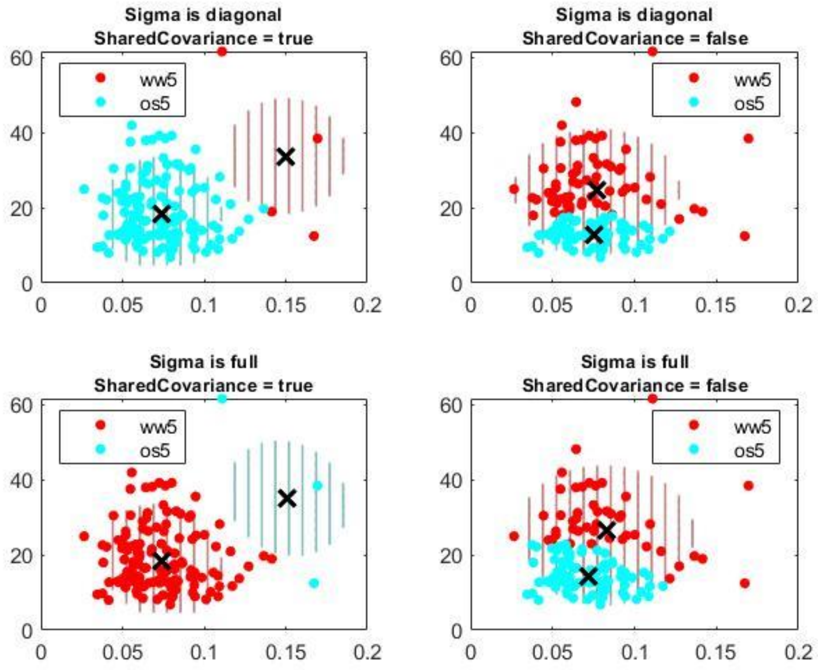
Appendix 2.4: Length vs. Surface Area



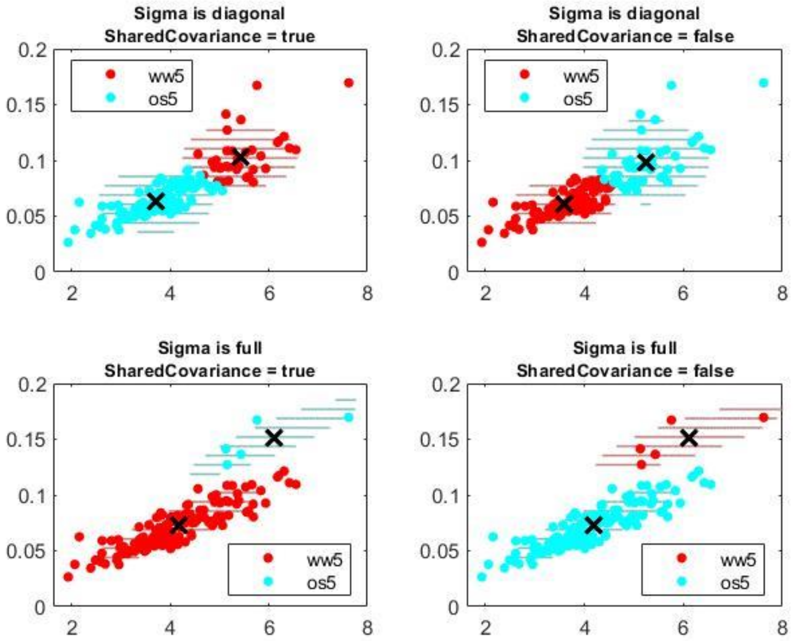
Appendix 2.5: Length vs. Number of Tips



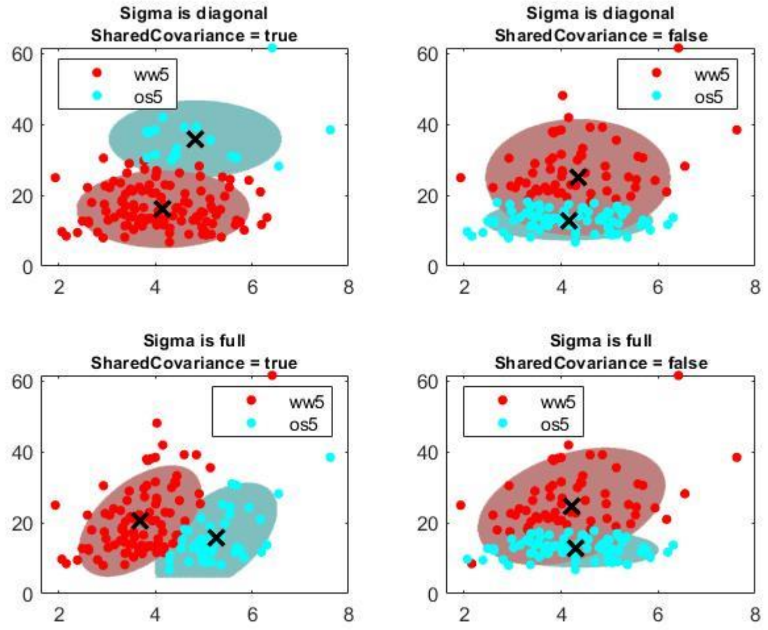
Appendix 2.6: Root Volume vs. Number of Tips



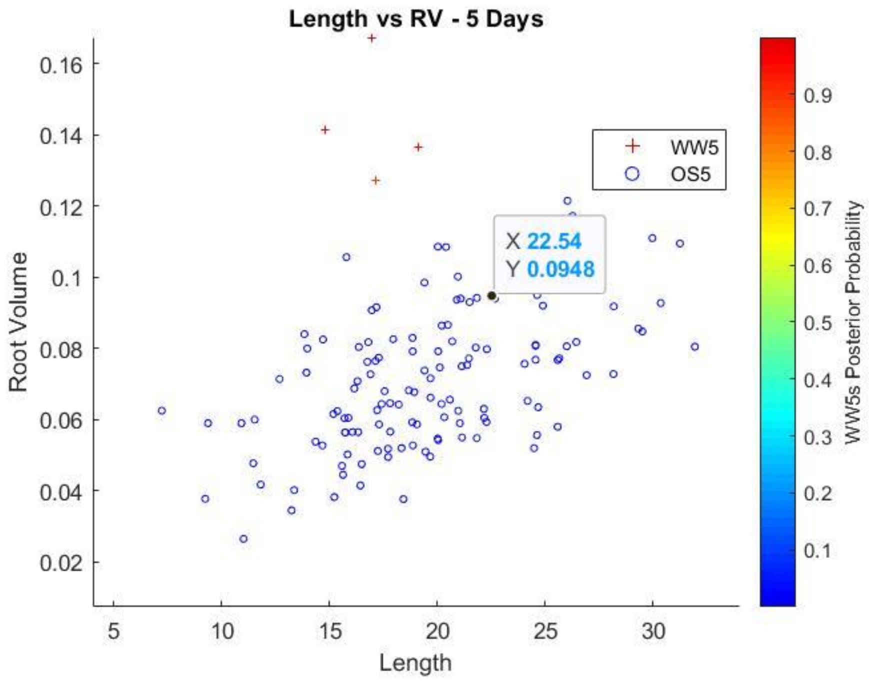
Appendix 2.7: Surface Area vs. Root Volume



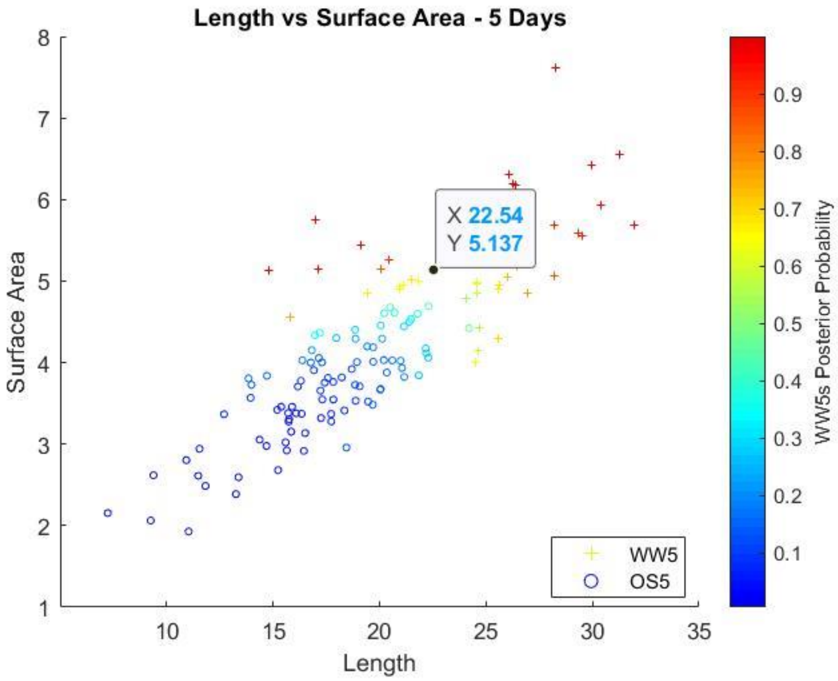
Appendix 2.8: Surface Area vs. Number of Tips



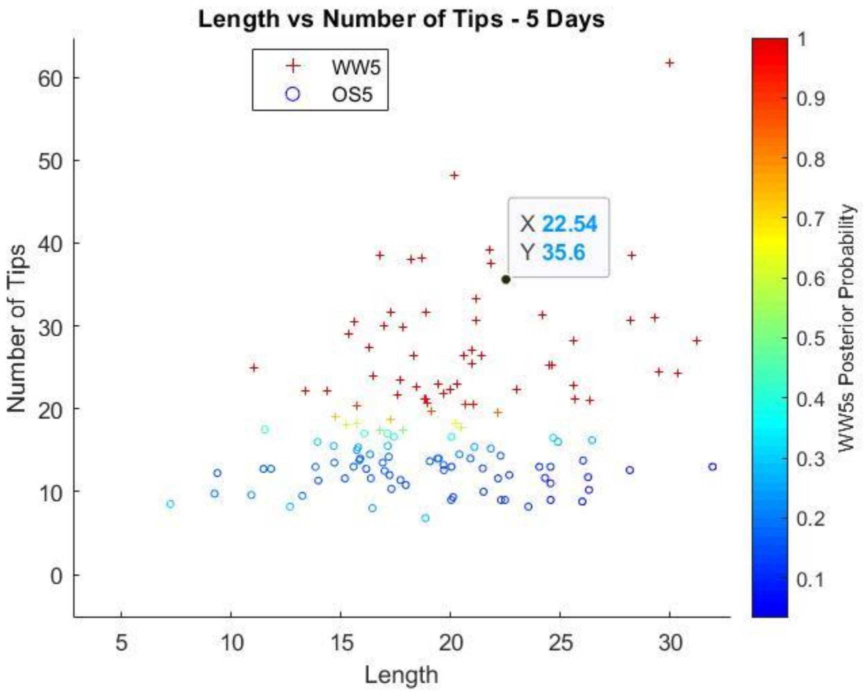
Appendix 2.9: AS37



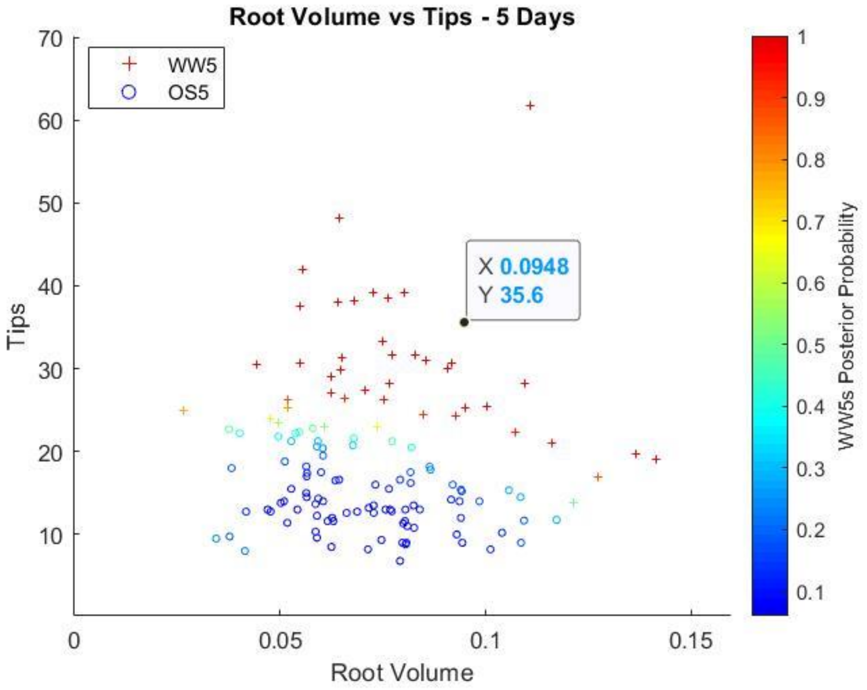
Appendix 2.10: AS37



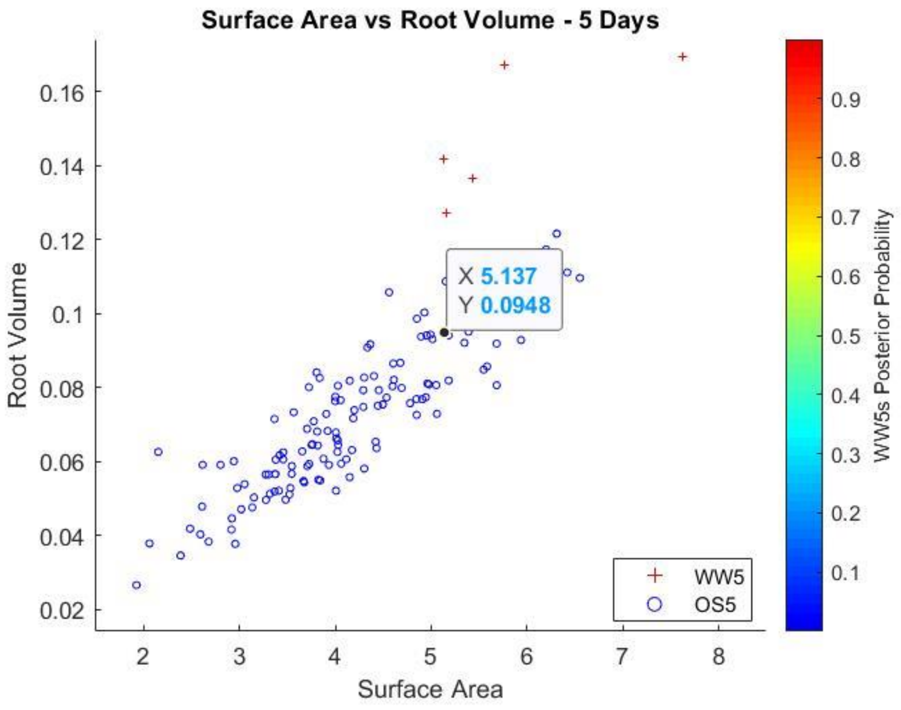
Appendix 2.11: AS37



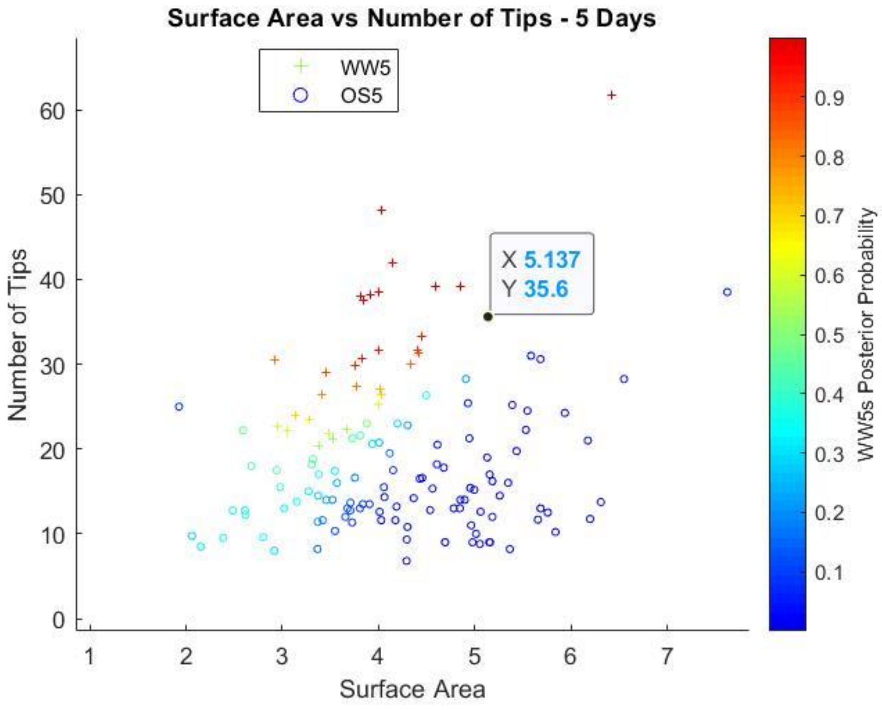
Appendix 2.12: AS37



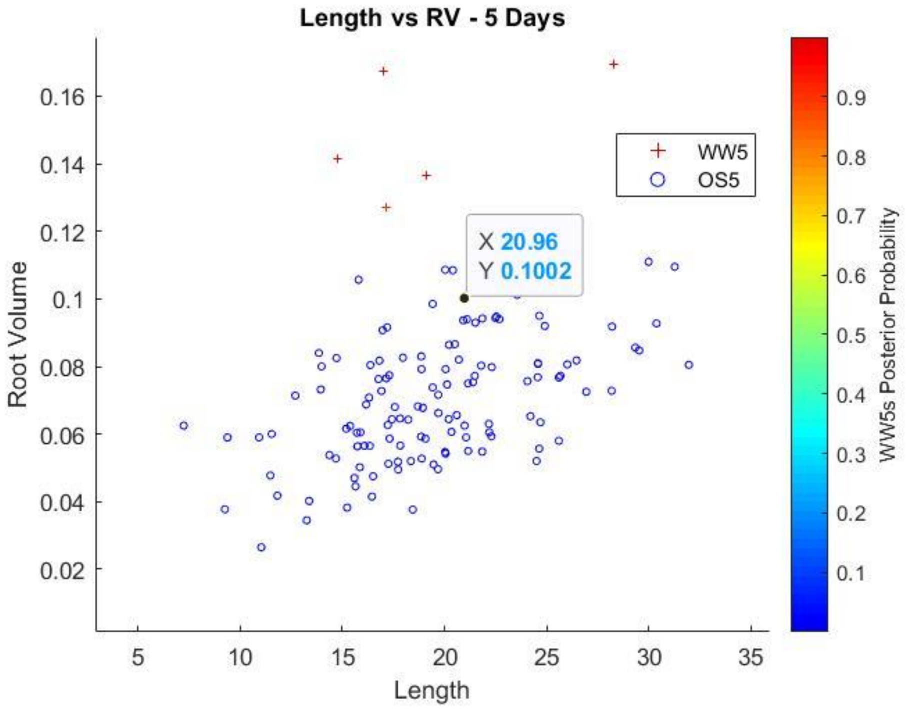
Appendix 2.13: AS37



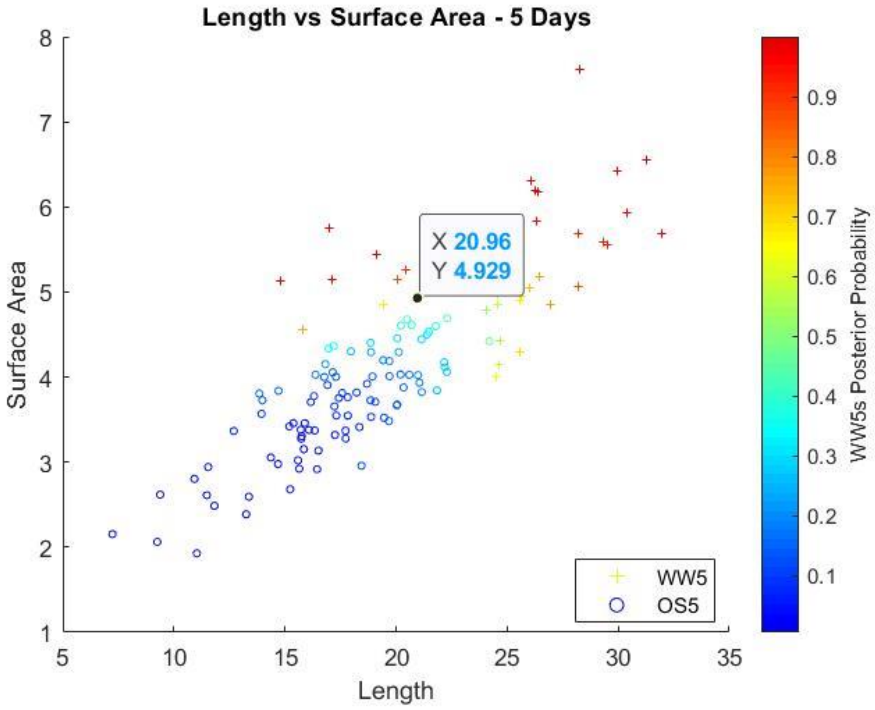
Appendix 2.14: AS37



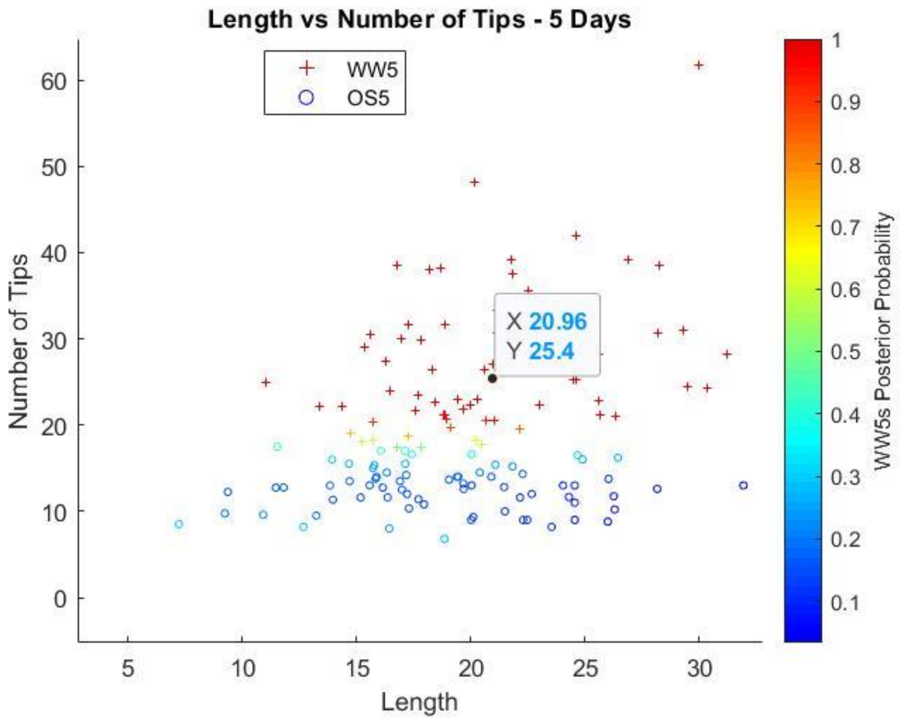
Appendix 2.15: AS74



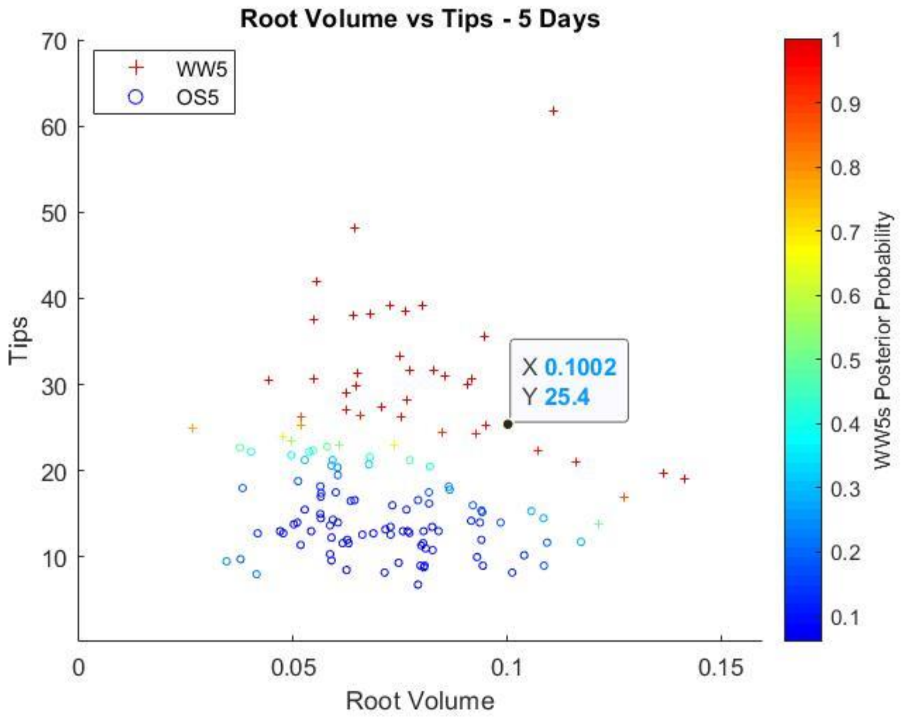
Appendix 2.16: AS74



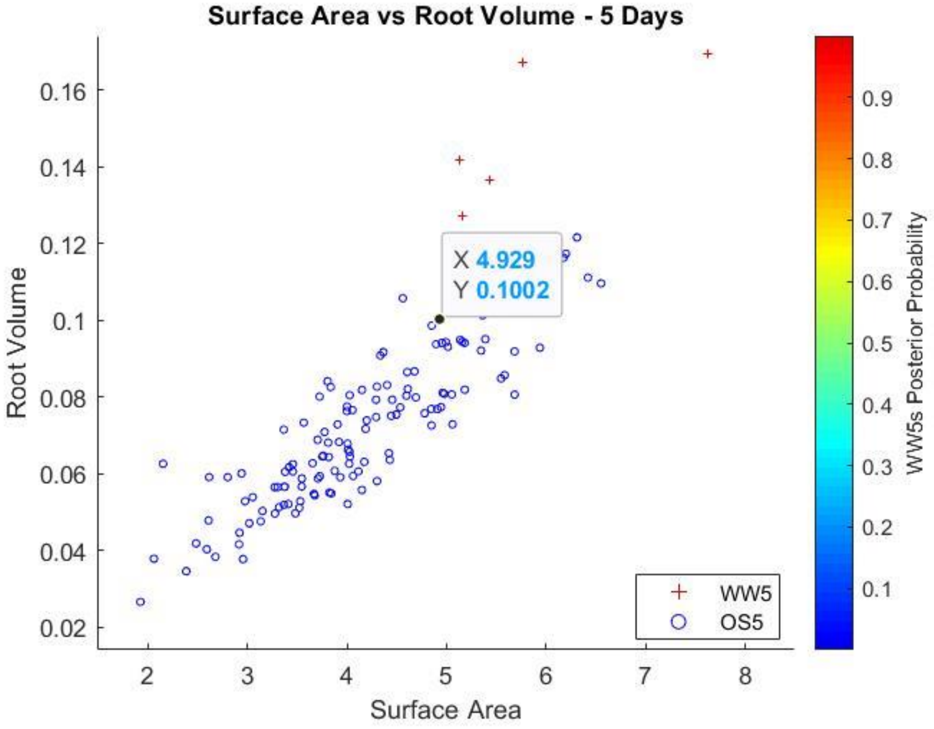
Appendix 2.17: AS74



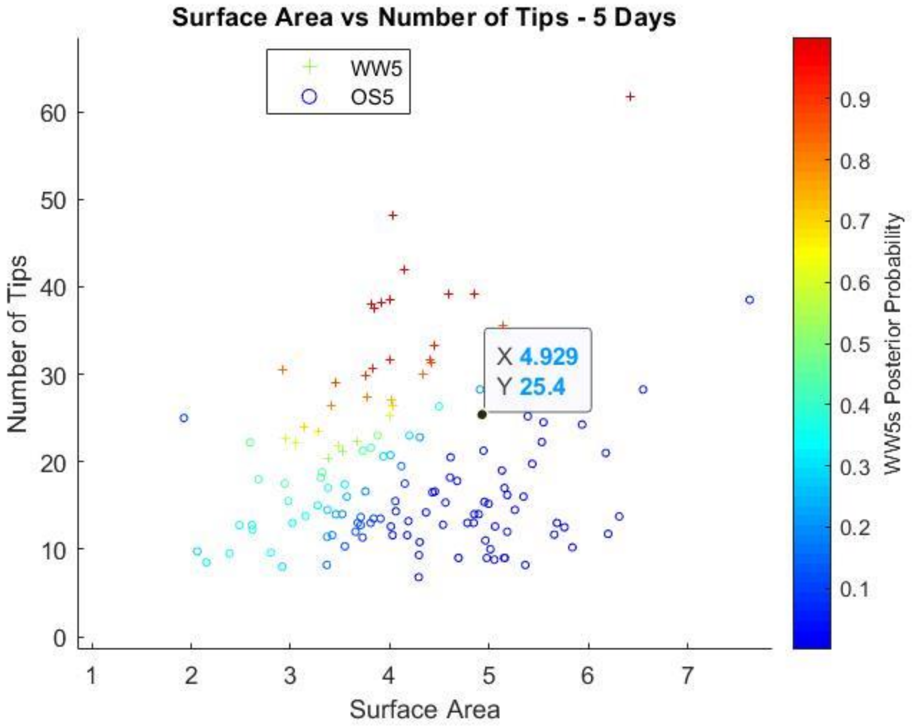
Appendix 2.18: AS74



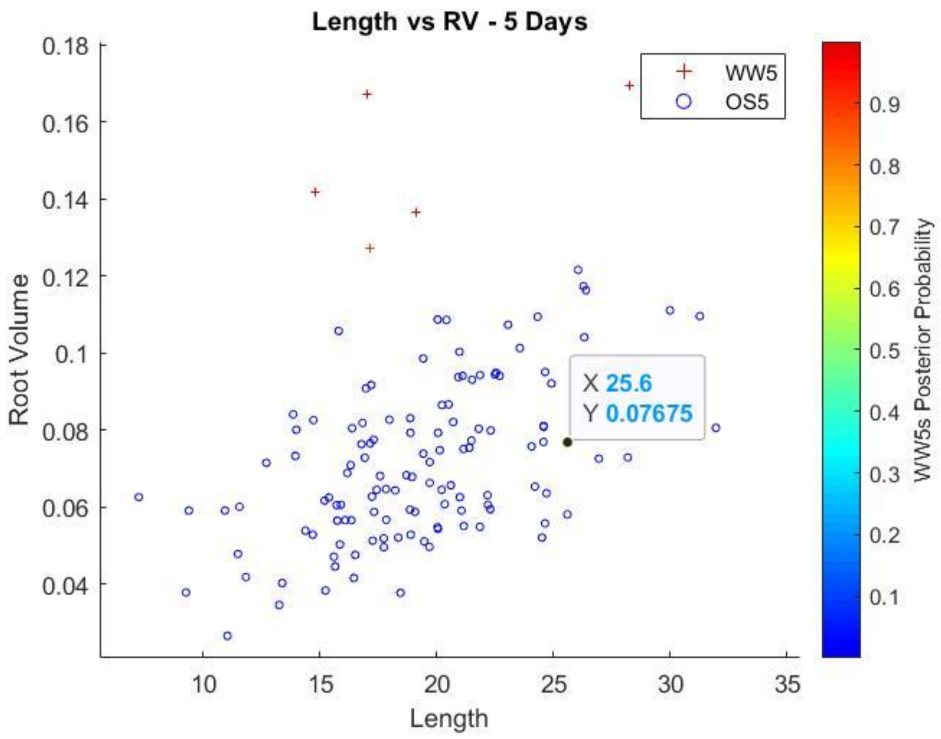
Appendix 2.19: AS74



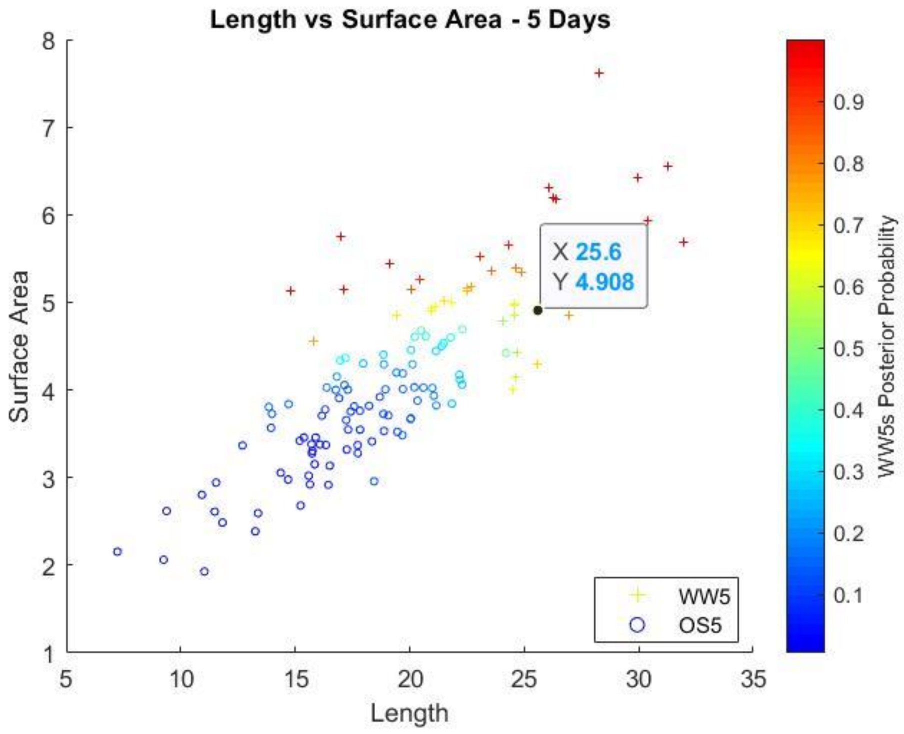
Appendix 2.20: AS74



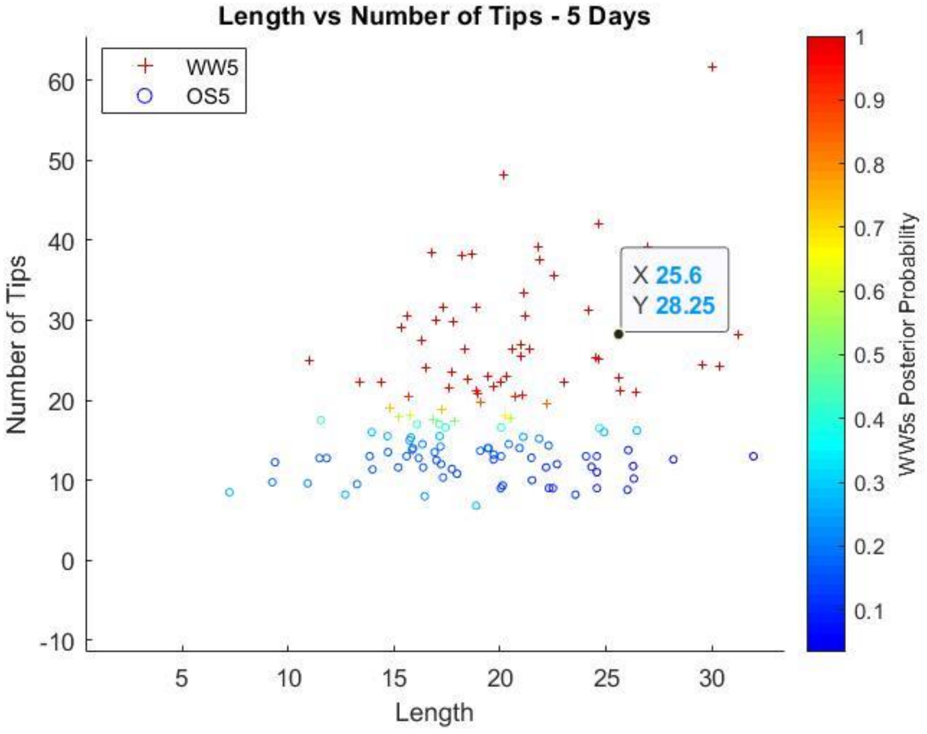
Appendix 2.21: AS131



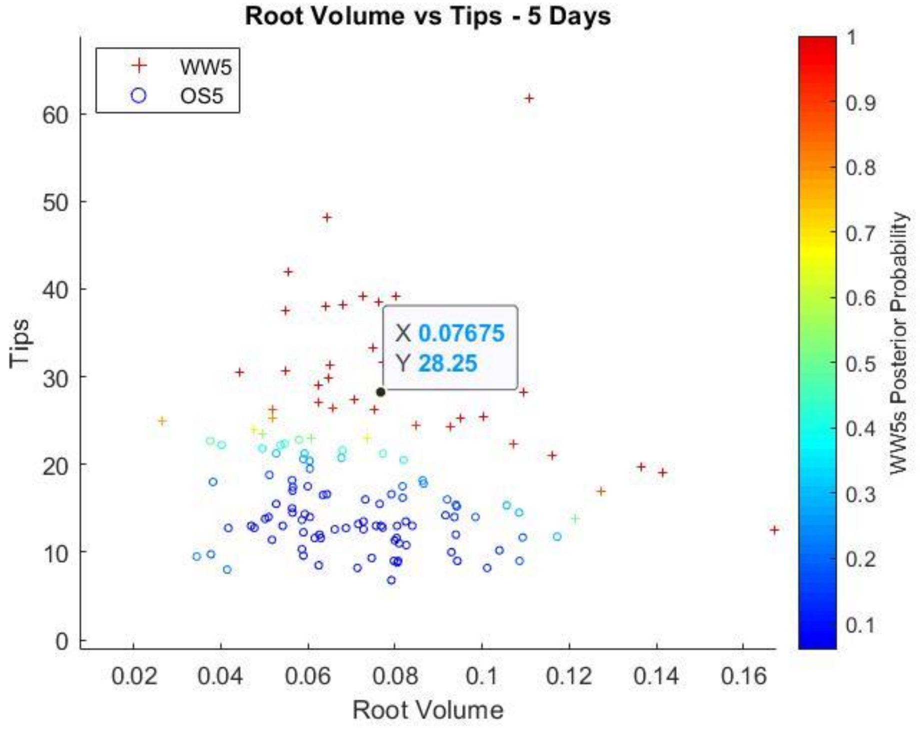
Appendix 2.22: AS131



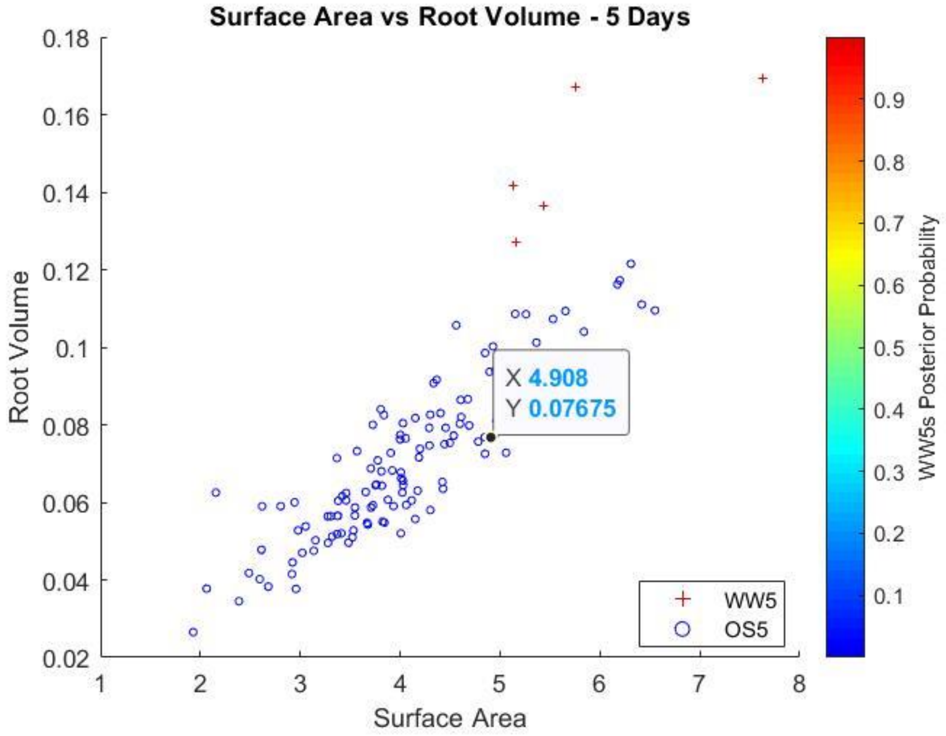
Appendix 2.23: AS131



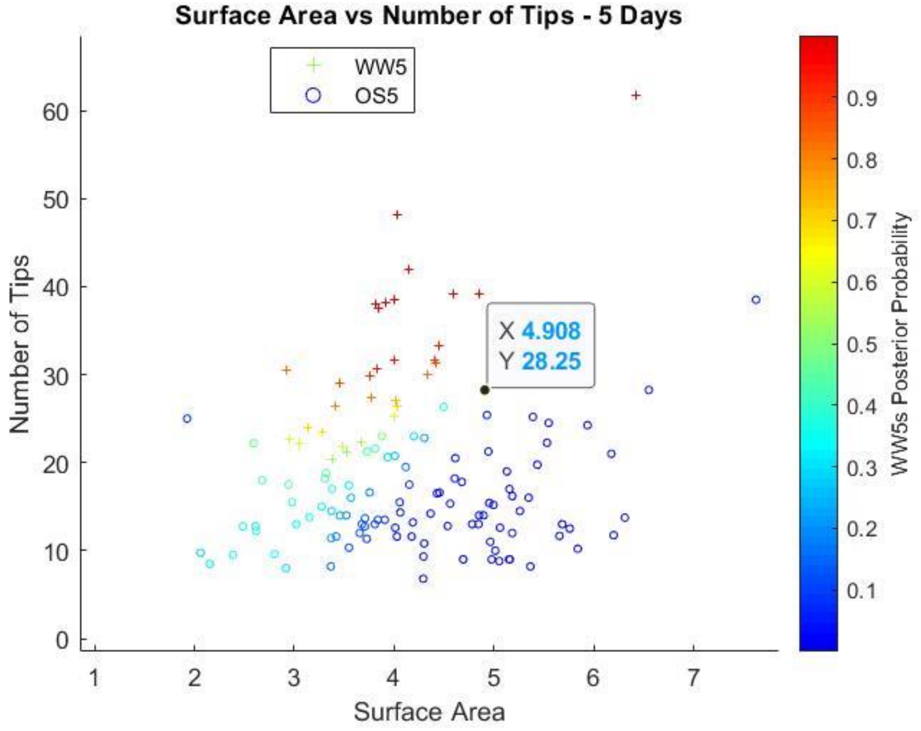
Appendix 2.24: AS131



Appendix 2.25: AS131



Appendix 2.26: AS131


Surface Area vs Number of Tips - 5 Days
Number of Tips
Surface Area
WW5
OS5
X 4.908
Y 28.25
WW5s Posterior Probability