Universidade Federal de Alagoas
Instituto de Computação
Laboratório de Computação Científica e Análise Numérica

Research report

Student: Danilo Fernandes Costa
Professor: Alejandro Frery

January
2020

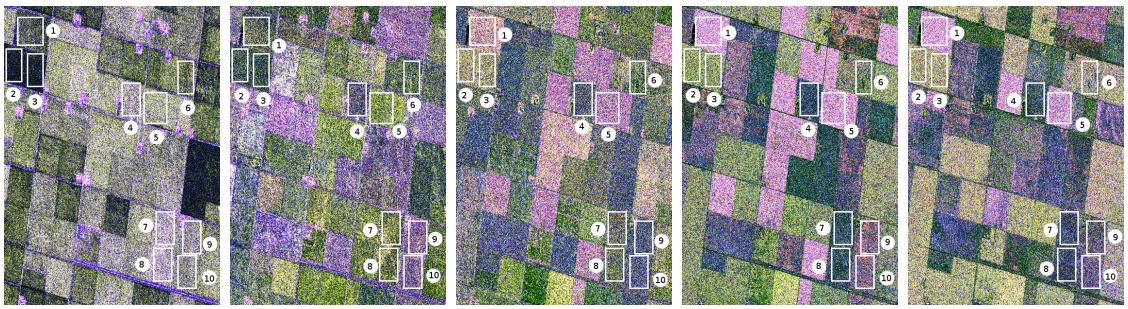
Abstract

In this report, we investigate the statistical properties of the distances between classified polarimetric synthetic aperture radar (PolSAR) data samples from crop regions over time and scatterers using the Geodesic Distance. This investigation can be justified by the possibility of using these properties to characterise these regions in terms of distances, and allowing the degree of vegetation and the way it has varied over time to be inferred.

1 Introduction

We show in this report results of the analysis of samples from plantation regions observed over time. The first observation was made on 16 May 2016, which was followed by four others at time intervals of 24 days. Those regions consist of three soybeans crops, three wheat, two oats, and two canola and are shown in Figs. 2a to 2e. Those samples were obtained using the classification of the regions given in Fig. 1. These dataset and classification were disponibilized by Prof. Avik Bhattacharya and his research group.

Figure 1: Classification of the regions on the PolSAR image



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 2: Samples analyzed over time: 1 to 10 corresponding, respectively, to Canola 43, Soybeans 231, Soybeans 232, Wheat 225, Canola 224, Soybeans 101, Oats 102, Oats 103, Wheat 105 and Wheat 104

2 Fitting the Beta Distribution

We fit the Beta distribution to histograms of the geodesic distances of the samples to the trihedral and random volume scatterers. Figs. 3 to 22 show the histograms of the distances between the scatterer and the pixels of the sample most similar to it. The number of those pixels are in Table 1, in which TR and RV indicate, respectively, trihedral and random volume. The parameters of the Beta distribution were estimated by maximum likelihood.

Table 2 shows the p-values of the goodness-of-fit as assessed by the Komolgorov-Smirnov test. Largest and smallest p-values are highlighted in bold.

Table 1: Number of pixels more similar to trihedral and random volume

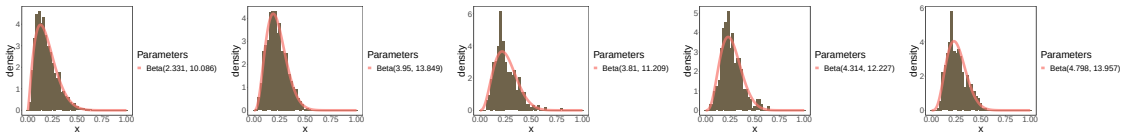
	16 May 2016		09 June 2016		03 July 2016		27 July 2016		20 Aug. 2016	
	TR	RV	TR	RV	TR	RV	TR	RV	TR	RV
SB 101	1481	71	867	306	429	633	397	743	319	829
SB 231	1285	148	656	381	449	706	615	649	508	775
SB 232	1275	132	649	387	474	649	596	630	518	811
WT 104	1618	144	478	861	161	842	133	1257	140	1099
WT 105	1488	222	482	852	196	842	154	1240	160	1117
WT 255	1552	205	567	746	180	917	85	829	157	904
CN 43	1964	155	1002	625	485	1195	168	1413	207	1395
CN 224	2106	87	1716	209	234	1298	162	1457	212	1334
OT 102	1441	246	1087	430	354	817	121	842	92	1054
OT 103	1429	266	1153	386	388	806	110	833	99	1045

3 Parameters evaluation

When observing the regions referring to Soybeans 231 and 232 along its samples in the Fig. 2, which are respectively indexed by 2 and 3 , it can be assumed that there was a gradual increase in the degree of vegetation of these regions.

Table 2: p-values of the Kolmogorov-Smirnov goodness-of-fit test of the distances to trihedral an random volume

	16 May 2016		09 June 2016		03 July 2016		27 July 2016		20 Aug. 2016	
	TR	RV	TR	RV	TR	RV	TR	RV	TR	RV
SB 101	0.065	0.517	0.947	0.758	**0.059**	0.195	0.452	0.109	0.401	0.144
SB 231	0.775	0.242	0.573	0.166	0.314	0.275	0.239	0.114	0.416	0.070
SB 232	0.244	0.340	0.968	0.328	0.713	0.070	0.422	0.357	0.163	0.630
WT 104	0.178	0.715	0.421	0.094	0.514	0.779	0.062	0.369	0.602	0.919
WT 105	0.231	0.090	0.069	0.139	0.557	0.613	0.108	0.195	0.192	0.252
WT 255	0.235	0.513	0.270	0.375	0.628	0.279	0.653	0.069	0.437	**0.993**
CN 43	0.238	0.406	0.217	0.202	0.930	0.318	0.623	0.732	0.262	0.747
CN 224	0.184	0.116	0.128	0.333	0.298	0.714	0.813	0.409	0.305	0.391
OT 102	0.289	0.191	0.243	0.532	0.384	0.212	0.710	0.370	0.928	0.396
OT 103	0.096	0.139	0.139	0.186	0.265	0.079	0.936	0.079	0.989	0.489



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

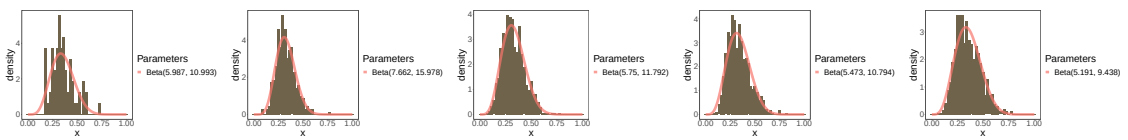
Figure 3: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Soybeans 101 most similar to trihedral

In order to relate this to the variation of the information contained in the distances of the data to the trihedral scatterer, in Figs. 23a and 23b we show, for each observation, a boxplot of the means of the distances between trihedral and the subregions generated by dividing a region into 45 subregions of size 7×6. In addition, all boxplots were connected by the mean of their means. It can be observed that the median in the first two samples of both regions is different at the confidence level of 0.95.

We adjusted the mean as a function of time for both regions with the following function:

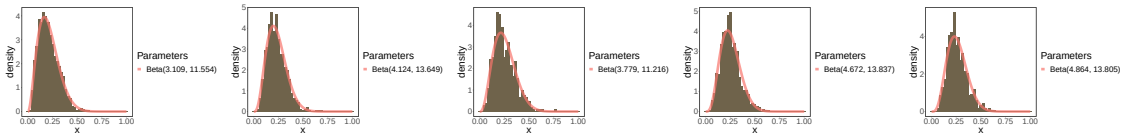
$$f(t) = -\frac{a}{bt + c} + d, \tag{1}$$

in which $a = 4.741$, $b = 2.415$, $c = 67.565$, $d = 0.276$ and t is the number of days since the first observation ($t = 0$). We checked lack of fit with ANOVA; Tables 3



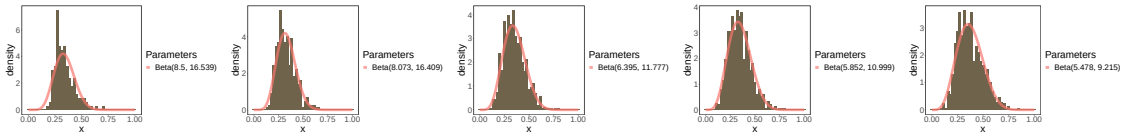
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 4: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Soybeans 101 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 5: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Soybeans 231 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 6: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Soybeans 231 most similar to random volume

and 4 show the results. We conclude that the proposed model is acceptable at the significance level of 0.1.

4 Separability test

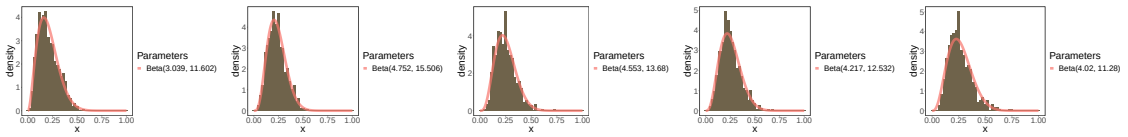
We also performed a separability test based on the Helinger distance between the distances of different images from Soybeans 231 to trihedral assuming the Beta distribution. Table 5 shows the p-values of the null hypothesis that each pair comes from the same law. We observe that at level 0.05, the only null hypothesis that cannot be rejected is that the data from the two last dates come from the same law.

4.1 Geodesic Purity Index

In this section, we analyse the geodesic purity index computed for the samples from Soybeans 231 region, which images are 30×65 pixels. The roadmap for the data analysis of these images is:

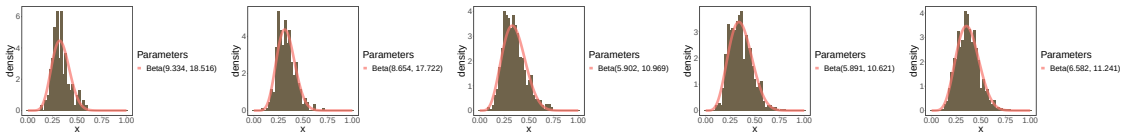
1. Obtaining the geodesic purity index of the data;

2. Making a descriptive analysis of the data with histograms and boxplots;



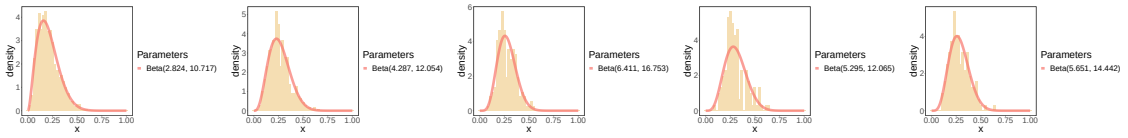
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 7: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Soybeans 232 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 8: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Soybeans 232 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 9: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Wheat 104 most similar to trihedral

3. Fitting the data with models;

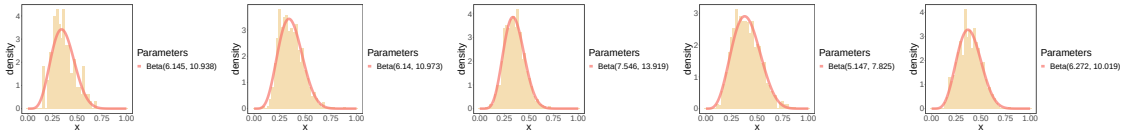
4. Making separability tests.

A first inspection of the purity indexes suggested that Beta distributions may be a good explanatory model. The data look crammed in their original space, though.

After this initial qualitative analysis, we decided to transform the data applying the logarithmic function. Fig. 24a shows the histograms and boxplots. The closeness to Gaussian distributions is remarkable. The QQ-Plots shown in Fig. 24b provide more evidence of such good adherence. Table 6 provides the p-values of the Shapiro-Wilk test of goodness-of-fit to the Gaussian distribution. There is no evidence, thus, to reject the hypothesis that the log-transformed purity data follows Gaussian distributions.

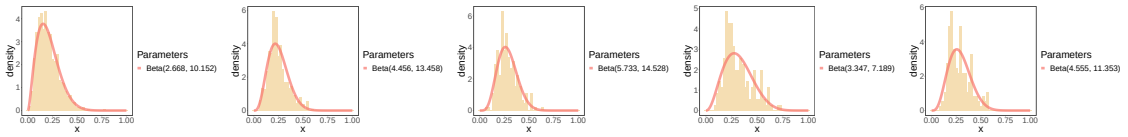
5 Classifier for vegetation regions

We propose a classifier based on results of the analysis of a subregion of Soybeans 231 with dimensions 15×15 pixels. We fitted a Beta distribution to its geodesic distances to the left and right helices on the first and last observation of this subregion, which belong to the regions indicated by index 2 in the figures 2a and 2e,



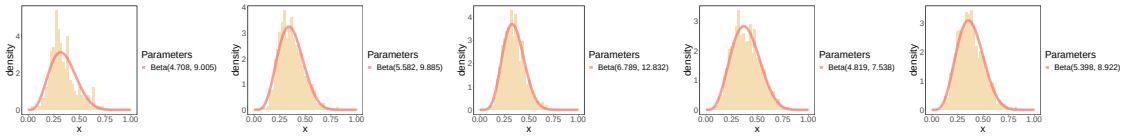
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 10: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Wheat 104 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 11: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Wheat 105 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

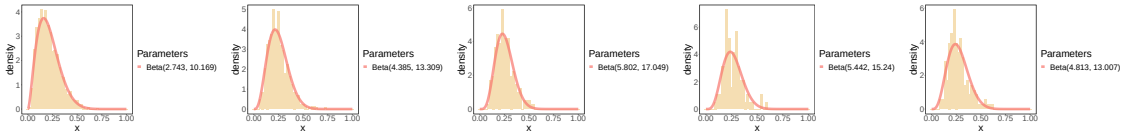
Figure 12: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Wheat 105 most similar to random volume

respectively. Figs. 25a and 25a show the histograms and the fitted densities. In adiction, the Komolgorov-Smirnov test for goodness-of-fit was performed and returned p-values are in the table 7.

Since the distances to the left helix and right helix are independent, because these are orthogonal, it is possible to define $d = 0.912$ as cutoff point for the densities Beta(21, 1.6) and Beta(10, 1.85) and obtain the joint probabilities shown in the table 8, where D_{lh} and D_{rh} are respectively the distance to left helix and right helix. These joint probabilities allow evaluate the error related in the separation of populations by this cutoff point.

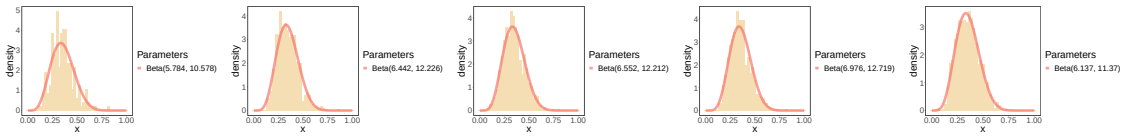
Assume that the pixels in the first and last observation as poor and rich in vegetation, respectively. Then, it is possible classify pixels with $d_{lh} > 0.912$ and $d_{rh} > 0.912$ as poor in vegetation and those with $d_{lh} \leq 0.912$ and $d_{rh} \leq 0.912$ as rich in vegetation. By this rule, there is 0.09 probability that a poor pixel in vegetation be classified rich and vice-versa. However, this approach allows classifying only 56 % of both populations.

For classify the other 44 % of the population, the components of the data in the



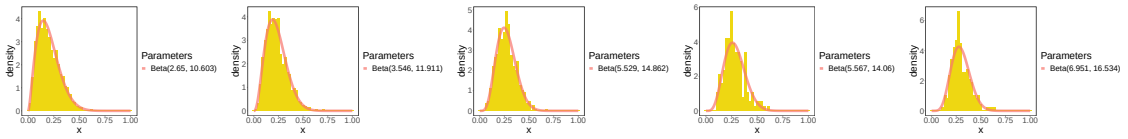
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 13: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Wheat 225 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 14: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Wheat 225 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 15: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Canola 43 most similar to trihedral

direction of these elementary scatterers are removed as follows:

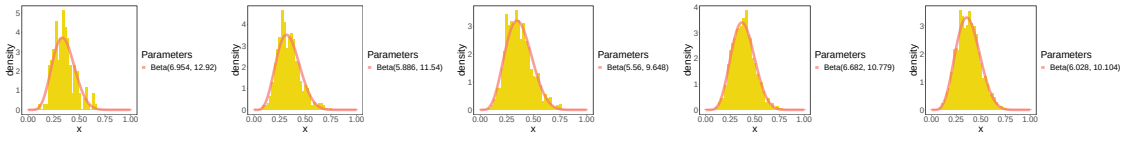
$$v'_{data} = v_{data} - \frac{\langle v_{data}, v_{lh}\rangle}{\|v_{lh}\|^2} v_{lh} - \frac{\langle v_{data}, v_{rh}\rangle}{\|v_{rh}\|^2} v_{rh}, \tag{2}$$

where v_{data}, v_{lh} and v_{rh} are respectively data, left helix and right helix in Kennaugh form. Then, we compute the followin distance:

$$D'_d = \frac{1}{\pi}\cos^{-1}\frac{\langle v'_{data}, v_d\rangle}{\|v'_{data}\|\|v_d\|} = \frac{1}{2}GD(v'_{data}, v_d), \tag{3}$$

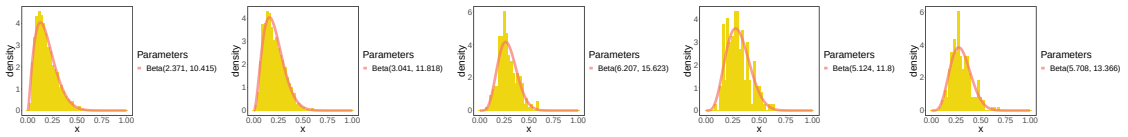
where $D'_d \in [0,1]$ and v_d is dihedral elementary scatterer in Kennaugh form. This distance was used because $GD(v'_{data}, v_d) \in [0,2]$ for analysed subregion. The histograms of this distance between dihedral and analysed samples are shown in the figure 26 and the p-values from Komolgorov-Smirnov goodness-of-fit test for first and last observation are respectively 0.127 and 0.105.

Since the procedure described makes D'_d independent of D_{lh} and D_{rh}, the fitted distributions can be used to classify the remaining population. For this, define $d_d = 0.462$ and $d_d = 0.512$, which are inserctions between densities, as cutoff points, whose related probabilities are shown in the table 9. With this, a pixel unclassified by d_{lh} and d_{rh} and with $d'_d < 0.462$ or $d'_d > 0.512$ can be classified as rich in vegetation



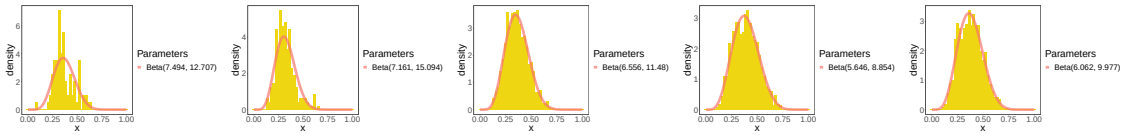
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 16: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Canola 43 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 17: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Canola 224 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 18: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Canola 224 most similar to random volume

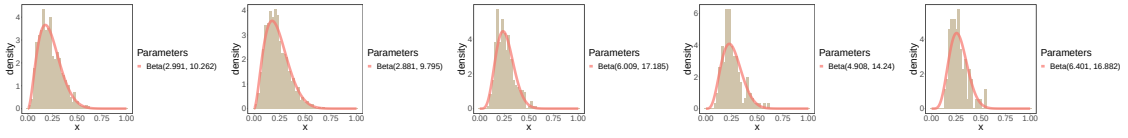
and, in opposite case, as poor in vegetation. In this situation, the probability of classify a poor pixel as rich is 0.13 and the opposite with probability 0.44.

This classifier can be evaluated by analysing of your confusion matrix. The table 10 shows the theoretical confusion matrix (in percentage) for this model and the table 11 shows the corresponding values of accuracy, coverage and precision.

The tables 12 and 13 show respectively the confusion matrices (in percentage) obtained by applying the model to the samples from the analysed subregion and Soybeans 231 region. In addiction, the tables 14 and 15 show their corresponding values of accuracy, coverage and precision. When comparing the theoretical values with the values obtained by applying the classifier to the data in these tables, the proximity between them is evident, which suggests that the model is suitable for the data.

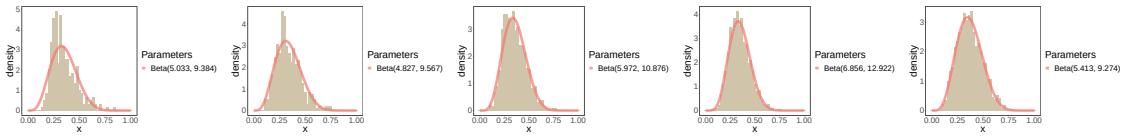
6 Conclusion

We concluded that the Beta distribution can be used to model the distances from crop regions to the trihedral and random volume scatterers and that their parameters vary when the degree of vegetation changes. Additionally, the lognormal distribution proved to be a possible model for the geodesic purity index. These



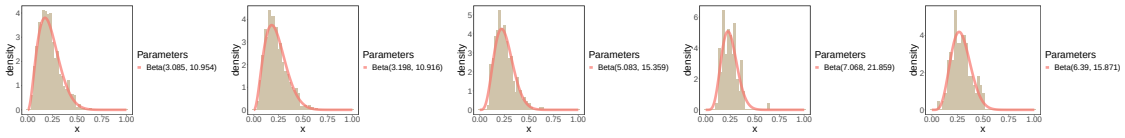
(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 19: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Oats 102 most similar to trihedral



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

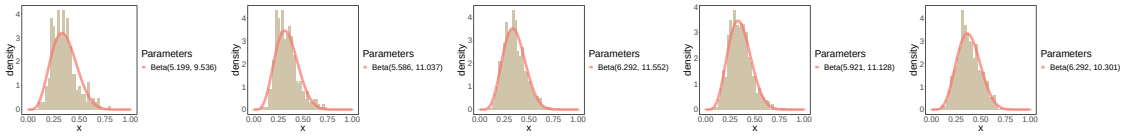
Figure 20: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Oats 102 most similar to random volume



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

Figure 21: Histograms of the Geodesic Distances between trihedral and the pixels of the sample extracted from Oats 103 most similar to trihedral

conclusions highlight the relevance of conducting a complete investigation to characterize vegetation images based on Geodesic Distance that can result in applications of deforestation and crop monitoring.



(a) 16 May 2016 (b) 09 June 2016 (c) 03 July 2016 (d) 27 July 2016 (e) 20 Aug. 2016

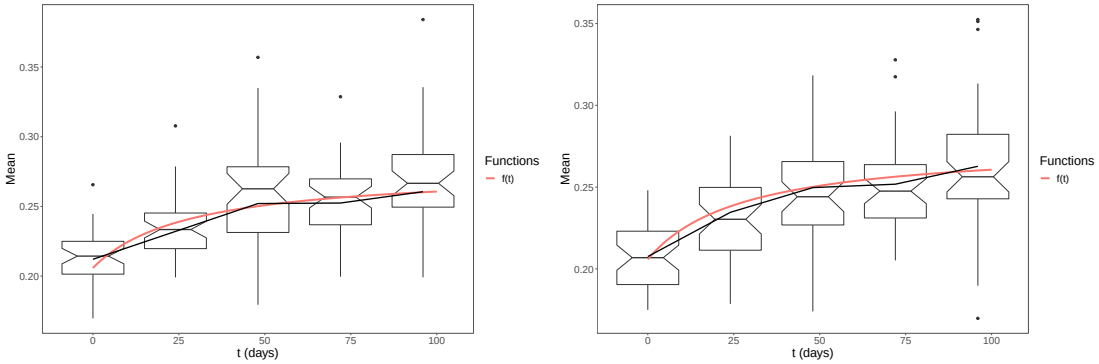
Figure 22: Histograms of the Geodesic Distances between random volume and the pixels of the sample extracted from Oats 103 most similar to random volume

Table 3: ANOVA for lack of fit on Soybeans 231

	Degree of freedom	Sum of squared errors	Mean squared error	Fisher statistics	p-value
Residual	223	0.1912	0.0008		
Lack od fit	3	0.0043	0.0014	1.7087	0.1661
Pure error	220	0.1859	0.0008		

Table 4: ANOVA for lack of fit on Soybeans 232

	Degree of freedom	Sum of squared errors	Mean squared error	Fisher statistics	p-value
Residual	223	0.1836	0.0008		
Lack od fit	3	0.0020	0.0007	0.7973	0.1965
Pure error	220	0.1819	0.0008		



(a) Soybeans 231 (b) Soybeans 232

Figure 23: Mean of the distances between trihedral and samples extracted from Soybeans 231 and 232 over time

Table 5: p-values from Separability test

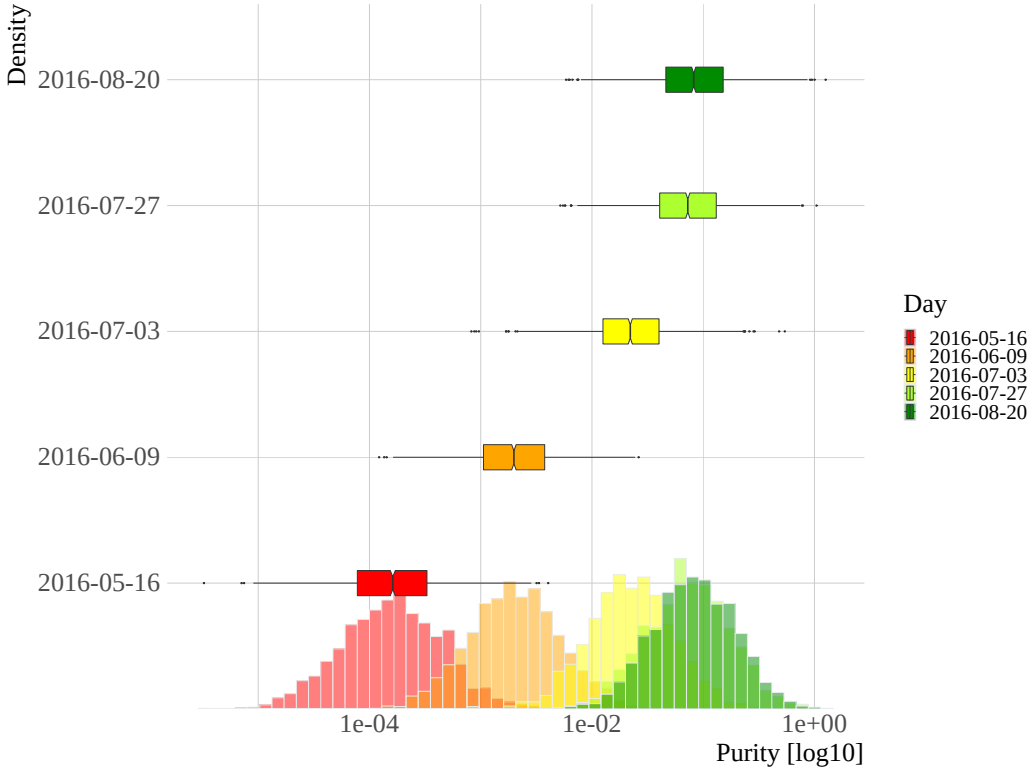
	09 June	**03 July**	**27 July**	**20 Aug.**
16 May 2016	7.467×10^{-8}	9.223×10^{-12}	5.159×10^{-21}	1.311×10^{-24}
09 June 2016	—	2.640×10^{-3}	4.551×10^{-4}	1.757×10^{-6}
03 July 2016	—	—	4.318×10^{-2}	1.072×10^{-2}
27 July 2016	—	—	—	3.642×10^{-1}

Table 6: p-values from Shapiro-Wilk Test

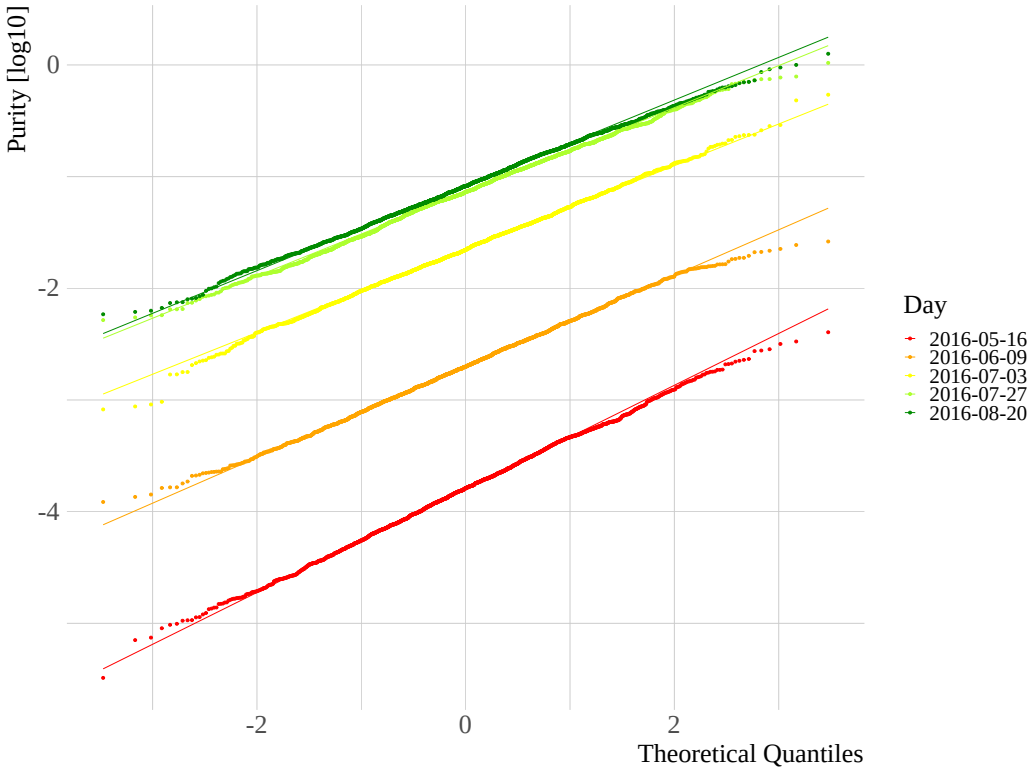
Day	**16 May**	**09 June**	**03 July**	**27 July**	**20 Aug.**
***p*-value**	0.4963	0.0650	0.3494	0.0585	0.3919

Table 7: p-values of the Kolmogorov-Smirnov goodness-of-fit test of the distances to left and right helix

	Left helix	Right helix
First sample	0.406	0.172
Last sample	0.940	0.817

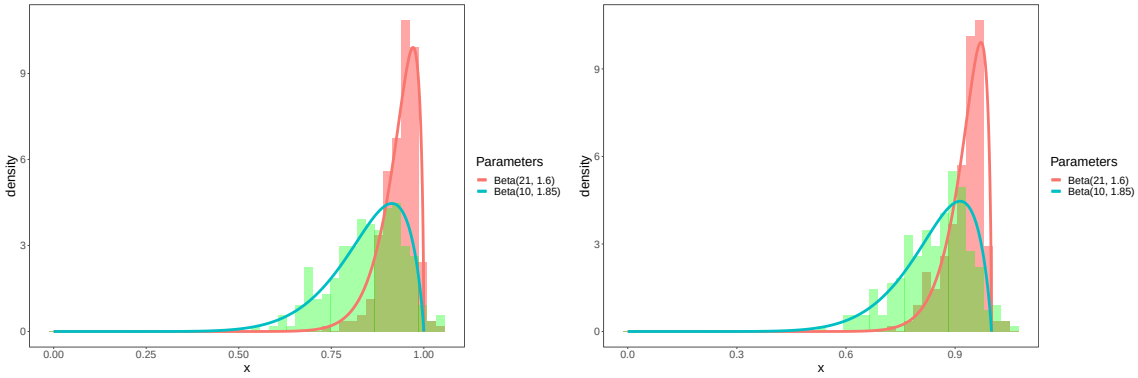


(a) Histograms



(b) QQPlots

Figure 24: Descriptive analysis of the logarithm purity values for each image



(a) Histogram of the distances to left helix (b) Histogram of the distances to right helix

Figure 25: Histograms of the distances between subregion extracted from Soybeans 231 and elementary scatterers

Table 8: Joint probabilities for distance to left and right helix

	$D_{rh} \leq 0.912$		$D_{rh} > 0.912$	
	$D_{rh} \leq 0.912$	$D_{rh} > 0.912$	$D_{rh} \leq 0.912$	$D_{rh} > 0.912$
First sample	**0.09**	0.21	0.21	**0.19**
Last sample	**0.19**	0.21	0.21	**0.09**

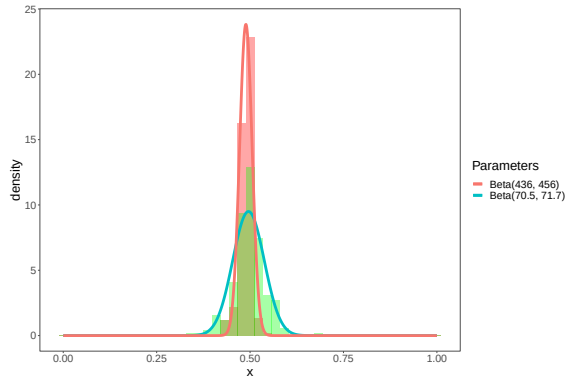


Figure 26: Modified geodesic distance between dihedral and samples

Table 9: Probabilities for modified geodesic distance to dihedral

	$D'_d < 0.462$	$0.462 \leq D'_d \leq 0.512$	$D'_d > 0.512$
First sample	0.05	0.87	0.08
Last sample	0.21	0.44	0.35

Table 10: Theoretical confusion matrix

	Poor in vegetation	Rich in vegetation
Poor in vegetation	0.855	0.145
Rich in vegetation	0.275	0.725

Table 11: Theoretical accuracy, coverage and precision

	Accuracy	Coverage	Precision
Poor in vegetation	0.790	0.855	0.757
Rich in vegetation	0.790	0.725	0.833

Table 12: Confusion matrix obtained by applying the model to the analysed subregion of Soybeans 231

	Poor in vegetation	Rich in vegetation
Poor in vegetation	0.827	0.173
Rich in vegetation	0.333	0.667

Table 13: Confusion matrix obtained by applying the model to the Soybeans 231 region

	Poor in vegetation	Rich in vegetation
Poor in vegetation	0.788	0.212
Rich in vegetation	0.333	0.667

Table 14: Accuracy, coverage and precision for the model applied to the analysed subregion

	Accuracy	Coverage	Precision
Poor in vegetation	0.747	0.827	0.713
Rich in vegetation	0.747	0.667	0.794

Table 15: Accuracy, coverage and precision for the model applied to the Soyebeans 231 region

	Accuracy	Coverage	Precision
Poor in vegetation	0.727	0.788	0.703
Rich in vegetation	0.727	0.667	0.759