UC Berkeley
UC Berkeley Previously Published Works

Title
Rock Mass Characterization of Discontinuities by Unsupervised Machine Learning

Permalink
https://escholarship.org/uc/item/96k9x61p

Journal
Geosciences, 16(9)

ISSN
2076-3263

Authors
Russo, Brittany M
Kayen, Robert E

Publication Date
2026-09-01

DOI
10.3390/geosciences16090347

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

geosciences

MDPI

Article

Rock Mass Characterization of Discontinuities by Unsupervised Machine Learning

Brittany M. Russo [1] and Robert E. Kayen [2,*]

[1] Department of Geological Sciences, Bridgewater State University, Bridgewater, MA 02324, USA; b1russo@bridgew.edu
[2] Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA
* Correspondence: rkayen@berkeley.edu

Abstract

Rock slope landslide potential, tunnel design, and the engineering of underground spaces critically require an analysis of rock joint orientation and spacing. Joint set measurements are typically determined by hand in the field using a compass–clinometer to measure the orientation of the geologic planes with respect to north and the dip of the plane with respect to the horizontal. Unmanned aerial vehicles (UAVs) can be utilized to capture hundreds of photos to create high-resolution 3D models of a rock outcrop, capturing visible joint set discontinuities on a faceted surface of a triangular irregular network (TIN). Computing methods of facets and facet normals from a point cloud allow for the characterization of discontinuity orientations without the need for manual measurements in the field. However, these methods used to calculate facets and facet normals result in the addition of noise in the dataset, which increases the difficulty of analysis. A two-stage filtering process employing density-based spatial clustering of applications with noise (DBSCAN), and the second derivative of the remaining clusters, removes data that are not representative of a discontinuity within the intact rock mass. Finally, an unsupervised clustering algorithm, K-Means Clustering, is applied to the dataset to extract the dip and dip direction of discontinuities. The methodology used to identify the orthogonal joint sets on a dam spillway demonstrated good performance across all identified joint sets, with the estimated variability in dip and dip direction broadly comparable to that observed in the manual field dataset. This indicates that this newly developed proof-of-concept approach can reliably capture the orientation and variability of orthogonal joint sets from large datasets.

Keywords: rock mechanics; kinematics; machine learning; unsupervised clustering

Check for updates

Academic Editors: Qingsheng Bai, Mohamed Shahin and Chaofan Chen

Received: 1 July 2026
Revised: 24 August 2026
Accepted: 28 August 2026
Published: 31 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Estimation of the potential of instability of rock masses is challenging, as strength is not the most important factor. Rather, typically, the frictional properties of pre-existing discontinuities control the deformation behavior of rocks. A variety of discontinuities can be present in a rock mass including joints, faults, bedding planes, and fractures [1]. The presence of these discontinuities allows for movement at resistive strengths much lower than the intact rock strength. Thus, frictional discontinuities and their orientations often have greater importance than intact rock strength, which can affect engineering design [2,3]. Therefore, understanding the properties of discontinuities such as roughness, orientation, fill material, and spacing are vitally important. These characteristics can be

difficult to measure since they need to be observed from natural outcroppings, boreholes, or excavations [3]. These field properties are normally recorded manually with measuring tapes, compass–clinometers, and hand lenses [4].

Characterizing rock masses and the properties of discontinuities is vital for many aspects of geosystems engineering including slope stability, foundation design, mining, seismic site response, and understanding groundwater flow and contaminant transport [5]. The properties of discontinuities are used as inputs for a range of analyses including rock slope kinematic analyses, key block analysis, and discontinuous deformation analysis. It is important to have representative data of the rock mass discontinuities for analysis, which can be difficult to obtain. Collecting field data to characterize rock mass discontinuities has many challenges including human bias in sampling measurements as well as access to outcrops and rock faces. Manual field mapping can be costly and time-consuming [1].

New field data collection methodologies have been developed to help reduce the challenges involved with characterizing rock masses. Scanline surveys were developed to characterize the frequency and trace length of discontinuities and the rock quality designation (RQD) of the rock mass [6]. Using techniques that provide a more complete understanding of the rock mass, such as LIDAR, allows for a better understanding of the natural variability of discontinuities. LIDAR produces detailed 3D point clouds that have been used to extract discontinuity characteristics [7–9].

The use of unmanned aerial vehicles (UAVs) has become increasingly popular for imaging rock outcrops. Hundreds of photos of the outcrop are processed by digital photogrammetric methods to create detailed 3D models of the exposed rock surface through the structure from motion (SfM) technique. The use of 3D models of outcrops from UAV data has been used to produce point clouds in order to map discontinuity traces, spacing, and frequency [9,10]. UAV data have also been shown to be effective at extracting discontinuity spatial data when combined with infrared thermography, as it helps to evaluate persistence and weathering of fractures and joint sets [11,12]. Other work has combined UAV-collected data with machine learning algorithms to estimate the orientation and spacing of joint sets [13–15].

This paper aims to create a proof-of-concept methodology that uses surface dip and dip direction measurements from facets and facet normals collected by the SfM method using UAVs to semi-automatically filter and extract the joint set information for orthogonal joint sets in granitic outcrops. Filtering of the dataset to remove vegetation, blocks that have been displaced, and noise was performed using a two-stage filtering process involving density-based spatial clustering of applications with noise (DBSCAN) and the second derivative of the clusters detected in DBSCAN. Unsupervised K-Means Clustering was used on the filtered dataset to detect the number of joint sets at the site and the dip and dip direction of each joint set.

The novelty of this methodology lies in the use of a two-stage filtering method to remove noise in a very large dataset. With a dataset as large as one produced using UAV or LIDAR, noise can be introduced into the data during data collection, data processing, or from objects in the field that are not useful to the analysis, such as vegetation or concrete walls. It can be difficult to isolate all sources of noise in the dataset using only one stage of filtering.

The novelty of this methodology also introduces a step to account for the natural variation in dip direction in steeply dipping joints, which is typically not addressed with joint sets. Steeply dipping joints with dip directions around 0° and 360° have similar orientations in geologic space but are treated as numerically separate in analysis. This is accounted for in the “Same Cluster Rule” applied in Step 4 of the methodology.

This methodology aims to address the challenges of manual field mapping by removing human bias created from selective measurements in the field to better represent the natural variability of joint set orientations found in nature and increasing the time efficiency of collecting, processing, and analyzing the data. The methodology framework is visualized in Figure 1.

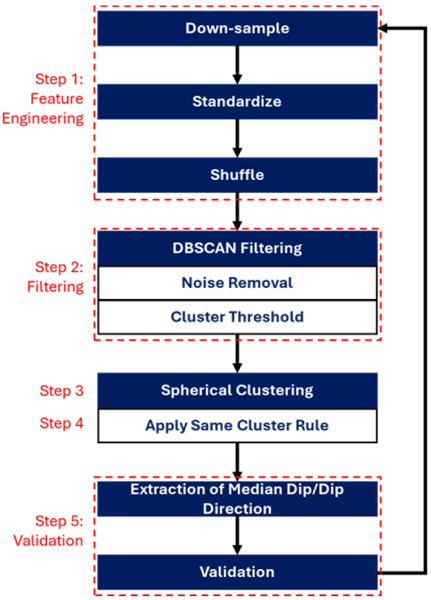


Figure 1. Procedure for the determination of joint set orientations.

2. Materials and Methods

2.1. Data Collection

The rock discontinuity orientations were collected on a dam spillway of a concrete gravity dam used for hydropower in Nevada County, California, USA. The dam is in the Sierra Nevada Mountains with the purpose of water supply, recreation, and power generation. This study was conducted on the auxiliary dam spillway composed of granite. The spillway is subjected to episodic discharges from the reservoir and is actively eroding with flooding events. The granite at this site has three major orthogonal joint sets. Figure 2 shows a plan view model of the spillway.

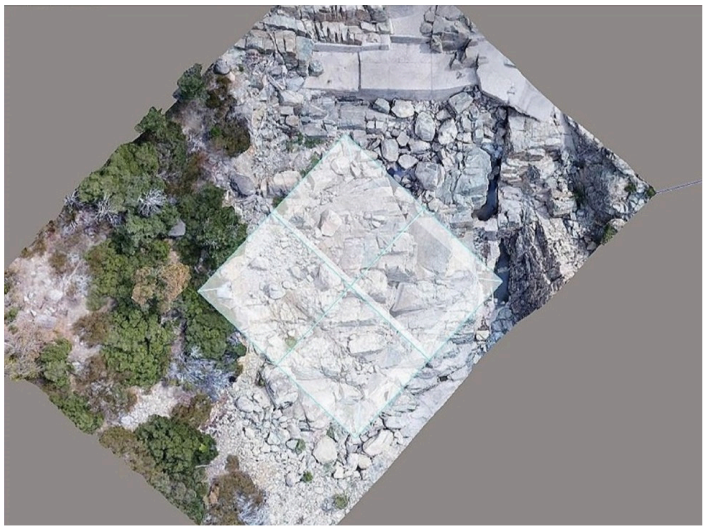


Figure 2. Plan view 3D model of the dam spillway in California. The area shaded represents the point cloud selection for determination of joint orientation.

Three datasets were measured on the spillway of Spaulding Dam. The first dataset contained 100 measurements and was collected in September 2022 by 30 students in the CE281 "Engineering Geology" course at the University of California, Berkeley. Students manually measured the joint set orientations using the StereonetMobile 4.2 and Clino 3.1.3 apps on their smartphones. The stereonet and data with respect to dip and dip direction are shown in Figure 3a,d.

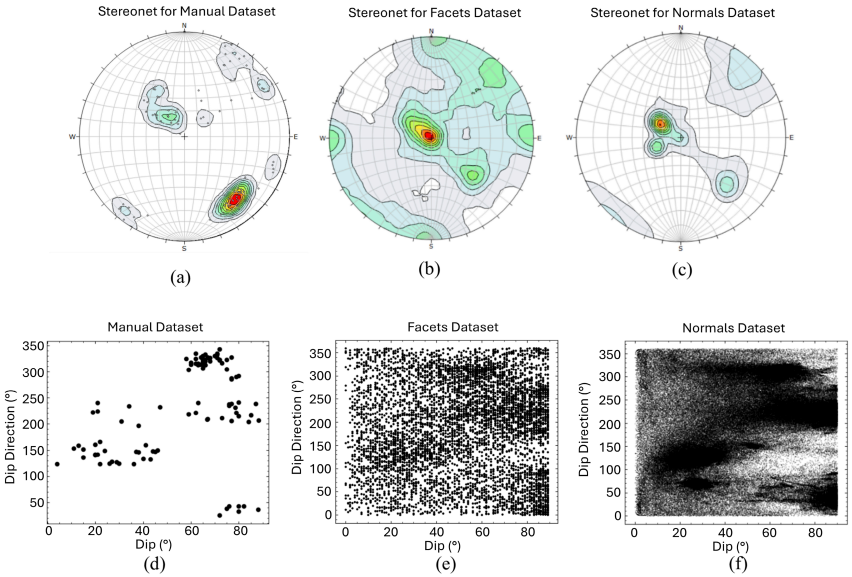


Figure 3. Stereonets showing the pole concentrations of the joint sets with high density shown in red and low density in gray for the (**a**) the manual dataset, (**b**) the facets dataset, and (**c**) the normals datasets with the data with respect to dip and dip direction for (**d**) the manual dataset, (**e**) facets dataset, and (**f**) facet-normals dataset.

The second and third datasets were collected using an unmanned aerial vehicle (UAV). The UAV flown was a DJI Mavic 2 Pro (DJI, Shenzhen, China) with the image collection parameters shown in Table 1. The Mavic DJI camera lens (DJI, Shenzhen, China) was calibrated to remove spherical distortions.

Table 1. UAV Data Collection Parameters.

Setting Type	Setting Parameter
Georeferencing	RTK GPS with Position Accuracy of ~10 mm
Flight Altitude	20, 40, and 60 m
Front Image Overlap	80%
Side Image Overlap	75%
Point Density of Scanner	Function of distance away from the laser scanner but typically below ~10 mm

The data were processed using Bentley's Context Capture version 23.00.001317 following the methodology described in [9]. CloudCompare v2.6.2 [16] was the open-source program used to produce facets and facet normals from point clouds of the spillway rock mass. From the point cloud produced by the SfM method, a triangular irregular network (TIN) was produced using point triangulation. Then, the facet surfaces were extracted using the kd-Tree (Octree method), and their normals represent the TIN. The facets and facet normals represent the surfaces of all the exposed materials at the site, including rock outcrops, vegetation, detached blocks, and noise. The dip and dip directions of all the surfaces were exported and plotted as stereonets, shown in Figure 3b,c. The facets

dataset contained 6727 data points while the normal dataset produced from the point cloud contained 275,426 data points.

2.2. Machine Learning Methods

Two unsupervised machine learning models were used in this methodology. The first was DBSCAN. This is a density-based clustering algorithm which groups high-density regions surrounded by low-density regions as clusters. One of the advantages of this method is that clusters can have arbitrary shapes and sizes.

This algorithm has two hyperparameters that need to be defined first. The hyperparameters characterize the density of points needed to be classified as a cluster. The first hyperparameter is ε, which represents the radius of the neighborhood to check for density. The second hyperparameter is minPts, which represents the number of points needed to be in the neighborhood defined by ε to be classified as a cluster. The minPts hyperparameter is defined as double the dimensionality of the dataset [17]. Since this dataset contained only two variables, dip and dip direction, the minPts parameter was set as 4. The ε hyperparameter is determined through an elbow rule where the average Euclidean distance (E) between each point in the dataset is calculated to its nearest neighbor through Equation (1):

$$E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \quad (1)$$

where x_1 and y_1 are the coordinates of the first point, and x_2 and y_2 are the coordinates of the second point.

The elbow rule is visualized by first sorting the average distances in ascending order and plotting them with respect to ε. The inflection point in the graph determines the optimal ε value [18].

The second unsupervised machine learning model used in this methodology was K-Means Clustering. K-Means Clustering is a distance-based algorithm which groups data into clusters based on how close data points are to a centroid. The algorithm inputs used in this study included k-means++ for the method of initialization to select the initial cluster centers, 300 maximum iterations, 10 as the number of initializations with different centroid locations, and a random seed of 252. This algorithm has one hyperparameter, k, which defines the number of clusters used in the analysis. The k parameter was determined through two methods, the elbow rule and the silhouette rule, as shown in Figure 4. The elbow rule for K-Means Clustering is calculated as the within-cluster sum of squares (WCSS) for a variety of clusters using Equation (2):

$$WCSS = \frac{1}{N}\sum_{i=1}^{N} E_i, \quad (2)$$

where N is the total number of points within a cluster [19]. The optimal cluster is chosen at the maximum curvature of the plot.

The silhouette rule is another method for determining k. This rule calculates a silhouette average (S) through Equation (3) [20]:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad (3)$$

where i represents a data point, a(i) is the average distance between i and all other data points in a cluster which is part of i, and b(i) is the average distance to the closest cluster of which i is not a part of. The average distance was calculated using the Euclidean distance.

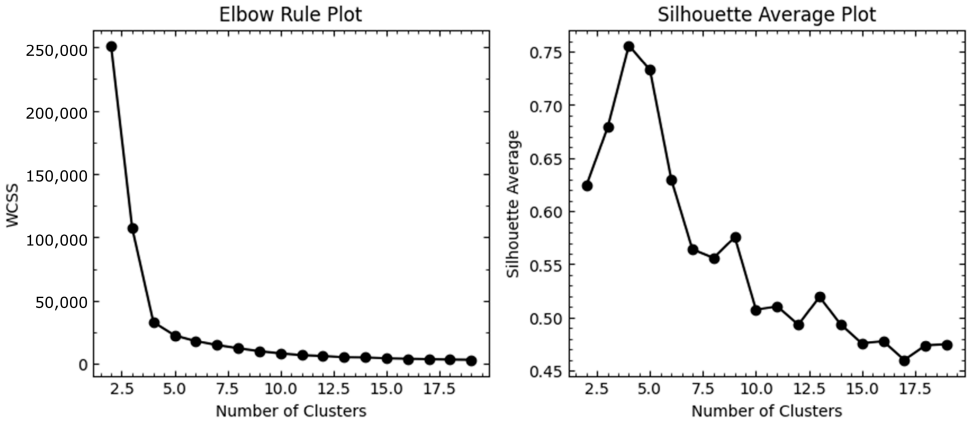


Figure 4. Clustering hyperparameter optimization showing the (**a**) elbow rule and (**b**) the silhouette rule.

The silhouette average ranges between −1 and 1 and represents how similar data points are within a cluster compared to data points in the other clusters, with a value of 1 being ideal. The optimal k is chosen as the cluster that maximizes the silhouette average. All machine learning algorithms were written in Python 3.12 using the SciKit Learn 1.7.0 libraries for K-Means and the silhouette score [21]. The results of the machine learning algorithms were displayed on stereonets using the Mplstereonet 0.6.3. Python library [22].

3. Results and Discussion

3.1. Step 1: Feature Engineering

Raw datasets created from UAV data collection methods can contain hundreds of thousands of data points. Large datasets can be challenging to work with, as they are computationally expensive in terms of storage as well as the reduction in data visualization and interpretation. To reduce these effects on the methodology, feature engineering is needed. Feature engineering is the process of manipulating data to be usable inputs into machine learning algorithms and may include sampling data, standardizing or normalizing data, selecting features for analysis, or removing data from the dataset.

One feature engineering method used in this procedure was down-sampling on the normals dataset. With over 250,000 data points, the algorithms were slow to run and would crash when there was not enough memory. To reduce the computational cost, the normals dataset was down-sampled by randomly removing data points. The number of data points removed was analyzed by creating datasets that contained 2.4% to 40% of the original dataset.

Removing data points allowed for a significant decrease in runtime of the methodology while not affecting the hyperparameter optimization of the number of clusters needed for the analysis. Using the full dataset resulted in the code crashing due to computer memory issues. By randomly removing 60% (165,256) of the data points using a random seed of 65, the methodology successfully completed in 2 h. To understand how down-sampling affected the performance of the methodology, the bias and dispersion of the distribution of the dip and dip direction were calculated between the manual, or reference, dataset and the down-sampled datasets. The bias of the distribution represents the lateral shift between the distributions and was calculated using Equations (4) and (5):

$$R_i = Z_i - Z_{known,i}, \quad (4)$$

where R_i is the residual and compares the known value ($Z_{known,i}$) of the reference dataset with the modeled value (Z_i) of the facet and normals datasets:

$$\mu = \sum_{i=1}^{n} \frac{R_i}{n}, \quad (5)$$

where n is the number of points in the dataset, and μ is the mean of the residuals and represents the sample bias. When the bias is large (positive or negative), there is a greater lateral shift between the distributions being compared.

The distributions between the reference and modeled distributions can also be evaluated based on the dispersion. The dispersion describes how similar the spread of values is between two distributions and is calculated using Equation (6):

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(R_i - \mu)^2}{n-1}}, \quad (6)$$

where σ is the standard deviation of the residuals and represents the dispersion between the reference and modeled distributions.

Equations (4) through (6) were applied to the dip and dip direction probability density functions for each joint set from the manual dataset as well as the probability density functions for the normals dataset when removing 60% of the data and 97.6% of the data. The comparison between the bias and dispersion (in degrees) between the manual dataset and the normals dataset is summarized in Table 2, with the probability density functions of dips and dip directions shown in Figure 5. Table 2 also shows the bias and dispersion for each cluster normalized to 90° for the dip and 360° for the dip direction.

Table 2. Bias and Dispersion when Down-Sampling.

	Cluster 1		**Cluster 2**		**Cluster 3**	
	μ (°)	**σ (°)**	**μ (°)**	**σ (°)**	**μ (°)**	**σ (°)**
Dip						
60%	0	8	−15	12	12	8
97.6%	0	9	−13	12	11	9
Dip Direction						
60%	−23	34	−36	39	9	81
97.6%	−25	36	−30	38	9	80
	Normalized for Dip and Dip Direction					
	μ (−)	**σ (−)**	**μ (−)**	**σ (−)**	**μ (−)**	**σ (−)**
Dip						
60%	0.00	0.09	−0.17	0.13	0.13	0.09
97.6%	0.00	0.10	−0.14	0.13	0.12	0.10
Dip Direction						
60%	−0.06	0.09	−0.10	0.11	0.03	0.23
97.6%	−0.07	0.10	−0.08	0.11	0.03	0.22

With respect to dip, the methodology produced relatively low bias across all three clusters in both down-sampled datasets. Cluster 1 had no bias (0°) in either dataset, indicating that its mean dip matched that of the reference dataset. Cluster 2 showed a consistent negative bias, −13° and −15°, suggesting a slight underestimation in the dip, while Cluster 3 had a positive bias that ranged from 11° to 12°, reflecting a slight overestimation in the dip. Dip dispersion values ranged from 8° to 12° across all clusters, showing moderate variability in the estimated orientations. These wider spreads are expected, as collecting UAV-based data will capture all surfaces, unlike manual measurements, which are limited to more accessible rock exposures.

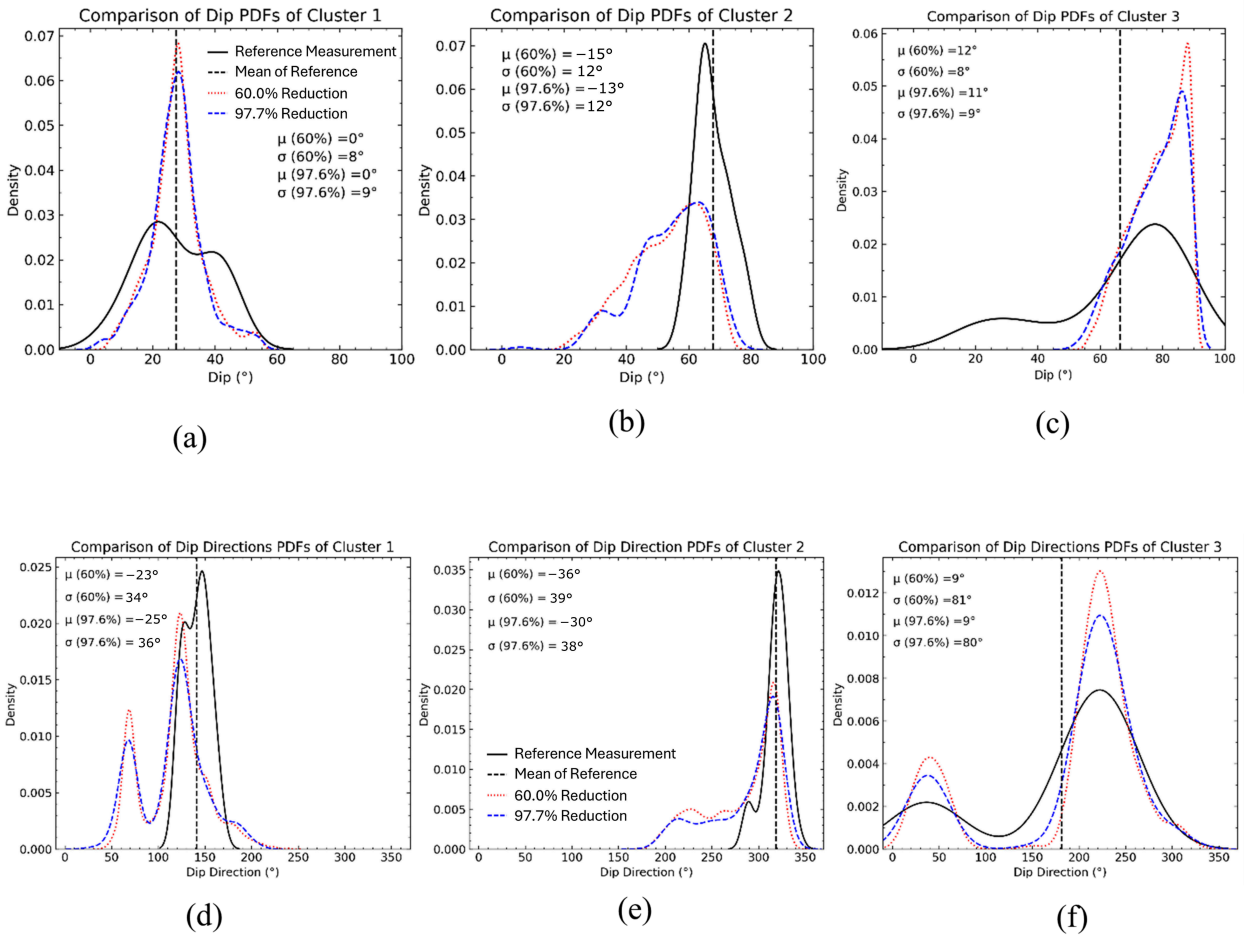


Figure 5. Comparison of dip and dip direction probability density functions for each joint set between the manual (reference) datasets and the down-sampled datasets: (**a**) comparison of dip for joint set 1, (**b**) comparison of dip for joint set 2, (**c**) comparison of dip for joint set 3, (**d**) comparison of dip direction for joint set 1, (**e**) comparison of dip direction for joint set 2, (**f**) comparison of dip direction for joint set 3.

The dip direction showed greater variability than dip. Cluster 3 had the lowest bias with 9° across both datasets but the largest dispersion (80–81°), which suggests that Cluster 3 was centered around the mean of the reference dataset but had higher variability in dip direction. Clusters 1 and 2 had larger biases in the range of −23° to −36°, indicating greater underestimation of dip direction, but had lower dispersions of 34° to 39°. Even though clusters 1 and 2 showed greater bias, their orientations were more tightly distributed than those of Cluster 3.

The 60% and 97.6% down-sampled datasets produced similar bias and dispersion values for both the dip and dip direction. Differences in bias between the datasets showed a maximum of 2° for dip and a maximum of 6° for dip direction. The differences in dispersion were a maximum of 1° for dip and a maximum of 2° for dip direction. These minimal differences between down-sampled datasets indicate that randomly reducing the size of the dataset does not significantly affect the identification of joint sets, supporting the robustness of this method across different data densities.

After the dataset was down-sampled, the next step in feature engineering was to randomly shuffle the dataset. Shuffling the dataset removes implicit bias created with the development of the dataset, which allows for a more robust methodology.

After the shuffling, the next step was to standardize the dataset. Standardization ($\overline{X}$) of the data was done using the mean and standard deviation (std) of the dataset through Equation (7):

$$\overline{X} = \frac{X - \text{mean}}{\text{std}}, \quad (7)$$

where X is a single point in the dataset. Standardizing creates a new dataset with a mean of zero and a standard deviation of 1. This is a useful feature engineering tool when more than one variable is used in the analysis, as it reduces the bias towards variables at different scales.

The methodology was tested on the facets and normals datasets with and without standardization on datasets with 60% and 97.6% of the data points removed (from the normals datasets only) to understand how standardizing affected the joint orientation classification results.

When standardizing both the facets and normals datasets, regardless of down-sampling, the joint set orientations were grouped into three clusters which had similar orientations to the manual dataset.

When standardization was not applied, the facets dataset produced three joint set orientations while the normals dataset, regardless of down-sampling, produced four joint set orientations. The fourth joint set in the raw dataset from the normals dataset had a median dip of 33° and a median dip direction of 333°, depicted in purple in Figure 6a.

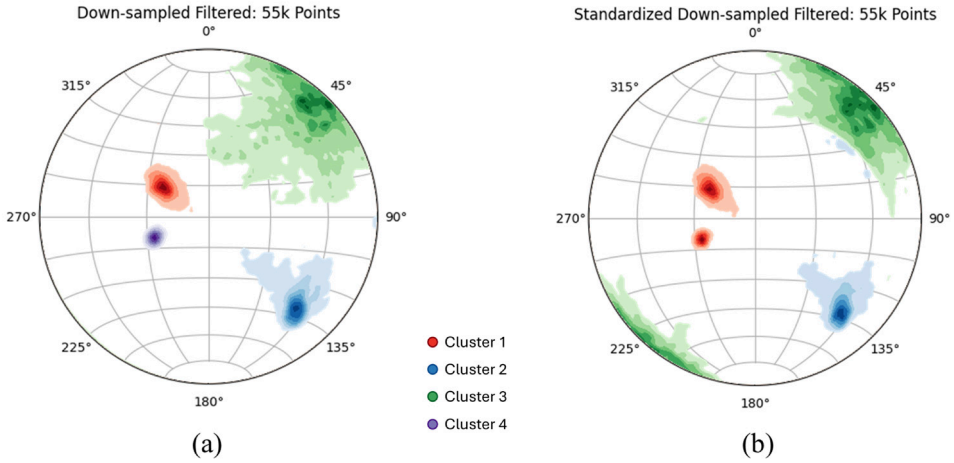


Figure 6. Comparisons between the methodology using raw data and the standardized data for the normals dataset showing (**a**) cluster variability using raw data and (**b**) cluster variability using standardized data.

Figure 6a,b also show the disappearance of the green cluster on the opposite side of the stereonet when standardization was not applied to the normals dataset, for all down-sampling datasets. It was assumed for this methodology that if data were on opposite sides of the stereonet, the data would be within the same cluster to account for natural variation in the joint orientation dip direction. The process of this rule is discussed in more detail in Step 4: Same Cluster Rule.

3.2. Step 2: Filtering

Filtering the dataset is an important step in this procedure as it prepares the dataset for identifying joint set orientations. This procedure uses a two-step filtering process. The first step is to use DBSCAN to filter the data based on density, defined by the optimized ε and minPts. For this study, the hyperparameters used for DBSCAN filtering are included in Table 3.

Table 3. Hyperparameters for DBSCAN Filtering.

Dataset	minPts	ε
Facets	4	0.045
Normals (60%)	4	0.015
Normals (97.6%)	4	0.06

Using a density-based filtering method is beneficial, as it assumes that a high-density area of data points represents a known joint surface while a low-density area of data points represents other surfaces at the site that do not repeat extensively such as vegetation or blocks that have been displaced.

The feature engineered dataset is fed into the DBSCAN algorithm to separate each point into a cluster or an outlier, as shown in Figure 7a. The data points that are colored were assigned clusters from DBSCAN while the gray data points were assigned as outliers or noise. The first stage of filtering removes all data points that are classified as outliers. This stage of filtering removes most of the random noise in the dataset that represents surfaces captured by the UAV that are not extensively repeating surfaces.

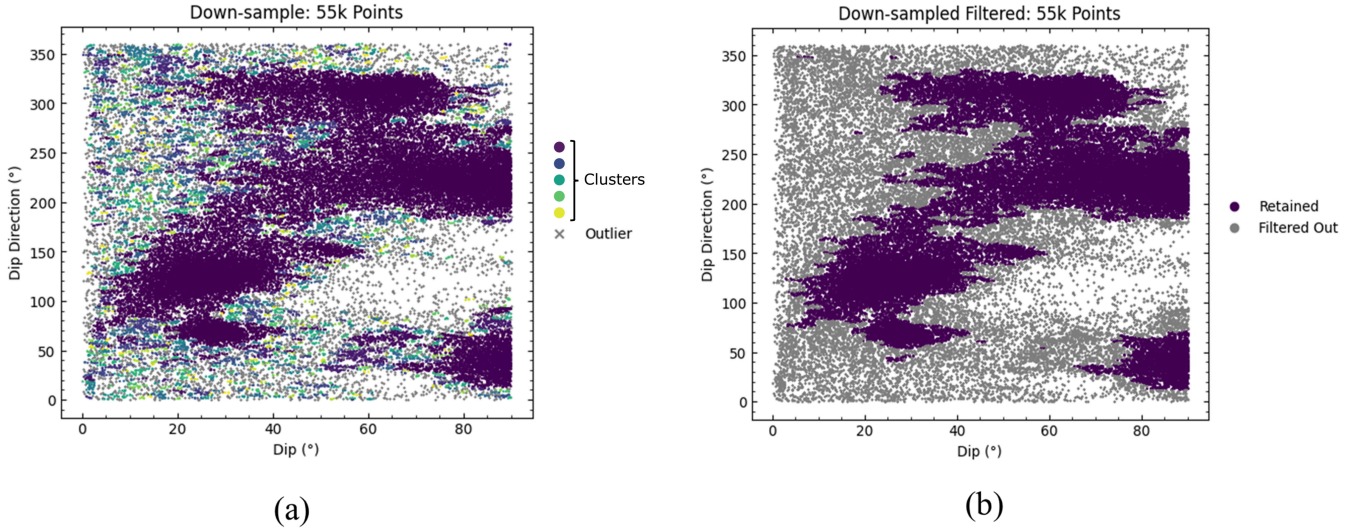


Figure 7. Two stages of filtering: (**a**) DBSCAN filtering removing the outlier points, shown in gray with clusters highlighted as colors and (**b**) second derivative filtering removing extraneous clusters, shown in gray with the retained clusters shown in purple.

The data points that are classified into a cluster that satisfy the minPts requirement in DBSCAN undergo the second stage of filtering. This stage of filtering is necessary because the DBSCAN filtering does not remove all the noise. The lower clusters from DBSCAN, shown in purple in Figure 7a, represent the higher density areas in the dataset. As a result, the lower numbered clusters are where the known joint orientation information is stored. To determine which clusters represent joint set orientation information and not noise, the number of data points in each cluster was summed and represented in a histogram, shown in Figure 8a.

The histogram in Figure 8a shows that the count of clusters exponentially decreases with increasing cluster number, with most of the data falling within the lower numbered clusters. In the case of Figure 8a, most of the data falls within clusters 1 through 15. Above cluster 15, very few data points are assigned to that cluster, representing a low density of data points and hence not a real joint set.

A threshold is needed to determine which cluster represents real joint information versus noise. The threshold chosen needs to represent where clusters have a high count compared to

where clusters have a low count. Many methods were explored to determine an appropriate threshold. The results of these thresholds are shown on a stereonet in Figure 9.

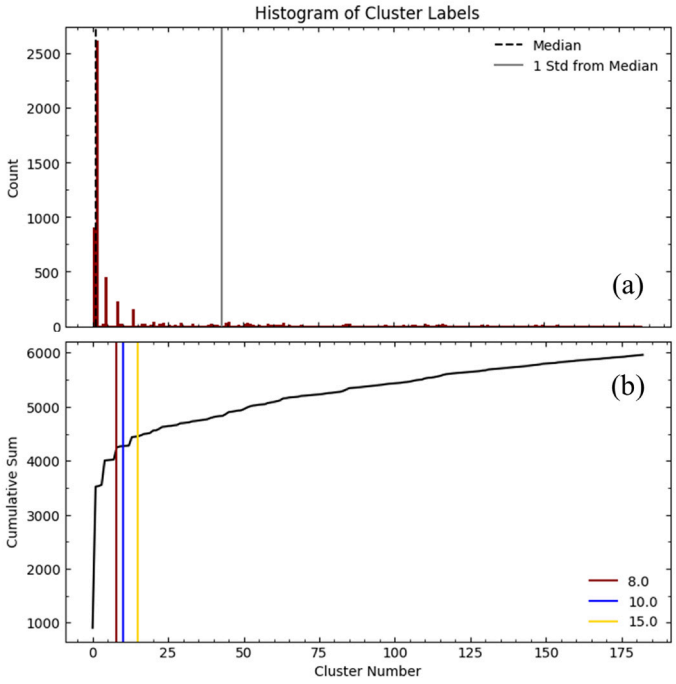


Figure 8. (**a**) Histogram of cluster numbers from DBSCAN shown as red bars and (**b**) cumulative sum of the histogram counts, shown in black, showing different elbow picks for noise threshold.

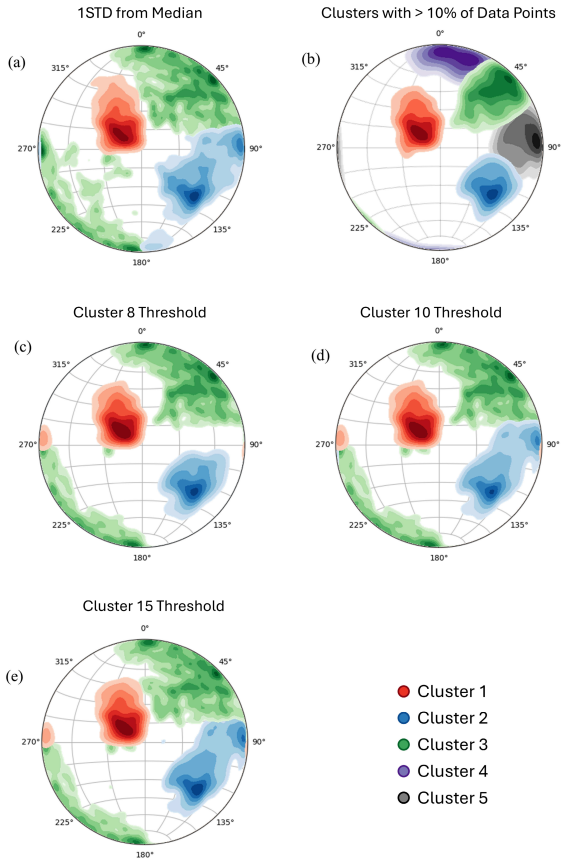


Figure 9. Stereonets showing clusters from (**a**) 1 STD from the median threshold, (**b**) a threshold that contained more than 10% of the cluster information, (**c**) cluster 8 threshold, (**d**) cluster 10 threshold, and (**e**) cluster 15 threshold.

The different thresholds that were looked at included the following:

- Using one standard deviation from the mean or median, shown in Figure 9a. This did not remove enough noise from the dataset, and when two standard deviations away from the mean or median were used, too much information was removed.
- Arbitrarily choosing a threshold based on counts in a cluster. For example, only clusters that contained more than 10% of all the data points would be considered real joint information. This method introduced bias and resulted in significant amounts of information loss, as shown in Figure 9b.

To determine the threshold, the cumulative sum of the counts of each cluster was calculated, shown as the solid black line in Figure 8b. The cumulative sum line is steep at low cluster numbers and then flattens out with increasing clusters. The threshold that separates real joint information from noise is chosen at the elbow, or the point of maximum change in slope of the cumulative sum as it separates clusters with high counts and low counts.

The elbow of the cumulative sum is found by calculating the second derivative of the cumulative sum line. Next, the angles between consecutive line segments are calculated. These angles represent the changes in slope of the cumulative sum line. Calculated angles that have high values represent significant changes in the slope of the cumulative sum line. These angles represent the elbow in the cumulative sum line and should be chosen for the threshold.

The cumulative sum line is made from discrete counts of clusters. As a result, when the angles are calculated between adjacent line segments, many angles with high values can emerge. Figure 8b shows three possible thresholds based on angles with high values. They are located at cluster number 8, cluster number 10, and cluster number 15. The thresholds are selected somewhat arbitrarily by the user; therefore, some judgment is needed.

Once the threshold is applied to the dataset, the remaining noise in the dataset is removed, as shown in Figure 7b. The data points that were retained are shown in purple, while the data points that were filtered out between the two stages of filtering are shown in gray. When compared to Figure 7a, the second stage of filtering removed the data points that were clusters with low-density points.

3.3. Step 3: Clustering

With the noise filtered out, the next step of the process is to identify the joint sets in the dataset. Joint sets are defined by dips and dip directions in the 2D space. As a result, dips and dip directions can be used to separate different joint sets. This was done in this procedure through the unsupervised clustering algorithm, K-Means Clustering.

K-Means Clustering first involves optimization of the hyperparameter, k, through the elbow rule or silhouette rule, as discussed in the Machine Learning Methods section. The dataset that has undergone feature engineering and filtering should be used to optimize k. The optimized hyperparameters for the facets and normals dataset are presented in Table 4. The optimized k, dip, and dip direction after the two-stage filtering process are used as the inputs into the K-Means Clustering algorithm.

Table 4. Optimized Hyperparameters for Clustering.

	k—Elbow Rule	k—Silhouette Rule
Facets Dataset	4	4
Normals Dataset (60% Reduction)	4	4
Normals Dataset (97.6% Reduction)	4	4

When the data are clustered, each data point will be assigned a numerical value ranging from 1 to k, in increments of 1. These values represent the cluster the data point is assigned to. The number of clusters the data will be sorted into is defined by the optimized k. These assigned clusters represent different joint sets found in the rock, as all points in a singular joint set have similar dip and dip direction properties.

3.4. Step 4: Same Cluster Rule

The clusters can be represented on a stereonet to graphically represent the dip and dip direction of each cluster, shown in Figure 10a, based on the clustering. Each joint set is represented by a different color. The first joint set is represented by the red points and has a mean dip of 28° with a standard deviation of 12° and a mean dip direction of 141° with a standard deviation of 13°. The second joint set, represented by the blue points, has a mean dip of 68° with a standard deviation of 6° and a mean dip direction of 319° with a standard deviation of 13°. The third joint set, represented by the green points, has a mean dip of 63° with a standard deviation of 22° and a mean dip direction of 222° with a standard deviation of 14°. A fourth joint set was identified and represented by the purple points. This joint set has a mean dip of 79° with a standard deviation of 5°, a mean dip direction of 037°, and a standard deviation of 6°.

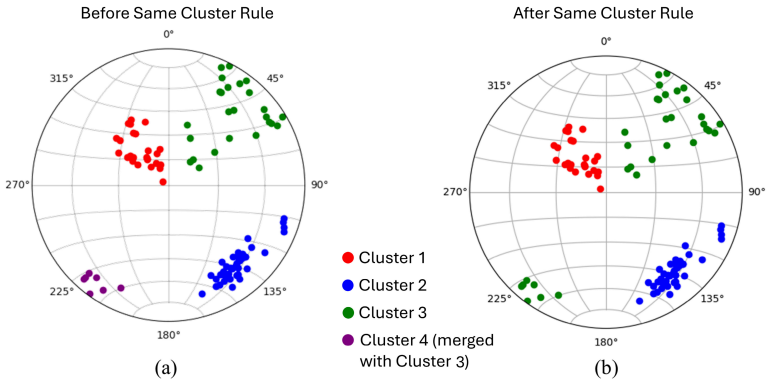


Figure 10. Same Cluster Rule comparison: (**a**) before the Same Cluster Rule was applied and (**b**) after the Same Cluster Rule was applied.

Due to the natural variation in dip and dip direction of joint sets in rock, a steeply dipping joint set may have some joint dip directions dipping in the opposite direction. These steeply dipping joint sets would appear around the edge of the stereonet, 180° apart in dip direction. This is evident in Figure 10a. Joint set 3 and joint set 4 are steeply dipping with dip directions roughly 180° apart. It is assumed that these two clusters are part of the same joint set. To account for this, a rule can be applied to merge the two clusters together. The steps to the Same Cluster Rule are outlined in Figure 11.

Clustered data allow us to calculate the median dip and dip direction of each grouping. For every combination of clusters, the dip directions will be subtracted to determine which are 180° apart. To account for natural variability in rock, the standard deviation of the dip direction is used to assess if the clusters are 180° apart. If the clusters are either less than 180° apart or greater than 180° apart, those two clusters are not combined.

If the two clusters are found to have dip directions within 180° of each other, the dips are checked to see if they are steeply dipping. It was assumed that joints dipping 65° or greater are dipping steeply. This cutoff was based on field experience at the site with the dipping discontinuities in the outcrop. The cutoff can be changed based on knowledge of the local geology.

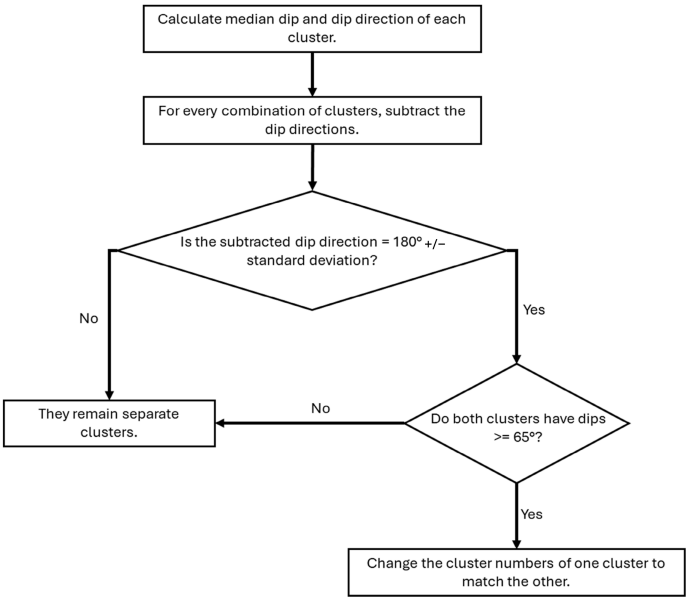


Figure 11. Same Cluster Rule procedure.

In the rule, if both clusters do not show steep dips, the clusters remain as separated joint sets. If both clusters were found to be steeply dipping, the clusters were combined into one, as it was assumed they represent the same joint set.

The stereonet from Figure 10a after the Same Cluster Rule was applied to the data is shown in Figure 10b. When compared to the original clustering, the cluster that represented a dip of 79° and a dip direction of 037° merged with the cluster with a dip of 63° and dip direction of 222°. A total of three joint sets were identified in the dataset that was collected manually. After the Same Cluster Rule is applied, the median and standard deviation of the dip and dip direction of each cluster can be used to describe the joint set orientations.

3.5. Step 5: Validation

To validate the methodology, the clustering results from the facets and normals dataset were compared to the dip and dip direction distributions from the manual dataset. The manual dataset was considered as the field reference dataset, as it contains known joint sets measured in the field. Comparing the results from the facets and normals datasets with the results from the manual dataset can be used to understand how well the joint set identification methodology worked. To understand how similar the distributions of dip and dip direction from the modeled and reference datasets were, the bias and dispersion were calculated using Equations (4) through (6). The comparison between the bias and dispersion for the manual dataset and the facets and normals dataset (97.6% of the data removed) is summarized in Table 5 with the probability density functions of dips and dip directions shown in Figure 12. Table 5 also shows the bias and dispersion for each cluster normalized to 90° for the dip and 360° for the dip direction.

There were varying levels of bias and dispersion across the three clusters when comparing the facets dataset and the normals dataset. Overall, when analyzing the dip of the joint sets, the facets dataset tends to have higher dispersion than the normals datasets, resulting in more variability in the dataset compared to the reference dataset. In contrast, the facets dataset had lower bias than the normals dataset, except for cluster 1. Even though in cluster 1 the normals dataset had the lower bias, there was only a 3° difference. This shows that the facets dataset provided more consistent dip estimates even with the larger variability.

Table 5. Bias and Dispersion for the Facets and Normals Datasets.

	Cluster 1		Cluster 2		Cluster 3	
	μ (°)	σ (°)	μ (°)	σ (°)	μ (°)	σ (°)
Dip						
Facets	3	14	−8	16	4	16
Normals (97.6%)	0	9	−13	16	11	9
Dip Direction						
Facets	−4	23	−22	24	−27	84
Normals (97.6%)	−25	36	−30	38	9	80
	Normalized for Dip and Dip Direction					
	μ (−)	σ (−)	μ (−)	σ (−)	μ (−)	σ (−)
Dip						
Facets	0.03	0.16	−0.09	0.18	0.04	0.18
Normals (97.6%)	0.00	0.10	−0.14	0.18	0.12	0.10
Dip Direction						
Facets	−0.01	0.06	−0.06	0.07	−0.08	0.23
Normals (97.6%)	−0.07	0.10	−0.08	0.11	0.03	0.22

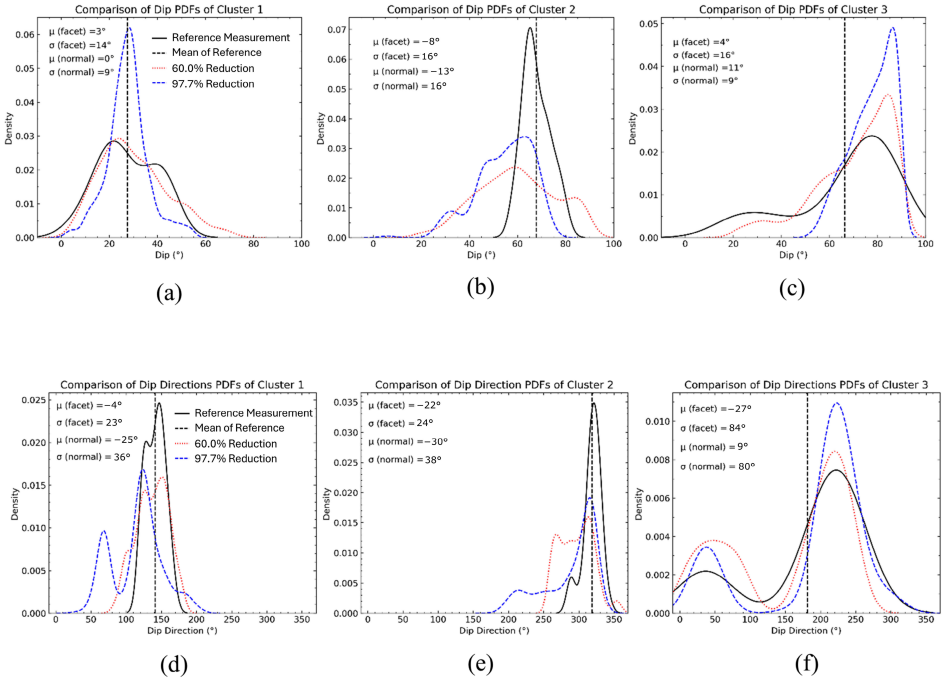


Figure 12. Comparison of dip and dip direction probability density functions for each joint set between the reference dataset and the facets and normals dataset: (**a**) comparison of dip for joint set 1, (**b**) comparison of dip for joint set 2, (**c**) comparison of dip for joint set 3, (**d**) comparison of dip direction for joint set 1, (**e**) comparison of dip direction for joint set 2, and (**f**) comparison of dip direction for joint set 3.

For the dip direction, the facets dataset tends to have lower bias and dispersion compared to the normals dataset, except for cluster 3. For cluster 3, the difference in dispersion is negligible with a difference of 4°, while the bias shows a difference of 36° between the facets and normals datasets. This could be attributed to cluster 3 having the most geometrically complex joint set with the highest variability. Cluster 3 is representative of the joint set with a mean dip of 66° with a standard deviation of 21° and a mean dip direction of 182° with a standard deviation of 78°. This is the cluster that is characterized by a steeply dipping joint set that was combined using the Same Cluster Rule described in Step 4.

The methodology used to identify the joint sets in both datasets demonstrated good performance across all clusters, as evidenced by consistently low bias and dispersion in both dip and dip direction. Natural joint sets inherently have substantial variation in orientation; as a result, some degree of bias and dispersion is expected. At the site, the manual dataset indicates a natural variation in dip and dip direction, respectively, of 12° and 13° for cluster 1, 6° and 12° for cluster 2, and 21° and 73° for cluster 3. The observed variation in dip and dip direction across all three joint sets is comparable to the degrees of bias and dispersion calculated in this study. This indicates that this newly developed proof-of-concept approach can reliably capture the orientation and variability of orthogonal joint sets in granitic outcrops.

4. Limitations and Future Work

The methodology proposed in this study was a proof-of-concept algorithm that showed how large datasets from LIDAR or UAVs can be processed to detect joint set orientations in rock outcrops. This study was conducted using only one site that contained orthogonal joint sets in granite.

Future directions for this study include applying this methodology to other geological settings with discontinuities. Using the different geologic settings stated below with this proposed methodology would provide an additional source of validation for the methodology.

The proposed methodology should be tested on geologic sites that are required to be mapped for stability analysis, including:

- Rock outcrops along roadways and railways.
- Shorelines.
- Tunnel walls and underground excavations.

This study should also be replicated in rock masses with variably oriented discontinuity sets such as:

- Sedimentary rocks with bedding and joint sets.
- Metamorphic rocks with wavy or banded foliation.
- Igneous rocks with exfoliation jointing or stress fractures.

Since the proof-of-concept methodology was tested for only one site, the current threshold for defining a steeply dipping discontinuity is based on the joint sets on the dam spillway. Recommended future work for this step would include compiling discontinuity orientation information from a variety of geologic sites under a variety of geologic conditions and performing a sensitivity analysis to choose an appropriate threshold for the majority of geologic conditions.

Other future work that can supplement this study include changes to the procedure developed including:

- Using other means to analyze the angular joint set orientation data such as repeating the study using the poles of the joint sets or Spherical K-Means Clustering.
- Developing a procedure to automatically select the DBSCAN filtering threshold for removing noise based on the second derivative of the cumulative sum of the count of data points in each DBSCAN cluster.

5. Conclusions

The dip and dip direction of joint sets are characterized using the machine learning procedures DBSCAN, taking the second derivative of the remaining clusters, and K-Means Clustering. The procedure starts with feature engineering, which involves down-sampling the dataset and standardizing the dip and dip directions. The size of down-sampling has

little effect on the robustness of the methodology to identify joint sets. Next, a two-stage filtering process using DBSCAN and the second derivative of the remaining clusters is used to remove noise in the dataset caused by surfaces that are not intact discontinuity planes. Finally, the procedure uses clustering, K-Means Clustering, to assign each data point to a cluster that is representative of a joint set in the outcrop. The number of clusters found corresponds to the number of joint sets in the rock. The bias and dispersion for the normals and facets datasets were calculated with respect to the manual dataset for each cluster. The bias and dispersion for the dip and dip direction for each joint set were comparable to the large variability observed in the manual field dataset, indicating that the probability density functions of the modeled datasets were unbiased. These results indicate that this proof-of-concept methodology is an efficient way to characterize orthogonal joint set data in granitic outcrops from large datasets collected by UAVs while removing the human bias accompanied with manual measurements in the field and increasing the representation of natural joint variability in nature.

Author Contributions: Conceptualization, B.M.R. and R.E.K.; methodology, B.M.R. and R.E.K.; software, B.M.R.; validation, B.M.R.; formal analysis, B.M.R.; investigation, B.M.R.; resources, R.E.K.; data curation, R.E.K. and B.M.R.; writing—original draft preparation, B.M.R.; writing—review and editing, B.M.R. and R.E.K.; visualization, B.M.R.; supervision, R.E.K.; project administration, R.E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and code presented in this study will be publicly available at the following GitHub repository: https://github.com/b1russo/RockMassCharacterization (accessed on 27 August 2026).

Acknowledgments: The authors are grateful to everyone at PG&E who supports the continual education of UC Berkeley Geosystems Engineering Master students as they learn how to take strike and dip measurements of rock outcrops. Without their continual support, the data collection for this study would not have been possible. The authors would also like to acknowledge the UC Berkeley MS Class of 2022 for aiding in taking hand measurements of the rock to be used in this analysis.

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

Abbreviations

The following abbreviations are used in this manuscript:

LIDAR	Light detection and ranging
RQD	Rock quality designation
UAV	Unmanned aerial vehicle
SfM	Structure from motion
DBSCAN	Density-based spatial clustering of applications with noise
TIN	Triangulated irregular network
E	Euclidean distance
WCSS	Within-cluster sum of squares
S	Silhouette average

References

1. Kemeny, J.; Post, R. Estimating three-dimensional rock discontinuity orientation from digital images of fracture traces. *Comput. Geosci.* **2003**, *29*, 65–77. [CrossRef]
2. Zhang, L.; Inokoshi, M.; Batuk, M.; Hadermann, J.; Naert, I.; Van Meerbeek, B.; Vleugels, J. Chapter 7—Strength. In *Engineering Properties of Rocks*, 2nd ed.; Butterworth-Heinemann: Oxford, UK, 2017.
3. Bieniawski, Z. *Engineering Rock Mass Classifications: A Complete Manual for Engineers and Geologists in Mining, Civil and Petroleum Engineering*; John Wiley & Sons: New York, NY, USA, 1989.

4. Coe, A.L. *Geologic Field Techniques*; Wiley-Blackwell: Milton Keynes, UK, 2010.
5. Hoek, E.; Brown, E.T. *Underground Excavations in Rock*; CRC Press: Boca Raton, FL, USA, 1990.
6. Priest, S.; Hudson, J. Estimation of discontinuity spacing and trace length using scanline surveys. *Int. J. Rock Mech. Min. Sci.* **1981**, *18*, 183–197. [CrossRef]
7. Slob, S.; van Knapen, B.; Hack, R.; Turner, K.; Kemeny, J. Method for Automated Discontinuity Analysis of Rock Slopes with Three-dimensional Laser Scanning. *Transp. Res. Rec.* **2005**, *1913*, 187–194. [CrossRef]
8. Tiruneh, H.; Stetler, L.D.; Oberling, Z.A.; Morrison, D.R.; Connolly, J.L.; Ryan, T.M. Discontinuity mapping using Ground-Based LiDAR: Case study from an open pit mine. In Proceedings of the 47th US Rock Mechanics/Geomechanics Symposium, San Francisco, CA, USA, 23–26 June 2013.
9. Kayen, R.; Russo, B.M. Computational-Rock Mechanics in Pedagogy and Practice. In Proceedings of the 58th US Rock Mechanics/Geomechanics Symposium, Golden, CO, USA, 23–26 June 2024.
10. Kong, D.; Saroglou, C.; Wu, F.; Sha, P.; Li, B. Development and application of UAV-SfM photogrammetry for quantitative characterization of rock mass discontinuities. *Int. J. Rock Mech. Min. Sci.* **2021**, *141*, 104729. [CrossRef]
11. Mineo, S.; Calio, D.; Pappalardo, G. UAV-Based Photogrammetry and Infrared Thermography Applied to Rock Mass Survey for Geomechanical Purposes. *Remote Sens.* **2022**, *14*, 473. [CrossRef]
12. Herrero, M.J.; Perez-Fortes, A.P.; Escavy, J.I.; Insua-Arevalo, J.M.; De la Horra, R.; Lopez-Acevedo, F.; Trigos, L. 3D model generated from UAV photogrammetry and semi-automated rock mass characterization. *Comput. Geosci.* **2022**, *163*, 105121. [CrossRef]
13. Mammoliti, E.; Di Stefano, F.; Fronzi, D.; Mancini, A.; Malinverni, E.S.; Taziolo, A. A Machine Learning Approach to Extract Rock Mass Discontinuity Orientation and Spacing, from Laser Scanner Point Clouds. *Remote Sens.* **2022**, *14*, 2365. [CrossRef]
14. Liu, Y.; Hua, W.; Chen, Q.; Liu, X. Characterization of Complex Rock Mass Discontinuities from LiDAR Point Clouds. *Remote Sens.* **2024**, *16*, 3291. [CrossRef]
15. Pola, A.; Herrera-Diaz, A.; Tinoco-Martinez, S.R.; Macias, J.L.; Soto-Rodriguez, A.N.; Soto-Herrera, A.M.; Sereno, H.; Avellan, D.R. Rock characterization, UAV photogrammetry and use of algorithms of machine learning as tools in mapping discontinuities and characterizing rock masses in Acoculco Caldera Complex. *Bull. Eng. Geol. Environ.* **2024**, *83*, 260. [CrossRef]
16. CloudCompare. *CloudCompare [GPL Software]*, version 2.6.2; CloudCompare: Grenoble, France, 2016. Available online: https://www.cloudcompare.org/ (accessed on 8 November 2023).
17. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996.
18. Hahsler, M.; Piekenbrock, M.; Doran, D. dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **2019**, *91*, 1–30. [CrossRef]
19. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. [CrossRef]
20. Aranganayagi, S.; Thangavel, K. Clustering categorical data using silhouette coefficient as a relocating measure. In Proceedings of the IEEE International Conference on Computational Intelligence and Multimedia Applications, Las Vegas, NV, USA, 13–15 December 2007.
21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. Kington, J. Mplstereonet 0.6.3, 2023. Available online: https://pypi.org/project/mplstereonet/#description (accessed on 14 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.