The Report Committee for Sareh Kouchaki
Certifies that this is the approved version of the following report:

Clustering Pavement Aggregate Particles Based on Shape and Texture Properties

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor: ______________________________
Peter Mueller

____
Jorge Prozzi

Clustering Pavement Aggregate Particles Based on Shape and Texture Properties

by

Sareh Kouchaki

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin
December 2017

Abstract

Clustering Pavement Aggregate Particles Based on Shape and Texture Properties

Sareh Kouchaki, M.S. Stat

The University of Texas at Austin, 2017

Supervisor: Peter Mueller

Aggregates are the major component of pavements. Physical characteristics of aggregates significantly affect the properties of pavements. Different pavement construction projects may require different characteristics of aggregate. Proper selection of aggregate with consistent shape properties ensures high performance of pavements. The available test methods for evaluating the aggregate physical properties and classifying them are laborious, time-consuming, and subjective. This study presents the development of an objective system which evaluates the shape properties of aggregate particles and classifies them into distinct groups regarding their sphericity, form, angularity and texture features. By using this system, the heterogeneity in an aggregate sample based on a given feature could be assessed. This system includes a laser scanner developed at the University of Texas at Austin to scan aggregate particles. Total of 1398 aggregate particles, from eight different quarries in the state of Texas, were scanned. The scanned data were analyzed using a MATLAB algorithm for measuring the sphericity, form, angularity, and texture of particles. All the measurements were stored in an Excel file and were imported to another algorithm developed in R software and OpenBUGS package to cluster the aggregate particles. Several methods of clustering were reviewed and finally, model-based clustering approach was selected. The model-based cluster analysis was applied to the measurements aiming to detect subclasses in aggregate particles based on each feature. This study shows how to use this clustering approach to group the particles based on their sphericity, form, angularity, and texture features.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Aggregate is a key component in asphalt and concrete pavements. Asphalt and concrete pavements consist of about 80% and 90% of aggregate, accordingly. Therefore, it is very important that the engineer takes careful consideration of the properties of the aggregates. From a pavement engineering point of view, the parameters of shape and surface of aggregates play a significant role on the performance of concrete and asphalt pavements because of the way they affect the interactions and bonds between aggregates and binder (Sun 2014).

Literature have shown that different pavement mixtures require aggregates with various physical characteristics (Herrin et al. 1954, Field 1958, Meininger 1998). For instance, in order to increase the fatigue life and stiffness of the thick pavements, it is recommended to use the rough textured aggregates. On the other hand, for thin pavements, it is recommended to use smooth textured aggregates to have a less stiff mixture to increase the fatigue life (Monismith 1970). It is very important to use aggregate particles with desirable properties in a pavement project. Poorly selected aggregates can cause early deteriorations of pavement structure. However, selecting proper aggregate will ensure the pavement performance.

For pavement construction, it is desired to use the local materials due to the economical and logistical considerations. But, one should consider that all the produced materials out of one quarry might not be consistent in terms of shape and texture properties. Therefore, more attention is needed when evaluating the materials extracted from a quarry. Currently, transportation agencies use different methods, such as ASTM D4791 and ASTM D5821, to evaluate the aggregate shape characteristics to select desirable ones for a project; however, these methods are subjective, laborious, and time-consuming. An improved methodology being consistent, repeatable, and objective in evaluating the aggregates based on shape and texture properties can benefit transportation agencies and industries. A classification system is required which puts aggregate particles in distinct clusters based on the shape and texture properties. One of the advantages of a classification system is that the uniformity of the aggregates in a pavement project could be controlled with respect to a desired feature. In addition, this system can help engineers to select the appropriate type of aggregate for different applications to improve the pavement performance and decrease the maintenance cost.

GOAL AND OBJECTIVES

The goal of this study is to develop a classification framework by using the statistical clustering analysis that can evaluate a sample of aggregate particles based on their shape and texture properties. To develop this framework, an automated measurement system of aggregate shape properties is required. In recent years, studies have turned to laser and image technologies to measure the aggregate shape properties more accurately and quickly. These studies showed that these new technologies could provide reliable and precise measurements. Accordingly, this study uses a laser scanning tool developed at the University of Texas at Austin. The following steps were accomplished to fulfill the purpose of this study. First, a literature review on aggregate shape evaluation methodologies was conducted. Then, samples of aggregate particles in varied sizes from different sources were selected and scanned to collect data. A computer algorithm was developed to measure the aggregate particles features followed by studying different common clustering methods and choosing the appropriate method to group the scanned aggregates based on the shape and surface features.

BACKGROUND

Characterization of Aggregate Shape Properties

Researchers have defined three different geometric properties of aggregate. These properties are independent and are able to describe the physical characteristics of a particle sufficiently. The properties are including the shape (or form), angularity (or roundness), and surface texture (Barrett 1980). Figure 1 shows these properties on a particle schematically. The disparities in the proportion of aggregate particles are represented by the shape factor. The variations on the particle corners are reflected by the factor of Angularity. The small-scale surface irregularities, that do not affect the shape of the particle, are reflected by surface texture.

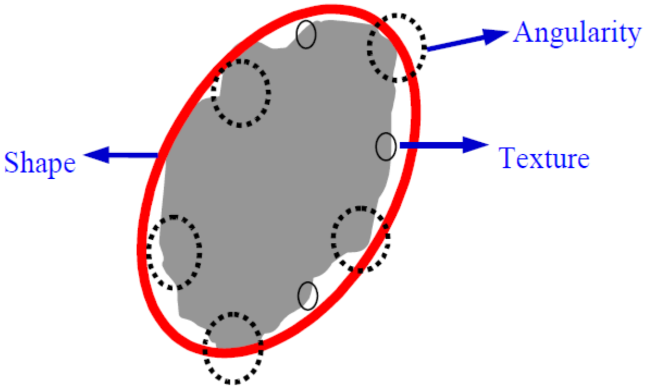


Figure 1: Schematic of the Aggregate Properties: Shape, Angularity, and Texture (Masad et al. 2003).

Recent advances in image-based analysis and laser scanning techniques have led to feasible and cost-effective methods for measuring aggregate shape properties. In an effort, Masad developed the Aggregate Imaging System (AIMS), which is capable of analyzing the shape properties of coarse and fine aggregates. He developed several indices for measuring form, sphericity, and angularity of aggregate particles. In this study, the shape indices developed by Masad in 2003 were used. To characterize the texture of aggregate particles the index provided by ASME B46 was considered. These indices are explained further in the following:

SPHERICITY INDEX

Masad in 2003 proposed the computation of sphericity factor using the Equation 1. Sphericity is a parameter based on the three-dimensional analysis which describes the overall form of an aggregate particle (Masad 2003).

$$\text{Sphericity} = \sqrt[3]{\frac{d_S.d_I}{{d_L}^2}} \qquad \text{Eq.1}$$

Where:

d_L = The Longest dimension of an aggregate particle;

d_I = The Intermediate dimension of an aggregate particles, and

d_s = The Shortest dimension of an aggregate particle.

Equation 1 shows that, as the form of an aggregate particle becomes spherical, the sphericity number reaches 1, whereas the sphericity of flat particles reaches 0.

FORM INDEX

This index (defined in Equation 2) was developed based on the two-dimensional measurement (2D image) of an aggregate. The form of an aggregate is calculated using the sum of incremental change in the aggregate particle radius (Masad 2003). The graphical representations of the form computations are illustrated in Figure 2.

$$\text{Form Index } = \sum_{\theta=0}^{360-5} \frac{|R_{\theta+5} - R_{\theta}|}{R_{\theta}} \qquad \text{Eq.2}$$

Where:

R = Radius of an aggregate particle at a given direction;

θ = Directional angle.

From the Equation 2 for round aggregates, the form index becomes zero since there is no change in radius.

ANGULARITY INDEX

Masad (2001) proposed a method based on the two-dimensional analysis to measure the angularity of an aggregate. In this method the difference between the radius of the particle at a given angle and that of an equivalent ellipse is calculated and summed over different angles. The mathematical computation of this index is provided in Equation 3. The graphical representations of the form and angularity computations are illustrated in Figure 2. It is to be noted that the equivalent ellipse has the same major and minor axes as the particle, but has no angularity.

$$AI = \sum_{\theta=0}^{355} \frac{R_{\theta} - R_{EE\theta}}{R_{EE\theta}} \qquad \text{Eq.3}$$

Where:

AI: Angularity Index

R_{θ}: Radius of the particle at an angle of θ

$R_{EE\theta}$: Radius of the equivalent ellipse at an angle of θ

According to the angularity index, particles with angular corners must have higher values of angularity index compared to well-rounded particles (Masad 2003).

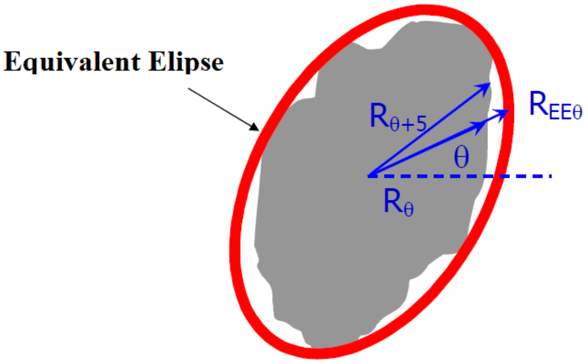


Figure2: Illustration of the radius of the particle and equivalent ellipse used in the Form and Angularity computation (Masad et al. 2003).

TEXTURE INDEX

Advanced technologies such as non-contact laser scanners allow direct measurement of the texture profiles with higher resolution. In a research study conducted by Kouchaki et al. it was reported that the developed LLS prototype showed promising results regarding scanning the surface texture of aggregates (Kouchaki et al. 2017). The collected data from the developed LLS prototype can be used to compute various profile statistics, such as slope variance (SV). Equation 4 represents the slope variance calculation method (Mora 2003). This index shows how the height values change along a profile. As shown in Figure 3, a profile that fluctuates widely results in the high value of SV and a profile which amplitudes change slowly regarding space gives a low value of SV (ASME B46.1). The SV can be computed for several profiles on the surface of an aggregate and subsequently, the mean of them could provide a criterion to compare the texture of aggregate particles.

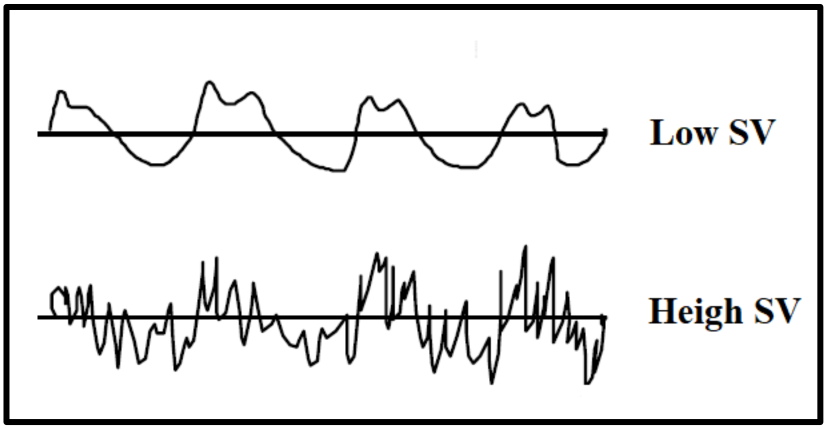


Figure 3: Example of profiles with low SV and high SV.

$$SV = \sqrt{\frac{1}{N-1}(\sum_{i=1}^{N-1}(S_i)^2 - \frac{1}{n}(\sum_{i=}^{N-1} S_i)^2} \qquad \text{Eq.4}$$

Where:

SV: Slope Variance

N: Number of points on height profile

S_i: Slope between i+1th and ith points

METHODOLOGY

Based on the literature, the aggregate shape features are significant factors affecting the pavement performance. In this study, attempts were made to develop a reliable framework to measure aggregate shape features and classify them based on those features. Several steps were considered for establishing this framework. First, aggregate particles from different quarries and various sizes were collected. In the second step, a laser scanner prototype developed at the University of Texas at Austin was used to scan the aggregate particles. This is followed by developing a MATLAB algorithm to measure the aggregate shape features from the collected scan data. As the last step, a clustering algorithm was created to group aggregate particles upon the extracted features. These steps are explained in more details in the following sections.

Data Collection

In this study, the aggregate of eight quarries which are commonly being used for asphalt mixtures in Texas were selected. These aggregates were provided by Texas Department of Transportation (TxDOT) Construction Division in Austin, TX. Total of 1398 particles, including two hundred particles in different range of sizes (greater than 2 mm) from each quarry, were randomly selected and prepared for examination. The aggregates were first washed in order to remove any dust and undesirable particles. Then, the aggregate particles were oven-heated to 160°C for 24 hours followed by four hours of regulated air temperature and humidity to reach air-dry condition.

A 3-D laser scanner prototype (Figure 4), called LLS, was utilized to scan the selected aggregates. As shown in Figure 4, this prototype constitutes a line laser scanner (LLS) and a motion controller with which the LLS can move over an aggregate particle and scan it. This device is a non-contact laser sensor that projects blue light in a horizontal line. The blue light is emitted from a source in the LLS and then the reflection of the light from the surface of aggregate is captured by a detector. If the angles between the projected light and the reflection are known then, by using triangulation, the system can calculate the height profile of the surface relative to the system's reference line. Small changes in the height due to the texture irregularities can be captured using this scanner system (KEYENCE 2017). Since the mechanism of the laser is based on projecting a light and capturing its reflection, all tests were performed in a laboratory with constant light condition to avoid any noise associated to the light variation. Along with the light, the room temperature and humidity were also kept constant during the experiments. The LLS is connected to a computer and the scanned data are collected in the computer in Excel. The collected data is used as an input to another algorithm to measure the aggregate features.

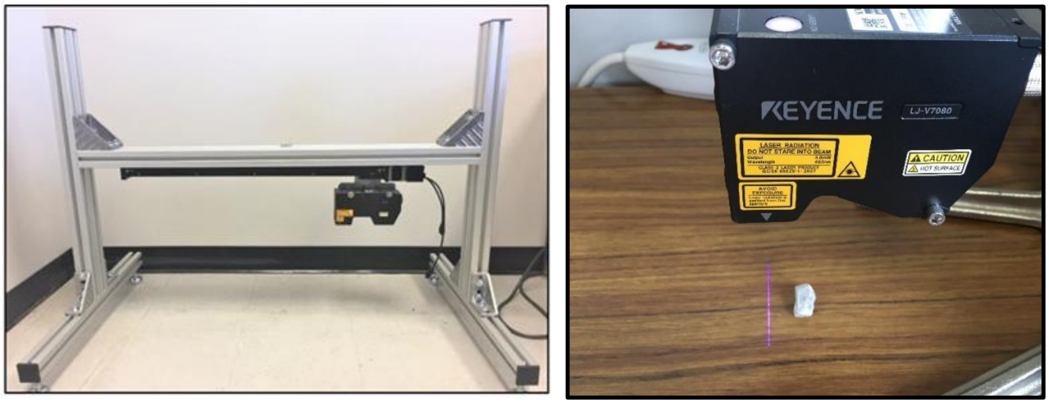


Figure 4: Left) The LLS prototype including the frame and the mounted laser, Right) the laser lane and an aggregate particle in the scanning area.

Feature Extraction

The data collected by the LLS prototype were analyzed in the MATLAB software to extract the shape features of the particles. Before extracting the shape features, the collected data were examined for noise or missing values. As mentioned previously, the LLS scanned the particles based on the triangulation system. Due to this system of scanning (which is shown in Figure 5) the reflection of the laser line from some areas around the edge of aggregate particles cannot be seen by the camera and the data are missed. Before extracting the aggregate features, these missing data (the white areas in Figure 5) must be taken care of. In this case, linear interpolation was used to substitute missing data. Then, the scanned data were pre-processed such that in each scan data, data associated with the aggregate particles were separated from the background and prepared for feature analysis.

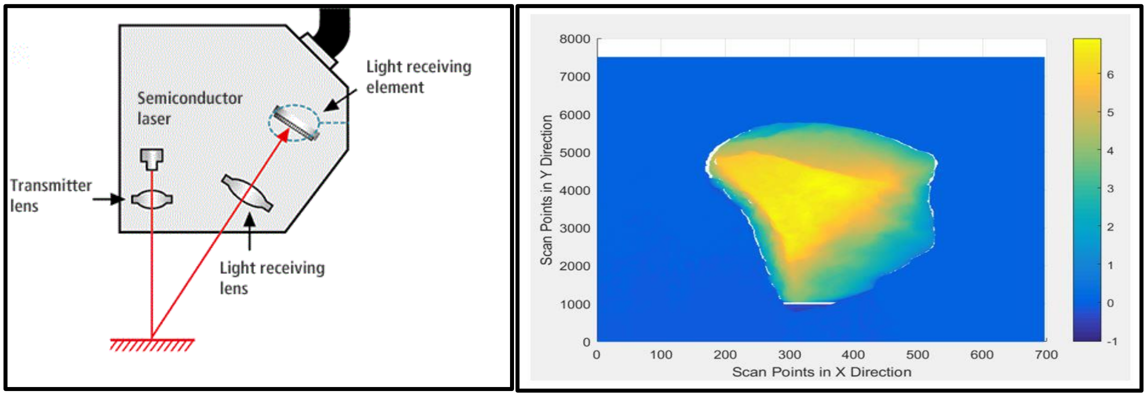


Figure5: Left) The triangulation system, Right) Scan result of an aggregate particle including the missing area.

Regarding the Equations, 1 through 3, three dimensions, length, width, and thickness of an aggregate particle are required to calculate the shape indices. At this step, a MATLAB algorithm was developed to measure the particles dimensions and calculates the four features All the features data were collected in an excel sheet and used for clustering the aggregate particles.

Clustering Analysis

Clustering analysis divides data into distinct groups such that quite similar observations are placed in one group, while different observations are placed in separate groups. Therefore, the structure of clusters is a set of groups of observations with high-within-group similarity and low-between-group-similarity. It should be noted that this segmentation is only based on the information found in data and there are no given labeled classes (James et al. 2013).

CLUSTERING ANALYSIS METHODS

Clustering analysis is popular in different areas such as biology, medicine, business, marketing, etc. Researchers are usually curious about finding subgroups in their observations to better understand the heterogeneity in their data. Because of this popularity, different methods have been developed to do the clustering. For example, hierarchical clustering is a widely used approach. This method starts with considering each observation as one cluster. Then it merges similar observations to reduce the number of clusters (James et al. 2013, Tan 2005). For example, Euclidean distance concept is used to seek the similarity among observations. This measurement of the similarity is defined in the Equation 5. The smaller the value of D_{ij}, the closer the observations i and j.

$$D_{ij} = \sqrt{\sum_{q=1}^{Q}(X_{qi} - X_{qj})^2} \qquad \text{Eq.5}$$

Where

D_{ij} : distance between observation i and j

Q: Number of features in multi-dimensional observations

X_{qi} : value of feature q for observation i

X_{qj} : value of feature q for observation j

In the next step, hierarchical clustering locates similar clusters to combine them. To do so, the similarity between two clusters needs to be defined. Three similarity criteria can be considered: single linkage, average linkage, or complete linkage. These three criteria are graphically depicted in Figure 6. Pairwise Euclidean distance, which is the distance between every member of clusters, is the basis of these three criteria. In the single-linkage, the minimum pairwise Euclidean distance is considered for merging. In the average linkage, the average of all computed pairwise distances is the factor to determine the similarity. For the complete linkage, the maximum pairwise distance indicates the similarity between two clusters (James et al. 2013, Tan 2005).

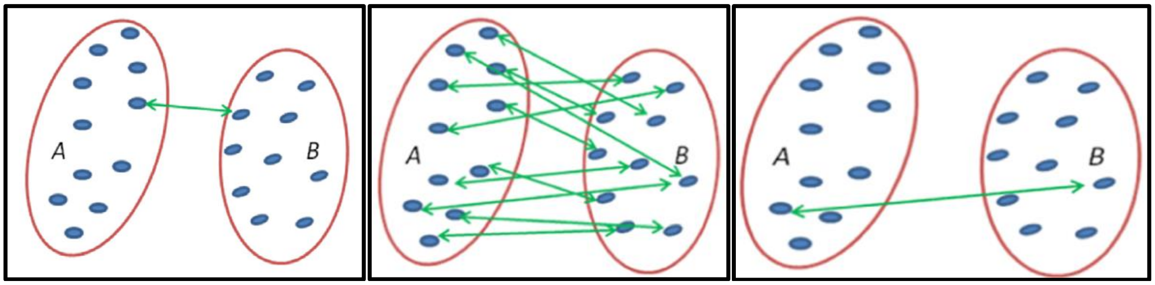


Figure 6: Left: single linkage, middle: average linkage, and right: complete linkage. (Saxena 2017).

The result of hierarchical clustering can be displayed as a tree and is called dendrogram (James et al. 2013, Tan 2005). An example of the dendrogram resulted from clustering of nine observations is provided in Figure 7. In the first level, all nine observations were separate clusters. In the next level, the hierarchical approach found out that X9 and X4 are similar to each other and merged those together. In parallel, X3 and X5 were combined because of their high similarity. In the next step, the cluster X1 was identified similar to the cluster resulted from the combination of X9 and X4. This process was continued to finally all observations combined in one cluster. Note that the clusters those are merged at a lower level of the dendrogram are much more similar than those are merged at higher levels. Using the dendrogram and based on the level of similarity, one can decide on the number of clusters in a dataset.

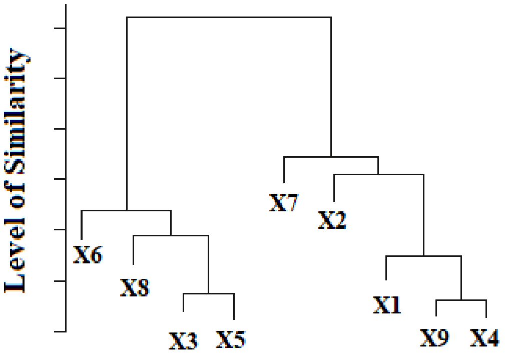


Figure 7: An example of dendrogram from hierarchically clustering of nine observations.

Ward’s clustering is another type of hierarchical clustering methods. The difference between Ward’s method and other hierarchical clustering methods is in their similarity criteria. Rather than using the Euclidean distance to merge the clusters, Ward’s approach uses the within-cluster sum of squares to combine two clusters (Shalizi 2009). For example, as shown in Equation 6, to combine two clusters A and B, the within-cluster sum of squares of each cluster individually and the within-cluster sum of squares of their combination are required. This method combines those two clusters with the minimum $W_{A,B}$.

$$W_{A,B} = \sum_{x_i \in A}(x_i - C_A)^2 + \sum_{x_i \in B}(x_i - C_B)^2 - \sum_{x_i \in AB}(x_i - C_{AB})^2 \qquad \text{Eq.6}$$

Where

$W_{A,B}$: Ward's criteria

x_i: ith observation

C_A: centroid of cluster A

C_B: centroid of cluster B

C_{AB}: centroid of cluster AB

* The centroid is the vector of feature means in a cluster.

K-mean clustering is another commonly used method that divides a dataset into k different and non-overlapped clusters. In contrast with hierarchical approaches, the number of clusters (k) must be specified at the beginning of analysis (James et al. 2013). Figure 8 graphically represents the steps of K-mean clustering algorithm. In the first step, the observations are randomly assigned to k clusters. In the next step, the centroid of each cluster is computed. The centroid of each cluster is a vector of the feature means of the multidimensional observations in that cluster. In step 3, each observation is reassigned to a cluster with the closest centroid (based on Euclidean distance) and the centroid of clusters is updated. These four steps are repeated until a steady result is obtained.

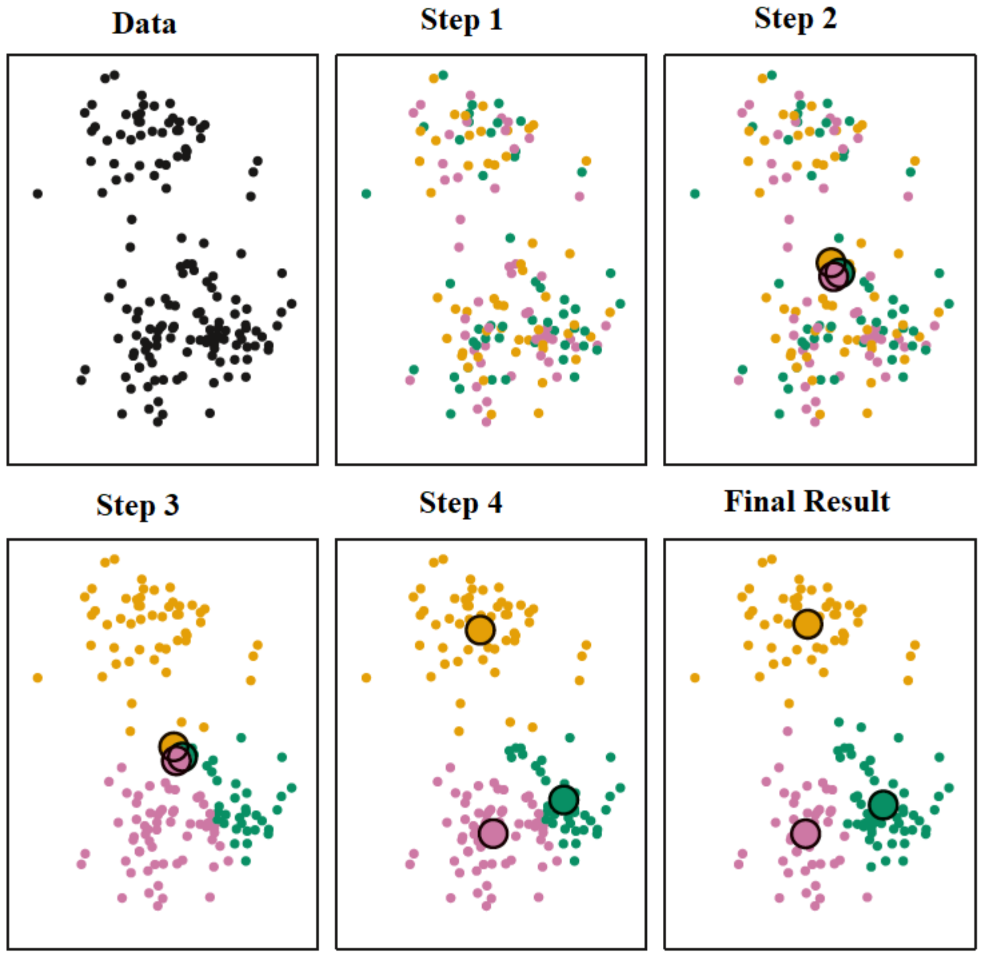


Figure 8: Graphically illustration of the K-mean clustering steps for k=3 (James et al. 2013).

DETERMINISTIC ALGORITHMS VS. PROBABILITY MODEL BASED CLUSTERING

Hierarchical and K-mean clustering techniques are deterministic algorithms. The uncertainty of clustering cannot be evaluated through these algorithms. A model-based approach is a suitable alternative way. This clustering method employs statistical concepts and views the observations as random variables generated from a probability distribution. Next section explains this clustering approach in more details (Chamroukhia 2015, Faranzen 2006).

MODEL-BASED CLUSTERING

Assume $y = \{y_1, \ y_2, \ldots, y_n\}$ is a sample of observations of size n, where $y_i \ (for \ i \ = 1,2, \ldots, n)$ is a q-dimensional variable. Model-based clustering assumes that each observation is drawn from a probability distribution. The density of data is a mixture of those probability distributions with possible random number of term K. Based on this analysis, the probability distribution of observations is expressed as the Equation 7:

$$f(y|\psi) \ = \ \sum_{k=1}^{K} w_k f_k(y|\theta_k) \qquad\qquad i = 1,2, \ldots, n \qquad\qquad \text{Eq.7}$$

Where

y: q-dimensional observations

ψ: mixture model parameter, $\psi \ = \ (\theta_1, \theta_2, \ldots, \theta_k, w_1, w_2, \ldots, w_k)$

w_k: mixing proportion or weight of each distribution such that $0 < w_k < 1$ and $\sum_{k=1}^{K} w_k \ = 1$

$f_k(y|\theta_k)$: mixture components

K: number of components

θ_k: parameter vector associated to f_k

Most often, normal distributions are used as the mixture components. Each normal distribution has its own mean (μ) and variance (ε). Accordingly, in the Equation 7, the parameter vector θ_k represents a vector of the mean (μ_k) and variance (ε_k). It is to be noted that each distribution represents a cluster of data where the centers of these clusters are defined by μ_k, and their shapes and sizes are described by ε_k. The ε_k can be considered either the same, or different across all components (Zhihui 2010, Faranzen 2008, Schafer 2015, Fraley 2007).

The Equation 7 can alternatively be written as a hierarchical model:

$$\begin{cases} y_i|\psi\,, c_i = k \;\sim\; f_k(y|\theta_k) \\ \quad p(c_i = k|w) = w_k \end{cases} \qquad \text{Eq.8}$$

Equation 8 implicitly defines a random variable of $c_i = k$ to the parameters as the cluster membership indicator where $c_i = k$ implies that the observation y_i is classified into cluster k. Therefore, the probability that y_i is in cluster k is equal to the mixing weight of that cluster, $p(c = k)$ = w_k (Diebolt 1994).

The complete collection of all parameters μ_k, ε_k, and w_k is presented by ψ. This parameter vector is required to be estimated. Different methods have been developed to estimate these parameters. For instance, the method of the moments is one of the earliest methods that was used in this regard. Nowadays, maximum likelihood estimator (MLE) and Bayesian estimator are common methods. The MLE considers the unknown parameters as fixed variables and finds their point estimates by maximizing the log likelihood function of parameters. The MLE is not necessarily a good estimator for parameters since it might capture the local maximum instead of the global maximum. In addition, this method only provides point-estimates of parameters without any estimation about the uncertainty of the parameters. However, Bayesian estimator views the unknown parameters as random. This estimator generates a probability distribution for each unknown parameter through which not only a point-estimate is obtained but also the uncertainty of the estimation could be studied (Faranzen 2008, Hoff 2009).

BAYESIAN INFERENCE

Bayesian estimation is implemented based on the Bayes theorem (as shown in Equation 9). According to this theorem, there is an initial guess on the probability distribution of the parameter ψ which is called prior distribution, $p(\psi)$. The prior reflects our knowledge of the parameter before observing the data. This prior distribution is updated after observing the data and turns into a new probability distribution which is called posterior distribution, $p(\psi|y)$ (Faranzen 2008, Heller 2007).

$$p(\psi|y) = \frac{p(\psi)p(y|\psi)}{p(y)} \qquad \text{Eq.9}$$

Where

$p(\psi)$: prior probability of ψ

$p(y|\psi)$: likelihood function of the parameters

$p(y)$: marginal distribution of data which can be calculated as $\sum p(\psi)p(y|\theta)$ when ψ is discrete or $\int p(\psi)p(y|\psi)d\psi$ when ψ is continuous.

$p(\psi|y)$: posterior distribution of ψ

The selection of the prior distribution is critical due to its impact on the posterior distribution. However, in some cases, like cluster analysis based on mixture models, where the number of clusters and the parameters of the model are unknown, the prior should have minor impact on the posterior and data primarily identify the posterior distribution. These priors are called vague or objective priors. In these cases, using conjugate priors in which the prior, likelihood function and therefore the posterior follow the same probability families is a suitable solution (Everitt 2011).

The goal of Bayesian inference is to maintain a full posterior probability distribution over the set of unknown parameters. Any features of that posterior distribution such as moments, quantiles, etc. are of interest to get an efficient summary of posterior distribution. However, for high-dimensional posterior distributions, finding these features provides no simple solution and the computation process is complex. Therefore, a method is required to approximate these quantities. In such cases, the Markov chain Monte Carlo (MCMC) algorithm is a possible solution to compute those posterior quantities of interest. Commonly used MCMC samplers for finite mixture models are the Gibbs sampler.

A possible problem in inference for mixture models is label switching. To get a better understanding of this problem, let's consider a mixture model of two components A and B which is shown in Equation 10. The mixture model is invariant under the permutation of the components label. Therefore, the same mixture model (shown in Equation 9) could be developed by switching the label of two components A and B.

$$\text{Mixture Model } 1 = wA + (1-w)B \qquad \text{Eq.10}$$
$$\text{Mixture Model } 2 = wB + (1-w)A$$

The likelihood function and the posterior distribution of these two mixture models are also invariant under the permutation of the labels. For a k-component mixture model, there is K! mixture models over which the likelihood and therefore the posterior distribution are identical. This label switching might occur in different iterations of the MCMC sampling and lead to a problem in identifying the parameters of the components. In the Bayesian analysis of finite mixture models, the label switching problem must be considered and one common solution is to impose constrain on the components parameters like mean of components. For example, we can say that, $\mu_1 < \mu_2 < \ldots < \mu_k$ (Everitt 2011, Jasra 2005).

RESULTS AND DISCUSSIONS

In this section, first, the results of aggregate features extraction are provided. This is followed by the results of the mixture-modeled clustering of aggregates along with the implementation of the Bayesian inference.

Preliminary Evaluation

Table 1 shows the results of the feature analysis for six different aggregate particles. These particles which are visually different in shape were selected to evaluate the capability of the new system in measuring the shape features of particles and differentiating them. By visual inspection, we found out that the characteristic feature of the particle # 1 is its angularity and it should have the highest value of angularity compared to the others. As can be seen in Table 1, the angularity results of feature analysis indicate that particle #1 has the highest angularity. Accordingly, the system was able to discriminate this aggregate from the others based on its angularity. Particle #2 and particle #3 were selected from flat and elongated samples respectively. Based on the defined form and sphericity indices, lower values of sphericity along with higher values of form indicate that a particle is flat/elongated. The results of this analysis (as provided in Table 1) show that the sphericity values of particle 1 and particle 2 are small compared to sphericity values of other particles. In addition, the form values of these two particles are the biggest.

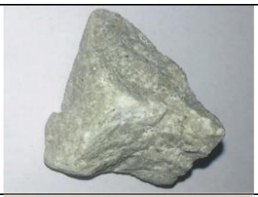
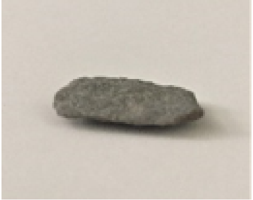
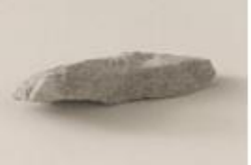
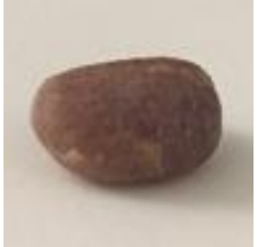
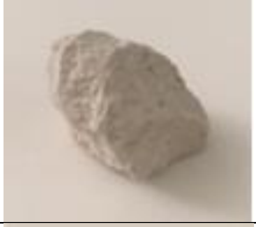
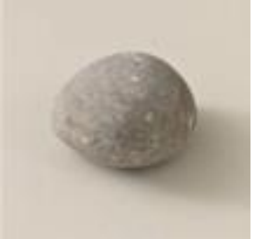
Sample #	Picture	Sphericity	Form	Angularity
1		0.824	4.687	10.982
2		0.459	5.115	2.626
3		0.515	6.386	7.598
4		0.716	4.037	1.993
5		0.881	3.679	3.943
6		0.748	3.321	2.640

Table 1: Preliminary evaluation results.

The texture index which was considered in this study was only applied to the aggregates with sizes greater than 9.5 mm. To evaluate the texture index, a sample of five coarse aggregate particles was selected and first evaluated using the sense of touch. The particles 1 and 2 were selected from completely rounded particles and were found to be smoother than others. However, by using the sense of touch, it is difficult to determine how rough the surface of an aggregate is. Therefore, the SV index was used in this study as a mean to discriminate the texture of aggregate particles. The second column of Table 2 shows the mean of SV calculated for 100 height profiles on the surface of aggregate particles. The third column presents the standard deviation of the SV values of 100 profiles. The outcomes of the texture analysis were found to be consistent with our observations. The results show that particle 4 has the highest value of SV.

Sample #	Mean	SD
1	0.2597	0.0088
2	0.2416	0.0151
3	0.5215	0.0203
4	1.2788	0.3096
5	0.5361	0.0578

Table 2: Results of Texture analysis.

Clustering Analysis Results

The feature data of 1398 aggregate particles were used for this analysis. As the first step in this study, the aggregate particles were clustered based on one feature at each time. This means that each observation is a one-dimensional variable. The Histogram of observations for each dataset with a kernel density plots are shown in Figure 9. The apparent skewness in the plot suggested that each dataset could be fit with a mixture of several univariate normal distributions. Note that for the mixture model, data are standardized by subtracting the mean and dividing by the standard deviation.

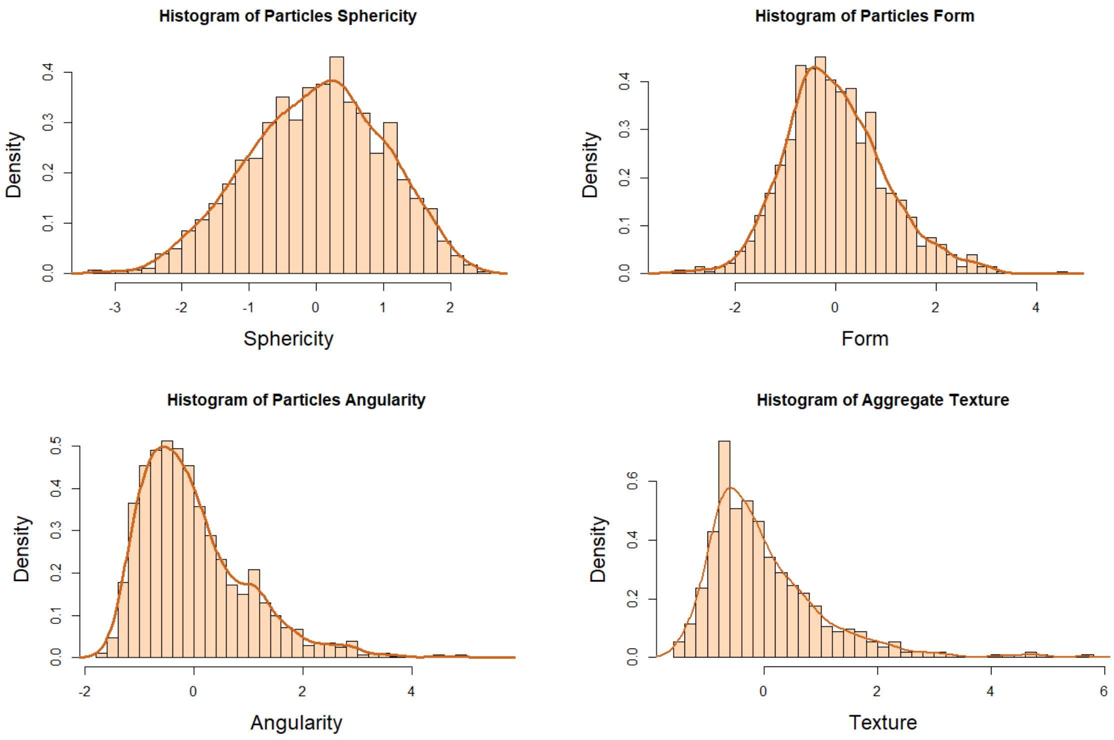


Figure 9: Histogram of observations.

To start the clustering analysis, it was decided to fit three normal distributions with different means and variances to these observations. To estimate the unknown parameters in the mixture model and perform Bayesian inference, following conjugate priors were given to the parameters c, w_k, and μ_k, ε_k.

1. The latent variable (c) has the categorical distribution. c ~ Cat [W]

2. The prior distribution of vector W is a Dirichlet distribution $(w_1, w_2, \dots, w_k) \sim D(\alpha_1, \alpha_2, \dots, \alpha_k)$.

*** The Dirichlet distribution is a multivariate probability distribution that describes $x_1, x_2, \dots, x_n$ where $n \geq 2$ and each $x_i \in (0,1)$ and $\sum_{i=1}^{n} x_i = 1$. It is parameterized byvector parameters $\alpha = (\alpha_1, \dots, \alpha_n)$ which are positive real numbers. We will use a symmetric Dirichlet distribution with $\alpha_1 = \alpha_2 = \dots = \alpha_k$

3. The prior distribution for μ_k is the normal distribution with mean of zero and fixed variance.

$\mu_k \sim N(0, \beta)$.

*** To avoid the label switching problem, the means of clusters were restricted in the following way:

$\mu_1 \sim N(0, \beta)$

$\mu_2 = \mu_1 + \xi_1$ where ξ_1 is a positive random variable following $N(0, \beta)$.

$\mu_3 = \mu_1 + \xi_1 + \xi_2$ where ξ_2 is a positive random variable following $N(0, \beta)$

The parametrization with the positive ξ_k avoids problems related to the label switching (shown in Figure 10).

4. The gamma distribution with a fixed shape parameter (a) and a fixed scale parameter (b) is the prior distribution of ${\varepsilon_k}^{-1}$(precision).

${\varepsilon_k}^{-1} \sim gamma(a, b)$

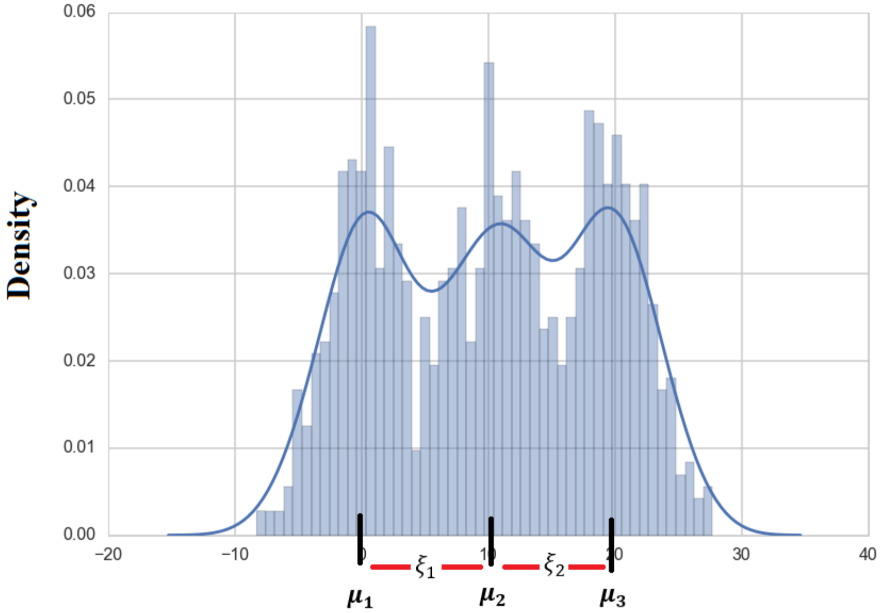


Figure 10: Restricting $\xi_k \geq 0$ rules out any issue of switching labels of cluster means.

Different packages such as OpenBUGS, WinBUGS are available to perform the Bayesian inference using Gibbs Sampling. In this study, OpenBUGS was used. In OpenBUGS, a statistical model including the relationship between variables and prior distribution of those variables need to be coded. The simulation began by providing initial values of the parameters to the model. By choosing initial values of the parameters closer to their target values, the MCMC would converge faster. Hierarchical clustering with complete linkage could be an effective way to generate a subset of initial values for the parameters μ_{k} and w_{k}. The initial values of μ_{k} were obtained from the mean of clusters created by hierarchical clustering and the initial values of w_{k} were calculated by dividing the number of observations labeled as cluster k to the total number of observations. The difference between μ_{k} values provides initial values of ξ_1 and ξ_2. Table 3 provides the results of hierarchical clustering for each dataset. The number of observations within each cluster indicates that the clusters have different variances. Hence, ε_k was considered different across the clusters.

Dataset	μ_1	μ_2	μ_3	w_1	w_2	w_3
Sphericity	-0.3377	0.5907	0.8638	785/1398 = 0.56	494/1398 = 0.35	119/1398 = 0.09

Table3: Initial values based on hierarchical clustering.

The described model was coded in the OpenBugs as follows. This model is only for the sphericity of the particles. The same model was developed for the other features.

```
model;

constant
N = 1398;                    # Number of Observations
K= 3;                        # Number of clusters

Variables
Sphericity[N],               # Observations
C[N],                        # The cluster attribution for each observation
lambda[C[i]],                # Mean of cluster
M,                           # Scaled positive shift between mean of clusters
lambdaTau[C[i]],             # Precision of clusters
sigma,                       # Standard deviation of clusters (1/tau)
W[];                         # mixing weight


{
     for(i in 1 : N) {
     Sphericity[i] ~ dnorm(mu[i], tau[i])     # distribution of observations

     mu[i] <- lambda[C[i]]

     tau[i] <- lambdaTau[C[i]]

     C[i] ~ dcat(W[])

}
W[1:3] ~ ddirch(alpha[])                # Dirichlet distribution

alpha[1] <- 10                         # prior parameter for mixing weights
alpha[2] <- 10
alpha[3] <- 10

lambda[1] <-z

z ~ dnorm(0.0, 1.0E-6)                  # hyperparameters for means
lambda[2] <- z + M1
lambda[3] <- z + M1 + M2
M1 ~ dnorm(0.0, 1.0E-6)I(0, )       # theta1 is a positive number
M2 ~ dnorm(0.0, 1.0E-6)I(0, )       # theta2 is a positive number


lambdaTau[1] ~ dgamma(0.01, 0.01)        # hyperparameters for precision
lambdaTau[2] ~ dgamma(0.01, 0.01)
lambdaTau[3] ~ dgamma(0.01, 0.01)


sigma[1] <- 1 / sqrt(lambdaTau[1])       # standard deviation of Normal dist
sigma[2] <- 1 / sqrt(lambdaTau[2])
sigma[3] <- 1 / sqrt(lambdaTau[3])

}

#initials
list(z = -0.338, theta=c(0.929, 0.273), W =c(0.56,0.35,0.09), lambdaTau =
c(100,100,100))
```

The process of sampling using the MCMC algorithm should be repeated for a large enough number of times. At each time, one sample is drawn. At the early iterations, the samples may not be drawn from the actual posterior distribution, but the MCMC algorithm guarantees that after a number of iterations the distribution approaches the stationary situation which is the actual posterior distribution. So, for an inadequate number of iterations, the simulations might be unrepresentative of the posterior distribution. In addition, the early iterations must be discarded from the analysis since they are influenced mainly by the initial values rather than the posterior distribution. These discarded samples are known as “burn-in” period. To determine the number of iterations and burn-in, for each dataset different numbers were tried, and the convergence of the sampling was evaluated.

Convergence is assumed when the sampling of all parameters seems to have reached the stationary regime. Based on the history plot of Markov chain for parameters, one can judge practical convergence of the chain. A Markov chain which has the appearance of a “fat hairy caterpillar” presents a stationary situation.

In the following, the Bayesian mixture model results for the sphericity data are provided. The same procedure was performed for the other datasets and the results are available in the appendix. For this study, 100000 iterations were used, and 10000 samples were discarded. Figures 11 and 12 provide the history plots of the parameters after the burn-in period. It is to be noted that in the OpenBUGS model, lambda[1], lambda[2], and lambda[3] represent the mean of three clusters and sigma[1], sigma[2], and sigma[3] represent the standard deviation of three clusters. The histories of the weights, means, and standard deviations look like hairy caterpillars. Hence, there is an evidence for convergence of the chains to the associated posterior distribution.

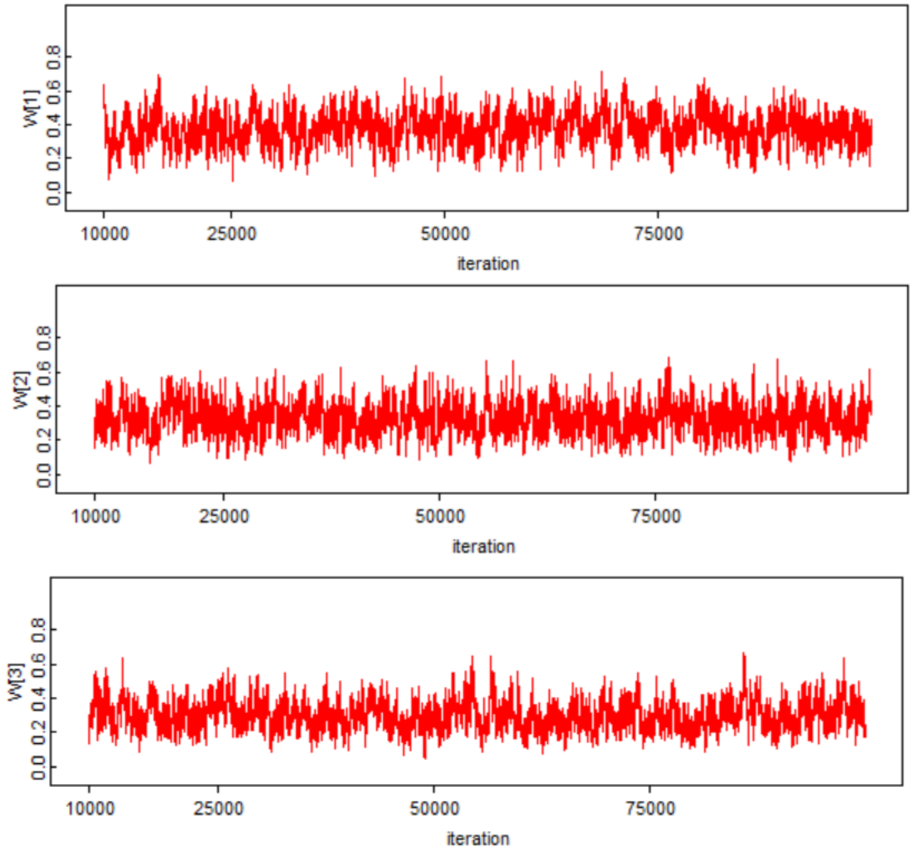


Figure 11: History pot of mixing weights.

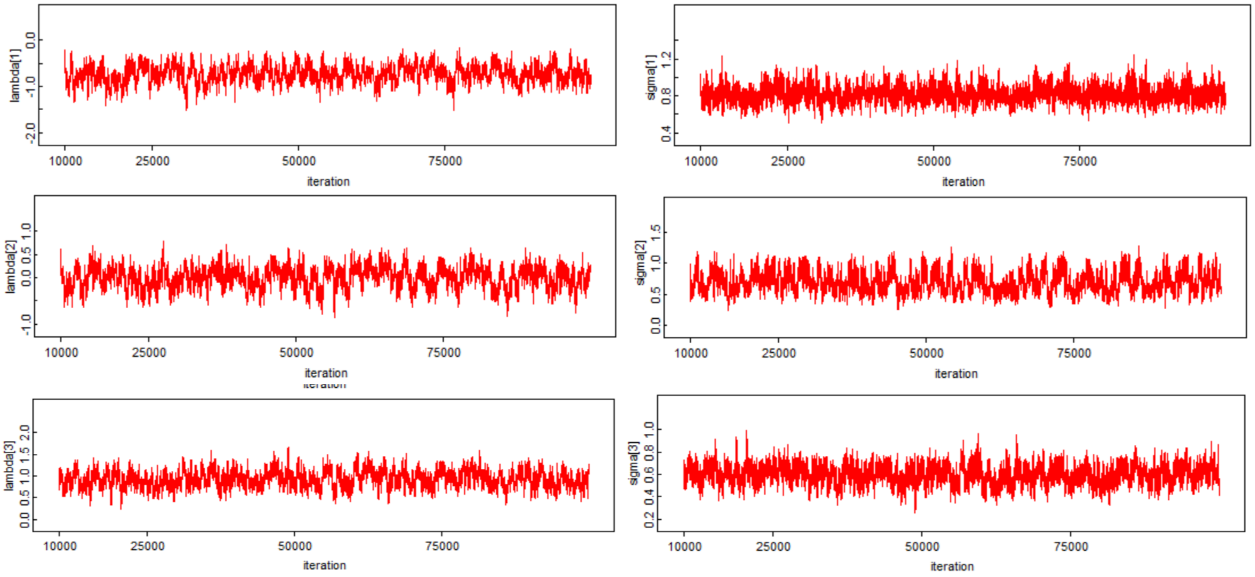


Figure 12: Left) History plots of means and right) standard deviations.

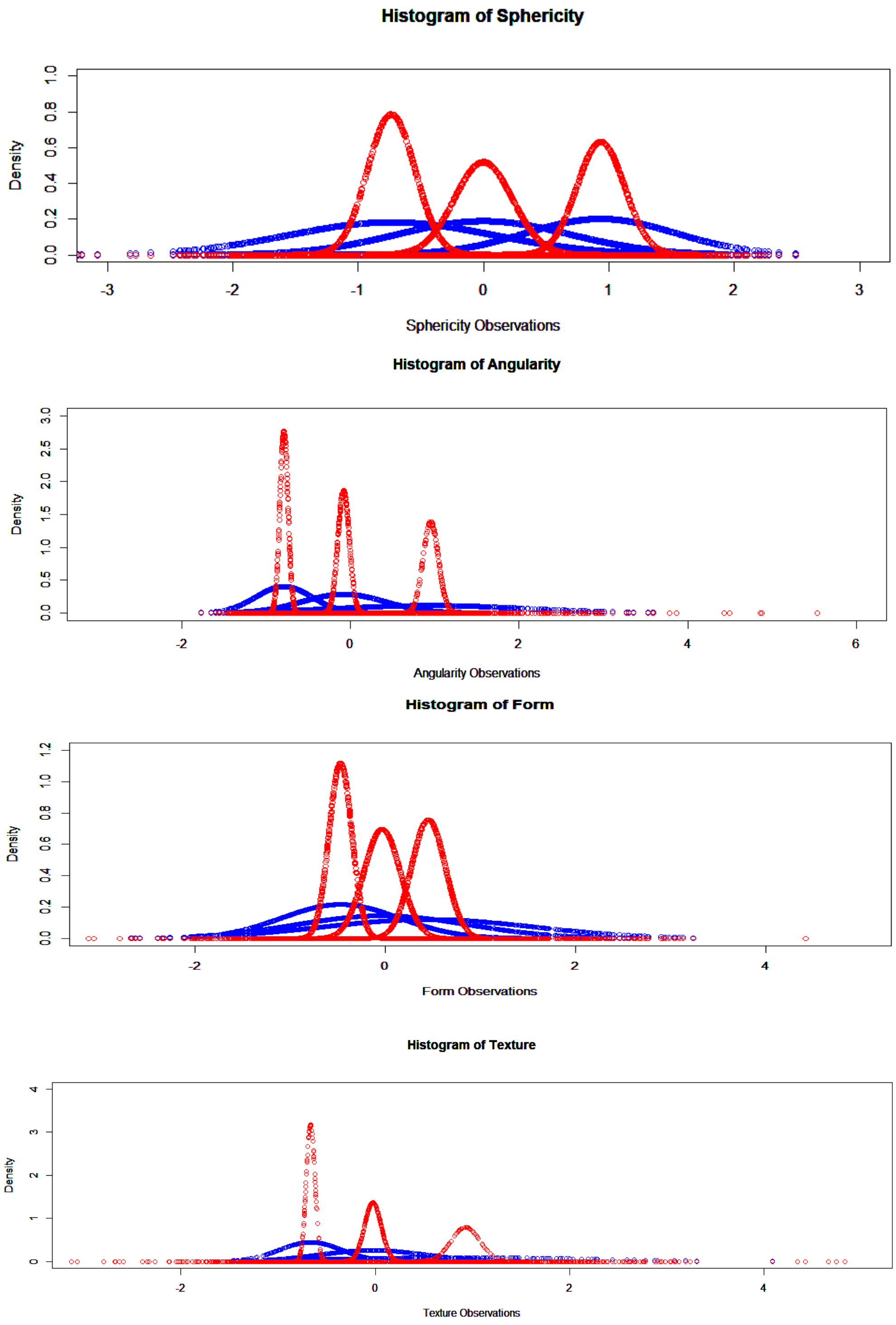
The Bayesian estimations of the parameters based on MCMC simulations are provided in Table 4. The results show that the weights are nearly identical for three groups. In addition, the results indicate that the standard deviations are slightly different over the clusters. For any parameter, the MCMC error in estimation is less than 5% of the sample standard deviation (sd) for that parameter. This confirms that the MCMC sampling used in this study was enough for this model.

Parameter	Mean	Sd	MCMC-error	Lower 2.5%	Upper 2.5%	Sample
W[1]	0.3733	0.0928	0.004	0.197	0.556	90000
W[2]	0.3263	0.0897	0.0035	0.1648	0.5121	90000
W[3]	0.3004	0.0835	0.0037	0.1575	0.4766	90000
lambda[1]	-0.7363	0.1897	0.0087	-1.112	-0.384	90000
lambda[2]	-0.0019	0.2313	0.011	-0.4686	0.4087	90000
lambda[3]	0.9340	0.19	0.0086	0.5769	1.295	90000
sigma[1]	0.8163	0.0804	0.003	0.6699	0.9826	90000
sigma[2]	0.6892	0.1626	0.0073	0.4165	1.005	90000
sigma[3]	0.5956	0.08	0.0033	0.4414	0.7456	90000

Table4: Estimation of parameters based on MCMC simulation.

Figure 13 shows the three fitted normal distributions, $p(y_i|c_i = k, \widehat{\mu_k}, \widehat{\varepsilon_k})$, along with the posterior distribution of cluster means $p(\mu_1|y)$. As seen in the sphericity plot, the first cluster, with the mean value of -0.74 models the flat particles which have low values of sphericity. About 37% of the particles belong to this cluster. The second cluster with the mean value of -0.002 includes around 33% of the particles with moderate sphericity. The third cluster with almost 30% of particles includes the particles with high values of sphericity.

Regarding the form of particles, as seen in the plot of form, three clusters with means of -0.48, 0.04, and 0.45 and mixing weights of 0.33, 0.33, and o.34 were found in the observations. These three clusters represent round, semi-round, and elongated particles, respectively. Three subgroups also were found in the angularity observations where the observations in the first group have the lowest values of angularity. The particles in the second group have moderate values of angularity. The third group represents angular particles. In terms of the texture of aggregates, as mentioned previously only the coarse aggregate of 1398 scanned particles were evaluated. Three clusters representing smooth, moderate rough, and rough particles were obtained in the observations.


Histogram of Sphericity

Histogram of Angularity

Histogram of Form

Histogram of Texture

Figure 13: Plot of fitted models.

Model Checking

In our model, the parameter α of Dirichlet distribution, and the number K of clusters were fixed. This following analysis was performed to understand how the changes in α and k affect the results. This analysis was performed in two parts. First, we examined different numbers of α and choose the best model. For the second part, the focus was on the number of clusters. Four different values of α and k were tested, $\alpha = \{0.5,1,10,100\}$ and $k = \{2,3,4,5\}$. In Bayesian analysis, the deviance information criterion (DIC) is a common method to check the goodness of fit and the complexity of models. However, OpenBUGS does not calculate DIC if the parameter set contains a discrete parameter. In this case, rather than DIC, the deviance statistic was calculated and compared between models. The deviance is defined by Equation 11. As seen in Equation 11, the deviance is a function of the unknown parameter and using the MCMC sampling of its posterior distribution can be generated. The posterior mean of the deviance could be used to measure the overall goodness of fit for a given model.

$$D(\psi) = -2log[p(y|\psi)] \qquad \text{Eq.11}$$

Where

$D(\psi)$: deviance

$p(y|\psi)$: likelihood function

Let $\overline{D}$ denote the posterior mean of $D(\psi)$. We note that the reported $\overline{D}$ can easily be mapped to the DIC by subtracting $D(\bar{\psi})$ where $\bar{\psi}$ are the posterior mean point estimates (Spiegelhalter et al. 2002). Table 5 provides the deviance of different models. In the first part of the analysis, the model with k=3 and α=10 has the lowest deviance. In the second part, α was kept fixed at 10 and the deviance was calculated for different numbers of k. As seen in Table 5, the model with k=5 and α=10 has the lowest deviance.

k	α	Deviance
First Part		
3	0.5	3398
3	1	3355
3	10	2990
3	100	3032
Second Part		
2	10	3320
3	10	2990
4	10	2800
5	10	2666

Table5: Results of Model Checking.

The same analysis was performed for other features. The results of analysis are provided in the appendix. It was found for form observations that the model with k=5 and α=10 has the lowest deviance. Regarding the angularity observations, the model with k=5 and α=100 has the lowest deviance. In addition, the results showed that a model with k=5 and α=10 fits better to texture observations. It is to be noted that by changing from k=3 to k=4 and 5, the deviance decreases slightly.

SUMMARY AND FURTHER DEVELOPMENT

Since we know each feature of aggregate (sphericity, form, angularity and texture) affect significantly on pavement performance, we can use cluster analysis to put aggregates in separate groups regarding their features and then for each application select those which lead to high-performance pavements. Accordingly, this study aimed to develop a classification system of aggregate particles based on four features, sphericity, form, angularity, and texture. This study was undertaken using a laser scanning tool (known as LLS) developed at the University of Texas at Austin. Several samples of aggregate particles obtained from different quarries in Texas were prepared and scanned. a MATLAB algorithm was developed which uses the data collected by the LLS as an input and delivers the measurement of sphericity, form, angularity, and texture. The measurements of scanned aggregates were used in another algorithm (developed using R software and the OpenBUGS package) to find clusters in the scanned aggregate particles regarding each feature. To cluster aggregate particles, mixture model-based clustering approach using Bayesian inference was used. This study reviewed the mixture model approach and showed the implementation of the Bayesian mixture model in order to find subgroups in aggregate particles.

This study is a part of an ongoing project. The extension to this study would be developing a clustering algorithm which considers a mixture of multivariate normal distributions in order to cluster particles based on all features simultaneously. In addition, in the next step, the clustering approach will be used to develop a classification table with specific levels for the aggregate features.

Appendix

Results of hierarchical clustering for $k = 2$

Dataset	μ_1	μ_2	w_1	w_2
Sphericity	-1.085	0.593	494/1398 = 0.35	904/1398 = 0.65
Form	-0.3408	1.5574	1147/1398 = 0.82	251/1398 = 0.18
Angularity	-0.3790	1.6590	1138/1398 =0.81	260/1398 = 0.19
Texture	-0.0502	4.7154	564/570 = 0.99	6/570 = 0.01

Results of hierarchical clustering for $k = 3$

Dataset	μ_1	μ_2	μ_3	w_1	w_2	w_3
Sphericity	-0.3377	0.5907	0.8638	785/1398 = 0.56	494/1398 = 0.35	119/1398 = 0.09
Form	-1.8398	-0.2329	1.5574	77/1398 = 0.05	1070/1398 = 0.77	251/1398 = 0.18
Angularity	-0.3790	1.5967	4.8408	1138/1398 =0.81	255/1398 = 0.18	5/1398 = 0.01
Texture	-0.5532	0.9559	4.7154	376/570 = 0.66	188/570 = 0.33	6/570 = 0.01

Results of hierarchical clustering for $k = 4$

Dataset	μ_1	μ_2	μ_3	μ_4	w_1	w_2	w_3	w_4
Sphericity	-2.3742	-0.9988	0.4253	1.6993	0.02	0.33	0.56	0.09
Form	-1.8398	-0.2329	1.3882	2.7585	0.06	0.76	0.16	0.02
Angularity	-0.3790	1.2735	2.6243	4.8408	0.81	0.14	0.04	0.01
Texture	-0.5532	0.8597	2.8702	4.7154	0.66	0.31	0.02	0.01

Results of hierarchical clustering for $k = 5$

Dataset	μ_1	μ_2	μ_3	μ_4	μ_5	w_1	w_2	w_3	w_4	w_5
Sphericity	-2.3742	-0.9988	0.1565	0.9971	1.6993	0.02	0.33	0.38	0.18	0.09
Form	-1.8398	-0.6919	0.2982	1.3882	2.7585	0.06	0.41	0.35	0.16	0.02
Angularity	-1.0363	-0.1208	1.2735	2.6243	4.8408	0.23	0.58	0.14	0.04	0.01
Texture	-0.5532	0.5357	1.6299	2.8702	4.7154	0.66	0.22	0.09	0.02	0.01

Results of Model Checking – Form

k	α	Deviance
First Part		
3	0.5	3602
3	1	3605
3	10	3497
3	100	3497
Second Part		
2	10	3615
3	10	3497
4	10	3385
5	10	3248

Results of Model Checking – Angularity

k	α	Deviance
First Part		
3	0.5	2387
3	1	2391
3	10	2501
3	100	2310
Second Part		
2	100	2826
3	100	2310
4	100	1989
5	100	1725

Results of Model Checking – Texture

k	α	Deviance
First Part		
3	0.5	940.7
3	1	941.7
3	10	929.3
3	100	967.9
Second Part		
2	10	1117
3	10	929.3
4	10	877.5
5	10	829.8

Reference

ASME B46.1. Surface Texture, Surface Roughness, Waviness and Lay. American Society of Mechanical Engineers, NY, New York 10017, 1995.

Barrett, P. J. The Shape of Rock Particles, A Critical Review. Sedimentology, Vol. 27, 1980, pp. 291-303.

Chamroukhia, F., M. Bartcusa, and H. Glotina. Dirichlet Process Parsimonious Mixtures for clustering. Journal of Elsevier, 2015.

Diebolt, J., and C. P. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 56, 1994, pp. 363-375.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. Cluster Analysis, 5th Edition. Wiley Series in Probability and Statistics, 2011.

Faranzen, J. Bayesian Cluster analysis. PhD dissertation, Department of Statistics, University of Stockholm, 2008.

Faranzen, J. Bayesian Inference for a Mixture Model using the Gibbs Sampler. Department of Statistics, University of Stockholm, 2006.

Field, F. Effect of Percent Crushed Variation in Coarse Aggregates of Bituminous Mixes. Association of Asphalt Paving Technologists Proceedings, Vol. 27, 1958, pp. 294-322.

Fraley, Ch., Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Journal of Classification, Vol. 24, 2007, pp. 155-181.

Heller, K. A. Efficient Bayesian Methods for Clustering. PhD Dissertation, University College London, 2007.

Herrin, M., and Goetz, W. H. Effect of Aggregate Shape on Stability of Bituminous Mixes. Highway Research Board Proceedings. Washington D.C. 1954, pp. 293-308.

Hoff, P. D. A First Course in Bayesian Statistical Methods. Springer Texts in Statistics, 2009.

James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning. Springer Science and Business Media New York, 2013.

Jasra, A., C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. Journal Statistical Science, Vol. 20, 2005, pp. 50-67.

KEYENCE, Measurement Sensors, http://www.keyence.com/products/measure/index.jsp, Accessed on January 1, 2017.

Kouchaki, S., H. Roshani, J.A. Prozzi, and J.B Hernandez. Evaluation of aggregates surface micro-texture using spectral analysis. Journal of Construction and Building Materials, Vol. 156, 2017, pp. 944-955.

Masad, E. A. The Development of a Computer Controlled Image Analysis System for Measuring Aggregate Shape Properties. Final Report for Highway-IDEA Project 77, Washington state University, 2003.

Masad, E. A., D. N. Little, L. Tashman, S. Saadeh, T. Al-Rousan, and R. Sukhwani. Evaluation of aggregate Characteristics Affecting HMA Concrete Performance. Report No. ICAR 203-1. 2003.

Meininger, R. C. Aggregate Test Related to Performance of Portland Cement Concrete Pavement. National Cooperative Highway Research Program Project 4-20A Final Report. Transportation Research Board, National Research Council, Washington, D.C., 1998.

Monismith, C. L. Influence of shape, size, and surface texture on the stiffness and fatigue response of asphalt mixtures. Highway Research Board 109 Special Report. Transportation Research Board, National Research Council, Washington, D.C., 1970.

Mora, M. L. International Roughness Index (IRI) and Slope Variance Models. Master Project in Civil Engineering. Brigham Young University, 2003.

Saxena, A., M. Prasad, A. Gupta, and N. Bharill. A Review of Clustering Techniques and developments. Journal of Neurocomputing, Vol. 267, 2017, pp. 664-681.

Schafer, M. Finite Bayesian Mixture Models with Applications in Spatial Cluster Analysis and Bioinformatics. PhD dissertation, Dortmund University, 2015.

Shalizi, C. Distances between Clustering, Hierarchical Clustering. Online Lecture Notes. Carnegie Mellon University, 2009. http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf. Accessed on November 22, 2017.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). J. Roy. Statist. Soc. B. Vol. 64, 2002, pp. 583-640.

Sun, W. Quantification of Morphological Characteristics of Aggregates at Multiple Scales. PhD Dissertation in Civil Engineering. Virginia Polytechnic Institute and State University, 2014.

Tan, P. N., M. Steinbach, and V. Kumar. Introduction to Data Mining. Pearson, 2005.

Zhihui, L. Bayesian Mixture Models. Master Thesis, McMaster University, 2010.