



HAL
open science

Probabilistically Sampled and Spectrally Clustered Plant Species using Phenotypic Characteristics

Aditya A Shastri, Kapil Ahuja, Milind B Ratnaparkhe, Yann Busnel

► **To cite this version:**

Aditya A Shastri, Kapil Ahuja, Milind B Ratnaparkhe, Yann Busnel. Probabilistically Sampled and Spectrally Clustered Plant Species using Phenotypic Characteristics. PeerJ, 2021, 9, pp.e11927. 10.7717/peerj.11927 . hal-03289961

HAL Id: hal-03289961

<https://imt-atlantique.hal.science/hal-03289961v1>

Submitted on 19 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Probabilistically Sampled and Spectrally 2 Clustered Plant Species using Phenotypic 3 Characteristics

4 Aditya A Shastri¹, Kapil Ahuja¹, Milind B Ratnaparkhe², and Yann Busnel³

5 ¹Math of Data Science & Simulation (MODSS) Lab, Indian Institute of Technology
6 Indore, Indore, India

7 ²ICAR-Indian Institute of Soybean Research, Indore, India

8 ³Network Systems, Cybersecurity and Digital Law Department, Institut Mines-Telecom
9 Atlantique, Rennes, France.

10 Corresponding author:

11 Kapil Ahuja¹

12 Email address: kahuja@iiti.ac.in

13 ABSTRACT

14 Phenotypic characteristics of a plant specie refer to its physical properties as cataloged by plant biologists
15 at different research centers around the world. Clustering species based upon their phenotypic character-
16 istics is used to obtain diverse sets of parents that are useful in their breeding programs. The Hierarchical
17 Clustering (HC) algorithm is the current standard in clustering of phenotypic data. This algorithm suffers
18 from low accuracy and high computational complexity issues. To address the accuracy challenge, we
19 propose the use of Spectral Clustering (SC) algorithm. To make the algorithm computationally cheap,
20 we propose using sampling, specifically, Pivotal Sampling that is probability based. Since application of
21 samplings to phenotypic data has not been explored much, for effective comparison, another sampling
22 technique called Vector Quantization (VQ) is adapted for this data as well. VQ has recently given
23 promising results for genotypic data.

24 The novelty of our SC with Pivotal Sampling algorithm is in constructing the crucial similarity matrix for
25 the clustering algorithm and defining probabilities for the sampling technique. Although our algorithm can
26 be applied to any plant species, we test it on the phenotypic data obtained from about 2400 Soybean
27 species. SC with Pivotal Sampling achieves substantially more accuracy (in terms of Silhouette Values)
28 than all the other proposed competitive clustering with sampling algorithms (i.e. SC with VQ, HC with
29 Pivotal Sampling, and HC with VQ). The complexities of our SC with Pivotal Sampling algorithm and
30 these three variants are almost same because of the involved sampling. In addition to this, SC with
31 Pivotal Sampling outperforms the standard HC algorithm in both accuracy and computational complexity.
32 We experimentally show that we are up to 45% more accurate than HC in terms of clustering accuracy.
33 The computational complexity of our algorithm is more than a magnitude less than that of HC.

34 1 INTRODUCTION

35 Genetic diversity has been an important foundation of plant breeding from the inception of agriculture
36 since it helps develop new plants to meet the growing food demand globally. The breeding process is a
37 complex combination of multiple stages (1). The first stage involves discovery of the native characteristics
38 where the selection of diverse parent donors is of paramount importance (2). One way plant genetic
39 diversity can be studied is by using their phenotypic characteristics (physical characteristics). This kind
40 of analysis can be relatively easily done because a sufficiently large amount of data is available from
41 different geographical areas. In the phenotypic context, which is our first focus, a few characteristics that
42 play an important role are Days to 50% Flowering, Days to Maturity, Plant Height, 100 Seed Weight,
43 Seed Yield Per Plant, Number of Branches Per Plant, etc.

44 Cluster analysis is an important tool to describe and summarize the variation present between different
45 plant species (3). Thus, clustering can be used to obtain diverse parents which, as mentioned above, is of
46 utmost importance. It is obvious that after clustering, the species present in the same cluster would have

47 similar characteristics, while those present in different clusters would be diverse. Phenotypic data for the
48 species of different plants (e.g., Soybean, Wheat, Rice, Maize, etc.) usually have enough variation for
49 accurate clustering. However, if this data is obtained for the species of the same plant, then clustering
50 becomes challenging due to less variation in the data, which forms our second focus.

51 Hierarchical Clustering (HC) is a traditional and standard method that is currently being used by
52 plant biologists for grouping of phenotypic data (3; 7; 8). However, this method has a few disadvantages.
53 First, it does not provide the level of accuracy required for clustering similar species (9). Second, HC is
54 based upon building a hierarchical cluster tree (also called dendrogram), which becomes cumbersome and
55 impractical to visualize when the data is too large. The second most common clustering algorithm that is
56 being currently used widely is Unweighted Pair Group Method using Arithmetic Mean (UPGMA). This
57 algorithm is a variant of HC, and hence, has the same two disadvantages as discussed above.

58 To overcome these two disadvantages, in this paper, we propose the use of the Spectral Clustering (SC)
59 algorithm. SC is mathematically sound and is known to give one of the most accurate clustering results
60 among the existing clustering algorithms (10). For genotypic data, we have recently shown substantial
61 accuracy improvements by using SC as well (11). Furthermore, unlike HC, SC does not generate the
62 intermediate hierarchical cluster tree. To the best of our knowledge, this algorithm has not been applied to
63 phenotypic data in any of the previous works (see the Literature Review section below).

64 HC, as well as SC, both are computationally expensive. They require substantial computational
65 time when clustering large amounts of data (10; 12). Hence, we use sampling to reduce this complexity.
66 Probability-based sampling techniques have recently gained a lot of attention because of their high
67 accuracy at reduced cost (13). Among these, Pivotal Sampling is most commonly used, and hence, we
68 apply it to phenotypic data (14). Like for SC, using Pivotal Sampling for phenotypic data is also new.
69 Recently, Vector Quantization (VQ) has given promising results for genotypic data (11). Hence, here we
70 adapt VQ for phenotypic data as well. This also serves as a good standard against which we compare
71 Pivotal Sampling.

72 To summarize, in this paper, we develop a modified SC with Pivotal Sampling algorithm that is
73 especially adapted for phenotypic data. The novelty of our work is in constructing the crucial similarity
74 matrix for the clustering algorithm and defining the probabilities for the sampling technique. Although
75 our algorithm can be applied to any plant species, we test it on around 2400 Soybean species obtained
76 from Indian Institute of Soybean Research, Indore, India (15). In the experiments, we perform four sets of
77 comparisons. *First*, we show that use of Pivotal Sampling does not deteriorate the cluster quality. *Second*,
78 our algorithm outperforms all the proposed competitive clustering algorithms with sampling in terms of
79 the accuracy (i.e. modified SC with VQ, HC with Pivotal Sampling, and HC with VQ). The computational
80 complexities of all these algorithms are similar because of the involved sampling. *Third*, our modified SC
81 with Pivotal Sampling doubly outperforms HC, which as earlier, is a standard in the plant studies domain.
82 In terms of the accuracy, we are up to 45% more accurate. In terms of complexity, our algorithm is more
83 than a magnitude cheaper than HC. *Fourth* and *finally*, we demonstrate the superiority of our algorithm by
84 comparing it with two previous works that are closest to ours.

85 The rest of this paper is organized as follows. Section 2 provides a brief summary of the previous
86 studies on phenotypic data. The standard algorithms for Pivotal Sampling and SC are discussed in Section
87 3. Section 4 describes the crucial adaptations done in Pivotal Sampling and SC for phenotypic data. The
88 data description, validation metric, and the experimental set-up are presented in Section 5. Section 6 gives
89 the experimental results. Finally, conclusions and future work are provided in Section 7.

90 2 LITERATURE REVIEW

91 In this section, we present some relevant previous studies on phenotypic data and the novelty of our
92 approach. Broadly, these studies can be classified into two categories. The first category consists of the
93 works that identify relationships between the different phenotypic characteristics (for example, lower
94 plant height may relate to lower plant yield or vice versa). These works are discussed in Section 2.1. The
95 second category consists of the studies that identify the species with dissimilar phenotypic characteristics
96 for the breeding program. These studies are discussed in Section 2.2. Finally, we present a set of works
97 that belong to both the categories in Section 2.3.

98 **2.1 First Category Previous Studies**

99 Immanuel et al. (16) in 2011 measured nine characteristics of 21 Rice species. Grain Yield (GY) was
 100 kept as the primary characteristic, and its correlations with all others were obtained. It was observed
 101 that characteristics like Plant Height (PH), Days to 50% Flowering (DF), Number of Tillers per Plant
 102 (NTP), Filled Grains per Panicle (FGP) and Panicle Length (PL) were positively correlated with GY. The
 103 remaining characteristics were negatively correlated with GY.

104 Divya et al. (17) in 2015 recorded 21 characteristics of two Rice species. The authors investigated the
 105 association between Infected Leaf Area (ILA), Blast Disease Susceptibility (BDS), Number of Tillers
 106 per Plant (NTP), Grain Yield (GY) and others. The authors concluded that, for example, (a) ILA had a
 107 significant positive correlation with leaf's BDS, (b) NTP exhibited the highest association with GY.

108 Gireesh et al. (15) in 2015 analyzed eight characteristics of 3443 Soybean species. The authors
 109 sampled the species using two methods, and correlations of all the characteristics with each other for
 110 both the samples were estimated. It was observed that, for example, Days to 50% Flowering (DF) was
 111 positively correlated with Days to Pod Initiation (DPI) in both the samples, while Number of Pods Per
 112 Plant (NPPP) showed a negative correlation with Nodes Per Plant (NPP).

113 Huang et al. (18) in 2018 studied six characteristics of 206 Soybean species. These characteristics
 114 were correlated with the three types of leaves; elliptical leaves, lanceolate leaves and round leaves. The
 115 authors deduced that Soybean plants with lanceolate leaves had maximum average Plant Height (PH),
 116 Number of Pods per Plant (NPP), Number of Branches per Plant (NBP), and 100-Seed Weight (SW),
 117 while Soybean plants with other two types of leaves had lower values of these characteristics.

118 Carpentieri-Pipolo et al. (19) in 2019 investigated 45 phenotypic characteristics of a Soybean
 119 specie. The authors then studied the effect of 20 bacteria isolated from roots, leaves, and stems on
 120 these characteristics (i.e. whether the bacteria had positive or negative activity on (correlation with) the
 121 45 characteristics). For example, *Enterobacter Ludwigii* (EL) bacteria, which is isolated from leaves,
 122 showed a positive correlation with 25 characteristics (e.g., Plant Growth Promotion (PGP)) and a negative
 123 correlation with remaining 20 characteristics (e.g., Phenylacetic Acid (PAC) assimilation). For better
 124 exposition, the above five studies are summarized in Table 1.

Studies	Plant	# of Species	Inferred Relationship
Immanuel et al. (2011)	Rice	21	PH, DF, NTP, FGP, PL \implies GY
Divya et al. (2015)	Rice	2	ILA \implies BDS
Gireesh et al. (2015)	Soybean	3443	DF \implies DPI and NPPP $\not\Rightarrow$ NPP
Huang et al. (2018)	Soybean	206	Lanceolate leaves \implies max avg PH, NPP, NBP and SW
Carpentieri-Pipolo et al. (2019)	Soybean	1	EL \implies PGP and EL $\not\Rightarrow$ PAC

Table 1. Summary of first category previous studies. Here, \implies represents positive correlation and $\not\Rightarrow$ represents negative correlation.

125 **2.2 Second Category Previous Studies**

126 Sharma et al. (7) in 2014 performed clustering of 24 synthetic Wheat species. Cluster analysis was
 127 performed using HC, and the species were grouped into three clusters using the polymorphic Inter Simple
 128 Sequence Repeat (ISSR) markers. The authors argued that species belonging to different clusters were
 129 diverse in terms of heat tolerance, and could be used to develop better heat tolerant specie.

130 Kahraman et al. (8) in 2014 analyzed the field performance of 35 Common Bean species by grouping
 131 them. The authors used HC, and the species were clustered into three groups based upon the matrix of
 132 relationship between the species. The species belonging to different clusters were considered diverse, and
 133 were used to select promising species for breeding.

134 Painkra et al. (3) in 2018 performed clustering of 273 Soybean species. Here, the authors used HC,
 135 and the species were grouped into seven clusters using Pearson Correlation Coefficient. According to

136 the authors, the species belonging to the distant clusters were more diverse such that choosing them
 137 maximized heterosis¹ in cross-breeding.

138 Islam et al. (20) in 2020 clustered ten Upland Rice species. Here, HC was used and the species were
 139 grouped into three clusters using a similarity coefficient between the species. The authors identified the
 140 two best species that could be used to obtain new species having higher plant yield. As earlier, here also,
 141 we summarize the above four studies in Table 2 below.

Studies	Plant	# of Species	Clustering Algorithm	# of Clusters	Development of Better Species
Sharma et al. (2014)	Wheat	24	HC	3	Heat Tolerant
Kahraman et al. (2014)	Common Bean	35	HC	3	Promising Species for Breeding
Painkra et al. (2018)	Soybean	273	HC	7	Improved Characteristics
Islam et al. (2020)	Rice	10	HC	3	Higher Plant Yield

Table 2. Summary of second category previous studies.

142 2.3 Both Categories Previous Studies

143 Fried et al. (21) in 2018 analyzed 11 characteristics of 49 Soybean species. The authors determined
 144 correlations between the root characteristics and other phenotypic characteristics. For example, Shoot
 145 Dry Weight (SDW) and Chlorophyll Index (CI) were positively correlated with Total Root Length (TRL)
 146 and Total Root Surface Area (TRSA), while Plant Height (PH) was negatively correlated with TRSA and
 147 Average Root Diameter (ARD). In this work, Principal Component Analysis (PCA) biplot was used to
 148 separate the species into seven clusters. According to the authors, this research was critical for Soybean
 149 improvement programs since it helped select species with the improved root characteristics.

150 Stansluos et al. (22) in 2019 analyzed 22 phenotypic characteristics for 11 Sweet Corn species. For
 151 example, the authors showed a positive and significant correlation of Yield of Marketable Ear (YME)
 152 with Ear Diameter (ED) and Number of Marketable Ear (NME), while a negative correlation between
 153 YME and Thousand Kernel Weight (TKW). Cluster analysis was performed using HC, and the corn
 154 species were grouped into four clusters using the Ward Linkage. The authors inferred substantial variation
 155 in morphological and agronomic capabilities of different species. Again, we summarize the above two
 156 studies in Table 3 below.

Studies	Plant	# of Species	Inferred Relationship	Clustering Algorithm	# of Clusters	Development of Better Species
Fried et al. (2018)	Soybean	49	SDW, CI \implies TRL, TRSA PH $\not\Rightarrow$ TRSA, ARD	PCA	7	Improved Root Characteristics
Stansluos et al. (2019)	Sweet Corn	11	YME \implies ED, NME YME $\not\Rightarrow$ TKW	HC	4	Better Morphological Capabilities

Table 3. Summary of both categories previous studies.

157 With the focus on the study of genetic diversity using phenotypic data, we have multiple novel
 158 contributions as below.

- 159 1. We focus on the second category above, and perform grouping of several thousand species as
 160 compared to a few hundred in the papers cited above. Note that from the first category, Gireesh et
 161 al. (15) did work with about three thousand species, and we do compare one aspect of our work
 162 with this previous work (more on this in the point 2a below).

¹Heterosis refers to the phenomenon in which a hybrid plant exhibits superiority over its parents in terms of Plant Yield or any other characteristic.

163 2. Clustering becomes computationally expensive when the size of the data is very large. Hence,
 164 sampling is required to make the underlying algorithm scalable. Thus, we perform clustering on
 165 the sampled data rather than the full one, which is not done in any of the papers above. We have
 166 two more innovations in this aspect as below.

- 167 (a) We use a probability-based sampling technique (Pivotal Sampling as mentioned earlier) that
 168 is highly accurate, and forms a completely new contribution. We demonstrate the superiority
 169 of our sampling by comparing it with the one done in Gireesh et al. (15). This comparison
 170 is discussed towards the end of the Results section. Please note that Gireesh et al. only
 171 performed sampling and did not cluster their data.
- 172 (b) HC, which is the most common clustering algorithm (and some other sporadically used
 173 algorithms like k -means and UPGMA), do not provide the level of accuracy needed. Again,
 174 as earlier, we develop a variant of the SC algorithm, which is considered highly accurate,
 175 especially for phenotypic data. Use of SC in this context is also completely new. We show
 176 the dominance of our clustering algorithm over the one proposed in the most recent past work
 177 by Islam et al. (20) towards the end of the Results section. Again, please note that Islam et al.
 178 only performed clustering and did not sample their data.

179 3 SAMPLING AND CLUSTERING ALGORITHMS

180 In this section, we briefly discuss the standard algorithms for Pivotal Sampling and SC in the two
 181 subsections below.

182 3.1 Pivotal Sampling

183 This is a well-developed sampling theory that handles complex data with unequal probabilities. The
 184 method is attractive because it can be easily implemented by a sequential procedure, i.e. by a single
 185 scan of the data (23). Thus, the complexity of this method is $\mathcal{O}(n)$, where n is the population size. It is
 186 important to emphasize that the method is independent of the density of the data.

187 Consider a finite population U of size n with its each unit identified by a label $i = 1, 2, \dots, n$. A
 188 sample S is a subset of U with its size, either being random ($N(S)$) or fixed (N). Obtaining the inclusion
 189 probabilities of all the units in the population, denoted by π_i with $i = 1, 2, \dots, n$, forms an important aspect
 190 of this unequal probability sampling technique.

191 The pivotal method is based on a principle of contests between units (13). At each step of the method,
 192 two units compete to get selected (or rejected). Consider unit i with probability π_i and unit j with
 193 probability π_j , then we have the two cases as below.

- 194 1. **Selection step** ($\pi_i + \pi_j \geq 1$): Here, one of the units is selected, while the other one gets the residual
 195 probability $\pi_i + \pi_j - 1$ and competes with another unit at the next step. More precisely, if (π_i, π_j)
 196 denotes the selection probabilities of the two units, then

$$(\pi_i, \pi_j) = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{with probability } \frac{1 - \pi_j}{2 - \pi_i - \pi_j} \\ (\pi_i + \pi_j - 1, 1) & \text{with probability } \frac{1 - \pi_i}{2 - \pi_i - \pi_j} \end{cases} \quad (1)$$

- 197 2. **Rejection step** ($\pi_i + \pi_j < 1$): Here, one of the units is definitely rejected (i.e. not selected in
 198 the sample), while the other one gets the sum of the inclusion probabilities of both the units and
 199 competes with another unit at the next step. More precisely,

$$(\pi_i, \pi_j) = \begin{cases} (0, \pi_i + \pi_j) & \text{with probability } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) & \text{with probability } \frac{\pi_i}{\pi_i + \pi_j} \end{cases} \quad (2)$$

200 This step is repeated for all the units present in the population until we get the sample of size $N(S)$ or
 201 N . The worst-case occurs when we obtain the last sample (i.e. N^{th} sample) in the last iteration.

3.2 Spectral Clustering

Clustering is one of the most widely used techniques for exploratory data analysis with applications ranging from statistics, computer science, and biology to social sciences and psychology etc. It is used to get a first impression of data by trying to identify groups having “similar behavior” among them. Compared to the traditional algorithms such as k -means, SC has many fundamental advantages. Results obtained by SC are often more accurate than the traditional approaches. It is simple to execute and can be efficiently implemented by using the standard linear algebra methods. The algorithm consists of four steps as below (10).

1. The first step in the SC algorithm is the construction of a matrix called the similarity matrix. Building this matrix is the most important aspect of this algorithm; better its quality, better the clustering accuracy (10). This matrix captures the local neighborhood relationships between the data points via similarity graphs and is usually built in three ways. The first such graph is a ε -neighborhood graph, where all the vertices whose pairwise distances are smaller than ε are connected. The second is a k -nearest neighborhood graph, where the goal is to connect vertex v_i with vertex v_j if v_j is among the k -nearest neighbors of v_i . The third and the final is the fully connected graph, where each vertex is connected with all the other vertices. Similarities are obtained only between the connected vertices. Thus, similarity matrices obtained by the first two graphs are usually sparse, while the fully connected graph yields a dense matrix.

Let the n vertices of a similarity graph be represented numerically by vectors a_1, a_2, \dots, a_n , respectively. Here, each $a_i \in \mathbb{R}^m$ is a column vector for $i = 1, \dots, n$. Also, let a_i^l and a_j^l denote the l^{th} elements of vectors a_i and a_j , respectively, with $l = 1, \dots, m$. There exist many distance measures to build the similarity matrix (24). We describe some common ones below using the above introduced terminologies.

- (a) **City block distance:** (24) It is the special case of the Minkowski distance

$$d_{ij} = \sqrt[p]{\sum_{l=1}^m |a_i^l - a_j^l|^p} \quad (3)$$

with $p = 1$.

- (b) **Euclidean distance:** (24) It is the ordinary straight line distance between two points in the Euclidean space. It is again the special case of the Minkowski distance, where the value of p is taken as 2. Thus, it is given by

$$d_{ij} = \sqrt{\sum_{l=1}^m (a_i^l - a_j^l)^2}. \quad (4)$$

- (c) **Squared Euclidean distance:** (24) It is the square of the Euclidean distance, and is given by

$$d_{ij} = \sum_{l=1}^m (a_i^l - a_j^l)^2. \quad (5)$$

- (d) **Cosine distance:** (24) It measures the cosine of the angle between two non-zero vectors, and is given by

$$d_{ij} = 1 - \frac{a_i \cdot a_j}{\|a_i\| \|a_j\|}, \quad (6)$$

where, $\|\cdot\|$ denotes the Euclidean norm of a vector.

- (e) **Correlation distance:** (25) It captures the correlation between two non-zero vectors, and is given by

$$d_{ij} = 1 - \frac{(a_i - \bar{a}_i)^t (a_j - \bar{a}_j)}{\sqrt{(a_i - \bar{a}_i)^t (a_i - \bar{a}_i)} \sqrt{(a_j - \bar{a}_j)^t (a_j - \bar{a}_j)}}, \quad (7)$$

236 where, \bar{a}_i and \bar{a}_j are the means of a_i and a_j multiplied with a vector of ones, respectively, and
 237 t signifies the transpose operation.
 238 (f) **Hamming distance:** (26) It measures the number of positions at which the corresponding
 239 values of two vectors are different, and is given by

$$d_{ij} = \frac{\#(a_i^t \neq a_j^t)}{n}, \quad (8)$$

240 (g) **Jaccard distance:** (27) It again measures the number of positions at which the corresponding
 241 values of two vectors are different excluding the positions where both the vectors have zero
 242 values, and is given by

$$d_{ij} = \frac{\#[(a_i^t \neq a_j^t) \cap ((a_i^t \neq 0) \cup (a_j^t \neq 0))]}{\#[(a_i^t \neq 0) \cup (a_j^t \neq 0)]}. \quad (9)$$

243 2. Next, a matrix called the Laplacian matrix is constructed. This matrix is either non-normalized or
 244 normalized. The non-normalized Laplacian matrix is defined as

$$L = D - W, \quad (10)$$

245 where W is the similarity matrix and D is a diagonal matrix whose elements are obtained by adding
 246 together the elements of all the columns for every row of W .

247 Normalized Laplacian matrix is again of two types: the symmetric Laplacian (L_{sym}) and the random
 248 walk Laplacian (L_{rw}). Both these matrices are closely related to each other and are defined as

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}. \quad (11)$$

$$L_{rw} = D^{-1} L = I - D^{-1} W. \quad (12)$$

249 Henceforth, the non-normalized Laplacian matrix is referred to as the Type-1 Laplacian, L_{sym} as
 250 the Type-2 Laplacian, and L_{rw} as the Type-3 Laplacian. In the literature, it is suggested to use
 251 the normalized Laplacian matrix instead of the non-normalized one, and specifically the Type-3
 252 Laplacian (10).

253 3. Once we have the Laplacian matrix, we obtain the first k eigenvectors u_1, \dots, u_k of this matrix, where
 254 k is the number of clusters.

255 4. Finally, these eigenvectors are clustered using the k -means clustering algorithm.

256 4 IMPLEMENTING PIVOTAL SAMPLING AND MODIFIED SPECTRAL CLUS- 257 TERING FOR PHENOTYPIC DATA

258 Here, we first present the application of Pivotal Sampling to obtain the samples from phenotypic data.
 259 Subsequently, we implement our modified SC algorithm on the same data. Consider that the phenotypic
 260 data of a plant consists of n species with each specie evaluated for m different characteristics/ traits. These
 261 characteristics may have categorical (non-numerical) or numerical values. Hence, we need to convert
 262 the categorical values into numerical ones. For this, we use the label encoder method (28). This method
 263 transforms non-numerical labels into numerical values between 0 and (number of categories) – 1. For
 264 example, if a characteristic has three possible labels; poor, good, and very good, we use 0, 1, and 2 to
 265 represent them, respectively.

266 As discussed in Section 3.1, Pivotal Sampling requires that the inclusion probabilities (i.e. π_i for
 267 $i = 1, \dots, n$), of all the species in the population U , be computed before a unit is considered for a contest.

268 The set of characteristics associated with a specie can be exploited in computing these probabilities. To
 269 select a sample of size N , where $N \ll n$, we obtain these probabilities as (23)

$$\pi_i = N \frac{\varkappa_i}{\sum_{i \in U} \varkappa_i}, \quad (13)$$

270 where \varkappa_i can be a property associated with any one characteristic (or a combination of them) of the i^{th}
 271 specie. Obtaining π_i in such a way also ensures that $\sum_{i=1}^n \pi_i = N$, i.e. we get exactly N selection steps,
 272 and in-turn, exactly N samples.

273 In our implementation, we use the deviation property of the species, which is discussed next. Since
 274 different characteristics have values in different ranges, we start by normalizing them as below (29; 30).

$$(\mathcal{X}_j)_i = \frac{(x_j)_i - \min(x_j)}{\max(x_j) - \min(x_j)}. \quad (14)$$

275 Here, $(\mathcal{X}_j)_i$ and $(x_j)_i$ are the normalized value and the actual value of the j^{th} characteristic for the i^{th}
 276 specie, respectively with $j = 1, \dots, m$ and $i = 1, \dots, n$. Furthermore, $\max(x_j)$ and $\min(x_j)$ are the maximum
 277 and the minimum values of the j^{th} characteristic among all the species. Now, the deviation for the i^{th}
 278 specie is calculated using the above normalized values as

$$dev_i = \sum_{j=1}^m \max(\mathcal{X}_j) - (\mathcal{X}_j)_i. \quad (15)$$

279 Here, $\max(\mathcal{X}_j)$ denotes the maximum normalized value of the j^{th} characteristic among all the species.
 280 Practically, a relatively large value of dev_i indicates that the i^{th} specie is less important, and hence, its
 281 probability should be small. Thus, the inclusion probability of a specie is calculated by taking $\varkappa_i = \frac{1}{dev_i}$
 282 in Eq. (13) or

$$\pi_i = N \frac{\frac{1}{dev_i}}{\sum_{i \in U} \frac{1}{dev_i}}. \quad (16)$$

283 Thus, if the sum of probabilities of two species under consideration is greater than or equal to 1, we follow
 284 the selection step as discussed in Section 3.1. On the other hand, we follow rejection step when this sum
 285 is less than 1. This process is repeated until we obtain N species.

286 Next, we discuss the clustering of these N species into k clusters. Similar to the standard SC algorithm
 287 discussed in Section 3.2, the first step in our modified SC is to obtain the similarity matrix. As mentioned
 288 earlier, this is the most important aspect of this algorithm since the better the matrix quality, the better
 289 the clustering accuracy. For this, we consider these N species as the vertices of a graph. Let vector p_i
 290 contain the normalized values of all the characteristics (m) for the i^{th} specie. Thus, we have N such
 291 vectors corresponding to the N species selected using Pivotal Sampling. That is, $p_i = [(\mathcal{X}_1)_i, \dots, (\mathcal{X}_m)_i]^T$
 292 for $i = 1, \dots, N$. In our implementation, we use a fully connected graph to build the similarity matrix, i.e.
 293 we obtain similarities among all the N species.

294 We define the similarity between the vectors p_1 and p_2 (without loss of generality, representing the
 295 species 1 and 2, respectively) as the inverse of the distance between these vectors obtained by using the
 296 distance measures mentioned in Section 3.2. This is intuitive because smaller the distance between any
 297 two species, larger the similarity between them and vice versa. We denote this distance by $d_{p_1 p_2}$. We
 298 build this matrix of size $N \times N$ by obtaining the similarities among all the N species.

299 The next step is to compute the Laplacian matrix, which when obtained from the above-discussed
 300 similarity matrix, generates poor eigenvalues,² and in-turn poor corresponding eigenvectors that are
 301 required for clustering³. Thus, instead of taking only the inverse of $d_{p_1 p_2}$, we also take its exponent, i.e.

²Zero/ close to zero and distinct eigenvalues are considered to be a good indicator of the connected components in a similarity matrix. Thus, eigenvalues are considered poor when they are not zero/ not close to zero or indistinct (10).

³For some distance matrices (like Euclidean distance), the eigenvalues don't even converge.

302 we define the similarity between the species 1 and 2 as $e^{-d_{p_1 p_2}}$ (31; 32). This, besides fixing the poor
 303 eigenvalues/ eigenvectors problem, also helps perform better clustering of the given data. Further, we
 304 follow the remaining steps as discussed in Section 3.2.

305 Above, we discussed the clustering of N sampled species into k clusters. However, our goal is to
 306 cluster all n species and not just N . Hence, there is a need to reverse-map the remaining $n - N$ species to
 307 these k clusters. For this, we define the notion of average similarity, which between the non-clustered
 308 specie \tilde{p} and the cluster C_l is given as

$$\mathcal{A} \mathcal{S}(C_l, \tilde{p}) = \frac{1}{\#(C_l)} \sum_{q \in C_l} e^{-d_{\tilde{p}q}}. \quad (17)$$

309 Here, $\#(C_l)$ denotes the number of species present in C_l and q is a specie originally clustered in C_l
 310 by our modified SC algorithm with Pivotal Sampling. We obtain the average similarity of \tilde{p} with all the
 311 k clusters (i.e. with C_l for $l = 1, \dots, k$), and associate it with the cluster with which \tilde{p} has the maximum
 312 similarity.

313 Next, we perform the complexity analysis of our algorithm. Since Pivotal Sampling and SC form the
 314 bases of our algorithm, we discuss the complexities of these algorithms before ours.

315 1. Pivotal Sampling (n : number of species, N : sample size)

316 (a) Obtaining Samples: $\mathcal{O}(n)$

317 2. SC (n, m : number of characteristics)

318 (a) Constructing Similarity Matrix: $\mathcal{O}(n^2 m)$

319 (b) Obtaining Laplacian Matrix: $\mathcal{O}(n^3)$

320 3. Our Algorithm (n, N, m)

321 (a) Obtaining Samples: $\mathcal{O}(n)$

322 (b) Constructing Similarity Matrix: $\mathcal{O}(N^2 m)$

323 (c) Obtaining Laplacian Matrix: $\mathcal{O}(N^3)$

324 (d) Reverse Mapping: $\mathcal{O}((n - N)N)$

325 Thus, the overall complexity of our algorithm is $\mathcal{O}(nN + N^3 + N^2 m)$. Here, we have kept three
 326 terms because any of these can dominate (here, $n \gg N, m$).

327 When we compare complexity of our algorithm with that of HC, which is $\mathcal{O}(n^3)$, it is evident that we are
 328 more than a magnitude faster than HC. We revisit this complexity analysis after discussing data in the
 329 next section, which supports our claim further.

330 5 METHODOLOGY

331 In this section, we first briefly discuss the data used for our experiments. Next, we check the goodness of
 332 our sampling technique by estimating a measure called the population total. The hypothesis related to this
 333 is as follows: for a particular sampling technique, if the estimate (or approximation) of the population
 334 total using the samples is close to the actual population total, then that sampling technique is considered
 335 good in an absolute sense. Finally, we describe the clustering set-up, where the below are analyzed.

336 (a) **The Validation Metric.** It is hypothesized that a good clustering is one where clusters are compact
 337 and well-separated.

338 (b) **The Ideal Number of Clusters.** The hypothesis related to this is as follows: given a set of
 339 eigenvalues of the Laplacian matrix, we can exploit the differences between these eigenvalues to
 340 obtain the ideal number of clusters.

341 (c) **The Suitable Distance Measures.** For building the similarity matrix and the Laplacian matrix, it
 342 is hypothesized to chose those matrices that give the best value for the validation metric.

5.1 Data Description

As mentioned in Introduction, our techniques can be applied to any plant data, however, here we experiment on phenotypic data of Soybean species. This data is taken from Indian Institute of Soybean Research, Indore, India, and consists of 29 different characteristics/ traits for 2376 Soybean species (15). Among these, we consider the following eight characteristics that are most important for higher yield: Early Plant Vigor (EPV), Plant Height (PH), Number of Primary Branches (NPB), Lodging Score (LS), Number of Pods Per Plant (NPPP), 100 Seed Weight (SW), Seed Yield Per Plant (SYPP) and Days to Pod Initiation (DPI). Out of these, EPV and LS have categorical values, while the remaining characteristics have numerical values. Hence, we convert these two categorical values into numerical ones using the label encoder method discussed in the previous section. A snapshot of this phenotypic data for a few Soybean species is given in Appendix A. Here, we also perform validation of this data by comparing it with a similar dataset.

Next, we compare the complexities of our algorithm and HC using the selected data; see Table 4. It is evident from this table that our algorithm achieves substantial savings.

# of Species (n)	# of Characteristics (m)	Sample Size (N)	Our Algorithm ($nN + N^3 + N^2m$)	HC (n^3)
2376	8	500	$(2376 \times 500) + (500)^3 + (500)^2 \times 8$ $= 1.28 \times 10^8$	$(2376)^3 = 1.34 \times 10^{10}$
2376	8	300	$(2376 \times 300) + (300)^3 + (300)^2 \times 8$ $= 2.84 \times 10^7$	$(2376)^3 = 1.34 \times 10^{10}$

Table 4. Computational complexity comparison for the given data.

5.2 Sampling Discussion

To inspect the quality of our sampling techniques, we estimate a measure called the population total, which is the addition of values of a particular characteristic for all the n units (species here) present in the population U . For example, if “Plant Height (PH)” is the characteristic of interest, then the population total is the addition of PH values for all the n species. Mathematically, the exact (or actual) population total for a characteristic of interest x_j is given as

$$Y = \sum_{i \in U} (x_j)_i, \quad (18)$$

where, as earlier, $(x_j)_i$ is the value of the j^{th} characteristic for the i^{th} specie and U is the set of all species. By the definition of this measure (and also for two more measures listed below in this section), we work with original (non-normalized) values of the characteristics rather than normalized ones. Also, based upon the same argument, we work with only those characteristics that are originally numerical.

In this work, we use two different estimators to compute an approximation of the population total from the sampled data. Closer the value of an estimator to the actual value, better the sampling. First is the Horvitz-Thompson (HT)-estimator (also called π -estimator), which is defined as (33)

$$Y'_{HT} = Y'_\pi = \sum_{i \in S} \frac{(x_j)_i}{\pi_i}, \quad (19)$$

where, π_i is the inclusion probability of the i^{th} specie as evaluated in Section 4 and S is the set of sampled species. Another estimator that we use is the Hájek-estimator. It is usually considered better than the HT-estimator and is given as (34)

$$Y'_{Hájek} = n \frac{\sum_{i \in S} \frac{(x_j)_i}{\pi_i}}{\sum_{i \in S} \frac{1}{\pi_i}}, \quad (20)$$

here, as earlier, n is the total number of species.

374 The actual population total and the values of the above two estimators for six characteristics (that have
375 numerical values) when using Pivotal Sampling and 500 samples are given in Table 5 (see columns 3, 4,
376 and 6, respectively). From this table, it is evident that the approximate values of the population total are
377 very close to the corresponding actual values. Thus, Pivotal Sampling works well in an absolute sense.
378 Here, we also compute the values of the two estimators when using VQ (see columns 5 and 7). We can
379 notice from these results that VQ also works reasonably well, but Pivotal Sampling is better.

Sr. No.	Characteristics	Actual Population Total	Pivotal Sampling (HT)	VQ (HT)	Pivotal Sampling (Hájek)	VQ (Hájek)
1	PH	121773.05	122507.84	123407.80	123716.09	113168.90
2	NPB	8576.56	8585.28	9669.29	8669.95	8867.05
3	NPPP	99712.72	100193.53	114465.66	101181.70	104968.67
4	SW	20073.32	19907.10	20966.86	20103.44	19227.28
5	SYPP	10048.04	10137.57	10536.08	10237.55	9661.92
6	DPI	136810	135309.78	149242.17	136644.29	136859.84

Table 5. HT and Hájek estimators values for Pivotal Sampling and VQ as compared to the actual population total with $N = 500$ as the sample size.

380 5.3 Clustering Setup

381 Here, *first*, we describe the criteria used to check the goodness of the generated clusters. There are two
382 categories of metrics available for the validation of clustering algorithms. One category includes the
383 metrics that require prior knowledge of the cluster labels (35). On the other hand, metrics from the second
384 category do not have this requirement (35; 36). In this work, the ideal cluster labels are not available, and
385 hence, we use a metric called Silhouette Value (from the second category) for validation of our clustering
386 algorithms (36).

387 Clustering is considered good if the obtained clusters are compact and well-separated. Silhouette
388 Value captures both these aspects well by computing the intra-cluster similarity and the inter-cluster
389 similarity. Consider that we have k clusters represented as C_1, \dots, C_k , and we want to obtain the Silhouette
390 Value of the i^{th} data point present in the cluster C_1 . For this, we compute the average distance between this
391 data point and all the other points in the cluster C_1 . This distance is denoted as $a(i)$. Next, we compute
392 the average distance between the i^{th} data point and all the other points in clusters C_2, \dots, C_k . This distance
393 is denoted as $b(i)$. Then, for this point, Silhouette Value is computed as below (36).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (21)$$

394 As evident from the above discussion, the intra-cluster similarity is captured by $a(i)$, and the inter-
395 cluster similarity is captured by $b(i)$. This value usually lies between minus one to plus one because the
396 denominator of Eq. (21) is always greater than its numerator. Silhouette Value for the overall clustering is
397 obtained by averaging the Silhouette Values of all the data points. If this value tends towards a positive
398 one, then the clustering is considered to be good. On the other hand, if this value tends towards a negative
399 one, then the clustering is considered poor.

400 *Second*, we determine the ideal number of clusters by using the eigenvalue gap heuristic (10; 37).
401 If $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of the matrix used for clustering (e.g., the Laplacian matrix), then
402 often the initial set of eigenvalues, say k , have a considerable difference between the consecutive ones
403 in this set. That is, $|\lambda_i - \lambda_{i+1}| \not\approx 0$ for $i = 1, \dots, k - 1$. After the k^{th} eigenvalue, this difference is usually
404 approximately zero. According to this heuristic, this k gives a good estimate of the ideal number of
405 clusters.

406 For this experiment, without loss of generality, we build the similarity matrix using the Euclidean
407 distance measure on the above discussed phenotypic data. As mentioned earlier, it is recommended to use
408 the Type-3 Laplacian matrix (10). Hence, we use its eigenvalues for estimating k . Figure 1 represents the
409 graph of the first fifty smallest eigenvalues (in absolute terms) of this Laplacian matrix. On the x -axis, we
410 have the eigenvalue number, and on the y -axis its corresponding value.

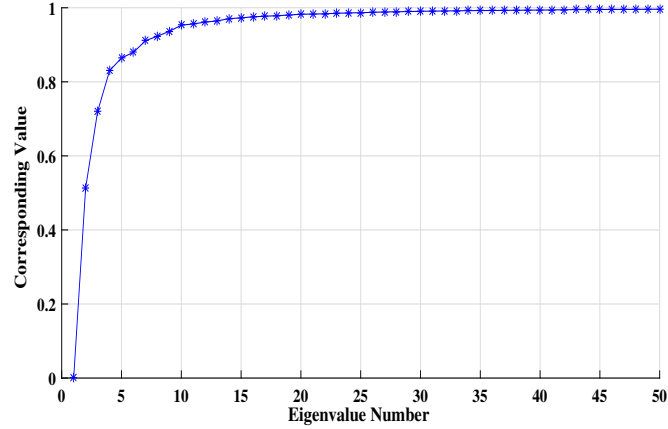


Figure 1. Fifty Smallest Eigenvalues of the Type-3 Laplacian Matrix Obtained from the Euclidean Similarity Matrix (for estimating the ideal number of clusters).

411 From this figure, we can see that there is a considerable difference between the first ten consecutive
 412 eigenvalues. After the tenth eigenvalue, this difference is very small (tending to zero). Hence, based upon
 413 the earlier argument and this plot, we take k as ten. To corroborate this choice more, we experiment with
 414 k as twenty and thirty as well. As expected, and discussed in-detail later in this section, Silhouette Values
 415 for these numbers of clusters are substantially lower than those for ten clusters.

416 *Third*, and final, we perform experiments to identify the suitable similarity measures to build the
 417 similarity matrix, and also verify that, as recommended, the Type-3 Laplacian matrix is the best. Table 6
 418 below gives Silhouette Values of our modified SC for all seven similarity measures and three Laplacians
 419 when clustering the earlier presented phenotypic data into 10, 20, and 30 clusters.

Sr. No.	Similarity Measure	Number of Clusters (k)	Type-1 Laplacian	Type-2 Laplacian	Type-3 Laplacian
1.	Euclidean	10	0.0828	-0.0273	0.2422
		20	0.0455	-0.1096	0.2069
		30	0.0887	-0.1536	0.1783
2.	Squared Euclidean	10	0.0815	-0.0555	0.3836
		20	-0.0315	-0.1809	0.2612
		30	0.0354	-0.2367	0.1538
3.	City-block	10	0.0687	0.2375	0.2647
		20	-0.0356	0.1347	0.2082
		30	-0.0870	0.0866	0.1887
4.	Cosine	10	0.1737	-0.1408	0.0694
		20	0.0359	-0.1973	0.0277
		30	0.0245	-0.2456	-0.0316
5.	Correlation	10	0.1926	-0.1259	0.3426
		20	0.0970	-0.2198	0.2313
		30	0.2383	-0.2604	0.1556
6.	Hamming	10	0.0643	0.0706	0.0775
		20	0.0683	0.0311	0.0382
		30	0.0715	0.0283	0.0229
7.	Jaccard	10	0.0716	0.0303	0.0458
		20	0.0446	0.0276	0.0236
		30	0.0279	0.0298	0.0318

Table 6. Silhouette Values for modified SC with seven similarity measures and three Laplacian matrices for $k = 10, 20,$ and 30 . Silhouette Values in bold represent good clustering.

420 From this table, it is evident that Silhouette Values for the Euclidean, Squared Euclidean, City-block
 421 and Correlation similarity measures and the Type-3 Laplacian matrix are the best. Hence, we use these
 422 four similarity measures and this Laplacian matrix. Also, as mentioned earlier, Silhouette Values decrease
 423 for twenty and thirty cluster sizes.

424 6 RESULTS AND DISCUSSION

425 Using the earlier presented dataset, and sampling-clustering setups, we compare our proposed algorithm
 426 (i.e. modified Spectral Clustering (SC) with Pivotal Sampling) with the existing variants in four ways.
 427 Again, as earlier, we use Silhouette Values for comparison. Quantifying statistical difference between
 428 different Silhouette Values is a hard task. In general, the more closer these values are to one, the better is
 429 the clustering (see Section 5.3).

430 First, we demonstrate that use of sampling with modified SC does not deteriorate the quality of clus-
 431 tering. Second, we compare our algorithm with modified SC with Vector Quantization (VQ), Hierarchical
 432 Clustering (HC)⁴ with Pivotal Sampling and HC with VQ for a sample size of 500. Since the results
 433 for modified SC with VQ come out to be closest to our algorithm, next, for broader appeal we compare
 434 these two algorithms for a sample size of 300. Third, we compare our algorithm with the current best
 435 in literature for this kind of data (i.e. HC without sampling) for both the sample sizes of 500 and 300.
 436 Fourth and finally, as discussed in the Literature Review section, we compare our sampling with that in
 437 Gireesh et al. (15) and our clustering with the one in Islam et al. (20).

438 *Initially*, we calculate the loss of accuracy incurred because of Pivotal Sampling in our algorithm.
 439 This loss for both the sample sizes and cluster size ten is listed in Table 7. Columns 1 and 2 give the
 440 sample sizes and the similarity measures chosen, respectively. Columns 3 and 4 give the Silhouette Values
 441 for modified SC without sampling (from Table 6) and our algorithm, respectively. The last column gives
 442 the percentage loss of accuracy. We can observe from this data that the loss of accuracy for one type of
 443 similarity measure (Correlation) is almost as low as -2% for both the sample sizes. This is considered
 444 acceptable because we are still better than the existing best algorithm (HC without sampling; please see
 445 Table 10 and its accompanying discussion below).

Sample Size	Similarity Measure	modified SC	modified SC with Pivotal Sampling	Percentage Loss of Accuracy
N = 500	Euclidean	0.2422	0.2152	-11.15%
	Squared Euclidean	0.3836	0.3362	-12.36%
	City-block	0.2647	0.2369	-10.50%
	Correlation	0.3426	0.3367	-1.72%
N = 300	Euclidean	0.2422	0.2104	-13.13%
	Squared Euclidean	0.3836	0.3280	-14.49%
	City-block	0.2647	0.2392	-9.63%
	Correlation	0.3426	0.3368	-1.69%

Table 7. Loss of accuracy because of Pivotal Sampling in modified SC for cluster size ten.

446 Here, we also perform a statistical test to support the above conjecture that using Pivotal Sampling
 447 does not substantially deteriorate the accuracy of our modified SC. For this, we use the ANOVA (analysis
 448 of variance) test (38). This test uses the variance between the different groups and the variance within
 449 each group to compute a value called the F-value, which is then compared with a standard estimate called
 450 F-critical. If F-value is less than F-critical, then it is inferred that the means of all the groups are equal.

451 The two groups for us refer to the modified SC results (column 3) and the modified SC with Pivotal
 452 Sampling results (column 4). The F-values here (using the Silhouette Values of the two groups) come
 453 out to be 0.3432 and 0.4202 for N = 500 and N = 300, respectively. Both these values are less than the
 454 F-critical value given in the F-distribution table of (39), which is 5.9873. Thus, using the above mentioned
 455 ANOVA test theory, we infer that that the mean Silhouette Value of modified SC is similar to the mean
 456 Silhouette Value of modified SC with Pivotal Sampling for both the sample sizes.

457 The results for the *second* set of comparisons are given in Table 8. Columns 2 and 3 give the similarity
 458 measures and the number of clusters chosen, respectively. Columns 4 and 5 give Silhouette Values of

⁴HC also requires building a similarity matrix.

459 modified SC with Pivotal Sampling and VQ, respectively, while columns 6 and 7 give Silhouette Values
 460 of HC with Pivotal Sampling and VQ, respectively.

Sr. No.	Similarity Measure	# of Clusters (k)	modified SC		HC	
			Pivotal Sampling	VQ	Pivotal Sampling	VQ
1.	Euclidean	10	0.2152	0.2061	0.2105	-0.1040
		20	0.1905	0.1448	0.2263*	-0.1620
		30	0.1741	0.1021	0.1933*	-0.2874
2.	Squared Euclidean	10	0.3362	0.2969	0.2634	-0.2096
		20	0.2469	0.1522	0.3726*	-0.5899
		30	0.1658	0.0440	0.2933*	-0.6083
3.	City-block	10	0.2369	0.2354	0.1703	-0.2278
		20	0.2019	0.1870	0.1879	-0.2398
		30	0.1752	0.1524	0.1988*	-0.2868
4.	Correlation	10	0.3367	0.2560	0.2582	-0.0060
		20	0.2291	0.0899	0.0867	-0.4120
		30	0.1742	-0.0349	0.0998	-0.7018

Table 8. Silhouette Values for modified SC and HC with Pivotal Sampling and VQ for $N = 500$. Silhouette Values in bold represent good clustering. Silhouette Values marked with * represent inflated values.

460

461 When we compare our algorithm (values in the fourth column, and highlighted in bold) with other
 462 variants, it is evident that we are clearly better than modified SC with VQ and HC with VQ (values in the
 463 fifth and the seventh columns); all our values are higher than those from these two algorithms.

464

465 When we compare our algorithm with HC with Pivotal Sampling (values in the sixth column), we
 466 again perform better for many cases. However, for some cases, our algorithm performs worse than HC
 467 with Pivotal Sampling (highlighted with a *). Upon further analysis (discussed below), we realize that
 468 segregation of species by HC with Pivotal Sampling into fewer clusters than practically observed, results
 469 in these set of Silhouette Values getting wrongly inflated.

470

471 To further assess the quality of the proposed technique, we present the distribution of species into
 472 different clusters (after reverse-mapping) for HC with Pivotal Sampling and our algorithm. Without loss
 473 of generality, this comparison is done using the Squared Euclidean similarity measure and cluster size
 474 thirty. In both the figures, on the x -axis, we have the cluster number and on the y -axis, the number of
 475 species present in them.

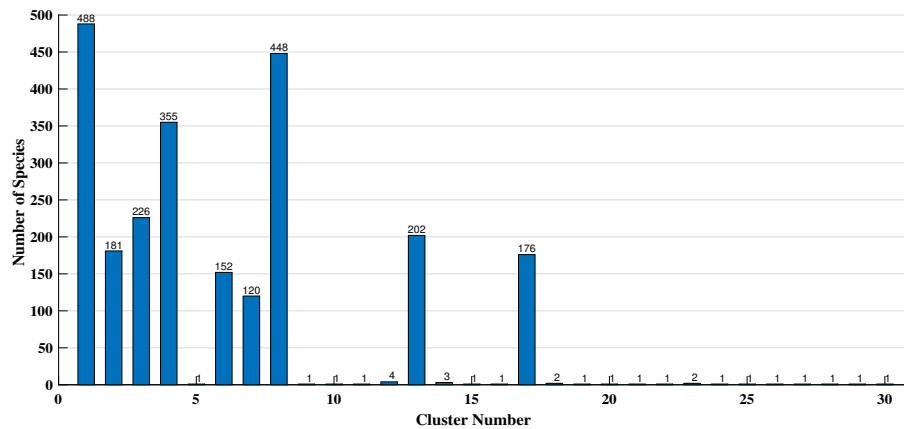


Figure 2. Distribution of Species (HC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size thirty.

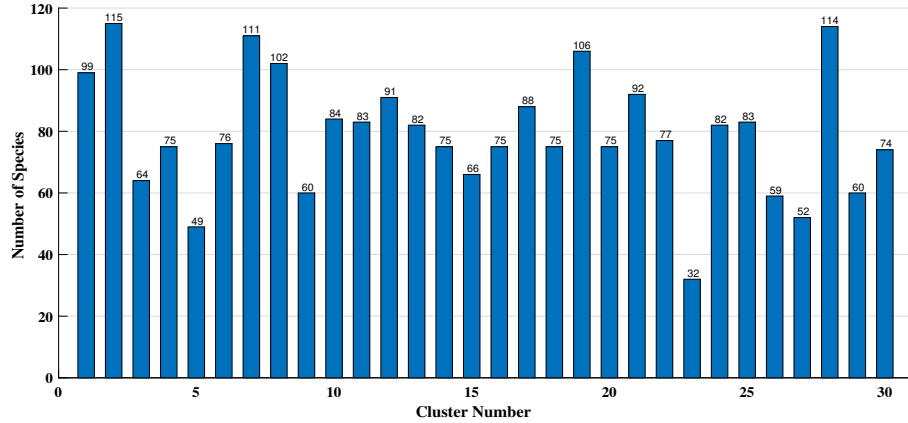


Figure 3. Distribution of Species (modified SC with Pivotal Sampling) for Squared Euclidean similarity measure and cluster size thirty.

475 As evident, Figure 2 depicts a very skewed distribution, i.e. most species are segregated into only a
 476 few clusters, while the remaining clusters contain only one or two species. At a broader level, this biased
 477 distribution of species obtained by HC with Pivotal Sampling is correct since all species belong to the
 478 same plant. On the contrary, the distribution in Figure 3 is fairly equal. That is, our algorithm equally
 479 distributes all species between the different clusters. At a finer level, this distribution is better since our
 480 algorithm is able to perform a more detailed clustering, i.e. it splits the bigger clusters into multiple
 481 smaller ones, which better captures the similarity between species.

482 This is also the reason for the inflation of Silhouette Values of HC with Pivotal Sampling in Table 8
 483 since the intra-cluster similarity for solitary specie is zero leading to its respective Silhouette Value to
 484 become one (the maximum possible; see Eq. (21)). Thus, our algorithm also outperforms HC with Pivotal
 485 Sampling, which from Table 8 was not very evident.

486 Next, as mentioned earlier, to further demonstrate the applicability of our work, we also present the
 487 results with a sample size 300. Since modified SC with VQ turns out to be our closest competitor, we
 488 compare our algorithm with this one only. This comparison is given in Table 9, with its columns mapping
 489 the respective columns of Table 8. As evident from Table 9, our modified SC with Pivotal Sampling
 490 substantially outperforms modified SC with VQ (see values in columns 4 and 5).

Sr. No.	Similarity Measure	# of Clusters (k)	modified SC	
			Pivotal Sampling	VQ
1.	Euclidean	10	0.2104	0.1833
		20	0.1968	0.0955
		30	0.1743	0.0722
2.	Squared Euclidean	10	0.3280	0.2589
		20	0.2424	0.1322
		30	0.1613	0.0044
3.	City-block	10	0.2392	0.2157
		20	0.1990	0.1696
		30	0.1752	0.1373
4.	Correlation	10	0.3368	0.2229
		20	0.2312	0.0336
		30	0.1725	-0.0788

Table 9. Silhouette Values for modified SC with Pivotal Sampling and VQ for $N = 300$.

491 As earlier, *third*, we compare the results of our algorithm (modified SC with Pivotal Sampling) with
 492 the currently popular clustering algorithm in the plant studies domain (i.e. HC without sampling). For this

493 set of experiments, without loss of generality, we use the cluster size of ten. The results of this comparison
 494 are given in Table 10, where the first four columns are self-explanatory (based upon the data given in
 495 Tables 8 and 9 earlier). In the last column of this table, we also evaluate the percentage improvement
 496 in our algorithm over HC. As evident from this table, our algorithm is up to 45% more accurate than
 497 HC for both the sample sizes. As earlier, our algorithm also has the crucial added benefit of reduced
 498 computational complexity as compared to HC.

Sample Size	Similarity Measure	modified SC with Pivotal Sampling	HC	Percentage Improvement
$N = 500$	Euclidean	0.2152	0.2173	-0.97%
	Squared Euclidean	0.3362	0.3257	3.22%
	City-block	0.2369	0.2135	10.96%
	Correlation	0.3367	0.2307	45.95%
$N = 300$	Euclidean	0.2104	0.2173	-3.28%
	Squared Euclidean	0.3280	0.3257	0.71%
	City-block	0.2392	0.2135	12.04%
	Correlation	0.3368	0.2307	45.99%

Table 10. Silhouette Values of modified SC with Pivotal Sampling and HC for cluster size ten.

499 *Fourth and finally*, as mentioned in the Literature Review section, we also compare our work with two
 500 previous works that are closest to ours. With the dataset almost the same as used by us, that is, a slightly
 501 larger phenotypic data for Soybean species, Gireesh et al. (15) performed Principal Component and Power
 502 Core based samplings to identify relationships between the different phenotypic characteristics (first
 503 category as in Section 2). We compare our sampling results with the best from (15) in Appendix B, which
 504 demonstrates the superiority of our sampling method. Islam et al. (20) performed HC on phenotypic data
 505 for Rice species (second category as in Section 2). In Appendix C, we apply modified SC on this dataset
 506 to again demonstrate that our clustering technique is better.

507 7 CONCLUSIONS AND FUTURE WORK

508 We present the modified Spectral Clustering (SC) with Pivotal Sampling algorithm for clustering plant
 509 species using their phenotypic data. We use SC for its accurate clustering and Pivotal Sampling for its
 510 effective sample selection that in-turn makes our algorithm scalable for large data. Since building the
 511 similarity matrix is crucial for the SC algorithm, we exhaustively adapt seven similarity measures to build
 512 such a matrix. We also present a novel way of assigning probabilities to different species for Pivotal
 513 Sampling.

514 We perform four sets of experiments on about 2400 Soybean species that demonstrate the superiority of
 515 our algorithm. *First*, we compare the Silhouette Values of modified SC without and with Pivotal Sampling,
 516 and show that the difference between these values is not significant. *Second*, when compared with
 517 the competitive clustering algorithms with samplings (SC with Vector Quantization (VQ), Hierarchical
 518 Clustering (HC) with Pivotal Sampling, and HC with VQ), Silhouette Values obtained when using our
 519 algorithm are higher. *Third*, our algorithm doubly outperforms the standard HC algorithm in terms of
 520 clustering accuracy and computational complexity. We are up to 45% more accurate and an order of
 521 magnitude faster than HC. *Fourth and finally*, we illustrate the excellence of our algorithm by comparing
 522 it with two previous works that are closest to ours.

523 Since the choice of the similarity matrix has a significant impact on the quality of clusters, in the future,
 524 we intend to adapt other ways of constructing this matrix such as Pearson χ^2 , Squared χ^2 , Bhattacharyya,
 525 Kullback-Liebler etc. (24). Furthermore, we also plan to observe the performance of Cube Sampling,
 526 which is another probabilistic sampling technique with data analysis properties complementary to Pivotal
 527 Sampling (13). Both Pivotal and Cube belong to the balanced sampling category, i.e. they satisfy $Y \approx Y'_{HT}$
 528 and $Y \approx Y'_{H\acute{a}jek}$ (recall Eqs. (18), (19), and (20)). Cube Sampling automatically obtains the samples
 529 (without specifying the sample size), which does not happen in Pivotal. As mentioned earlier, our
 530 algorithm is developed to work well for phenotypic data of all plant species. This is because different
 531 species vary only in the number of characteristics and the type of characteristics, both of which do not

532 affect our algorithm. We have preliminarily discussed this aspect for Maize and Rice in Appendix C, with
 533 extensive experiments for these two plants planned for future.

534 ACKNOWLEDGMENTS

535 The authors would like to thank Mr. Mohit Mohata, Mr. Ankit Gaur and Mr. Suryaveer Singh (IIT Indore,
 536 India) for their help in preliminary experiments, which they did as part of their undergraduate degree
 537 project. We would also like to sincerely thank Dr. Vangala Rajesh and Dr. Sanjay Gupta (Indian Institute
 538 of Soybean Research, Indore, India) for their help in generating the experimental data.

539 APPENDIX A

540 Here, we first present phenotypic data of the Soybean species used for our experiments. Please see Table
 541 A1 below. Next, we validate this data. For this, we compare our species data with a similar Soybean
 542 species data from (15) for the common set of phenotypic characteristics; Plant Height (PH), Number
 543 of Pods Per Plant (NPPP), and Days to Pod Initiation (DPI). This comparison is done using standard
 544 statistical metrics and is given in Table A2 below.

545 From this table, it is evident that the Standard Deviation (SD), Coefficient of Variance (CV), and
 546 Mean of our data and the data from the previous work are very close (for all three characteristics of PH,
 547 NPPP, and DPI). The slight variation in the metrics between the two data for all the characteristics is due
 548 to the difference in the ranges of the respective characteristics (due to the slightly differing selection of
 549 the species by the two works).

Species	EPV	PH	NPB	LS	NPPP	SW	SYPP	DPI
1	Poor	54	6.8	Moderate	59.8	6.5	2.5	65
2	Poor	67	3.4	Severe	33	6.2	3.9	64
3	Good	60.8	4	Moderate	34.6	6.1	3	65
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>n</i>	Very Good	89.6	5	Severe	32.6	7.3	3.4	62

Table A1. Phenotypic data of the Soybean species used for experiments. EPV: Early Plant Vigor, PH: Plant Height, NPB: Number of Primary Branches, LS: Lodging Score, NPPP: Number of Pods Per Plant, SW: 100 Seed Weight, SYPP: Seed Yield Per Plant, DPI: Days to Pod Initiation.

Parameter	Work	PH	NPPP	DPI
Standard Deviation (SD)	Our Work	16.61	20.16	7.85
	Previous Work (15)	18.6	24.1	8
Coefficient of Variance (CV)	Our Work	31.80	47.13	13.62
	Previous Work (15)	30.9	55.2	17.8
Mean	Our Work	52.24	42.78	57.60
	Previous Work (15)	60.3	43.6	54.7
Range	Our Work	13-102	4.33-197.66	24-80
	Previous Work (15)	5.4-118.8	1.33-301	30-98

Table A2. Comparison of SD, CV, mean, and range for our phenotypic data and similar previous data. Here, for comparison purposes, we have to work with original (non-normalized) values of the characteristics.

550 APPENDIX B

551 Here, we compare our sampling technique with those proposed by Gireesh et al. (15) for a similar dataset.
 552 As earlier, we do Pivotal Sampling on 2376 Soybean species while Gireesh et al. performed the Principal

553 Component Score (PCS) and the Power Core (PC) samplings on 3443 Soybean species. Since the samples
 554 obtained by the PC method are better, we compare our results with this sampling only.

555 This comparison is done using the statistical metrics of Standard Deviation (SD), Coefficient of
 556 Variance (CV) and Mean, and is given in Table A3 below. Since the metrics of our sampled data are more
 557 closer to our respective full data as compared to the metrics of the previous works' sampled data to its
 558 respective full data, our sampling is better.

Parameters	Work	Population	PH	NPPP	DPI
Standard Deviation (SD)	Our Work	Overall	16.61	20.16	7.85
		Sampled	17.34	18.90	7.42
	Previous Work (15)	Overall	18.6	24.1	8
		Sampled	22.15	45.33	11.73
Coefficient of Variance (CV)	Our Work	Overall	31.80	47.13	13.62
		Sampled	31.91	43.97	13.03
	Previous Work (15)	Overall	30.9	55.2	17.8
		Sampled	39.86	91.06	25.46
Mean	Our Work	Overall	52.24	42.78	57.60
		Sampled	54.34	42.99	56.94
	Previous Work (15)	Overall	60.3	43.6	54.7
		Sampled	55.57	49.78	56.65

Table A3. Comparison of Pivotal Sampling and Power Core method for three characteristics. Here, for comparison purposes, we have to work with original (non-normalized) values of the characteristics.

559 APPENDIX C

560 Since in the manuscript, we have demonstrated the usefulness of our algorithm on the species of the
 561 Soybean plant, here we demonstrate our algorithms' applicability to the species of the other two plants
 562 (Maize and Rice). The phenotypic data for the Maize species is given in Table A4, and for the Rice
 563 species is given in Table A5.

Species	DS	PH	EH	ED	EL	SW
1	77	75	33	3.2	11.6	2.3
2	98	45	14	2.7	8.1	1.6
3	68	132	80	3.7	16.2	3.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>n</i>	70	50	35	3.1	10.6	2.6

Table A4. Phenotypic data of the Maize species (40). DS: Days to Silking, PH: Plant Height, EH: Ear Height, ED: Ear Diameter, EL: Ear Length, SW: 100 Seed Weight.

Species	TN	PH	PN	PL	SW	BDR
1	6.8	124.2	5.5	25.6	22.1	Resistant
2	6.5	121.6	6.8	24.8	23.1	Moderately Resistant
3	7.2	126.4	4.5	26.1	19.5	Moderately Susceptible
⋮	⋮	⋮	⋮	⋮	⋮	⋮
<i>n</i>	7.1	131.4	5.1	25.9	18.5	Susceptible

Table A5. Phenotypic data of the Rice species (20; 41). TN: Tiller Number, PH: Plant Height, PN: Panicle Number, PL: Panicle Length, SW: 100 Seed Weight, BDR: Blast Disease Resistance.

564 We can observe from Tables A1, A4, and A5 that there is a set of common phenotypic characteristics
 565 for the three plant species. Also, the values of all the characteristics are either categorical or numerical.

566 As mentioned earlier, the categorical values can be easily converted to numerical ones. Since the input to
 567 our algorithm is a matrix built using the phenotypic data for given species, it can be applied to any of
 568 these plants.

569 To demonstrate the usefulness of our algorithm to the two new plant species, without loss of generality,
 570 we perform clustering of Rice species using our modified SC. For this, we use the data from Islam et
 571 al. (20), where the authors have used HC to cluster ten Rice species into three clusters. Hence, we also
 572 cluster these ten species into three clusters using our modified SC. In (20), the output is in the form of a
 573 hierarchical tree, which is non-numerical, and hence, difficult to compare. Thus, we compute Silhouette
 574 Values for our modified SC and HC. This data for the four similarity measures are given in Table A6. As
 575 evident from this table, our algorithm substantially outperforms HC.

Similarity Measure	modified SC	HC
Euclidean	0.2743	0.0076
Squared Euclidean	0.3276	0.0253
City-block	0.2561	0.0219
Correlation	0.3265	0.0433

Table A6. Silhouette Values of modified SC and HC for three clusters of ten Rice species.

576 REFERENCES

- 577 [1] Louwaars NP, Plant breeding and diversity: a troubled relationship, *Euphytica*, **214(7)**:1–9, 2018.
 578 [2] Swarup S, Cargill E, Crosby K, Flagel L, Kniskern J and Glenn K, Genetic diversity is indispensable
 579 for plant breeding to improve crops, *Crop Science*, **61**:839–852, 2020.
 580 [3] Painkra P, Shrivatava R, Nag SK and Markam NK, Clustering analysis of soybean germplasm (*Glycine*
 581 *max* L. Merrill), *The Pharma Innovation Journal*, **7(4)**:781–786, 2018.
 582 [4] Ingvarsson PK and Street NR, Association genetics of complex traits in plants, *New Phytologist*,
 583 **189(4)**:909–922, 2011.
 584 [5] Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, Fang C, Shen Y, Liu T,
 585 Li C, Li Q, Wu M, Wang M, Wu Y, Dong Y, Wan W, Wang X, Ding Z, Gao Y, Xiang H, Zhu B, Lee
 586 S, Wang W and Tian Z, Resequencing 302 wild and cultivated accessions identifies genes related to
 587 domestication and improvement in soybean, *Nature Biotechnology*, **33(4)**:408–414, 2015.
 588 [6] Subramanian S, Ramasamy U and Chen D, VCF2PopTree: a client-side software to construct popula-
 589 tion phylogeny from genome-wide SNPs, *PeerJ*, **7**:e8213, 2019.
 590 [7] Sharma P, Sareen S, Saini M, Verma A, Tyagi BS and Sharma I, Assessing genetic variation for heat
 591 tolerance in synthetic wheat lines using phenotypic data and molecular markers, *Australian Journal of*
 592 *Crop Science*, **8(4)**:515–522, 2014.
 593 [8] Kahraman A, Onder M and Ceyhan E, Cluster analysis in common bean genotypes (*Phaseolus vulgaris*
 594 L.), *Turkish Journal of Agricultural and Natural Sciences*, **1**:1030–1035, 2014.
 595 [9] Rokach L, A survey of clustering algorithms, in Maimon O, Rokach L (eds.), *Data Mining and*
 596 *Knowledge Discovery Handbook*, Springer, Boston, MA, pp. 269–298, 2009.
 597 [10] Luxburg UV, A tutorial on spectral clustering, *Statistics and Computing*, **17(4)**:395–416, 2007.
 598 [11] Shastri AA, Ahuja K, Ratnaparkhe MB, Shah A, Gagrani A and Lal A, Vector quantized spectral
 599 clustering applied to whole genome sequences of plants, *Evolutionary Bioinformatics*, **15**:1–7, 2019.
 600 [12] Mullner D, fastcluster: fast hierarchical, agglomerative clustering routines for R and Python, *Journal*
 601 *of Statistical Software*, **53(9)**:1–8, 2013.
 602 [13] Tille Y, Sampling algorithms, Springer-Verlag New York, 2006.
 603 [14] Chauvet G, On a characterization of ordered pivotal sampling, *Bernoulli*, **18(4)**:1320–1340, 2012.
 604 [15] Gireesh C, Husain SM, Shivakumar M, Satpute GK, Kumawat G, Arya M, Agarwal DK and Bhatia
 605 VS, Integrating principal component score strategy with power core method for development of core
 606 collection in Indian soybean germplasm, *Plant Genetic Resources*, **15(3)**:230–238, 2015.
 607 [16] Immanuel SC, Pothiraj N, Thiyagarajan K, Bharathi M and Rabindran R, Genetic parameters of
 608 variability, correlation and path-coefficient studies for grain yield and other yield attributes among
 609 rice blast disease resistant genotypes of rice (*Oryza sativa* L.), *African Journal of Biotechnology*,
 610 **10(17)**:3322–3334, 2011.

- 611 [17] Divya B, Robin S, Biswas A and Joel JA, Genetics of association among yield and blast resistance
612 traits in rice (*Oryza sativa*), *Indian Journal of Agricultural Sciences*, **85(3)**:354–360, 2015.
- 613 [18] Huang F, Gan Y, Zhang D, Deng F and Peng J, Leaf shape variation and its correlation to phenotypic
614 traits of Soybean in northeast China. *Proc. of the 6th Int. Conf. on Bioinformatics and Computational
615 Biology*, Association for Computing Machinery, New York, pp. 40–45, 2018.
- 616 [19] Carpentieri-Pipolo V, de Almeida Lopes KB and Degrassi G, Phenotypic and genotypic characteriza-
617 tion of endophytic bacteria associated with transgenic and non-transgenic soybean plants, *Archives of
618 Microbiology*, **201(8)**:1029–1045, 2019.
- 619 [20] Islam SS, Anothai J, Nualsri C and Soonsuwon W, Genetic variability and cluster analysis for
620 phenological traits of Thai Indigenous Upland Rice (*Oryza sativa* L.), *Indian Journal of Agricultural
621 Research*, **54(2)**:211–216, 2020.
- 622 [21] Fried HG, Narayanan S and Fallen B, Characterization of a soybean (*Glycine max* L. Merr.) germplasm
623 collection for root traits, *PLoS ONE*, **13(7)**:e0200463, 2018.
- 624 [22] Stansluos AA, Ozturk A, Kodaz S, Pour AH and Sylvestre H, Genetic diversity in sweet corn (*Zea
625 mays* L. *saccharata*) cultivars evaluated by agronomic traits, *Mysore Journal of Agricultural Sciences*,
626 **53(1)**:1–8, 2019.
- 627 [23] Deville JC and Tille Y, Unequal probability sampling without replacement through a splitting method,
628 *Biometrika*, **85(1)**:89–101, 1998.
- 629 [24] Cha SH, Comprehensive survey on distance/similarity measures between probability density functions,
630 *International Journal of Mathematical Models and Methods in Applied Sciences*, **4(1)**:300–307, 2007.
- 631 [25] Szekely GJ, Rizzo ML and Bakirov NK, Measuring and testing dependence by correlation of distances,
632 *The Annals of Statistics*, **35(6)**:2769–2794, 2007.
- 633 [26] Norouzi M, Fleet DJ and Salakhutdinov RR, Hamming distance metric learning, *Proc. of the 25th Int.
634 Conf. on Advances in Neural Information Processing Systems*, pp. 1061–1069, 2012.
- 635 [27] Matlab Documentation, Pdist - pairwise distance between pairs of observations, 2006.
- 636 [28] Hancock J and Khoshgoftaar T, Survey on categorical data for neural networks, *Journal of Big Data*,
637 **7**:1–41, 2020.
- 638 [29] Jain A, Nandakumar K and Ross A, Score normalization in multimodal bio-metric systems, *Pattern
639 Recognition*, **38(12)**:2270–2285, 2005.
- 640 [30] Shastri AA, Tamrakar D and Ahuja K, Density-wise two stage mammogram classification using
641 texture exploiting descriptors, *Expert Systems with Applications*, **99**:71–82, 2018.
- 642 [31] Ng AY, Jordan MI and Weiss Y, On spectral clustering: analysis and an algorithm, *Proc. of the 15th
643 Int. Conf. on Advances in Neural Information Processing Systems*, pp. 849–856, 2002.
- 644 [32] Nemade V, Shastri AA, Ahuja K and Tiwari A, Scaled and projected spectral clustering with vector
645 quantization for handling big data. *Proc. of the Symposium Series on Computational Intelligence
646 (SSCI)*, pp. 2174–2179, 2018.
- 647 [33] Horvitz DG and Thompson DJ, A generalization of sampling without replacement from a finite
648 universe, *Journal of the American Statistical Association*, **47**:663–685, 1952.
- 649 [34] Hájek J, Comment on “An essay on the logical foundations of survey sampling, part one”, in Godambe
650 VP, Sprott DA (eds.), *Foundations of Statistical Inference*, Rinehart and Winston, Toronto, 1971.
- 651 [35] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Fofou S and Bouras A, A survey of
652 clustering algorithms for big data: taxonomy and empirical analysis, *IEEE Transactions on Emerging
653 Topics in Computing*, **2(3)**:267–279, 2014.
- 654 [36] Rousseeuw PJ, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,
655 *Journal of Computational and Applied Mathematics*, **20**:53–65, 1987.
- 656 [37] Kong W, Sun C, Hu S and Zhang J, Automatic spectral clustering and its application, *Proc. of the 3rd
657 Int. Conf. on Intelligent Computation Technology and Automation*, IEEE, pp. 841–845, 2010.
- 658 [38] Rutherford A, ANOVA and ANCOVA: a GLM approach, Wiley, 2011.
- 659 [39] Beyer WH, Handbook of tables for probability and statistics, CRC Press, 2019.
- 660 [40] Belalia N, Lupini A, Djemel A, Morsli A, Mauceri A, Lotti C, Khelifi-Slaoui M, Khelifi L and Sunseri
661 F, Analysis of genetic diversity and population structure in Saharan maize (*Zea mays* L.) populations
662 using phenotypic traits and SSR markers. *Genetic Resources and Crop Evolution*, **66(1)**:243–257,
663 2019.
- 664 [41] Kim B, Classifying *Oryza sativa* accessions into *Indica* and *Japonica* using logistic regression model
665 with phenotypic data, *PeerJ*, **7**:e7259, 2019.