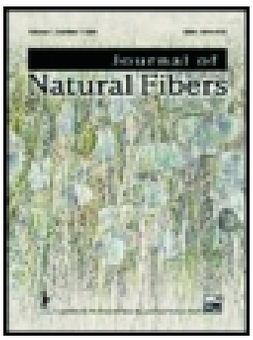

Journal of Natural Fibers

ISSN: 1544-0478 (Print) 1544-046X (Online) Journal homepage: http://www.tandfonline.com/loi/wjnf20

Clustering and Classification of Cotton Lint Using Principle Component Analysis, Agglomerative Hierarchical Clustering, and K-Means Clustering

Edwin Kamalha , Jovan Kiberu, Ildephonse Nibikora, Josphat Igadwa Mwasiagi & Edison Omollo

To cite this article: Edwin Kamalha , Jovan Kiberu, Ildephonse Nibikora, Josphat Igadwa Mwasiagi & Edison Omollo (2017): Clustering and Classification of Cotton Lint Using Principle Component Analysis, Agglomerative Hierarchical Clustering, and K-Means Clustering, Journal of Natural Fibers, DOI: 10.1080/15440478.2017.1340220

To link to this article: http://dx.doi.org/10.1080/15440478.2017.1340220

JOURNAL OF NATURAL FIBERS
https://doi.org/10.1080/15440478.2017.1340220

Taylor & Francis
Taylor & Francis Group

Clustering and Classification of Cotton Lint Using Principle Component Analysis, Agglomerative Hierarchical Clustering, and K-Means Clustering

Edwin Kamalha [a,b,c], Jovan Kiberu[a], Ildephonse Nibikora[a], Josphat Igadwa Mwasiagi[d], and Edison Omollo[e]

[a]Department of Textile and Ginning Engineering, Busitema University, Tororo, Uganda; [b]ENSAIT-GEMTEX Laboratory, Lille 1 University of Science & Technology, Roubaix, France; [c]Department of Applied Sciences and Technology, Politecnico di Torino, Corso Duca degli Abruzzi, Torino, Italy; [d]School of Engineering, Moi University, Eldoret, Kenya; [e]Department of Fashion and Textiles, Technical University of Kenya, Nairobi, Kenya

ABSTRACT

Cotton from the three cotton growing regions of Uganda was characterized for 13 quality parameters using the High Volume Instrument (HVI). Principal Component Analysis (PCA), Agglomerative Hierarchical Clustering (AHC) and k-means clustering were used to model cotton quality parameters. Using factor analysis, cotton yellowness and short fiber index were found to account for the highest variability. At 5% significance level, the highest correlation (0.73) was found between short fiber index and yellowness. Based on Cotton Outlook's world classification and USDA Standards, the cotton under test was deemed of high and uniform quality, falling between *Middling* and *Good Middling* grades. Our suggested classification integrates all lint quality parameters, unlike the traditional methods that consider selected parameters.

KEYWORDS

Agglomerative hierarchical clustering (AHC); classification; cotton quality; high volume instrument (HVI); k-means clustering; principal component analysis (PCA)

摘要

从乌干达三棉产区棉花的特点 13 质量参数使用高容量仪器（ HVI ）。主成分分析（ PCA ），凝聚层次聚类（ AHC ）和 k 均值聚类模型用于棉花质量参数。采用因子分析、棉花枯黄、短纤维指数都占最高的变异。在 5% 的显著水平的相关性最高（ 0.73 ）发现短纤维指数和黄色之间。根据棉花展望世界分类标准和美国农业部标准，被测棉花质量高，质量均匀，中等和中高档之间。我们建议的分类集成所有皮棉质量参数，不像传统的方法考虑选定的参数。

关键词

凝聚层次聚类（AHC）; 分类; 棉花质量; 高容量仪器（HVI）; k-均值群集; 主成分分析（PCA）

Introduction

Cotton Quality and HVI measurement

Cotton is the most globally traded fiber in conventional textile use. Some reasons for apparel preference have been discussed(Kamalha et al. 2013; Norum and Ha-Brookshire 2011). Quality of cotton fiber is judged on many factors, such as staple length, maturity, fineness, cleanliness, stickiness and strength, to mention but a few. Cotton quality characteristics are of importance to farmers, traders, researchers, and cotton spinners. Cotton fiber properties are influenced by a number of factors, including; type of breed or species, farming and harvesting methods, environmental and climatic profiles, processing methods (such as ginning), storage and handling among others. These fiber characteristics have inherent effect on

CONTACT Edwin Kamalha edwinkam11@gmail.com Department of Textile and Ginning Engineering, Busitema University, P.O Box 236, Tororo, Uganda.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/WJNF.

processing capabilities or requirements, yarn properties (such as evenness, strength and fineness) and fabric quality (such as dyeing quality, strength, and appearance)(Judith and Davidonis 2000). With precise measurement and classification techniques, standardization is achieved at different levels in the cotton supply chain(Judith and Davidonis 2000; Sharma 2014). Data from HVI measurement is usually of wider dimension, offering difficulty in interpretation. Often, cotton fiber parameters show non-linear relationships, hence robust data visualization and analysis is necessary to present such complex data for decision making. Common statistical methods such as linear regression and measure of central tendency are inferior for this cause(Mwasiagi, Wang, and Huang 2009).

Decades past, human classers were the basis for determining cotton quality. Often, classers had limitations in evaluating some properties like strength and elongation, maturity, short fiber content and fineness. Of the many systematic cotton measurement systems through the 1940's to the 70's, the HVI has been named as the most versatile, reliable and dependable testing system for most important cotton quality characteristics(Schleth, Furter, and Ghorashi 2006; The United States Department of Agriculture/ USTER 2006). The High Volume Systems have evolved to cover over ten cotton parameters, which include length, maturity, strength, trash percentage, color, fineness and Spinning consistency index (SCI). The US Department of Agriculture (USDA) exclusively uses this measurement system, among which is the USTER HVI 1000, which is the world reference for cotton classification(Furter 2009).

There are four cotton growing zones in Uganda; Northern, West Nile, Eastern and Kazinga channel (Western). Uganda's cotton quality classification is monitored and graded by the Cotton Development Organization (CDO), with reference to the World Cotton Outlook(Government of Uganda 2014; International Trade Centre 2011; Lubwama 2012). All Uganda's cotton is handpicked, and over 96% of this cotton is roller ginned. Our study focused on lint picked from three regions, which account for the largest proportion of Uganda's cotton yield. In the following sections, we briefly introduce fundamentals of the multivariate techniques used in our study.

Principal component analysis (PCA), Agglomerative hierarchical clustering (AHC) and k-means clustering

PCA analyzes data in which observations are described by several inter-correlated quantitative dependent variables. PCA extracts the most useful information and presents it in a new space, as a set of linear and *orthogonal* variables called *principal components* (also called factors) (F1+F2. . ..+Fn): where n is the total number of variables. Each variable or observation is represented on each principle component by a geometric projection known as a *factor score or factor loading*(Dray 2008; Jolliffe 2002). Closely related variables essentially load similarly on a principal component. A variable's significance to a factor is represented by the percentage contribution or the factor loading of the variable on the particular component(Abdi 2007). Correlation plots, scree plots and bi-plots among others can be used to visualize PCA relationships. The *squared cosine* represents the component's contribution to the squared distance of a variable or observation to the origin(Dray 2008; Kruskal 1978). Variables or observations with larger *squared cosines* are significantly important to a component and so is their importance to the total variability.

Hierarchical (connectivity) clustering establishes a hierarchy of clusters of objects on a set of quantitative attributes, yielding multiple levels of abstraction of the original data set. Unlike *divisive* clustering, AHC algorithms are a "bottom up" iterative classification technique, where observations start in their own clusters, and pairs of clusters are merged up the hierarchy (Addinsoft 2015a). Clustering of objects is based on combinations that minimize a given agglomeration criterion. Often, a metric, indicating the distance between pairs of observations, is used together with a linkage criterion which determines the distance between sets of observations. Commonly used metrics include: *Manhattan distance, Euclidean distance*, and s*quared Euclidean distance*. While, linkage criterion include: *minimum within class variance, mean linkage clustering, weighted pair group mean*, and *centroid linkage clustering*(Addinsoft 2015a; MacKay 2003; O'Connor 1987; SAS INSTITUTE 2008; Ward 1963). A binary clustering tree known as a

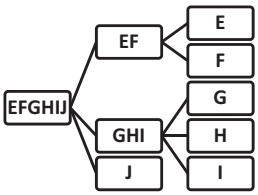

Figure 1. A sample dendrogram from AHC of objects EFGHIJ.

dendrogram is obtained (Figure 1), from which appropriate clusters may be selected. Graphically, the y-axis of the *dendrogram* represents the dissimilarity distance, while the x-axis represents items or observations. In our study, we performed AHC utilizing the squared Euclidean distance and the weighted pair-group average.

Like AHC, k-means clustering is an iterative method which, converges on a solution, but solutions obtained by k-means can vary for different starting points(Mwasiagi, Wang, and Huang 2009). The definitive k-means algorithm works with in-memory information, yet it could be effectively stretched out for out-of-memory occupant datasets. For a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$ …, $\mathbf{x}_n$), k-means clustering aims at optimally dividing the n observations into k ($\leq n$) sets; $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ while minimizing the within-cluster sum of squares. This sum of squares is the squared Euclidean distance; hence, it is the "nearest" mean. Each observation is allocated to a cluster whose mean gives the least within-cluster sum of squares (Addinsoft 2015b; MacKay 2003). Algorithms locate nearest centers and clusters by distance measures. Some classification criteria include; *Trace (W), Determinant (W), Wilks lambda*, and *Median* among others (Addinsoft 2015b; Deza and Deza 2009; MacKay 2003; Mwasiagi, Wang, and Huang 2009). Our study algorithm was based on k-means clustering using the *Trace (W)/Median*.

Both AHC and k-means have merits and drawbacks with regard to efficiency (in computation), and effectiveness (in application). Unlike connectivity clustering (as for AHC), k- means is a centroid based clustering system that requires a preset number of clusters as input and are nondeterministic. Hierarchical clustering on the other hand does not require users to preset the number of clusters and most hierarchical algorithms are deterministic. Also, the AHC output hierarchical structure is more informative than the unstructured set of clusters returned by flat clustering of k-means. Additionally, k-means calculations require probing over the whole dataset on each cycle, and it will only focalize to a quality solution after a series of cycles. Finally, k-means clustering is associated with a linear complexity compared to the quadratic complexity common with hierarchical clustering algorithms.

A recent study was performed to elucidate cotton characteristics based on different harvesting systems(Kazama et al. 2015), using PCA, and hierarchical cluster methods. Earlier, (Mwasiagi, Wang, and Huang 2009) used k-means clustering and artificial neural network to classify cotton lint quality. We did not find further relevant literature indicating the use of these three multivariate tools in the relationship analysis of cotton quality. In our study, we performed PCA to transform cotton quality parameters into factorial axes and consequently analyzed samples-variables relationships. Particularly, we used PCA to elucidate, for possible quality control, the most contributing parameters to the quality variability. We also draw comparison between k-means and AHC profiled classes, integrating all HVI fiber quality characteristics in one analysis. Finally, our study compounds the comparison between cotton growing regions, albeit in one country. These studies can be expounded to compare growing countries and seasons among others and classify them accordingly.

Materials and methods

Materials

A total of 60 cotton lint samples were realized, from bales harvested in the 2013/2014 cotton growing season, representing the Northern, Eastern and Western regions of Uganda. Sampling was done in

accordance with ASTM D1441-12- standard practice for sampling cotton fibers for testing. The samples and datasets were labeled according to regions: SN for Northern, SW for Western and SE for Eastern. Following data cleaning, and pre-processing, only 42 samples were retained for further study. Prior o testing, the samples were then conditioned under standard atmosphere (21°C, 65% RH) for 24 hours following ASTM D1776/D1776M- standard practice for conditioning and testing textiles.

Methods

The USTER® HVI 1000 (USTER Technologies, Switzerland) was used to characterize cotton fiber quality (Table 1). Following normal distribution fitting, and Grubbs' test (Two-tailed test) for outliers(Grubbs 1969, 1950), we excluded outlying samples. Correlation analysis and factor analysis were carried out using PCA to elucidate relationships within the data. A dissimilarity analysis was also done for samples using the Euclidean distance measure, to ascertain the most dissimilar cotton based on regional clustering. PCA was then used to find the most important parameters explaining the variation in cotton quality. We finally performed AHC and k-means clustering to classify and profile cotton samples. PCA and K-means clustering were performed using R- 3.2.3 software (The R Foundation, Austria). We then performed AHC using XLSTAT 2014.5.03 (Addinsoft, USA) and obtained consequent quality profiles and clusters, which we compared with k-means clustering results. In our study, AHC was computed utilizing the squared Euclidean distance as metric measure of dissimilarity and the weighted pair-group average as the linkage criteria (Addinsoft 2015a; SAS INSTITUTE 2009, 2008).

Results and discussion

Descriptive statistics for samples

Grubbs test for outliers at 99% confidence interval, significance 5%, yielded p-values and critical Z-scores to detect presence of outliers. We retained 42 (70%) samples whose summary statistics are recorded in Table 1. Fiber strength, elongation, trash count and short fiber index had the highest deviations from CDO's quality reference values. Coefficients of variation, CV (%) values of the samples, suggest that trash indicators presented the highest variance, followed by short fiber index. This is obvious as these variables are mostly dependent on human and processing factors. Maturity, uniformity index and whiteness were the least variant.

Table 1. Descriptive summary of cotton quality measurements from USTER® HVI 1000.

Quality attribute	Abbr	Unit	Min	Max	Mean	Stdev	Std error	CDO reference	CV %
Spinning consistency index	SCI	%	128	142	135.3	3.63	0.56	60 (min)	2.7
Micronaire	Mic	(-)	3.68	4.52	4.2	0.21	0.03	3.8–4.2	5.2
Maturity	Mat	%	0.86	0.88	0.87	0.007	0.00	0.85 (min)	0.8
Upper half mean length	UHML	mm	26.8	30	28.7	0.98	0.15	27 (min)	3.4
Uniformity index	UI	%	83.	85.9	84.2	0.78	0.12	85 (min)	0.9
Short fiber index	SFI	%	6.1	8.8	7.2	0.77	0.12	6 (max)	10.7
Fiber strength	Str	g/tex	26.6	30.4	28.2	0.95	0.15	30 (min)	3.4
Fiber Elongation	Elg	%	4.3	6.2	5.3	0.44	0.07	6.5 (min)	8.3
Color grade: whiteness	Rd	%	73	76.2	74.4	0.79	0.12	74–76	1.1
Color grade: Yellowness	+b	(-)	9.7	12.4	10.5	0.80	0.12	7–10	7.6
Trash particle (count)	TrCnt	(-)/g	17	49.0	29.5	8.36	1.29	20 (max)	28.4
Trash (area)	TrAr	%/g	0.16	0.78	0.44	0.17	0.03	2 (max)	38.4
Leaf grade	Lfgd	(-)	1.0	7.0	3	1	0.18	4 (max)	35.2

**Abbr=Abbreviation; Min=Minimum; Std error=Standard error.

Table 2. Pearson correlation coefficients between variables.

Variables	SCI	Mic	Mat	UHML	UI	SFI	Str	Elg	Rd	+b	TrCnt	TrAr	Lfgd
SCI	**1.00**	0.22	0.03	0.03	−0.20	0.03	0.05	0.07	−0.18	−0.09	0.26	**0.38**	0.09
Mic	0.22	**1.00**	0.25	0.13	0.17	**−0.60**	−0.15	**0.39**	−0.16	**−0.53**	**0.33**	0.25	0.26
Mat	0.03	0.25	**1.00**	−0.19	0.19	**−0.39**	0.10	0.05	0.10	−0.20	0.24	0.14	0.07
UHML	0.03	0.13	−0.19	**1.00**	0.10	−0.21	−0.08	0.16	0.13	−0.18	0.11	0.01	0.09
UI	−0.20	0.17	0.19	0.10	**1.00**	**−0.37**	0.03	**0.43**	−0.12	−0.27	0.08	−0.19	0.09
SFI	0.03	**−0.60**	**−0.39**	−0.21	**−0.37**	**1.00**	−0.09	**−0.52**	−0.15	**0.73**	**−0.48**	**−0.38**	**−0.32**
Str	0.05	−0.15	0.10	−0.08	0.03	−0.09	**1.00**	−0.02	0.23	−0.05	0.14	0.17	−0.15
Elg	0.07	**0.39**	0.05	0.16	**0.43**	**−0.52**	−0.02	**1.00**	−0.17	**−0.58**	**0.33**	0.27	**0.35**
Rd	−0.18	−0.16	0.10	0.13	−0.12	−0.15	0.23	−0.17	**1.00**	−0.11	0.26	0.23	0.15
+b	−0.09	**−0.53**	−0.20	−0.18	−0.27	**0.73**	−0.05	**−0.58**	−0.11	**1.00**	**−0.54**	**−0.44**	**−0.42**
TrCnt	0.26	**0.33**	0.24	0.11	0.08	**−0.48**	0.14	**0.33**	0.26	**−0.54**	**1.00**	**0.75**	**0.55**
TrAr	**0.38**	0.25	0.14	0.01	−0.19	**−0.38**	0.17	0.27	0.23	**−0.44**	**0.75**	**1.00**	**0.50**
Lfgd	0.09	0.26	0.07	0.09	0.09	**−0.32**	−0.15	**0.35**	0.15	**−0.42**	**0.55**	**0.50**	**1.00**

Values in bold are different from 0 with a significance level alpha=0.05.

Factor analysis of variables and samples on principal components

We normalized data for PCA using the zero-mean normalization-$(xi - meanA)/StdevA$; x_i is an entry in a column A. Based on Pearson (n) correlation computation, at significance level 0.05 (Table 2), the 13 principal components/factors F1-F13 (Table 3) are presented, with varying eigenvalues and explanation rates. Generally, no single factor explained the variability significantly, suggesting that at least five principle factors were important. The first component, F1 explains the maximum variability (0.3097), corresponding to an eigenvalue of 4.03. For further analysis, we retained only components F1 to F5, which had eigenvalues ≥ 1.0 (accounting for 73.6% of variability). From the eigenvector matrix, we computed a correlation matrix; variables/factors (factor loadings matrix), from which, a matrix of squared cosines of variables was computed (summary in Table 4). The squared cosines represent the significance of each variable to the principle components, and so to the overall variability. The highest correlation was between trash particle count and trash area (0.75), followed by short fiber index and yellowness (0.73). There was also moderate correlation among trash indicators; trash area, leaf grade and trash particle count. Micronaire and short fiber index had the highest negative correlation.

Generally, most parameters were strongly independent without obvious variation in respect to a host of other variables. From Figure 2, Figure 3, and Table 4, we can elucidate that cotton yellowness accounts for the highest variability in fiber quality, closely followed by short fiber index and trash particle count. Of successive importance were: fiber length, uniformity index, trash area, strength and elongation, in sequence.

Spinning consistency index and leaf grade were of least pertinence to the PCA model. This finding implies that critical control of yellowness and short fiber index would lead to control of several other parameters. This can be said to be true, since trash indications practically have influence on most physical properties such as length, strength, elongation, and colour. High trash content lowers the quality grade of cotton.

Table 3. PCA explanation rate for variability (Eigenvalues).

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
Eigenvalue	4.03	1.82	1.45	1.27	1.00	0.90	0.65	0.43	0.41	0.37	0.30	0.23	0.15
Variability (%)	30.97	13.97	11.18	9.76	7.68	6.92	4.97	3.30	3.16	2.85	2.30	1.75	1.17
Cumulative %	30.97	44.94	56.12	65.88	73.56	80.5	85.5	88.8	91.9	94.77	97.1	98.8	100

Table 4. Squared cosines of the variables as representative contribution to total variability.

Variable	+b	SFI	TrCnt	UHML	UI	TrAr	Str	Elg	Mic	Rd	Mat	Lfgd	SCI
Largest Squared cosine	0.7	0.68	0.61	0.48	0.46	0.46	0.43	0.43	0.42	0.39	0.39	0.38	0.36
Component	F1	F1	F1	F4	F2	F1	F5	F1	F1	F3	F4	F1	F3

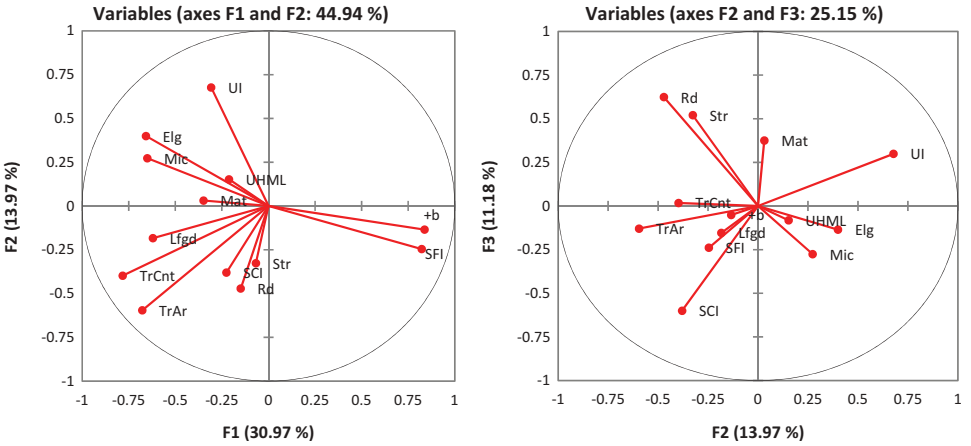

Figure 2. Cotton quality variables' correlation plots for factors F1, F2 and F3.

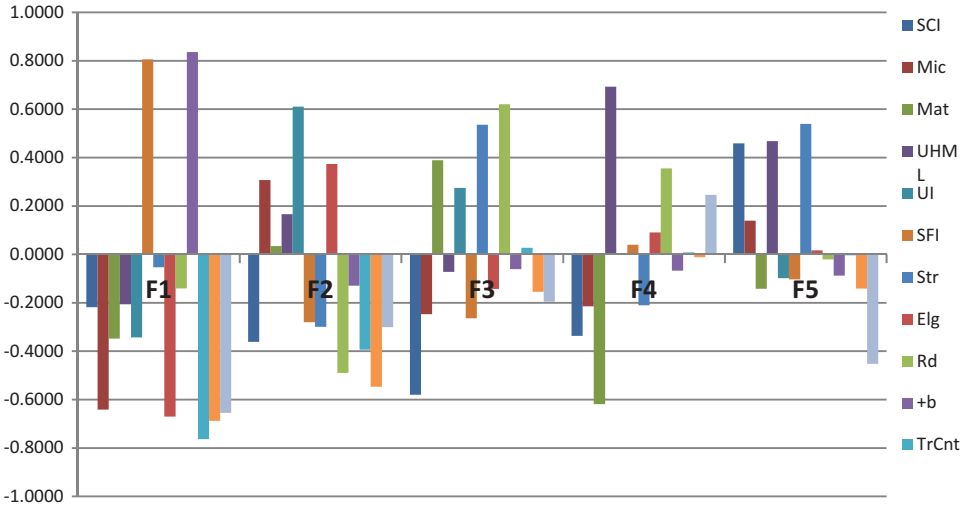

Figure 3. Plot of factor loadings for cotton quality parameters on F1-F5.

We also observed that F1 chiefly defines +b, SFI, Trcnt, TrAr, Elg, Mic and Lfgd; F2 defines UI; F3 defines Rd and SCI; F4 defines UHML and Mat; and F5 defines Str (Figure 3 and Table 4). This taxonomy of cotton quality parameters can be rightly coded and classified accordingly. It was found that parameters defining trash and fineness loaded similarly, while maturity loads with fiber length. The loading of samples (Figure 4) presents a clustering map between regions. Generally, samples from the Western and Nothern regions were clustered closer and load similarly compared to samples from the Eastern region.

In addition, there was higher variance within samples from the Eastern region compared to within other regions. The Northern region cotton was the most closely and uniformly clustered. Considering the Euclidean distance between regions, the highest dismilarity was found between samples from the Western and Eastern regions. The Western and Northern regions' cotton was closest in quality similarity. Cotton from Eastern Uganda had more short fibers and a yellowness index compared to the other regions (Figure 3 and Figure 4)

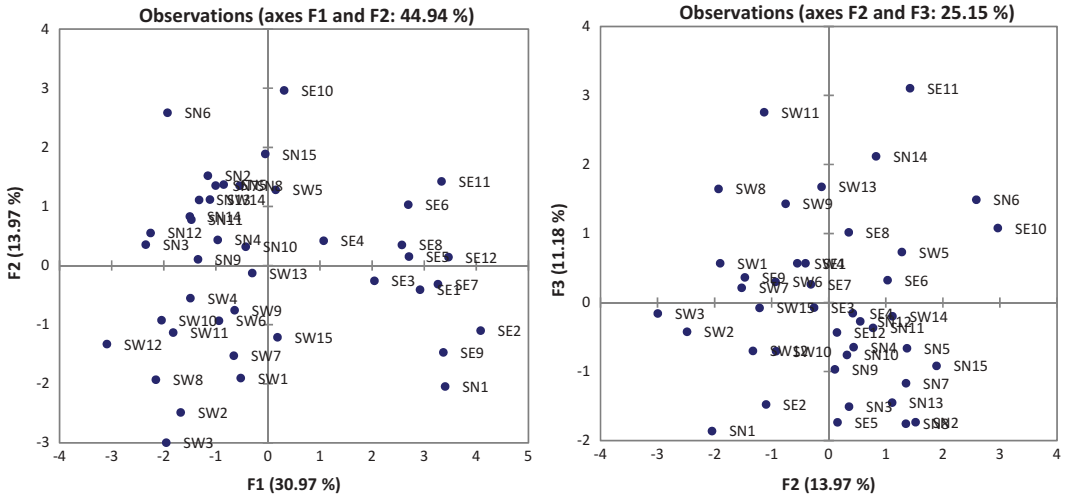

Figure 4. Plots of samples on principal components (factor scores) F1, F2 and F3.

Also, PCA clustering depicts that the Northern region cotton registered the highest values for micronaire, length, elongation, uniformity, and maturity than the counterparts. While, the Western region had cotton with the highest trash count/area, leaf grade (worst) and spinning consistency index. Variations in these properties can be a result of variations in soils, climate, farming and processing. Through efficient blending and mixing of fibers from such different regions, yarn spinners would achieve better quality yarn.

AHC and k-means Cotton Quality classification

From AHC and k-mean clustering algorithms, we elicited three classes to compare samples contained, and profiles corresponding (Table 5, Table 6, and Figure 5).

Under k-means, we pre-specified three (3), as the number of classes based on AHC returned optimum upon convergence. The metrics and linkage criterion have already been mentioned in the methods section.

Table 5. AHC classes and clusters of cotton samples for n = 3.

Class	1		2					3
No. of samples	11(26.2%)		30(71.4%)					1(2.4%)
	SN1	SE6	SN2	SN8	SN14	SW5	SW11	SE11
	SE1	SE7	SN3	SN9	SN15	SW6	SW12	
	SE2	SE8	SN4	SN10	SW1	SW7	SW13	
	SE3	SE9	SN5	SN11	SW2	SW8	SW14	
	SE4	SE12	SN6	SN12	SW3	SW9	SW15	
	SE5		SN7	SN13	SW4	SW10	SE10	

Table 6. k-means classes and clusters of cotton samples for n = 3.

Class	1	2					3	
No. of samples	6(14%)	26(62%)					10(24%)	
	SN1	SN2	SN11	SW3	SW11	SE10	SN4	SW12
	SE2	SN3	SN12	SW4	SW13	SE11	SN8	SE1
	SE5	SN5	SN14	SW6	SW14		SN10	SE3
	SE7	SN6	SN15	SW7	SW15		SN13	SE4
	SE9	SN7	SW1	SW9	SE6		SW5	
	SE12	SN9	SW2	SW10	SE8		SW8	

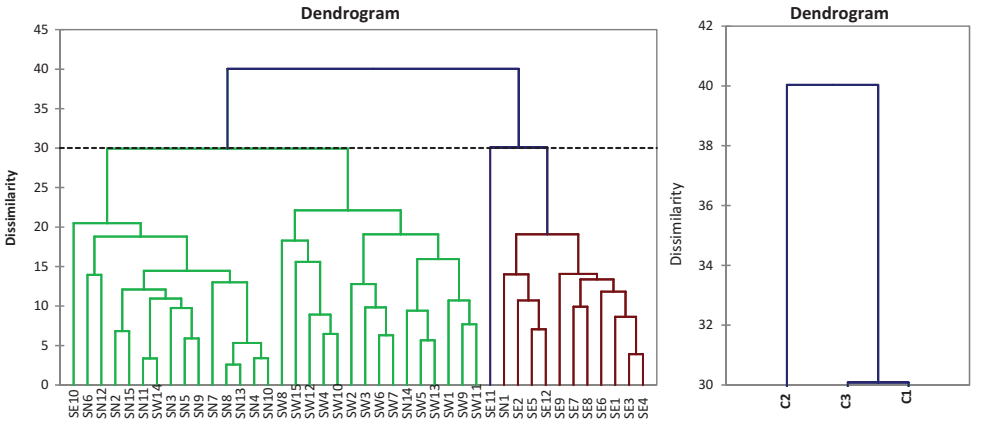

Figure 5. Schematic plot of AHC dendrogram of the 42 samples for classes 1, 2 and 3.

Data normalization was used to suppress the dominance of variables defined by larger magnitudes, since the clustering methods are magnitude responsive. With Normalization accords equal importance to each variable, while recognizing the ranks of individual entries through scores (weight).

AHC and k-means profiles, composition and clusters were slightly related. These slight differences arise from the differences in metrics and linkage criterion already presented. Most samples fell within class 2; comprising mainly of the Northern and Western regions' cotton. This clustering is in agreement with our PCA results. Save for a few samples, the Eastern region cotton shows a detachment from the other two regions in terms of quality clustering. The maximum number of possible classes is equivalent to the number of objects. However, higher numbers of classes, lead to low quality of clustering despite the fact that the "within class variance" also lowers with increasing number of classes. For both AHC and k-means clustering, we established that centroids of classes 2 and 3 are the closest, giving more credence that cotton from the Northern and Western Uganda have closely related quality compared to cotton from Eastern Uganda (Table 7 and Table 8). The AHC dendrogram (Figure 5) shows the clustering and closeness of classes.

Sample SE11 most portrays an outlying property. We computed the centroid values (Table 7 and Table 8) and a profile plot (Figure 6) for the three classes, for quality classification and profiling. The class centroids indicate expected thresholds within the different classes of cotton fiber quality. Class 2 represents the vast of the cotton samples, hence adopted for a general representation of Ugandan's cotton for the 2013/2014 season.

Using this classification, and basing on CDO quality standards (Table 1) matched with the World Cotton Outlook values, Uganda's cotton is generally of high spinning consistency index, high

Table 7. AHC class centroids as indicative values of cotton quality classification.

Class	SCI	Mic	Mat	UHML	UI	SFI	Str	Elg	Rd	+b	TrCnt	TrAr	Lfgd
1	135	4.03	0.86	28.5	83.7	8.2	27.8	4.8	74.3	11.6	21	0.30	3
2	136	4.28	0.87	28.8	84.4	6.8	28.2	5.5	74.4	10.1	33	0.50	4
3	129	3.68	0.87	28.8	85.5	7.7	29.9	5.1	74.5	11.6	17	0.16	1

Table 8. k-means class centroids as indicative values of cotton quality classification.

Class	SCI	Mic	Mat	UHML	UI	SFI	Str	Elg	Rd	+b	TrCnt	TrAr	Lfgd
1	140	3.7	0.86	26.78	83.3	8.7	27.1	4.8	73.9	12.3	27	0.41	4
2	137	4.51	0.87	28.44	85.1	6.9	27.9	5.3	73.5	10.4	35	0.34	3
3	135	4.33	0.86	30.45	84.1	6.6	26.8	5.4	73.5	10.2	25	0.43	4

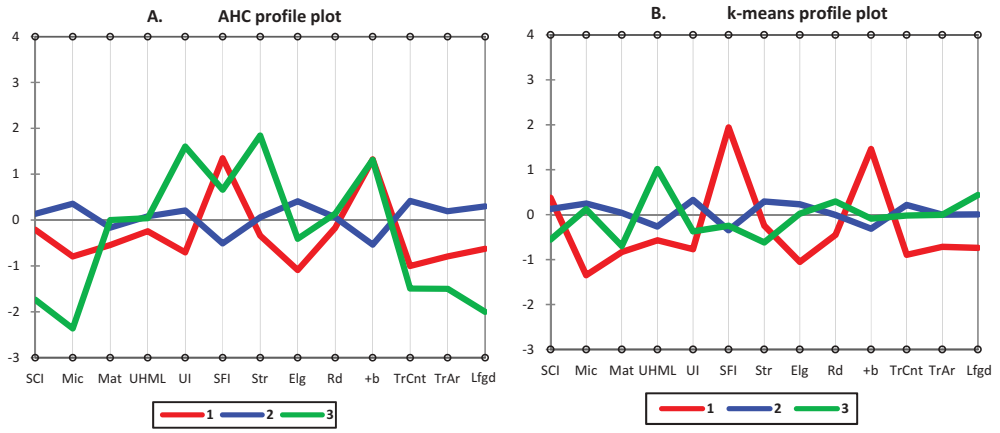

Figure 6. Cotton quality profile plots: **A**- AHC; **B**-k-means, for n = 3.

micronaire, high maturity, long staple length, high uniformity, normal short fiber fraction, normal strength, normal elongation, high reflectance, normal yellowness, high trash particle count but very low trash area, and average leaf grade. Internationally, this cotton qualifies for *Middling* to *Good Middling* grading (USDA equivalent standards). These characteristics suggest that Uganda's cotton falls in the premium range of fibers often listed as basis for the Cotlook A' Index computations.

The quality profile plots (Figure 6) indicate variation of quality attributes between classes. For instance, considering AHC, class 1 is defined by the lowest values of length, uniformity index, and elongation, and highest for short fiber index, and yellowness. Class 2 has the highest values for trash indicators, elongation, length, micronaire and spinning consistency index; however with the lowest yellowness and short fiber index. With k-means clustering, the trend is slightly different, with class 2 having higher spinning consistency index and trash count. Our k-means results are similar to those obtained by Mwasiagi's team(Mwasiagi, Wang, and Huang 2009) indicating that longer cottons often have higher values of micronaire, maturity, spinning consistency index, strength, uniformity, elongation and reflectance, but lower values for yellowness, short fiber index and trash measurements values.

Our AHC and k-means results are slightly different, although with some similarities. This is partly explained by the different metrics used in each method, and the nature of convergence of each algorithm. For cotton classification, the use of AHC would be more representative, since AHC compares each sample to all samples within the subset, linking closely related objects. On the contrary, k-means collects objects to a nearest centre drawn by a criterion. Hence, for k-means, objects could belong to more than one centre (class) depending on criteria, and re-allocation. Depending on interests or preference of a cotton buyer or spinner, and different applications, cotton fiber quality classification can also be tailored to include selected characteristics. Our suggestion for AHC over k-means is not absolute; k-means clustering has been found exceptionally handy for a corpus of commonsense settings and widely used applications. Many researchers have focused on increasing the performance of the algorithm by reducing the amount of passes needed for 5-means(Bottou and Bengio 1995). Hence, improvements have been introduced, with new derivatives for targeted applications(Bottou and Bengio 1995; Kulis and Jordan 2011). However, these methodologies often give surmised results, with possibility of deterministic or probabilistic limits. A key preference of 5-means is it's convergence to a local minimum, which does not hold accurate for the estimated versions.

Conclusion

In view of the high dimensionality of HVI data, multivariate statistics are a versatile tool in mining and analysis. PCA, AHC and k-means clustering were effectively applied to the data in question, and

relationships were drawn among quality characteristics. Of particular importance is to use PCA along with any of the two clustering techniques. We established that cotton from Uganda's Eastern region is slightly different in quality from that produced in the North and West of the country, for the particular cotton season studied. Three unique classes of cotton quality were drawn by AHC and k-means clustering. It is upon the relativity in application and effectiveness of computation that one would opt for one method over the other. We assert that AHC is a better method in view of the connectivity relationship that fixes an object to one class, for similar iterations, which is not true for k-means clustering. Using these methods together with the USDA standards, this batch of Uganda's cotton can be classified as *middling to Good Midlling*. We also established that cotton yellowness and short fiber index closely account for much of the variability in cotton fiber quality, and that HVI quality parameters are less related, and are highly independent of each other. The highest correlation (0.73) was found between cotton yellowness and short fiber index.

Our future focus is to analyse Uganda's cotton quality data from preceding and current seasons, for comparative purposes and to affirm findings in the regional cotton differences.

Acknowledgment

The authors recognize the assistance of Cotton Development Organization (CDO), Uganda for the HVI equipment.

ORCID

Edwin Kamalha http://orcid.org/0000-0002-2923-8886

References

Abdi, H. 2007. Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition(GSVD). In *Encyclopedia of Measurement and Statistics*. ed. Neil Salkind, Thousand Oaks, CA: Sage Publications.

Addinsoft. 2015a. Agglomerative Hierarchical Clustering (AHC). https://www.xlstat.com/en/solutions/features/agglomerative-hierarchical-clustering-ahc.

Addinsoft. 2015b. K-Means clustering. https://www.xlstat.com/en/solutions/features/k-means-clustering.

Bottou, L., and Y. Bengio. 1995. Convergence properties of the K-Means Algorithms. *In Advances in Neural Information Processing Systems* 7:585–92.

Deza, E., and M. M. Deza. 2009. *Encyclopedia of distances*, 1st ed. Berlin, Heidelberg: Springer-Verlag.

Dray, S. 2008. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics & Data Analysis* 52 (4):2228–37. doi:10.1016/j.csda.2007.07.015.

Furter, R. 2009. *The standardazation of quality characteristics in the textile supply chain*. USTER Technologies Application Report- Quality Management. Sonnenbergstrasse, Zurich, Switzerland.

Government of Uganda. 2014. *Investment opportunities in Uganda's cotton sector.* Uganda Investment Authority (UIA), Working Report. Kampala, Uganda.http://www.ugandainvest.go.ug/uia/images/Download_Center/SECTOR_PROFILE/Cotton_Sector_Profile.pdf.

Grubbs, E. F. 1950. Sample criteria for testing outlying observations. *Annals of Mathematical Statistics* 21 (1):27–58. doi:10.1214/aoms/1177729885.

Grubbs, E. F. 1969. Procedures for detecting outlying observations in samples. *Technometrics* 11 (1):1–21. doi:10.1080/00401706.1969.10490657.

International Trade Centre. 2011. *Cotton from Uganda.* Agricultural Commodities Programme. Kampala, Uganda. http://www.coton-acp.org/sites/default/files/documents/downloads/final_uganda_brochure_october_2011.pdf.

Jolliffe, I. T. 2002. *Principal component analysis springer*. New York, NY: Springer-Verlag.

Judith, M. B., and H. G. Davidonis. 2000. Quantitation of fiber quality and the cotton production-processing interface: A physiologist's perspective. *The Journal of Cotton Science* 4:34–64.

Kamalha, E., Y. Zeng, M. I. Josphat, and S. Kyatuheire. 2013. The comfort dimension; a review of perception in clothing. *Journal of Sensory Studies* 28 (6):423–44. doi:10.1111/joss.12070.

Kazama, H. E., M. F. Ferreira, P. R. Da Silva, R. B. A. Da Silva, and A. D. Fiorese. 2015. Multivariate analysis of fiber characteristics of dense cotton in different harvest systems. *Australian Journal of Crop Science* 9 (11):1075–81.

Kruskal, J. B. 1978. Factor analysis and principal component analysis: Bilinear methods. In *International Encyclopedia of Statistics.* Eds. Kruskal, W. H., and Tanur, J. M. New York, NY: The Free Press.

Kulis, B., and M. I. Jordan. 2011. Revisiting K-Means: New algorithms via Bayesian nonparametrics. *arXiv* 14. http://arxiv.org/abs/1111.0352.

Lubwama, D. 2012. *Current status of cotton research and production trends in Uganda.* 11th meeting of the International Cotton Advisory Committee (ICAC), Southern and Eastern Africa Cotton Forum (SEACF). Nyeri, Kenya. 27-29 August, 2012. Kampala, Uganda. https://www.icac.org/wp-content/uploads/2012/09/Uganda_Report1.pdf.

MacKay, D. 2003. An example inference task: Clustering. In *Information Theory, Inference and learning algorithm*. ed. David J.C. MacKay. 284–92. Cambridge, UK: Cambridge University Press.

Mwasiagi, I. J., X. H. Wang, and X. B. Huang. 2009. The use of K-Means and artificial neural network to classify cotton lint. *Fibers and Polymers* 10 (3):379–83. doi:10.1007/s12221-009-0379-z.

Norum, P. S., and J. E. Ha-Brookshire. 2011. Consumer trade-off analysis and market share estimation for selected socially responsible product attributes for cotton apparel. *Clothing and Textiles Research Journal* 29 (4):348–62. doi:10.1177/0887302X11425956.

O'Connor, C. 1987.*An introduction to multivariate statistical analysis. Computers & Mathematics with applications*, Vol. 14, 3rd ed., Huboken, New Jersey: A John Wiley & Sons, Inc.

SAS INSTITUTE. 2008. The distance procedure: Proximity measures. In *SAS/STAT® 9.2 user's guide*, 1483–531. Cary, NC: SAS Institute Inc.

SAS INSTITUTE. 2009. The cluster procedure: Clustering methods. In *SAS/STAT 9.2 users guide* 2nd ed., 7745. Cary, NC: SAS Institute Inc.

Schleth, A., R. Furter, and H. Ghorashi. 2006. *The role of cotton classification in the* textile industry. *The Fiber Classification and Analysis System. USTER HVI 100 Application Report*. Sonnenbergstrasse, Zurich, Switzerland.

Sharma, S. K. 2014. Cotton yarn : Quality depends on mixing strategy. *The Indian Textile Journal* 124 (6):35–42.

The United States Department of Agriculture/USTER. 2006. *Quality characteristics used for cotton classification*. USTER HVI 100 Application Report. Sonnenbergstrasse, Zurich, Switzerland.

Ward, H. J. 1963. Hierarchical grouping to optimize an objective function. *American Statistical Association* 58 (301):236–44. doi:10.1080/01621459.1963.10500845.