3(3): 1-13, 2019; Article no.AJPAS.48316

Statistical Monitoring of Tobacco Moisture Using the Particle Size Distribution Data-a Multidimensional Scaling Audit

Chisimkwuo John1* and Chukwuemeka O. Omekara1

1 Department of Statistics, Michael Okpara University of Agriculture, Umudike, Nigeria.

Authors' contributions

This work was carried out in collaboration between both authors. Author CJ designed the study, performed the statistical analysis, wrote the protocol, wrote the first draft of the manuscript and managed the analyses of the study. Authors CJ and COO managed the literature searches. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2019/v3i330096 *Editor(s):* (1) Dr. S. M. Aqil Burney, Department of Computer Science, University of Karachi, Pakistan. *Reviewers:* (1) Yahaya Shagaiya Daniel,Kaduna State University, Nigeria. (2) Janilson Pinheiro De Assis, Federal University Of Semi-Arid (UFERSA), Brazil. Complete Peer review History: http://www.sdiarticle3.com/review-history/48316

Original Research Article

Received: 25 January 2019 Accepted: 04 April 2019 Published: 16 April 2019

Abstract

Tobacco manufacturers see the tobacco moisture content as one of the determining factors in the quality of the finished tobacco product. During primary processing stage, the Particle Size Distribution (PSD) of the cut tobacco is a good measure of the tobacco moisture content. This paper presents statistical analyses of a two month PSD data using graphical techniques from noteworthy statistical multidimensional scaling (MDS) approaches in characterizing the tobacco moisture quality ratio. At the end, the evaluation within the investigated months fosters an indicative process audit, control and predictive monitoring that is capable of providing valuable impacts to future production.

Keywords: Metric MDS; procrustes analysis; correspondence analysis; PCA biplot; profile analysis; boxplots.

1 Introduction

Many statistical techniques have found remarkable usefulness in real life applications. In a fully integrated process, a careful exploratory procedure on the multivariate setting is capable of revealing historical

^{}Corresponding author: E-mail: john.chisimkwuo@mouau.edu.ng;*

activities in the manufacturing system. Exploring this intensively with statistical approaches such as modern multidimensional (MDS) approaches might serve as a useful statistical process improvement technique. Such critical checks could be referred to as process audit. In some manufacturing companies, the term auditing is only synonymous with financial activities. Most times, process activities are left out. Also, appraisals and efficiency are built around financial management without much credence to outstanding process performance and improvements.

However, it is important to note that a bad process performance will leave the industry with low income or in some cases, gross loss and debt. Thus, augmenting a financial audit with a proper statistical process audit may be reasonable. This could be carried out quarterly or annually as in the case of a financial audit and entails the full involvement of a high level Statistician with good knowledge of the contemporary process. Such approach would tend to bring outstanding process operational excellence in the manufacturing. The overall aim is not geared to penalize or police workers but to critically check possible ways of avoiding process failures and explore new ways of improvement from non failure dimensions of the manufacturing processes over time. Statistical Process Control (SPC) guidelines and systematic reporting approaches need to be followed in achieving this audit technique. Such reports might comprise duly and transparent activities and fault log keeping by the operators and engineers.

Also, MDS approaches need to be incorporated since they play an important role in exploring salient and useful information in a multivariate setting with little or no statistical data assumptions. Such useful MDS approaches usually use multivariate plots, which appeal more to manufacturers, to explore the historical relationships that have existed in the process over time. Interpretation of these plots backed by a suitable process experience readily brings out the needed information. In recent times, many literary articles have explored these MDS techniques in process monitoring and improvement. For instance, Yunus and Zhang used classical MDS and a Procrustes approach in process monitoring [1]; Aldrich et al*.* [2] used a biplot approach in monitoring metallurgical data sets; Wang et al*.* [3] used PCA and other classical methods to monitor batch processes, to mention but a few. Some of the useful MDS approaches include the Biplots, Correspondence Analysis Plots, Profile Analysis Plots, Procrustes Analysis Plots etc. Several authors have given in-depth methods of applying these approaches. Detailed references to some of these authors' guides are found in many literary books, for example: Gower et al*.*, Greenacre, Cox and Cox and lots of other books [4,5 and 6].

2 The PSD Problem, Data Collection and Description

Information for the evaluation of the tobacco moisture variability can be defined by considering the particle size distribution (PSD). This also provides a good description of the overall performance of the tobacco thrashing process within the primary manufacturing section. Moisture variability has been shown to have a relatively vital relationship with the PSD. The moisture importance to the PSD plays a big role since too dry tobacco will be cut as dust while a too wet tobacco will readily be cut as lumps of laminas and stems.

2.1 Design and data collection

In July and in August 2009, the PSD data of the tobacco grade (B305E) from a tobacco manufacturing plant in Nigeria were measured at the two sampling points for 14 different operation runs [7]. A reason for choosing this grade is because its raw material was supplied and stored the same time and does seem to be homogenous within the two months considered. See Appendix A for the Ex-ITM and Ex-ADD data sets. A systematic sampling design without replacement enhanced by the company's electronic capturing of primary data sets was utilized.

Samples at Ex.ITM and Ex.ADD sampling points are taken within an operation run. Samples were captured automatically by precision instruments every four seconds (fixed intervals) during a run constituting sample size $(n=15)$ for each minute and the mean per minute calculated to represent a run. This is then transferred to the sieving machine for a simple sieve separation and measurements (categorized by sieve diameters sieve1, sieve2, sieve3, sieve4 and Fines) are taken thereof as showcased in Appendix A. Note that the Fines are more or less, the dust and are not considered. Also, since the Sieve machine design gives more priority to sieve1 and sieve2 PSDs, as they mark the Quality Ratio (QR), only sieve1 and sieve2 data sets would be considered in this analysis.

2.2 The problem pathway

Two line managers in the primary manufacturing section operated during the two months (July and August 2009) data set considered here and management needs to appraise the line managers on an evidence of a better QR. This appraisal is of importance since the management might want to use this as an audit approach to critically explore how the managers performs and more importantly, reveal details at what level improvements were made. This later aspect is crucial as this would help to foster control and further improvements. Graphical analytical representations programmed using the R software (obtainable from http://www.cran.r-project.com) will be used in the analysis that will follow.

All the July PSD values will appear in red while all the August PSD values will appear in blue. The sum of sieve1 and sieve2 gives the Quality Ratio, and this would be compared over two months.

3 A Graphical Descriptive Comparison - The Profile Result

A first level approach would be to employ a profile analysis to explain the multivariate profile performance. Profile analysis pertains to situations in which a battery of p treatments are administered to two or more group of subjects with all responses expressed in the same units. Also, an assumption is that the different groups are independent of one another. One might pose the question, are the population mean vectors the same? This question of equality of means is divided into several possibilities. Of these possibilities, Johnson & Wichern enumerates that first is to check whether the means are parallel. And if they are parallel, are they coincident? And finally, if they are coincident, are they level [8]? The level test might not be of great importance in this analysis since the management had already pointed that appraisal must be made on the two managers. Suggesting that they had presumed variabilities and on this, they placed more importance on the differences. The basic mathematics suitable for programming the Profile Analysis is summarized in [8 p.323-328].

Consider the population means $\mu_A^T = [\mu_{A1}, \mu_{A2}, \mu_{A3}, \mu_{A4}]$ representing the average responses to the four sieve values for the July PSD and $\mu_B^T = [\mu_{B1}, \mu_{B2}, \mu_{B3}, \mu_{B4},]$ for the August PSD values, a profile plot is presented as shown in Fig. 1. The two profiles seem to be the same and the non rejection of the null hypothesis in the case of the test for parallelism (with calculated value = 1.26 and associated p-value of 0.31) as well as in the case of the coincidence test (with calculated value = 0.39 and associated p-value of 0.54) supports this.

As noted earlier, the management is looking beyond the average PSD values and would like to consider also the variability within these two months. A simple graphical way of showcasing this is using the Notched Box and Whisker plots presented in Figs. 2 and 3.

The Box and Whisker plots presented in Fig. 2 presents a simultaneous graphical display of the PSD data set. The notches can be regarded as nonparametric approximate 95% confidence intervals for the respective medians. Since all four notches show some degree of overlap, it can be concluded that the four medians do not differ significantly. However, observe the huge differences in the variability displayed by both the July Ex-ITM and Ex-ADD PSDs (plotted in red) when compared to those of their August counterparts (plotted in blue). One or some of the operation runs tend to differ from the others. This is readily justifiable as some of the July's Ex-ITM and Ex-ADD PSD confidence intervals extend outside the box. As such, this causes some notches to go outside the hinges (box). Another observation is to notice that the mean (denoted by a star

within the box) for July is not central within the box when compared to their August counterparts although the August Ex-ITM shows little of such property with lower variability.

QR Profile for July&Aug09 PSD dataset

Fig. 2. Box plots for July and August Ex-ITM and Ex-ADD

Fig. 3. Combined box plot for Ex-ITM and Ex-ADD

Furthermore, a pooled effect of the PSD values between the two months can be considered by adding the Quality Ratio of the Ex-ITM and Ex-ADD for the individual months. The Box and Whisker plots of the combined Ex-ITM and Ex-ADD data sets are presented in Fig. 3.

The plot in Fig. 3 reveals that the overall variability across the two measuring points in August is considerably lower than that in July. This suggests that the August operations tend to perform better although there is no significant difference between the medians as suggested by the overlap of the notches.

4 Tracing the Atypical Runs – The Procrustes Application

Meanwhile, the management has greater interest on the gains to the company if there are evidences on process changes made at distinct operations and more importantly, if there are exploratory ways to recognize those operations that were highly dissimilar to the July production. This would help to see what went right or wrong in those flagged operations. In other words, management needs to use this approach to learn more on some of the feasible improvement strategies in the manufacturing. The Box and Whisker plots of Fig. 3 have been used to affirm that the August variability is appreciably lower. Thus, there is need to know what operations went wrong in July when compared with those in August using the same run order of this blend.

A Procrustes analysis is a procedure for comparing two or more MDS configurations. Cox & Cox enumerates the use of the Procrustes analysis in the rotation, reflection, dilation, and translation of one configuration to fit a target configuration optimally [6 p.133]. The mathematical background for programming the Procrustes analysis using the R software is obtainable in [6 p.133 - 149].

One could possibly start the application by considering a metric MDS approach. For instance, suppose there are *n* objects with dissimilarity $\{\delta_{rs}\}\$. Metric MDS attempts to find a set of points in a space where each point represents one of the objects and the distances between points $\{d_{rs}\}\$ are such that $d_{rs} = f(\delta_{rs})$.

The function f is required to be a continuous parametric monotonic function. It can be the identity function or a function that attempts to transform the dissimilarities into a distance like form.

To perform the metric MDS, the data matrix $X: n \times 4$ of Sieve1.ITM, Sieve2.ITM, Sieve1.ADD and Sieve2.ADD was used to construct an $n \times n$ distance matrix of Euclidean distances. These are displayed in Figs. 4a and 4b for the July and August PSD data sets respectively.

Salient observations from Fig. 4a are to notice within the July PSD that run 10 and run 07 tend to be different from the other operation runs within the configuration. Also, it is apparent that the closeness between runs 4, 8 and 14 shows that they form a distinct group from the others runs. The August operation showcases an entirely clustered configuration for all the runs except for run 11 as could be seen in Fig. 4b. At this stage, these operations constitute the major runs that perform differently for both months.

However, a more appropriate insight to depicting the configurations that really gives a better fit would be to match or transform each of the months' PSD runs configurations to the other maintaining the same scaling approach in a Procrustes analysis. Thus, a Procrustes analysis approach is sought to compare the metric MDS configurations presented in Figs. 4a and 4b.

The results of these Procrustes analyses are shown in Figs. 5a and 5b for the July and August PSD runs respectively.

In Fig. 5a, a matching of the July PSD on the August PSD is made and the R^2 value of 0.0046 is considerably a minimum. The huge variabilities of run 07 and run 10 of the July data (in red) are obvious. Observe the homogeneity displayed by the August runs (in blue) in general except for its run 11. Also in Fig. 5b, a matching of the August PSD on the July PSD is made. Variability of run 11 is evident as well. Meanwhile, this approach has evidently from this configuration flag the operations 7 and 10 of July for further investigation. Although the August operation performs marginally better (with smaller variability as judged from the Notched Box & Whisker plots of Figs. 2 and 3), similar investigation needs to be launched for its operation run 11 for further improvement.

July 2009 PSD

Fig. 4a. Metric MDS for July PSD

Aug 2009 PSD

Procrustes analysis

 $R^2 = 0.0046$

Fig. 5a. Procrustes analysis for PSD (July matched to August)

Procrustes analysis

 $R^2 = 0.0046$

Fig. 5b. Procrustes Analysis for PSD (August matched to July)

5 Row to Column Characterization – A Correspondence Analysis Application

The Profile Analysis, the Notched Box and Whisker plots, the Metric MDS and the Procrustes analysis has so far pointed evidences of possible variabilities and this arouses the need to look at the sieves' data itself. The process engineering arm of the industry is believed to have an in-depth knowledge of the process behaviour in practice. Based on this, a characterization of the runs (rows) for each of the sieve values (columns) from the two sampling points (Ex-ITM and Ex-ADD) could help the process engineer to apply a pre-knowledge of the process performance in judging what might possibly be an improvement procedure. Thus, this suggests the application of the correspondence analysis. Correspondence analysis is an exploratory data analytic technique designed to analyze simple two-way tables containing some measure of correspondence between the rows and columns. As opposed to traditional hypothesis testing designed to verify a priori hypotheses about relations between variables, exploratory data analysis is used to identify systematic relations between variables when there are no (or rather incomplete) *a* priori expectations as to the nature of those relations; details are seen in [5 p.201-211]

Exploiting the exploratory capabilities of this approach, Figs. 6 and 7 presents the correspondence analysis (symmetric map) of the July and August PSD data sets respectively.

Note that the word *'symmetric map'* used in the captions of Figs. 6 and 7 implies that the map displays both the rows and column points in principal coordinates.

Firstly, a total Inertia of 95.28%, which is quite close to unity strongly suggests that the display in Fig. 6 nicely fits within the two dimensional configuration. The distinctness of operations 7 and 10 in Fig. 6 could easily be depicted and this is comparable to the observations of the Metric MDS of Fig. 4a and the Procrustes map of Fig. 5a. Meanwhile, added information is to notice the closeness of the operations 1 and 7 to Sieve1.ADD and the closeness of operation 10 to Sieve1 ITM. Also, observe that 9 operation runs are on the left hand side of the plot leaving only 5 operations dispersed on the right. With only Sieve1.ADD on the right hand side, the variability problems observed in the July PSD might easily be attributed to be majorly a Sieve1 problem. Meanwhile, key information is to observe that the concerned operations 7 and 10 are closely interlocked with Sieve1.ADD and Sieve1.ITM. This suggests that the problem might be a Sieve1 issue. Hence, the engineer needs to check and interpret Sieve1 engineering signals as a possible cause of the July issues. Critical study of the signals with current process data would foster more improvement.

Fig. 6. Correspondence analysis plot for July PSD data set (symmetric map)

Fig. 7. Correspondence analysis plot August PSD data set (symmetric map)

Fig. 7 has a total Inertia of 97.32% and again this is close to unity suggesting that the display nicely fits within the two dimensional configuration. Observe at first glance that all Sieves 1 and Sieves 2 values appeared on the opposite sides of the plot showing a sort of proportionality and randomness, which was not ascertainable in the case of the July's map of Fig. 6.

Furthermore, notice the random distribution of the runs within the configuration. Tracing the earlier identified operation 11 according to the MDS of Fig. 4b, this operation's closeness to Sieve1.ITM is obvious. Thus, the process engineer needs to track this situation in the light of the Sieve1.ITM problem to foster future improvement. On a general note, a suggestion for future improvement using this approach might be to look at all Sieve1 signals. This turns out to be a vital improvement strategy.

6 What About the Future – The Predictive PCA Biplot Application

A rich multidimensional graphical analytic approach to explore the relationship between the runs (rows) and the corresponding axes (the sieves) could be employed using the Principal Component Analysis (PCA) biplots. Given that the QR is a linear combination problem as it is obtained by adding Sieves1 and Sieves2, the PCA is suitable here since the PCA transforms a correlated and dependent data set into linearly independent principal components or latent values. Although this latent values may be meaningless and thus, the reason for the application of the Biplots which back transforms the data to its original scale and then displays it in a two dimensional configuration showing the rows and the columns in a single calibrated plot. Additional theories on biplots are obtained in [9] and [10] and other literary books in multivariate statistics.

Interestingly, the different month's psd process data could be reviewed for a better production. A method to perform this can utilize the predictivity power of the PCA biplot as discussed in [4 p.94] to predict the runs with highest and lowest QR within the calibrated configuration. A guideline on what will be valuable to the process engineer would be to make predictions of the maximum minimum QR's knowing that operations with could be tuned accordingly to produce sieve values that could yield the maximum QR. Knowing the bounds within which a contemporary process is operating at or could be tuned to will always be a great knowledge to the process operator for optimum QR.

Figs. 8 and 9 showcases the July and August QR.ADD data sets biplots with predictions of their highest and lowest QR. The predicted values are shown in Table 1 for the July and August data sets respectively. It is obvious from the quality of displays of Figs. 8 and 9 of 88.65% and 86.65% respectively is appreciably high enough for a reasonable reliance on any information obtained from the 2D configuration.

Fig. 8. PCA biplot for July PSD data set (Quality of Display = 88.65%)

| | July EX.ADD | | August EX.ADD | |
|------------------------|--------------------------|-----------------------|----------------------|------------------------|
| | Lowest r ₂ | Highest r 7 | Lowest r 05 | Highest r 11 |
| | | | | |
| sievel.itm | 36.369 | 40.2958 | 37.9384 | 41.965 |
| sieve _{2.itm} | 29.104 | 28.4277 | 29.8087 | 25.3358 |
| sievel.add | 23.553 | 47.7599 | 33.8037 | 26.8321 |
| sieve _{2.add} | 38.134 | 27.0336 | 35.7132 | 35.6822 |
| OR.ADD | 61.686 | 74.7935 | 69.5169 | 62.5143 |

Table 1. July and August EX.ADD predicted values using Figs. 8 and 9

Fig. 9. PCA biplot for August PSD data set (Quality of Display = 85.65%)

7 Conclusion

Multidimensional Scaling (MDS) approaches like Classical scaling, Profile Analysis, Procrustes and Correspondence analysis has proved to be a reasonable indicative statistical monitoring and improvement approaches on the case study considered. This is justifiable as the approach uses a statistical audit technique to access the moisture content of the primary tobacco process by looking at the Particle Size Distribution (PSD) data set. In summary, individual runs were traced to their inherent source problems by attributing them to their Sieve diameter sizes, and this proves to work in practice as the operators noted and redressed the indicative changes. Furthermore, highest and lowest predictions showing the limits of the QR obtainable in the contemporary process variables were made and this enhances the engineering control decision.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Yunus MYM, Zhang J. A Multidimensional Scaling based Process Monitoring, 20th European Symposium of Computer Aided Process Engineering, Elsevier B.V. 2010;28:1333-1342.

- [2] Aldrich C, Gardner S, Le Roux NJ. Monitoring of metallurgical Process Plants by using biplots. AIChE Journal. 2004;50.
- [3] Wang X, Kruger U, Irwin G. Process monitoring approach using fast moving window PCA. Industrial & Engineering Chemistry Research. 2005;44(14):5691-5702.
- [4] Gower JC, Lubbe GS, Le Roux NJ. Understanding biplots. Chichester: Wiley; 2011.
- [5] Greenacre M. Correspondence analysis in practice. 2nd Edition, New York: Chapman & Hall/CRC; 2007.
- [6] Cox TF, Cox MAA. Multidimensional scaling. 2nd Edition, New York: Chapman & Hall/CRC; 2001.
- [7] John C. Industrial statistical process controls a case study: Tobacco moisture variability reduction. Unpublished MComm/MSc thesis, University of Stellenbosch; 2011.
- [8] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th Edition, Upper Saddle River, NJ: Pearson International Edition; 2007.
- [9] La Grange A, Le Roux NJ, Gardner-Lubbe S. Biplot GUI: Interactive biplots in R. Journal of Statistical Software. 2009;30(12).
- [10] Gower JC, Hand DJ. Biplots, monographs of statistics and applied probability, Chapman and Hall: UK, In: LA GRANGE, A., LE ROUX, N.J. & GARDNER-LUBBE, S. (2009), Biplot GUI: Interactive Biplots in R, Journal of Statistical Software. 1996;30(12).

Appendix - A

Used Data Sets

___ *© 2019 John and Omekara; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Peer-review history:

The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)

http://www.sdiarticle3.com/review-history/48316 *http://www.sdiarticle3.com/review-history/48316*