

STATISTICAL MONITORING OF TOBACCO MOISTURE USING THE PARTICLE SIZE DISTRIBUTION DATA-A MULTIDIMENSIONAL SCALING AUDIT

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 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 (2010) used classical MDS and a Procrustes approach in process monitoring; Aldrich et al. (2004) used a biplot approach in monitoring metallurgical data sets; Wang et al. (2005) 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 the books, for example: Gower *et al.* (2011), Greenacre (2007), Cox and Cox (2001), Johnson and Wichern (2007), and lots of other books.

2. The PSD Problem, Data Collection & 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.

Design & 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 (John, 2011). 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.

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 (2007) 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? 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 Johnson & Wichern (2007).

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 Figure 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.

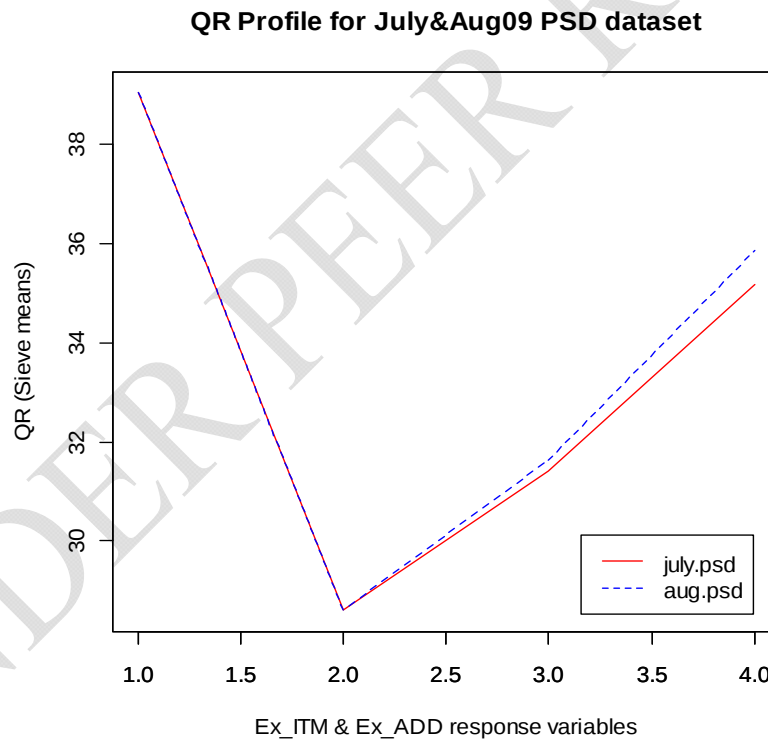


Figure 1: Profile plot for PSD data set

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 Figures 2 and 3.

The Box and Whisker plots presented in Figure 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.

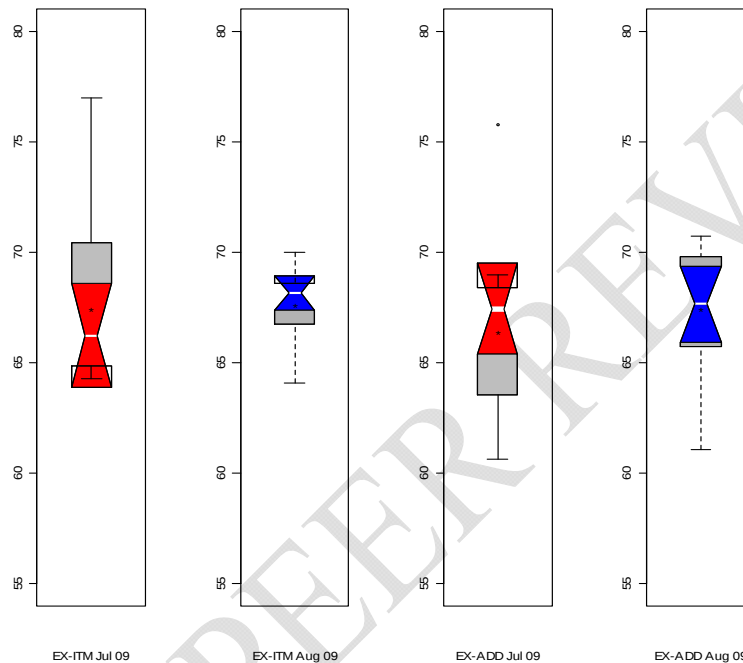


Figure 2: Box plots for July and August Ex-ITM and Ex-ADD

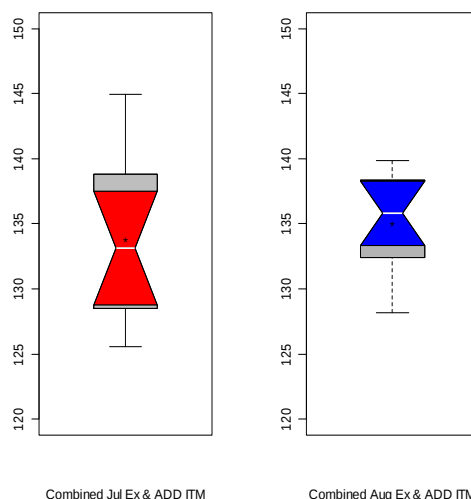


Figure 3: Combined box plot for Ex-ITM and Ex-ADD

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 Figure 3.

The plot in Figure 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 Figure 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.

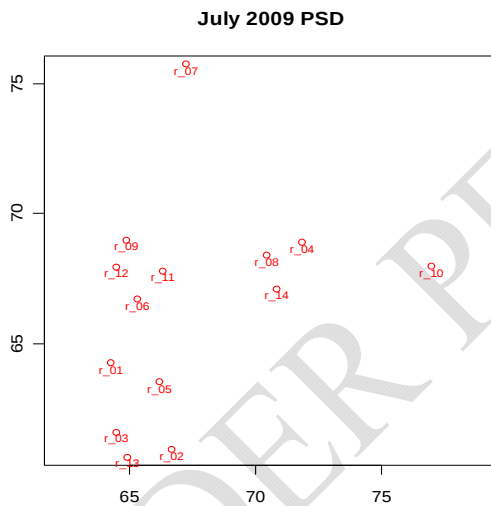


Figure 4a: Metric MDS for July PSD

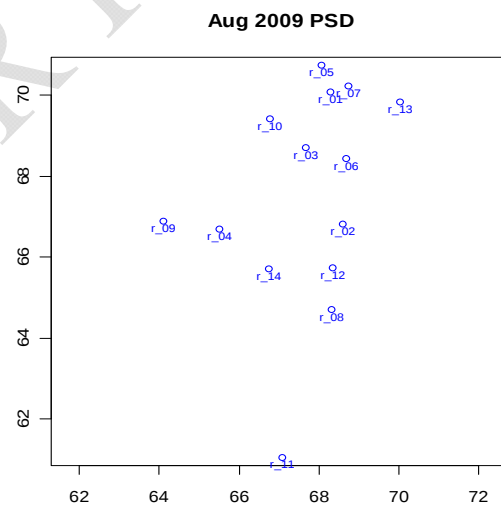


Figure 4b: Metric MDS for August PSD

A Procrustes analysis is a procedure for comparing two or more MDS configurations. Cox & Cox (2001) enumerates the use of the Procrustes analysis in the rotation, reflection, dilation, and translation of one configuration to fit a target configuration optimally. The mathematical background for programming the Procrustes analysis using the R software is obtainable in Cox & Cox (2001).

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 $d_{rs} = f(\delta_{rs})$.

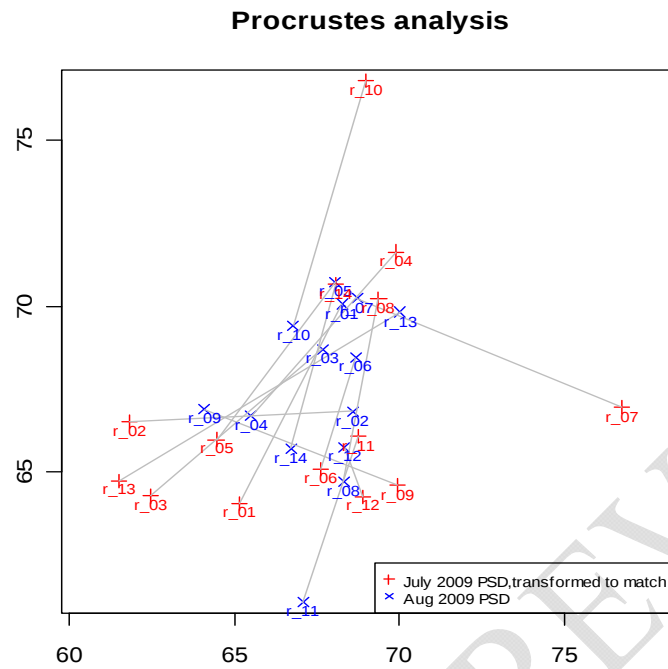


Figure 5a: *Procrustes analysis for PSD (July matched to August)*

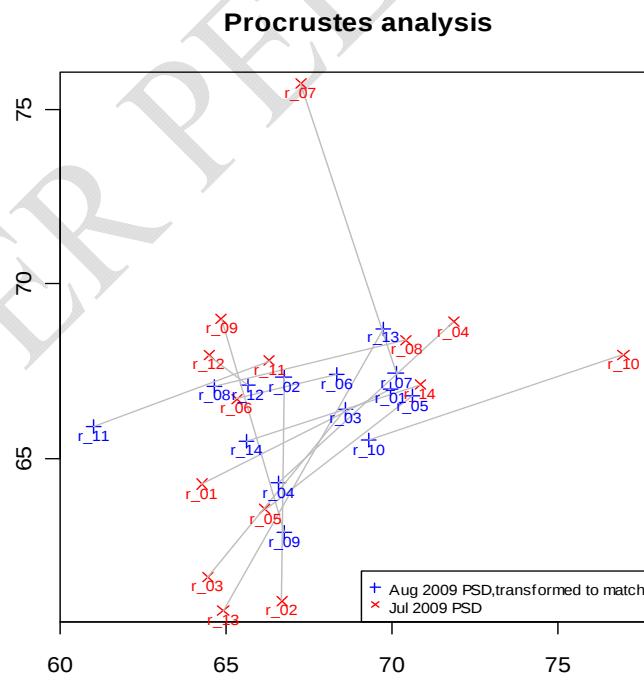


Figure 5b: *Procrustes Analysis for PSD (August matched to July)*

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. See mathematical details in Cox and Cox (2001).

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 Figures 4a and 4b for the July and August PSD data sets respectively.

Salient observations from Figure 4a is 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 Figure 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 Figures 4a and 4b.

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

In Figure 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 Figure 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 Figures 2 and 3), similar investigation needs to be launched for its operation run 11 for further improvement.

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 as discussed by Greenacre (2007). 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 Greenacre (2007).

Exploiting the exploratory capabilities of this approach, Figures 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 Figures 6 and 7 implies that the map displays both the rows and column points in principal coordinates.

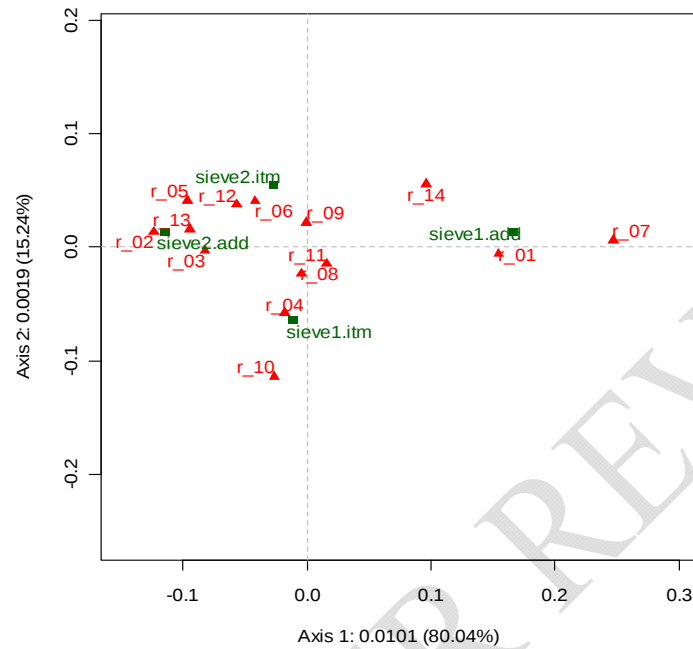


Figure 6: Correspondence analysis plot for July PSD data set (*symmetric map*)

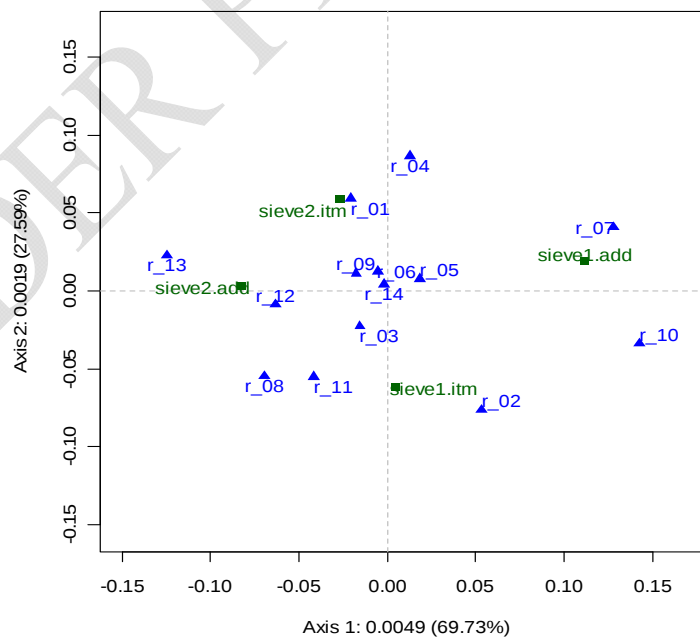


Figure 7: Correspondence analysis plot August PSD data set (*symmetric map*)

Firstly, a total Inertia of 95.28%, which is quite close to unity strongly suggests that the display in Figure 6 nicely fits within the two dimensional configuration. The distinctness of operations 7 and 10 in Figure 6 could easily be depicted and this is comparable to the observations of the Metric MDS of Figure 4a and the Procrustes map of Figure 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.

Figure 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 Figure 6.

Furthermore, notice the random distribution of the runs within the configuration. Tracing the earlier identified operation 11 according to the MDS of Figure 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. Theories on biplots are obtained in La Grange et al. (2009), Gower et al. (2011), Greenacre (2010), Gower and Hand (1996) 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 Gower et al. (2011) 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.

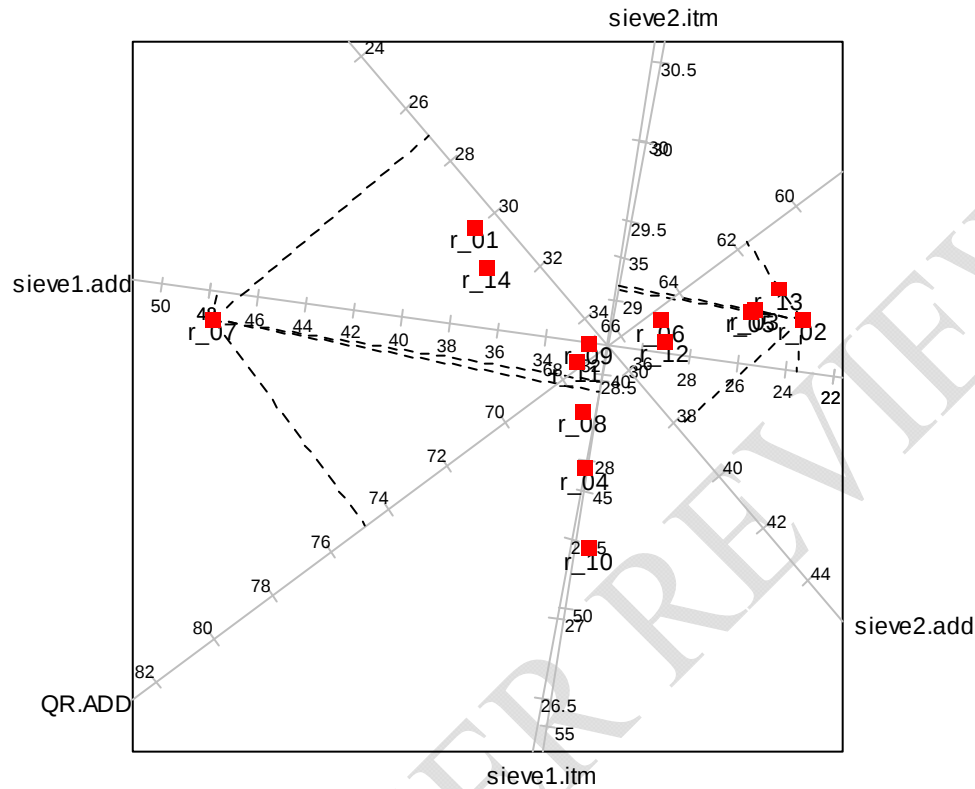


Figure 8: PCA biplot for July PSD data set (Quality of Display = 88.65%)

Figures 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 Figures 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.

Table 1: July and August EX.ADD Predicted Values using Figures 8 and 9.

	July EX.ADD		August EX.ADD	
	Lowest	Highest	Lowest	Highest
	r_2	r_7	r_05	r_11
sieve1.itm	36.369	40.2958	37.9384	41.965
sieve2.itm	29.104	28.4277	29.8087	25.3358
sieve1.add	23.553	47.7599	33.8037	26.8321
sieve2.add	38.134	27.0336	35.7132	35.6822
QR.ADD	61.686	74.7935	69.5169	62.5143

References

- ALDRICH, C., GARDNER, S. & LE ROUX, N.J. (2004), Monitoring of Metallurgical Process Plants by using Biplots, *AIChE Journal*, 50.
- COX, T.F. & COX, M.A.A. (2001), *Multidimensional Scaling*, 2nd edition, New York: Chapman & Hall/CRC.
- GOWER, J.C., LUBBE, G.S. & LE ROUX, N.J. (2011), *Understanding Biplots*, Chichester: Wiley.
- JOHN, C. (2011), *Industrial Statistical Process Controls A Case Study: Tobacco Moisture Variability Reduction*, Unpublished MComm/MSc thesis, University of Stellenbosch.
- GREENACRE, M. (2007), *Correspondence Analysis in Practice*, 2nd edition, New York: Chapman & Hall/CRC.
- GREENACRE, M. (2010), *Biplots in Practice*, 2nd edition, Barcelona: Fundacion BBVA.
- GOWER, J.C. & HAND D.J. (1996), *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*, 30(12).
- JOHNSON, R.A. & WICHERN D.W. (2007), *Applied Multivariate Statistical Analysis*, 6th edition, Upper Saddle River, NJ: Pearson International Edition.
- MASON, R.L. & YOUNG, J.C. (2002), *Multivariate Statistical Process Control with Industrial Applications*, Alexandria Virginia: ASA-SIAM.
- LA GRANGE, A., LE ROUX, N.J. & GARDNER-LUBBE, S. (2009), Biplot GUI: Interactive Biplots in R, *Journal of Statistical Software*, 30(12).
- R DEVELOPMENT CORE TEAM (2009). *R: Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3 900051-07-0, URL <http://www.R-project.org/>.
- WANG, X., KRUGER, U., & IRWIN, G. (2005), Process Monitoring Approach Using Fast Moving Window PCA, *Industrial & Engineering Chemistry Research*, 44 (14), pp. 5691-5702.
- YUNUS, M.Y.M. & ZHANG, J. (2010), A Multidimensional Scaling based Process Monitoring, 20th European Symposium of Computer Aided [Process Engineering](#), Elsevier B.V, 28, pp. 1333-1342.

Appendix A - Used Data Sets

AUG 2009 - EX-ITM PSD Data

OPN	RN	sieve1	sieve2	sieve3	sieve4	fines	QR
9605	r_01	36.51	31.77	25.93	2.47	3.33	68.28
9608	r_02	43.47	25.12	23.66	2.26	5.49	68.6
9611	r_03	40.15	27.53	24.92	0.8	6.6	67.68
9613	r_04	33.92	31.59	25.1	4.05	5.34	65.51
9616	r_05	39.24	28.83	24.67	2.44	4.82	68.08
9620	r_06	38.97	29.71	25.08	3.51	2.72	68.68
9629	r_07	38.6	30.13	25.64	3.11	2.52	68.73
9631	r_08	41.3	27.01	23.62	2.17	5.9	68.31
9667	r_09	36.59	27.5	27.15	1.24	7.52	64.09
9668	r_10	41.29	25.49	25.73	0.95	6.54	66.78
9670	r_11	40.51	26.58	25.06	1.04	6.8	67.1
9639	r_12	39.27	29.07	24.99	1.35	5.32	68.34
9685	r_13	38.46	31.56	24	1.18	4.81	70.01
9705	r_14	38.22	28.51	24.21	3.26	5.79	66.73

AUG 2009 - EX-ADD PSD Data

OPN	RN	sieve1	sieve2	sieve3	sieve4	fines	QR
9605	r_01	32.2	37.87	21.45	3.52	4.96	70.1
9608	r_02	33.35	33.47	25.53	2.38	5.27	66.8
9611	r_03	31.2	37.5	25.41	0.59	5.3	68.7
9613	r_04	32.45	34.24	25.63	2.47	5.21	66.7
9616	r_05	33.95	36.79	23.87	1.85	3.55	70.7
9620	r_06	31.91	36.53	26.7	2.58	2.28	68.4
9629	r_07	39.42	30.81	23.97	4.83	0.97	70.2
9631	r_08	26.83	37.87	23.82	3.15	8.32	64.7
9667	r_09	30.45	36.44	27.75	1.21	4.15	66.9
9668	r_10	38.95	30.47	24.53	0.53	5.51	69.4
9670	r_11	26.66	34.4	30.62	2.2	6.12	61.1
9639	r_12	27.89	37.84	23.57	2.39	8.31	65.7
9685	r_13	26.77	43.06	24.28	0.13	5.77	69.8
9705	r_14	30.79	34.92	26.86	2.72	4.71	65.7

JUL 2009 - EX-ITM PSD Data

OPN	RN	sieve1	sieve2	sieve3	sieve4	fines	QR
9485	r_01	37.29	26.97	26.32	2.2	7.21	64.27
9487	r_02	37.14	29.54	25.68	4.39	3.26	66.67
9491	r_03	36.96	27.49	26.36	3.8	5.39	64.45
9494	r_04	44.06	27.79	22.43	2.03	3.7	71.85
9496	r_05	35.88	30.29	24.28	4.27	5.28	66.17
9503	r_06	35.82	29.5	27.23	1.38	6.08	65.32
9504	r_07	39.35	27.9	25.17	1.69	5.89	67.25
9510	r_08	41.53	28.9	23.34	2.39	3.83	70.44
9512	r_09	36.75	28.1	26.39	3.74	5.01	64.86
9519	r_10	49.4	27.58	16.25	2.83	3.94	76.98
9575	r_11	39.08	27.23	24.99	0.92	7.77	66.32
9580	r_12	35.7	28.79	27.97	1.63	5.91	64.49
9585	r_13	36.19	28.72	27.04	2.44	5.61	64.91
9599	r_14	37.57	33.28	24.28	3.55	1.31	70.85

JUL 2009 - EX-ADD PSD Data

OPN	RN	sieve1	sieve2	sieve3	sieve4	fines	QR
9485	r_01	37.4	26.85	26.06	3.41	6.28	64.3
9487	r_02	23.52	37.38	29.57	5.33	4.2	60.9
9491	r_03	25.52	36.06	28.15	3.1	7.18	61.6
9494	r_04	31.64	37.26	24.96	1.64	4.5	68.9
9496	r_05	25.79	37.74	28.96	4.15	3.36	63.5
9503	r_06	29.49	37.2	26.45	2.28	4.58	66.7
9504	r_07	47.8	27.96	18.96	1.4	3.89	75.8
9510	r_08	32.1	36.29	25.51	3.05	3.05	68.4
9512	r_09	32.28	36.71	25.12	2.31	3.59	69
9519	r_10	30.98	36.98	24.46	1.71	5.88	68
9575	r_11	32.59	35.21	25.23	1.1	5.87	67.8
9580	r_12	29.22	38.73	26.92	1.33	3.8	68
9585	r_13	24.7	35.92	30.57	2.3	6.51	60.6
9599	r_14	36.8	30.3	24.55	5.56	2.79	67.1