

PRINCIPAL COMPONENT CHART FOR MULTIVARIATE STATISTICAL PROCESS CONTROL

Gafar Matanmi OYHEYEMI

Department of Statistics, University of Ilorin, Ilorin, Nigeria

email: gmoyeyemi@gmail.com

Abstract: Multivariate statistical process control technique (Hotelling T^2 chart) was used to monitor four correlated quality characteristics (active detergent, moisture content, bulk density and ph level) of detergent produced by a company which indicated out-of-control signal. Principal Component Chart is used as a follow-up to out-of-control signal of the Multivariate Control Chart, to identify the quality characteristic(s) that contributed to the signal. The component scores obtained from the principal component analysis of the four quality characteristics measured were used to identify the quality characteristic(s) that contributed to the out-of-control signaled by the Hotelling T^2 chart. The chart of the first component which accounted for 96.7% of the total variability and has moisture content highly loaded in it is out-of-control, which implied that moisture content of the detergent produced by the company is out-of-control.

Keywords: principal components, out-of-control, quality characteristics, control limits, eigen-values

Introduction

Statistical process control is based on a number of basic principles which apply to all processes, including batch and continuous processes of the type commonly found in the manufacture of bulk chemicals, pharmaceutical products, specialist chemicals, processed foods and metals. The principles apply also to all processes in service and public sectors and commercial activities, including forecasting, claim processing and many financial transactions. One of these principles is that within any process variability is inevitable (Chanda, 2001).

Generally there are two groups of statistical process control (SPC), i.e. univariate statistical process control (USPC) and multivariate statistical process control (MSPC), which are used for different scenarios. The process of monitoring and control primarily apply to the systems or processes from the univariate perspective, which has only one process output variable or quality characteristic measured and tested. If a process is to meet or exceed customer expectations, generally it should be produced by a process that is stable or repeatable. More precisely, the process must be capable of operating with little variability around the target or nominal dimensions of the producer's quality characteristics.

Typically process monitoring applies to systems or processes in which only one variable is measured and tested. There are many processes in which the simultaneous monitoring or control of two or more quality characteristics is necessary. Process monitoring problems in which several variables are of interest are called Multivariate Statistical Process Control (MSPC). One of the disadvantages of a univariate monitoring scheme is that for a single process, many variables may be monitored and even controlled. MSPC methods overcome this disadvantage by monitoring several variables simultaneously. Using multivariate statistical process control methods, engineers and manufacturers who monitor complex processes may monitor the stability of their process.

The first original study in multivariate quality control was introduced by Hotelling (1947). Three of the most popular multivariate control statistics are Hotelling T^2 , Multivariate Exponentially-Weighted Moving Average (MEWMA) and the Multivariate Cumulative Sum (MCUSUM). The multivariate charts mentioned above take the correlations among the variables into account in monitoring the mean vector or variance-covariance matrix (Runger and

Montgomery, 1997). Multivariate charts are less popular than univariate charts because of the following reasons;

- The difficulty involved in their computation.
- Unlike univariate case, the scale of the values displayed on the multivariate chart is not related to the scales of any of the monitored variables.
- Once an out-of-control signal is given by the multivariate chart, it may be difficult to identify which of the variables caused the out-of-control signal. More complicated operations are required to determine the cause of the signals.

Multivariate control charts, such as Hotelling's-T² Control chart, Multivariate Exponential Weighted Moving Average (MEWMA) chart, Multivariate Cumulative Sum (MCuSum) chart, are used for monitoring several quality characteristics measured simultaneously on a product or process. Multivariate charts are also useful for monitoring quality profiles as discussed by Woodall et al. (2004). The objective of multivariate control charts is in two phases;

To identify shifts in the mean vector that might distort the estimation of the in-control mean vector and variance covariance matrix, and

- To identify and eliminate multivariate outliers. (Williams et al. 2006)

Alt (1995) defined two phases in constructing multivariate control charts, with Phase I divided into two Stages. In the retrospective Stage 1 of Phase I, historical data (observations) are studied for determining whether the process was in control and to estimate the in-control parameters of the process. The Hotelling's-T² Control chart is utilized in this stage (Alt and Smith, 1998, Tracy et al. 1992, and Wieda, 1994). In phase II, control charts are used with future observations for detecting possible departures from the process parameters estimated in Phase I. In Phase II, one uses charts for detecting any departure from the parameter estimates, which are considered in the in-control process parameters (Vargas, 2003).

An important aspect of the Hotelling's-T² Control chart is how to determine the sample variance-covariance matrix used in the calculation of the chart statistics (UCL and LCL). When rational subgroups are taken, the implication is that the appearance of a special cause of variation within a subgroup is unlikely, so that all observations within a subgroup share a common distribution. Thus, the regular sample variance-covariance matrix is useful and taking the average over all the subgroups is the common procedure, unless there are special causes that alter the variance-covariance matrix. If subgroups are taken and the population parameters are known then the Hotelling's T² statistic, T_i^2 , is $\chi_{\alpha,p}^2$ distributed, where p is the number of variables and α is the probability of false alarm. In the event that the population parameters are unknown (that is, the mean vector and the variance-covariance matrices are unknown), the estimates are obtained from the sample and the Hotelling's T² statistic, T_i^2 , has an F or Beta distribution (Kolarik, 1999).

Construction of Hotelling's T² Control Chart

A set of n sub-samples of m observations collected are used in computing the control limits of the Hotelling's T² Control chart. For the observation x_{ij} , $i = 1, 2, 3, \dots, m$ and $j = 1, 2, 3, \dots, n$, the Hotelling's T² statistic is obtained as follows; If the true parameters of a probability distribution are known, the χ^2 distribution is appropriate. Suppose that $x_1, x_2, x_3, \dots, x_p$, are the variables from a normal distribution process and $\mu_1, \mu_2, \mu_3, \dots, \mu_p$ are the population means of the variables. $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_p$ are the sample means, under assumption of knowing the variance-covariance matrix Σ , where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & - & - & - & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & - & - & - & \sigma_{2p} \\ - & - & - & - & - & - & - \\ - & - & - & - & - & - & - \\ - & - & - & - & - & - & - \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & - & - & - & \sigma_{pp} \end{bmatrix} = \sigma_{ij}$$

The statistic $T^2 = n(\bar{x} - \mu)' \Sigma^{-1}(\bar{x} - \mu)$ follows a chi-square (χ_p^2) with p degrees of freedom. Monitoring and detecting the out-of-control points depends on constructing correct control limits. The upper control limit (UCL) for the chart is given;

$$UCL = \chi_{\alpha,p}^2$$

When the true population values are not known, the statistics are computed using the estimates of the population parameters. The variance-covariance matrix Σ is estimated by the simple average of the m samples variance-covariance matrices;

$$S = \frac{S_1 + S_2 + \dots + S_m}{m}$$

$$S_i = \begin{bmatrix} s_{11} & s_{12} & s_{13} & - & - & - & s_{1p} \\ s_{21} & s_{22} & s_{23} & - & - & - & s_{2p} \\ - & - & - & - & - & - & - \\ - & - & - & - & - & - & - \\ - & - & - & - & - & - & - \\ s_{p1} & s_{p2} & s_{p3} & - & - & - & s_{pp} \end{bmatrix} = s_{ij}$$

$$T_i^2 = n(\bar{x}_i - \bar{x})' S^{-1}(\bar{x}_i - \bar{x}) \text{ for } i = 1, 2, 3, \dots, m$$

The upper control limit (UCL) of the Hotelling T^2 chart is given by;

$$UCL = \frac{(m-1)^2}{m} B_{\left(\alpha, \frac{p}{2}, \frac{m-p-1}{2}\right)}, \text{ where } \alpha \text{ is the probability of false alarm for each}$$

point plotted on the control chart, and $B_{\left(\alpha, \frac{p}{2}, \frac{m-p-1}{2}\right)}$ is the (1- α) percentile of the beta distribution

with parameters u_1 and u_2 (Tracy et al. 1992, Wierda, 1994). The Hotelling T or F-distribution table may be use to obtain the Upper Control Limit (UCL). The Lower Control Limit (LCL) is always set to zero. The values of T_i^2 are plotted on the Hotelling T^2 chart, and if one of or more of the m points are out-of-control, special causes of variation are sought.

Although the knowledge of the statistical distribution of the control chart statistic is needed to calculate the upper and the lower control limits of the chart and estimate the control chart performance which are unknown in most cases. If the exact distribution is unknown or intractable, most especially when there are no subgroups, the upper control limit, UCL can be calculated from either an approximate distribution or from a Monte Carlo simulation (Williams et al. 2006).

Although the T^2 chart is the most popular, easiest to use and interpret method for handling multivariate process data, and is beginning to be widely accepted by quality engineers and operators, it is not a panacea. First, unlike the univariate case, the scale of the values displayed on the chart is not related to the scales of any of the monitored variables. Secondly, when the T^2 statistic exceeds the upper control limit (UCL), the user does not know which particular variable(s) caused the out-of-control signal.

With respect to scaling, we strongly advise to run individual univariate charts in tandem with the multivariate chart. This will also help in honing in on the culprit(s) that might have caused the signal. However, individual univariate charts cannot explain situations that are a result of some problems in the covariance or correlation between the variables. This is why a dispersion chart must also be used.

Another way to analyze the data is to use *principal components*. For each multivariate measurement (or observation), the principal components are linear combinations of the standardized p variables (to standardize subtract their respective targets and divide by their standard deviations). The principal components have two important advantages:

The new variables are uncorrelated (or almost)

Very often, a few (sometimes 1 or 2) principal components may capture most of the variability in the data so that we do not have to use all of the p principal components for control.

Unfortunately, there is one big disadvantage: The identity of the original variables is lost! However, in some cases the specific linear combinations corresponding to the principal components with the largest *eigenvalues* may yield meaningful measurement units. What is being used in control charts are the principal factors. A principal factor is the principal component divided by the square root of its eigenvalue.

Principal Component Chart

One of the problems of multivariate control chart (Hotelling T^2 chart) is the problem of identifying the variable(s) that cause out-of-control signal in the chart. Because of its complexity in nature it is difficult to identify the variable(s) caused out-of-control signal, except constructing univariate control chart for each of the variables which poses another problem in the sense that the studied quality characteristics are believed to be highly correlated. Principal components can be used to investigate which of the p variables in the multivariate control chart are responsible for out-of-control signal. The most common practice is to use the first k most significant components, if Hotelling T^2 control chart gave an out-of-control signal, for further investigation.

The basic idea is that the first k principal components can be physically interpreted, and named. Consequently, if the Hotelling T^2 chart gives an out-of-control signal and, for instance, the second principal component chart also gives an out-of-control signal, then from the interpretation of this component, a direction to the variables which are suspect to be out-of-control can be deduced (Jackson, 1991). The discovery of the assignable cause of the problem, with this method, demands a further knowledge of the process itself, from the practitioner. The basic problem is that the principal components do not always have a physical interpretation.

The principal components are those uncorrelated linear combinations Y_1, Y_2, \dots, Y_p , whose variances are as large as possible. Suppose the original dataset X is a p dimensional normal vector with mean and variance-covariance matrix given by μ and Σ respectively. The density of X is constant on the μ centered ellipsoids

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = c^2$$

Which have axes $\pm c\sqrt{\lambda_i}e_i$, $i = 1, 2, \dots, p$, where the pair (λ_i, e_i) are the eigenvalue-eigenvector pairs of Σ . The c statistic is well-known in the literature as the Hotelling T^2 statistic. According to Johnson & Wichern (2002)

$$c^2 = x^T \Sigma^{-1} x = \frac{1}{\lambda_1} (e_1^T x)^2 + \frac{1}{\lambda_2} (e_2^T x)^2 + \dots + \frac{1}{\lambda_p} (e_p^T x)^2$$

Where $e_1^T x, e_2^T x, \dots, e_p^T x$ are the principal components of X . Setting $y_1 = e_1^T x, y_2 = e_2^T x, \dots, y_p = e_p^T x$, we have

$$c^2 = \frac{1}{\lambda_1} (y_1)^2 + \frac{1}{\lambda_2} (y_2)^2 + \dots + \frac{1}{\lambda_p} (y_p)^2$$

In this way, the Hotelling T^2 statistic can be expressed as a function of the X values or as a function of the Y values (The components). Recommendations for selecting an appropriate number of principal component variables for multivariate statistical process control are typically the same as those proposed for traditional Principal Component Analysis in which the objective is to summarize a complex dataset. Following some guidelines suggested by Runger & Alt (1996):

- Choose k such that $\sum_{i=1}^k \lambda_i \geq 0.9 \sum_{j=1}^p \lambda_j$
- Increment k such that $\lambda_i \geq \lambda_m$, where $\lambda_m = \frac{\sum_{j=1}^p \lambda_j}{p}$ and $i = 1, 2, \dots, k$
- Plot λ_i against i and select k at the “knee” in the curve.

Although these are useful guidelines in general, process control has a different objective than a summary of variation in a random sample of in-control data. Because the goal of statistical process control is to detect assignable cause in a stream of data collected over time, an approach to principal component analysis is to investigate the performance of a control chart as a function of k . Assuming we want 99.7% confidence interval, the Upper Control Limit (UCL), Center Line (CL) and the Lower Control Limit (LCL) are given in equation (1).

$$\begin{aligned}
 UCL &= +3\sqrt{\lambda_k} \\
 CL &= 0 \\
 LCL &= -3\sqrt{\lambda_k}
 \end{aligned}
 \tag{1}$$

Multivariate Control Chart

The used is a secondary data and it consisted thirty-five single samples of detergent produced by a detergent company were randomly taken at regular interval. Four quality characteristics of the detergent were measure and they are x_1 (active detergent), or x_2 (moisture content) or x_3 (bulk density) or x_4 (ph level). Since single sample was taken at a time, that is, the sub-sample size $n = 1$; $m=35$. Hotelling T^2 control chart for individual observation is used to monitor the four quality characteristics of detergent produced by the company. The Mahalanobis

distances are obtained using equation (2), the distances are then plotted to obtain the control chart for the detergent produced.

$$d_{1,i}^2 = (x_{1i} - \bar{x}_1)^t S_1^{-1} (x_{1i} - \bar{x}_1) \quad i = 1, 2, 3, \dots, 48 \quad \text{----- (2)}$$

The summary statistics are given as follows;

$$\bar{x} = \begin{bmatrix} 23.3377 \\ 3.4197 \\ 309.4537 \\ 10.4520 \end{bmatrix} \quad S = \begin{bmatrix} 1.3021 & -0.1217 & 1.5289 & -0.0222 \\ -0.1217 & 0.1066 & -1.5065 & -0.0418 \\ 1.5289 & -1.5065 & 147.4957 & 0.0783 \\ -0.0222 & -0.0418 & 0.0783 & 0.0865 \end{bmatrix}$$

$$R = \begin{bmatrix} 1.0000 & -0.1217^* & 0.1103^* & -0.0663 \\ & 1.0000 & -0.3798^{**} & -0.4349^{**} \\ & & 1.0000 & 0.0219 \\ & & & 1.0000 \end{bmatrix}$$

* Significant at 0.1 ** significant at 0.05

The correlation matrix shows that there exists inter-correlation among the four quality characteristics hence the reason for using multivariate control chart. The data as well as the mahalanobis distances are shown in appendix I. The Upper Control Limit (UCL) for the chart is given as follows;

$$UCL = \frac{(m-1)^2}{m} B_{\left(\alpha, \frac{p}{2}, \frac{m-p-1}{2}\right)} = 8.7181; \text{ where } m = 35 \text{ and } p = 4.$$

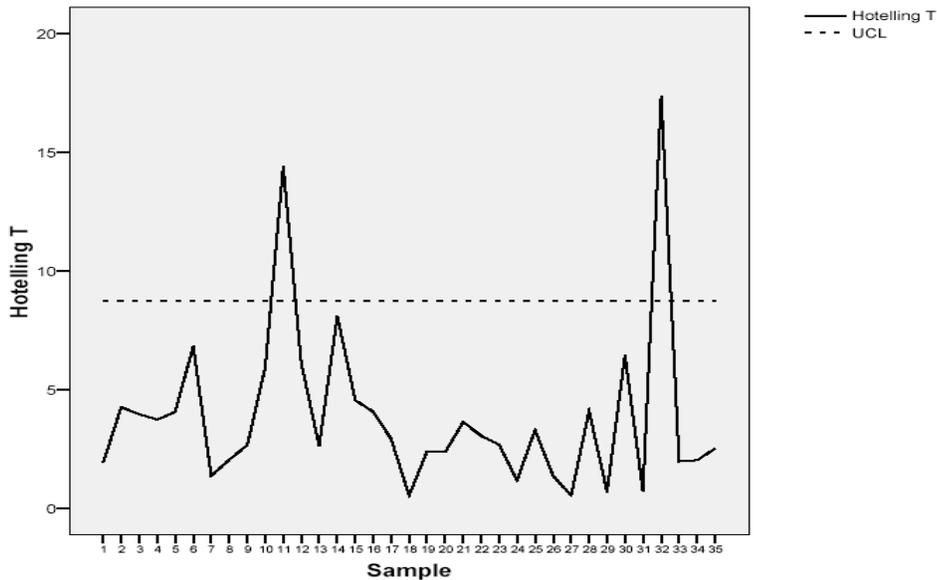


Figure 1. Hotelling T² chart for detergent produced by the company

Principal Component Charts

According to the multivariate control chart (Hotelling T^2 chart), the process producing the detergent was found to be out-of-control at 0.05 level of significance as shown in figure 1. One of the drawbacks of multivariate control charts is the identification of variable(s) that contributed to the out-of-control signal. From the chart, it is not possible to categorically say that either x_1 (active detergent), or x_2 (moisture content) or x_3 (bulk density) or x_4 (ph level) or any combinations of the four quality characteristics contributed to the out-of-control signal.

To identify which of the quality characteristics responsible for the out-of-control situation in the Hotelling T^2 chart, principal component analysis was use to obtain new components (PC1, PC2, PC3 and PC4) for the dataset (equation 3). These four components, PC1, PC2, PC3 and PC4 are linear combinations of the original variables (quality characteristics) and they are uncorrelated with one another. The component scores were then obtained from the linear combinations (components).

$$\begin{aligned}
 PC1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \\
 PC2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 \quad \dots\dots\dots(3) \\
 PC3 &= a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 \\
 PC4 &= a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4
 \end{aligned}$$

The component scores are treated as expected observations for each of the datasets, which are used to obtain a control chart. The quality characteristics will have different weight (eigen-vector) in each of the four components and if such component is out-of-control when plotted, then the variable that has highest weight will be concluded to have contributed to the out-of-control signal experienced in the Hotelling T^2 chart. The control limits of the principal component charts are given in equation (1), where λ_k is the eigenvalue of each of the components.

The Component Charts

Because of the different units in measuring the four quality characteristics, the characteristics were standardized by making use of their correlation matrix. The eigen-values and components matrix of the principal components analysis for the four quality characteristics given below;

$$\begin{aligned}
 \lambda &= [3.8695, 1.0906, 0.8739, 0.3511] \\
 &\begin{bmatrix} PC1 & PC2 & PC3 & PC4 \\ 0.3640 & -0.6209 & 0.5983 & 0.8520 \\ -0.6941 & -0.0710 & -0.0708 & 0.7128 \\ 0.4572 & -0.2825 & -0.7716 & 0.3404 \\ 0.4203 & 0.7277 & 0.2042 & 0.5021 \end{bmatrix}
 \end{aligned}$$

It can be seen that while the second quality characteristic, x_2 (moisture content), is highly loaded in the first component (PC1), it is x_4 (ph level) for the second component (PC2), it is third quality characteristic, x_3 (bulk density) in the third component (PC3) and x_1 (active detergent) in the fourth component (PC4). Since the first component (PC1) accounted for about 96.7% total variability, it is therefore sufficient to make use of it in principal component chart. The component scores are obtained using the component matrix and the 99.7% control confidence limits (control limits) are obtained as follows;

$$UCL = +3\sqrt{3.8695} = 3.8939$$

$$CL = 0$$

$$LCL = -3\sqrt{3.8695} = -3.8939$$

The principal component chart is shown in figure 2



Figure 2. Principal component chart for the first component (PC1)

Discussion and Conclusion

The correlation analysis shows that there exist inter-correlation among the four quality characteristics monitored, hence the need for multivariate statistical process control technique. Hotelling T^2 control chart signaled an out-of-control for the product (detergent) based on the four quality characteristics by having samples 11 and 32 above the upper control limit. The principal component chart for the first component (PC1) also signaled an out-of-control with sample 32 falling below the lower control limit. The only characteristic that is highly loaded though negatively in the first component is x_2 (moisture content) can therefore be concluded that contributed to the out-of-control signal by the multivariate control chart.

The use of principal component chart has assisted in identifying variable(s) that contributed to the out-of-control signal given by the Hotelling T^2 chart. An individual Shewart control chart for each of the four quality characteristics would have been needed to monitor the four quality characteristics. And since the four quality characteristics are significantly related, the individual Shewart control chart will definitely not give convincing results. It is shown in this work that when a multivariate control chart signals an out-of-control, principal component chart can be used to identify which of the monitored quality characteristic(s) contribute to the out-of-control signal.

References

Alt, F. B. (1995). Multivariate Quality Control in Encyclopedia of Statistical Sciences 6 New York, John Wiley & Sons.

Alt, F. B. & Smith, N. D. (1998). Multivariate Process Control. Handbook of Statistics, P. R. Krishnaiah and C. R. Rao (eds). North-Holland Elsevier Science Publishers B. V., 7, 333-351

Chanda, M. J. (2001). Statistical Quality Control. CRC Press, LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431

Hotelling, H. (1947). Multivariate Quality Control. Techniques of Statistical Analysis. McGraw-Hill

Jackson, J. E. (1991). A User Guide to Principal Components. John Wiley & Sons, N.Y.

Johnson, R. A. & Wichern, D. W. (2002). Applied Multivariate Statistical Analysis. Prentice Hall.

Kolarik, W. J. (1999). Creating Quality: Process Design for Results. McGraw-Hill International Edition. Singapore, McGraw-Hill Book Company.

Runger, G. C. & Alt, F. B. (1996): Choosing principal components for multivariate statistical process control. Communications in Statistics: Theory and Methods, 25, 909-922.

Runger, G.C. & Montgomery, D.C. (1997) Multivariate and univariate process control: geometry and shift directions. Quality and Reliability Engineering International, 13, 153-158.

Tracy, N. D., Young, J. C. & Mason, R. L. (1992). Multivariate Control Charts for Individual Observations. Journal of Quality Technology, 24, 88-95.

Vargas, J. A. (2003). Robust Estimation in Multivariate Control Charts for Individual Observations. Journal of Quality Technology, 35(4), 367-376

Wieda, S. J. (1994). Multivariate Statistical Process Control- Recent Results and Directions for Future Research, Statistica Neerlandica, 48, 147-168.

Williams, J. D., Woodall, W. H., Birch, J. B. & Sullivan, J. E. (2006). Distribution of Hotelling's T^2 Statistic Based on the Successive Differences Estimator. Journal of Quality Technology, 38(3), 217-229.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C. & Gupta, S. (2004). Using Control Charts to Monitor Process and Product Quality Profiles. Journal of Quality Technology, 36, 309-320

Appendix

sample	detergent	moisture	density	ph level	distance
1	22.23	3.65	315.45	10.47	1.9110911
2	22.27	3.90	295.05	10.03	4.2494428
3	22.34	3.67	289.64	10.24	3.9452862
4	21.49	3.53	314.67	10.70	3.7194520
5	22.26	3.43	325.58	10.70	4.0488209
6	23.23	3.46	334.35	10.66	6.8160043
7	23.58	3.22	322.86	10.41	1.3518301
8	23.44	3.12	324.20	10.41	2.0275244
9	23.35	3.24	329.04	10.43	2.6469698
10	23.73	3.01	330.92	10.16	5.9341587
11	25.81	3.75	300.94	10.80	14.4046455
12	25.27	3.28	305.63	10.86	6.1193951
13	22.53	3.20	300.87	10.72	2.6055144
14	22.05	3.23	300.15	11.21	8.0851934
15	22.34	3.16	299.69	10.94	4.5429742
16	22.12	3.32	301.42	10.94	4.0531837
17	22.70	3.23	296.70	10.61	2.9029115
18	23.65	3.48	302.00	10.47	0.5197853
19	23.16	2.97	318.79	10.63	2.3841190
20	23.31	2.96	319.00	10.60	2.3574296
21	25.41	3.22	305.60	10.43	3.6242317
22	25.21	3.35	306.80	10.48	3.0371474
23	24.69	3.24	303.08	10.29	2.6625240
24	23.99	3.29	303.97	10.35	1.1557459
25	23.61	3.16	301.34	10.29	3.2802867
26	23.57	3.28	301.16	10.40	1.3501598
27	23.38	3.57	306.78	10.25	0.5282338
28	24.06	3.92	297.73	10.10	4.0879491
29	24.01	3.49	308.99	10.47	0.6733271
30	23.81	3.47	336.54	10.32	6.4455951
31	22.59	3.57	309.78	10.32	0.6917973
32	20.44	4.71	291.88	10.01	17.3669976
33	23.51	3.54	312.13	10.05	1.9585475
34	23.67	3.50	307.60	10.06	2.0075102
35	24.01	3.57	310.55	10.01	2.5042148