Journal of Industrial and Systems Engineering
Vol. 2, No. 4, pp 250-258
Winter 2009

JISE

A Comparison of the Mahalanobis-Taguchi System to A Standard Statistical Method for Defect Detection

Elizabeth A. Cudney[1], David Drain[2], Kioumars Paryani[3*], and Naresh Sharma[4]

[1,2,4] Missouri University of Science and Technology, Rolla, Missouri 65409 USA
[1] cudney@mst.edu, [2] draind@mst.edu, [4] nkscnc@mst.edu

[3] Lawrence Technological University, Southfield, Michigan, 48075 USA
kparyani@ltu.edu

ABSTRACT

The Mahalanobis-Taguchi System is a diagnosis and forecasting method for multivariate data. Mahalanobis distance is a measure based on correlations between the variables and different patterns that can be identified and analyzed with respect to a base or reference group. This paper presents a comparison of the Mahalanobis-Taguchi System and a standard statistical technique for defect detection by identifying abnormalities. The objective of this research is to provide a method for defect detection with acceptable alpha (probability of type I) and beta (probability of type II) errors.

Keywords: Mahalanobis distance, Mahalanobis-Taguchi System, Multivariate, Diagnosis, Alpha (Probability of Type I) Error, Beta (Probability of Type II) error, Forecasting.

1. INTRODUCTION

Mahalanobis-Taguchi System (MTS) is a pattern information technology, which has been used in different diagnostic applications to help in making quantitative decisions by constructing a multivariate measurement scale using data analytic methods. In the MTS approach, Mahalanobis distance (MD, a multivariate measure) is used to measure the degree of abnormality of patterns and principles of Taguchi methods are used to evaluate accuracy of predictions based on the scale constructed. The advantage of MD is that it considers correlations between the variables, which are essential in pattern analysis.

Well-known statistician Professor P.C. Mahalanobis introduced Mahalanobis distance (MD) in 1930 to distinguish patterns of a certain group from another group. Dr. Genichi Taguchi led the development of MTS by providing a means to define the reference group and measure the degree of abnormality of individual observations (Taguchi and Jugulum, 2000). MTS is a very economical approach for multidimensional pattern recognition systems. Pattern recognition is the study of how to observe and distinguish patterns of interest, and make reasonable decisions about the categories of pattern (Taguchi and Jugulum, 2002).

In multidimensional systems, one can simplify descriptions of the system by neglecting variables that have little or no effect on the measurement function. There are standard statistical approaches

[*] Corresponding Author

for doing this such as principal component analysis, linear discriminant analysis, and logistic regression decision trees (Jain et al., 2000). In this article we compare performance of a statistical method for pattern recognition and dimension recognition to the Mahalanobis-Taguchi System.

2. REVIEW OF RELEVANT LITERATURE

Considerable research is available utilizing Mahalanobis distance to determine similarities of values from known and unknown samples. Existing research also uses the Mahalanobis-Taguchi System for prediction and diagnosis which illustrates the methodology's accuracy and precision. However, little is available to compare the use of MTS to determine outliers with other methodologies such as standard statistical methods or neural networks.

Taguchi utilized the Mahalanobis-Taguchi System for diagnosis and pattern recognition. His research discussed a case study involving liver disease diagnosis in Tokyo, Japan using fifteen variables (Taguchi, 2000). In his research, he developed an eight-step procedure entitled "Mahalanobis Distance for Diagnosis and Pattern Recognition System Optimization Procedure" (Taguchi, 2000).

Lande conducted research using Mahalanobis distance to evaluate potential habitats for large carnivores in Scandinavia. The species involved included bears, wolves, lynx, and wolverines. The variables used included land cover, human density, infrastructure, and prey density (Lande, 2003). The results of the study were used to determine which areas were suitable for each species. This research considered a different field for application with respect to habitats and the environment.

Hayashi et al. used Mahalanobis distance to maximize productivity in a new manufacturing control system. The research applied Mahalanobis distance as a core to their manufacturing control system because of the methods ability to recognize patterns (Hayashi et al., 2001). The new system detected deviations from normal productivity much earlier and enabled root cause identification and prioritized resolution.

Asada used the Mahalanobis-Taguchi System to forecast the yield of wafers. Yield of wafers is determined by the variability of electrical characteristics and dust. The research focused on one wafer product that had a high yield. Mahalanobis distances were calculated on various wafers to compare the relationship between yield and distance (Asada, 2001). The signal-to-noise ratios were used to indicate the capability of forecasting and the effect of the parameters. This research showed the applicability of Mahalanobis distance to forecasting defective components.

Pattern recognition using Mahalanobis distance was demonstrated in the work of Wu. In this research, pattern recognition was used to diagnose human health. The results of tests from a regular physical check-up were used as the characteristics. The correlation between different tests was shown. Mahalanobis distance was used to summarize the multi-dimensional characteristics into one scale. In this research the base point was difficult to define because it was a healthy person. People who were judged to be healthy for the past two years were considered to be healthy (Wu, 1996). The research considered diagnosis of liver function with the objective to forecast serious disease until the next check-up. The approach provided a more efficient method that also avoided inhuman treatment that previously used double blind tests.

Statistical methods for treating these problems have a long history, beginning with R.A. Fisher's application of linear discriminant analysis to classify Iris plants into species through the measurement of four easily observable physical characteristics (Fisher, 1936). Dimension reduction

is achieved through the use of principal components analysis (PCA). For a complete description of the PCA techniques, including supporting theory, see the reference (Johnson and Wichern, 1992).

PCA begins by selecting the linear combination of the variables in the data accounting for the most observed variance; this linear combination is called the *first principal component*. The second principal component is that linear combination that is orthogonal to the first principal component and also accounts for the most variance in the raw data which is not accounted for by the first principal component. PCA can find as many principal components as variables in the data, but usually only the first few components are actually necessary to provide an adequate description of the data. Consider for example the scree plot from a PCA of a 17-variable data set as shown in Figure 1 (Minitab, 2005). The eigenvalue axis measures the proportion of variance accounted for by each principal component (on the horizontal axis). After six principal components are computed, there is relatively little variance left unexplained, so the remaining eleven principal components are unnecessary.

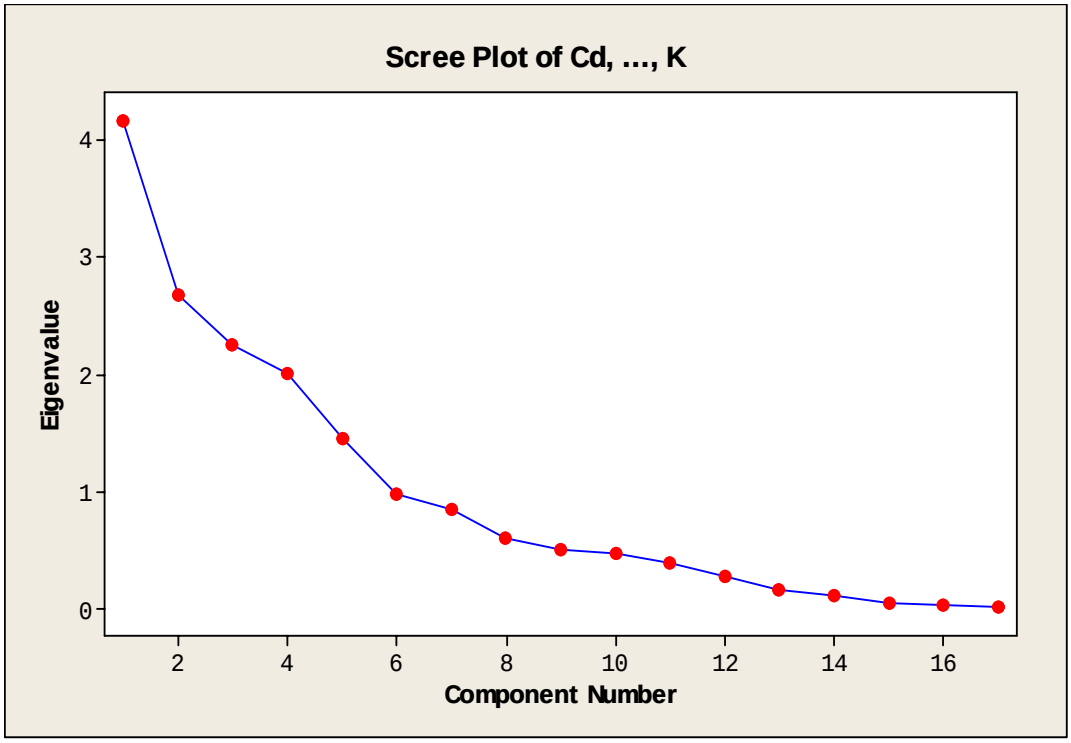

Figure 1 Scree Plot for PCA

3. MULTIDIMENSIONAL SYSTEMS

Multivariate analysis considers multiple related random variables simultaneously. Multivariate data consist of observations on several different variables for a number of objects or individuals. Initially each variable is considered equally important in the analysis (Manly, 1994). There are various types of multivariate analysis; however, the underlying theme for most of multivariate analysis is simplification. The objective is to summarize a large amount of data with relatively few parameters. Multivariate techniques are typically exploratory in that their purpose is to seek a general hypotheses rather than performing tests (Chatfield and Collins, 1980)

Multivariate analysis includes various techniques such as principal component analysis, factor analysis, discriminant analysis, cluster analysis, canonical correlation analysis, and multidimensional scaling. Any system, by its nature, is multidimensional. To improve or optimize any system's performance, it is necessary to understand the variables and noise conditions that affect the system's performance. All system's variables do not equally contribute to the system's performance; therefore, it is critical to identify those few important variables that have the most influence on the system's performance. This smaller set of variables can then be used for diagnosis and prediction purposes.

Generally speaking, a multidimensional system can be illustrated as in Figure 2. The input signal, *M*, is typically given by the consumer. Noise factors arise from the changes in consumer use of the product, wear and deterioration, and manufacturing variation. The control factors or system's variables are used to provide information in making a decision about the system. The output, *y*, should ideally closely match the input signal and represents a measurable characteristic.

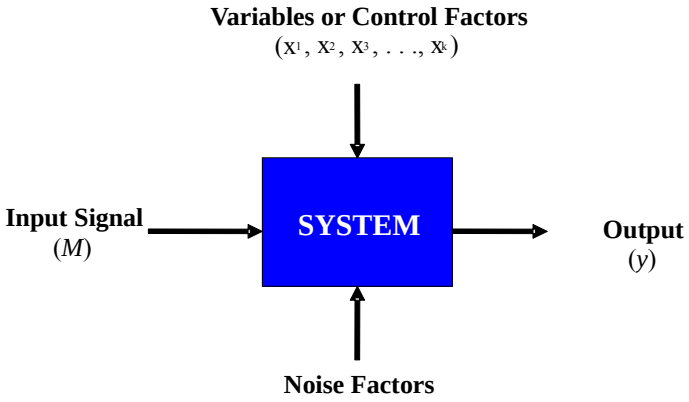

Figure 2 Multidimensional System

The Mahalanobis-Taguchi System can be used to minimize the number of variables (or control factors) required for diagnosis and to predict the performance of a system. The calculation of the Mahalanobis distance is outlined in section 5.

4. STATISTICAL APPROACH

The data utilized in the standard statistical approach consisted of a simulated data set of 500 observations of each of 17 variables. The data was simulated to represent an inspected stable process. The data were used to set the control limits. The approach used all 17 variables – one vector of measurements per group, and a known covariance matrix for the purpose of simulation only.

The 500 observations were simulated to assume production by a stable process, free of outliers. In actual practice, this means that the observations would be taken from a fairly short period of production during which all process indicators and maintenance logs would indicate the process would be operating as usual, and that an engineer would check each individual measurement to assure that there would be no outliers present. This simulated set of data serves as the basis to establish limits for a Hotelling's T^2 chart and accompanying generalized variance chart (Mason et al., 1992 and 1995, Montgomery, 1985). This procedure is a standard approach in multivariate

statistical process control; it accounts for correlations among predictor variables and allows for user-defined alpha risk choice. The limits determined from this stable process data were used when evaluating the performance of the charts on four simulated sets of data.

The 500 observations were simulated from a multivariate normal distribution with excursions for the same frequency but four different sizes: 0.5, 1.5, 2.5, and 3.5 standard deviations. Excursions could be of one of four types: vector outlier, individual variable outlier, vector shift or individual variable shift. A vector outlier is an outlier resulting from unusual values for some set of variables (more than one) all simultaneously varying from target at one time. A variable outlier is a one-time excursion in one variable alone. Shifts are persistent differences from target rather than one-time occurrences. Because this was a simulation practice, we were able to compare the decision made with the control chart to the decision that should have been made.

Alpha risk, also known as the probability of Type I error, is often referred to as a false positive. It is defined as the probability of rejecting a true null hypothesis. It is also known as the producer's risk because it represents the risk of rejecting a lot with acceptable product. In this case, it would mean to conclude that a part is defective when it is truly a good part.

Beta risk, also known as the probability of Type II error, is the probability of failing to reject the null hypothesis when it is false. It is also known as the consumer's risk because it represents the risk of accepting a lot with unacceptable quality. In this case, it would mean to conclude that a part is good when it is actually defective; therefore, it would be passed on to the consumer. The performance of the chart in terms of the resulting alpha and beta risks is shown in Table 1. As expected, larger shifts in the production data are easier to detect than smaller shifts (shifts are expressed as multiples of standard deviation of the production process distribution). Even the smaller shifts were detected about three fourths of the time, so this method seems sufficiently sensitive for industrial usage, especially when we consider that some of the simulated excursions here were rather subtle: a one-time jump of a single variable of the 17 from target, for example.

Table 1 Alpha and beta risk using a standard statistical approach

	Alpha	Beta
Production data	3.33%	N/A
0.5 sigma shift	2.05%	21.10%
1.5 sigma shift	2.03%	26.67%
2.5 sigma shift	2.29%	10.67%
3.5 sigma shift	1.89%	10.93%

5. MAHALANOBIS DISTANCE

The Mahalanobis-Taguchi System (MTS) is a pattern recognition technology that aids in quantitative decisions by constructing a multivariate measurement scale using a data analytic method. The main objective of MTS is to make accurate predictions in multidimensional systems by constructing a measurement scale (Taguchi and Jugulum, 2002). The patterns of observations in a multidimensional system depend strongly on the correlation structure of the variables in the system. One can make the wrong decision about the patterns if each variable is looked at separately without considering the correlation structure. To construct a multidimensional measurement scale, it is important to have a distance measure. The distance measure is based on the correlation

between the variable and the different patterns that could be identified and analyzed with respect to a base or reference point.

MD is very useful in determining the similarity of a known set of values to that of an unknown set of values. This method has successfully been applied to a broad array of cases mainly because it is very sensitive to inter-variable changes in the data. MTS is a technique for comparison that has the ability to handle a large number of variables with relative ease. MTS enables a reduction in dimensionality and the ability to develop a scale based on MD values (Cudney et al., 2006, 2007, 2007, and 2008).

The Mahalanobis distance methodology differs from other classical statistical approaches in several aspects. First, MD considers the variance and covariance of the measured variables rather than just the average value. It weights the differences by the variability range in the sample point direction. This accounts for natural variation within a data set. Second, it also accounts for the ranges of acceptability between the variables. It compensates for the interactions between variables. This is useful since most systems are not just comprised of independent variables rather dependent variables are also impacting the system's performance. Besides, it calculates distances in units of standard deviation from the group mean. By using standard deviation, the attribute under consideration is given in the original units, which relates the results to the original data. The calculated circumscribing ellipse around a cluster of data defines one standard deviation boundary of that group. Normally distributed variables can be converted into probabilities using the χ^2 (Chi square) probability density function. The probability of a group being normal or abnormal can be calculated for a confidence interval.

In the MTS, the Mahalanobis space (MS, the reference group) is obtained using the standardized variables of healthy or normal data. The Mahalanobis space can be used to discriminate between normal and abnormal objects. Once this MS is established, the number of attributes is reduced using orthogonal array (OA) and signal-to-noise ratio (SN) by evaluating the contribution of each attribute. Each row of the OA determines a subset of the original system by the including and excluding that attribute of system.

The same data used in the standard statistical approach was used to detect defects using the Mahalanobis distance (MD). Using exactly the same data consisting of the 500 observations with seventeen variables, Mahalanobis distance was calculated to determine which observations were considered defective. Based on the data, the MD was calculated. The initial production data consisting of 500 data sets was used to develop the threshold value of 1.6995. In MTS/MTGS, the threshold is essentially a safety factor. The threshold is typically used in MTS during the final stage to monitor conditions using the measurement scale, for more on MTS/MTGS and their applications in industrial problems see (Cudney et al., 2006, 2007, 2007, and 2008).

In this research, a breakthrough procedure for identifying outliers using MTS and a threshold value was developed. Observations (values for certain vehicles attribute) above the threshold value are classified as outliers. The threshold is used for the purpose of identifying outliers because it determines the general condition of the data. The threshold value of 1.6995 was used in the consecutive analysis of the data sets with 0.5, 1.5, 2.5, and 3.5 sigma shifts. Figure 3 illustrates the MD values for the 500 observations containing the 0.5 sigma shift data.

Based on a threshold value of a MD value of 1.6995, 15 observations were determined to be defective. The MD values for the 15 observations determined to be defective are given in Table 2.

The performance in terms of the resulting alpha and beta risk using the Mahalanobis distance is shown in Table 3.

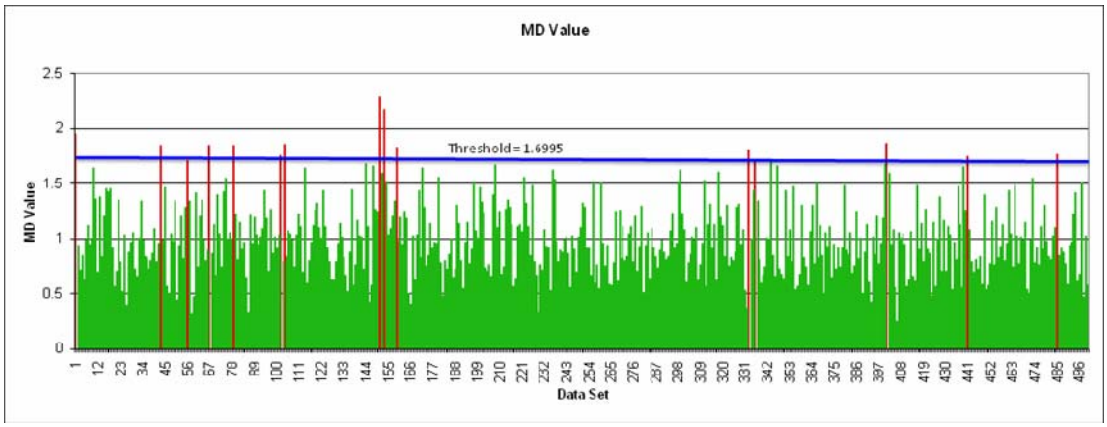

Figure 3 MD values for observations with 0.5 sigma shift

Table 2 Observations Determined Defective

Observation	MD Value
1	1.9511
43	1.8425
56	1.7007
67	1.8461
79	1.8420
102	1.7514
104	1.8576
151	2.2838
153	2.1694
159	1.8293
333	1.8014
336	1.6995
401	1.8642
441	1.7445
485	1.7559

Table 3 Alpha and beta risk using Mahalanobis Distance

	Alpha	Beta
Production data	3.35%	N/A
0.5 sigma shift	2.00%	21.20%
1.5 sigma shift	2.60%	0.00%
2.5 sigma shift	4.40%	0.20%
3.5 sigma shift	1.80%	0.60%

6. CONCLUSION

A standard statistical approach and Mahalanobis distance result in similar alpha risks for the production data and the four levels of sigma shift. However, the Mahalanobis distance results in a

steeper decrease in beta risks as the sigma shifts increase while showing an increase in alpha risks in some cases. The two approaches will generally show similar risks if they are "tuned" to produce the same alpha risk using a data set from an epoch of stable process operation. The advantage of the statistically based technique is that the "tuning" is automatic; a percentile from a distribution is selected. Choosing the threshold level from the MTS method is left to the expertise of the user.

PCA is similar to MTS in that it creates new axes to measure the distance of each point to its projections. PCA also uses a threshold value. The threshold value in PCA is used to determine which eigenvalues can be discarded as they contribute little to the data (Mardia et al., 1979). Principal components are uncorrelated linear combinations of the original variables which are similar to orthogonal Gram-Schmidt vectors (Cudney et al., 2006 and 2008). Therefore, Mahalanobis distances could be calculated using principal components. However, this is not recommended because the objectives of PCA and MTS are different and this requires additional calculations. PCA is useful for explaining the variance-covariance through a smaller number combination of the original variables. However, the entire original set of variables is required to calculate a principal component. PCA does not provide a methodology for reducing the dimensionality in terms of the original variables (Taguchi and Jugulum, 2002). MTS does provide such a methodology by constructing a Mahalanobis space from samples.

In PCA, the selection of the principal components depends on the correlation matrix. The correlation is based on the entire population. In contrast, MTS provides a measurement scale for multivariate systems using the Mahalanobis space which contains means, standard deviations and correlations. The dimensionality reduction in MTS is based on the ability of the scale to measure conditions outside of the Mahalanobis space. Another key difference between PCA and MTS is the use of orthogonal arrays in MTS to reduce the dimensionality in terms of the original variables (Taguchi and Jugulum, 2002). The reduction in dimensionality is based on Mahalanobis distances and signal-to-noise (S/N) ratios.

REFERENCES

[1] Asada M. (2001), Wafer yield prediction by the Mahalanobis-Taguchi system; *IIE Transactions*; 25-28.

[2] Chatfield C., Collins A.J.(1980), Introduction to Multivariate analysis; Chapman and Hall, London.

[3] Cudney E., Hong J., Jugulum R., Paryani K., Ragsdell K., Taguchi G. (2007), An evaluation of Mahalanobis-Taguchi system and neural network for multivariate pattern recognition; *Journal of Industrial and Systems Engineering* 1(2); 139-150.

[4] Cudney E., Paryani K., Ragsdell K. (2006), Applying the Mahalanobis-Taguchi system to vehicle handling; *Concurrent Engineering: Research & Applications* 14(4); 343-354.

[5] Cudney E., Paryani K., Ragsdell K. (2007), Applying the Mahalanobis-Taguchi system to vehicle ride; *Journal of Industrial and Systems Engineering* 1(3); 251-259.

[6] Cudney E., Paryani K., Ragsdell K.(2008), Identifying useful variables for vehicle braking using the adjoint matrix approach to the Mahalanobis-Taguchi System; *Journal of Industrial and Systems Engineering* 1(4); 281-292.

[7] Fisher R.A. (1936), The Use of Multiple Measurements in taxonomic problems; *Annals of Eugenics* 7(2); 179-188.

[8] Hayashi S., Tanaka Y., Kodama E.(2001), A new manufacturing control system using Mahalanobis distance for maximizing productivity; *IEEE Transactions*; 59-62.

[9] http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc543.htm

[10] Jain Anil K., Duin Robert P.W., Mao Jianchang (2000), Statistical pattern recognition: A review; *IEEE Transaction on Pattern Analysis and Machine Intelligence* 22(1); 4-37.

[11] Johnson R.A. Wichern D.W. (1992), Applied multivariate statistical analysis; 3rd edition, Prentice Hall, Englewood Cliffs, New Jersey.

[12] Lande, U. (2003), Mahalanobis distance: A theoretical and practical approach; http://biologi.uio.no/fellesavdelinger/finse/spatialstats/Mahalanobis%20distance.ppt.

[13] Manly Bryan F.J. (1994), Multivariate statistical methods: A primer; Chapman & Hall, London.

[14] Mason R.L., Tracy N.D., Young J.C. (1992), Multivariate control charts for individual observations; *Journal of Quality Technology* 24; 88-95.

[15] Mason R.L., Tracy N.D., Young J.C. (1995), Decomposition of T^2 for multivariate control chart interpretation; *Journal of Quality Technology* 27; 899-108.

[16] Mardia K.V., Kent J.T., Bibby J.M. (1979), Multivariate analysis; Academic Press, New York, NY.

[17] Minitab Corporation (2005), Minitab Statistical Software, Release 14.20, State College Pennsylvania.

[18] Montgomery D.C. (1985), Statistical quality control; John Wiley & Sons.

[19] Taguchi G., Jugulum R. (2000), New trends in multivariate diagnosis; *Indian Journal of Statistics* 62(B); 233-248.

[20] Taguchi G., Jugulum R. (2002), The Mahalanobis-Taguchi strategy: A pattern technology system; John Wiley & Sons Inc.

[21] Taguchi S. (2000), Mahalanobis Taguchi system; *ASI Taguchi Symposium*.

[22] Wu Y. (1996), Pattern recognition using Mahalanobis distance; *TPD Symposium*; 1-14.