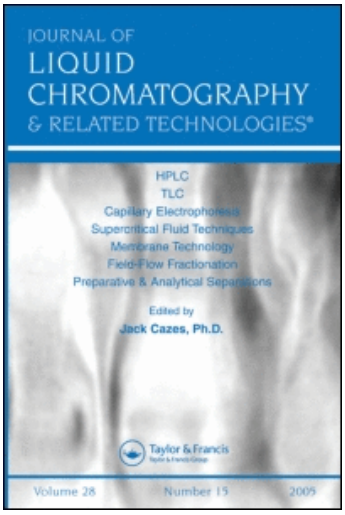

Journal of Liquid Chromatography & Related Technologies
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597273

New Statistical Treatment of the Raw Data from a Gel Permeation Chromatography Analysis with Applications to Thermoplastic Polyurethanes

Dennis J. Keller[a]; Edmond G. Kolycheck[a]
[a] The BFGoodrich Chemical Company Specialty Polymers and Chemicals Division Avon Lake Technical Center, Avon Lake, Ohio

To cite this Article Keller, Dennis J. and Kolycheck, Edmond G.(1990) 'New Statistical Treatment of the Raw Data from a Gel Permeation Chromatography Analysis with Applications to Thermoplastic Polyurethanes', Journal of Liquid Chromatography & Related Technologies, 13: 10, 2035 — 2047
To link to this Article: DOI: 10.1080/01483919008049011
URL: http://dx.doi.org/10.1080/01483919008049011

JOURNAL OF LIQUID CHROMATOGRAPHY, 13(10), 2035-2047 (1990)

NEW STATISTICAL TREATMENT OF THE RAW DATA FROM A GEL PERMEATION CHROMATOGRAPHY ANALYSIS WITH APPLICATIONS TO THERMOPLASTIC POLYURETHANES

DENNIS J. KELLER
AND EDMOND G. KOLYCHECK
The BFGoodrich Chemical Company
Specialty Polymers and Chemicals Division
Avon Lake Technical Center
P. O. Box 122
Moore & Walker Roads
Avon Lake, Ohio 44012

ABSTRACT

The molecular weight distribution of thermoplastic polyurethanes can be adequately and accurately described by a mixture of two ln-Normal distributions. A new statistical procedure has been developed that utilizes data in a Gel Permeation Chromatography area slice table to estimate the five distribution-characterizing parameters. The statistical method is described in a step-by-step approach and the interpretation of the resultant parameter estimates is explained in detail. Method reproducibility is quantified as is the accuracy of the new method.

INTRODUCTION

A new statistical treatment of the data in a Gel Permeation Chromatography (GPC) area slice table has been developed that generates five new summary statistics for quantifying the molecular weight distribution of thermoplastic polyurethanes (TPU's). At first glance, the molecular weight distribution of any TPU appears to possess what is commonly referred to as a "low

molecular weight tail". From a statistical viewpoint, however, this is not a "tail" of a distribution as characterized by a significant amount of skewness, but rather, is a separate distribution mixed into the main molecular weight distribution. Consequently, this new technique approaches the quantification of the molecular weight distribution of TPU's as a problem of characterizing the two distributions in the mixture, estimating the resulting distributions' characterizing parameters and estimating the mixing parameter (i.e. the weight fraction of material belonging to the low molecular weight distribution).

In the case of TPU's, the underlying molecular weight distribution is adequately described by a simple mixture of two ln-Normal distributions (Johnson and Kotz [1]). Hence, only five parameters/statistics are needed to completely characterize the molecular weight distribution of TPU's, namely:

a) $\mu 1$ and $\sigma 1$, the mean and standard deviation of the ln-Normal distribution associated with the small, low molecular weight distribution.
b) $\mu 2$ and $\sigma 2$, the mean and standard deviation of the ln-Normal distribution associated with the main molecular weight distribution, and
c) p, the mixing fraction.

This new technique is applicable to any polymeric molecular weight distribution where there is a mixture of two distributions, each of which is transformable Normal.

EXPERIMENTAL

Materials

The material used in this study was a thermoplastic polyurethane based on poly(tetramethylene adipate)glycol 1,4-butanediol and 4,4'-diphenylmethane diisocynate.

Procedure

Typically polyurethanes are run on a Waters component system (see equipment) using THF stabilized with 250 ppm BHT. The flow rate is 1.0 ml/minute with the sample concentration nominally at 0.15 percent by weight. The injection amount is 100 microliters. The five Pl-Gel 10 micron columns are connected in series in the following manner: 10**5 Å, 10**4 Å, 10**3 Å, 500 Å, 10**6 Å. The

columns are heated and held constant at 40°C. The Refractive Index Detector is also held at a constant temperature of 40°C.

Equipment

All GPC components used were manufactured by Waters:

Model 590 HPLC Pump

Model 712 Wisp Automatic Sampler

Model 410 Differential Refractometer

GPC data manipulations are carried out using the Nelson Analytical, Inc. Model 2600 Chromatography Software Package, on an IBM Personal Computer AT.

Statistical Software

All statistical analyses were conducted on BFGoodrich-written FORTRAN programs. These programs, five in all, employ several IMSL subroutines in various calculations. The programs were run on a VAX 11/750 VMS.

NEW METHOD

The method involves two stages in order to achieve the desired parameter estimates. The first stage uses a simple graphical statistical technique that provides estimates which are close to the "optimal" parameter estimates. "Optimal" will be discussed later. The second stage uses the initial parameter estimates from stage one as starting values in a nonlinear regression (Marquardt [2], [3]). These two stages will be described briefly below. The first stage consists of simply plotting the cumulative weight fraction values from the area slice table versus their corresponding ln-transformed molecular weight values on Normal probability paper (Montgomery [4]). If the underlying distribution of the TPU were actually just a simple ln-Normal distribution, the result of this plot would be a straight line. However, for all TPU's investigated to date, the result is a typical pattern that indicates a mixture of two ln-Normal distributions. See Figure 1.

This plot can be used to estimate p, the mixing fraction parameter. Through a substantial amount of mathematical effort it can be shown that a rough approximation to p can be obtained from the inflection point of this curve. (Goodness of this initial estimate depends upon the degree of overlap or contamination of the two distributions involved in the mixture.) Let this estimate of p be denoted by $\hat{p}$.

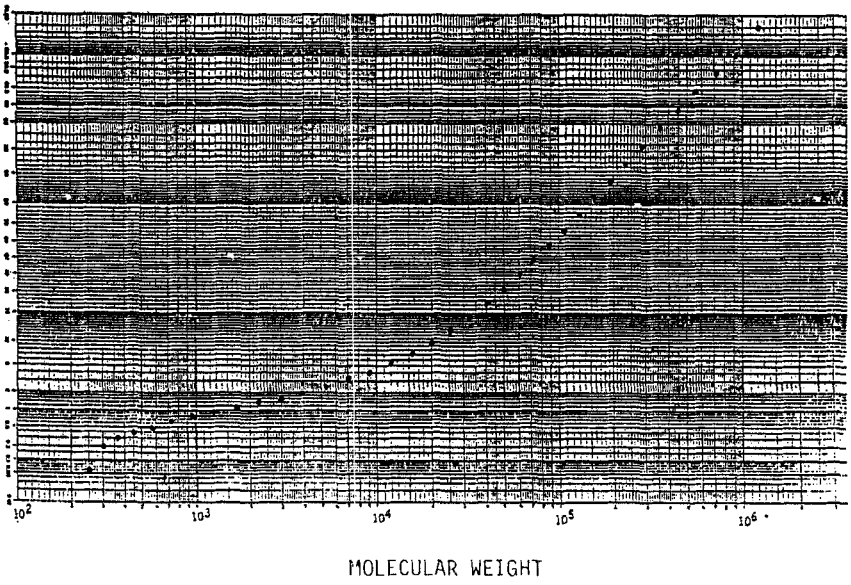

FIGURE 1

Typical Probability Plot for TPU's

Next $\hat{p}$ is used to split the original data set into two portions, those whose cumulative weight fraction is less than $\hat{p}$ and those whose cumulative weight fraction is greater than $\hat{p}$. These two data sets will be referred to as data set 1 and data set 2, respectively.

Now $\hat{p}$ is used again to rescale the cumulative weight fractions in each data set so that the cumulative weight fractions for each data set once again span the range of nominally 0 to 1. This is accomplished using equations 1 and 2 below for data sets 1 and 2, respectively.

Eq. (1) Data Set 1 Rescaled Cumulative Weight Fraction Values = Original Cumulative Weight Fraction Values/$\hat{p}$

Eq. (2) Data Set 2 Rescaled Cumulative Weight Fraction Values = (Original Cumulative Weight Fraction Values - $\hat{p}$)/(1-$\hat{p}$)

Lastly in stage one, the rescaled cumulative weight fraction values are plotted versus their corresponding ln-transformed molecular weight values on Normal probability paper for each data set separately. In each case the result should nominally be a

straight line. Let them be denoted as L1 for data set 1 and L2 for data set 2. Using L1, then, one can simply employ the 50% rule to obtain an estimate of $\mu 1$ and the 84% rule to obtain an estimate of $\mu 1 + \sigma 1$[1]. Subtracting the estimate of $\mu 1$ from the estimate of $\mu 1 + \sigma 1$ provides the desired estimate of $\sigma 1$. Let the estimates of $\mu 1$ and $\sigma 1$ be denoted $\hat{\mu 1}$ and $\hat{\sigma 1}$, respectively. Similarly, the 50% and 84% rules can be applied to L2 to obtain estimates of $\mu 2$ and $\sigma 2$ to be denoted $\hat{\mu 2}$ and $\hat{\sigma 2}$.

1: For normally distributed data, the 50th and 84th percentiles can be used to estimate μ and $\mu + \sigma$, respectively. (See Montgomery [4], p. 281.)

As mentioned above, stage two of the new method employs a nonlinear regression technique to "fine tune" the estimates from stage one. Nonlinear regression requires three essential inputs:

1. data, which in this instance are the (molecular weight, cumulative weight fraction) pairs from the GPC area slice table.
2. a model, which is given in equation 3.
3. initial parameter estimates, which are supplied from stage one, namely, $\hat{\mu 1}$, $\hat{\sigma 1}$, $\hat{\mu 2}$, $\hat{\sigma 2}$ and $\hat{p}$.

Eq. (3)

$$F^{-1}(\text{Cum. Fract}) = \int_{-\infty}^{\ln MW} [\, p\, f(x;\mu 1,\sigma 1) + (1-p)\, f(x;\mu 2,\sigma 2)\,]\, dx$$

$$\text{where } f(x;\mu,\sigma) = (2\pi\sigma)^{-1/2} \exp\left[-1/2 \left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\text{and } F(x) = \int_{-\infty}^{x} f(t;0,1)dt$$

Careful inspection of equation 3 reveals that the model being fit to the original (molecular weight, cumulative weight fraction) pairs is comprised of two parts. The left hand side represents cumulative weight fraction values being "plotted" on

Normal probability paper. (It can be shown that F^{-1} is the proper transformation for linearizing the highly nonlinear vertical axis on Normal probability paper.) The right hand side of equation 3 represents the cumulative probability for a mixture of two ln-Normal molecular weight distributions (Johnson and Leone [5]).

The nonlinear regression technique, then, searches in a user-specified neighborhood around the stage one estimates, $\hat{\mu}1$, $\hat{\sigma}1$, $\hat{\mu}2$, $\hat{\sigma}2$ and $\hat{p}$ to find the "best" combination of the five parameter values that "best" fit the (cumulative weight fraction, molecular weight) data pairs. The statistical criterion for "best" in this case is that combination of parameter values that minimizes the squared distances between the actual cumulative weight fraction values and the model predicted cumulative weight fraction values summed over all the original data points. (It should be noted that these actual-minus-observed differences are in the F^{-1} metric.)

The final result of conducting stage two of this method is a new "fine tuned" set of parameter estimates. Because the estimates from stage one are no longer needed, the new estimates from stage two will simply replace them and can, therefore, be denoted as were the stage one estimates, $\hat{\mu}1$, $\hat{\sigma}1$, $\hat{\mu}2$, $\hat{\sigma}2$ and $\hat{p}$.

The resultant output from this new method consists of:

1. the updated estimates, $\hat{\mu}1$, $\hat{\sigma}1$, $\hat{\mu}2$, $\hat{\sigma}2$ and $\hat{p}$
2. a computer probability plot of the original (cumulative weight fraction, molecular weight) data pairs with the curve of model predicted cumulative weight fraction values. See Figure 2.

PARAMETER INTERPRETATION

Recall from above the definition of each of five parameters/summary statistics:

a) $\hat{\mu}1$ and $\hat{\sigma}1$ are the estimated mean and standard deviation of the ln-Normal distribution associated with the small, low molecular weight distribution.

b) $\hat{\mu}2$ and $\hat{\sigma}2$ are the estimated mean and standard deviation of the ln-Normal distribution associated with the main molecular weight distribution, and

c) $\hat{p}$ is the mixing fraction.

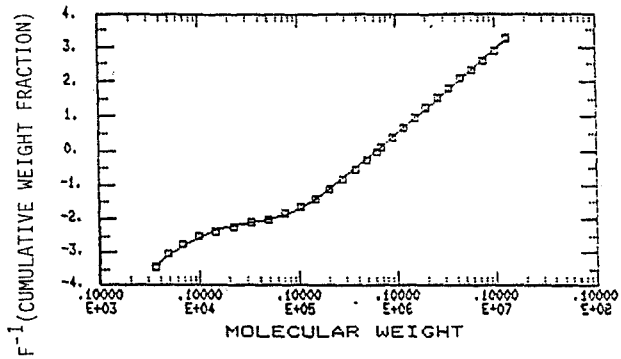

FIGURE 2

Typical Output Plot from New Method

The interpretation of $\hat{p}$ is straightforward. It is the estimated weight fraction of material belonging to the low molecular weight distribution. Obviously, then, $1-\hat{p}$ is the estimated weight fraction of material belonging to the main molecular weight distribution. For the three BFGoodrich products examined, as well as for several other companys' TPU's, the range of $\hat{p}$ values obtained was approximately 0.01 to 0.10 or 1-10%.

The interpretation of the other four summary statistics is also simple with one minor wrinkle and that is that they are in natural logarithmic units. To backtransform the estimated means, $\hat{\mu}1$ and $\hat{\mu}2$, to estimated means in original molecular weight units requires the use of equations 4 and 5.

Eq. (4) Backtransformed $\hat{\mu}1 = e^{\hat{\mu}1}\, e^{\hat{\sigma}1^2/2}$

Eq. (4) Backtransformed $\hat{\mu}2 = e^{\hat{\mu}2}\, e^{\hat{\sigma}2^2/2}$

Unlike the two means, the two standard deviations cannot be directly backtransformed to meaningful counterparts in original molecular weight units. They can, however, be extremely useful in making accurate and meaningful statements about the two molecular weight distributions. For example, two standard

TABLE 1

95% Test Reproducibility for the New Summary Statistics

	$\hat{\mu}1$	$\hat{\sigma}1$	$\hat{\mu}2$	$\hat{\sigma}2$	$\hat{p}$
S test	0.2610	0.1075	0.0196	0.0190	0.0046
95% Test Rep.	$\pm$0.5319	$\pm$0.2191	$\pm$0.0399	$\pm$0.0387	$\pm$0.0094

statements about a distribution (other than its mean) are its 95% probability limits and its 99% probability limits. These can be generated in original molecular weight units for each of the two ln-Normal distributions using equations 6, 7, 8 and 9.

Eq. (6) Backtransformed 95% probability limits for the small, lower molecular weight distribution = $EXP(\hat{\mu}1 \pm 1.960\hat{\sigma}1)$

Eq. (7) Backtransformed 95% probability limits for the main molecular weight distribution = $EXP(\hat{\mu}2 \pm 1.960\hat{\sigma}2)$

Eq. (8) Backtransformed 99% probability limits for the small, lower molecular weight distribution = $EXP(\hat{\mu}1 \pm 2.576\hat{\sigma}1)$

Eq. (9) Backtransformed 99% probability limits for the main molecular weight distribution = $EXP(\hat{\mu}2 \pm 2.576\hat{\sigma}2)$

REPRODUCIBILITY

A statistically designed study was conducted to determine the test reproducibility of the new method. The same GPC machine was used throughout the study as was the same operator. Hence, the components of variance included in the estimate of test reproducibility were machine error, sample preparation error, day-to-day variation and method error. The resulting 95% test reproducibility figures are presented in Table 1.

ACCURACY

To check the accuracy of the new method, an in-depth simulation study was conducted in which a wide variety of combinations of parameter values was investigated. The specific combinations were chosen according to a statistical experimental design to investigate the effects of distribution overlap (i.e. contamination of one ln-Normal distribution into the other) on parameter estimation accuracy.

TABLE 2

Results from Simulation Study
for Assessing Parameter Estimation Accuracy

Exp	Actual Parameter Values					Predicted Parameter Values				
#	$\mu 1$	$\sigma 1$	$\mu 2$	$\sigma 2$	p	$\hat{\mu 1}$	$\hat{\sigma 1}$	$\hat{\mu 2}$	$\hat{\sigma 2}$	$\hat{p}$
1	5	1.5	14	1.5	0.20	4.969	1.491	14.00	1.499	0.1994
2	5	1.5	14	1.0	0.01	4.960	1.487	14.00	1.001	0.0099
3	5	1.5	10	1.5	0.01	4.629	1.410	9.97	1.513	0.0078
4	5	1.0	12	1.5	0.50	5.010	1.004	12.02	1.490	0.5017
5	5	0.5	14	0.5	0.01	5.000	0.500	14.00	0.500	0.0100
6	5	0.5	10	0.5	0.01	5.000	0.500	10.00	0.500	0.0100
7	5	0.5	10	0.5	0.50	5.001	0.500	10.00	0.500	0.5000
8	5	1.5	10	1.0	0.20	4.848	1.453	9.96	1.018	0.1885
9	5	0.5	10	1.0	0.10	5.002	0.501	10.01	0.992	0.1007
10	5	1.0	14	0.5	0.01	4.998	0.100	14.00	0.500	0.0100
11	5	0.5	12	1.0	0.20	4.999	0.500	11.10	1.003	0.2000
12	5	0.5	14	0.5	0.50	5.002	0.501	14.00	0.501	0.5000
13	5	1.0	12	1.0	0.10	4.993	0.996	12.01	0.994	0.1001
14	5	1.5	10	1.5	0.50	4.967	1.488	10.06	1.481	0.5008
15	5	1.5	14	0.5	0.50	5.002	1.502	14.00	0.502	0.5000
16	5	1.5	14	1.5	0.10	4.961	1.487	14.00	1.498	0.9947
17	5	1.0	14	1.5	0.01	4.969	0.990	14.00	1.498	0.0099
18	5	1.0	10	1.5	0.20	4.983	0.992	10.07	1.472	0.2073
19	5	0.5	12	1.5	0.10	5.012	0.506	12.00	1.502	0.9922
20	5	1.0	14	1.0	0.50	5.003	1.001	14.00	1.002	0.5000
21	5	2.0	10	2.0	0.01	3.534	1.682	9.93	2.024	0.0038
22	5	2.0	10	2.0	0.10	4.740	1.944	9.93	2.027	0.0867
23	5	2.0	10	2.0	0.50	4.961	1.981	10.15	1.951	0.5058
24	5	3.0	10	3.0	0.01	-3.619	1.094	9.88	3.072	0.0001
25	5	3.0	10	3.0	0.10	2.000	2.355	9.67	3.111	0.0241
26	5	3.0	10	3.0	0.50	4.976	2.978	10.36	2.900	0.5381

TABLE 3

Candidate Distance Criteria

$$DC1 = \frac{(\mu 2-\mu 1)}{(\sigma 1+\sigma 2)}$$

$$DC2 = p \times \frac{(\mu 2-\mu 1)}{(\sigma 1+\sigma 2)}$$

$$DC3 = p^{1/2} \times \frac{(\mu 2-\mu 1)}{(\sigma 1+\sigma 2)}$$

$$DC4 = \frac{1}{(1-p)} \times \frac{(\mu 2-\mu 1)}{(\sigma 1+\sigma 2)}$$

Each simulation was conducted as follows. First an "exact" data set was generated using the specific combination of parameter values under investigation in conjunction with the model shown in equation 3 above. Then this data set was analyzed using the new method as though it came directly from a GPC area slice table. Comparisons of the actual parameter values versus new method parameter estimates are given in Table 2.

Scenarios 1-20 were part of the original simulation experimental design. Notice that of these 20 simulations, scenario 3 yields the poorest agreement between actual and estimated parameter values. This is because scenario 3 represents the most severe distribution overlap since $\mu 1$ and $\mu 2$ are close and $\sigma 1$ and $\sigma 2$ are large. To further test the estimation procedure, scenarios 21-26 were added. Scenarios 21-23 represent an extension of scenario 3 with larger standard deviations at three different levels of p. Scenarios 24-26 investigate a further extension of scenario 3 with even larger standard deviations again at three different levels of p.

Visual inspection of the simulation results suggests that, in general, the new method provides very accurate parameter estimates except where there is severe contamination of one distribution into the other AND where p is small. Hence, if p is small AND $\mu 1$ is "close" to $\mu 2$ relative to $\sigma 1$ and $\sigma 2$, then parameter estimates may not be accurate.

TABLE 4

Candidate "Goodness-of-Parameter-Estimation" Criteria

$$GC1 = \left[\left(\frac{\mu1-\hat{\mu1}}{\mu1}\right)^2 + \left(\frac{\sigma1-\hat{\sigma1}}{\sigma1}\right)^2 + \left(\frac{\mu2-\hat{\mu2}}{\mu2}\right)^2 + \left(\frac{\sigma2-\hat{\sigma2}}{\sigma2}\right)^2 + \left(\frac{p-\hat{p}}{p}\right)^2 \right]^{1/2} \times 100$$

$$GC2 = \left[\frac{|\mu1-\hat{\mu1}|}{\mu1} + \frac{|\sigma1-\hat{\sigma1}|}{\sigma1} + \frac{|\mu2-\hat{\mu2}|}{\mu2} + \frac{|\sigma2-\hat{\sigma2}|}{\sigma2} + \frac{|p-\hat{p}|}{p} \right] \times 100$$

$$GC3 = |\mu1-\hat{\mu1}| + |\sigma1-\hat{\sigma1}| + |\mu2-\hat{\mu2}| + |\sigma2-\hat{\sigma2}| + |p-\hat{p}|$$

INDEX OF PARAMETER ESTIMABILITY

Based on the above simulation results from the statistically designed study, an index of parameter estimability was concocted. The purpose of such an index is to define in a single statistic, a measure of distribution overlap (i.e. a simultaneous measure of "smallness" of p and "closeness" of $\mu1$ and $\mu2$ relative to $\sigma1$ and $\sigma2$). Several candidate "distance-between-distributions" criteria were explored as were several candidate "goodness-of-parameter-estimation" criteria. These are listed in Tables 3 and 4, respectively.

Statistical analysis revealed that GC1 (the standard sum of squared deviations) was the best descriptor of goodness of parameter estimation and that DC4 was the best index of distribution overlap that also incorporated some effect of p. It was also determined that a highly statistically significant relationship existed between these two entities on a ln-ln basis. See Figure 3.

The usefulness of such an index as DC4 may not be readily apparent. Realize that in actual practice, one does not know what the "true" parameter values are. One only has the estimates, $\hat{\mu1}$, $\hat{\sigma1}$, $\hat{\mu2}$, $\hat{\sigma2}$ and $\hat{p}$ that are output from this new method. These estimates can be used to calculate a DC4 index value, which in turn can be used to predict a GC1 value. Notice from Figure 3 above, that if the resultant DC4 is greater than

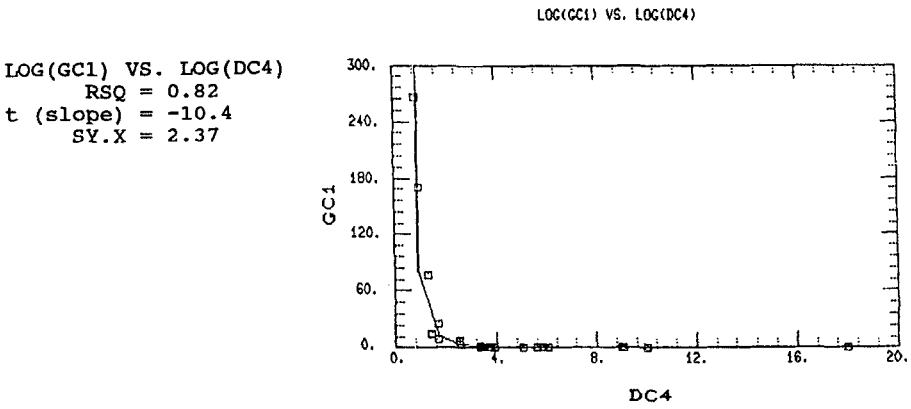

FIGURE 3

Relation Between Goodness Criteria and Distance Criteria

1.6, then GC1, the "combined" relative errors in $\hat{\mu}1$, $\hat{\sigma}1$, $\hat{\mu}2$, $\hat{\sigma}2$ and $\hat{p}$, is no worse than roughly 10%.

CONCLUSIONS

A new method of analyzing the (cumulative weight fraction, molecular weight) data pairs from a GPC area slice table has been developed. This new method provides five new summary statistics that not only completely characterize the molecular distribution of TPU's, but also offer an added advantage of ease of interpretation as follows:

a) $\hat{\mu}1$ and $\hat{\sigma}1$ are the estimated mean and standard deviation of the ln-Normal distribution associated with the small, low molecular weight distribution.

b) $\hat{\mu}2$ and $\hat{\sigma}2$ are the estimated mean and standard deviation of the ln-Normal distribution associated with the main molecular weight distribution, and

c) $\hat{p}$ is the estimated fraction of total TPU in the small, low molecular weight distribution.

As explained above, these estimates can be used with ease to make a multitude of conclusions about the TPU molecular weight distribution under investigation.

It has been demonstrated that these new summary statistics are accurate and reproducible. Only under cases of severe distribution overlap does the accuracy of parameter estimation fall apart.

An index of distribution overlap, DC4, has been provided that can be related to an index of goodness of parameter estimability, GC1. Hence, a resultant output set of $\hat{\mu}1$, $\hat{\sigma}1$, $\hat{\mu}2$, $\hat{\sigma}2$ and $\hat{p}$ can be used to calculate a DC4 and ultimately, to predict a GC1 to ensure accuracy of parameter estimation.

ACKNOWLEDGMENTS

The authors would like to thank BFGoodrich for allowing the public release of this information. The authors would also like to thank the following people for their support and contributions: J. Berrettoni, B. Cross, N. Drooby, D. Harmon, B. Kantola, R. Khattree, T. Lapp, A. Lehman, M. Mehta, J. Nelder, G. Perkins, M. Ritland, V. Smith, F. Wehrfritz, J. Christ.

REFERENCES

[1] Johnson, N.L. and Kotz S., Continuous Univariate Distributions - 1, John Wiley & Sons, New York, 1970, p. 112.

[2] Marquardt, D. W., Solution of Nonlinear Chemical Engineering Models, Chemical Engineering Progress, Vol. 55, No. 6, p. 65, June, 1959.

[3] Marquardt, D. W., An Algorithm for Least-Squares Estimation of Nonlinear Parameters, Journal of Society of Industrial Applied Mathematics, Vol. 11, No.2, p. 431, June, 1963.

[4] Montgomery, D. C., Introduction to Statistical Quality Control, John Wiley & Sons, New York, 1985, p. 280.

[5] Johnson, N. L. and Leone, F. C., Statistics and Experimental Design in Engineering and the Physical Sciences, John Wiley & Sons, New York, 1964, 1977, p. 157.