Computer-Assisted Data Treatment in Analytical Chemometrics III. Data Transformation

[a]M. MELOUN and [b]J. MILITKÝ

[a]*Department of Analytical Chemistry, Faculty of Chemical Technology, University Pardubice, CZ-532 10 Pardubice*

[b]*Department of Textile Materials, Technical University, CZ-461 17 Liberec*

Received 30 April 1993

In trace analysis an exploratory data analysis (EDA) often finds that the sample distribution is systematically skewed or does not prove a sample homogeneity. Under such circumstances the original data should often be transformed. The power simple transformation and the Box—Cox transformation improves a sample symmetry and also makes stabilization of variance. The Hines—Hines selection graph and the plot of logarithm of the maximum likelihood function enables to find an optimum transformation parameter. Procedure of data transformation in the univariate data analysis is illustrated on quantitative determination of copper traces in kaolin raw.

When exploratory data analysis shows that the sample distribution strongly differs from the normal one, we are faced with the problem of how to analyze the data. Raw data may require re-expression to produce an informative display, effective summary, or a straightforward analysis [1—10]. We may need to change not only the units in which the data are stated, but also the basic scale of the measurement. To change the shape of a data distribution, we must do more than change the origin and/or unit of measurement. Changes of origin and scale mean linear transformations, and they leave shape alone. Nonlinear transformations such as the logarithm and square root are necessary to change shape.

This paper brings a description of the power transformation and the Box—Cox transformation and a re-expression of statistics for transformed data. The procedure of the power transformation and the Box—Cox transformation is illustrated on a practical example of the quantitative determination of copper traces in kaolin raw.

THEORETICAL

Examining data we must often find the *proper transformation* which leads to symmetrizing data distribution, stabilizes the variance or makes the distribution closer to normal. Such transformation of original data x to new variable value $y = g(x)$ is based on an assumption that the data represent a nonlinear transformation of normally distributed variable $x = g^{-1}(y)$.

i) *Transformation for variance stabilization* implies ascertaining the transformation $y = g(x)$ in which the variance $\sigma^2(y)$ is constant. If the variance of the original variable x is a function of the type $\sigma^2(x) = f_1(x)$, the variance $\sigma^2(y)$ may be expressed by

$$\sigma^2(y) \approx \left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right)^2 f_1(x) = C \qquad (1)$$

where C is a constant. The chosen transformation $g(x)$ is then the solution of the differential equation

$$g(x) = C\int \frac{\mathrm{d}x}{\sqrt{f_1(x)}} \qquad (2)$$

In some instrumental methods of analytical and physical chemistry, the relative standard deviation $\delta(x)$ of the measured variable is constant. This means that the variance $\sigma^2(x)$ is described by a function $\sigma^2(x) = f_1(x) = \delta^2(x)\, x^2 = \text{const}\, x^2$. The substitution into eqn (2) will be $g(x) = \ln x$, so that an optimal transformation of original data is the logarithmic transformation. This transformation leads to the use of a geometric mean.

When the dependence $\sigma^2(x) = f_1(x)$ is of power nature, the optimal transformation will also be a power transformation. Since for a normal distribution the mean is not dependent on a variance, a transformation that stabilizes the variance makes the distribution closer to normal.

ii) *Transformation for symmetry* is carried out by a simple power transformation

$$y = g(x) = \begin{Bmatrix} x^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln x & \text{for parameter } \lambda \neq 0 \\ -x^{-\lambda} & \text{for parameter } \lambda < 0 \end{Bmatrix} \quad (3)$$

which does not retain the scale, is not always continuous and is suitable only for positive x. Optimal estimates of parameter $\hat{\lambda}$ are sought by minimizing the absolute values of particular characteristics of asymmetry. In addition to the classical estimate of a skewness $\hat{g}_1(y)$, the robust estimate $\hat{g}_{1,R}(y)$ is used

$$\hat{g}_{1,R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{(\tilde{y}_{0.75} - \tilde{y}_{0.25})} \quad (4)$$

The robust estimate of asymmetry $\hat{g}_P(y)$ may be also expressed with the use of a relative distance between the arithmetic mean $\bar{y}$ and the median $\tilde{y}_{0.50}$ by

$$\hat{g}_P(y) = \frac{\bar{y} - \tilde{y}_{0.50}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} \quad (5)$$

as for symmetric distributions it is equal to zero, $\hat{g}_P(y) \approx 0$.

iii) *Transformation leading to the approximate normality* may be carried out by the use of family of *Box—Cox transformation* defined as

$$y = g(x) = \begin{Bmatrix} (x^{\lambda} - 1)/\lambda & \text{for parameter } \lambda \neq 0 \\ \ln x & \text{for parameter } \lambda = 0 \end{Bmatrix} \quad (6)$$

where x is a positive variable and λ is real number. Box—Cox transformation has the following properties:

a) The curves of transformation $g(x)$ are monotonic and continuous with respect to parameter λ because

$$\lim_{\lambda \to 0} \frac{(x^{\lambda} - 1)}{\lambda} = \ln x \quad (7)$$

b) All transformation curves share one point [$y = 0$, $x = 1$] for all values of λ. The curves nearly coincide at points close to [0, 1]; *i.e.* they share a common tangent line at that point.

c) The power transformations of exponent -2; $-3/2$; -1; $-1/2$; 0; 1/2; 1; 3/2; 2 have equal spacing between curves in the family of Box—Cox transformation graph.

The Box—Cox transformation defined by eqn (*6*) can be applied only on the positive data. To extend this transformation means to make a substitution of x values by $(x - x_0)$ values which are always positive. Here x_0 is the threshold value $x_0 < x_{(1)}$.

An excellent diagnostic tool enabling estimation of parameter λ is represented by the Hines—Hines selection graph [8]. It is based on the equation

$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2 \quad (8)$$

valid for distribution symmetrical around a median. For the cumulative probability $P_i = 2^{-i}$, the letter values F, E, $i = 2, 3$ are usually chosen.

To compare empirical dependence of experimental points with the ideal one, ideal curves for various values of parameter λ are drawn in a selection graph. These curves λ represent a solution of the equation $y^{\lambda} + x^{-\lambda} = 2$ in the range $0 \leq x \leq 1$ and $0 \leq y \leq 1$:

1. For $\lambda = 0$ the solution is a straight line $y = x$.
2. For $\lambda \leq 0$ the solution is in a form $y = (2 - x^{-\lambda})^{1/\lambda}$.
3. For $\lambda \geq 0$ the solution is in a form $x = (2 - y^{\lambda})^{-1/\lambda}$.

The estimate $\hat{\lambda}$ is guessed from a selection graph, according to the location of experimental points near to the various ideal curves.

To estimate the parameter λ in Box—Cox transformation, the method of maximum likelihood may be used because for $\lambda = \hat{\lambda}$ a distribution of transformed variable y is considered to be normal, $N(\mu_y, \sigma^2(y))$. The logarithm of the maximum likelihood function may be written as

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \quad (9)$$

where $s^2(y)$ is the sample variance of transformed data y. The function $\ln L = f(\lambda)$ is expressed graphically for a suitable interval, for example, $-3 \leq \lambda \leq 3$. The maximum on this curve represents the maximum likelihood estimate $\hat{\lambda}$.

The asymptotic $100(1 - \alpha)$ % confidence interval of parameter λ is expressed by

$$2\left[\ln L(\hat{\lambda}) - \ln L(\lambda)\right] \leq \chi^2_{1-\alpha}(1) \quad (10)$$

where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This interval contains all values λ for which it is true that

$$\ln L(\lambda) \geq \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1) \quad (11)$$

This Box—Cox transformation is less suitable if confidence interval for λ is too wide. When the value $\lambda = 1$ is also covered by this confidence interval, the transformation is not efficient.

After an appropriate transformation of the original data $\{x\}$ has been found, so that the transformed data give approximately normal symmetrical distribution with constant variance, the statistical measures of location and spread for the transformed data $\{y\}$ are calculated. These include the sample mean $\bar{y}$, the sample variance $s^2(y)$, and the confidence interval of the mean $\bar{y} \pm t_{1-\alpha/2}\,(n-1)s(y)/(n)^{1/2}$. These estimates must then be recalculated for original data $\{x\}$. Two different approaches to re-expression of the statistics for transformed data can be simply used:

1. Rough re-expressions represent a single reverse transformation $\bar{x}_R = g^{-1}(y)$. This re-expression for a simple power transformation leads to the general mean

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^\lambda}{n}\right]^{\frac{1}{\lambda}} \qquad (12)$$

where for $\lambda = 0$, ln x is used instead of x^λ and e^x instead of $x^{1/\lambda}$. The re-expressed mean $\bar{x}_R = \bar{x}_{-1}$ stands for the *harmonic mean*, $\bar{x}_R = \bar{x}_0$ for the *geometric mean*, $\bar{x}_R = \bar{x}_1$ for the *arithmetic mean,* and $\bar{x}_R = \bar{x}_2$ for the *quadratic mean*.

2. The *more correct re-expressions* are based on the Taylor series expansion of the function $y = g(x)$ in a neighbourhood of the value $\bar{y}$. The re-expressed mean $\bar{x}_R$ is then given

$$\bar{x}_R \approx g^{-1}\left\{\bar{y} - \frac{1}{2}\frac{d^2g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right\} \qquad (13)$$

For variance it is then valid

$$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y) \qquad (14)$$

where individual derivatives are calculated at the point $x = \bar{x}_R$. The $100(1-\alpha)$ % confidence interval of the re-expressed mean for the original data may be defined as

$$\bar{x}_R - l_L \le \mu \le \bar{x}_R + l_U \qquad (15)$$

where

$$l_L = g^{-1}\left[\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right] \qquad (16a)$$

$$l_U = g^{-1}\left[\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right] \qquad (16b)$$

$$G = -\frac{1}{2}\frac{d^2g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y) \qquad (17)$$

On the basis of the (known) actual transformation $y = g(x)$ and the estimates $\bar{y}$, $s^2(y)$ it is easy to calculate re-expressed estimates $\bar{x}_R$ and $s^2(\bar{x}_R)$:

1. For a logarithmic transformation (when $\lambda = 0$) and $g(x) = \ln x$ the re-expressed mean and variance are calculated by eqns (*18*) and (*19*)

$$\bar{x}_R \approx \exp\left[\bar{y} + 0.5s^2(y)\right] \qquad (18)$$

and

$$s^2(x_R) \approx \bar{x}_R^2\, s^2(y) \qquad (19)$$

2. For $\lambda \neq 0$ and the Box—Cox transformation (*7*) the re-expressed mean $\bar{x}_R$ will be represented by one of the two roots of the quadratic equation

$$\bar{x}_{R,1,2} = \left[0.5(1+\lambda\bar{y}) \pm \right.$$
$$\left. \pm\, 0.5\sqrt{1+2\lambda(y+s^2(y))+\lambda^2(y^2-2s^2(y))}\right]^{1/\lambda} \qquad (20)$$

which is closest to the median $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. If $\bar{x}_R$ is known the corresponding variance may be calculated from

$$s^2(x) = \bar{x}_R^{(-2\lambda+2)} s^2(y) \qquad (21)$$

COMPUTATION

Procedure POWER TRANSFORM in package ADSTAT [11] searches parameters of simple power transformation and parameters of normalized Box—Cox transformation of data. It enables the exploratory data analysis of transformed data. For the transformation (*3*) different measures of symmetry (*4*) and (*5*) are calculated and the sample curtosis in the range $-3 \le \lambda \le 3$ with a step 0.1 and the optimal values of these measures are printed. The selection graph is drawn as well as the points of optimal values of λ. From this graph the value of λ can be estimated. Using transformed data the mean $\bar{y}$, the variance $s^2(y)$, the skewness $\hat{g}_1(y)$, and the curtosis $\hat{g}_2(y)$ are calculated. These computations can be repeated for various values of λ. For the transformation (*6*) the estimate $\hat{\lambda}$ maximizing ln $L(\lambda)$ defined by eqn (*9*) is calculated. Different measures of symmetry (eqns (*4—6*)) and the sample curtosis are searched. Search is obviously realized in the range $-3 \le \lambda \le 3$ with a step 0.1. Optimal values of $\hat{\lambda}$ and corresponding measures are printed. The graph of ln L *vs.* λ with the 95 % confidence interval (*10*) is drawn. From the ln $L = f(\lambda)$ plot the λ value is estimated. Selected $\hat{\lambda}$ is used in calculation of estimates $\bar{y}$, $s^2(y)$, $\hat{g}_1(y)$, and $\hat{g}_2(y)$. Then from these estimates, the re-expressed estimates of original variables $\bar{x}_R$ (*13*), $s^2(\bar{x}_R)$ (*14*), and the 95 % confidence interval of the re-expressed variable μ are calculated.

RESULTS

Study Case 1. Determination of copper trace in kaolin

In a standard sample of kaolin the content of copper trace was determined in ppm and the values were arranged in increasing order. The type of a sample distribution and measures of location and scale were examined.

Data: the copper content *w*/ppm in increasing order gives a set: 4, 5, 7, 7, 7, 8, 8.3, 8.4, 9.4, 9.5, 10, 10.5, 12, 12.8, 13, 22, 23.

Solution: Applying an analysis of basic assumptions about data the following conclusions were met:

a) *Combined sample skewness and curtosis test* leads to statistic $C_1 = 7.908 > \chi^2(0.95, 2) = 5.992$ and therefore a normality of data distribution was rejected.

b) Interval of both *Hoaglin's outer bounds* [– 3.191; 22.191] does not contain one observation and therefore this point $x_{(17)}$ may be denoted as an outlier. The measures of location, scale and distribution shape for data without 1 outlier are $\bar{x} = 9.619$, $s(\bar{x}) = 4.170$, $\hat{g}_1(x) = 1.610$, and $\hat{g}_2(x) = 6.340$.

c) *Test of sample elements independence* leads to statistic $t_{17} = 1.036$, $t_{0.975}\ (18) = 2.101$ and therefore an independence is accepted.

Examining the first part of the EDA diagnostics following sample properties were found: the jittered dot diagrams and the box-and-whisker plots (Fig. 1) indicate two outliers which can be accepted if the distribution is skewed.

The nonparametric kernel estimation of probability density function (Fig. 2) indicate that the distribution is skewed towards higher values. The quantile-quantile (rankit) plot (Fig. 3) with convex increasing shape confirms that the distribution is skewed to higher values.

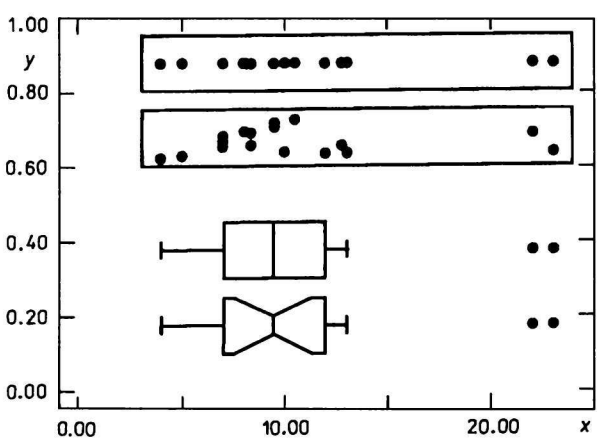

Fig. 1. The jittered diagrams and the box-and-whisker plots of original sample.

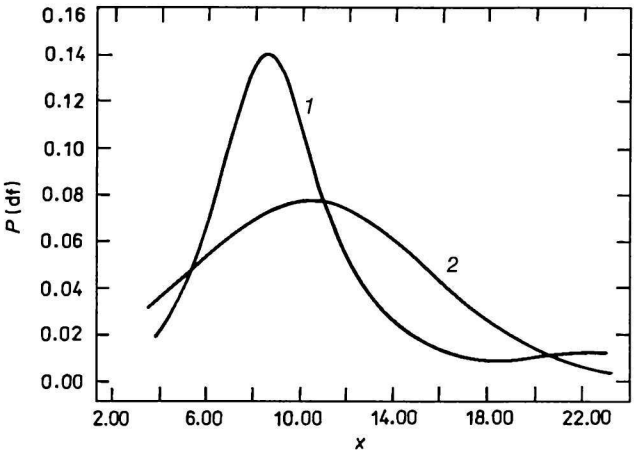

Fig. 2. The kernel estimation of the probability density function of original sample: *1*. robust, *2*. classical.

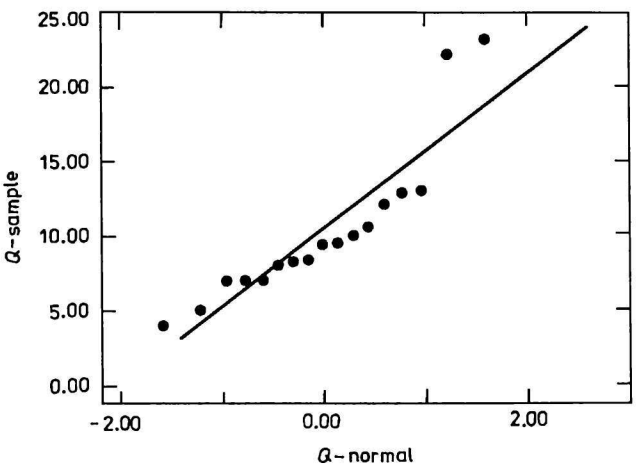

Fig. 3. The quantile-quantile (rankit) plot of original sample.

The second part of EDA concerns the search for a suitable symmetric transformation of the data. The selection graph (Fig. 4) shows that the optimal power reaches a value above – 0.5 in the range near zero which corresponds to a logarithmic transformation.

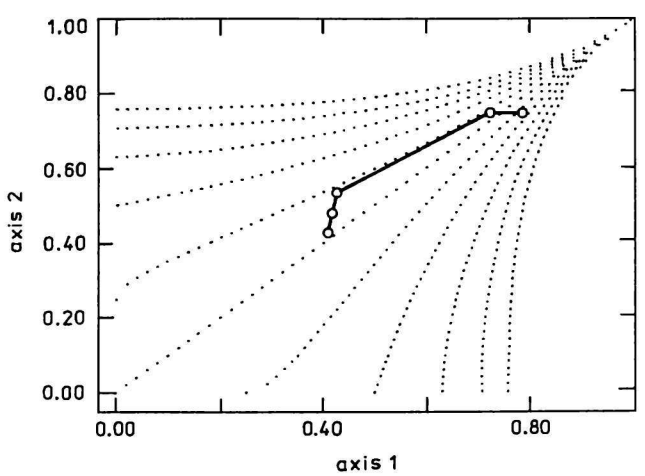

Fig. 4. The Hines—Hines selection graph.

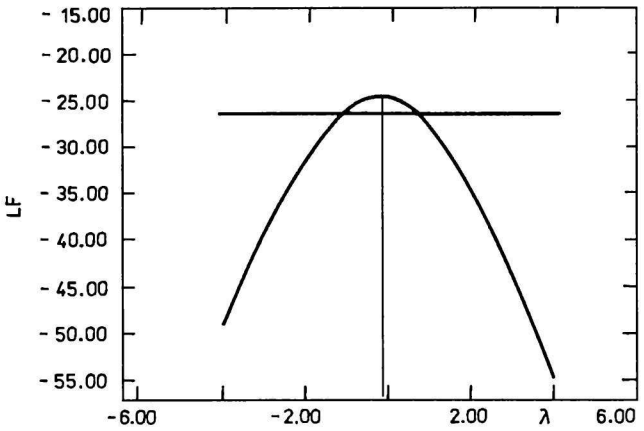

Fig. 5. The plot of logarithm of likelihood function.

From the plot of the logarithm of the likelihood function (Fig. 5) for the Box—Cox transformation the maximum of the curve is at $\lambda = -0.2$. The corresponding 95 % confidence interval does not contain the value $\lambda = 1$, so this transformation is statistically significant. The rankit plot (Fig. 6*a*—*c*) shows that there is a significant improvement in the distribution symmetry for transformation $\hat{\lambda} = -0.27$.

The measures of location, spread and shape for the original data have values of mean $\bar{x} = 10.406$, standard deviation $s(x) = 5.180$, skewness $\hat{g}_1(x) = 1.399$, and curtosis $\hat{g}_2(x) = 4.272$. After a logarithmic transformation ($\hat{\lambda} = 0$) the values are 2.243, 0.203, 0.304, and 3.070, and after a power transformation ($\hat{\lambda} = -0.27$) they are 0.5536, 0.065, 0.048, and 3.071 while the Box—Cox transformation ($\hat{\lambda} = -0.27$) leads to values 1.674, 0.246, −0.048, and 3.071.

By the *rough re-expression* (*12*) $\bar{x}_R = \exp(\bar{x}^*) = 9.337$. The corresponding confidence limits are $l_L = 7.742$ and $l_U = 11.878$ (eqns (*16a*, *16b*)). Quantile $t_{0.975}(17 - 1) = 2.12$.

By the *more correct re-expression* (*13*) there is $\bar{x}_R = 9.187$ with $l_L = 8.272$ and $l_U = 13.147$ (eqn (*20*)).

In comparison of the sample distribution with a theoretical exponential one, the correlation coefficient r_{xy} of the *Q*-*Q* plot is found to be 0.967, while for the log-normal one r_{xy} is 0.961.

The assumption of the log-normal distribution is acceptable. Because of the small sample size it is difficult to be certain whether there are outliers in the sample, or if the sample distribution is of skewed log-normal or of skewed exponential nature.

CONCLUSION

Often, the chemical data are less ideal and do not fulfill all basic assumptions. Original data are then transformed to improve a symmetry of data distribution and a variance stabilization. Statistical measures of transformed data are re-transformed to get these unbiased and rigorous measures for original data.

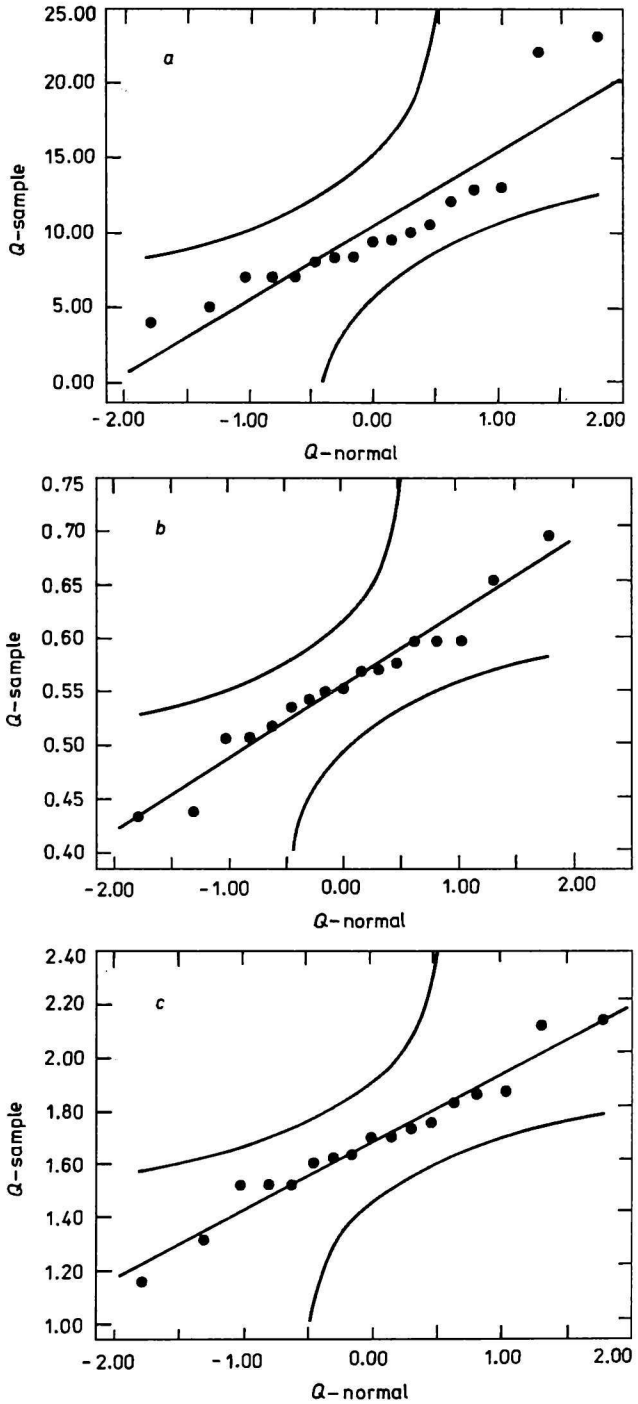

Fig. 6. The quantile-quantile plot indication of improvement of a distribution symmetry of *a*) original data when *b*) the power transformation, and *c*) the Box—Cox transformation are applied.

REFERENCES

1. Tukey, J. W., *Exploratory Data Analysis*. Addison Wesley, Reading, Massachusetts, 1977.

2. Chambers, J., Cleveland, W., Kleiner, W., and Tukey, P., *Graphical Methods for Data Analysis*. Duxbury Press, Boston, 1983.
3. Hoaglin, D. C., Mosteler, F., and Tukey, J. W., *Exploring Data Tables, Trends and Shapes*. Wiley, New York, 1985.
4. Scott, D. W. and Sheater, S. J., *Commun. Statist. 14*, 1353 (1985).
5. Lejenne, M., Dodge, Y., and Koelin, E., *Proceedings of the Conference COMSTAT'82 Toulouse*. P. 173 (Vol. *III*).
6. Hoaglin, D. C., Mosteler, F., and Tukey, J. W. (Editors), *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.
7. Kafander, K. and Spiegelman, C. H., *Comput. Stat. Data Anal. 4*, 167 (1986).
8. Hines, W. G. S. and Hines, R. J. H., *Am. Statist. 41*, 21 (1987).
9. Hoaglin, D. C., *J. Am. Statist. Assoc. 81*, 991 (1986).
10. Stoodley, K., *Applied and Computational Statistics*. Ellis Horwood, Chichester, 1984.
11. *Statistical package ADSTAT 2.0*. TriloByte, Pardubice, 1992.

Translated by M. Meloun

The Testing of Carbon Sorbent for Preconcentration of Volatile Organic Trace Compounds

[a]S. ŠKRABÁKOVÁ, [a]E. MATISOVÁ, [a]M. ONDEROVÁ, [b]I. NOVÁK, and [b]D. BEREK

[a]*Department of Analytical Chemistry, Faculty of Chemical Technology, Slovak Technical University, SK-812 37 Bratislava*

[b]*Polymer Institute, Slovak Academy of Sciences, SK-842 38 Bratislava*

Received 30 April 1993

Carbon sorbent Carb I (prepared by controlled pyrolysis of saccharose) was tested for preconcentration of volatile organic compounds from the gas phase. The model mixture of hydrocarbons (n-alkanes and aromatics) and mixture of aromatics with low-boiling polar solvents was used. For desorption of compounds several solvents were utilized, carbon disulfide was found to be the best. Adsorption—desorption process was studied in the concentration range of components in nitrogen 0.03—15 μg dm^{-3}. Chromatographic measurements were performed on gas chromatograph with on-column and splitless injection, fused silica capillary columns with chemically bonded stationary phases under temperature programmed conditions and flame ionization detector. The recovery of n-alkanes and aromatics was found to be around 90 %, the recovery of low-boiling solvents, particularly of polar character was low.

The main part of toxicological analyses and analyses that have been required with regard to the control of the environment is concentrating upon compounds with very low concentration. In many cases the concentrations of contaminating components are so low, *e.g.* in air, that the common detectors used, *e.g.* in gas chromatography, do not detect them. Organic contaminating components in environmental samples generally occur in ng kg^{-1} to μg kg^{-1} as a part of a complex matrix [1]. Besides that, samples are usually not compatible with chromatographic system, therefore analysis with direct sample injection is not possible.

It is therefore necessary to perform sample pretreatment before an analysis. It is mainly preconcentration of components of interest, isolation of determined analytes from the matrix and removal of potential interferences. For this reason there have been utilized each time more special qualities of sorbent materials [2—4].

In the selection of a proper sorbent it is necessary to take into consideration general characteristics as functional groups at the surface, chemical and thermal stability, as well as inertness and catalytic properties, mechanical resistance, pores diameter and volume, specific surface area, size and shape of particles.

Affinity of sorbents towards various organic compounds depends on the type of functional groups bound on the surface of a sorbent and on their orientation on the surface.