Available online at www.sciencedirect.com
SCIENCE DIRECT®
Environmental Pollution 137 (2005) 273–280
ENVIRONMENTAL POLLUTION
www.elsevier.com/locate/envpol

The analysis of soil cores polluted with certain metals using the Box–Cox transformation

Milan Meloun [a,*], Milan Sáňka [b], Pavel Němec [b], Soňa Křítková [a], Karel Kupka [c]

[a] *Department of Analytical Chemistry, University of Pardubice, CZ532 10 Pardubice, Czech Republic*
[b] *Central Institute for Supervisiting and Testing in Agriculture Division of Agrochemistry, Soil and Plant Nutrition, Hroznová 2, CZ656 06 Brno - Pisárky, Czech Republic*
[c] *Trilobyte Statistical Software Ltd., CZ530 02 Pardubice, Czech Republic*

Received 21 July 2004; accepted 28 January 2005

A new procedure of statistical analysis, with exploratory data diagnostics and Box–Cox transformation was used.

Abstract

To define the soil properties for a given area or country including the level of pollution, soil survey and inventory programs are essential tools. Soil data transformations enable the expression of the original data on a new scale, more suitable for data analysis. In the computer-aided interactive analysis of large data files of soil characteristics containing outliers, the diagnostic plots of the exploratory data analysis (EDA) often find that the sample distribution is systematically skewed or reject sample homogeneity. Under such circumstances the original data should be transformed. The Box–Cox transformation improves sample symmetry and stabilizes spread. The logarithmic plot of a profile likelihood function enables the optimum transformation parameter to be found. Here, a proposed procedure for data transformation in univariate data analysis is illustrated on a determination of cadmium content in the plough zone of agricultural soils. A typical soil pollution survey concerns the determination of the elements Be (16 544 values available), Cd (40 317 values), Co (22 176 values), Cr (40 318 values), Hg (32 344 values), Ni (34 989 values), Pb (40 344 values), V (20 373 values) and Zn (36 123 values) in large samples.

Keywords: Data transformation; Exploratory analysis; Soil pollution; Risk element contents

1. Introduction

High concentrations of some (specifically heavy) metals in soils can cause long-term harm to ecosystems and humans. As society becomes increasingly concerned about the hazard posed by polluted soil it wants to know how much of any pollutant there is in the ground. Soil survey, monitoring and inventarization programs are the inevitable tools used to define soil properties for a given area or country, including its pollution status. The Register of Contaminated Sites is one such program, established as part of the Fertilisers Act (Act No. 156/98 S.B., as amended) and connected decrees in the Czech Republic. Within the framework of the Register, a survey of the risk element (Cd, Pb, Cr, Hg) content of agricultural soils on a 1-km^2 grid was

* Corresponding author. Tel.: +42 40 603 7026; fax: +42 40 603 7068.
E-mail addresses: milan.meloun@upce.cz (M. Meloun), pavel.nemec@ukzuz.cz (P. Němec), kupka@trilobyte.cz (K. Kupka).
URL: http://meloun.upce.cz

0269-7491/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envpol.2005.01.027

implemented from 1990 to 1993. The four elements were successively complemented by analyses of Be, Co, Ni, V and Zn. This survey established a database which has since been continuously filled out by the results of supplementary sampling. Each sample in the batch is identified by geographical co-ordinates and the number of the plot in the relevant agricultural enterprise. The results of risk element content quantification in 2 M nitric acid extract or aqua regia extract are related. A batch of over 40 000 soil samples has been analysed for the Register database, and the exact sample sizes for each element (2 M HNO_3 extraction) are Be: 16 544 values available, Cd: 40 317 values, Co: 22 176 values, Cr: 40 318 values, Hg: 32 344 values, Ni: 34 989 values, Pb: 40 344 values, V: 20 373 values and Zn: 36 123 values.

When an exploratory data analysis (Tukey, 1977; Chambers et al., 1983) indicates that the sample distribution strongly differs from a normal one, the problem arises as to how to analyze the soil sample data. Raw data may require re-expression to produce an informative display, effective summary, or straightforward analysis (Meloun et al., 1992). Difficulties may arise because the raw data have (i) a strong asymmetry, or (ii) batches at different levels with a widely differing spread. By altering the shape of the batch or batches these problems may be alleviated. The data are transformed by applying a single mathematical function to all of the raw data values. It is not only the units in which the data are stated that may need to be changed but also the basic scale of the measurement. Changes of origin and scale mean linear transformations, and these leave shape alone; nonlinear transformations such as the logarithm and square root are necessary to change shape. Changing the scale of measurement is natural because it provides an alternative means of reporting the information. Batch symmetry is often a desirable property, as many estimates of location work best, and are best understood, when the data come from a symmetric distribution. The Box–Cox transformation eliminates heteroscedasticity, and the reconstructed mean and standard deviation are the estimates for the corrected distribution of the data. But heteroscedasticity has nothing in common with the outliers, objects which do not follow the distribution of data majority. The geochemical and pedological data sometimes arrive in several batches at different levels and a systematic relationship between spread and level is often found: increasing level usually brings increasing spread. When this relationship is strong there are several reasons for transforming the data in a way that reduces or eliminates the dependence spread on level: the transformed data will be better suited for comparison and visual exploration, and may be better suited for common confirmatory techniques, while individual batches become more nearly symmetric and have fewer outliers.

This paper provides a description of the Box–Cox transformation and a re-expression of statistics for transformed data. The procedure of Box–Cox transformation is illustrated on a typical soil pollution survey case study concerning a determination of cadmium content (mg kg^{-1}) and other elements (Be, Cd, Co, Cr, Hg, Ni, Pb, V and Zn).

2. Materials and methods

2.1. Sampling

The batches of soil samples for risk element analyses were taken from agricultural areas across the whole Czech Republic using the following sampling technique: one sample from arable land and grassland represents an area of 7–10 ha, from hop gardens and orchards 3 ha and from vineyards 2 ha. One composite sample consists of 30 individual probes to depths of 30 cm or 15 cm on arable land and grassland, respectively. The spatial distribution of the composite samples is arranged so as to achieve approximately 1 soil composite sample per 100 ha of agricultural soil. The following parameters were measured in the soil samples: soil texture (determined according to the maps of the Complex Soil Survey in the categories light, medium and heavy), pH exchangeable in KCl extraction, As, Be, Cd, Co, Cr, Cu, Mo, Ni, Pb, V, Zn determined by the AAS or ICP method in 2 M HNO_3 extraction, and total Hg content. Samples from different areas were analysed for the selected range of elements. The detection limit $\hat{x}_D$ [mg kg^{-1}] and quantification limit $\hat{x}_Q$ [mg kg^{-1}] for the quantitative determination of elements are, respectively, for As 1.307 and 4.619, for Be 0.06 and 0.197, for Cd 0.061 and 0.196, for Co 0.658 and 2.203, for Cr 0.598 and 2.179, for Cu 0.515 and 1.821, for Hg 0.02 and 0.06, for Mo 0.1 and 0.297, for Ni 0.574 and 1.992, for Pb 0.957 and 2.916, for V 1.611 and 5.764, and for Zn 1.37 and 3.979.

2.2. Proposed procedure of statistical data treatment

Step 1 *Survey of descriptive statistics*: the statistical software for an actual sample batch usually calculates a survey of parameters of location and spread.

Step 2 *Basic diagnostic plots in the exploratory data analysis* for a graphical visualization of data, diagrams and simple plots are used i.e.
- (a) the dot diagram and the jitter dot diagram,
- (b) the box-and-whisker plot and the notched box-and-whisker plot,
- (c) the quantile plot, and

(d) the symmetry plot (cf. Meloun et al., 1992, pp. 45–67).

Step 3 *Determination of sample distribution*: the sample distribution represented by symmetry, skewness and kurtosis is examined by

(e) the kernel density estimator of the probability density function

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left[\frac{x-x_i}{h}\right], \tag{1a}$$

where h is bandwidth, which controls the smoothness of $\hat{f}(x)$, and $K(x)$ is the kernel function, which is symmetric around zero, and also has the properties of a frequency function. The actual choice of shape for the kernel function is not important, so here a bi-quadratic kernel estimate is used

$$K(x)=\begin{cases}0.9375(1-x^2)^2 & \text{for } -1\le x\le 1\\ 0 & \text{for } x \text{ outside } [-1;\ 1]\end{cases} \tag{1b}$$

The quality of the kernel estimate $\hat{f}(x)$ is controlled mainly by the selection of parameter h. If h is too small, the estimate is too rough; if it is too large, the shape of $\hat{f}(x)$ is flattened too much (cf. Meloun et al., 1992, pp. 58).

(f) the quantile–quantile plot, which is used for comparison of the actual with the theoretical sample distribution.

Step 4 *Tests of basic assumptions about the sample* (cf. Meloun et al., 1992, pp. 78–82): applying an analysis of basic assumptions about the data, the following examinations were applied:

(a) Examination for the independence of sample elements,

(b) Examination for the normality of the sample distribution,

(c) Examination of sample homogeneity. All of the algorithms used are available on the internet (Kupka, 2004).

Step 5 *Data transformation*: when most of the EDA diagnostic plots exhibit an asymmetric distribution in the original sample data, data transformation seems to be the most convenient of several possible techniques to apply. For the transformation the estimate λ maximizing $\ln L(\lambda)$ is calculated. The selected λ is used in the calculation of estimates $\bar{y}$, $s^2(y)$, $\hat{g}_1(y)$, and $\hat{g}_2(y)$. From these estimates, the re-expressed estimates of the original variables $\bar{x}_R$, $s^2(\bar{x}_R)$, and the 95% confidence interval of the re-expressed variable μ are then calculated (cf. Meloun et al., 1992, pp. 70–77).

2.3. Measurement of location using data transformation

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

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 $\sigma^2(x)=f_1(\mu_x)$ where μ_x is the population mean of the original data, the variance $\sigma^2(y)$ may be expressed by

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

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

$$g(x)\approx C\int\frac{\mathrm{d}x}{\sqrt{f_1(x)}}.$$

When the dependence $\sigma^2(x)=f_1(x)$ is of a power (exponent) 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. Transformation leading to approximate normality may be carried out by the use of the *Box–Cox transformation family* (Box and Cox, 1964) defined as

$$y=g(x)=\left\{\begin{array}{ll}\left(|x|^{\lambda}-1\right)/\lambda & \text{for parameter } \lambda\neq 0\\ \ln|x| & \text{for parameter } \lambda=0\end{array}\right\}, \tag{2}$$

where x is a positive variable and λ is a real number.

The Box–Cox transformation can be applied only to 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 $x_0<x_{(1)}$. To estimate the parameter λ in a Box–Cox transformation the method of profile likelihood may be used, because for $\lambda=\hat{\lambda}$ the distribution of the transformed variable y is considered to be normal, $N(\mu_y, \sigma^2(y))$. The logarithm of the profile likelihood function may be written as

$$\ln L(\lambda)=-\frac{n}{2}\ln s^2(y)+(\lambda-1)\sum_{i=1}^{n}\ln x_i, \tag{3}$$

where $s^2(y)$ is the sample variance of the transformed data y (Box and Cox, 1964). The function $\ln L=f(\lambda)$ is

expressed graphically for a suitable interval, for example, $-3 \le \lambda \le 3$. 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 $2\left[\ln L(\hat{\lambda}) - \ln L(\lambda)\right] \le \chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This interval contains all λ values for which it is true that $\ln L(\lambda) \ge \ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1)$. This Box–Cox transformation is less suitable if the confidence interval for λ is too wide – and if the sample size is small then the confidence interval for the parameter will be wide. When the value $\lambda = 1$ is also covered by this confidence interval, the transformation is not efficient.

2.4. Re-expression of the statistical measurements after data transformation

After an appropriate transformation of the original data $\{x\}$ has been found, such that the transformed data give an approximately normal symmetrical distribution with constant variance, the statistical measurements 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 $\bar{y} \pm t_{1-\alpha/2}(n-1)s(y)/\sqrt{n}$. These estimates must then be recalculated for the original data $\{x\}$. Two different approaches to the re-expression of the statistics for transformed data can be used without difficulty:

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

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^\lambda}{n}\right]^{1/\lambda}, \tag{4}$$

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.

(b) 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 by

$$\bar{x}_R \approx g^{-1}\left\{\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right\}. \tag{5}$$

The variance is then expressed as

$$s^2(\bar{x}_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y),$$

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

$$\bar{x}_R - I_L \le \mu \le \bar{x}_R + I_U, \tag{6}$$

where

$$I_L = g^{-1}\left[\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right], \tag{7}$$

$$I_U = g^{-1}\left[\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right], \tag{8}$$

and

$$G = -\frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y). \tag{9}$$

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

$$\bar{x}_R \approx \exp\left[\bar{y} + 0.5\, s^2(y)\right], \tag{10}$$

and

$$s^2(\bar{x}_R) \approx \bar{x}_R^2 s^2(y). \tag{11}$$

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

$$\bar{x}_{R,1,2} = \left[0.5(1+\lambda\bar{y}) \pm 0.5 \times \sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2\left(\bar{y}^2 - 2s^2(y)\right)}\right]^{1/\lambda}, \tag{12}$$

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

$$s^2(x) = \bar{x}_R^{(-2\lambda+2)} s^2(y). \tag{13}$$

3. Results

Many statistical programs offer a list of the estimates of various point parameters of location and spread, but they rarely help the user to choose the

statistically adequate parameter for an actual sample batch. Exploratory data analysis and an examination of sample assumptions will find an answer to this question. The first case study with this methodology runs on typical geochemical sample data and will illustrate a rigorous procedure of the statistical treatment of univariate data with exploratory data analysis.

Properly processed analytical data can be used in research, government and legislation. For example, (a) results may serve as a national database characterising the degree of pollution of agricultural soils; (b) the appropriate parts of such a database may be distributed to local offices, (environmental sections) to be available to the regional and local government (e.g. in urban planning, privatisation projects, changes in land use, the application of sewage sludge or sediment to agricultural soil); (c) results may be used in the process of constructing legislative measures concerning the limit values of harmful substances in soil; and (d) a database can serve as one source for calculating critical loads and balances of risk elements in agro-ecosystems.

(1) *Survey of descriptive statistics*: ADSTAT statistical software calculates an actual sample batch, a survey of parameters of location and spread for $n = 40\,317$ (for an elucidation of the statistics cf. Meloun et al., 1992). On the basis of EDA, the user should select the most convenient parameter of location from the following available estimates: the arithmetic mean $\bar{x} = 0.238 \pm 0.003\ \mathrm{mg\,kg^{-1}}$, the median $\hat{x}_{0.5} = 0.19\ \mathrm{mg\,kg^{-1}}$ and the following trimmed means $\bar{x}(10\%) = 0.210 \pm 0.001\ \mathrm{mg\,kg^{-1}}$, $\bar{x}(20\%) = 0.203 \pm 0.001$, $\bar{x}(40\%) = 0.194 \pm 0.001\ \mathrm{mg\,kg^{-1}}$ (calculated with ADSTAT, TriloByte Statistical Software, Pardubice, Czech Republic), the standard deviation $s = 0.300\ \mathrm{mg\,kg^{-1}}$, and the parameters of shape are the skewness $\hat{g}_1 = 30.7$ and the kurtosis $\hat{g}_2 = 2123$, proving strongly skewed asymmetric distribution with a very sharp peak.

(2) *Basic diagnostic plots in the EDA* are used for a graphical visualization of the data: in Fig. 1 all of the exploratory diagnostic graphs prove a strong deviation from a normal distribution. The box-and-whisker plot (Fig. 1a) indicates many outliers at high values, and the quantile plot (Fig. 1b) an asymmetric, skewed distribution.

(3) *Determination of sample distribution in the EDA*: the sample distribution represented by symmetry, skewness and kurtosis is examined by the kernel density estimator of the probability density function (Fig. 1c). This plot shows that most sample points are located in one class and the plot indicates a strongly skewed sample distribution. The normal probability plot (Fig. 1d) checking a normal distribution does not exhibit close agreement of the sample points with a straight line.

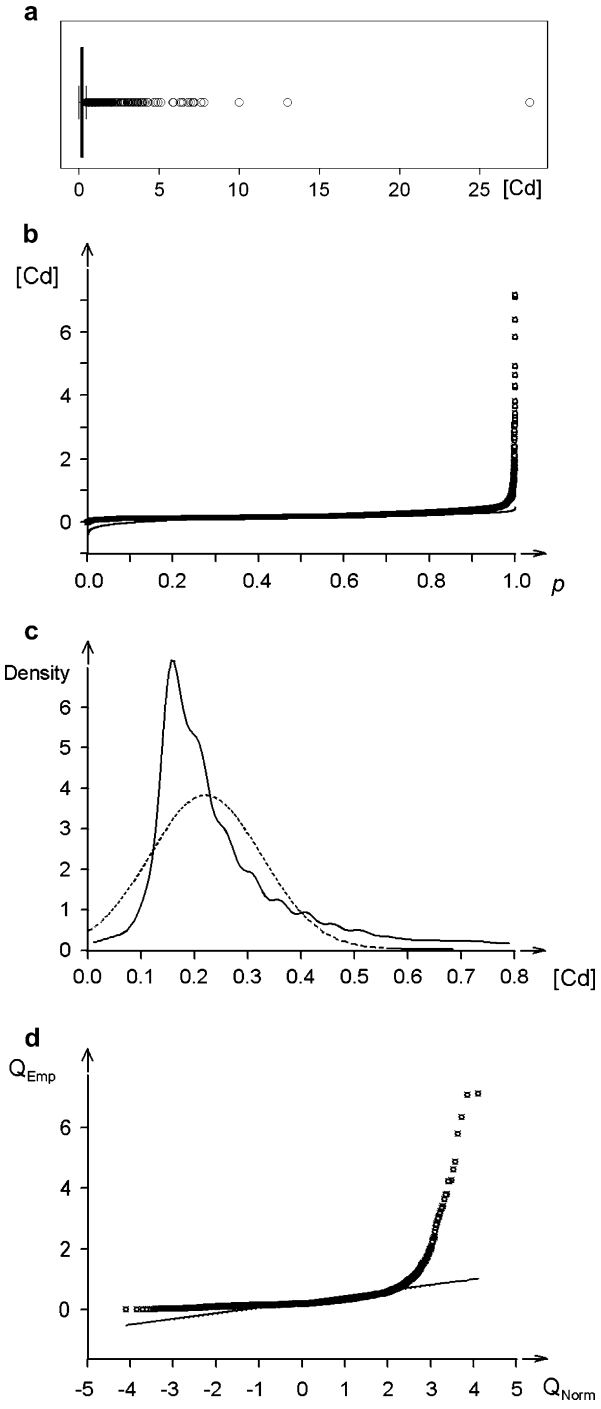

Fig. 1. (a) The box-and-whisker plot of the Cd sample data. (b) The quantile plot of the Cd sample data. (c) The kernel density estimator of the probability density function of the Cd sample data. (d) The quantile–quantile plot (for normal distribution called the Rankit plot) of the Cd sample data.

(4) *Basic assumptions about the sample* (cf. Meloun et al., 1992, pp. 78–82): applying an analysis of basic

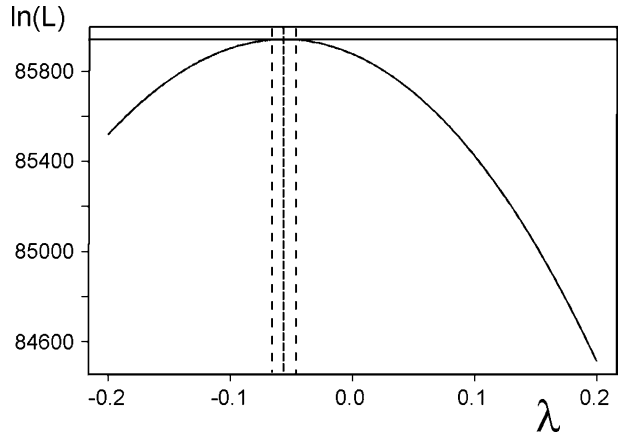

Fig. 2. The plot of the logarithm of the maximum likelihood for the Cd sample data with Box–Cox transformation.

assumptions about the data the following conclusions were derived: sample elements are independent and homogeneous. A combined sample skewness and kurtosis test leads to the test statistic

$$C_1 = \frac{\hat{g}_1^2(x)}{s^2(\hat{g}_1(x))} + \frac{[\hat{g}_2(x) - 3]^2}{s^2(\hat{g}_2(x))},$$

is $291.06 > \chi^2(0.95,\ 2) = 5.992$ and therefore normality of data distribution was rejected.

(5) *Data transformation*: as most diagnostic plots of the EDA exhibit an asymmetric distribution of original sample data, data transformation is a convenient technique to employ. In the case of the Box–Cox transformation the true mean value of a sample distribution with both confidence limits I_L and I_U is calculated. From the plot of the logarithm of the likelihood function for the Box–Cox transformation (Fig. 2) the maximum of the curve is at $\lambda = -0.0556$.

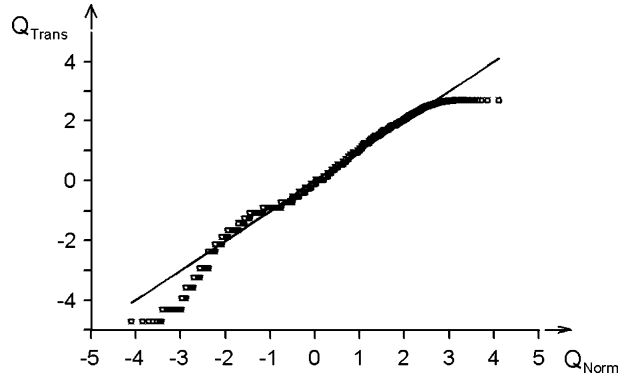

Fig. 3. The quantile–quantile plot for the Cd sample data with Box–Cox transformation.

The corresponding 95% confidence interval does not contain the exponent value $\lambda = 1$, so all transformations are statistically significant. The normal probability plot (also called the Rankit plot) on Fig. 3 shows that the Box–Cox transformation brings more accurate results.

While classical measures of location, spread and shape for the original data are $\bar{x} = 0.238\ \mathrm{mg\,kg^{-1}}$, $s(x) = 0.300\ \mathrm{mg\,kg^{-1}}$, the skewness $\hat{g}_1(x) = 30.74$ and curtosis $\hat{g}_2(x) = 2123.04$ are out of statistical significance and may be taken as false estimates of location. The Box–Cox transformation ($\hat{\lambda} = -0.0556$) calculates the corrected mean $\bar{x}_R = 0.187 \pm 0.001\ \mathrm{mol\,dm^{-3}}$.

(6) *Conclusion*: all EDA display techniques prove that the sample distribution is skewed with many outliers, and does not come from a population with a normal distribution. For the best estimate of a location parameter the arithmetic mean does not represent an objective measure of location, $0.238 \pm 0.003\ \mathrm{mg\,kg^{-1}}$, and cannot be used. On the basis of the quantile–quantile plot the Box–Cox transformation is considered the most rigorous technique to estimate a measure of location, with the corrected mean value $\bar{x}_R = 0.187 \pm 0.001\ \mathrm{mol\,dm^{-3}}$.

4. Conclusions

Often, chemical data are less than ideal and do not fulfill all basic assumptions. Original data can be transformed to improve the symmetry of data distribution and variance stabilization. Statistical measures of the transformed data are re-transformed to get these rigorous measures for the original data. Table 1 shows a survey of summary statistics for the elements beryllium, cadmium, cobalt, chromium, mercury, nickel, lead, vanadium and zinc. This survey includes classical and robust measures of central tendency, measures of variability, and measures of shape. Of particular interest here are sample size, the minimum and maximum values in a large sample, and both quartiles. The classical measures $\bar{x}$ and s are strongly corrupted with outliers and cannot be used here while the robust measures seem to be more accurate. Since the exploratory data analysis, the skewness and kurtosis and the quantile measures of location prove that the sample distribution strongly differs from a normal one, the data should be examined to find the proper transformation leading to symmetric distribution, stabilizing variance and making the distribution closer to normal. The most rigorous estimate of location is represented by the re-transformed mean $\bar{x}_R$ after Box–Cox transformation of original data. This

Table 1
Survey of summary statistics for the elements Be, Cd, Co, Cr, Hg, Ni, Pb, V and Zn including classical and robust measures of central tendency, measures of variability, and measures of shape

Estimate of	Beryllium	Cadmium	Cobalt	Chromium	Mercury	Nickel	Lead	Vanadium	Zinc
Sample size n	16 544	40 317	22 176	40 318	32 344	34 989	40 344	20 373	36 123
Minimum x_1	0	0	0.2	0.1	0	0.1	0.17	0.37	0.7
Maximum x_n	9.33	28.1	110.5	1577.4	69.086	662.0	1121.0	86.0	2070.0
Lower quartile F_L	0.32	0.14	3.9	3.2	0.06	3.0	11.7	7.0	12.0
Upper quartile F_U	0.57	0.27	6.7	6.9	0.11	7.3	19.4	13.0	22.0
Interquartile range $F_U - F_L$	0.25	0.13	2.8	3.7	0.05	4.3	7.7	6.0	10.0
Classical estimates of location, scale and shape									
Sample mean $\bar{x}$	0.470 ± 0.004	0.238 ± 0.003	5.593 ± 0.039	7.104 ± 0.170	0.105 ± 0.006	6.033 ± 0.081	18.637 ± 0.299	10.878 ± 0.083	19.354 ± 0.234
Standard deviation s	0.264	0.300	2.930	17.35	0.534	7.728	30.594	6.015	22.73
Skewness $\hat{g}_1$	5.99	30.74	4.19	40.09	107.88	34.49	19.77	2.16	34.20
Kurtosis $\hat{g}_2$	119	2123.1	89.85	2608.52	12 963.7	2298.8	528.2	12.41	2265.0
Robust estimates of location									
Median $\tilde{x}_{0.5}$	0.43 ± 0.01	0.19 ± 0.00	5.0 ± 0.0	4.60 ± 0.05	0.08 ± 0.00	4.70 ± 0.05	14.90 ± 0.05	9.60 ± 0.10	16.0 ± 0.05
Trimmed mean $\bar{x}(10\%)$	0.449 ± 0.003	0.210 ± 0.001	5.356 ± 0.033	5.361 ± 0.040	0.086 ± 0.001	5.320 ± 0.039	15.860 ± 0.067	10.320 ± 0.074	17.446 ± 0.089
Trimmed mean $\bar{x}(20\%)$	0.443 ± 0.003	0.203 ± 0.001	5.264 ± 0.032	5.072 ± 0.033	0.083 ± 0.001	5.109 ± 0.037	15.548 ± 0.063	10.050 ± 0.072	17.020 ± 0.085
Trimmed mean $\bar{x}(40\%)$	0.438 ± 0.003	0.194 ± 0.001	5.150 ± 0.030	4.795 ± 0.030	0.081 ± 0.001	4.883 ± 0.036	15.214 ± 0.061	9.752 ± 0.067	16.531 ± 0.084
Jarque–Berra normality test, critical value for $\alpha = 0.05$ is $\chi^2_{0.95}(2) = 5.99$									
Testing criterion C_1	157.1	291.1	145.9	311.1	382.0	294.3	259.3	111.4	294.9
Normality is	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected	rejected
Homogeneity test									
Number of outliers	265	2095	496	2285	1180	1128	1359	486	961
Box–Cox transformation									
Re-transformed mean $\bar{x}_R$	0.427 ± 0.003	0.187 ± 0.001	5.078 ± 0.030	4.922 ± 0.023	0.082 ± 0.001	4.797 ± 0.020	15.172 ± 0.050	9.611 ± 0.050	16.360 ± 0.050

Of particular interest here are sample size, minimum and maximum values within the sample, and both quartiles. The most rigorous estimates of location are re-transformed means after Box–Cox transformation.

estimate can be taken as the best for each element studied.

Acknowledgements

The financial support of the Grant Agency of the Ministry of Education of the Czech Republic (Grant No MSM0021627502) is gratefully acknowledged.

References

Box, G.E.P., Cox, D.R., 1964. Journal of Royal Statistical Association B26, 211–252.

Chambers, J., Cleveland, W., Kleiner, W., Tukey, P., 1983. Graphical Methods for Data Analysis. Duxbury Press, Boston.

Kupka, K., 2004. <http://www.trilobyte.cz/eda>.

Meloun, M., Militký, J., Forina, M., 1992. Chemometrics for analytical chemistry. In: PC-Aided Statistical Data Analysis, vol. 1. Ellis Horwood, Chichester.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison Wesley, Reading, MA.