PRELIMINARY DATA ANALYSIS

Werner Gurker and **Reinhard Viertl**
Vienna University of Technology, Wien, Austria

Keywords: Box-plots, data sets (uni-, bi-, multivariate), data transformations, exploratory data analysis (EDA), graphical representations, mean, measures of location, measures of spread, outliers, probability plots, quantiles. regression, variance

Contents

1. Univariate Data Sets
1.1 Graphical Displays
1.1.1 Frequencies
1.1.2 Cumulative Frequencies
1.2. Measures of Location
1.3. Depths
1.4. Measures of Spread
1.5 Outliers
1.6. Seven-point summaries
1.7 Box-Plots
1.8. Data Transformation
1.9. Probability Plots
1.9.1 P-P-Plots
1.9.2 Q-Q-Plots
2. Bivariate Data Sets
2.1 Graphical Displays
2.2 Numerical Characteristics
2.3 Regression
3. Multivariate Data Sets
3.1 Data Matrix and Summary Statistics
3.2 Data Transformations
3.3 Graphical Displays
4. Concluding Remarks
Glossary
Bibliography
Biographical Sketches

Summary

A careful and critical analysis of a data set is an essential step in the course of a statistical analysis. The final conclusions cannot be more reliable than the data set they rely on. This chapter deals with the case of given data sets; the important issue of planning and designing an investigation or experiment, however, is beyond the scope of the chapter. Revealing as much as possible about the structure of a data set, its limitations and peculiarities, is a good starting point for further statistical analysis and modeling. Note that frequently a careful preliminary data analysis is all that is needed, answering the questions posed. Several simple graphical devices are considered, various

measures of 'location', 'spread' and 'skewness' of the data set are introduced, some important data transformations discussed. In real situations frequently some irregularities such as 'outliers' occur, or the 'fuzziness' of the data may be too strong to be ignored. A detailed discussion of these crucial topics, however, is given in other chapters. Likewise, for the practically important case of multivariate data sets only some basic analytical and graphical devices are given; the two-dimensional case, however, is discussed in some detail.

1. Univariate Data Sets

1.1 Graphical Displays

1.1.1 Frequencies

Small numerical data sets can be conveniently presented in a *frequency table.* For a data set consisting of n values the distinct values a_j, $j = 1,...,k$, are tabulated with their *frequency* $H_n(a_j)$ or, more often, with their *relative frequency* $h_n(a_j) = H(a_j)/n$. The table can be represented by a *line graph,* plotting the distinct values against their frequencies, or by a *frequency polygon,* which connects the plotted points with straight lines. What is graphically more appealing is a *bar graph,* a line graph with thicker lines, centered around a_j.

Example 1.1 As an example we consider the distribution of families in Austria for two years (1991 and 1997) with respect to the number of their children. (*Source of data: Statistisches Jahrbuch für die Republik Österreich 1998*, Wien, 1998.)

Note that class '4'really is an open class which gathers all families with 4 or more children (Figure 1).

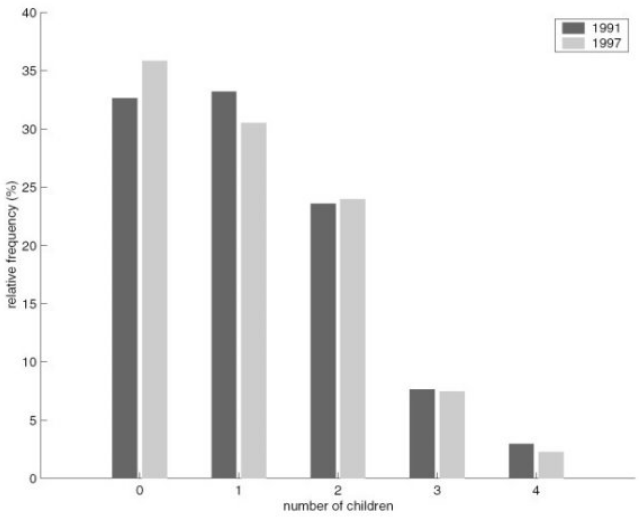

Figure 1: Families in Austria by the number of their children (Example 1.1)

For *categorical* data set a *pie chart* is often used, where the relative frequencies are represented by the areas of slices of a circular pie. Graphical analysis however is generally more important with measurement (numerical) data than with non-numerical (categorical) data.

For large (numerical) data sets, when the number of distinct values is large a useful procedure is to group the data into *class intervals* and to represent the (relative) *class frequencies* $H_j(h_j)$ with a bar graph. Displays like that are called *(frequency) histograms.* For cognitive reasons it is important to use the area of the bars, and not their heights as the representation criterion, especially when the lengths of the classes are different. Whenever possible, however, one should use (adjacent) intervals of equal length covering the data. To make things unique at the class boundaries we use the *left-end convention,* which means the intervals contain their left ends but not their right ends (other choices are possible of course).

The appropriate number of classes is largely a subjective choice, some guidelines are given by $\sqrt{n}$ or $\log_{10}(n)$, where n is the number of data values.

Example 1.2 As an example for histograms we consider the distribution of lightings in Austria for 1999 with regard to the polarity (negative or positive) of their charge (measured in kA). (*Source of data: ALDIS* - Austrian Lighting Detection & Information System; located in Vienna.)

Because there are so many (over 200.000 negative, and over 40.000 positive lightings per year) suitable classes have to be chosen (here intervals of equal length =1 kA were used). Negative flashes with more than 100 kA rarely are registered, whereas positive flashes with more than 100 kA and high as 200 kA can be observed rather frequently.

First we consider the histogram for negative flashes (Figure 2). The histogram for positive flashes (Figure 3) is quite similar in appearance, although it is a little more 'peaked' while having a more heavy upper tail.

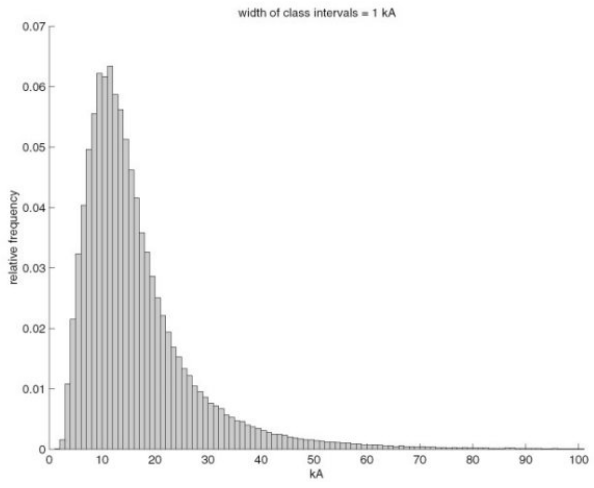

Figure 2: Negative lightings in Austria 1999(Example 1.2)

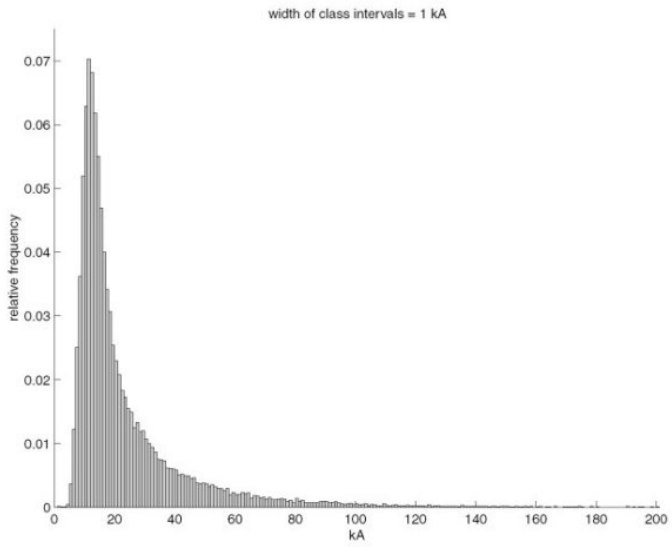

Figure 3: Positive lightings in Austria 1999 (Example 1.2)

Note that histograms constructed like that, using the area as representation criterion, are *densities*, meaning they integrate to one. This is a rather convenient feature if some (approximate) probability statements are to be made. Adopting this idea one can give a more 'smooth' density-representation of the data by the use of so-called *kernels*, being similar to the application of 'windows' in signal analysis. In principle, every nonnegative, bounded and symmetrical function $k(u)$ fulfilling

$$\int_{-\infty}^{\infty} k(u)\ du\ =\ 1, \qquad \int_{-\infty}^{\infty} k^2(u)\ du\ <\infty, \ \left|\frac{k(u)}{u}\right| \to 0 \text{ for } |u| \to \infty \tag{1}$$

could be used. Some frequently used kernels are

(1) Uniform kernel

$$k(u) = \begin{cases} 1 & |u| \ \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

(2) Epanechnikow kernel

$$k(u) = \begin{cases} \frac{3}{4}(1-u^2) & |u| \ \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

(3) Normal kernel

$$k(u) = \frac{1}{\sqrt{2\ \pi}} e^{-u^2/2}, \qquad -\infty < u < \infty \tag{4}$$

The *kernel estimator* of the density is given by

$$\hat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} k\left(\frac{x - x_i}{b}\right), \quad -\infty < x < \infty \tag{5}$$

where b is the *bandwidth* and $x_1, ..., x_n$ are the observed data. The appropriate choice of the bandwidth depends on the degree of 'smoothness' that is to be achieved (the larger the bandwidth the smoother the density).

In practice it is a good choice to try several b's. (Note that typical 'side lobe' effects may be present if the bandwidth is not chosen appropriately.)

Locally, $\hat{f}(x)$ is determined by the x_i's nearby x. The uniform kernel gives equal weight to all data values in the interval $[x - b/2,\ x + b/2]$, the Epanechnikow kernel uses data values in the interval $[x - b,\ x + b]$ but gives heavier weight to values near x, and the normal kernel uses all the data values giving heavier weight to values near x.

Example 1.3 Considering again the lightnings in Austria for 1999 (cf. Example 1.2) we chose in both cases a normal kernel and a rather large bandwidth of 2 kA.

This cuts off the peaks a little bit, but otherwise the fit seems to be very good, especially in the upper parts (Figure 4 and 5).

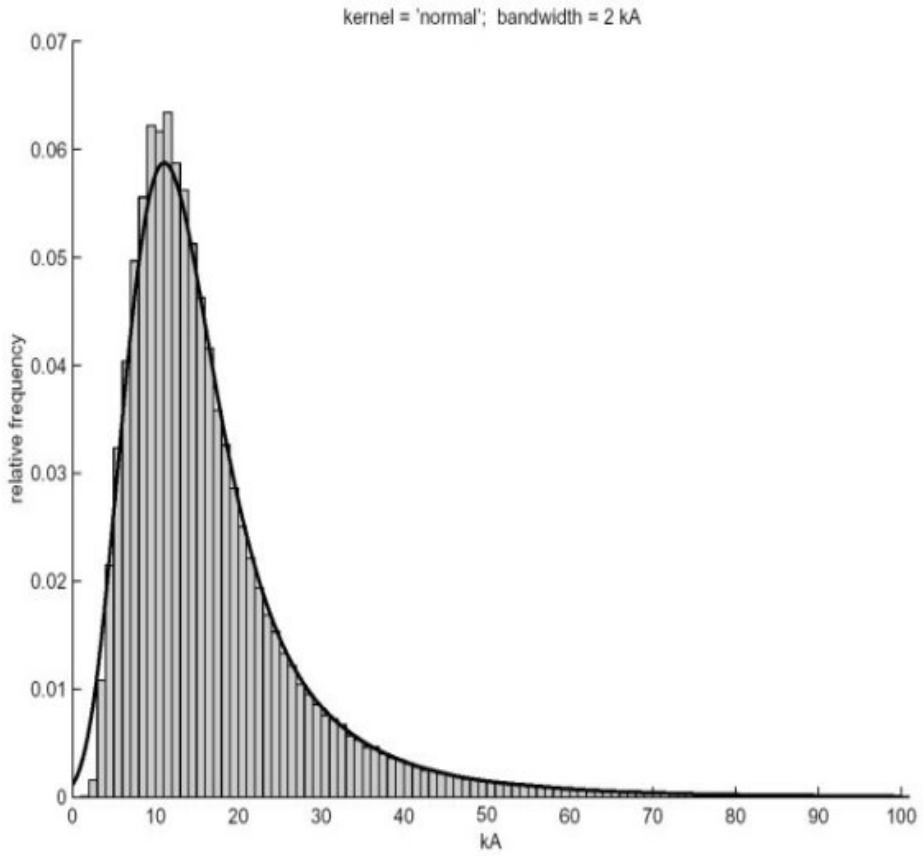

Figure 4: Negative lightnings in Austria 1999 (Example 1.3)

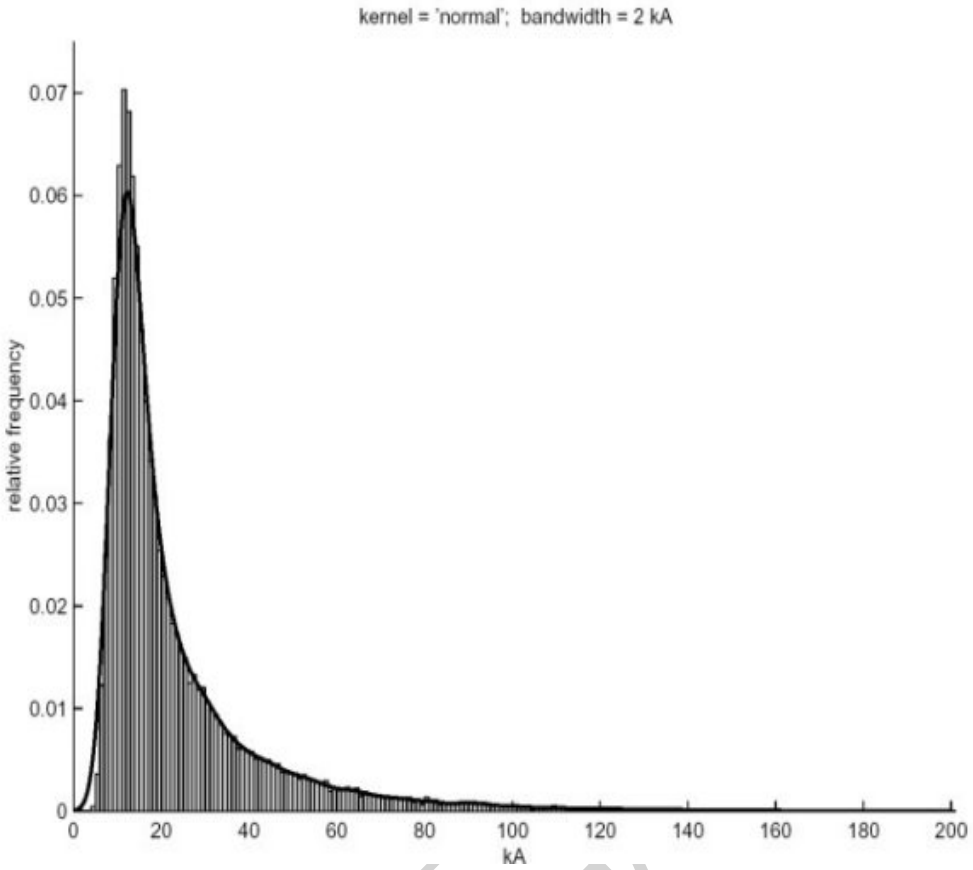

Figure 5: Positive lightnings in Austria 1999 (Example 1.3)

The display of a grouped frequency distribution by *gray intensity levels* can also be an appealing method, because the human eye is rather sensitive in detecting small differences in color or gray-levels. The method is even more appealing for two dimensions as an alternative to two-dimensional histograms, which are sometimes quite difficult to 'read' (cf. Section 2.1).

The darkness of the gray-level of a (rectangular) cell is set proportional to the (relative) class-frequency. Remove the cell boundaries and/or interpolate between the gray-levels of the cell-vertices to provide a more 'smooth' impression.

Example 1.4 As an example of gray-intensity strips we consider the distribution of nitrate in ground- and spring-water (measured in mg/liter) in Austria, for the whole territory and for the 'Länder' (Wien, Vorarlberg, Tirol, Steiermark, Salzburg, Oberösterreich, Niederösterreich, Kärnten, Burgenland). (*Source of data UBA* (Umweltbundesamt), Wien, 1998.)

For the correct interpretation of the gray intensities a 'color-bar' (in %) is set aside. '*a*%' means, that *a*% of the measurements made (in Austria or in one of the Länder) resulted in nitrate-concentrations falling into the particular class (see Figure 6).

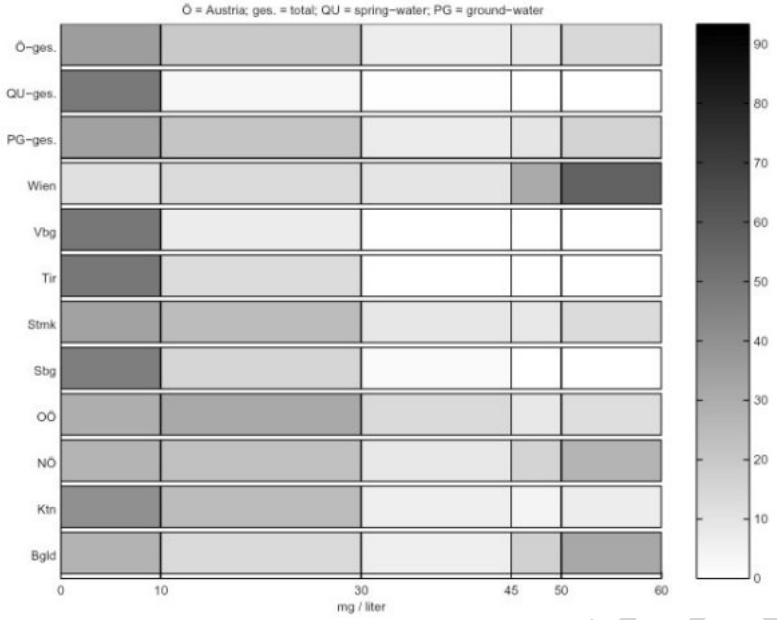

Figure 6: Nitrate in ground-and spring-water in Austria 1995-1997 (Example 1.4)

A quick and easy way to display small or moderate sized data sets is the *stem-and-leaf plot*, some sort of a special histogram, where the classes are defined by the data itself.
First each data value is split into two parts, the stem and its leaf. For example, if 548 is the observed life length in hours of some item, and the typical life length of those items is of the same order, one would probably take 5 as the stem and 48 as its leaf. This determines the *unit* of the plot, that is 5 is read as 500. As it is common (though not imperative) to take as leaf only one digit, 48 is cut off to 4 (cutting off is more common than rounding here) and the observation is written as 5 | 4. To obtain a more refined display the stems can be subdivided into several classes, usually 2 (double-stem) or 5 (five-stem). In the first case, in our example, 5^* gathers the life times from 500 to 549 hours and 5 the life times from 550 to 599 hours. The data value would be written as 5^* | 4.

In the second case, 5^* comprises the times from 500 to 519 hours, 5t(t = two, three) the times from 520 to 539, 5f(f = four, five) the times from 540 to 559, 5s(s = six, seven) the times from 560 to 579, and 5. the times from 580 to 599. The data value 548 would be written as 5f | 4. To make the computation of some parameters (like data quantiles) more easy, the (cumulated) *depths* of the data values can be written on the left side of the plot (see Section 1.3 for a more complete discussion of the 'depth' of a data value).

Example 15 As an illustration for stem-and-leaf plots we consider the life expectations (in years) of European and some non-European countries (Australia, Japan, Canada, New Zealand, Turkey, USA) for men and women at birth. (*Source of data: Statistisches Jahrbuch für die Republik Österreich 1998*, Wien, 1998.)
For the countries considered, Iceland, Sweden, Switzerland, and Japan have the highest expectations for men, and France, Switzerland, and Japan the highest for women. It has to be noted however, that the reference years for the different *life tables* are quite different (they vary between 1990 and 1997).

The 5-stem-and-leaf plots for men and women are as follows:

Depth		Men
1	5.	8
2	6*	1
7	6t	22233
9	6f	45
11	6s	67
16	6.	88999
20	7*	0001
(9)	7t	222233333
14	7f	4444445555
4	7s	6666

Depth		Women
2	6.	99
3	7*	1
5	7t	33
14	7f	444444555
19	7s	66777
(9)	7.	888999999
15	8*	000000001111
3	8t	222

In this case it is instructive to present the two plots *dos á dos* for a more direct comparison:

Men		Women
8	5.	
1	6*	
33222	6t	
54	6f	
76	6s	
99988	6.	99
1000	7*	1
333332222	7t	33
5555444444	7f	444444555
6666	7s	66777
	7.	888999999
	8*	000000001111
	8t	222

There are numerous other and well established ways to display frequencies, for example the *dot plot*, quite similar to the stem-and-leaf plot (with dots in place of digits), or the *rootogram*, some sort of a transformed histogram, where one tries (by a square root transformation) to achieve a more symmetrical (or even 'normal-shaped') histogram (see Section 1.8 for a discussion of data-transformations). This latter rootogram tries to achieve more than a simple preliminary graphical representation of the data (and it is more difficult to understand 'what is going on'). But one should be aware of the two main purposes of statistical graphs. Firstly, they can help *others* to better understand the structure of an already well understood data set; as such the presentation methods should be as simple as possible. Secondly, they can assist *you* in the *exploration* of the structure and peculiarities of a new data set. The latter purpose usually requires some more refined methods (for example in the detection of outliers), because seldom (or almost never) the collection of and the resulting data are 'clear cases'. For non-precise data a generalized histogram is obtained, for which the relative frequencies are 'fuzzy numbers'. (Cf. also *Statistical Inference with Non-Precise Data* for a discussion of 'fuzzy' data).

-
-
-

Bibliography

Ferschl, F. (1985): *Deskriptive Statistik*, 3rd. Ed., Würzburg: Physica Well written careful introduction into descriptive statistical methods with many examples (in German)]

Freedman, D., Pisani, R. and Purves, R. (1978): *Statistics*, New York: Norton. [Very well written, pedagogically oriented textbook on the more elementary aspects of statistics and probability, with many real life examples, still one of the best texts on the subject]

Hamilton, L.C. (1990):*Modern Data Analysis – A First Course in Applied Statistics*, Pacific Grove: Brooks/Cole. [Well written textbook on elementary statistics, carefully discussing various classical and some exploratory topics, with many examples]

Thiessen, H. (1997): *Measuring the Real World*, Chichester: Wiley. [An easy to read textbook on applied statistical methods with many examples]

Tukey, J.W. (1977): *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley. [Initializing textbook on various new and fresh approaches to statistical thinking, by one of the most influential statisticians of the last decades]

Viertl, R. (1996): Statistics with Non-precise Data, *Journal of Computing and Information Technology*, Vol. 4. [In this paper a generalized histogram based on non-precise data is explained]

Biographical sketches

Werner Gurker born in March 18, 1953, at Mauthen in Carinthia, Austria. Studies in engineering mathematics at the Technische Hochschule Wien. Receiving a Dipl.-Ing. degree in engineering mathematics in 1981. Dissertation in mathematics and Doctor of engineering science degree in 1988. Assistant professor at the Technische Hochschule Wien since 1995. Main interest and publications in statistical calibration and reliability theory.

Reinhard Viertl born March 25, 1946, at Hall in Tyrol, Austria. Studies in civil engineering and engineering mathematics at the Technische Hochschule Wien. Receiving a Dipl.-Ing. degree in engineering mathematics in 1972. Dissertation in mathematics and Doctor of engineering science degree in 1974. Appointed assistant at the Technische Hochschule Wien and promotion to University Docent in 1979. Research fellow and visiting lecturer at the University of California, Berkeley, from 1980 to 1981, and visiting Docent at the University of Klagenfurt, Austria in winter 1981 - 1982. Since 1982 full professor of applied statistics at the Department of Statistics, Vienna University of Technology. Visiting professor at the Department of Statistics, University of Innsbruck, Austria from 1991 to 1993. He is a fellow of the Royal Statistical Society, London, held the Max Kade fellowship in 1980, and is founder of the Austrian Bayes Society, member of the International Statistical Institute, president of the Austrian Statistical Society from 1987 to 1995. Invitation to membership in the New York Academy of Sciences in 1998. Author of the books *Statistical Methods in Accelerated Life Testing* (1988), *Introduction to Stochastics* in German language (1990), *Statistical Methods for Non-Precise Data* (1996). Editor of the books *Probability and Bayesian Statistics* (1987), *Contributions to Environmental Statistics* in German language (1992). Co-editor of a book titled *Mathematical and Statistical Methods in Artificial Intelligence* (1995), and co-editor of two special volumes of journals. Author of over 70 scientific papers in algebra,

probability theory, accelerated life testing, regional statistics, and statistics with non-precise data. Editor of the publication series of the Vienna University of Technology, member of the editorial board of scientific journals, organiser of different scientific conferences.