Proof of Pareto's 80/20 Law and Precise Limits for ABC-Analysis

Alfred Ultsch
Technical Report 2002/c
DataBionics Reseach Group
University of Marburg
35032 Marburg/Lahn
Germany
ultsch@informatik.uni-marburg.de

In many projects 20% of the total effort yields 80% of the total outcome. This phenomenon is usually termed Pareto's 80/20 law. In this paper we propose a theory to explain this empirical observation. The yield gained by the subdivision of a project into several tasks is measured. The requirements for such a yield lead to the axioms of Shannon Information. With the right adjustment of units for cost and yield this gives the definition of Entropic Yield. Pareto's 80/20 law thus results from an economic optimization of Entropic Yield in the form of minimizing unrealized potential. As an application of the theory we have derived precise limits for ABC-analysis. The outlined theory adds to Information Theory the consideration of production costs for information. Furthermore it sheds some light on the connection between the physical term Entropy and Shannon Information. The theory can be applied in statistical data analysis, e.g. cluster- and/or factor analysis, as well as in marketing, logistics or other business applications.

1. Pareto's 80/20-law

The Italian economist Vilfredo Pareto (1848-1923) observed in 19th century Italy that 20% of the population owned 80% of the usable land (Pareto 1935). Pareto found the same distribution in other economical and natural processes. As a general rule he formulated this finding as: "in any arbitrary set of elements, that try to achieve something a subset small in numbers will have the biggest effect" (Sombart 1967).

Nowadays many projects show the same 80/20 distribution of yield vs. costs. This is often attributed to Pareto's observation and is called Pareto's 80/20-law or the law of the trivial many and the critical few. Economic activities follow Pareto's 80/20-law quite frequently, in particular in marketing and quality management (Kimber et al. 1997), (Dyche 2001). Here a selection of the many occasion in which Pareto's 80/20-law is reported:

- 20 % of customers generate 80% of turnover
- 20% of products make 80% of turnover
- 20% of possibilities to make faults in production are responsible for 80% of product defects

- 80% of the decisions are made in 20% of the time in a meeting
- 20 % of products make 80% of profit
- 20 % of employees account for 80% of the time absent
- 80 % of results are achievable in 20% of working time if strategic time planning is used
- the best 80% of sellers are responsible for 80% of the profit of a firm
- 20% of the goods in a stock sum up to 80% of the stock worth
- 80% of the requests for stocked articles are on only 20% of the goods
- 80 % of the costs or losses of a business are caused by 20% of the problems

Pareto's 80/20-law is in particular observed in many software projects. The following table is taken from Arthur (Arthur 1992):

20 % of these		80 % of these
modules	consume	resources
modules	contribute	errors
modules	consume	execution time
errors	consume	repair costs
enhancements	consume	adaptive maintenance costs
tools	experience	tool usage

Pareto's 80/20-law is usually justified only by observation i.e. empirically (Arthur 1992). We do not know of any theory that would explain Pareto's 80/20-law. Since the 80/20 phenomena seems to be so ubiquitous the question is, whether there might be "law of nature" behind this observation. Pareto's 80/20-law is also used in the so called ABC-analysis used for the optimization of businesses and projects.

In the following we derive Pareto's 80/20-law from fundamental principles. We analyze the yield generated by the division of a project into subprojects. This elementary yield is measured in terms of information or entropy. Pareto's 80/20-law follows from a optimization of cost and yield of such a division. Using our theory on Pareto's 80/20-law we can deduct exact theoretical limits for the parameters of an ABC-analysis.

2. ABC-Analysis

When business processes are optimized the so called ABC-analysis is often applied (Gourdin 2001). This analysis is partly based on Pareto's 80/20-law and therefore often called Pareto-analysis (Juran (Ed.) 1998). ABC-analysis means to classify subprojects into three classes A, B, and C. Subprojects are ordered in decreasing order of yield. Class A should contain projects of high yield, class B projects of medium yield and C projects of low yield as follows:

Class A: subprojects with relative low costs returning an over proportional yield, i.e. the relatively few subprojects in this class should return a very high yield with low expense

Class B: subprojects with at least average ratio of yield to cost. Yields of projects in this class should be at least direct proportional to costs.

Class C: the rest of the subprojects, i.e. these subprojects generate low profit on high costs

Different rules of thumb are proposed for the borderlines in yield to determine the classes. Typical proposals for the limits of yield in class A range from 5% to 33%. Proposals for class B range from 15% to 33%, for class C from 25% to 50%. In many practical applications the borderlines are manually selected using so called Lorenz curves (Juran 1950). On the x-axis of Lorenz curves are the cumulated costs, on the y-axis the cumulated yields of the projects. The following Lorenz curve is generated from data in a basic textbook on statistics (Hartung 1997). It shows size of the land owned by farms in South Africa vs. the number of farms. Both axis units are in percent.

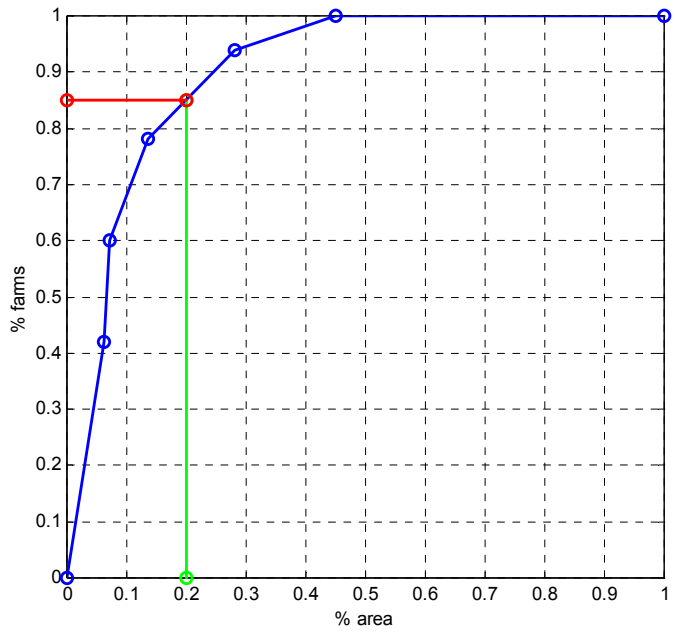

Figure 1 Lorenz curve of land area usage in South Africa

In the sequel we describe a derivation of boundaries derived from our theory on Pareto's 80/20-law.

3. Dividing Projects into Tasks

The set of steps necessary to solve a problem is called a project. The following is a very general approach to solve any kind of problem: first, divide the problem into a set of sub problems or tasks. Second, solve each sub problem separately and third, combine the partial solutions and the problem is solved. This approach to problem solving is known as the „Divide and Conquer" approach with reference to the Italian politician Macciavelli who is supposed to have recommended this principle as guideline for state leaders (Aho et al. 1982).

In order to measure the proportion of effort spent in a project we assume that a report is filed on all the subprojects. For each subproject a unique symbol, e.g. sign such as a letter or number, is used to indicate that a certain portion of cost is spent on that project. At regular intervals during a project these symbols may be written into a report. After the completion of a project, the number of different symbols are proportional to the costs spent for each subproject. The overall cost of the project is proportional to the total number of

letters in this type of report. The frequency of occurrence of the task's symbol in the report is proportional to the effort (cost) spent for this task.

4. Entropic Yield

Entropic Yield is a measure for the yield generated by the division of a project into tasks. Assume a problem can be divided in **n** tasks. Then there are **n** different symbols present in the project report. Each of the symbols occurring with frequencies p_1, ...$p_{\mathbf{n}}$ in the report. If only one subproject takes up all resources, it's symbol will be the only one present in the project's report. Nothing is gained by this extreme "division". The Entropic Yield of this division is set to zero. If a task's symbol is not contained in the project report, this means no effort is spent on this task. The Entropic Yield of this task should also be zero. If the project is divided such that the cost for each subproject is equal the total Entropic Yield should be maximal.

Shannon Information as defined in Information Theory measures the information of a message consisting of distinct signs (Shannon 1948). The information contained in a project report is calculated as the sum of the information for each task's sign. The information of a task's sign occurring with frequency p is calculated with the following formula:

$$\inf(p) = -c \cdot p \cdot \log_b(p) \,. \quad (1)$$

Where b is the base of the logarithm and c is a positive constant. In Information Theory c and b are adjusted such that the values of formula (1) are measured in bits. The same formula is used in physics to denote entropy (Boltzmann 2000). The formula for entropy can be derived from axiomatic principles which are consistent with an intuitive understanding of yield (see[10]).

For our purposes it is important, that cost and yield are measured in the same units: in percentages. We measure cost spent in a subproject as fraction of the total cost i.e. in the range 0% to 100%. Using elementary calculus it can be shown that the Shannon Information of a task is maximal at $p_{\max} = \frac{1}{e}$. With e = 2.1718..., the Euler number, this gives the maximal possible Shannon Information at a frequency of p_{max} = 37%. Thus we define Entropic Yield as $EY(p) = -e \cdot p \cdot \ln(p)$. Where e denotes the Euler number and ln the natural logarithm. Entropic Yield for $p = p_{max}$ becomes 100%. This means cost and yield are measured in the same range from 0% to 100%. Figure 2 depicts a graph of the Entropic Yield of a task as function of the cost.

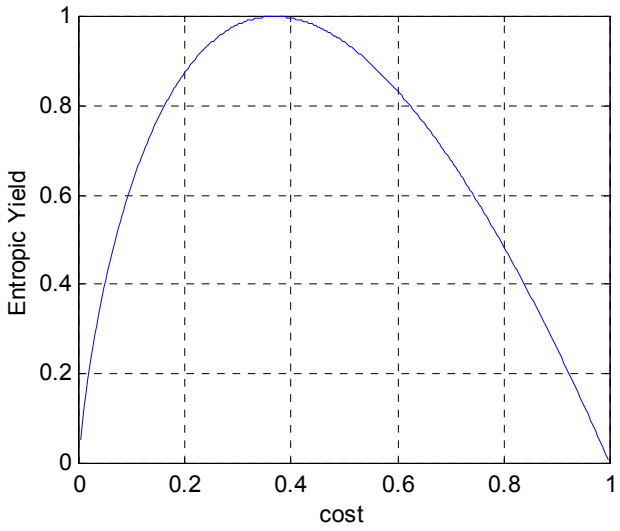

Figure 2 Entropic Yield of a subproject as function of the cost .

As can be seen in figure 2 the Entropic Yield of a subproject grows at first when more is spent on a subproject. For costs greater than 37% of the total cost the yield gained by the division into subprojects decreases. Too much effort goes into such a subproject. If all the effort is spent in one subproject there is no yield generated by the division.

5. Unrealized Potential

Economical consideration calls for projects producing a maximum of yield with costs as low as possible. The ideal situation of a project would be to produce 100% of yield with 0% of costs. Although this optimal situation O can hardly be found in practical projects it may serve as a starting point to measure how good a given cost/yield situation S of a project is. We define the unrealized potential URP of a project's situation as the Euclidian distance from the ideal situation O to a given project's situation S.

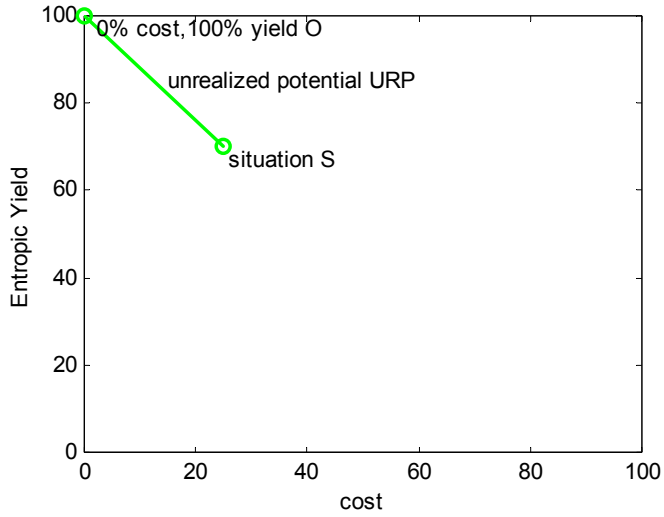

Figure 3 Unrealized Potential of a project's situation

Using Euclidian distance has the advantage that the measuring of unrealized potential is independent of translation and rotation of the coordinate system in which URP is measured. Furthermore path integration is possible for Euclidian distance. This means it is a distance

metric consistent with common sense measuring of distances in space. These properties brought Einstein to use Euclidian distance for the combination of space and time. Mathematically the unrealized potential URP of a project's situation. Thus URP is defined as the Euclidian distance from point O to a given project's situation S. The values of URP for all tasks with an Entropic Yield function can be seen in the following figure:

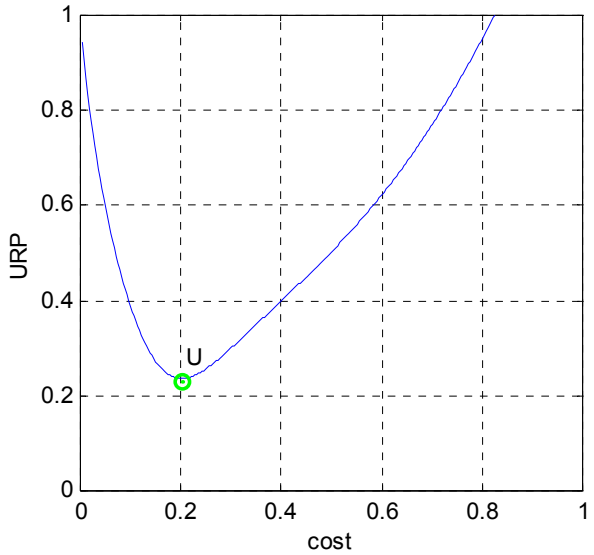

Figure 4 Unrealized potential of tasks with Entropic Yield

6. Pareto's 80/20 law

The central question is why in many projects there is one subproject that produces about 80% yield with 20% of the total costs. Consider the situation with the smallest unrealized potential, i.e. the task with the smallest distance to the ideal situation U. From figure 3 we can see that the cost for this situation is abut 20%, more precisely it is 20.13%. Measuring the Entropic Yield of this situation results in a yield of 87.77%. as can be seen in the following figure.

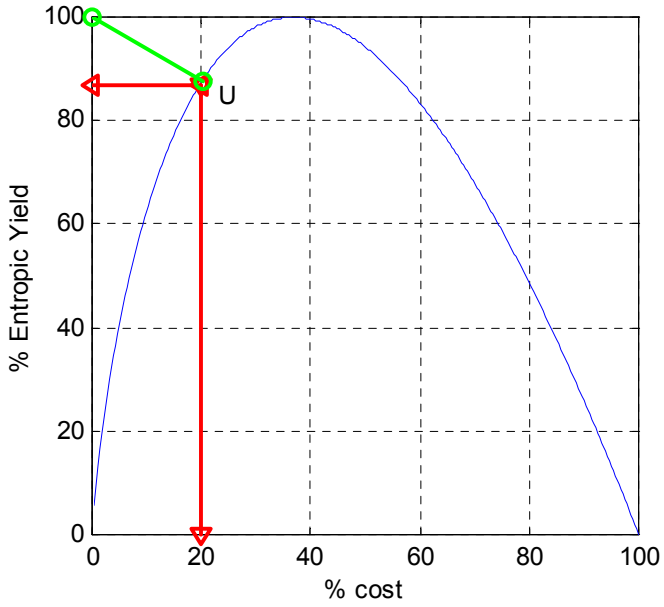

Figure 5 Project situation with best realized potential

Within error margins of 10% on yield the situation U with cost/yield at about (20%, 88%) may be regarded as the situation which is encountered so frequently and lead to the formulation of Pareto's law. An economical optimization of the information or entropy generated by the division of a project into subtasks functions as a theoretical foundation of Pareto's 80/20 law.

7. Properties of unrealized potential

In this chapter we cite some properties of the above defined unrealized potential function URP. It can be shown that minimizing URP means to optimize a project's situation in terms of cost and yield. For mathematical proofs the interested reader is referred to Ultsch 2001[10].

A URP of zero means that the situation is equal to O, the theoretically best situation. For all situations having a unrealized potential of u it can be shown that the maximum cost of such situations may be only as big as u. Furthermore the yield of these situations must be 100% - u or bigger (theorem 5 in (Ultsch 2001)). If the cost of a situation is fixed and u is decreases, then the yield of the situation increases (theorem 6 in (Ultsch 2001)). Besides cost and yield of a situation also the efficiency, i.e. cost per yield, may be considered. If u is minimized it can be shown that the minimal efficiency of a situation is increasing (theorem 8 in (Ultsch 2001)). Furthermore the yield of the situation with minimal efficiency increases. If the cost of a situation is constant and the unrealized potential decreases, then yield and efficiency increase. If the yield of a situation is given then a smaller unrealized potential results in a better efficiency. For each given u there is a smaller u' such that the efficiency of all situations with u' is definitely bigger (theorem 9 in (Ultsch 2001)).

In summary minimizing URP means an economic optimization of a project's situation. If the URP is minimized yield and/or efficiency increase while costs are optimized.

8. Precise limits for ABC-analysis

In this chapter we derive limits for ABC-analysis. Tasks in class A of an ABC-analysis should have maximal yield with minimal costs. This means the situation U is the best candidate for such a task. consequently a cost limit of 20.13% is the best limit for class A.

For tasks in subproject B the following situations would be good candidates:

(I) a situation with a maximal partial yield (100%) or

(II) a situation where the efficiency is bigger than 100%

It can be shown, that for Entropic Yield the conditions (I) and (II) are identical (Ultsch 2001). The best limiting cost for class B is therefore 37%. For class C remains the rest of the cost i.e. 100% - 20% - 37% = 43%.

A division with borderlines of 20%, between A and B and 57% between B and C is optimal and may be termed "Entropic ABC-analysis". For Entropic ABC-analysis there is a precise definition of the classes' meaning:

Class A: Cost as well as yield are optimized in the sense that cost are minimized and yield is maximized. Optimal subproject is the one which comes closest to 0% effort and 100% yield.

Class B: Yield should be bigger than costs. Efficiency of this project should at least 100%.

Class C: the rest of the subprojects having smaller efficiency.

9. Applying unrealized potential theory

In many projects yield is defined through the project's nature, for example in terms of money. For these projects the theory outlined above might be used as a guideline to measure the quality of subprojects using URP. The derived borderlines for entropic ABC-analysis may be used as starting points to identify useful classes of projects.

If an explicit definition of yield is not given, however, the Entropic Yield may be applied. This might, for example, be used in

- determining the ideal number of factors doing principal component analysis
- determining the ideal number of clusters in cluster analysis
- deciphering the meaning of genetic code
- and others more

As an example for an application thereof let us consider cluster analysis. Cluster analysis means the detection of homogenious groups in high dimensional data sets. For the data a distance measure has to be defined measuring the (dis-)similarity between data points. For each point costs may be defined proportional to the distance between points. Yield can be assumed to be the number of other points in a point's vicinity. Using point U of URP a local density may be defined. This can be used to find clusters as well as outliers (Ultsch 2002). Entropic ABC-analysis provides precise borders allowing to state if a point belongs to a cluster or not.

10. Discussion

In this work we pose the question might there be a deeper reason behind the ubiquote encounter of Parto's 80/20 law? For this we regard the division of a project into subtasks. A very general accounting on the effort spent on subtasks of a project is a project report having symbols (letters, numbers) for each partial effort spent on a task. We try to measure the yield generated by the division of the project into subprojects. The requirements for such an yield function lead naturally to the axioms of Shannon Information (see (Ultsch 2001)). I.e. the yield generated by the division into subprojects might be identified with the information or entropy contained in the project report. With the right adjustment of units for cost and yield this leads to the definition of Entropic Yield.

Pareto's 80/20 law results from an optimization of cost and Entropic Yield for subprojects. This theory for the foundation of Pareto's 80/20-law derives a value of 20.13% for cost. This is almost identical with Pareto's law. The theory, however, implies a bigger yield of 88% instead of 80%. Since Pareto's law was until now only a rule of thumb one might allow that within 10% of error our theory is consistent with Pareto's observation. It is plausible that for realistic projects effects like friction or precision of measurement have to be considered. Cases where these phenomena may be disregarded may serve to (in-) validate the theory in the future.

Interestingly in the example of the South African land usage, which was randomly chosen from a book on the author's bookshelf[9], the interpolation for an area of 20.13% comes very close to the theoretical point U (please refer to Figure 1).

In figure 4 it can be seen that U is at a relatively flat section of the graph for URP. For the discussion of error margins of U let us consider the following graph:

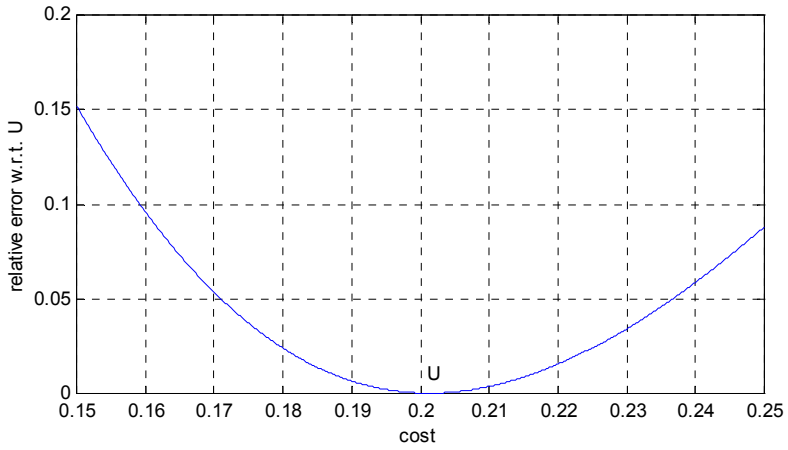

Figure 6 Relative error of Entropic Yield in the vicinity of U

In this graph URP(S) is related to U at 20.13%. It can be seen that a derivation of 10% from U i.e. the interval 18%..22% lies within 98% of the optimum. A derivation of 20%, i.e. the interval 16%..24%, still gives more than 90% of the optimal URP. 16% of cost results in 80% of Entropic Yield.

The outlined theory may be also important for Information- and Coding theory. Classical definitions of Shannon Information assume a source of information sending out distinguishable signs drawn from a finite set of signs. Our approach contributes to this theory the view that the production of a sign may be associated with costs. As a condition for the optimization of data streams the minimization of the number of signs while maximizing the transmitted information may be used. Optimal transmission frequency in this sense is $p_u = 20.13\%$.

Our theory also sheds some light onto the physical term Entropy according to Boltzmann (Boltzmann 2000). Entropy in physical systems might be considered as the probability of a system of many microscopic parts to be in a certain macroscopic state. Entropic Yield might be considered as the yield generated by solving a problem by the combination of random solutions.The usage of Shannon Information as some sort of yield is certainly justified in such situations where no other yield function is defined by the project. Entopic yield might also be used in project where the transmission of information is important, for example in marketing. As an application of the theory we have derived precise limits for an ABC-analysis. The partial yields for the classes derived are 32%, 37% and 36%. Within error margins this is consistent with the proposal of some practitioners of ABC-analysis to set the borderlines by dividing the total yield into three equal parts (33% yield for all classes).In many modern applications of data analysis like Data Mining or Knowledge Discovery heuristical rules are used to find abstractions of a mass of data, like principal components or clusters. Entropic ABC-analysis might serve as a theoretical sound justification to concentrate on a smaller but meaningful representation of the data (Class A).

11. Summary

In this work the often encountered occurrence of Pareto's 80/20 law is given a theoretical framework. Dividing a project into subtasks generates information which may be measured using Shannon Information. This information is considered as a form of yield. Cost is associated with the effort spent in a subtask. The optimization of cost and yield can be performed by minimizing a precise definition of the unrealized potential of a project's cost/yield situation. The optimal situation derived by this theory is 20% of cost yielding 88% of yield. The derived value of 20% for cost is identical with the Pareto's law. A yield of 88% is in the same order of magnitude as in Pareto's law. The excess yield might be encountered when friction is disregarded.

From our theory follow exact definitions of class borders for an ABC-analysis. This gives the classes a precise mathematical meaning. Our approach sheds some light on link between the physical definition of entropy and Shannon Information. Application of the theory in the field of Data Mining, Clustering and Bioinformatics are obvious.

12. Acknowledgement

The basis of this paper was developed between February to August 2001. I whish to thank Prof. Dr. Walter Seifritz, Prof Dr. Bruno Fritsch und Prof. Dr. Dirk Van den Poel for their comments on first versions. For textual corrections I thank Prof. Barbara Bahr.

13. Literature

Aho, A.V. ,Hopcroft, J.E., Ullman, J.: Data Structures and Algorithms, Addison Wesley,1982

Arthur L.J., Rapid Evolutionary Development - Requirements, Prototyping & Software Creation, John Wiley & Sons,1992

Boltzmann, L.: Entropy and Probability, Harri Deutsch, Frankfurt, 2000

Dyche, J.: The CRM Handbook: A Business Guide to Customer Relationship Management, Addison-Wesley, 2001

Gourdin,K.N., Global Logistics Management : a Competitive Advantage for the new Millennium, Oxford, Blackwell Publishers, 2001

Hartung, J.: Statistics, Oldenburg Verlag, München, 1997(in German)

Juran, J. M. (Editor), A. Blanton Godfrey (Editor), Juran's Quality Handbook, McGraw Hill, 1998

Juran, J.M., Pareto, Lorenz, Carnot, Bernoulli, Juran and Others, Industrial Quality Control, October 1950, p. 25

Kimber, R. J., Grenier, Robert W., Heldt,Jj.J., Quality Management Handbook, Marcel Dekker, York, NY, 1997

Pareto, V., Trattato di Sociologia Generale, Firenze,1916, Engl.: The Mind and Society, Dover, 1935

Shannon, C.E., A Mathematical Theory of Communication, The Bell System Technical Journal, Vol 27, pp 379-423, 1948

Sombart, W.: Die drei Nationalökonomien. Geschichte und System der Lehre von der Wirtschaft, Zeller Verlag, 1967 (in German)

Ultsch, A., Justification of Pareto's 80/20 law, Technical Report, 30, Department of Computer Science, University of Marburg, Germany, 2001(in German).

Ultsch, A., UPCluster: a Density Based Clustering Algorithm Based on a Proven Form of Pareto's 80/20 law, to appear, 2002.