Issues in Pareto analysis and their resolution

Article *in* Total Quality Management and Business Excellence · January 2012

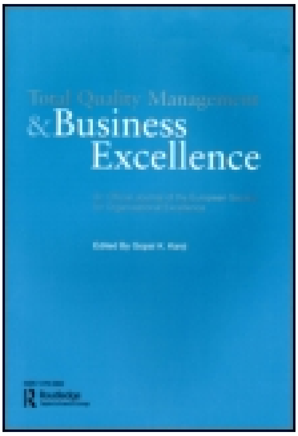
DOI: 10.1080/14783363.2012.704265

3 authors:

Ashok Sarkar
Indian Statistical Institute

Arup Ranjan Mukhopadhyay
Indian Statistical Institute

Sadhan Kumar Ghosh
Jadavpur University



Total Quality Management & Business Excellence

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ctqm20

Issues in Pareto analysis and their resolution

Ashok Sarkar [a], Arup Ranjan Mukhopadhyay [b] & Sadhan Kumar Ghosh [c]

[a] SQC & OR Unit , Indian Statistical Institute , Room No. 320 Old CGO Buildings, 101 MK Road, Mumbai , 400020 , India

[b] SQC & OR Unit , Indian Statistical Institute , Kolkata , India

[c] Department of Mechanical Engineering , Jadavpur University , India

Published online: 16 Aug 2012.

To cite this article: Ashok Sarkar , Arup Ranjan Mukhopadhyay & Sadhan Kumar Ghosh (2013): Issues in Pareto analysis and their resolution, Total Quality Management & Business Excellence, 24:5-6, 641-651

To link to this article: http://dx.doi.org/10.1080/14783363.2012.704265

Total Quality Management, 2013
Vol. 24, No. 5, 641–651, http://dx.doi.org/10.1080/14783363.2012.704265

Issues in Pareto analysis and their resolution

Ashok Sarkar[a*], Arup Ranjan Mukhopadhyay[b] and Sadhan Kumar Ghosh[c]

[a]*SQC & OR Unit, Indian Statistical Institute, Room No. 320 Old CGO Buildings, 101 MK Road, Mumbai 400020, India;* [b]*SQC & OR Unit, Indian Statistical Institute, Kolkata, India;* [c]*Department of Mechanical Engineering, Jadavpur University, India*

Vilfredo Pareto established the 'Pareto principle', which is also known as 'vital few, trivial many', to help in identifying 'vital few' errors for problem-solving. However, in many industrial applications, issues such as (a) the incorrect selection of the 'vital few' errors, (b) the interrelationship among errors and (c) the merging-up errors of different processes together need to be addressed. Otherwise, the chances are pretty high that the application of Pareto analysis will fail to correctly identify the 'vital few' errors, leading to an incorrect problem-solving approach. In this paper, the authors demonstrate the issues with the help of a real-life case study in a service scenario and suggest the appropriate remedial measures, for effectively separating out the 'vital few' causes from the 'trivial or useful many' causes to enhance the discriminating power of the Pareto graph.

Keywords: Pareto; correlation; discrimination; dendogram; quality improvement

1. Introduction

The 'Pareto principle', as stated by Vilfredo Pareto around 1895, is 'In any series of elements to be controlled, a selected small fraction, in terms of numbers of elements, always accounts for a large fraction in terms of effect' (Goodman, 2007, p. 370). Quality Management pioneer, Dr Joseph Juran, in the 1930s and 1940s recognised a universal principle applicable in problem-solving in quality improvement, which he called the 'vital few and trivial many' principle. The principle proposes that around 80% of the overall impact of errors in any industrial scenario is due to a small number of error types, termed the 'vital few' and 20% of the impact due to other error types, called the 'trivial many'. This was later popularised as 'Pareto analysis' (Juran, 1975). As the Pareto principle applies to most aspects of life, it is today popularly known as 'the 80/20 rule' (Koch, 1998).

'Pareto analysis' or the '80/20 rule' is a very useful tool in quality management, e.g. in implementing TQM, Six Sigma or Lean Six Sigma (Chen, Chang, & Huang, 2009; Wang & Chen, 2010), carrying out the design of experiment and problem-solving methods (Mukhopadhyay & Nataraja, 2004). In a comparative study, it has been found that Pareto analysis is the second ranked tool among 'Ten different Quality Control Technique' in Japan (Dahlgaard, Kanji, & Kristensen, 1990). In Six Sigma/Lean Six Sigma initiatives, Pareto analysis is used for the selection of problem/projects in the 'Define' phase and/or the identification of the 'vital few' errors in the 'Analyze' phase of Six Sigma. These 'vital few' errors are then analysed further for root causes and action is subsequently taken, either to prevent their occurrence or to reduce their impact. Hence, Pareto analysis guides the efforts to improve the process.

*Corresponding author. Email: sarkar.ashok@gmail.com

Pareto analysis is applied not only to guide solving quality problems, but also in different areas such as the identification of critical success factors for Six Sigma implementation (Karuppusami & Gandhinathan, 2006) and in aligning IT and business objectives (Huanga, Wub, & Chenc, 2012), identification of service quality parameters (Chatterjee & Chatterjee, 2007), etc. Pareto-like effects can also be seen to exist in multiple regression analysis (Chatterjee & Sorenesen, 1998). In the case of a parameter design, the effect of various factors can be understood quickly and easily by carrying out Pareto analysis on the contribution ratio (Park, 1995). In some instances, it may be better to look into different ways of carrying out Pareto analysis. For example, in a maintenance scenario, the question is whether to carry out Pareto analysis on the frequency of occurrence, or on the resultant downtime, of various types of failure. A better way to handle the scenario is to categorise the failures as either 'acute' or 'chronic' (Knights, 2001), but this does not address the cost-effectiveness of implementing a solution – it may be that the cost associated with removing a 'vital few' error is too high. In another case, it was found that Pareto analysis prioritised different types of error when the analysis was carried out with respect to the frequency of occurrence, to warranty cost, and to cost to the user (Radson & Boyd, 1997).

However, in real-life situations, the 80/20 pattern is often not observed, which makes the classification and differentiation of defects in 'vital few and trivial many' categories difficult. To verify the closeness of the empirical Pareto chart to the 'ideal' Pareto chart, an entropy-based approach was suggested by Grosfeld-Nir, Ronen, and Kozlovsky (2005), but the remedial measures were not addressed. The maintenance example cited above (Knights, 2001) is analysed for a number of occurrences and total downtime and presented in Figure 1. From Figure 1, we observe a flat-type Pareto chart for both cases, i.e. the occurrences of successive errors, is more or less the same. Many practitioners suggest that whenever the heights of the bars are more or less equal, then it is necessary to look for other ways to categorise the data, or look for different kinds of data, for the given problem.

It may be noted here that Pareto analysis results are useful only when the data come from a stable process. Where process stability is an issue, then the interpretation of Pareto analysis results may be erroneous (Bajaria, 1998). However, other than the stability issue, the authors feel that the major issues which need to be addressed for effective application of Pareto analysis are:

- criteria for selection of an error as being in the 'vital few' or 'trivial many' category, i.e. discrimination between the two types of errors;

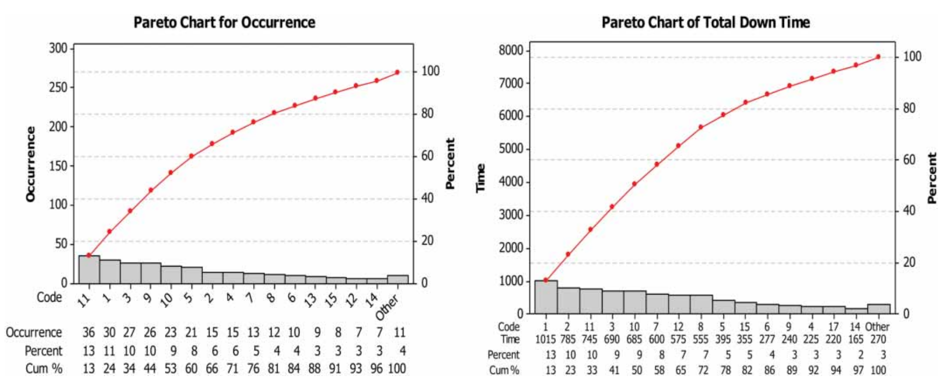


Figure 1. Pareto chart for downtime and occurrences.

- interrelationship between the errors;
- combining errors of different processes and solving them as one problem.

In this paper, the authors first demonstrate the above issues with a real-life case study, suggest analytical tools for the identification of the above issues and then suggest remedial measures. Finally, a methodology is proposed for carrying out Pareto analysis.

2. Problem, analysis and issues

The company where the issues are identified is engaged in medical transcription, where the input is 'recorded patient–doctor conversation in an audio file' and the output is a 'transcription document'. The company has four clients abroad and the files are processed in a Business Process Outsourcing organisation in India. Errors in conversions create major customer dissatisfaction and affect productivity. During July–September 2010, a 15% error rate was reported in the transcription process. Hence, we planned a study to identify the type of errors, their causes and possible remedial measures.

The errors which occurred during the transcription process are collected and presented in Table 1.

Pareto analysis is carried out as per standard text-book procedure (Ishikawa, 1976), and the results are presented in Figure 2. The 'vital few' and 'trivial many' errors are also identified and shown.

Table 1. Processing errors and their occurrences.

Error code	Details of the error	Occurrences
1	TranscriptionDate: the transcription date is not a valid date	220
2	PatientClass: the visit ID does not match the demographics data	158
3	PatientLocation: the visit ID does not match the demographics data	157
4	PatientType: the visit ID does not match the demographics data	154
5	VisitID: the visit ID does not match the demographics data	153
6	SigningPhysicianID: the signing physician ID is missing	142
7	DictatingPhysicianID: dictating physician ID and CPC code does not match	139
8	EditStatus: Edit status set as no dictation. Please mark job as no dictation	104
9	PatientName: the patient last name is missing	81
10	DictatingPhysicianID: dictating physician ID should never match the Vox Header Author ID. CPC code from the physician list needs to be entered	75
11	MRN: MRN cannot be blank	65
12	VisitID: visit ID should not be blank. If the ADT is not available then enter 000999999 in the visit ID field	52
13	DictatingPhysicianID: The dictating physician ID is missing	42
14	PatientName: the patient last name does not match the demographics data	31
15	NumberofCopies: please check the value entered in the number of copies filed	13
16	AttendingPhysician: attending physician field should not be blank	11
17	JobNoBookMarkCheck: the job number bookmark is missing	8
18	DictationDate: the dictation date is not a valid date	6
19	JobNumber: the job number is missing	4

Note: For easier interpretation and further analysis, errors are shown in the decreasing frequency order.

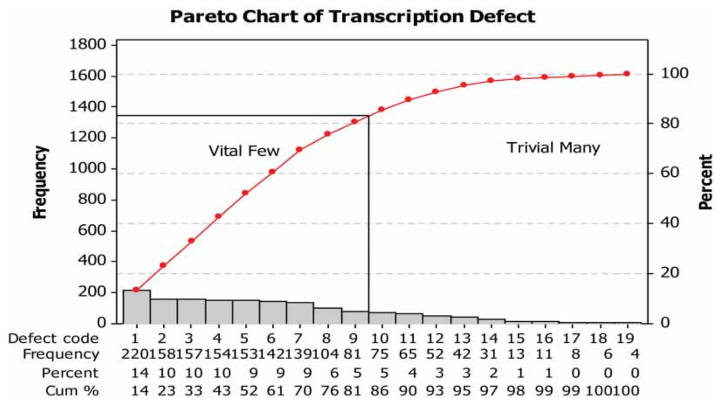


Figure 2. Pareto chart for transcription errors, and the 'vital few' errors.

2.1 *Issue 1: discrimination among error types*

From Figure 2, we can see that Error Type 9 is identified as 'vital few', whereas Error Type 10 is grouped into 'trivial many', but what differentiates the two error types is not clear. The errors in the 'vital few' category *should* have significantly higher occurrences than those in the 'trivial many' category. The 'two-sample Poisson rate' test was used to examine the difference of occurrences of Error Types 9 and 10, and the result is given in Table 2. The null hypothesis of equality of the occurrences of Error Types 9 and 10 was not rejected at the 5% level of significance. The corresponding alternate hypothesis is one-sided, i.e. the occurrence rate of Error Type 9 is higher than that of Error Type 10, and this finding proves that the discrimination of Error Types 9 and 10 into different categories is not justified.

Successive tests of the hypothesis were carried out to find the difference of occurrence rate of the last error type in the 'vital few' category and the first error type in the 'trivial many' category, and the results are presented in Table 3. The '$p < 0.05$' indicates that the occurrence rate of an error type is not the same and discrimination is justified. In other words, the 'p-value' indicates the validity of the discrimination in the selection.

From Table 3, we can see that the discrimination is significant between Error Types 1 and 2 ($p < 0.05$), as well as Error Types 7 and 8 ($p < 0.05$). As the inclusion of Error Type 7 gives a selection of 70%, we suggest identifying Error Types 1–7 as the 'vital few' and rest as the 'trivial many'.

2.2 *Issue 2: correlation among errors*

From the Pareto chart in Figure 2, we see that the heights of the bars for Error Types 2–7 are almost equal. The general approach to get over this problem is to look for ways to re-categorise the data and/or look for different kinds of data for the given problem. The issue

Table 2. Result of two-sample Poisson rate analysis.

Error type	Occurrences	Total occurrences	Occurrence rate	p-Value
9	81	1618	0.0500618	0.631
10	75	1618	0.0463535	

Table 3. Discrimination power of selection between 'vital few' and 'trivial many'.

Vital few		Trivial many		
Total count	Last entry error	Total count	First entry error	p-Value
1	1	18	2	0.001
2	2	17	3	0.955
3	3	16	4	0.865
4	4	15	5	0.954
5	5	14	6	0.522
6	6	13	7	0.858
7	7	12	8	0.025
8	8	11	9	0.091
9	9	10	10	0.631

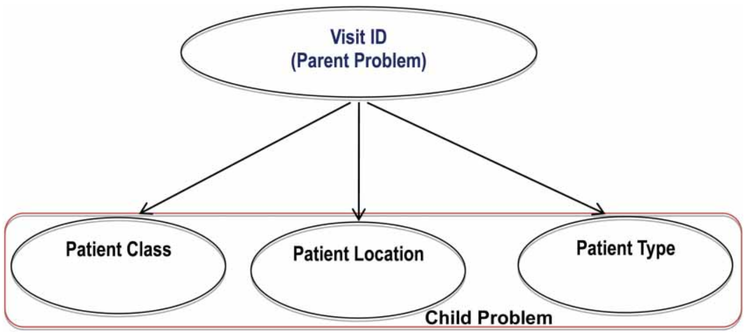


Figure 3. Parent–Child syndrome among error types.

is discussed with the process engineer, and a peculiar relationship has been found among the error types, which we term as the 'Parent–Child' syndrome, as shown in Figure 3.

The logic in this syndrome is that if there is an error type in the Parent category, then the other three error types belonging to the Child category will also occur. However, the opposite is not true; an error in the Child category will not create an error in the Parent category. Hence, we cannot say the error types are mutually exclusive. The 'Visit ID' is assigned to a patient when admitted, and this number is used for billing the patient at the time of discharge. To determine billing, the following information are captured along with the visit ID:

- Patient class, which determines the admission type of the patient and the intended management.
- Patient location, which determines the code identifying the location where the patient is receiving medical treatment.
- Patient type, which determines the code identifying the type of treatment the patient is receiving.

For the client, these are mandatory parameters necessary in patient billing associated with the visit ID. If visit ID is missing for a particular patient's stay, the patient class, location and type do not get captured.

The error checking system verifies details for visit ID or account number. If the visit ID field is missing or not available in the client ADT information, then the error checking

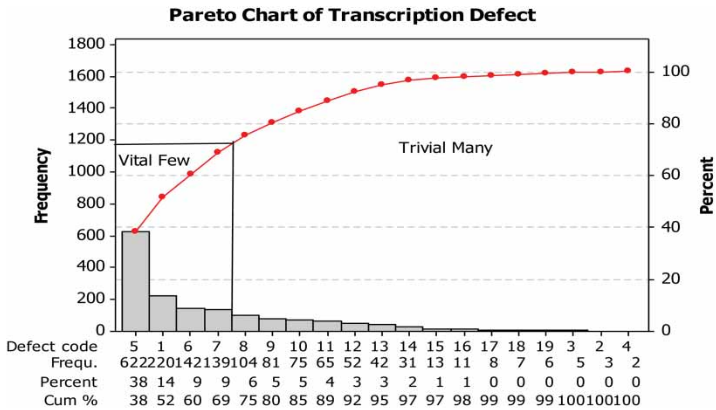


Figure 4. Modified Pareto chart for transcription error and the 'vital few' errors.

system will reject the particular job for patient class, patient location and patient type fields as well. If the 'visit ID field missing error' is resolved, then the file will not get rejected for the other three dependent fields.

After verification of the above, we decided to combine the four error types when they occur together with visit ID. However, the other error types which occurs independent of visit ID are kept separate. The Pareto chart is drawn again and shown in Figure 4.

Now if we look at the Pareto chart, we can select four error types as 'vital few', which contribute to 69% of the total, which meets the 'Pareto principle'.

It may be argued here that the error types should be mutually exclusive to carry out Pareto analysis. However, errors categorised here as 'Child' can occur independently, and hence there is no relation among them. Now the question is how to identify such patterns among the error types.

2.2.1 *Identification of correlation among error types*

The correlation among error types can be identified in two ways

(1) *Method 1*: *Supervised learning (through experience and knowledge)*: When defining the error types, the process experts should have thorough discussions about the relationships, and if they agree, then the categories should be defined accordingly.
(2) *Method 2*: *Unsupervised learning (through data analysis and interpretation)*: Carry out either cluster analysis or correlation among the error types and prioritise those which occur together. Confirm their independence through chi-square analysis.

Method 1 is totally dependent upon the experts. However, it is always better to go through Method 2 in addition to Method 1 to identify the relationship, and then to validate it through process knowledge. In order to describe Method 2, data have been collected from another transcription process, where the occurrences are as given in Table 4.

In order to identify the relationships between the error types, the data are entered in the following way. There are a total of 15 error types in the transcription process. If a particular error type occurs for a transcription, then we enter 1, otherwise we enter 0. The resulting correlation matrix is shown in Table 5.

Table 4. Error types and their occurrences in a transcription process.

Error code	Details of the error	Occurrences
A	AttendingPhysician: attending physician field should not be blank	11
B	DictatingPhysicianID: dictating physician ID and CPC code does not match	156
C	DictatingPhysicianID: dictating physician ID should never match the Vox Header Author ID. CPC code from the physician list needs to be entered	58
D	DictatingPhysicianID: the dictating physician ID is missing	42
E	DictationDate: the dictation date is not a valid date	3
K	PatientName: the patient last name is missing	127
G	EditStatus: edit status set as no dictation. Please mark job as no dictation	166
H	JobNoBookMarkCheck: the job number bookmark is missing	5
I	MRN: MRN cannot be blank	65
J	NumberofCopies: please check the value entered in the number of copies filed	13
F	AttendingPhysician: attending physician field should not be blank	35
L	SigningPhysicianID: the signing physician ID is missing	142
M	TranscriptionDate: the transcription date is not a valid date	221
N	VisitID: the visit ID does not match the demographics data	638
O	VisitID: Visit ID should not be blank. If the ADT is not available then enter 000999999 in the visitID field	36

It can be seen from Table 5 that the Error Types G and K, are highly related ($r = 0.86$), but difficult to identify. In order to identify the relationship between the variables, the easiest method is to draw a 'dendrogram'. A dendrogram is a visual representation of the correlation or association of variables. The individual variables are arranged along the bottom of the dendrogram (x-axis) and referred to as leaf nodes. On the y-axis, the similarity between two observations, or cluster, is shown. The similarity, s_{ij}, between two observations i and j is given by $s_{ij} = 100(1 - d_{ij})/d_{\max}$, where $d_{\max}$ is the maximum value in the distance matrix. The distance between two points is measured using various methods, e.g. in the Euclidean distance formula, the distance between observations i and j is measured as

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2)}$$

The higher the similarity, the more the two variables are correlated. A 'dendrogram' of the error types is presented in Figure 5.

From Figure 5, it can easily be seen that the Error Types G and K are correlated, as the similarity between the two is maximum. The independence of Error Types G and K has been verified through a chi-square test of independence, and the result is shown in Table 6. The null hypothesis of independence of error types is rejected at the 1% level of significance. The corresponding alternate hypothesis is that the error types are dependent.

The findings from Table 6 have been discussed with a process expert and their rationale has been explored. Later on, both the Error Types G and K have been combined and the modified Pareto analysis is presented in Figure 6. This verification is necessary when a

Table 5. Correlation coefficient matrix.

	A	B	C	D	F	G	I	J	K	L	M	N	O	H	E
A	1.00														
B	−0.03	1.00													
C	−0.02	−0.08	1.00												
D	−0.02	−0.07	−0.04	1.00											
F	−0.02	−0.06	−0.04	−0.03	1.00										
G	−0.03	−0.13	−0.05	0.04	−0.06	1.00									
I	−0.02	−0.08	−0.05	0.00	−0.04	0.11	1.00								
J	−0.01	−0.04	−0.02	−0.02	−0.02	−0.04	−0.02	1.00							
K	−0.03	−0.11	−0.06	0.04	−0.05	**0.86**	0.08	−0.03	1.00						
L	−0.03	−0.10	−0.04	0.04	−0.06	0.03	0.18	−0.03	0.00	1.00					
M	−0.04	−0.14	0.21	−0.08	−0.07	−0.16	−0.09	−0.04	−0.14	−0.13	1.00				
N	−0.09	−0.25	−0.20	−0.17	−0.16	−0.02	0.13	−0.10	0.03	−0.16	−0.42	1.00			
O	−0.02	−0.06	−0.04	−0.03	−0.03	−0.06	−0.04	−0.02	−0.05	−0.03	−0.07	−0.16	1.00		
H	−0.01	0.02	−0.01	−0.01	−0.01	−0.02	−0.01	−0.01	−0.02	−0.02	0.04	−0.06	−0.01	1.00	
E	0.00	−0.02	−0.01	−0.01	−0.01	−0.02	−0.01	−0.01	−0.02	−0.02	0.02	−0.05	−0.01	0.26	1.00

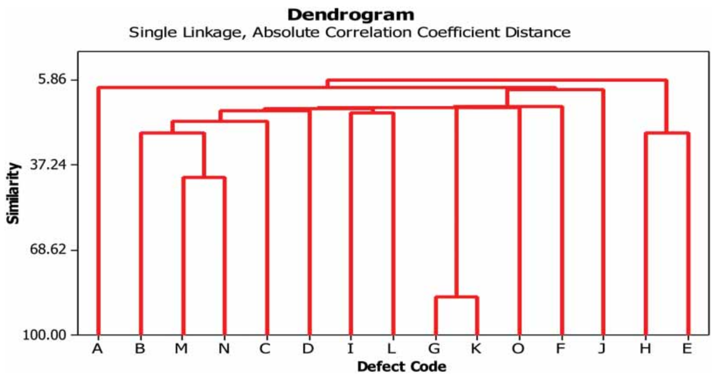


Figure 5. Dendrogram of the error type.

Table 6. Results of chi-square test for independence.

KG	0	1	Total
0	1170	0	1170
	1058.78	111.22	
	11.68	111.22	
1	39	127	166
	150.22	15.78	
	82.34	783.90	
Total	1209	127	1336

Notes: Expected counts are below observed counts. Chi-square contributions are below expected counts. $\chi^2 = 989.149$, DF $= 1$, $p = 0.000$.

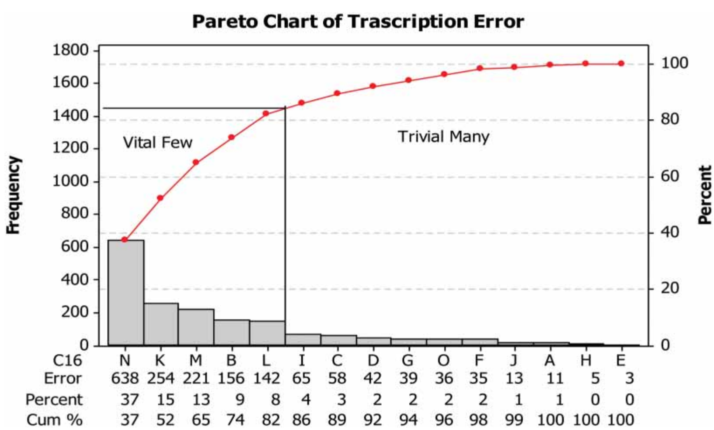


Figure 6. Modified Pareto for transcription error types.

product has multiple error types, as there is a possibility that some error types may be correlated.

The modified Pareto chart is now used for the prioritisation of error types and root cause analysis.

Table 7. Error type and its frequency in two processes.

	Frequency		
Error type	Client A	Client B	Total
B	156	69	225
C	58	45	103
D	42	22	64
F	35	157	192
G	166	65	231
I	65	46	111
J	13	35	48
K	127	22	149
L	142	245	387
M	221	36	257
N	638	369	1007
O	36	55	91

2.3 *Issue: combining two processes*

The general practice is to combine the errors of two or many processes together and then to carry out the Pareto analysis. However, it is possible that the two process practices are entirely different for various reasons. For demonstration purpose, we have taken data from two processes dealing with two clients and presented it in Table 7.

The independence of processes A and B is verified through the chi-square test of independence. The χ^2 value is 401.695 with 11*df* and the *p*-value is 0.000. The null hypothesis of independence of process type has been rejected at the 1% level of significance. The corresponding alternate hypothesis is that the error types are dependent. As the processes are independent, the Pareto analysis should be carried out separately instead of doing it for the combined data.

3. Conclusion

In the above sections, the authors have discussed the various issues related to carrying out Pareto analysis and their possible resolution. However, the challenge is the identification of correlation between error types, for which the authors suggest carrying out cluster analysis, and its visual representation through the use of dendrograms.

To carry out Pareto analysis properly, we suggest the following methods.

(1) Identify the various error types in a process and their relationships. If two error types have dependence, then combine them.
(2) Collect data for each error type as a binary variable. If a particular error type occurs, put 1, otherwise put 0.
(3) Construct a dendrogram using standard statistical software and visually identify the error types having dependence between them.
(4) Carry out chi-square analysis and verify their dependence or independence.
(5) If the two error types are found to be dependent, verify the finding in consultation with the pertinent process owners.
(6) Combine the error types, if necessary, and then draw the modified Pareto chart.
(7) If the error types belonging to two processes are found to be dependent, combine the processes after verifying their dependence through chi-square analysis.

(8) Verify the validity of discrimination for classification of the error types as 'vital few' or otherwise.
(9) Use the Pareto analysis results for root cause analysis and further improvement.

However, as mentioned earlier, steps 1–4 are applicable subject to the existence of multiple error types in a product.

Acknowledgements

The demonstrative case studies used for highlighting the major points of the paper are carried out at a leading company engaged in medical transcription. The authors acknowledge the contribution of designated Black Belt, Mr K.P. Singh.

References

Bajaria, H.J. (1998, May). *Vital few, trivial many – challenged*. Annual Quality Congress (Vol. 52, pp. 474–479). Philadelphia, PA.

Chatterjee, S., & Chatterjee, A. (2007). Prioritization of service quality parameters based on ordinal responses. *Total Quality Management*, *16*(4), 477–489.

Chatterjee, S., & Sorenesen, E. (1998). A Pareto-like effect in regression? *Total Quality Management & Business Excellence*, *9*(8), 681–687.

Chen, S.C., Chang, L., & Huang, T.H. (2009). Applying Six-Sigma methodology in the Kano quality model: An example of the stationery industry. *Total Quality Management & Business Excellence*, *20*(2), 153–170.

Dahlgaard, J.J., Kanji, G.K., & Kristensen, K. (1990). A comparative study of quality control methods and principles in Japan, Korea and Denmark. *Total Quality Management*, *1*(1), 115–132.

Goodman, T. (Ed.). (2007). *The Forbes book of business quotations*. New York, NY: Black Dog & Leventhal Publishers, Inc.

Grosfeld-Nir, A., Ronen, B., & Kozlovsky, N. (2005). The Pareto managerial principle: When does it apply? *International Journal of Production Research*, *45*(10), 2317–2325.

Huanga, S., Wub, M., & Chenc, L. (2012). Critical success factors in aligning IT and business objectives: A Delphi study. *Total Quality Management & Business Excellence*, 1–22, iFirst, doi:10.1080/14783363.2011.637785.

Ishikawa, K. (1976). *Guide to quality control* (Rev ed.). Tokyo: Asian Productivity Organization.

Juran, J.M. (1975). The non-Pareto principle; Mea Culpa. *Quality Progress*, *8,* 8–9.

Karuppusami, G., & Gandhinathan, R. (2006). Pareto analysis of critical success factors of total quality management. *The TQM Magazine*, *18*(4), 372–385.

Knights, P.F. (2001). Rethinking Pareto analysis: Maintenance application of logarithmic scatterplots. *Journal of Quality in Maintenance Engineering*, *7*(4), 252–263.

Koch, R. (1998). *The 80/20 principle*. New York, NY: Doubleday.

Mukhopadhyay, C., & Nataraja (2004). Improvement of piston ring quality: A case study. *Total Quality Management & Business Excellence*, *15*(1), 105–125.

Park, S.H. (1995). A new method of analysis for parameter design in quality engineering. *Total Quality Management*, *6*(1), 13–20.

Radson, D., & Boyd, A.H. (1997). The Pareto principle and rate analysis. *Quality Engineering*, *10*(2), 223–229.

Wang, F., & Chen, K. (2010). Applying Lean Six Sigma and TRIZ methodology in banking services. *Total Quality Management & Business Excellence*, *21*(3), 301–315.