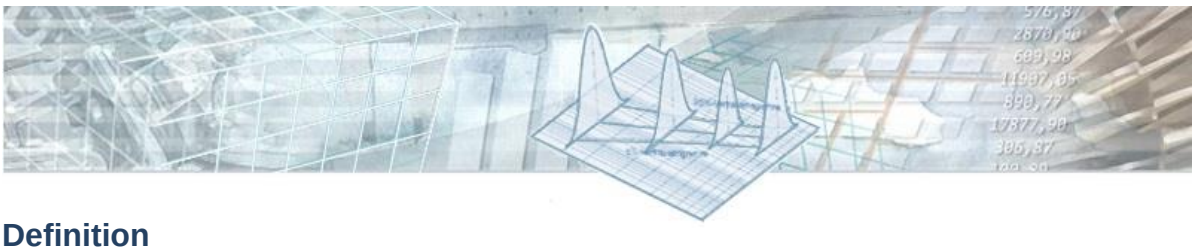
Häufigkeitsgruppen



Definition

Bei der Bildung von Häufigkeitsgruppen zählt man die vorkommenden Einträge bzw. Merkmale in einer Tabelle. Dieses Vorgehen dient dazu Kombinationen zu finden, die am meisten vorkommen, insbesondere bei der Fehlersuche. Es können große Datenmengen ausgewertet werden. Ähnliche Methoden sind unter dem Begriff Data Mining bekannt.

Ziel und Nutzen

Das Ziel ist es durch die Bestimmung von Häufigkeiten Muster zu erkennen, um Rückschlüsse auf Zusammenhänge zu finden, die nicht durch Regressionsverfahren aufgezeigt werden können. Mit Hilfe des sogenannten Chi²-Mehrfeldtests lassen sich signifikante Unterschiede bestimmen.

Grundlagen

Es liegt eine Tabelle vor mit Messwerten und Eigenschaften, etc. Im folgendem werden diese Merkmale genannt. In diesem Fall enthält die letzte Spalte eine Zielgröße Dichtheit.

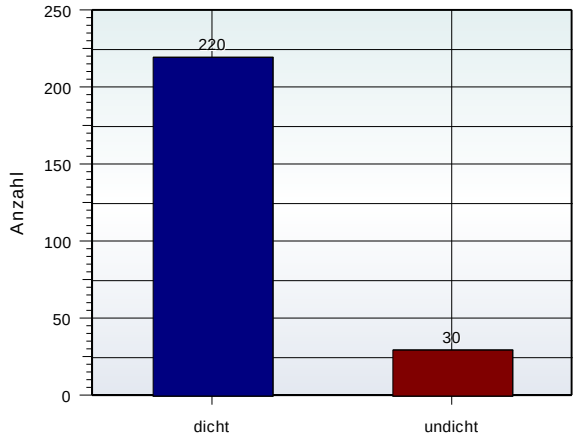
Es soll ermittelt werden, bei welchen Kombinationen das System am meisten undicht ist.

Der erste Schritt ist es die Anzahl dicht und undicht zu zählen. Dies stellt später die Bezugsgröße der „Hauptgruppen“ dar.

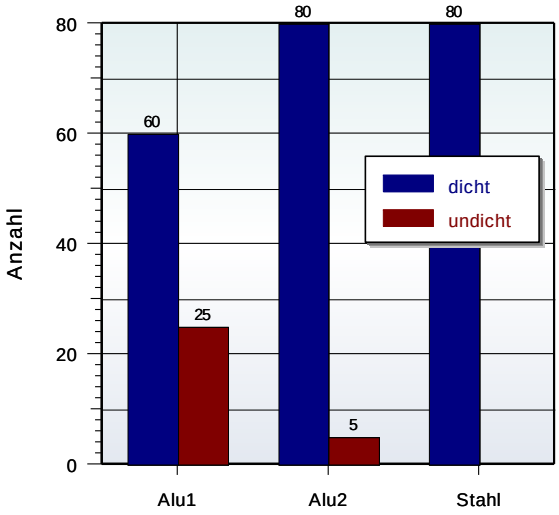
Parameterspalten | Zielgröße

	A	B	C	D	E
1	Temperatu	Druck	Druckplatte	Rohmaterial	Dichtheit
2	100	5	Stahl	EPDM	dicht
3	100	7	Alu2	FPM	dicht
4	80	7	Alu2	HNBR	dicht
5	60	5	Alu1	HNBR	dicht
6	60	7	Stahl	HNBR	dicht
7	100	6	Stahl	HNBR	dicht
8	60	7	Alu2	EPDM	dicht
9	80	7	Alu1	FPM	dicht
10	80	5	Stahl	FPM	dicht
11	60	6	Alu2	ACSM	dicht
12	80	6	Alu2	ACSM	dicht
13	80	6	Alu1	EPDM	dicht
14	100	7	Alu1	HNBR	undicht
15	100	5	Stahl	ACSM	dicht
16	100	7	Alu2	ACSM	dicht
17	100	7	Alu1	EPDM	undicht
18	60	7	Alu1	HNBR	dicht
19	100	5	Alu1	ACSM	dicht

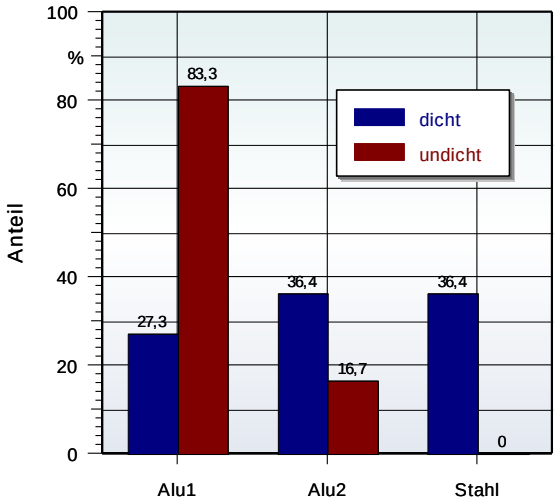
Bei den Parameterspalten werden nun die Häufigkeiten der Merkmale zwischen dicht und undicht aufgeteilt. Idealerweise sollte die Zielgröße, genauso viele „dicht“ wie „undicht“ aufweisen. In den meisten Fällen liegen aber deutlich weniger Fälle eines fehlerhaften Systems vor. Damit undicht auch bei möglichen Zweierkombinationen der Merkmale entdeckt werden kann, sollten mindestens für p-Merkmale p+p(p-1)/2 Zeilen mit undicht vorliegen. Oft liegen deutlich größere Datenmengen vor, insbesondere aus Felddaten. Umfänge von mehr als 10.000 Datensätzen bringen dagegen meist keine genaueren Aussagen mehr, verlangsamen aber die Auswertezeit.



Kommen Merkmale grundsätzlich unterschiedlich oft vor, so wird dies durch eine prozentuale Betrachtung auf dicht und undicht relativiert. Im Gegensatz zur absoluten Betrachtung (linkes Bild) stellt die relative Darstellung (rechtes Bild) die Verhältnisse besser dar:



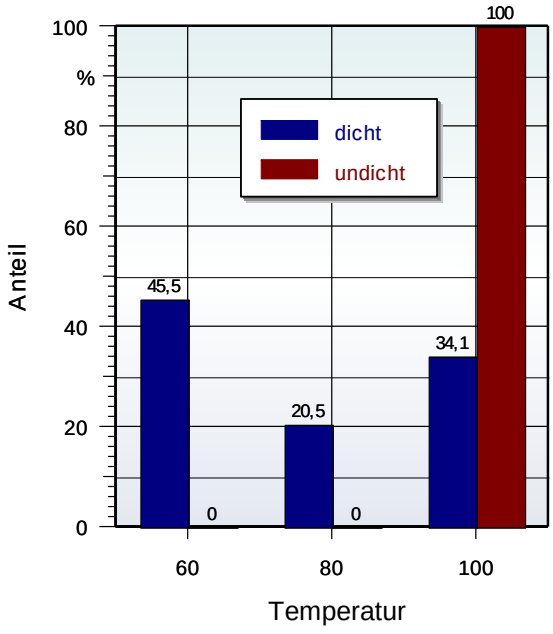
Absolute Anzahl für Material

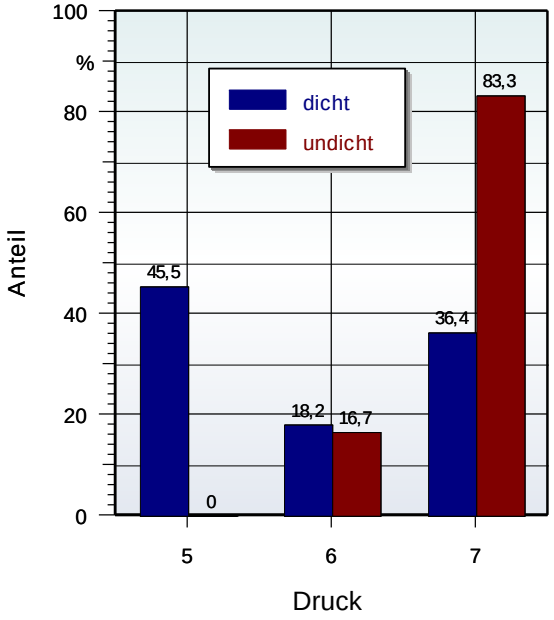


Relative Häufigkeit für Material

Gibt es beispielsweise die Druckplatte weniger in Stahl, gegenüber Alu, so kommt Stahl bei dicht und undicht im Verhältnis gleich oft vor. Ist aber eine prozentuale Häufung von Alu bei undicht vorhanden, so ist hier anzunehmen, dass undicht eher bei Alu auftritt. In diesem Beispiel gab es Alu1 bei dicht 12 mal. Insgesamt gibt es 44 mal dicht. Somit ist das Verhältnis dichter Einheiten mit Alu1 27,3%. Undichte mit Alu1 kommen jedoch 5 mal vor bei insgesamt 6 undichten Einheiten. Alu1 und undicht sind somit 83,3%.

Der nächste Vergleich betrifft die Temperatur. Undicht tritt offensichtlich erst bei 100°C auf. Ebenso ist es beim Druck, der möglichst hoch sein muss, insbesondere bei 7bar.





Der Fehler tritt also nur in Kombination auf. In der Praxis ist das dann meist seltene Betriebs- oder Umweltbedingungen. Ansonsten hätte man dies bei der Erprobung schneller festgestellt.

Häufigkeitsgruppen

In diesem Beispiel sind die Verhältnisse sehr eindeutig. Sind die Unterschiede nicht so stark, wie hier beim Rohmaterial, so hilft der bekannte χ^2 -Mehrfeldtest eine statistische Aussage zu treffen. Verglichen wird die Anzahl der Merkmale der jeweiligen Parameterspalte (Stichproben), bezogen auf den Gesamtumfang der Tabelle.

Die Nullhypothese H_o lautet: Die Verteilung der Merkmalen ist für die Zielgröße Dichtheit gleich. Die so genannte Kontingenztafel ist allgemein:

Gruppe i	1	2	...	k
Stichprobenumfang	n1	n2	...	nk
Anzahl Fehler	x1	x2	...	xk

In diesem Beispiel soll die Prüfung anhand des Rohmaterials erfolgen. Hier gibt es: ACSM, EPDM, FPM und HNBR. Jedes Material stellt eine Stichprobe dar mit folgenden dichten und undichten Einheiten:

Gruppe i		1	2	3	4
		ACSM	EPDM	FPM	HNBR
	dicht	55	50	60	55
Anzahl Fehler xi	undicht	10	10	5	5
Stichprobenumfang ni	Summe	65	60	65	60

Hinweis: Die Besetzungszahlen sollten nicht kleiner als 5 sein. Ist eine Besetzungszahl 0, so werden auf allen Felder 0,5 hinzuaddiert.

Die Prüfgröße bestimmt sich mit k=Anzahl Merkmale mit:

$$\chi^2 = \sum \frac{(Anzahl Fehler - Anzahl erwartet)^2}{Anzahl erwartet} = \sum_{i=1}^{k} \frac{\left(x_i - n_i \frac{x_{ges}}{n_{ges}}\right)^2}{n_i \frac{x_{ges}}{n_{ges}}}$$

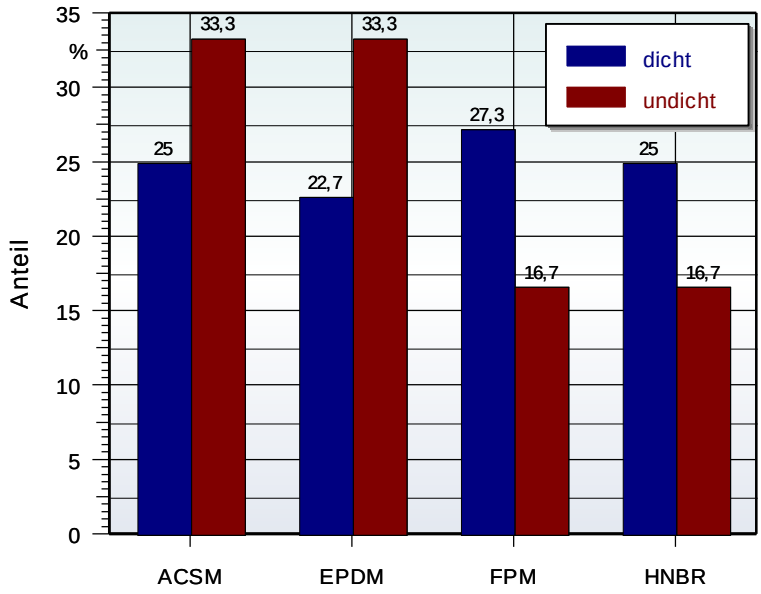
mit $x_{ges} = \sum_{i=1}^{k} x_i = 30$; $n_{ges} = \sum_{i=1}^{k} n_i = 250$

χ^2 berechnet die Summe der Abweichungsquadrate. Wird diese zu groß, wird die Nullhypothese abgelehnt, d.h. ein Unterschied zwischen den Merkmalen ist dann signifikant. In diesem Beispiel berechnet sich die Prüfgröße zu:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(x_i - n_i \frac{x_{ges}}{n_{ges}}\right)^2}{n_i \frac{x_{ges}}{n_{ges}}} = \frac{(10 - 65 \cdot 30/250)^2}{65 \cdot 30/250} + ... + \frac{(5 - 60 \cdot 30/250)^2}{60 \cdot 30/250} = 3{,}39$$

Für diese Prüfgröße wird die Wahrscheinlichkeit P aus der χ^2–Verteilung berechnet. Der hierfür benötigte Freiheitsgrat ist f = k-1. Die Wahrscheinlichkeit kann über die Excel-Formel =CHIQU.VERT(χ^2; f; WAHR) ermittelt werden. Hier ergibt sich P= 0,664. Die Irrtumswahrscheinlichkeit p_{value} = 1 – P = 1 - 0,664 = 0,336. Ist dieser

Wert kleiner als das definierte Signifikanzniveau von α = 0,05, wäre die Nullhypothese abzulehnen. Dies ist hier aber nicht der Fall, obwohl es Unterschiede der Materialien gibt, wie im folgendem Bild dargestellt.



Diese sind aber aufgrund des Testergebnisses als nicht signifikant anzusehen.

Hinweis: Für den χ^2- Mehrfeldtest sollten die Häufigkeiten in den Gruppen mindestens zweistellig sein.

Es ist aber so, dass es gleichzeitig noch andere Einflussparameter in den Gruppen gibt, von denen für die Anwendung des χ^2 -Tests vereinfacht angenommen wurde, dass sich diese gleichmäßig aufteilen und unabhängig sind.

Die tatsächlichen Anteile der Gruppenmerkmale sollen nun in einem zweiten Schritt bestimmt werden. Welche Zeilenkombinationen kommen mehrfach vor. Diese werden entfernt, sodass Zeilenkombinationen einmalig sind (Unikat). Die Zielgröße wird zuerst dargestellt, gefolgt von den Parametern mit der geringsten Anzahl Merkmalen, siehe rechte Tabelle:

Dichtheit	Temperatu	Druck[bar]	Druckplatt	Rohmaterial
dicht	100	5	Stahl	EPDM
Anz:220	Anz:75	Anz:45	Anz:10	ACSM
			Alu2	EPDM
			Anz:20	FPM
				HNBR
				ACSM
			Alu1	EPDM
			Anz:15	HNBR
				ACSM
		7	Stahl	EPDM
			Alu2	FPM
		6	Stahl	HNBR
		Anz:15	Anz:10	ACSM
			Alu2	EPDM
undicht	100	7	Alu2	Anz:15
Anz:30	Anz:30	Anz:25	Alu1	
			Anz:25	FPM
				HNBR
				ACSM
		6		Anz:10

Gut zu erkennen ist hier, was bei den Grafiken bereits herauskam, dass undicht mit 100°C, 7bar und Alu1 am meisten vorkommt (letzter Block mit 6 Zeilen). Das Rohmaterial ist eher indifferent.

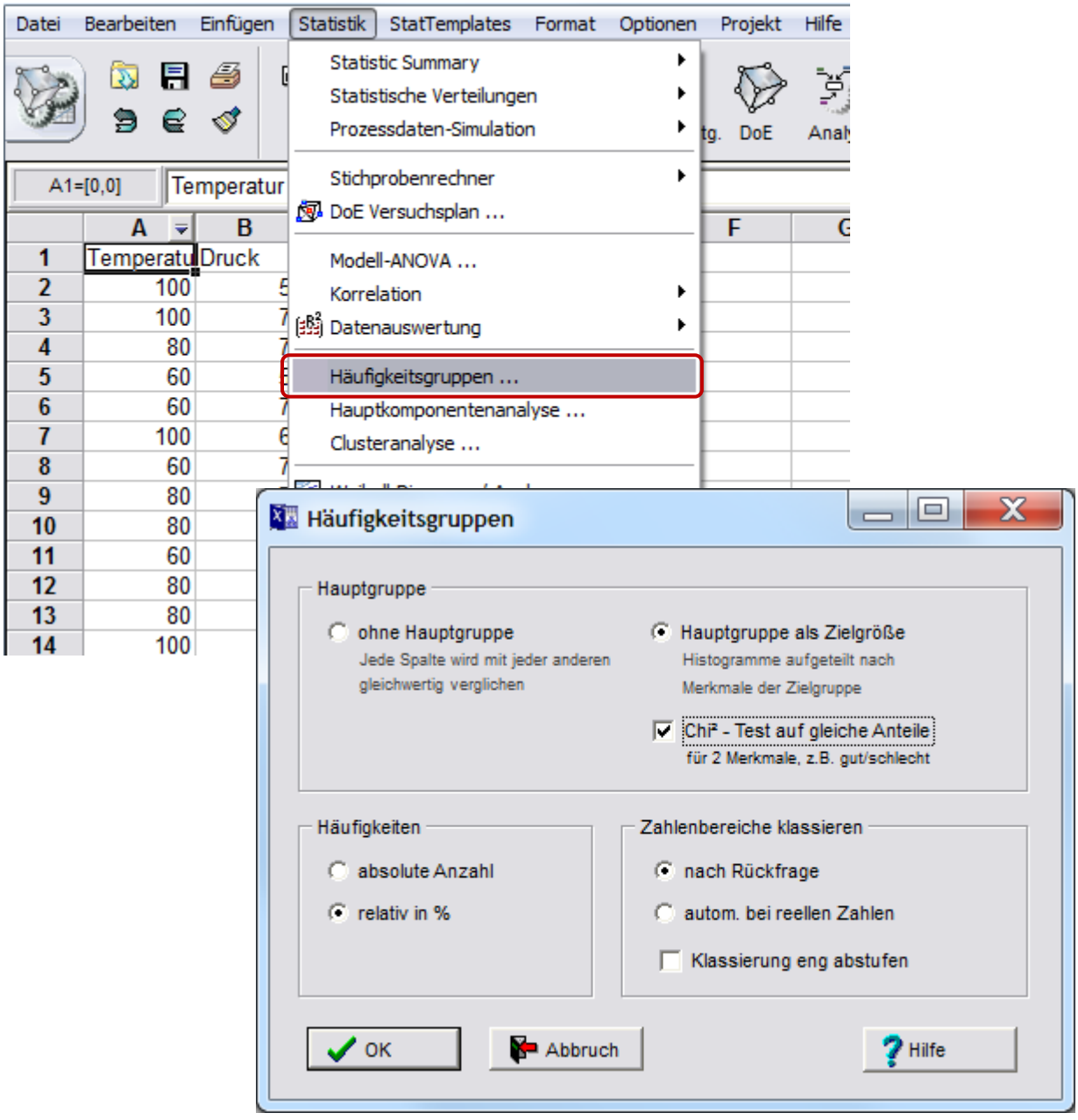
Besonders stark werden klare Zusammenhänge, wenn ab der Grenze zu undicht eine möglichst

durchgehende Trennlinie vorhanden ist. Diese wird hier erst ab Rohmaterial unterbrochen.

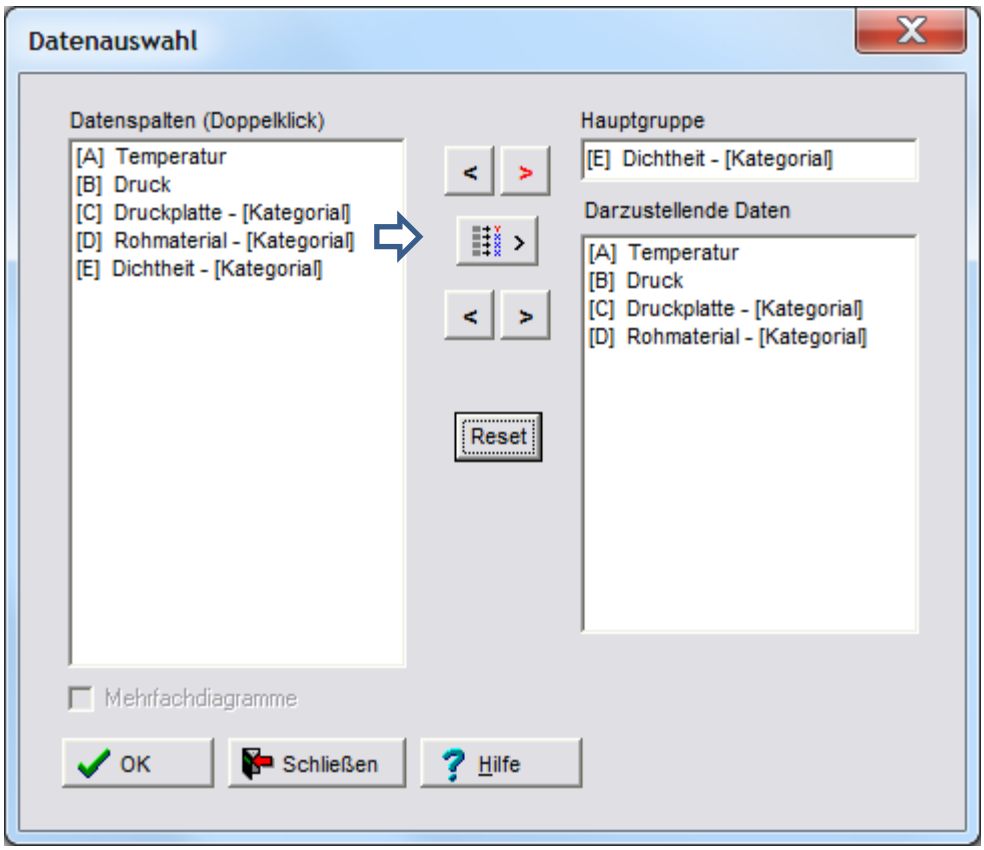
Der Vorteil diese Darstellung gegenüber einer hierarchischen Baumstruktur ist, dass man die Zusammenhänge quer im gesamten Überblick beibehält, während man in einer Baumstruktur „gedanklich“ immer einem Pfad folgt und die anderen Kombinationen verlässt.

Anwendung in Visual-XSel 14.0

www.crgraph.de/WebDownload.htm



Häufigkeitsgruppen



Dichtheit	Temperatu	Druck[bar]	Druckplatt	Rohmaterial
dicht	100	5	Stahl	EPDM
Anz:220	Anz:75	Anz:45	Anz:10	ACSM
			Alu2	EPDM
			Anz:20	FPM
				HNBR
				ACSM
			Alu1	EPDM
			Anz:15	HNBR
				ACSM
		7	Stahl	EPDM
			Alu2	FPM
		6	Stahl	
		Anz:15	Anz:10	
			Alu2	
undicht	100	7	Alu2	
Anz:30	Anz:30	Anz:25	Alu1	
			Anz:25	
		6