ColourMatrix: White Paper

Finding relationship gold in big data mines

One of the most common user tasks when working with tabular data is identifying and quantifying correlations and associations. Fundamentally, if two measures are associated, we have the opportunity for relationships to exist, and insight to be garnered.

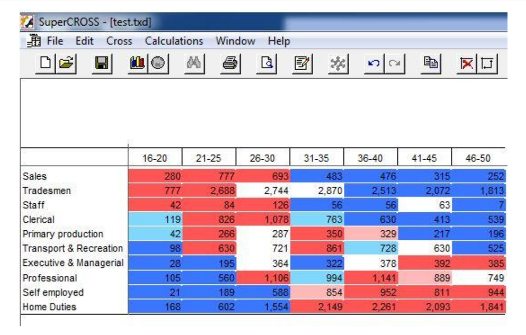
To find an association, we start by calculating what the data would look like in the absence of any pattern. That is, we determine the 'expected' number of counts in each cell, based on a simple assumption that the values are distributed homogeneously.

Associations will manifest as unexpected 'patterns' in the data: cells (or groups of cells) that are significantly higher (or lower) than expected.

There are a number of statistical tests that can be applied to tables to ask questions such as:

- Do the cell values differ from the homogeneous base case?
- Is there a pattern?
- How strong is the pattern?

For the end-user to glean understanding, we require a simple but robust mechanism to assess these questions. A positive result on a test for existence is merely based on the statistical likelihood of seeing such an association appear in the data by random chance. The strength of an association relates to the statistical effect size, and can be considered (to some extent) the predictability of the associative outcome.



	16-20	21-25	26-30	31-35	36-40	41-45	46-50
Sales	280	777	693	483	476	315	252
Tradesmen	777	2,688	2,744	2,870	2,513	2,072	1,813
Staff	42	84	126	56	56	63	7
Clerical	119	826	1,078	763	630	413	539
Primary production	42	266	287	350	329	217	196
Transport & Recreation	98	630	721	861	728	630	525
Executive & Managerial	28	195	364	322	378	392	385
Professional	105	560	1,106	994	1,141	889	749
Self employed	21	189	588	854	952	811	944
Home Duties	168	602	1,554	2,149	2,261	2,093	1,841

This white paper describes the new ColourMatrix feature available in SuperCROSS 9.0.

It outlines the methodology used to identify and cluster associations in dynamic cross tabulations.

ColourMatrix is a brand new SuperCROSS feature. It replaces ColourVIEW, which was available in previous SuperCROSS releases.

ColourVIEW was based on an **Expectation Ratio**: each cell value was divided by the expected cell value to determine how far it deviated from the expectation. However, this algorithm did not take into account the size of deviation from the expectation.

For example, having a value of 3 in a cell when expecting 2 (an overestimation of 1 unit) was given the same score as having an excess of 250 when expecting 500 (an over representation of 50% or an expectation ratio of 1.5). If there were an expectation of 750 and 500 were encountered then the expectation ratio would be 0.66. The relationship between these two differences is not readily apparent.

The new algorithm retains the simple, visual interface while being much easier to interpret and providing much more information about the size of any over or under representation.

It also conforms to simple, well documented statistical rules.

Algorithm Overview

The ColourMatrix algorithm is based on the χ^2 ("kye squared") test for homogeneity and independence. As this test is only uni-directional (it does not reflect under or over representation in the table), it is supplemented with a variant of analysis of standardised residuals (Haberman, 1973).

The ColourMatrix algorithm is based on the following assumptions:

- Tables consist of categorical/nominal (frequency) data in mutually exclusive categories.
- The data represents a random sample of *n* independent observations.
- The expected frequency in each cell is 5 or greater.

The third assumption is the subject of much debate in the community. How conservative (and stringent) should this assumption be?

Fundamentally, the association tests here rely on a smooth approximation to what is, in fact, a discrete distribution. Generally, an expectation of 5 or greater ensures that this approximation is acceptable. If cells have a lower expectation than 5, the chi-squared distribution of probabilities may not provide a truly accurate representation.

The algorithm first calculates the expected values for each cell (under the assumption of a completely homogeneous set).

It then undertakes three key statistical tests and provides the following feedback:

- Cells are coloured to indicate how close they are to the expected value.
- If the user selects the **Cluster** option, then the table is reordered to group together cells with similar levels of deviation.
- Textual results are provided for the tests for association existence (using the χ2 test) and the measure of association strength (using Cramer's ϕ', with Cohen's ***w*** as benchmarked values).

Notation

This white paper uses the following notation to convey techniques:

- *f* represents a cell count, with the dimensionality of the resident cube dictated by the number of subscripts. For example, f_{ij} comes from a 2-dimensional cross tabulation of *i* rows and *j* columns; whereas f_{ijk} has *i* rows, *j* columns, and *k* wafers.
- The marginals (totals and subtotals) are denoted by a dot in the relevant subscript. For example:
- $f_{.j}$ is the j^{th} column total.
- $f_{i..}$ is the wafer total of the i^{th} row.
- *f*… is the grand total of a 3 dimensional data cube.

Examples

Table 1: Example notation of an r x c table.

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	$f_{1.}$
f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	$f_{2.}$
f_{31}	f_{32}	f_{33}	f_{34}	f_{35}	$f_{3.}$
$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	$f_{.5}$	$f_{..}$

Table 2: Generalised r x c cross tabulation notation.

f_{ij}	$f_{i.}$
$f_{.j}$	$f_{..}$

Expected Cell Values

The Expected Value

Association is naturally expressed between only two variables. As such, examples are often based on a 2-dimensional cross table. Yet this ignores the natural hierarchy in modern data storage systems: the data cube. A data cube is essentially a series of cross-tables (wafers) layered on top of each other.

Describing the meaning of an association within a single layer is simple: an association is between the variable represented by the rows and the columns. However, describing the meaning of an association within a cube can be problematic.

It might be between the rows and columns with the wafer irrelevant. It might be between the column and the row/wafer. It could be weak, but across all three, or strong but only between two variables. As such, “strength” can lose meaning across cubes.

Standardised Residuals

Armed with an expectation value, we can generate a data cube of the same dimensions as the original cube, and propagate it with expressions for the deviation of the data, from the expectation. For each cell, the rule for this is

$$Z_{ij} = \frac{(f_{ij} - E_{ij})}{\sqrt{E_{ij}}}$$

This form is selected for a number of specific reasons:

- It will be standardised.
- It is negative for values that are less than expected, and positive for values that are over-represented.
- It is symmetric, and centred on zero.
- It is interpretable, in that the standardisation means that an independent selection of these values will conform to a Z-distribution.

Calculating the Expected Cell Value

Under an assumption of no association, either homogeneity or independence, the expected cell value (*E_{ij}*) is deduced from the relevant row, column, and wafer totals:

$$E_{ij} = \frac{f_{i\cdot}}{f_{\cdot\cdot}}\frac{f_{\cdot j}}{f_{\cdot\cdot}} f_{\cdot\cdot}$$

This relationship is expressed in a way to highlight its construction from simple joint probabilities. To extend to cubes (and dimensionally beyond), we can use the simple generalisation:

$$E_{ijk} = \frac{f_{i\cdot\cdot}}{f\ \ldots}\frac{f_{\cdot j\cdot}}{f\ \ldots}\frac{f_{\cdot\cdot k}}{f\ \ldots} f\ \ldots$$

The ColourMatrix algorithm uses this to generate the expected cell counts for a cube that will match the top two aggregated dimensions (i.e. grand total and wafer totals for cubes, grand total and row/column totals for cross tables will be satisfied identically).

Colouring the Cell

The standardised residual for each cell is used to colour the cell.

The colour is based on standard statistical interpretation of the number.

Typical values for the Z standardised variables are as follows:

- They will average 0
- The modal value will be 0.0
- 50% will be positive, 50% will be negative.
- ~70% will be between (-1,1)
- ~95% will be between (-2, 2)
- less than 1% will be outside (-2.6,2.6)

These relationships are well known, and can be derived directly via numerical function, or through the use of look-up tables. In the same way, the probability of exceeding values (p-values) can readily be derived across an entire cube.

Clustering the Table

The standardised residuals are configured such that any group should have a net sum of zero. Taking each wafer, row or column in turn, we can produce a measure of that sum as a metric. Those that are positive are generally over-represented in the set, while a negative score suggests under-representation. When a user selects the **Cluster** option, SuperCROSS automatically reorders the wafers, rows and columns to place the most positive scores in the upper left corner, generating further insights into the potential underlying association.

Testing the Association

Is there an Association?

If there is genuinely no association across the variables, then the standardised residuals can be assessed as a master set with predictable outcomes. The sum of their squares will conform to a χ^2 distribution with the relevant degrees of freedom:

$$X^2 = \sum \frac{\left(f_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum Z_{ij}^2$$

To test the existence of a possible association, we attempt to reject the possibility that a value could be as large as it is under chance alone. While we do not expect that every value in every cell will precisely match the expectation, each cell should be within the noise of the expectation. We also understand how an aggregate of squared Z distributed values should be distributed according to a χ^2 distribution. The critical value is the χ^2 value for the relevant degrees of freedom.

$$X_{table}^2 < X_{crit}^2 = X_{Dof}^2$$

The degrees of freedom relate to the construction of the cube and its ability to maintain the relevant and required marginal totals. If the amount of deviation we measure is less than this critical value, then there is no evidence to suggest that the table is significantly different to the one we would expect to see if there were no association between variables.

For a cross tabulation, the degrees of freedom is given by

$$DoF_{2D} = (r-1)(c-1) = rc - r - c + 1$$

where *r* is the number of rows, and *c* the number of columns. For a cube this becomes

$$DoF_{cube} = rcw - r - c - w + 2$$

where *w* is the number of wafers.

This can be generalised to N dimensions as

$$DoF_N = \Pi\, n_i - \Sigma n_i + N - 1$$

where *N* is the dimensionality of the cube, and n_i are the sub-dimension counts.

For the Z values, there are calculations and look-up tables for determining the critical values of χ^2, and reporting the likelihood of such a combination of residuals arising by chance (p–values).

Given a statistical likelihood that a χ^2 value should arise, if it is deemed likely that there is a true association within the cube or table then the standardised residuals guide the analysts in finding cause for why (or where) we were compelled to reject the hypothesis that there was no association within the table. The largest absolute residual values contribute the most to the finding of an association.

How Strong is the Association?

Modern data tools are designed to move and display large volumes of data. Consequently, they are capable of resolving very small effects to great statistical significance. It is therefore important to determine the potential size or strength of any detected association.

The use of a χ^2 value needs in some way to be corrected for the volume of data involved, determining strength. When there are only two variables at hand, the greatest association can be easily visualised (it is when all the data resides on the diagonal). We can go further than this, and deduce the maximum value that a χ^2 statistic can be for a given table. Scaling our determined metric by this value gives us a standardised metric of association strength.

This is known as Cramer's ϕ' (or ϕ_C)

$$\Phi_c = \sqrt{\frac{X^2}{n(k-1)}}$$

where *k* is the smaller of the number of columns or rows and *n* is the total, independent contributor count. This value depends on the size and shape of the original table. To some extent it can be standardised by reporting Cohen's ***w***.

$$w = \Phi_c \sqrt{(k-1)}$$

There is an accepted range for ***w***, as follows

Effect Size	Range (w)
Small	0.1≤w<0.3
Medium	0.3≤w<0.5
Large	0.5≤w

Healthcare Case Study

To demonstrate the power of the ColourMatrix feature, here is a sample cross tabulation of healthcare data, based on a table of patients who present with a single disease, tabulated against their age group.

Space-Time Research
Healthcare
Table 1
Age by Disease
for Patients

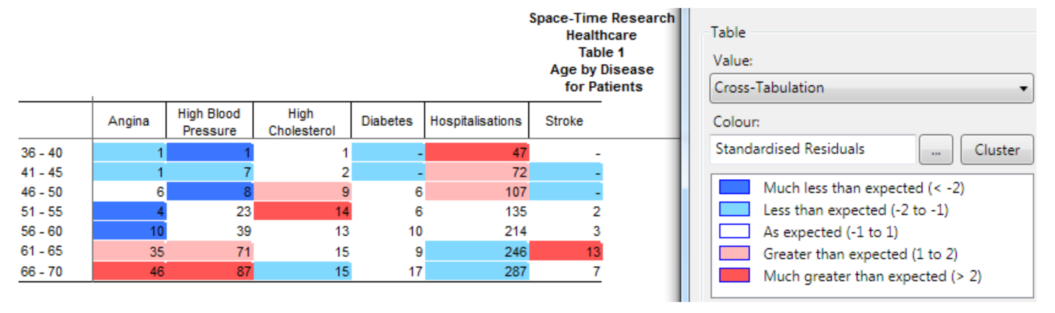
	Angina	High Blood Pressure	High Cholesterol	Diabetes	Hospitalisations	Stroke	Grand Total
36 - 40	1	1	1	-	47	-	50
41 - 45	1	7	2	-	72	-	82
46 - 50	6	8	9	6	107	-	136
51 - 55	4	23	14	6	135	2	184
56 - 60	10	39	13	10	214	3	289
61 - 65	35	71	15	9	246	13	389
66 - 70	46	87	15	17	287	7	459
Grand Total	103	236	69	48	1,108	25	1,589

Using the formulas above, it is possible to deduce that an association exists to better than 99% confidence ($\chi^2 = 96.98 > \chi^2_{crit} = 43.77$).

Based on this calculation, if there were no association then 99% of all permutations of this particular table would have a measured χ^2 value of less than 43.77. Therefore we conclude that, if there were no association, it is exceedingly unlikely we would see such a table.

The ϕ' value is 0.11, which for a table such as this corresponds to a “small” association (Cohen’s ***w*** = 0.25). So we can be confident that a small magnitude association exists.

Instead of manually calculating these values, SuperCROSS users can activate ColourMatrix to see a visual representation:



Space-Time Research
Healthcare
Table 1
Age by Disease
for Patients

	Angina	High Blood Pressure	High Cholesterol	Diabetes	Hospitalisations	Stroke
36 - 40	1	1	1	-	47	-
41 - 45	1	7	2	-	72	-
46 - 50	6	8	9	6	107	-
51 - 55	4	23	14	6	135	2
56 - 60	10	39	13	10	214	3
61 - 65	35	71	15	9	246	13
66 - 70	46	87	15	17	287	7

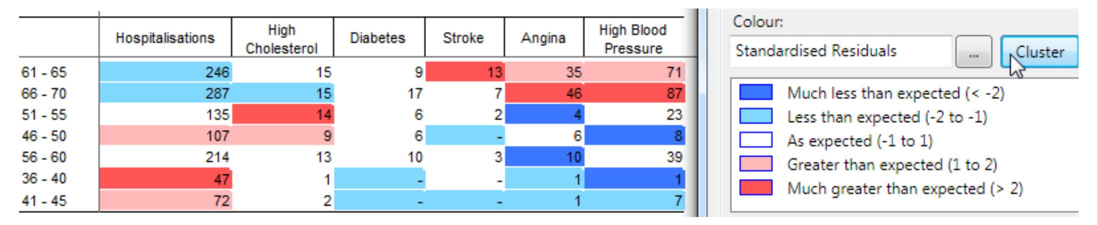
Table
Value: Cross-Tabulation
Colour: Standardised Residuals

- 🟦 Much less than expected (< -2)
- 🟦 Less than expected (-2 to -1)
- □ As expected (-1 to 1)
- 🟥 Greater than expected (1 to 2)
- 🟥 Much greater than expected (> 2)

Here, the darkest red and blue colours represent over and under representation outside the 95% confidence. Note how they are not related to their absolute values. For example, we can see that younger patients are over represented in hospitalisations, that angina appears to be loosely correlated with aging, as is high blood pressure. There is an unusual spike in stroke patients at age 61-65.

Users can also click the **Cluster** option. SuperCROSS automatically rearranges the table to highlight potential groups to the analyst.

In this example, high blood pressure has been grouped with angina and stroke. In these diseases, the under 60 year olds are underrepresented, and patients ages 60+ are over represented.

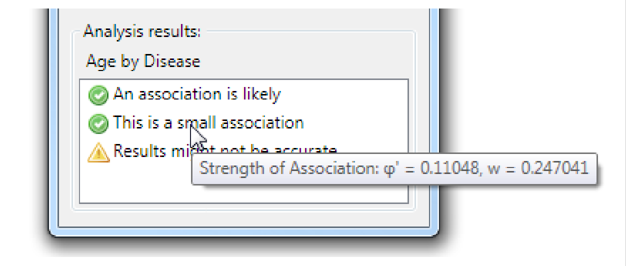


	Hospitalisations	High Cholesterol	Diabetes	Stroke	Angina	High Blood Pressure
61 - 65	246	15	9	13	35	71
66 - 70	287	15	17	7	46	87
51 - 55	135	14	6	2	4	23
46 - 50	107	9	6	-	6	8
56 - 60	214	13	10	3	10	39
36 - 40	47	1	-	-	1	1
41 - 45	72	2	-	-	1	7

Colour: Standardised Residuals

- 🟦 Much less than expected (< -2)
- 🟦 Less than expected (-2 to -1)
- □ As expected (-1 to 1)
- 🟥 Greater than expected (1 to 2)
- 🟥 Much greater than expected (> 2)

If we look at the age groups there are 2 or 3 clusters appearing. 36-40 year olds are underrepresented in angina and high blood pressure, yet over represented in hospitalisations. Similarly, 61-70 year olds are underrepresented in hospitalisations and high cholesterol, but over represented for stroke, angina and high blood pressure. There is an interesting pocket of 46-50 year olds who exhibit high cholesterol, which appears to rectify with age (or treatment).



Analysis results:
Age by Disease
✓ An association is likely
✓ This is a small association
⚠ Results might not be accurate
Strength of Association: φ' = 0.11048, w = 0.247041

SuperCROSS also displays information about the strength of the association. In this case an association is likely, but it is a small association: