Cognalysis MultiRate™
Version 2 Release Notes

Major new functionality in version 2:

- **Custom Geographic Territories** –Variable Gradient Geospatial Smoothing can now be applied to user-defined geographic territories. This is a powerful tool for measuring the predictive power of geographical location. The user can provide territory definitions, or MultiRate can group records together for you, using latitude and longitude, anywhere in the world.

- **Interaction Effects** – Cognalysis MultiRate™ now has the ability to scan combinations of predictive characteristics and point out potential interaction effects based on residual errors. It then aids the user in incorporating these interaction effects explicitly into their models.

- **t-statistic Credibility Option** – This option assigns credibility to observed results based on the consistency of the observations. The more consistent the relationship is across the observed data, the more credibility is assigned to that specific relationship.

Additional improvements in version 2:

- **Drill down in the Results Inspection** – By double-clicking on individual factors in the inspection panel, the user can now see much greater detail about the underlying calculation, giving greater insight into model results.

- **Convergence Acceleration** – Under certain circumstances the speed of model convergence can be improved by selecting this option.

- **Print Enhancements** – The report now contains a table of contents and the assumptions/field selections

- **SQL Permission Gap Identification** – MultiRate now identifies when permissions are missing on a particular SQL database or server, and notifies the user, so that these can be requested from the database administrator.

- **Faster opening** – Users should be able to notice significant improvement in the speed at which the software opens.

- Miscellaneous process improvements, bug fixes, and pre-analysis data validation
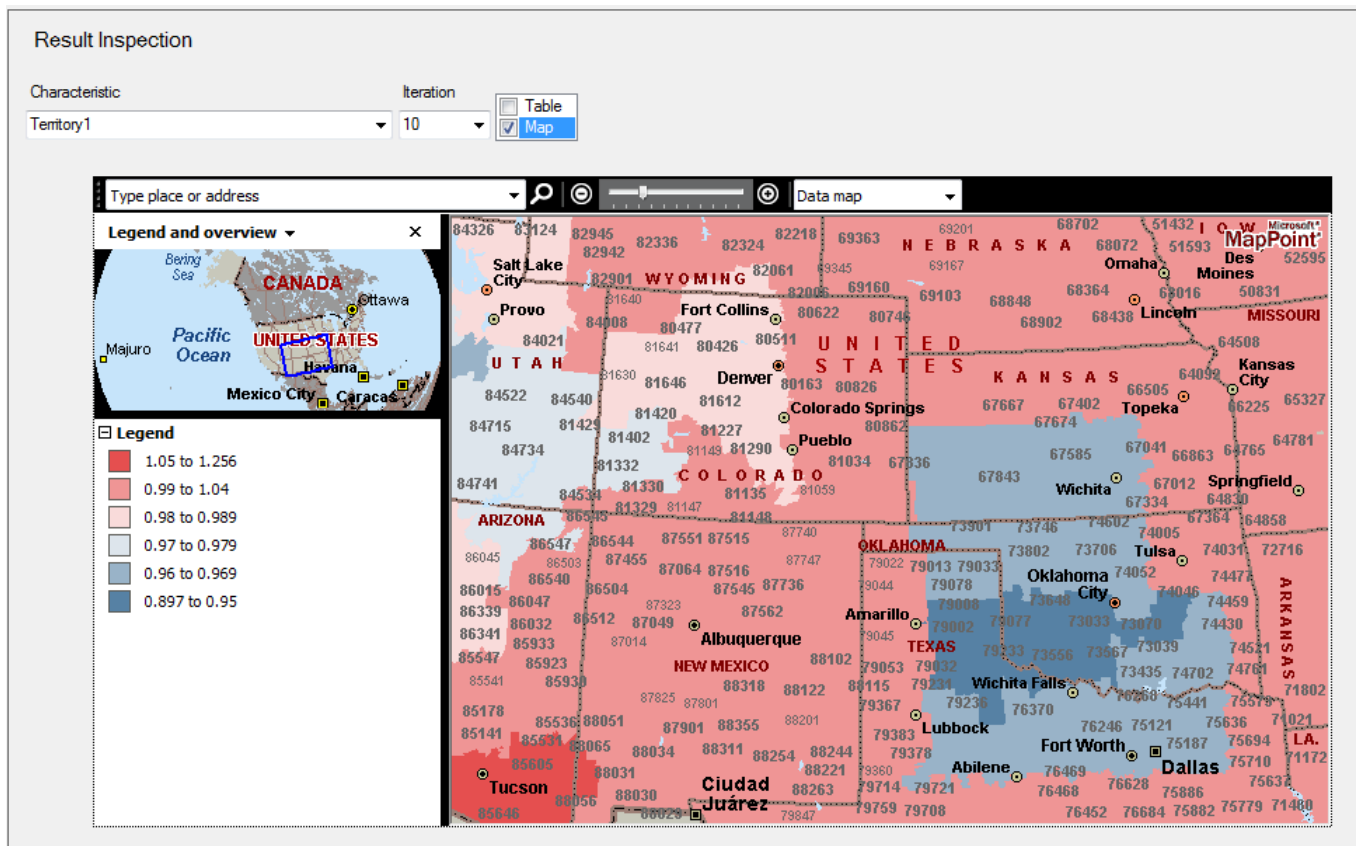
Technical Notes

Custom Geographic

The Custom Geographic data type is used for records that have been grouped into custom territories. This allows data from each territory to be credibility weighted with data from nearby territories, using Variable Gradient Geospatial Smoothing.

MultiRate can group records using latitude and longitude data (this can be done with the Create New Territory Column tool) or they can be grouped into territories by the user (see User Defined Territories).

If the territories are created by MultiRate or the user creates a Map Data Table, an approximate mapping of the the resulting Custom Geographic factors can be viewed in the Result Inspection Panel if the computer has Microsoft MapPoint installed.

Note: This map is created by displaying the factor for each Territory in all of the ZIP Codes whose centroid is contained in that Territory. As a result, the outlines of territories on the map will be approximate. An example of this can be seen below.

Create New Territory Column

Creating a new territory column allows the user to include latitude and longitude data in an analysis. MultiRate will group the data points into territories, attempting to assign a roughly equal amount of exposure to each territory (more information about this process can be found below). This will allow data from each territory to be credibility weighted with data from nearby territories, using Variable Gradient Geospatial Smoothing.

In order to create a territory column, first select the Training Data table containing latitude and longitude.

Next the points must be assigned to territories. This will be done by adding a territory column to the Training Table and information about the territories will be stored in the Custom Geographic Tables. To do this, select **Tools/Custom Geographic/Create New Territory Column**.

When the window below appears, select the columns containing the latitude, longitude, and exposure data. Latitude and longitude must be in decimal degrees format. Also, enter the desired number of territories to be created. As a result of the grouping algorithm, there will likely be a few more territories created (more information about this process can be found below).
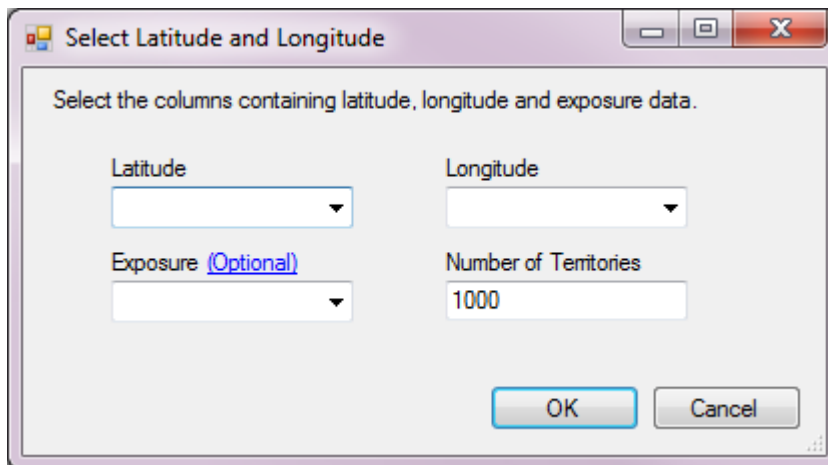
Note: If no exposure is selected in the "Select Latitude and Longitude" window, all points will be treated as if they have an exposure of 1, resulting in a roughly equal number of points in each territory.

Next the user will be prompted to enter a name for the column to be created. This name will be used for the column created in the Training Table, as well as the three Custom Geographic Tables created to store the territory information. To use this newly created field in an analysis, select the new territory column in the Select Fields window and Custom Geographic as the data type.

Technical Notes

Territory Grouping Process:

The first step for grouping the records into territories is to map each latitude and longitude point onto an XY grid to allow for calculations to be performed. This process sets the center of all of the latitude and longitude points as (0, 0) and then maps each point relative to that origin, with meters as the unit. This is the basis for the XY points that appear in the Custom Geographic tables. While this mapping is very useful for doing calculations concerning the relative position of the points, it is important to note that since the origin will be different for every unique dataset, these XY points are not useful for comparing points from different Training Files. Use latitude and longitude for that purpose.

Next the "Northwestness" of each point is calculated by taking each point's Y-coordinate and subtracting its X-coordinate (Y-X). (The green lines on the graph below represent sets of points with the same "Northwestness", with "Northwestness" increasing in the direction of the arrow.) The most northwest point is selected, and larger and larger squares are drawn around it (keeping the selected point in the center) until it contains $\frac{1}{\# \ of \ desired \ territories}$ of the exposure (i.e. 1/1000 of the total exposure if the default of 1000 territories was selected). Once this exposure threshold has been reached, all of the points in the box are placed in territory 1.
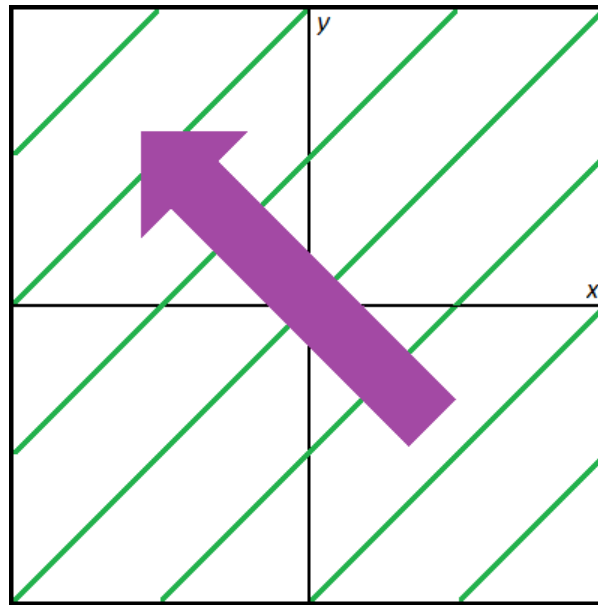

This process is repeated over and over, each time throwing out the points that have already been assigned to a territory and decreasing the number of territories in the denominator of the threshold by 1. So after creating territory 1, the most northwest point not in territory 1 will be selected, and larger and larger squares are drawn around it until it contains $\frac{1}{\# \ of \ desired \ territories \ remaining}$ of the remaining exposure (i.e. 1/999 of the total exposure not in territory 1 if the default of 1000 territories was selected). The process continues until all points have been assigned a territory.

Technical Notes

Note: This method should result in relatively equal amounts of exposure in each territory. However, this cannot always be achieved when there are a few points with very high exposures or when a territory must be divided (this is discussed in the next paragraph). This can result in territories with significantly more or less exposure than the threshold.

It is important to note that even though the square is drawn all the way around the selected point, there will never be any points located northwest of the point itself, since it was selected precisely because it was the most northwest. The majority of the territory will generally be southeast of the selected point, but there will often be points either north or west (but never both north and west) of the selected point included. For example, see box 4 in the diagram below where point A is the selected point.
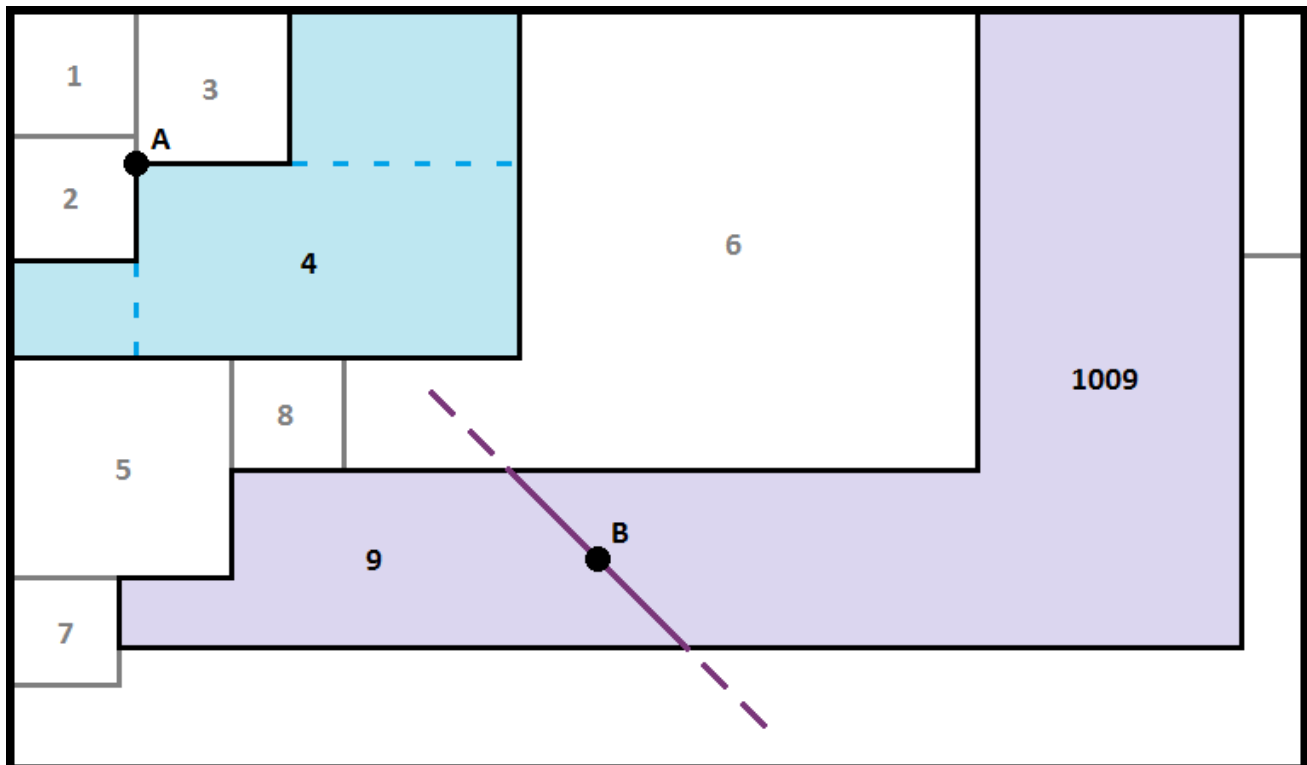
As a result of the way that the territories are being drawn, there is the potential for some of the territories to be shaped like a backwards "L", usually with high exposure areas in the ends of the "L". These territories are identified and then divided into two territories by drawing a line through the exposure weighted centroid, running 45° from northwest to southeast. The southwestern portion of the divided territory is assigned the same territory number as the original combined territory, and the northeastern portion is assigned the territory number of the original territory plus the number of desired territories. For example, in the diagram above territory 9 was divided into two territories (9 & 9 + 1000 = 1009, assuming the default of 1000 territories was selected) by drawing a 45° line through point B, the exposure weighted centroid.

After all of the records have been assigned to a territory, a column is added to the Training Data containing each record's territory number. Also, three Custom Geographic Tables are

Technical Notes

created in SQL, containing information about each territory, mapping information, and parameters used in the calculations.

Custom Geographic Tables

This page discusses the Custom Geographic Tables created when MultiRate is used to generate territory definitions from latitude/longitude data. If user defined territories are being used instead, see the User Defined Territories page.

These three tables are created when MultiRate Creates a Territory Column. The first table created is the Territory Table, which contains information about each territory, such as exposure, centroid, area, and other information used for calculations. The second table created is the Map Data Table, which contains a list of ZIP codes in the area of the original data and the territory which contains their centroids, and is used only for mapping Custom Geographic data. The final table created is the Parameters Table, which contains parameters used in the calculations.

Territory Table

Table Name: *TrainingTable_TerritoryColumnName*

Territory Table Layout:

Field Name	Data Type	Explanation
TerritoryColumnName	int	The territory number.
Exposure	float	The total exposure contained in each territory.
Centroid_X	int	The X value of the centroid of the territory. (Different LatLon to XY mapping used for each dataset.)
Centroid_Y	int	The Y value of the centroid of the territory. (Different LatLon to XY mapping used for each dataset.)
Centroid_Lat	numeric	The latitude of the centroid of the territory.
Centroid_Lon	numeric	The longitude of the centroid of the territory.
Divide	int	Number code indicating if box has been divide or not (used for intermediate calculations).
Orig_Centroid_X	int	The X value of the centroid before division (used for intermediate calculations).
Orig_Centroid_Y	int	The Y value of the centroid before division (used for intermediate calculations).
Orig_maxX	int	The maximum X value of the territory before division (used for intermediate calculations).
Orig_minY	int	The minimum Y value of the territory before division (used for intermediate calculations).
Area	numeric	The total area of the territory.

Map Data Table

Technical Notes

Table Name: *TrainingTable_TerritoryColumnName_*mapData

Map Data Table Layout:

Field Name	Data Type	Explanation
ZIP	varchar	The ZIP code.
TerritoryColumnName	int	The territory which contains the centroid of the ZIP code.

Parameters Table

Table Name: *TrainingTable_TerritoryColumnName_*Parameters

Parameters Table Layout:

Field Name	Data Type	Explanation
Territories	int	The number of territories created (before divisions).
LatCol	nvarchar	The name of the latitude column in the Training Data.
LonCol	nvarchar	The name of the longitude column in the Training Data.
onIntlDate	bit	A 1 or 0 indicating if the data points are near the International Dateline.
Lat0	numeric	The center of the range of latitude in the original data (used to convert LatLon to XY).
Lon0	numeric	The center of the range of longitude in the original data (used to convert LatLon to XY).
Lat0z	numeric	The center of the range of latitude after translating original data (used to convert LatLon to XY).
Lon0z	numeric	The center of the range of longitude after translating original data (used to convert LatLon to XY).
Lat1z	numeric	Standard parallel (used to convert LatLon to XY).
Lat2z	numeric	Standard parallel (used to convert LatLon to XY).

Technical Notes

User Defined Territories

It is possible for MultiRate to analyze data that has been grouped into territories by the user. This can be done by providing a few key pieces of information to MultiRate about these user defined territories. This data needs to be placed into the SQL database:

- **Territory Column:** A territory column must be added to the Training Table. This column indicates the territory for each record.

- **Territory Table:** A table must be added which contains a record for each territory and the XY centroid of that territory. As many or as few territories can be created as desired. MultiRate is expecting that these (X,Y) coordinates are map projections stated in meters. While it is not necessary that these coordinates are _actually_ in meters, the user must be aware that the Variable Gradient Geospatial Smoothing will be governed by radii that are stated in multiples of a thousand to whatever is used (normally this would be kilometers).

- **Map Data Table (Optional):** A table may be added if the user wants to see the territory factors on a map. It must include a list of ZIP Codes to be mapped and the territory where each ZIP Code is located. For most territory definitions, this will be an approximate mapping based on the ZIP Codes included in the table, but still can be very helpful for visualizing geographic factors. For example, when MultiRate is used to generate territories from latitude and longitude data (see Create New Territory Column), a list of the ZIP Codes in the area of the original data are assigned territories based on their centroids.

Naming Conventions:
The naming of these tables and columns must be exact. Follow the Table and Field Names shown in the table diagrams below.

Territory Table

Table Name: _TrainingTable_TerritoryColumnName_

Territory Table Layout:

Field Name	Data Type	Explanation
TerritoryColumnName	int	The territory number.
Centroid_X	int	The X value of the centroid of the territory.
Centroid_Y	int	The Y value of the centroid of the territory.

Map Data Table

Table Name: _TrainingTable_TerritoryColumnName_mapData

Technical Notes

Map Data Table Layout:

Field Name	Data Type	Explanation
ZIP	varchar	The ZIP code.
TerritoryColumnName	int	The territory which contains the centroid of the ZIP code.

For example, to add a territory column called 'Region' to a Training Table named 'InputData', do the following:

- Add a column named 'Region' to the 'InputData' table.
- Create a table named 'InputData_Region' containing the columns 'Region', 'Centroid_X', and 'Centroid_Y'.
- (Optional) Create a table named 'InputData_Region_mapData' containing the columns 'ZIP' and 'Region'.

Technical Notes

Interaction Effects

Relationships between the target variable and the characteristics are often not as simple as multiplicative factor relationships. Sometimes the relationship of the target variable to a specific characteristic depends on the value of a separate characteristic. This utility allows the user to investigate interaction effects and provides tools for handling them. For more information on these tools, see the Add Interaction Effect Fields screen.

After completing an analysis run, to investigate interactions, select the yellow braid button on the menu bar, .

A pop-up window will then open, allowing you to view:

- Interaction Statistic
- Actual / Modeled
- Exposure
- Modeled Target
- Residual

Interaction Statistics

The interaction statistic is designed to identify situations where consistent deviations from the strict multiplicative relationship are observed. It relies on the deviation of actual from modeled results, the level of credibility associated, and the scale of the observation.

The calculation is performed at the intersection of a bin for a specific characteristic with a bin of another characteristic. It is then aggregated at various levels depending on the view selected by the user. This summarized statistic can be viewed at three levels.

1. Each Characteristic vs. each other Characteristic
2. One Characteristic's bins vs. each other Characteristic
3. One Characteristic's bins vs. another Characteristic's bins

The interaction statistics are displayed within the matrix defined. To define the matrix columns and rows, select from the drop downs boxes in the top left hand corner of the Interaction Effects window.

For example, select 'All Characteristics' in the drop-down box labeled 'Row'. Each of the selected characteristics from the model will become the labels down the left hand side of the matrix. Or select a particular characteristic from the analysis from the 'Column' drop-down box, the different bins that characteristic takes will then become the labels across the top of the matrix.

Drilling into two specific characteristics and their bins can also be done by double clicking on the number in the cell that intersects the two desired characteristics.

How the Interaction Statistic is calculated:

Technical Notes

$$Interaction\ Statistic_{ij} = \frac{|Residual_{ij}| * \left(\frac{Exposure_{ij}}{Exposure_{ij} + Credibility\ Standard}\right) * \sqrt{Modeled\ Target_{ij}}}{\Sigma\ Modeled\ Target_{ij}}$$

$$Residual_{ij} = Actual\ Target_{ij} - \left[(Modeled\ Target_{ij}) * \left(\frac{\Sigma\ Actual\ Target_i}{\Sigma\ Modeled\ Target_i}\right) * \left(\frac{\Sigma\ Actual\ Target_j}{\Sigma\ Modeled\ Target_j}\right)\right]$$

The final two terms in the above equation are necessary to adjust the modeled amount to full Credibility. Otherwise what we are observing is often dominated by the overall difference between modeled and actual that results from assigning less than 100% credibility to what is observed. This is not an interaction effect, and so we attempt to exclude it here.

** The subscript 'i' describes the data for the records with the value 'i' for the row characteristic
** The subscript 'j' describes the data for the records with the value 'j' for the column characteristic
** The subscript 'ij' describes the data for the records with the value 'i' for the row characteristic and the value 'j' for the column characteristic.

Using an example:

		NULL	A	B
	NULL			
▶ 0.11740205066533	Bin 1		0.141	0.194
0.22201295550131	Bin 2		0.080	0.101
0.32799714069210	Bin 3		0.064	0.080
0.44193827289015	Bin 4		0.024	0.034
0.55642611969607	Bin 5		0.004	0.005
0.66988105650529	Bin 6		0.047	0.056
0.77975903689963	Bin 7		0.108	0.151
0.89098864531395	Bin 8		0.086	0.117
0.99987146438871	Bin 9		0.119	0.154

Interaction Effects

Char3 — Row
Char1 — Column

- Interaction Statistic
- Actual/Modeled
- Exposure
- Modeled Target
- Residual

Technical Notes

(The values of the grouped characteristic Char3 have been labeled with 'Bin 1', 'Bin 2', etc. for reference purposes.)

Consider specifically, the cell highlighted in blue, where:

Char1 = 'A'
Char3 = 'Bin 1'

In the formulation of the two equations above, the subscript 'i' refers to 'Bin 1', and 'j' refers to 'A'.

$$Interaction\ Statistic_{Bin\ 1:A} = \frac{|Residual_{Bin\ 1:A}| * \left(\frac{Exposure_{Bin\ 1:A}}{Exposure_{Bin\ 1:A} + Credibility\ Standard}\right) * \sqrt{Modeled\ Target_{Bin\ 1:A}}}{\sum Modeled\ Target_{Bin\ 1:A}}$$

$$Residual_{Bin\ 1:A} = Actual\ Target_{Bin\ 1:A} - \left[(Modeled\ Target_{Bin\ 1:A}) * \left(\frac{\sum Actual\ Target_{Bin\ 1}}{\sum Modeled\ Target_{Bin\ 1}}\right) * \left(\frac{\sum Actual\ Target_A}{\sum Modeled\ Target_A}\right)\right]$$

$Exposure_{Bin\ 1:A}$ is the total exposure of the records where Char 3 = 'Bin 1' and Char 1 = 'A'
$Modeled\ Target_{Bin\ 1:A}$ is the modeled target of the records where Char 3 = 'Bin 1' and Char 1 = 'A'
$Actual\ Target_{Bin\ 1}$ is the actual target of the records where Char 3 = 'Bin 1' and Char 1 = 'A'

A larger interaction statistic suggests possible interaction between the two bins or characteristics.

The following four data computations can only be calculated in a 'Bins v. Bins' manner.

Actual/Modeled

This matrix displays the ratio $\left(\frac{\sum Actual\ Target_{ij}}{\sum Modeled\ Target_{ij}}\right)$.

Exposure

This matrix displays the breakdown of the total Exposure in the data set. Each cell represents the sum of the Exposure for all records that have the combination of bins specified by the row and column. For example, referring to the screen shot below, the value '1,303' represents the sum of the Exposure for records with the value 'A' for the Char1 characteristic and the value 'X' for the Char2 characteristic.

Technical Notes

Modeled Target

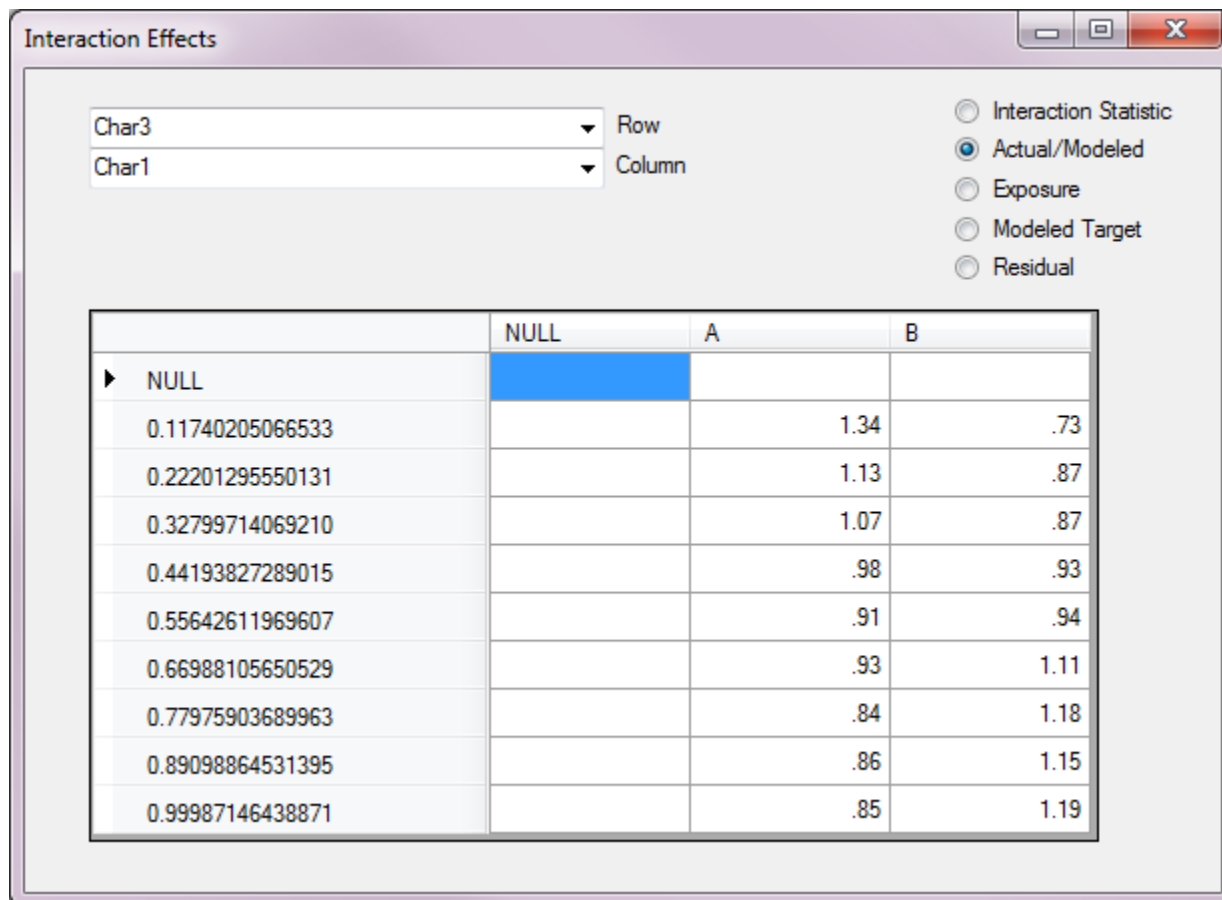This matrix displays, $\sum Modeled\ Target_{ij}$.

Residual

$$Residual_{ij} = Actual\ Target_{ij} - \left[(Modeled\ Target_{ij}) * \left(\frac{\sum Actual\ Target_i}{\sum Modeled\ Target_i} \right) * \left(\frac{\sum Actual\ Target_j}{\sum Modeled\ Target_j} \right) \right]$$

Interpreting the Statistics

To analyze the Actual/Modeled screen, look for patterns within the matrix for numeric Grouped Characteristics as well as looking for extreme values for categorical characteristics (Generic Characteristic).

Technical Notes

Interaction Effects			
Char3 ▾ Row			
Char1 ▾ Column			

○ Interaction Statistic
● Actual/Modeled
○ Exposure
○ Modeled Target
○ Residual

	NULL	A	B
▶ NULL			
0.11740205066533		1.34	.73
0.22201295550131		1.13	.87
0.32799714069210		1.07	.87
0.44193827289015		.98	.93
0.55642611969607		.91	.94
0.66988105650529		.93	1.11
0.77975903689963		.84	1.18
0.89098864531395		.86	1.15
0.99987146438871		.85	1.19

Technical Notes

t-Statistic Credibility Method

The t-Statistic Method assigns credibility based on the consistency of observed relationships and the selected Confidence Level, and is calculated as follows:

First, all of the records are sorted by the ratio of Target to Modified Exposure, smallest to largest. Records that take the same value for their Target (many values at zero, for example) are partitioned. Then, the partitions are assigned Z-scores from a standard normal distribution based on the amount of exposure in each partition, using the formula:

$$Z_i = \Phi^{-1}\left(\frac{\sum_{j=1}^{i} Exposure_j}{\sum_{j=1}^{n} Exposure_j}\right)$$

where Φ represents the CDF of the standard normal distribution.

Once these partitions have been calculated, the average values of Z and Z^2 over the partition are calculated using the formulae below, which can be derived by integration over the standard normal CDF[1]:

$$\bar{Z}_i = \frac{\left(\sum_{j=1}^{n} Exposure_j\right)\left(e^{-.5Z_{i-1}^2} - e^{-.5Z_i^2}\right)}{Exposure_i\sqrt{2\pi}}$$

$$\overline{Z_i^2} = 1 + \frac{\left(\sum_{j=1}^{n} Exposure_j\right)\left(Z_{i-1}e^{-.5Z_{i-1}^2} - Z_i e^{-.5Z_i^2}\right)}{Exposure_i\sqrt{2\pi}}$$

All *records* within a common partition (i.e. the same ratio of Target to Modified Exposure) are assigned these partition values of $\bar{Z}_i$ and $\overline{Z_i^2}$.

Next these values are averaged across records at the level at which a factor is being calculated. In the case where smoothing is not being applied, this means all records within the bin are averaged, weighted by exposure. In the case where smoothing is being applied, all records being used to calculate the factor are averaged, weighting by both the Exposure and the smoothing weights[2]. The calculated values of $\bar{Z}$ and $\overline{Z^2}$ corresponding to a specific bin are then used to calculate the standard deviation and standard error of $\bar{Z}$ for the bin, and finally a t-statistic. The formulae for these calculations are as follows:

$$SD(\bar{Z}) = \sqrt{\overline{Z^2} - (\bar{Z})^2}$$

Technical Notes

$$s.e.(\bar{Z}) = \frac{SD(\bar{Z})}{\sqrt{n}}$$

where n is the quasi-count of the number of observations in the sample, defined as:

$$quasi\text{-}count = min(\# \, of \, records, \frac{Exposure_i}{\sum_{j=1}^{n} Exposure_j} * total \, \# \, of \, records)$$

$$t = \frac{\bar{Z}}{s.e.(\bar{Z})}$$

Next, a critical t value, t_0, is calculated from an inverse t distribution, using the selected Confidence Level and degrees of freedom equal to the quasi-count minus 1:

$$df = quasi\text{-}count - 1$$

The credibility is then calculated as the percentage by which the the absolute value of the t statistic exceeds the critical t, as seen below:

$$Credibility = \frac{|t| - t_0}{|t|}$$

Finally, if the observed relationship and the z score have the same direction (i.e. z<0, raw factor<1 or z>0 and raw factor>1), the credibility above is used. However, if they have opposite direction, a credibility of 0 is assigned.

[1] For the first and last partitions, Z_{i-1} and Z_i have not been defined above. For the first partition, $Z_{i-1} = -\infty$, which will result in the first exponential term going to 0, and for the last partition, $Z_i = \infty$, which will result in the second exponential term going to 0.

[2] In cases where smoothing is being applied, the formula for the averages, weighted by both Exposure and the smoothing weights, is:

$$\bar{Z} = \frac{\sum_{i=1}^{n}(\bar{Z}_i * Exposure_i * Smooth \, Weight_i)}{\sum_{i=1}^{n}(Exposure_i * Smooth \, Weight_i)}$$

For $\overline{Z^2}$, simply replace $\bar{Z}_i$ with $\overline{Z_i^2}$.

Technical Notes

In the example below, the average for bin 1 is being calculated, and the Variable Gradient Radius used to calculate the smoothing weights is 1.

Key ID	Bin	$\overline{Z}$	$\overline{Z^2}$	Exposure	Smoothing Weight
1	2	-0.5	0.3	1	0.5
2	3	0	0.1	2	0.25
3	1	0.2	0.2	3	1
4	2	0.7	0.6	4	0.5
5	1	1.3	2.1	5	1

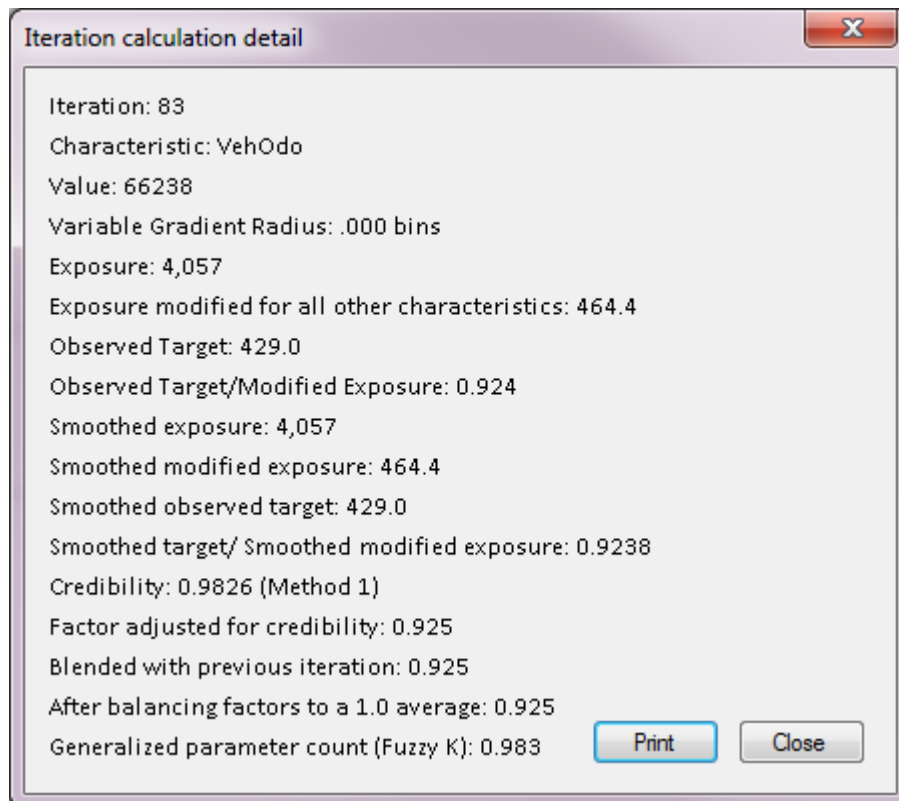$$\overline{Z} = \frac{(-0.5 * 1 * 0.5) + (0 * 2 * 0.25) + (0.2 * 3 * 1) + (0.7 * 4 * 0.5) + (1.3 * 5 * 1)}{(1 * 0.5) + (2 * 0.25) + (3 * 1) + (4 * 0.5) + (5 * 1)}$$

$$\overline{Z^2} = \frac{(0.3 * 1 * 0.5) + (0.1 * 2 * 0.25) + (0.2 * 3 * 1) + (0.6 * 4 * 0.5) + (2.1 * 5 * 1)}{(1 * 0.5) + (2 * 0.25) + (3 * 1) + (4 * 0.5) + (5 * 1)}$$
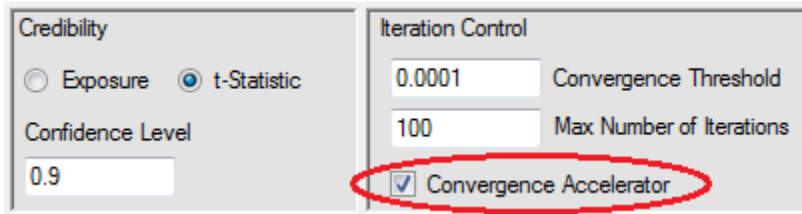
Technical Notes

Results Inspection Drill Down

From table view, double click on any of the rows to get more information about how the factor for a given bin was calculated:

Iteration calculation detail

Iteration: 83
Characteristic: VehOdo
Value: 66238
Variable Gradient Radius: .000 bins
Exposure: 4,057
Exposure modified for all other characteristics: 464.4
Observed Target: 429.0
Observed Target/Modified Exposure: 0.924
Smoothed exposure: 4,057
Smoothed modified exposure: 464.4
Smoothed observed target: 429.0
Smoothed target/ Smoothed modified exposure: 0.9238
Credibility: 0.9826 (Method 1)
Factor adjusted for credibility: 0.925
Blended with previous iteration: 0.925
After balancing factors to a 1.0 average: 0.925
Generalized parameter count (Fuzzy K): 0.983

Print   Close

Technical Notes

Convergence Accelerator

The Convergence Accelerator is an option that may help the model reach convergence faster and can be useful where analyses take a long time to run. To select this option, check the box in the Iteration Control section at the bottom of the Data and Analysis Control Panel.

MultiRate always performs Factor Blending between iterations in order to reduce the potential for oscillation, particularly with highly correlated characteristics. A natural effect of Blending is that it can slow the movement toward the final converged relationships.

Convergence acceleration, if selected by the user, attempts to measure consistent movement in the same direction by factors across iterations. If this is seen, the software applies an increasingly aggressive (faster) blending factor as the iterations progress.

However, this acceleration can result in some situations of non-convergence, particularly in the case for which Blending is designed (i.e. when there are several highly correlated characteristics in the analysis). If analyses do not converge while using the convergence accelerator, unselecting the Convergence Accelerator option is usually appropriate.