

In This Issue

- The Microsoft Corner:
*Excel Data Mining
Add-In*

Contact Us

stevemong@
marketconsulteks.com

Data Mining Utilities: *Excel Add-Ins*¹

By Steve Mong

This article begins the first in a series devoted to SQL Server data mining utilities. I consider these utilities a useful, fun and tangible means of getting introduced to data mining. SQL Server 2000 will be our immediate focus, as many of you may not upgrade to 2005 until after the first patch release. Rest assured, however, that the rich new series of features available in 2005 will be reviewed later.

We'll gently test the waters by first examining the Excel Data Mining Add-Ins. Though provided by Microsoft, they're unsupported. These utilities—which eventually take the form of four Excel menu items—can be downloaded on the second page of the following URL:

<http://www.sqlserverdatamining.com/DMCommunity/SQLServer2000/default.aspx>.

This download requires site registration.

Three files will be extracted from the ZIP executable you receive:

- Data Mining Add Ins.xla
- Census Test.xls
- Census Sales Data.xls

As with any Excel Add-In, you'll need to Open Excel, browse for the ".XLA" file you downloaded, and load the "Data Mining Add Ins", per Figure 1.

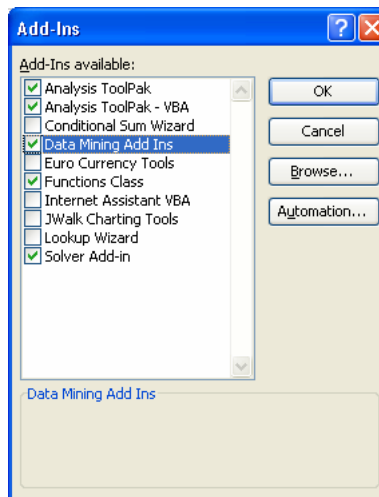


Figure 1: Loading Add-In

¹ This article was first published by the DW/BI SIG of the Professional Association for SQL Server in 2005.

Four new menu items appear under "Tools". As described by Microsoft in a 2003 posting:

- **Similarity Analysis** uses the clustering algorithm to group your Excel rows by self-similarity and produces a chart showing the population of each group. The description of each group is taken from the *first* three clauses in the description taken from the model content.
- **Primary Indicator** displays a message box indicating the column that is most predictive of the selected column. It uses the decision tree algorithm to create a model and then determines the column by navigating the model content.
- **Auto Fill** uses a decision tree to automatically fill missing values in a spreadsheet.
- **Anomaly Detection** flags as anomalies any cell not self-predicted by the decision tree algorithm

Similarity Analysis

Jamie MacLennan, Microsoft's development lead for SQL Server's data mining engine, has suggested you start by selecting all the data in the "**Changes**" sheet within "Census Sales Data.xls". Select range A1:Q126 and use the Tools/Similarity Analysis utility to group each row into the most likely cluster—i.e., distinct group—in which we it deserves to belong.

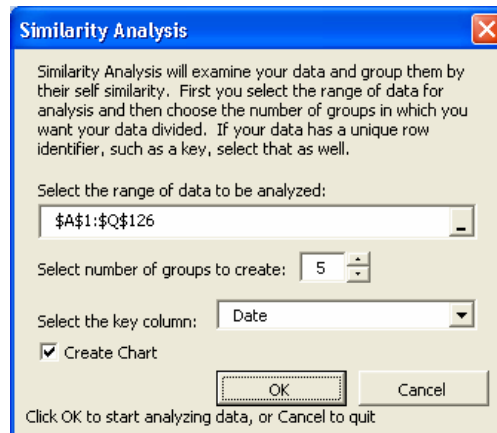


Figure 2: Similarity Analysis Dialog Box

Note in Figure 2 that the number of distinct clusters to attempt to create has been reduced from the default of ten to five. I prefer five as a starting point to hopefully gain a big picture without excessive clutter.

In the end, data mining success all boils down to data quality, model quality and an ability to interpret results. With regards to interpretation, Figure 3 shows three columns added by the utility. In the case of Group 5, we're essentially stating that each of the ten rows in Group 5 shares common traits partially described under column T.

S	T	U
Group Number	Group	Membership
1	-0.02 <= Furniture and Home Furnishings Stores <= 0.07,-0.03 <= Retail	38
2	-0.08 <= Furniture and Home Furnishings Stores <= 0.02,-0.07 <= Retail	34
3	Motor Vehicles and Parts Dealers > 0.26,Auto parts accessories and tires	22
4	Radio TV and other electrical stores > 0.63,Appliances TV and other elect	21
5	Furniture and Home Furnishings Stores <= -0.28, Home Furnishings Stores <= -0.40, -0.45 <= Appliances TV and other electrical stores <= -0.39	10

Figure 3: Cluster Summary

This boils down to probabilities, however, and it's easy to spot apparent discrepancies. For example, sales for Jan-94 changed more markedly for "Home Furnishings Stores" than would seem permissible by Group 5's cluster definition. Sales dropped by 36.7% while our cluster implied that group membership depended on sales drops of at least 40%.

Note, however, that sales changes *are* in line with respect to cluster expectations for "Furniture and Home Furnishing Stores" (-28.5%) and "Appliances TV and other electrical stores" (-43.3%). These general consistencies should hold true for the other column categories, too, of which only three are shown in the utility's group label.

Furniture and Home Furnishings Stores	Furniture Stores	Home Furnishings Stores	Appliances TV and other electrical stores
-22.0%	-16.2%	-29.2%	-40.5%
-28.5%	-21.9%	-36.7%	-43.3%
Furniture and Home Furnishings Stores <= -0.28, Home Furnishings Stores <= -0.40, -0.45 <= Appliances TV and other electrical stores <= -0.39			
-22.9%	-12.7%	-34.6%	-39.9%
-26.7%	-13.9%	-40.0%	-41.5%
-24.8%	-14.3%	-35.9%	-41.7%
-17.6%	-8.7%	-27.0%	-42.9%
-23.1%	-14.2%	-32.6%	-45.3%

Strictly speaking, any particular record is only loosely assigned to a cluster in a 'soft' manner—that is, with a probability. Per Jamie: "The numbers reported by AS are for "soft clustering" where each case is fractionally assigned to (potentially) each cluster." The following code can be used to firmly associate an underlying case (i.e., record) with a given cluster. "UniqueCaseID" is the primary—or 'case'—key for our input records:

```
SELECT model.Cluster(), t.UniqueCaseID
FROM model
PREDICTION JOIN OPENROWSET("...", "SELECT .... FROM ... as t")
```

"The above query is essentially "hard clustering" where you are assigning each case to one and only one cluster—in this situation, you are assigning it to the cluster it most likely belongs to."

Auto Fill

Our next example assumes we need to generate a best guess as to what a missing data value should be. For this, we'll use "Census Test.xls" and fill in missing values under the column "relationship".

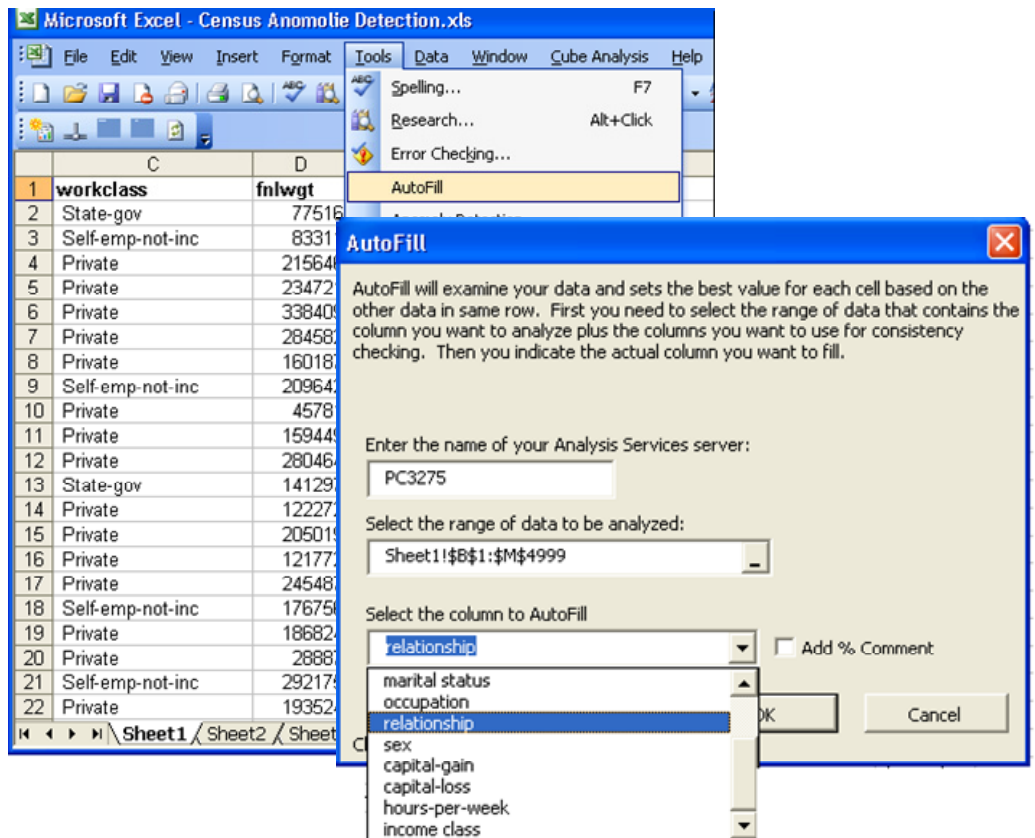


Figure 4: Auto Fill Set-Up

Interpretation of these decision tree results is more straightforward. Following the steps presented in Figure 4 (check "Add % Comment", too), results are shown in Figure 5. Again, we're dealing with varying probabilities of being right or wrong. Record 76 is given a rather weak 43% chance of having a "Not-in-Family" relationship.

	A	H	I	J
1	unique_id	relationship	sex	capital-gain
76	75	Other-relative	M	
77	76	Not-in-family	M	
78	77	Husband	M	
79	78	Husband	M	
80	79	Own-child	F	
81	80	Husband	Male	0
82	81	Own-child	Male	0
83	82	Husband	Male	0
84	83	Wife	Female	0
85	84	Husband	Male	0
86	85	Not-in-family	Female	14344
87	86	Not-in-family	Female	0
88	87	Husband	Male	0
89	88	Husband	Male	0
90	89	Not-in-family	Male	0
91	90	Not-in-family	Female	0

Figure 5: Auto Fill Results

Though not visible, a stronger 63% chance is given to record 86.

Hopefully, you'll enjoy reviewing the remaining two tools, **Anomaly Detection** and **Primary Indicator**. Don't confuse the latter with primary keys, of course. It's goal is to identify the best field that predicts the values of one you specify. Also, Anomaly Detection is particularly useful in this age of information overload. I find that far less use is made of exception reports and conditional formatting than should be within business today.

*Steve Mong is the DW/BI SIG Marketing Chair and president of Market Consulteks, Inc.
(www.marketconsulteks.com)*