Addendum to Segmentation of Customer Transactions

16.5 Addendum to Chapter 16

As stated in Chapter 16, the Time Series nodes in Enterprise Miner 7.1 are *experimental* and as such the nodes and their property sheets are in a state of flux until they are considered production worthy. The following table should outline the modifications from what is given in Chapter 16 so that the Similarity matrix can be given as well as clustering the transactions into similar groups. The steps are outlined in the Process Flow Table below.

Process Flow Table: Transaction Segmentation

Step	Process Step Description	Brief Rationale
1	Start SAS Enterprise Miner and create a new project called Transaction Segmentation.	Demonstrates how to perform segmentation of customer unit transactions.
2	Create a new project process flow diagram called Similarity.	
3	Add the data set called Customer_Account_Trans to the Data Sources folder. Set variables for the proper roles.	Adds a data set that contains both customer firmographics merged with unit purchases over time.
4	Add a TS Data Preparation node and connect the input data source to it.	Prepares the time series data using totals by quarters.
5	Add a TS Similarity node and connect the TS Data Preparation node to it.	Computes similarity metrics for all available time series.

Step 1: So, now let's create a new data mining project called Transaction Segmentation.

Step 2: Create a new process flow diagram called Similarity.

Step 3: Add a data set to the Data Sources folder called Customer_Account_Trans and note the following roles for the variables in this data set. Reject all variables except CUST_ID as an ID role, DATE as a Time ID role, US_REGION as a Cross ID role, and UNITS as a Target role. The variable UNITS represents the quantities of items in a time period.

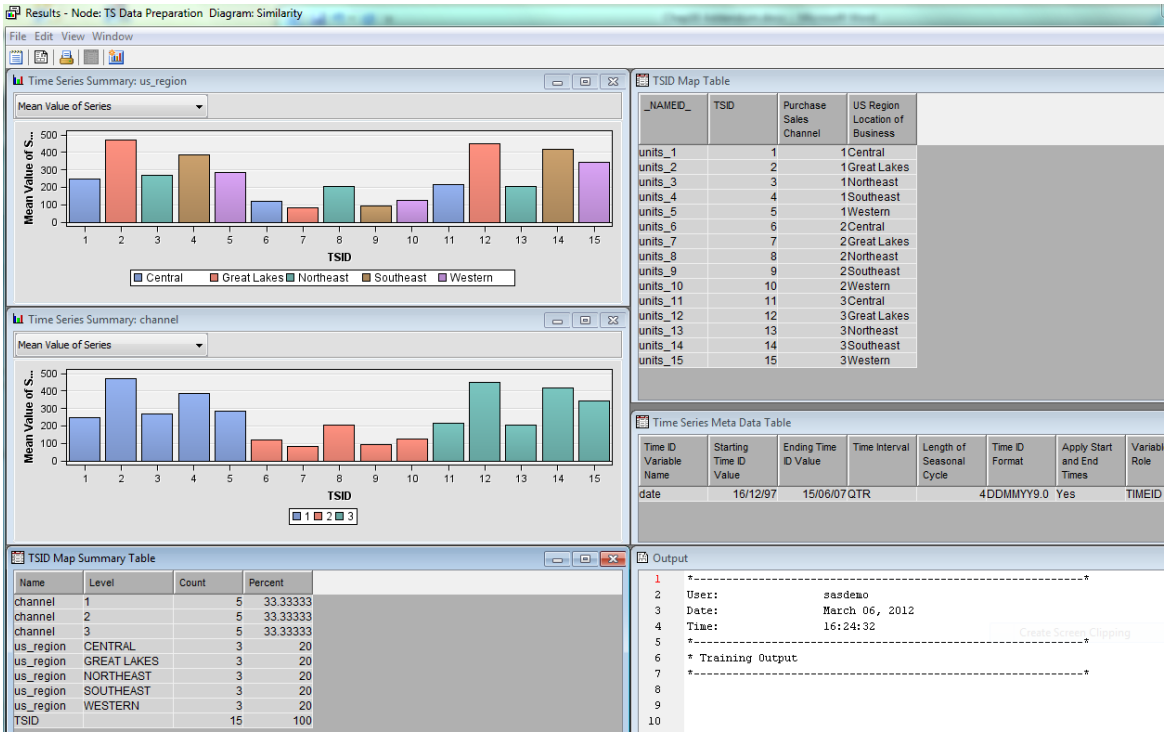
Step 4: On the TSDM tab of nodes, add the TS Data Preparation node to the process flow and connect the input node for Customer_Account_Trans to it. In the property sheet be sure to use the following settings as shown in Figure 16.5. These settings will specify that the time period of this data is quarterly and the accumulation method is Total. The variable UNITS, which is a Target role, will be summed by each time period for each customer group combination of the cross ID variables. A cross ID variable is a variable in which an aggregation is to be performed for the analysis. In the Missing Value property sheet, the set value of "Median" will cause any missing entries in each time unit to impute the median for that series of transactions for a customer ID. Be sure the Transpose property is set to Yes and the BY variable indicates "by TSID." The TSID is useful for similarity search, and the Time ID variable is useful for clustering purposes.

Figure 16.5 TS Data Preparation Node Property Settings

General	
Node ID	TSDP
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Input Data Type	Default
Time Interval	
Specify an Interval	Quarter
Seasonal Cycle Selection	Default
Length of Cycle	24
Start and End Time	Default
Date Time Selector	
Time of Day	Default
Accumulation	Total
Transformation Options	
Transformation	None
Box-Cox Parameter	0.0
Difference Options	
Apply Differencing	No
Difference Order	1
Seasonal Differencing	No
Missing Value	
Set Value	Median
Constant Value for Missing	0.0
Zero Missing	None
Transpose Options	
Transpose	Yes
By Variable	By TSID

Step 5: Now we will add a TS Similarity node and connect the Metadata node to it. The Similarity node will compute transaction similarity from the SIMILARITY procedure. For this analysis, we'll use the following property sheet settings: Similarity Measure (Squared Deviation), Sequence Sliding (None), Interval (Default), Normalization (Standard), and Accumulation (Total). Figure 16.7 shows all the property sheet settings.

Figure 16.6 TS Data Preparation Node Results



Step 6: Now we will add a TS Similarity node and connect the Metadata node to it. The Similarity node will compute transaction similarity from the SIMILARITY procedure. For this analysis, we'll use the following property sheet settings: Similarity Measure (Squared Deviation), Sequence Sliding (None), Scale (None), Normalization (Standard), and Accumulation (Total). The compression and expansion options are set to none for now. Figure 16.7 shows all the property sheet settings.

Figure 16.7 TS Similarity Node Property Sheet Settings

General	
Node ID	TSSIM
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Similarity Measure	Squared Deviation
Sequence Sliding	None
Accumulation	Total
Normalization	Standard
Scale	None
Distance Matrix Options	
Export Distance Matrix	Yes
Include Targets	No
Hierarchical Clustering	Yes
Compression Options	
Compress	None
Global Absolute Compression	0
Global Compression Percentage	0
Local Absolute Compression	0
Local Compression Percentage	0
Expansion Options	
Expansion	None
Global Absolute Expansion	0
Global Expansion Percentage	0
Local Absolute Expansion	0
Local Expansion Percentage	0
Report	
Similarity Plot Maximum	5
Preference of Similarity Plot	Most Similar

Figure 16.8a Directions of Paths for Metric Measures
(Diagonal – direct), (Horizontal – expansion), (Vertical – compression)

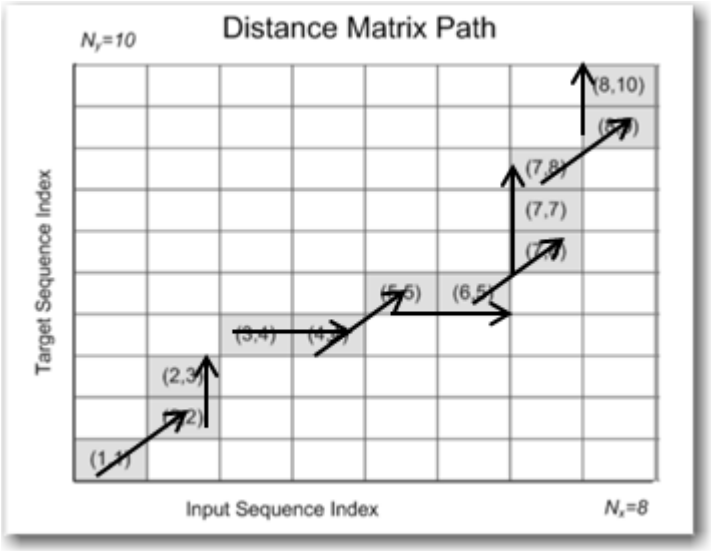
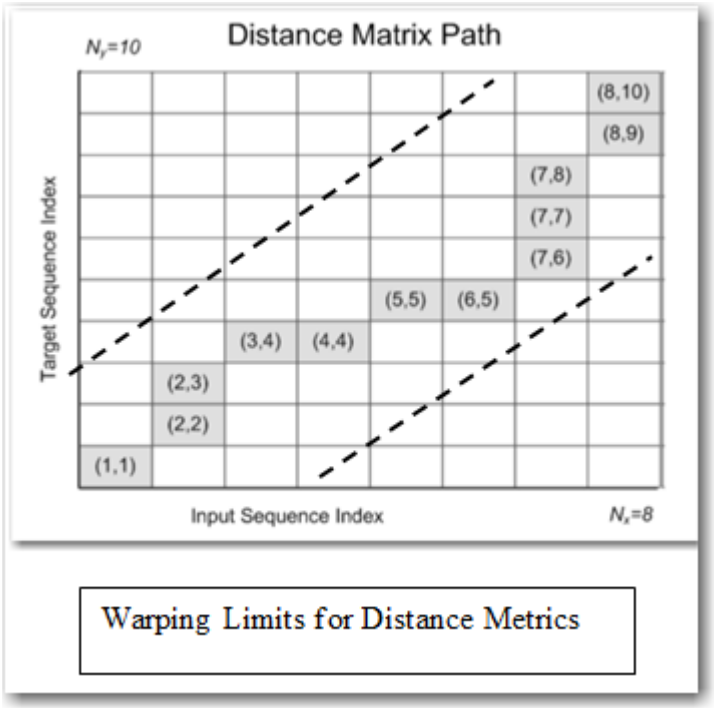


Figure 16.8b Warping Limits for Distance Metrics



The SIMILARITY procedure has a large number of possibilities for computing distance path metrics. The path matrix shown in Figure 16.4 gives the target versus input sequence. There are many paths that can be used to compute a distance metric. For example, in Figure 16.8a the arrows indicate how various direct (diagonal), expansion (horizontal), and compression (vertical) directions that distance measurements can be made. If limits are placed from the main diagonal (shown in dashed lines), then these indicate how far from deviation between target and input sequence and give boundaries for distance metric computations. Figure 16.8b shows the limits for distance computations. For each path taken through the matrix, the relative distance deviation encountered is measured and for a set of paths, statistics such as the minimum and maximum path averages, and cost functions that measure distances from warping limits to the path taken in the matrix. These statistical measures are used to define a distance metric that measures the overall similarity between transaction patterns in both the time and magnitude dimensions. This allows the metric to be used in other analytics such as clustering, segmentation, and as an input into other predictive models. Now, run the Similarity node and open the Results window.

In the first set of results in the Similarity node you'll see the Similarity Map. This color-coded map indicates by the time series ID the similarity metric to every other time series ID. The more blue the color, the more similar the two series are to each other; the more red, the less similar they are to each other. For example, in Figure 16.9 the Similarity Map matrix plot shows that Units_8 and Units_14 are not very similar whereas Units_8 and Units_9 are similar.

If you select the View pull-down menu and select **SAS Results→Train Graphs**, all of the selected Graphs items in the property sheet in Figure 16.7 will display. Figures 16.10a and 16.10b show the sequence plots between Units_8 versus Units_14 and Units_8 versus Units_9, respectively.

Figure 16.9 Similarity Map Matrix for Units Transactions

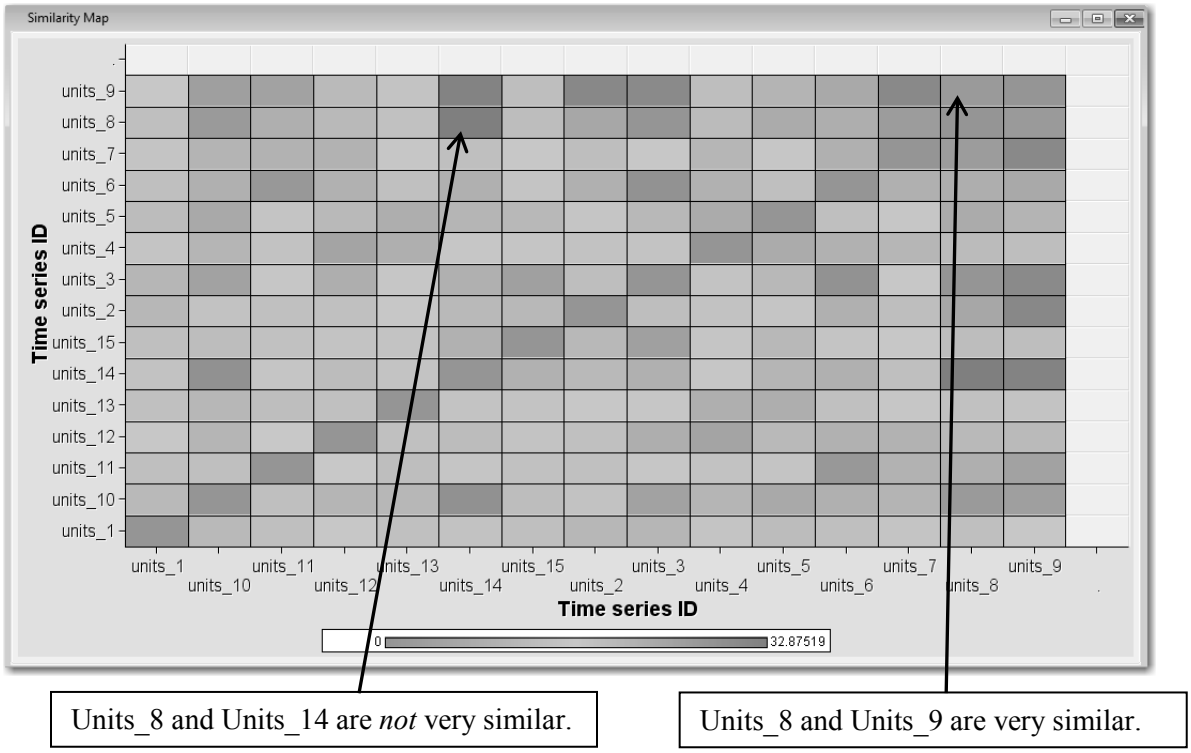


Figure 16.10a Units_8 and Units_14 Sequence Plot

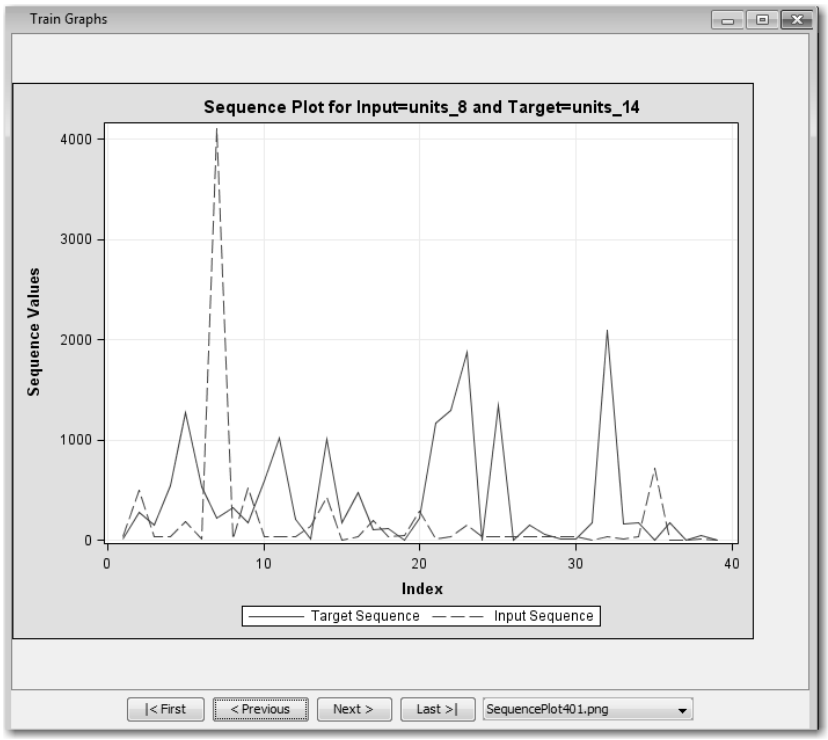


Figure 16.10b Units_8 and Units_9 Sequence Plot

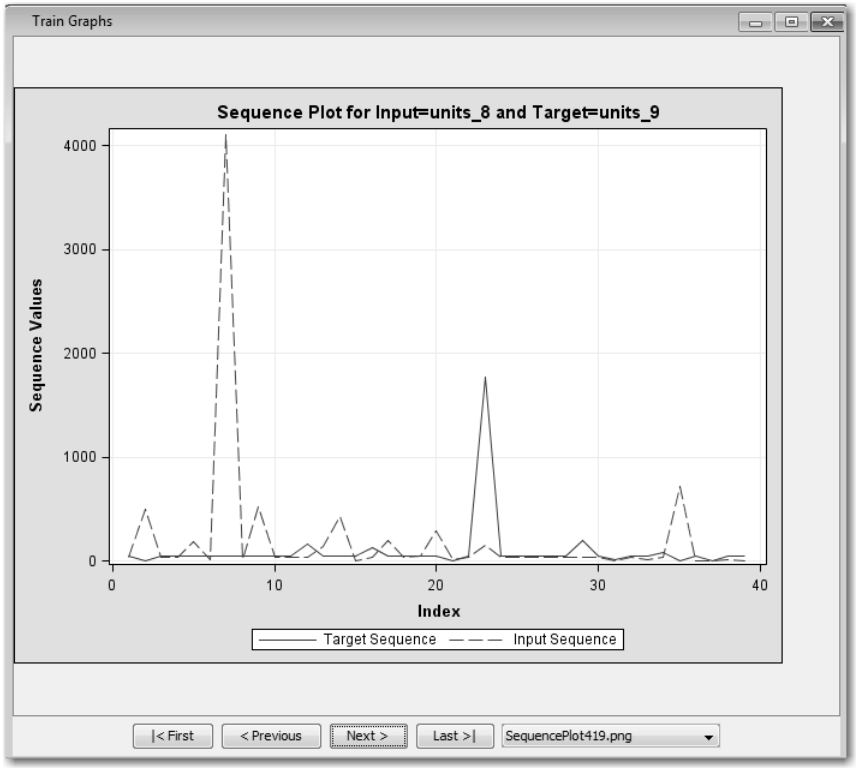


Figure 16.11 Clustering of the Similarity Metrics from the Units Transactions Data

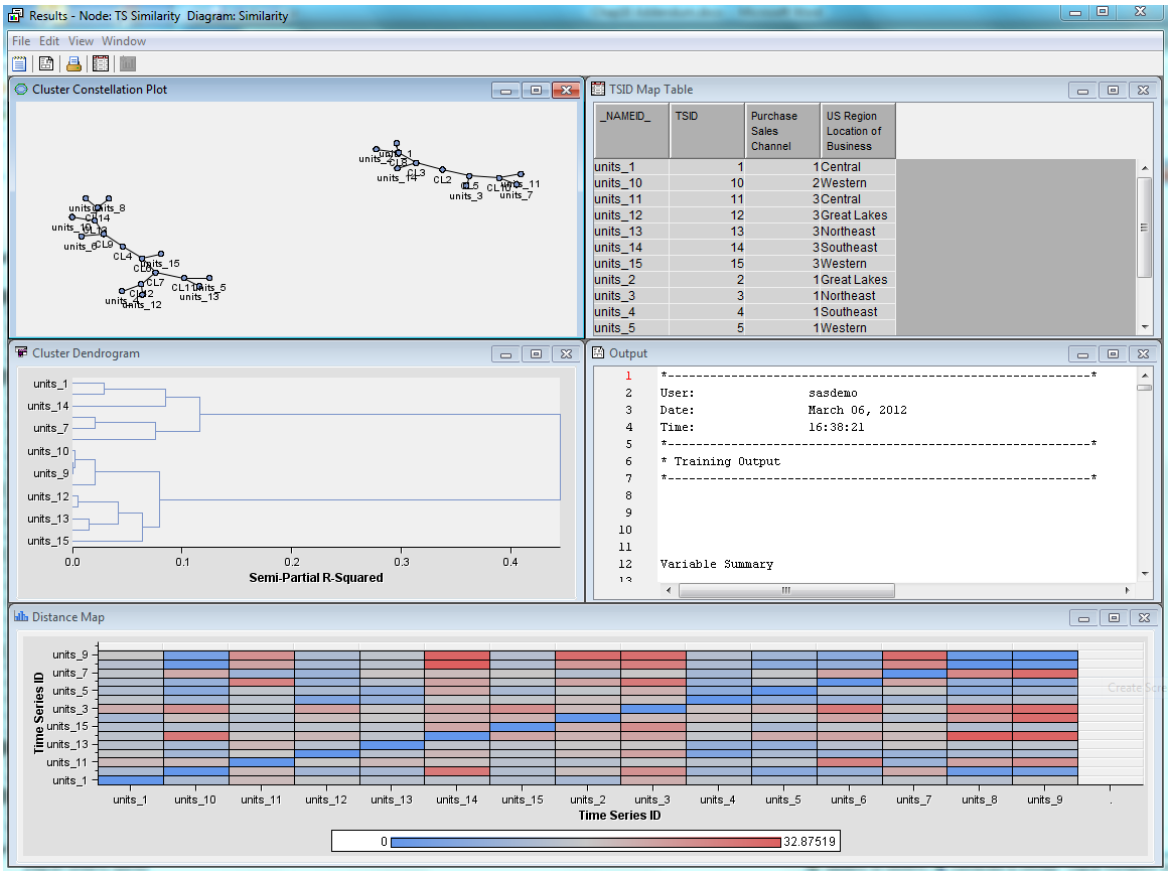


Figure 16.12 Completed Process Flow Diagram for Transaction Segmentation