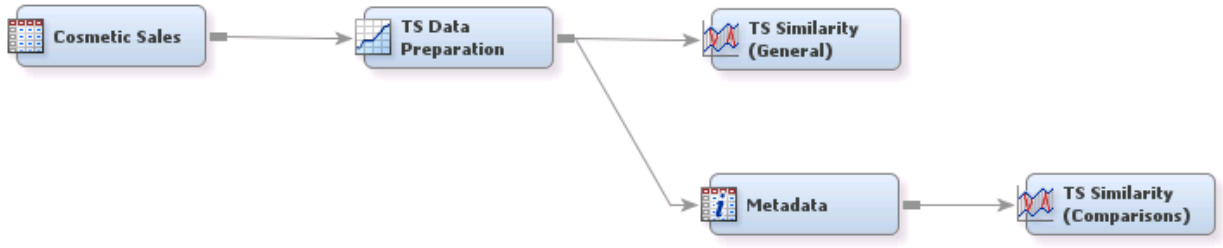
Time Series Exploration and Comparison Using SAS® Enterprise Miner™



Data:

The example uses the Cosmetic data set in the SAS library SAMPSIO to create the data source, which is generated by clicking **Help >> Generate Sample Data Sources >> Cosmetic Sales**. The Cosmetic data set contains time series data that show sales over time for various cosmetic products across different states. The following figure shows first few observations in the Cosmetic data set:

SAMPSIO.COSMETIC

	Time ID	Sales	CrossID: product	CrossID: group	CrossID: state
1	Jan 1, 1996	82971.0	54105	A	NC
2	Jan 1, 1996	82322.0	54105	A	GA
3	Jan 1, 1996	95834.0	54105	A	WI
4	Jan 1, 1996	98710.0	54105	A	MD
5	Jan 1, 1996	123089.0	54105	A	FL
6	Jan 1, 1996	86419.0	54105	B	NC
7	Jan 1, 1996	81668.0	54105	B	GA

Goal:

The goal is to explore a time series data set and organize it into a format for further analysis that compares different time series in the data.

The IDS (input data source) node, named Cosmetic Sales, is connected to the TS Data Preparation node. The TS Data Preparation node provides graphs and statistics of the time series data and prepares the time series data for the TS Similarity node.

TS Data Preparation Node:

The following properties are changed in the TS Data Preparation node:

Variables:

- The group variable is not used in the analysis.

Transpose options:

- Transpose: Yes

- By Variable: By TSID
- Keep Variable Role: No

The following output table shows that the TS Data Preparation node identifies 25 time series (_TS_01:_TS_25) variables in the data, one sales series for each combination of product and state.

TSID Map Table

Time Series ID	Original Variable Name	Variable Label	CrossID: product	CrossID: state
1	_TS_01	Sales 1	54105	FL
2	_TS_02	Sales 2	54105	GA
3	_TS_03	Sales 3	54105	MD
4	_TS_04	Sales 4	54105	NC
5	_TS_05	Sales 5	54105	WI
6	_TS_06	Sales 6	54321	FL
7	_TS_07	Sales 7	54321	GA
8	_TS_08	Sales 8	54321	MD
9	_TS_09	Sales 9	54321	NC
10	_TS_10	Sales 10	54321	WI
11	_TS_11	Sales 11	54551	FL
12	_TS_12	Sales 12	54551	GA
13	_TS_13	Sales 13	54551	MD
14	_TS_14	Sales 14	54551	NC
15	_TS_15	Sales 15	54551	WI
16	_TS_16	Sales 16	56771	FL
17	_TS_17	Sales 17	56771	GA
18	_TS_18	Sales 18	56771	MD
19	_TS_19	Sales 19	56771	NC
20	_TS_20	Sales 20	56771	WI
21	_TS_21	Sales 21	57998	FL
22	_TS_22	Sales 22	57998	GA
23	_TS_23	Sales 23	57998	MD
24	_TS_24	Sales 24	57998	NC
25	_TS_25	Sales 25	57998	WI

Time Series ID

The results of the TS Data Preparation node include many graphs and statistics (such as seasonal statistics plots, multiple time series comparison plots, and data tables) that describe the time series data from different aspects and enable you to visualize your data.

TS Similarity (General) Node:

The TS Similarity (General) node compares each time series to the target time series. For this analysis, the default properties of the TS Similarity node are used. Since none of the time series variables are assigned a Role of Target, the TS Similarity node performs pairwise comparisons among all of the 25 input time series variables.

The results of the TS Similarity node include the following Cluster Dendogram, which shows the hierarchical clustering of time series; a Cluster Constellation Plot, which shows the affinities among the

time series in addition to the structures of the different clusters; and a The Distance Map which shows the distances between two time series by different shades of blue and red. Note that for each series the closest distance (darkest blue) occurs at itself, because each series is at zero distance from itself. In this analysis, the TS Similarity node uses the default similarity measure, **Squared Deviation**, to calculate the distance between different time series. You can choose a different similarity measure by using the **Similarity Measure** property of the node.

TS Similarity (Comparisons) Node:

Whereas TS Similarity (General) node performs pairwise comparison between all 25 time series variables, because none of the variables in the input data set has the target role, the TS Similarity (Comparisons) node compares each time series to the target time series. Hence, the TS Similarity (Comparisons) node requires changes in the role of the input variables. This can be done by the Metadata node which enables you to use a single data source in multiple capacities in your flow diagram. In this analysis, the Metadata node specifies new target variable roles for the input time series variables _TS_01 and _TS_02. The role of each variable can be seen by clicking the Transaction property of the TS Similarity (Comparisons) node.

The results of the TS Similarity (Comparisons) node include graphs that compare each target time series to the five closest input series. For example the “Target: Sales 2 versus Input Series” graph shows that Sales 4, 5, 3, 20, and 18 (listed in order) are the closest to Sales 2. Notice that the two target time series variables are compared to each other. If you want to compare the target series to each other, then you need to set each series as a target one at a time.