Time Series Data Mining

Mihai Paunescu

Was ist Data Mining?

Verhalten = f(Merkmale)

Statische und dynamische Merkmale

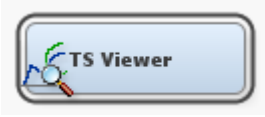
- Statische Merkmale:
  - Geschlecht, Adresse, … Branche, Rechtsform, ...
- Dynamische Merkmale
  - Umsatz, Konsum, Kontakte, Lagerbestand, Transaktionen, Klicks

KundenID	Geschlecht	Adresse	Umsatz 2010/02	Umsatz 2010/01	Umsatz 2009/12	Umsatz 2
1	M	1030	350	300	450	32
2	W	1070	50	73	91	25
…	…	…	…	…	…	..

KundenID	Geschlecht	Adresse	Umsatz Durchschnitt	Umsatz Wachstum	Umsatz Volatilität
1	M	1030	312	0.02	1.2
2	W	1070	33	- 0.01	2.3
…	…	…	…	…	…

Aggregation von Zeitreiheninformationen mit Time Series Data Mining nodes im EM 6.1

- TS Viewer



  - Provides users user's specific time series plots
- TS Data Preparation

  - Provides a tool of aggregation, differencing, summarization, etc.
- TS Merge

  - Merges a clustering output with the segment variable into the original format of input data
- TS Hierarchical Clustering

  - Provides time series hierarchical clustering
- TS Exponential Smoothing

  - Fits ESM to interval variables
- TS Seasonal Decomposition

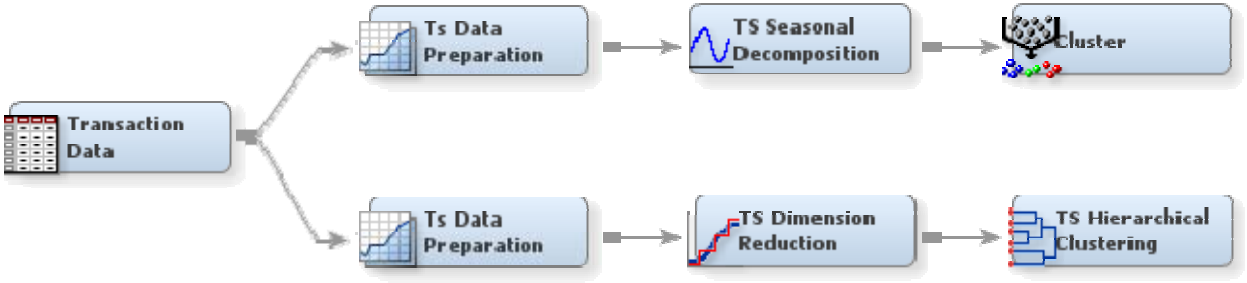
  - Provides users user's specific time series plots
- TS Dimension Reduction

  - Perform time series dimension reduction using SVD, Wavelet, and Fourier transformation etc.
- TS Cross Correlation

  - Compute lagged cross-correlation among time series
- TS Similarity

  - Computes several similarity measures among time series

Vorgehen bei Time Series Data Mining

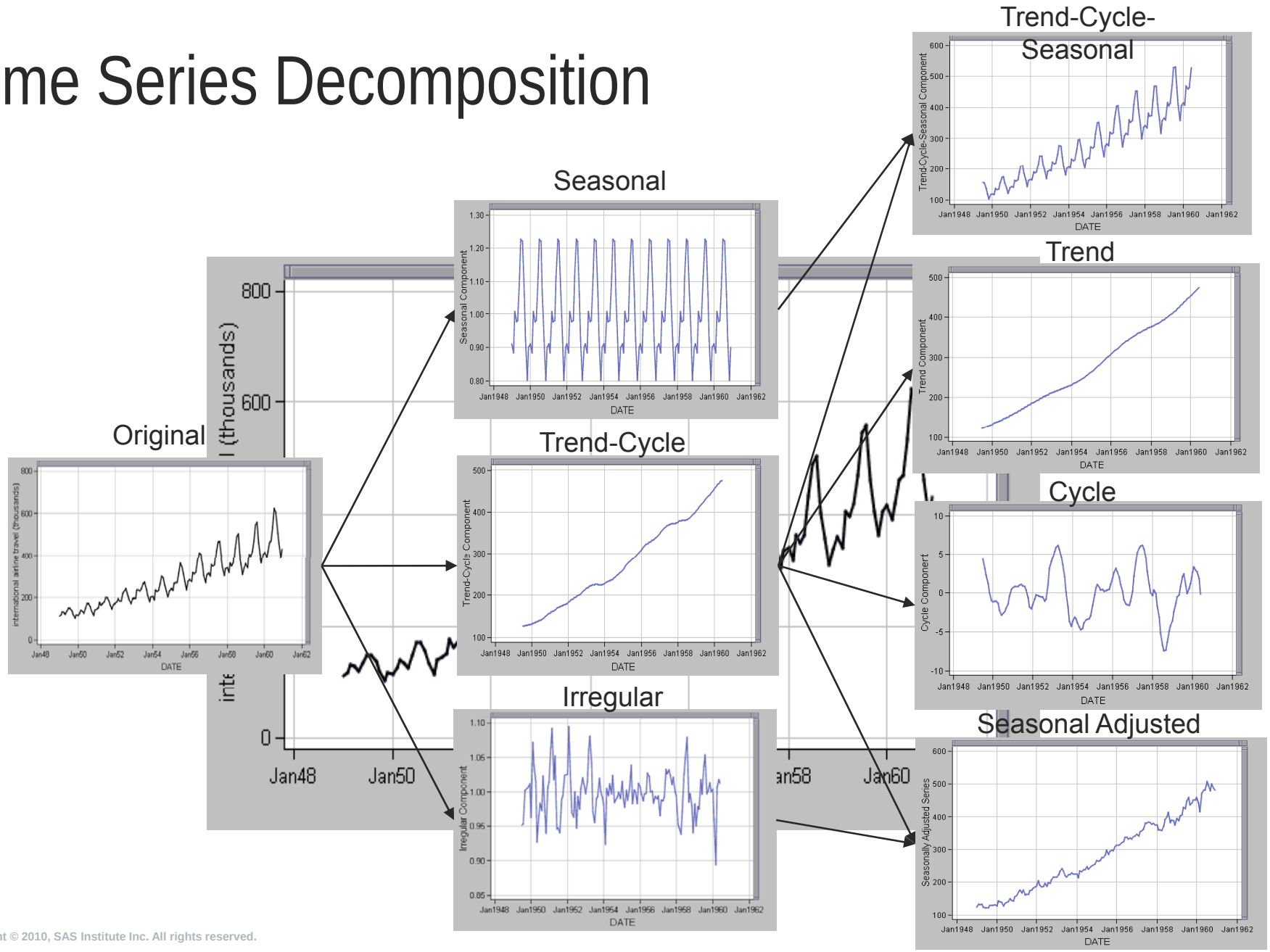


id	TCC:PERIOD1	TCC:PERIOD2	TCC:PERIOD3	TCC:PERIOD4	TCC:PERIOD5	TCC:PERIOD6	SC:PERIOD1	SC:PERIOD2	SC:PERIOD3
id	Segment...								
1	Cluster1	5988.931	5944.3	5869.333	5831.017	5858.418	0.87515	0.847697	0.847978
2	Cluster1	3760.046	3773.292	3792.917	3812.355	3825.415	0.846798	0.859126	0.844815
3	Cluster5	7172.849	7264.033	7332.42	.	.	0.880371	0.881523	0.888228
4	Cluster1	219274.7	218786.8	218132.7	217437.1	217043.5	0.887757	1.296663	0.891567
5	Cluster1	2149.886	2150.399	2163.092	2179.98	2180.278	0.848795	0.853559	0.862608
7	Cluster2	129674.1	130116.7	130656.7	130987.5	130920.1	0.861852	0.900875	0.92742
9	Cluster1	165894.4	166553.7	167550.9	169148	171674.3	0.877393	0.888616	0.908495
10	Cluster2	86313.64	86506.36	86790.27	87383.49	89033.07	0.794953	0.886591	0.844434
11	Cluster1	221551.3	222388.6	223591.8	224859	225615.9	0.899973	0.885815	0.910796

KundenID	Geschlecht	Adresse	Umsatz Segment
1	M	1030	Cluster1
3	W	1070	Cluster5
…	…	…	…

B...	id	time_id	Umsatz
34	1	Oct2003	.
35	1	Nov2003	.
36	1	Dec2003	.
37	1	Jan2004	5.254,01
38	1	Feb2004	5.067,19
39	1	Mar2004	5.063,56
40	1	Apr2004	5.078,08
41	1	May2004	4.868,30
42	1	Jun2004	9.639,65
43	1	Jul2004	4.892,91
44	1	Aug2004	4.892,91
45	1	Sep2004	4.932,32
46	1	Oct2004	4.921,43
47	1	Nov2004	10.247,49
48	1	Dec2004	5.482,49
49	1	Jan2005	5.473,15
50	1	Feb2005	5.202,70
51	1	Mar2005	4.863,19
52	1	Apr2005	5.037,15
53	1	May2005	5.013,07
54	1	Jun2005	9.722,94
55	1	Jul2005	5.013,07
56	1	Aug2005	5.002,18
57	1	Sep2005	5.144,97
58	1	Oct2005	5.898,19
59	1	Nov2005	10.873,40
60	1	Dec2005	5.861,41
61	1	Jan2006	5.255,25
62	1	Feb2006	5.192,79
63	1	Mar2006	4.958,57
64	1	Apr2006	4.712,37
65	1	May2006	4.712,37
66	1	Jun2006	9.822,15
67	1	Jul2006	4.712,37
68	1	Aug2006	4.803,24
69	1	Sep2006	4.985,16
70	1	Oct2006	4.986,84
71	1	Nov2006	9.985,54
72	1	Dec2006	.
73	1	Jan2007	.
74	2	Jan2001	2.504,89
75	2	Feb2001	2.497,62
76	2	Mar2001	2.501,25

Time Series Decomposition



Dimensionsreduktion

Methoden von Dimension reduction

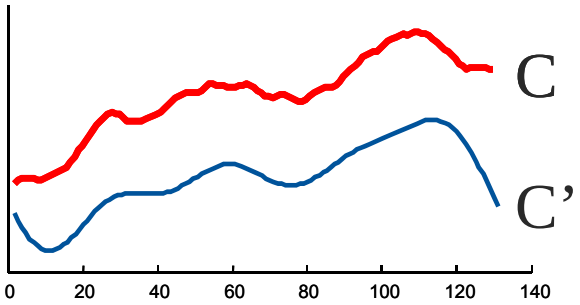
1. Discrete Fourier Transformation (DFT)
2. Discrete Wavelet Transformation (DWT)
3. Singular Value Decomposition (SVD)
4. Line Segment Methods (with Mean LSM, or Sum LSS)

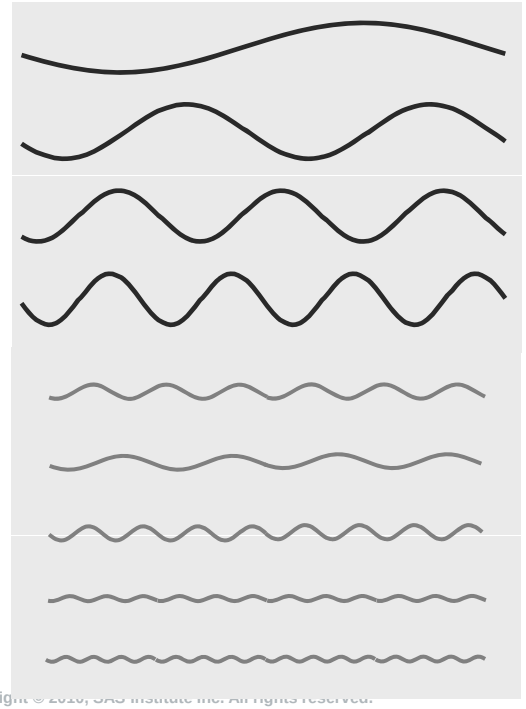
implentiert in

- TS Dimension Reduction 

Beispiel für die Methoden Dimensionsreduktion

$n = 128$

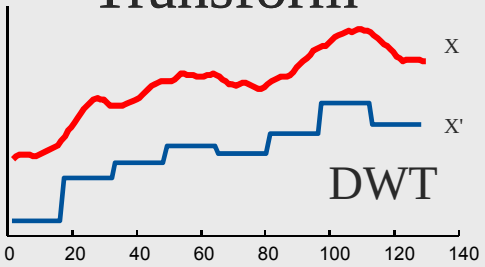


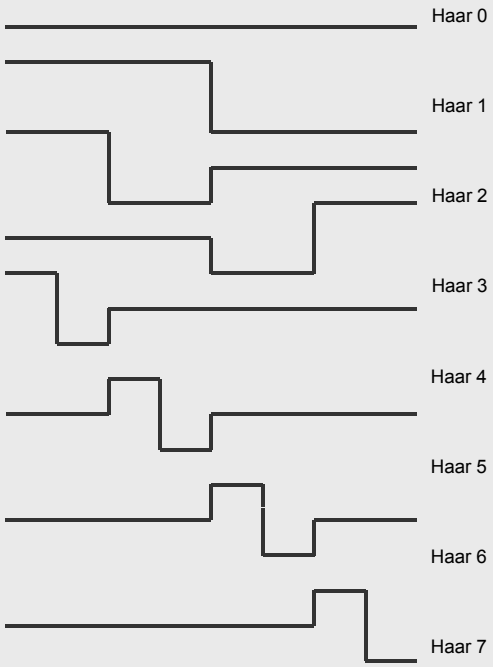


Zeitreihe	Fourier Koeffizienten	8 Fourier Koeffizienten
0.4995	1.5698	1.5698
0.5264	1.0485	1.0485
0.5523	0.716	0.716
0.5761	0.8406	0.8406
0.5973	0.3709	0.3709
0.6153	0.467	0.467
0.6301	0.2667	0.2667
0.6420	0.1928	0.1928
0.6515	0.1635	
0.6596	0.1602	
0.6672	0.0992	
0.6751	0.1282	
0.6843	0.1438	
0.6954	0.1416	
0.7086	0.14	
0.7240	0.1412	
0.7412	0.153	
0.7595	0.0795	
0.7780	0.1013	
0.7956	0.115	
0.8115	0.1801	
0.8247	0.1082	
0.8345	0.0812	
0.8407	0.0347	
0.8431	0.0052	
0.8423	0.0017	
0.8387	0.0002	
...	...	

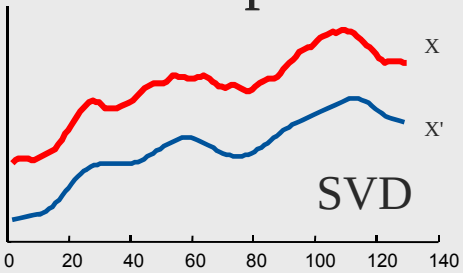
$N = 8$

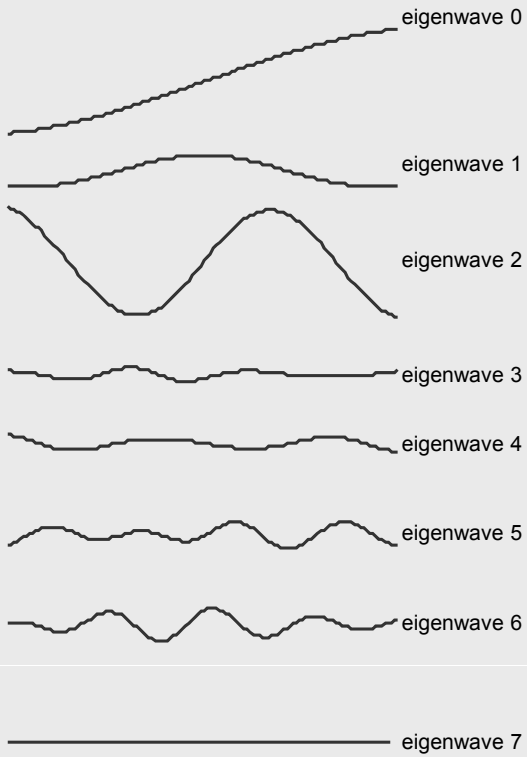
Discrete Wavelet Transform



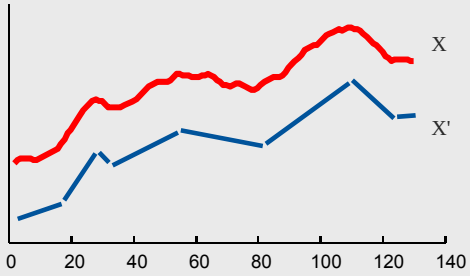


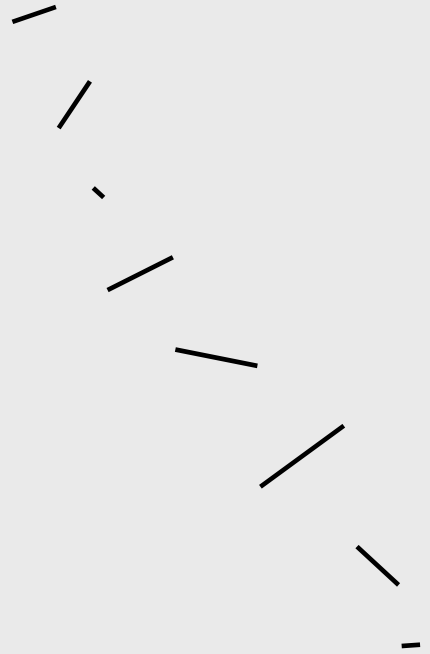
Singular Value Decomposition



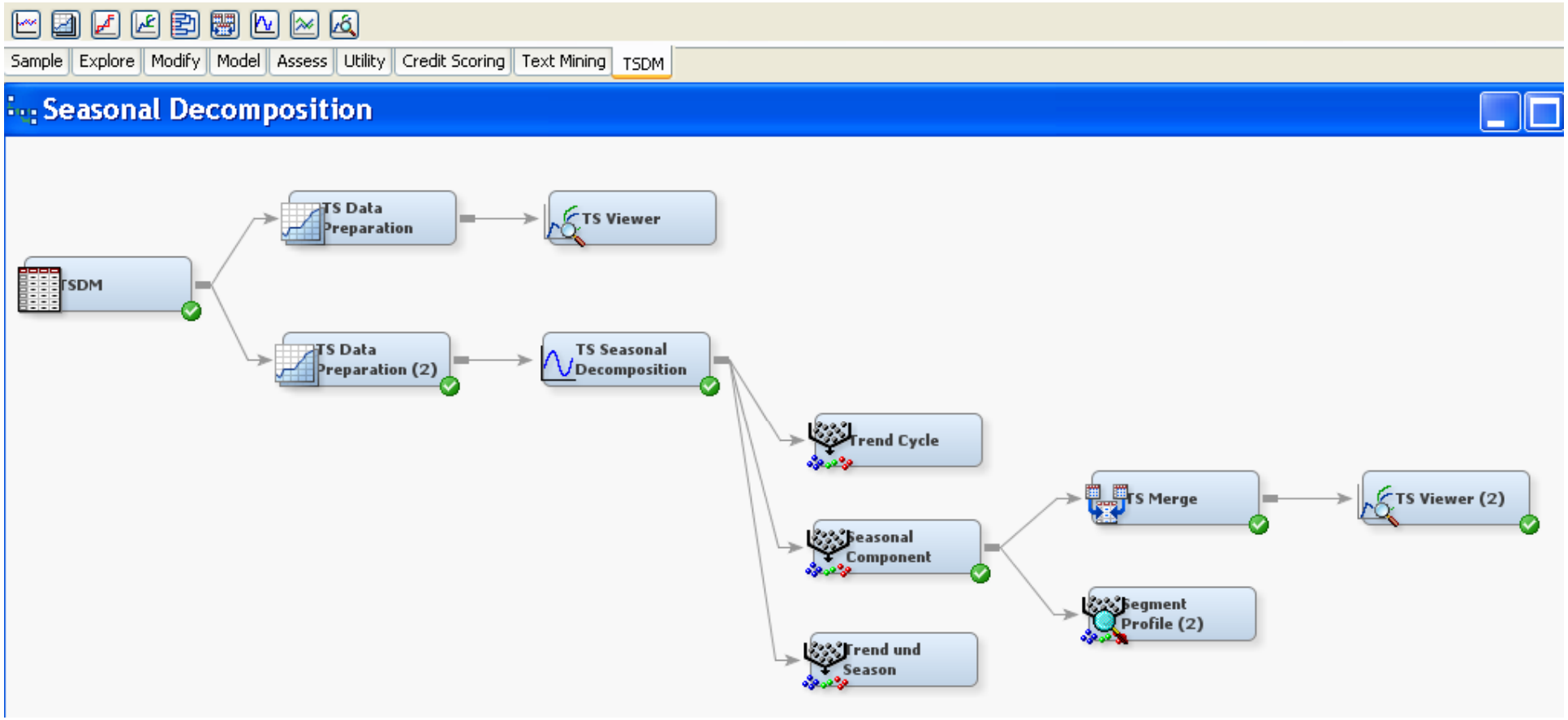


Piecewise Linear Approximation I





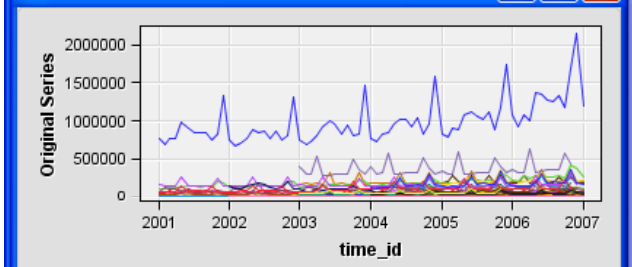
Demo für Seasonal Decomposition



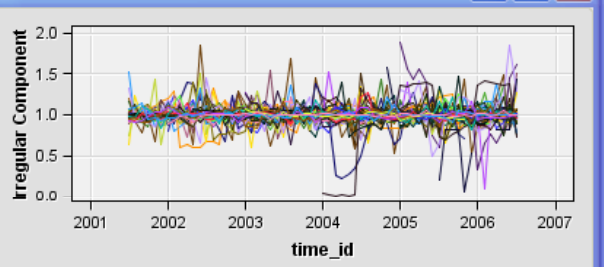
Results - Node: TS Seasonal Decomposition Diagram: Seasonal Decomposition

File Edit View Window

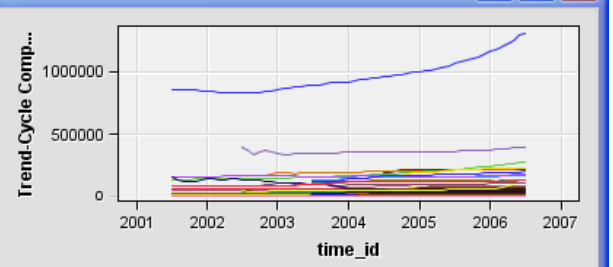
Original Time Series Plot



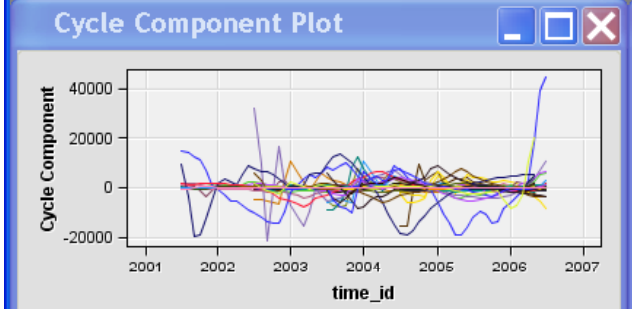
Irregular Component Plot



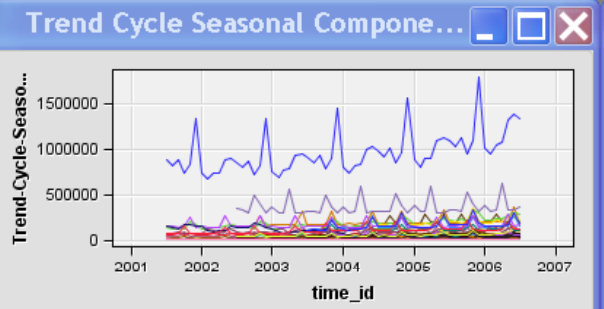
Trend Cycle Component Plot



Cycle Component Plot



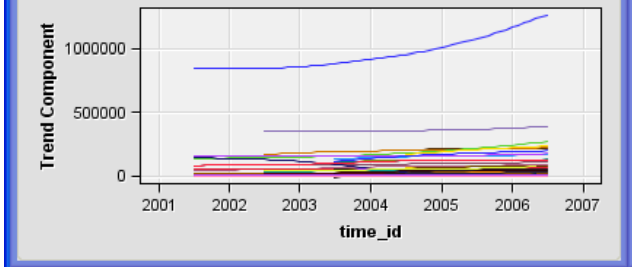
Trend Cycle Seasonal Compone...



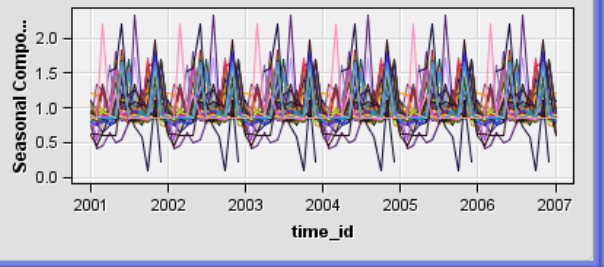
Exported Decomposed Compon...

TSID	id	TCC:PERIOD 1	TCC:PERIOD 2	TCC:PERIOD 3	TC 4
1	1	6003.879	5988.931	5944.3	
2	2	3747.453	3760.046	3773.292	
3	3	7589.044	7172.849	7264.033	
4	4	219805.7	219274.7	218786.8	
5	5	2199.395	2149.886	2150.399	
6	7	129742.2	129674.1	130116.7	
7	9	165258.5	165894.4	166553.7	
8	10	86305.74	86313.64	86506.36	

Trend Component Plot



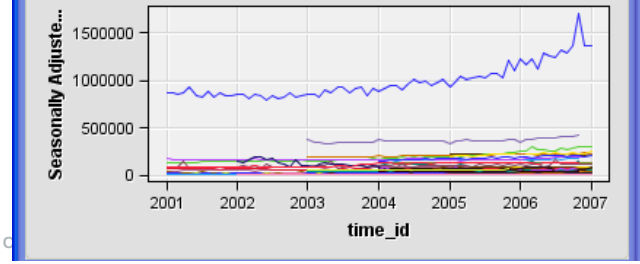
Seasonal Component Plot



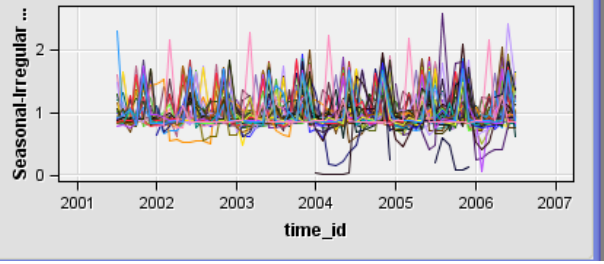
Output

```
1  *-------------------------------------------------
2  User:                  autmip
3  Date:                  05. Mai 2010
4  Time:                  10.31 Uhr
5  *-------------------------------------------------
6  * Training Output
7  *-------------------------------------------------
8
9
10
```

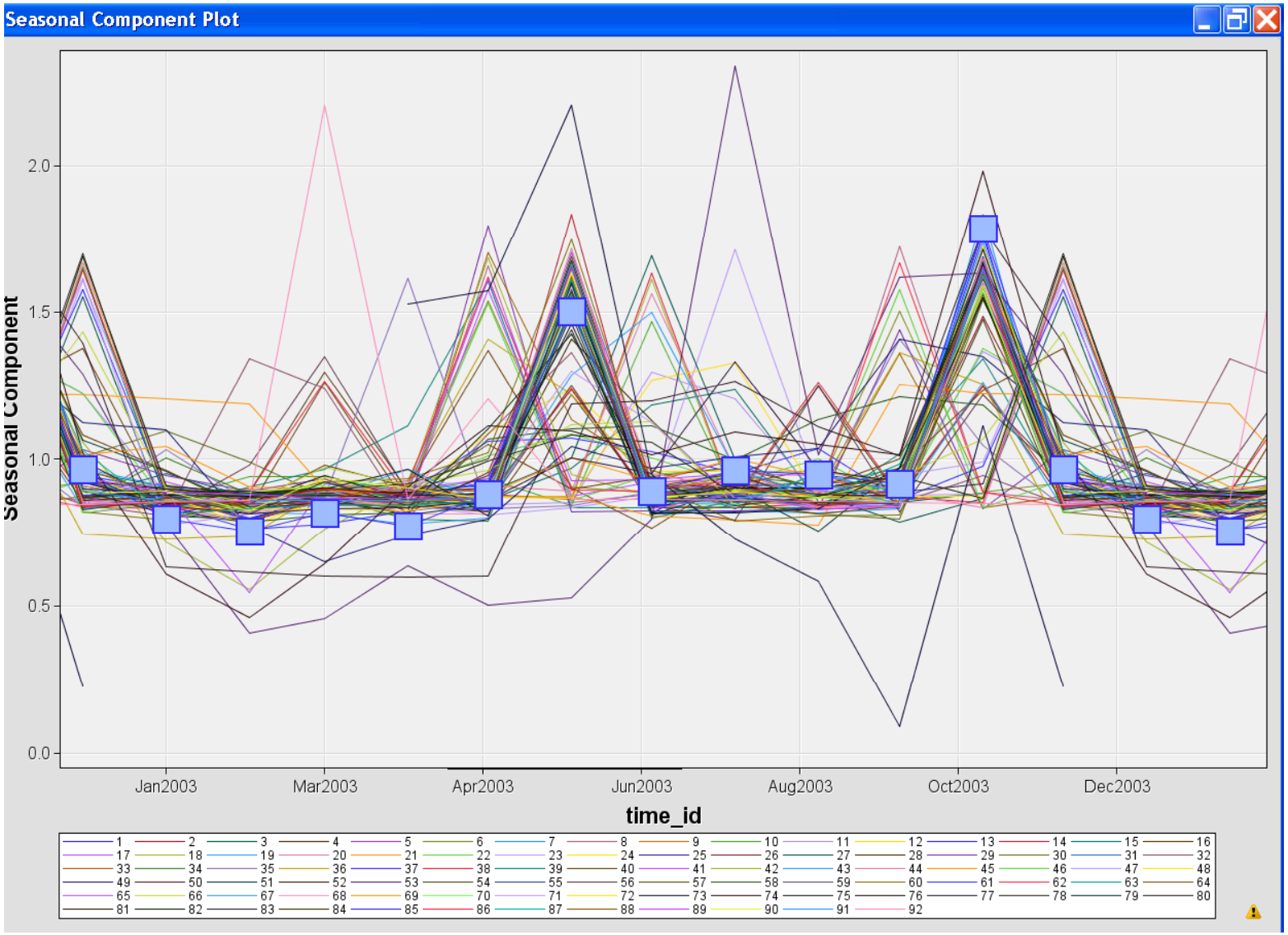
Seasonal Adjusted Plot

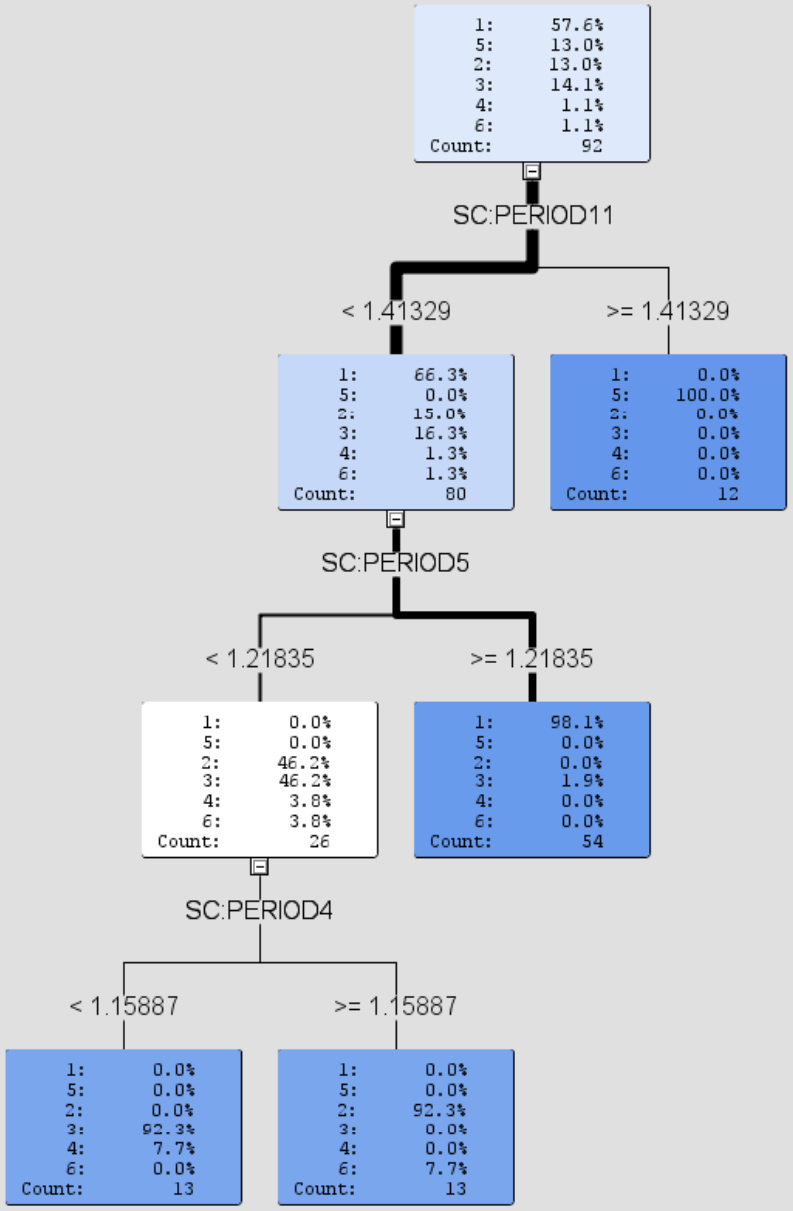


Seasonal Irregular Component ...



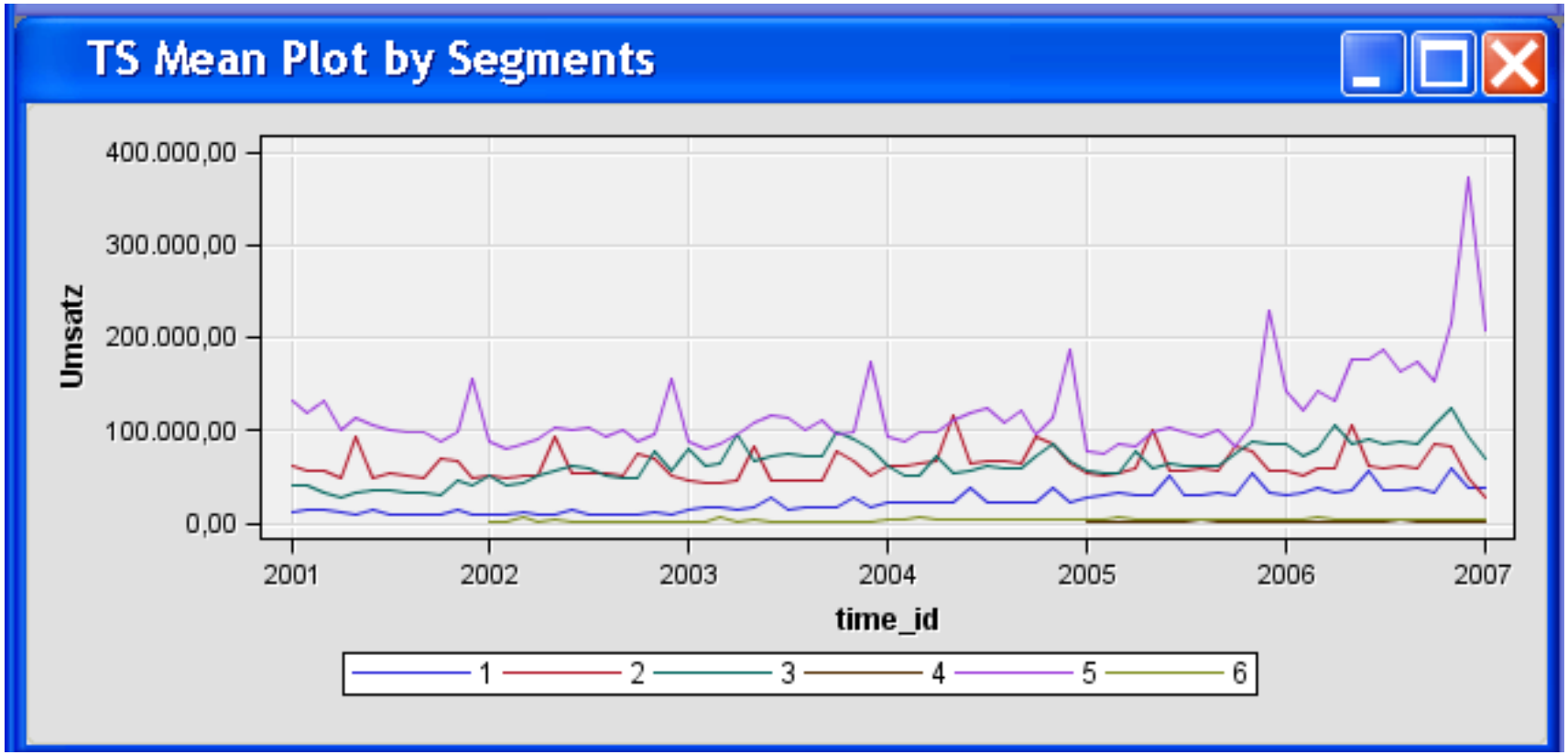
Seasonal Component





Entscheidungsbaum auf den Cluster für Saison Komponente

Plot der Mittelwerte für Cluster der Saison Komponente



Fazit

- Transaktionsdaten in Form von Zeitreihen liefern wertvolle Informationen für Vorhersagemodelle
- Time Series Data Mining aggregiert Zeitreihen mit minimalen Informationsverlust zu neuen Inputvariablen
  - Vorbereitung von Transaktionsdaten für Zeitreihenanalyse
  - Visualisierung der Zeitreihendaten für Analyse
  - Aggregation von Informationen auf der Grundlage von:
    - Zeitreihenzerlegung
    - Dimensionsreduktion
    - Ähnlichkeitsmaße
  - Zusammenfassung durch Clusteranalyse