Similarity Analysis – an Introduction, a Process, and a Supernova

Paper 2016-11884

David J Corliss, Wayne State University

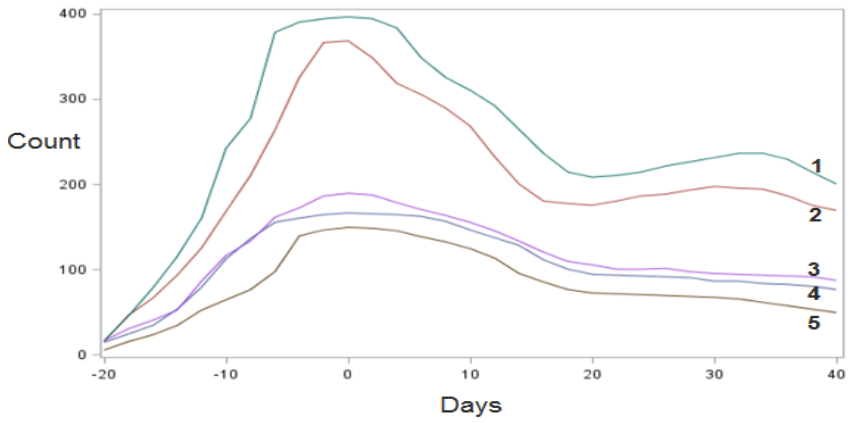
Abstract

Similarity analysis is used to classify an entire time series by type. In this method, a distance measure is calculated to reflecting difference in form between two ordered sequences, such as are found in time series data. The mutual differences between many sequences or time series can by calculated in SAS© using PROC SIMILARITY and then compiled in a similarity matrix, similar in form to a correlation matrix. When cluster methods are applied to a similarity matrix, time series with similar patterns are grouped together and placed into clusters distinct from others containing times series with different patterns. In this example, similarity analysis is used to classify supernovae by way the amount of light they produce changes over time. Additional examples demonstrate the use of similarity analysis in other areas, including biostatistics and econometrics.

Introduction to Similarity Analysis

Similarity analysis is an emerging statistical method for comparing and clustering different trends or patterns over time. It uses a distance measure to quantify the difference in form or shape between different time series: two series get a small similarity distance if their pattern is similar but receive a large number if the patterns are different. The mutual differences between many sequences or time series can be compiled in a similarity matrix, similar in form to a correlation matrix using multiple correlations. Cluster analysis performed using the similarity distance. This allows data captured over time to be classified into groups.

This plot shows five light curves – the amount of light produced by a star over time. Each shows the rise and fall in brightness from a Type Ia supernova in the SDSS Supernova Survey. Thanks are due to David Cinabro of Wayne State University for copy of the source data.

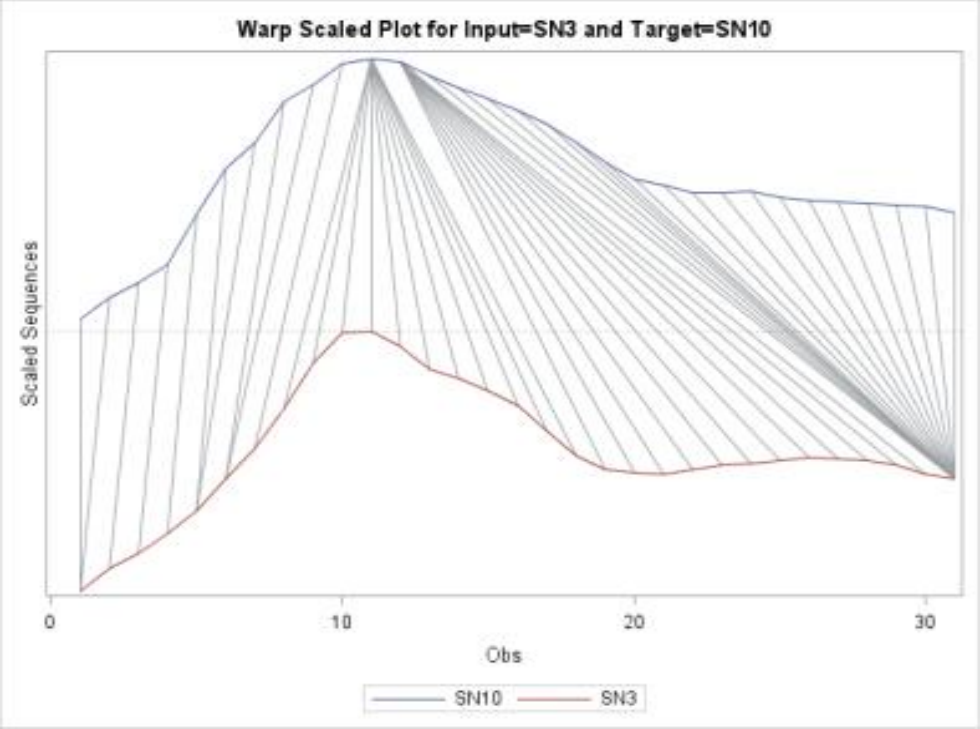


In this example, Similarity Analysis reads and quantifies the differences in shape between the five lines. Two groups are created. The first group, with lines 1 and 2, has a higher, faster peak and also a secondary peak later on. The second group, with lines 3, 4 and 5, has a slower initial growth rate and no secondary peak.

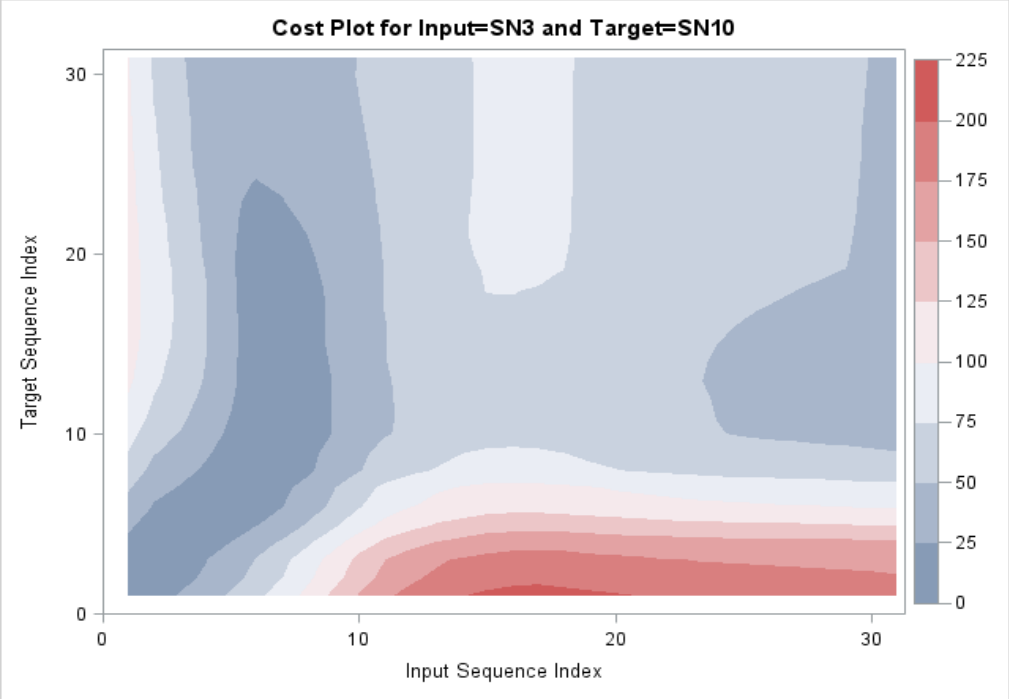
Dynamic Time Warping

In addition to calculating the similarity distance, PROC SIMILARITY provides several important plots to visualize the data in SAS to facilitate analysis and interpretation. The difference in shape can be captured and quantified by a process called Dynamic Time Warping (DTW). Scale lines are drawn between two time series, resulting in a need for additional lines when difference in shape occurs. The similarity distance is sum of the length of the lines, so shapes with only a few, minor differences have a smaller distance than time series with a greater difference in shape. The "warping" in the name refers to non-linear re-scaling of the time axis to find the optimal (shortest) path independent of changes in speed. For example, DTW works on EKG heartbeat time series data, even though heartbeats can different durations.

This Warp Plot shows the lines describing the difference in shape between two time series. The sum of the lengths of the line segments gives the Similarity Distance.



This Cost Plot shows a different way visualizing the difference in shape between two time series. Blue areas indicate areas of greater similarity while red indicates more difference – for example, at the end of these two time series where the upper one declines monotonically while the lower one shows a second peak.



Data Preparation

Some quantities are highly variable, making classification using cluster analysis more difficult. Fields with larger variances are given more weight in determining clusters. Also, if average value of a field steadily increases or decreases over time, the weight that field is given will also change. Use of PROC STANDARD before similarity analysis corrects for weighting by normalizing fields to the total amount of variation for each field in the data set.

Other issues are commonly found. Commodity prices can be subject to inflation over time, requiring normalization to some standard. The time series of some quantities can be autocorrelated. For example, values in the short-term future may be more strongly correlated to values in the recent past than to those in the more distant past. In this case, the time *rate of change* of the quantity may better characterize the behavior than values observed at one moment in time. A RETAIN statement can preserve the values for price from the previous interval, allowing the percent change in these values is calculated. The calculation of ordinal week from the date of a record is also included. Gasoline prices provide an excellent example of all of the data issues. They are highly volatile and often show autocorrelation over the short term and inflation effects over the long term. The following code demonstrates treatment of gas price data before using similarity analysis.

```
**** Normalize Dept of Energy (DOE) gas prices to annual mean ****;

data work.doe;
   infile "/home/doe_price.txt" dsd dlm='09'x truncover firstobs=2;
```

```
   input date  :mmddyy10. price :8.2;
   year = year(date);
   week = round((((date + 3) - mdy(1,1,year)) / 7),1);
run;

proc sort data=work.doe;
   by year week;
run;

proc univariate data=work.doe noprint;
   by year;
   var price;
   output mean=annual_mean_price out=work.annual;
run;

data work.doe;
   merge work.doe work.annual;
   by year;
   annualized_price_index = price / annual_mean_price;
run;

**** Calculate rate of change ****;

proc sort data=work.doe;
   by year week;
run;

data work.doe;
   set work.doe;
   by year week;
   retain pw_price_index; output;
   if first.week then pw_price_index = annualized_price_index;
run;

data work.doe;
   set work.doe;
   by year week;
   pct_change = (annualized_price_index - pw_price_index) * 100;
run;
```

PROC STANDARD manages highly volatile data by transforming the values of a field to the number of standard deviations above or below the mean. Apply PROC STANDARD to highly variable data before modeling can greatly improve the quality of model results.

```
proc standard data=work.doe mean=0 std=1 out=work.doe_stan;
   var week annualized_price_index pct_change;
run;
```

Calculating the Similarity Distance with PROC SIMILARITY

Classification using Similarity Analysis is performed with the following steps:

1. Use PROC SIMILARITY to measure the difference between the time series, calculate a Similarity Matrix and produce data visualizations supporting investigate the differences between time series.
2. Use PROC CLUSTER on the Similarity Matrix to identify clusters of time series with similar forms.

As an example, the process is performed on a set of supernova explosions captured by the Sloan Digital Sky Survey (SPSS). Each time series, known to astronomers as a Light Curve, records the changing brightness of the supernova over time.

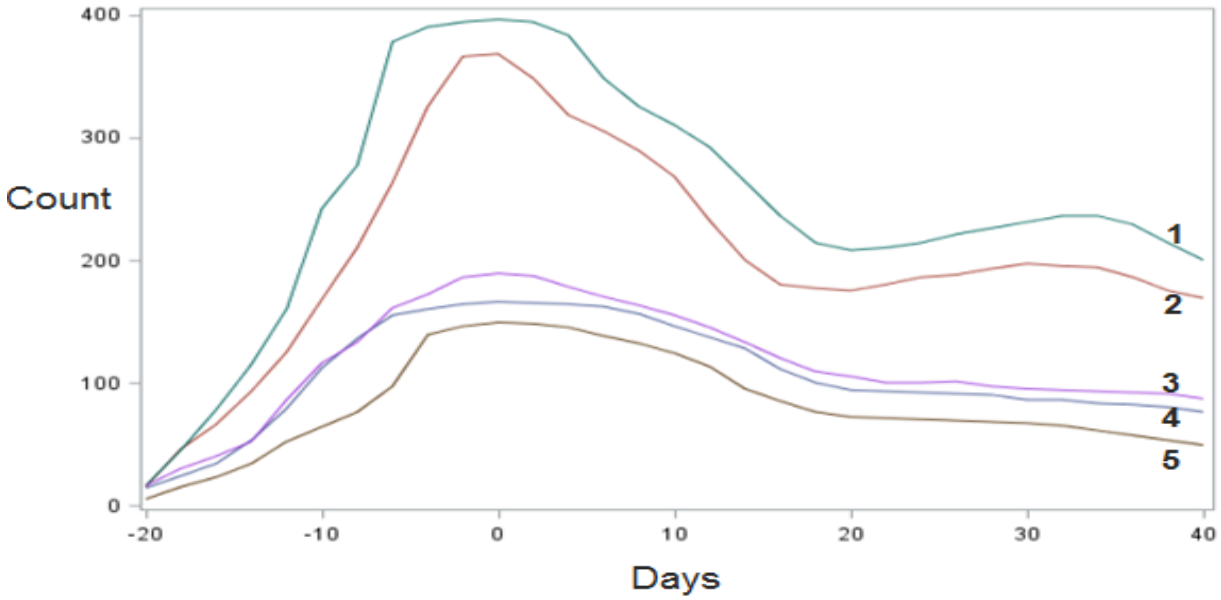
```
proc similarity data=sn_splines out=_null_ outsum=simmatrix;
   target sn1--sn39;
run;
```

The PROC SIMILARITY performs DTW and calculates the similarity distance for all time series included in the procedure. These are collected into a similarity matrix, as specified by statement OUTSUM=simmatrix:

	TS1	TS2	TS3	TS4	TS5	TS6
TS1	0	817	100	469	338	307
TS2	817	0	422	644	871	895
TS3	100	422	0	92	835	818
TS4	469	644	92	0	56	206
TS5	338	871	835	56	0	378

	TS1	TS2	TS3	TS4	TS5	TS6
TS6	307	895	818	206	378	0

The matrix shows the amount of difference in shape between any two time series. For example, TS4 and TS5 are very similar in shape, so they have small number (56). In contrast, TS4 is very different from TS1, resulting in a very large numbers (469). A number of visualization can be created as this stage as well, including warp and cost plots.

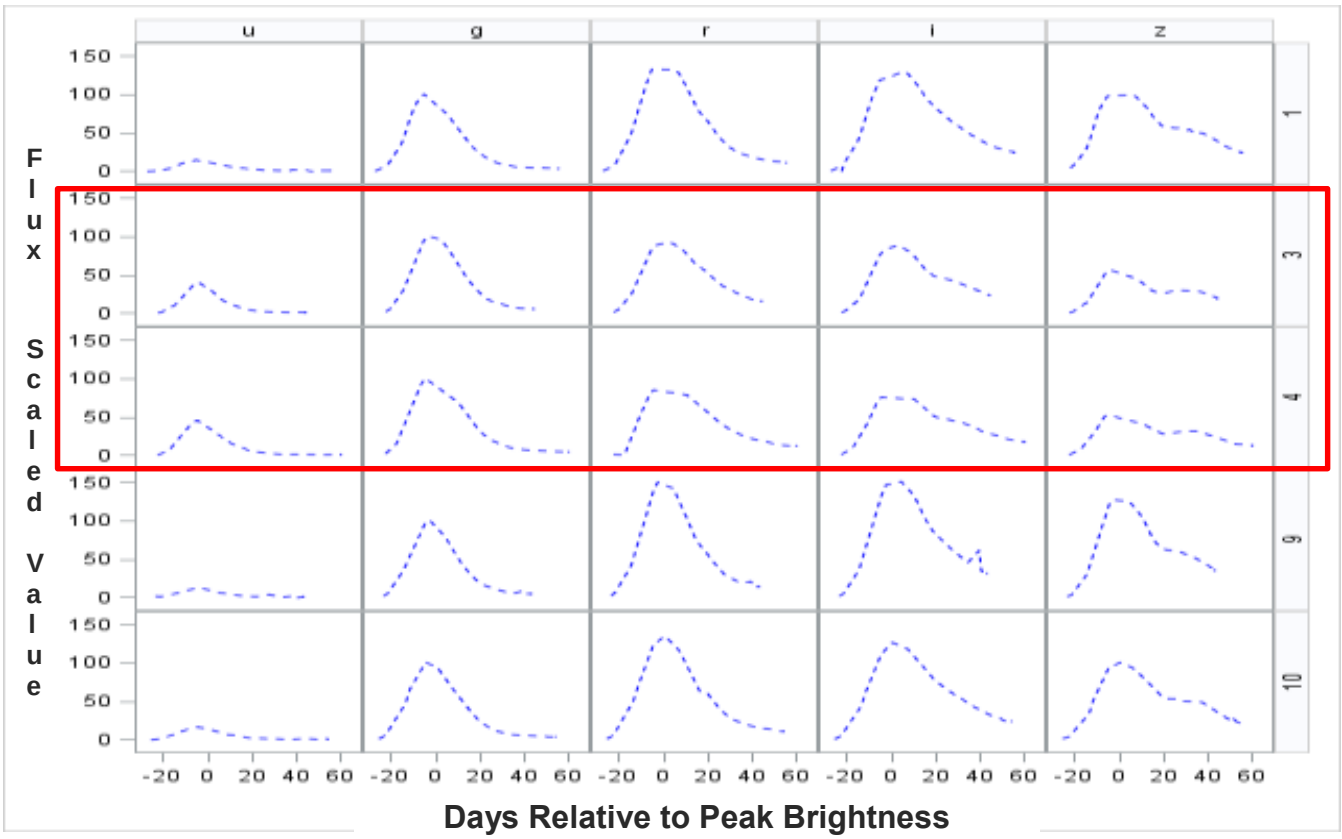


Clustering the Time Series by Similarity Distance

Once the similarity distances have been calculated, the time series can be clustered using standard methods. All time series with sufficient data to describe the shape in detail are used to develop the clusters – in this example, 78 time series in all.

```
proc cluster data=simmatrix(drop=_status_) method=ward noprint;
   id _input_;
run;

proc sgpanel data=all_filters;
   title 'SDSS Supernova Survey Light Curves';
   panelby sn flux_order / columns = 5 rows =5;
   series x=mjd_datetime y=flux;
   where sn in (1,3,4,9,10);
run;
```



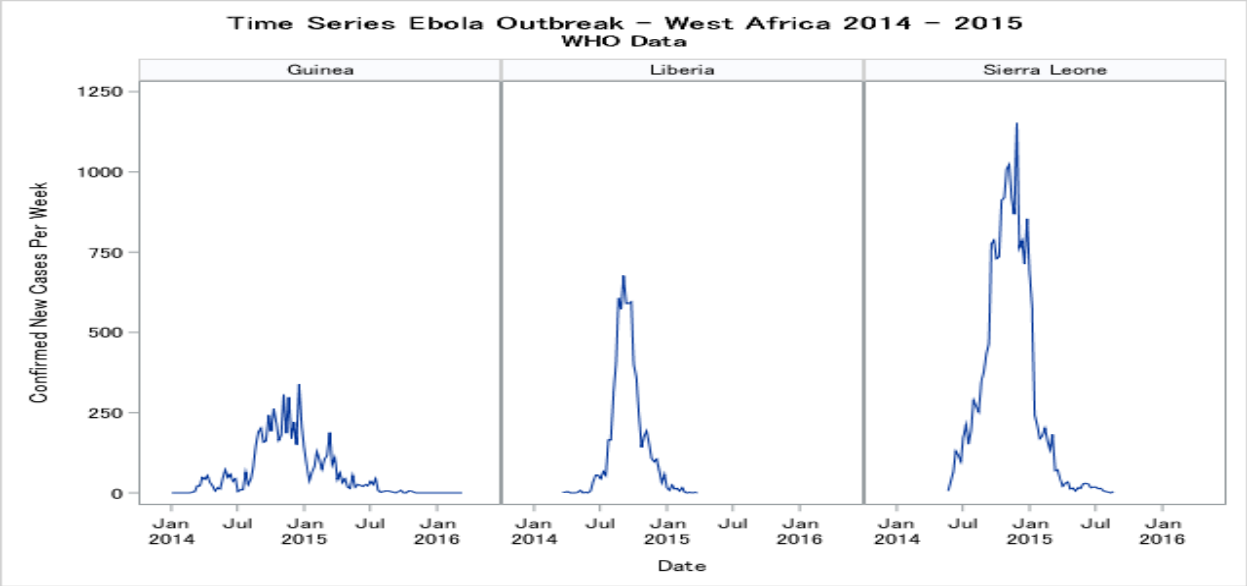
Supernova Classification Using Similarity Analysis

Once the clustering is complete, additional characteristics of the clusters may be identified. This trellis plot displays data for the five supernovae plotted earlier. In addition to the infrared light (z, in the column on the right) used for the similarity analysis, four other colors of light are plotted: u (ultraviolet), g (green), r (red), and I (near infrared). The similarity analysis using all the time series with sufficient data yields two clusters, one for those with a secondary peak in the far infrared one without. When all the data are plotted in the panel, additional characters of the clusters can be identified. In this case, the g, r, and i columns are very similar. However, the ultraviolet u column on the left shows a large peak in the first cluster with minimal increases in the u data for the other cluster. It may be that the mechanism or characteristics responsible for the secondary far infrared peak that first attracted attention is also responsible for the much stronger ultraviolet radiation seen in this sub-type.

An Example in Biostatistics

Similarity Analysis can be applied in many areas of research where there is a need to classify time series. One instance from biostatistics is found in epidemiology, where the spread of an epidemic can be compared to other previous of current events. The 2014 – 2015 Ebola outbreak in western African nations of Liberia,

Sierra Leone and Guinea provides a useful illustration of this method. While the Ebola outbreak affected all three nations, each displayed unique characteristics. Publicly available data from the World Health Organization was used in this example.




```
PROC SGPANEL DATA = ebola;
   PANELBY Country / novarname columns=3;
   SERIES X = date Y = Count;
   FORMAT date mmyy.;
   TITLE 'Time Series Ebola Outbreak - West Africa 2014 - 2015';
   TITLE2 'WHO Data';
RUN;

PROC SIMILARITY DATA=ebola OUT=ebola_out OUTSUM=similarity_outsum;
   ID date INTERVAL=week accumulate=total;
   INPUT Liberia SierraLeone Guinea;
RUN;
```

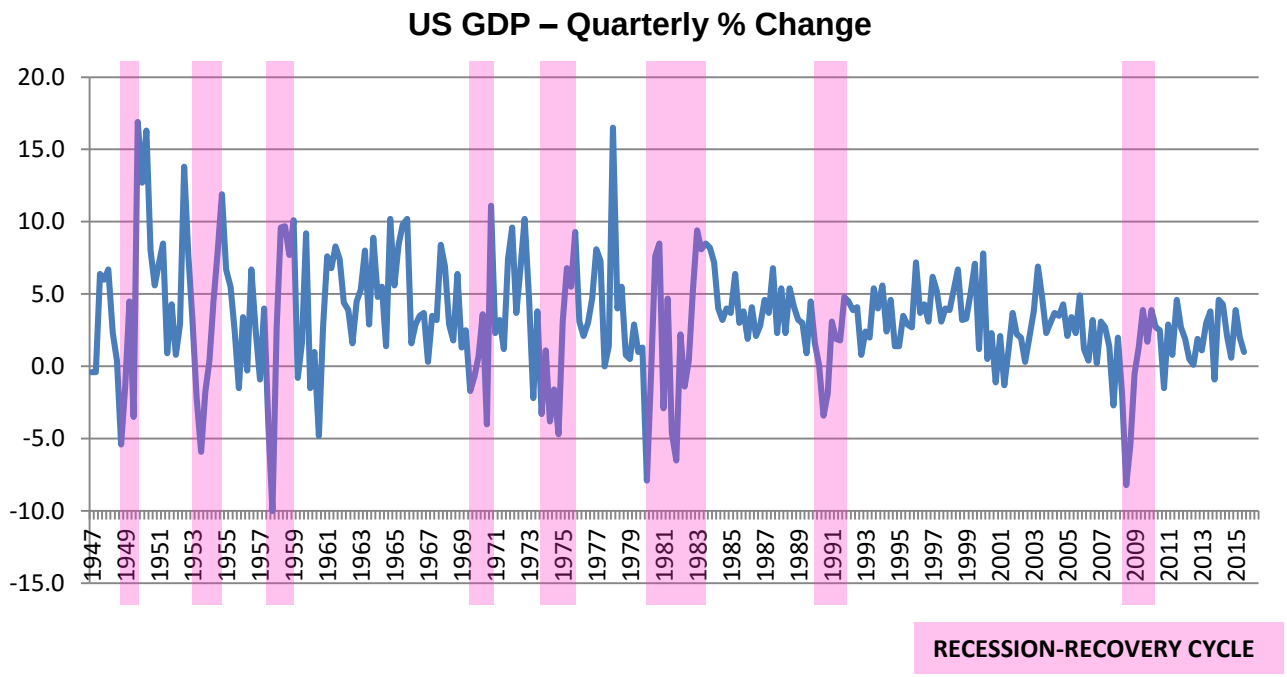
While superficially having much in common, similarity analysis reveals important differences. Here is the resulting similarity matrix:

	Liberia	Sierra Leone	Guinea
Liberia	0	841049	538459
Sierra Leone	841049	0	4632704
Guinea	538459	4632704	0

The time series of the unfolding epidemic in Liberia and Guinea are found to be much more similar to each other than to that in Sierra Leone. Careful study of the graph finds Liberia and Guinea have some noise at first, followed by a sudden, dramatic increase in the number of cases. This contrasts with the pattern in Sierra Leone, where a more gradual increase is seen at the start of the outbreak. A cost plot can be very useful in ascertaining exactly where the differences are greatest.

Econometrics: Time Series Patterns in Recession and Recovery

The same process can be used to consider questions in economics and econometrics. For examples, the United States economy has experienced eight recession and recovery cycles since 1945, defining a recession in the usual way as two consecutive quarterly decreases in GDP. The comparison and contrast provided by Similarity Analysis, utilizing official US government data reported by the Bureau of Economic Analysis (BEA), provides important insights into the nature of these events and their patterns over time. In this investigation, the Recovery part of the cycle is captured by including four consecutive quarters in increasing GDP following the last decrease in the preceding Recession.




```
PROC SIMILARITY DATA=gdp OUT=gdp_out OUTSUM=similarity_outsum;
   ID date INTERVAL=qtr accumulate=total;
   INPUT r1_gdp_change r2_gdp_change r3_gdp_change r4_gdp_change
         r5_gdp_change r6_gdp_change r7_gdp_change r8_gdp_change;
RUN;
```

The resulting similarity matrix provides a level of detail supporting comparison of the different economic cycles beyond what may be apparent from a plot of the raw numbers:

	Cycle1	Cycle2	Cycle3	Cycle4	Cycle5	Cycle6	Cycle7	Cycle8
Cycle1	0	193.67	185.09	184.12	222.47	341.67	392.48	532.5
Cycle2	193.67	0	71.34	233.86	34.16	280.74	79.89	115.73
Cycle3	185.09	71.34	0	229.11	57.79	287.56	121.01	160.87
Cycle4	184.12	233.86	229.11	0	176.47	317.19	108.08	133.58
Cycle5	222.47	34.16	57.79	176.47	0	281.18	35.96	85.93
Cycle6	341.67	280.74	287.56	317.19	281.18	0	256.74	219.47
Cycle7	392.48	79.89	121.01	108.08	35.96	256.74	0	65.64
Cycle8	532.5	115.73	160.87	133.58	85.93	219.47	65.64	0

With this similarity analysis, Cycles 2, 3, 5, and 7 form a cluster, with small numbers for the similarity distance between each other and generally larger distances between other cycles. This cluster, containing many members, indicates a common pattern among Recession – Recovery cycles. Cycles 1, 4, and 6 are unlike all other members. This is how outliers appear is a cluster analysis, presenting clusters with very few members. Cycle 6 is noteworthy for containing a “Double Dip” recession, the only example in these data. Similarity analysis has identified this outlier. The most recent cycle, including the “Great Recession”, is found to have similarities with the recession and recovery of the early 1990’s. This instance points out an important characteristic of similarity analysis: this method clusters time series by *shape, not intensity*. The 1990-1991 and the 2008-2010 recession were very different in terms of magnitude. However, they were similar in shape, with a sharp, rapid recession followed by a slow recovery in both cases.

Conclusions

Similarity analysis is an emerging method for the classification of time series. It uses a distance measure to quantify the difference in form or shape between the two series, with a lower number indicating greater similarity. The similarity distances for a set of time series can be arranged into a similarity matrix. Classification of time series can be achieved using standard clustering methods on the similarity distance.

In the example, SDSS light curves from a set of Type 1a supernovae indicates the presence of two sub-types, one with a fast initial growth rate and a secondary peak in the infrared z-range while the second sub-type has slower initial growth rate and no secondary infrared peak. Classification of the time series in this way led to the discovery of other differences between the two types, including a larger peak in in the ultraviolet in the first sub-type.

Summary of Best Practices for Similarity Analysis

- When clustering, examine both point in time variables and rate of change data.

- When data are “unevenly spaced”, that is, the time interval between observations is not constant, treat as an evenly spaced time series with missing data and impute the missing values.
- Where highly volatile variables are present, consider using PROC STANDARD before calculating the similarity distance with PROC SIMILARITY.
- Once the similarity distances have been calculated, standard clustering best practices still apply.
- Watch out for outliers, which will appear as time series clusters with very few members.

ACKNOWLEDGMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J Corliss
david.corliss@wayne.edu
dcorliss@gmail.com