SAS586-2017

Automatic Singular Spectrum Analysis and Forecasting

Michael Leonard and Bruce Elsheimer
SAS Institute Inc., Cary, NC, USA

ABSTRACT

The singular spectrum analysis (SSA) method of time series analysis applies nonparametric techniques to decompose time series into principal components. SSA is particularly valuable for long time series, in which patterns (such as trends and cycles) are difficult to visualize and analyze. An important step in SSA is determining the spectral groupings; this step can be automated by analyzing the *w*-correlations (weighted correlations) of the spectral components. This paper provides an introduction to singular spectrum analysis and demonstrates how to use SAS/ETS® software to perform it. To illustrate, monthly data on temperatures in the United States for about the last 100 years are analyzed to discover significant patterns.

INTRODUCTION

Time series data often contain trends, cycles, anomalies, and other components. For long time series, these patterns are often difficult to visualize and discover. Singular spectrum analysis (SSA) applies nonparametric techniques that adapt the commonly used principal component analysis (PCA) for decomposing time series data. The principal components can help you discover and understand the various patterns that the time series contains. After you understand each of these component series, you can model and forecast them separately; then you can aggregate the component series forecasts in order to forecast the original series under investigation. SSA requires grouping of the eigenspectrum. In the past, this grouping was performed manually. Based on *w*-correlation analysis, the spectral grouping can be performed automatically.

BACKGROUND

This section provides a brief theoretical background on singular spectrum analysis. It is intended to provide the analyst with motivation, orientation, and references. An introductory discussion of singular spectrum analysis can be found in Golyandina, Nekrutkin, and Zhigljavsky (2001) and in Elsner and Tsonis (1996). This section extends the discussion found in Leonard, Elsheimer, and Kessler (2010).

Traditional Singular Spectrum Analysis

Given a time series y_t for $t = 1, \dots, T$ and a window length $2 \leq L < T/2$, singular spectrum analysis decomposes the time series into spectral groupings by using the following steps:

1. **Embedding step:** Using the time series, form a $K \times L$ trajectory matrix $X = \{x_{k,l}\}_{k=1,l=1}^{K,L}$ such that $x_{k,l} = y_{(k-l+1)}$ for $k = 1, \dots, K$ and $l = 1, \dots, L$, where $K = (T - L + 1)$. By definition, $L \leq K < T$ because $2 \leq L < T/2$.

2. **Decomposition step:** Apply singular value decomposition to the trajectory matrix $X = UQV$, where U represents the $(K \times L)$ matrix that contains the left-hand-side (LHS) eigenvectors, Q represents the diagonal $(L \times L)$ matrix that contains the singular values, and V represents the $(L \times L)$ matrix that contains the right-hand-side (RHS) eigenvectors.

   Therefore, $X = \sum_{l=1}^{L} X^{(l)} = \sum_{l=1}^{L} u_l q_l v_l'$, where $X^{(l)}$ represents the $(K \times L)$ principal component matrix, u_l represents the $(K \times 1)$ left-hand-side (LHS) eigenvector, q_l represents the singular value, and v_l represents the $(L \times 1)$ right-hand-side (RHS) eigenvector that is associated with the *l*th window index.

3. **Grouping step:** For each group index, $m = 1, \dots, M,$ define a group of window indices $I_m \subset \{1, \dots, L\}$. Let $X_{I_m} = \sum_{l \in I_m} X^{(l)} = \sum_{i \in I_m} u_l q_l v_l'$ represent the grouped trajectory matrix for group I_m.

   Note that if groupings represent a spectral partition, $\cup_{m=1}^{M} I_m = \{1, \dots, L\}$, and $I_m \cap I_n = \emptyset$ for all $m \neq n$, then according to the singular value decomposition theory, $X = \sum_{m=1}^{M} X_{I_m}$.

4. **Averaging step:** For each group index, $m = 1, \ldots, M$, compute the diagonal average of

   $$X_{I_m} = \left\{x_{k,l}^{(m)}\right\}_{k=1,l=1}^{K,L}, \ \tilde{x}_t^{(m)} = \frac{1}{n_t}\sum_{l=s_t}^{e_t} x_{(t-l+1),l}^{(m)}$$

   where $s_t = 1, e_t = t, n_t = t$ for $(1 \le t < L)$
   $s_t = 1, e_t = L, n_t = L$ for $(L \le t \le (T - L + 1)$
   $s_t = (T - t + 1), e_t = L, n_t = (T - t + 1)$ for $((T - L + 1) < t \le T)$

   Note that if groupings represent a spectral partition, $\cup_{m=1}^{M} I_m = \{1, \ldots, L\}$, and $I_m \cap I_n = \emptyset$ for all $m \ne n$, then $y_t = \sum_{m=1}^{M} \tilde{x}_t^{(m)}$ by definition. Hence, singular spectrum analysis additively decomposes the original time series, y_t, into m component series: $\tilde{x}_t^{(m)}$ for $m = 1, \ldots, M$.

5. **Forecasting step (*optional*):** If the groupings represent a spectral partition, then each component series, $\tilde{x}_t^{(m)}$ for $m = 1, \ldots, M$, can be modeled and forecasted independently using an appropriate time series model (ARIMAX, unobserved component model, exponential smoothing model, and others), possibly using different time series models that include different input series (causal factors) and calendar events (interventions).

   Let $\hat{x}_t^{(m)}$ for $m = 1, \ldots, M$ represent the *component series forecasts* that are derived from the *m*th independent time series model. Then the forecast for the original time series, $\hat{y}_t$, can be derived by simply aggregating the component series forecasts: $\hat{y}_t = \sum_{m=1}^{M} \hat{x}_t^{(m)}$.

   The SSA forecasting step represents a clever forecast model combination technique.

Automatic Spectral Grouping

An important step in SSA is specifying the groups, $I_m \subset \{1, \ldots, L\}$ for $m = 1, \ldots, M$. In order to automate the SSA grouping step, the *w*-correlations are computed to form an $(L \times L)$ *w*-correlations matrix, assuming the maximal number of groups: $M = L$.

$$P^{(w)} = \left\{\rho_{i,j}^{(w)}\right\}_{i=1,j=1}^{L,L}$$

where $\rho_{i,j}^{(w)} = \frac{\left(\tilde{x}_t^{(i)}, \tilde{x}_t^{(j)}\right)_w}{\left\|\tilde{x}_t^{(i)}, \tilde{x}_t^{(i)}\right\|_w \left\|\tilde{x}_t^{(j)}, \tilde{x}_t^{(j)}\right\|_w}$, $\left(\tilde{x}_t^{(i)}, \tilde{x}_t^{(j)}\right)_w = \sum_{t=1}^{T} w_t \tilde{x}_t^{(i)}, \tilde{x}_t^{(j)}$, and $w_t = min(t, L, T - 1)$

The following steps are performed in SSA autogrouping:

1. Initially assume the maximal number of groups: $M = L$.
2. Diagonally average the groups as described previously: $\tilde{x}_t^{(m)}$ for $m = 1, \ldots, L$.
3. Compute the *w*-correlations between groups: $\rho_{i,j}^{(w)}$.
4. Choose the groups based on the *w*-correlations for which the absolute values are near 1. Or more formally,

   $I_m \subset \{1, \ldots, L\}$ such that $\left|\rho_{i,j}^{(w)}\right| \approx 1$ whenever $i, j \in I_m$

After the groups have been chosen based on the *w*-correlation analysis, group according to step 3, average diagonally according to step 4, and optionally forecast according to step 5.

SAS IMPLEMENTATION

The singular spectrum analysis described in the previous section can be performed using SAS/ETS software. This section describes how the TIMESERIES procedure analyzes timestamped and time series data.

PROC TIMESERIES Statement

The PROC TIMESERIES statement has the following options that are related to SSA (for options related to other analyses, see the *SAS/ETS User's Guide*):

OUTSSA=*SAS-data-set*

names the output data set to contain the singular spectrum analysis result series.

PLOTS=SSA

plots the singular spectrum analysis results.

PRINT=SSA

prints the singular spectrum analysis results.

SSA STATEMENT

You can use the new SSA statement in the TIMESERIES procedure to specify options that are related to singular spectrum analysis (SSA) of the accumulated time series. Only one SSA statement is allowed.

The SSA statement has the following syntax:

SSA < / options **> ;**

You can specify the following options in the SSA statement following the slash (/):

LENGTH=*number*

specifies the window length to be used in the analysis, where *number* represents the maximum lag used in the SSA trajectory matrix. The *number* must be greater than 1 and less than 1,000. When the SEASONALITY= option is specified or implied by the INTERVAL= option in the ID statement, the default window length is the smaller of two times the length of the seasonal cycle and one-half the length of the time series. When no seasonality value is available, the default window length is the lesser of 12 and one-half the length of the time series.

For example, the following SSA statement specifies a window length of 10:

```
ssa / length=10;
```

As another example, the following SSA statement specifies a window length of 24 if the INTERVAL=MONTH or SEASONALITY=12 option is specified:

```
ssa;
```

If the specified window length is greater than what is feasible based on one-half the length of the accumulated time series, the window length is reduced and a warning message is printed to the log.

THRESHOLD=*percentage*

specifies the threshold value used to determine the size of the last group based on the cumulative percentage of the singular values. The *percentage* must be greater than 0 and less than 100. The default is 90% (THRESHOLD=90).

For example, the following SSA statement specifies a threshold of 80%:

```
ssa / threshold=80;
```

As another example, the following SSA statement specifies a threshold of 90%:

```
ssa;
```

The size of the last group must be at least 1 but less than the window length, and the threshold is adjusted to achieve this requirement.

For example, the following SSA statement specifies a threshold of 0% and implies that the size of the last group is 1 less than the window length:

```
ssa / threshold=0;
```

As another example, the following SSA statement specifies a threshold of 100% and implies that the size of the last group is 1:

```
ssa / threshold=100;
```

GROUPS=*(numlist) …(numlist)*

specifies the list of groups of window lags to be stored in the OUTSSA= data set or plotted. The window lags must be separated by spaces or commas. For example, GROUPS=(1 3) (2 4) specifies that the first and third window lags form the first group and the second and fourth window lags form the second group. The default is to evenly divide the window into two groups based on the window length, which is specified in the LENGTH= option.

For example, the following SSA statement specifies three groups:

```
ssa / groups=(1 3)(2 4 5)(6);
```

The first group contains the first and third principal components; the second group contains the second, fourth, and fifth principal components; and the third group contains the sixth principal component.

For example, the following SSA statement specifies two groups:

```
ssa;
```

The first group contains the principal components whose spectra sum to greater than the threshold of 90%; the second group contains the remaining principal components.

GROUPS=AUTO*(number)*

specifies the maximal number of groups to be retained when automatic grouping is used. When the GROUPS=AUTO(*number*) option is specified, groups are created automatically based on the *w*-correlations.

Specifying the Window Length

You can explicitly specify the maximum window length, $2 \leq L \leq 1000$, by using the LENGTH= option in the SSA statement, or you can implicitly specify the window length by using the INTERVAL= option in the ID statement or the SEASONALITY= option in the PROC TIMESERIES statement.

In any case, the window length is reduced based on the accumulated time series length T to enforce the requirement that $2 \leq L < T/2$.

Specifying the Groups

You can explicitly specify the grouping $I_m \subset \{1,...,L\}$ by using the GROUPS= option in the SSA statement, or you can implicitly specify the grouping by using the THRESHOLD= option in the SSA statement. The THRESHOLD= option is useful for removing noise or less dominant patterns from the accumulated time series.

Let $(0 < \alpha < 1)$ be the cumulative percentage singular value threshold. Then I_M (the last group) is determined by the following threshold:

$$\min_{(l_\alpha - 1)} \left(\left(\sum_{l=1}^{l_\alpha} q_l \Big/ \sum_{l=1}^{L} q_l \right) \geq \alpha \right) \text{ where } I_M = \{l_\alpha,...,L\} \text{ where } 1 < l_\alpha \leq L$$

Using this rule, the last group $I_M = \{l_\alpha,...,L\}$ describes the least dominant patterns in the time series, and the size of the last group is at least 1 and is less than the window length, $L \geq 2$.

MANUAL SPECTRAL GROUPING EXAMPLE

To illustrate the use of SSA in SAS/ETS software, monthly data on US temperatures for about the last 100 years are analyzed to find significant patterns. The analysis of this example illustrates how spectral grouping is manually performed. This example is found in Leonard, Elsheimer, and Kessler (2010) and is repeated here for convenience.

BASIC TIME SERIES ANALYSIS

The monthly temperature anomaly (in degrees Celsius) for the United States over the last 128 years, which was provided by the National Oceanic Atmospheric Administration (NOAA), was analyzed. The temperature anomaly is seasonally adjusted by using the reference decade of the 1960s. First, the time series was plotted using the TIMESERIES procedure as follows:

```
proc timeseries data=NOAA out=_NULL_ plot=(SERIES CYCLES);
   id DATE interval=MONTH;
   var TEMPERATURE;
run;
```

DATA=NOAA specifies that the data set Work.NOAA contains the temperature anomaly records. The ID statement specifies that the time ID variable is DATE and the time interval is MONTH. The VAR statement specifies that the variable under analysis is TEMPERATURE. The PLOT=(SERIES CYCLES) option plots the series and the year-over-year monthly cycles.

Figure 1 illustrates the SERIES plot. The X axis represents the time ID (DATE), and the Y axis represents the temperature anomaly (TEMPERATURE). As you can see from this plot, it is difficult to see any patterns in the time series because of its length and variation.

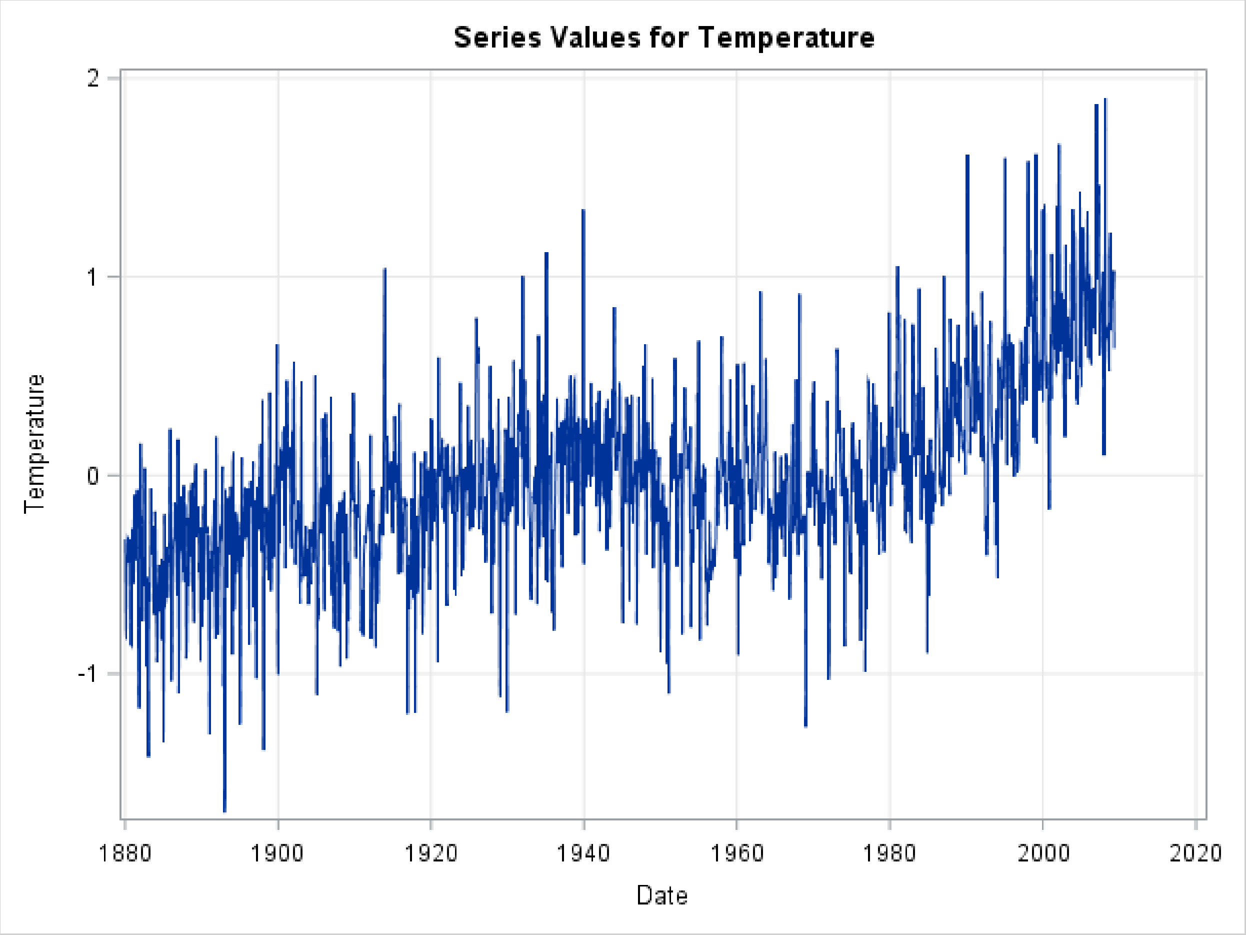

Figure 1. Monthly Series Plot of the Temperature Anomaly

Figure 2 illustrates the year-over-year monthly cycles plot (CYCLES). The X axis represents the monthly seasonal index (January=1, …, December=12), and the Y axis represents the temperature anomaly (TEMPERATURE). Each line represents one year (128 seasonal cycles). As you can see from this plot, the series has no discernible monthly pattern—as expected, because the time series is seasonally adjusted.

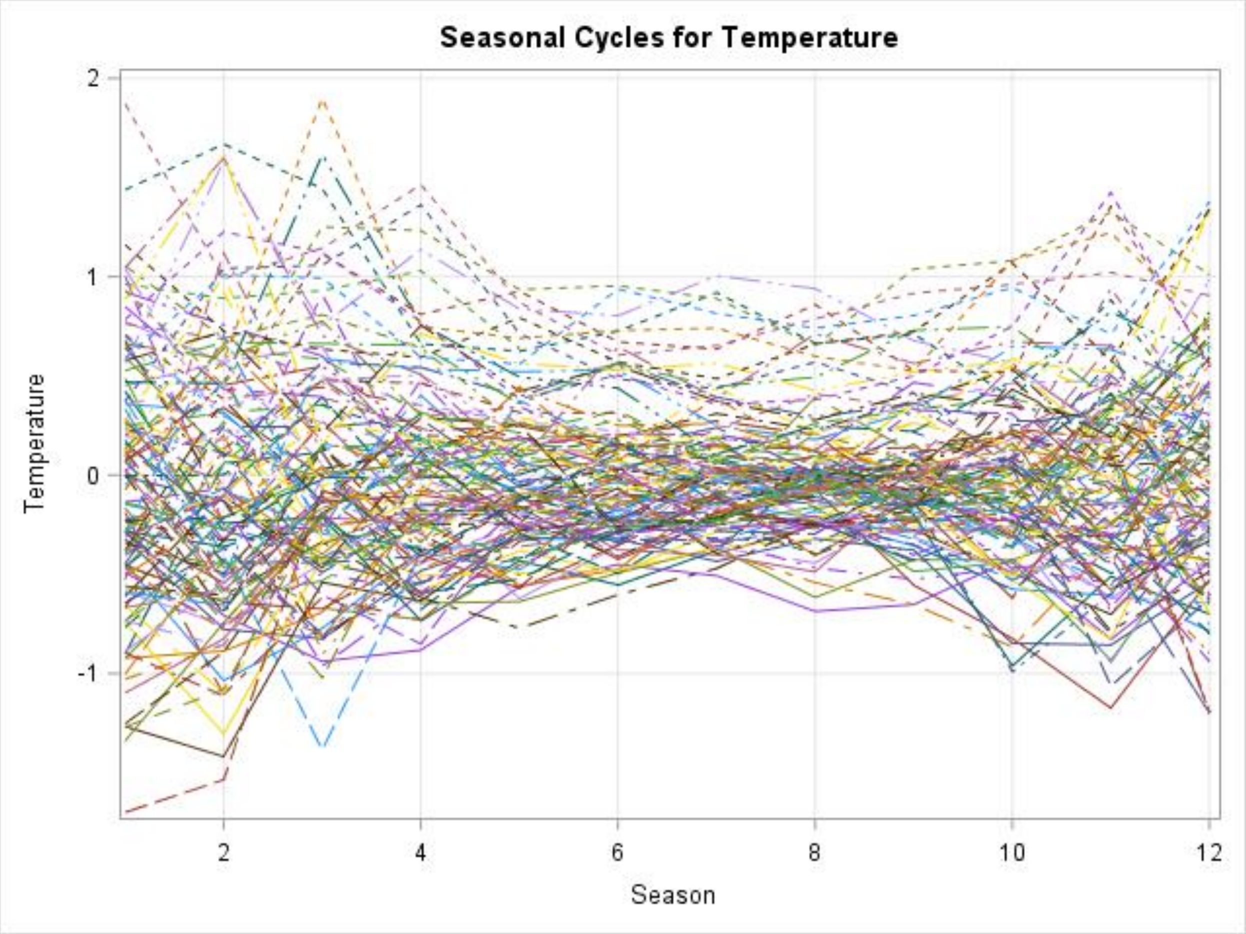

Figure 2. Seasonal Cycles Plot of the Temperature Anomaly

SINGULAR SPECTRUM ANALYSIS

Next, singular spectrum analysis is applied using a threshold value for the eigenspectrum. The time series is analyzed using the TIMESERIES procedure as follows:

```
proc timeseries data=noaa out=_NULL_ plot=(series cycles SSA);
   SSA / LENGTH=120 THRESHOLD=80;
   id date interval=month;
   var temperature;
run;
```

The LENGTH=120 option in the SSA statement specifies a window length of 120 (10 years), and the THRESHOLD=80 option specifies an eigenspectrum threshold value of 80%. Including SSA as one of the values in the PLOT= option plots the SSA analysis.

Figure 3 illustrates the eigenspectrum plot. The upper graph illustrates the eigenspectrum (log-scale), and the lower graph illustrates the cumulative percentage of the eigenspectrum on the Y axis. The common X axis represents the window lags. As you can see from the upper graph, the eigenspectrum decreases rapidly after the seventh lag. Close inspection reveals four “steps” of equal value in the eigenspectrum plot: (1)(2)(3 4)(5 6 7).

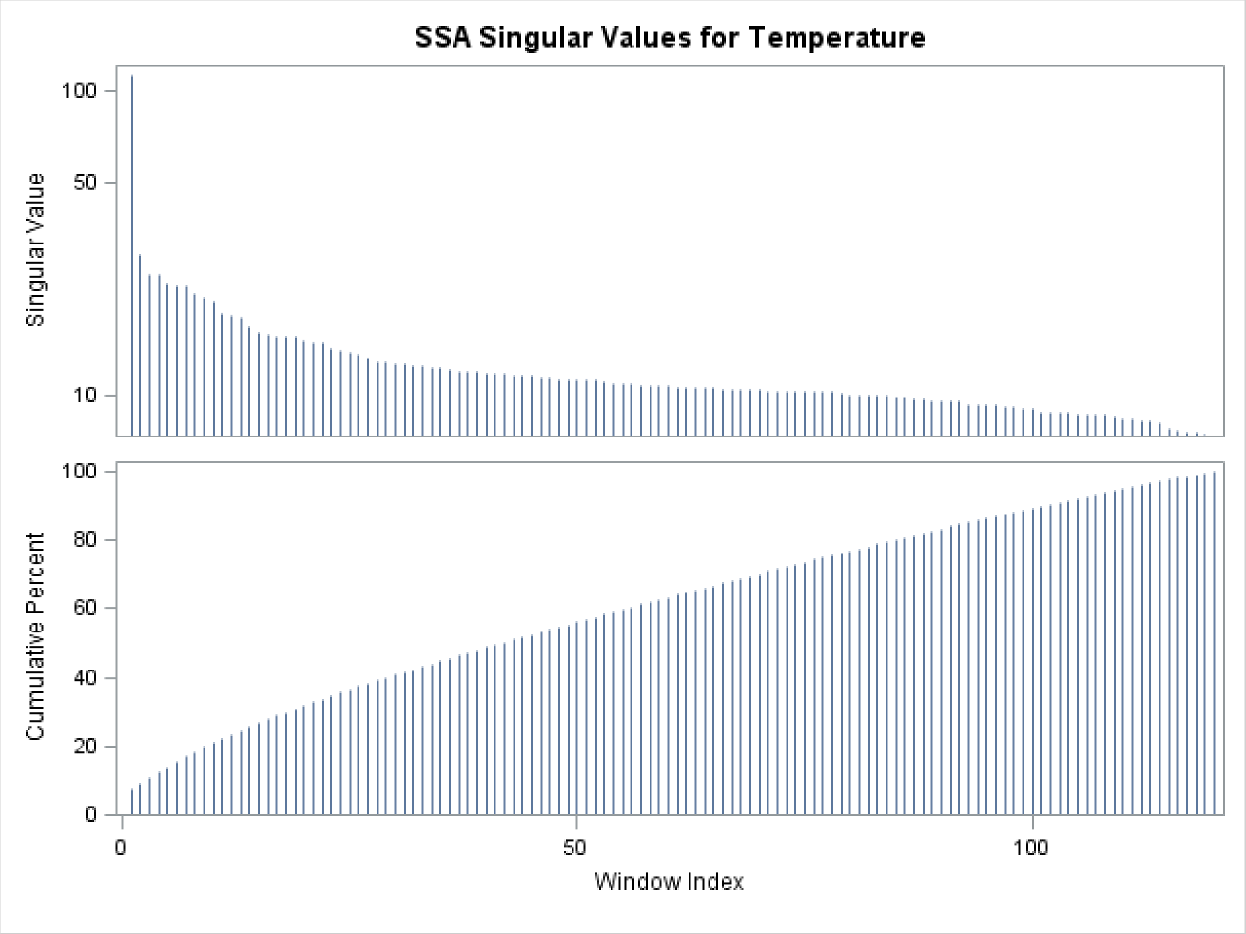

Figure 3. Eigenspectrum of the Temperature Anomaly

Next, singular spectrum analysis is applied using grouping of the eigenspectrum. The time series is analyzed using the TIMESERIES procedure as follows:

```
proc timeseries data=noaa out=_NULL_ plot=(series cycles ssa) OUTSSA=SSA;
   ssa / length=120 GROUPS=(1)(2)(3 4)(5 6 7);
   id date interval=month;
   var temperature;
run;
```

The SSA statement GROUP=(1)(2)(3 4)(5 6 7) decomposes the series into four spectral groups. The first group contains the first lag; the second group contains the second lag; the third group contains the third and fourth lags; and the fourth group contains the fifth, sixth, and seventh lags. The OUTSSA=SSA option stores the spectral grouping in the Work.SSA data set. Becuse three spectral groups are requested, the data set contains four variables (GROUP1, GROUP2, GROUP3, and GROUP4).

Figure 4 illustrates the first group. In the upper graph, the jagged blue line represents the original series, and the smooth blue line represents the first group. In the lower graph, the blue line represents the first group. As you can see from the plot, the first group represents the dominant trend in the temperature anomaly series. From these graphs, it appears that temperatures have increased about one degree over about the last 100 years.

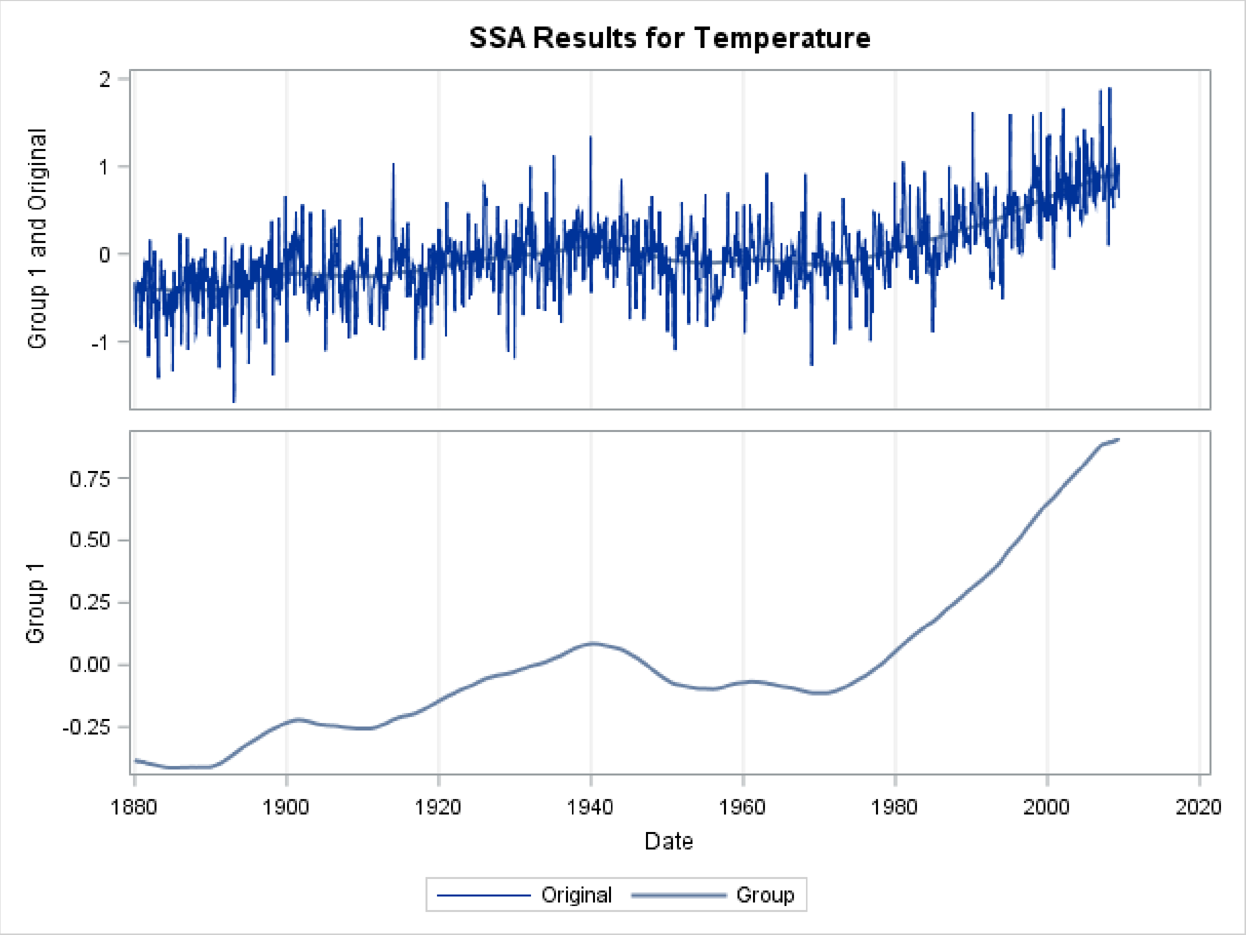

Figure 4. First Spectral Group of the Temperature Anomaly

Figure 5 illustrates the second group. As you can see from these graphs, the second group represents the dominant long-term cycle in the temperature anomaly series.

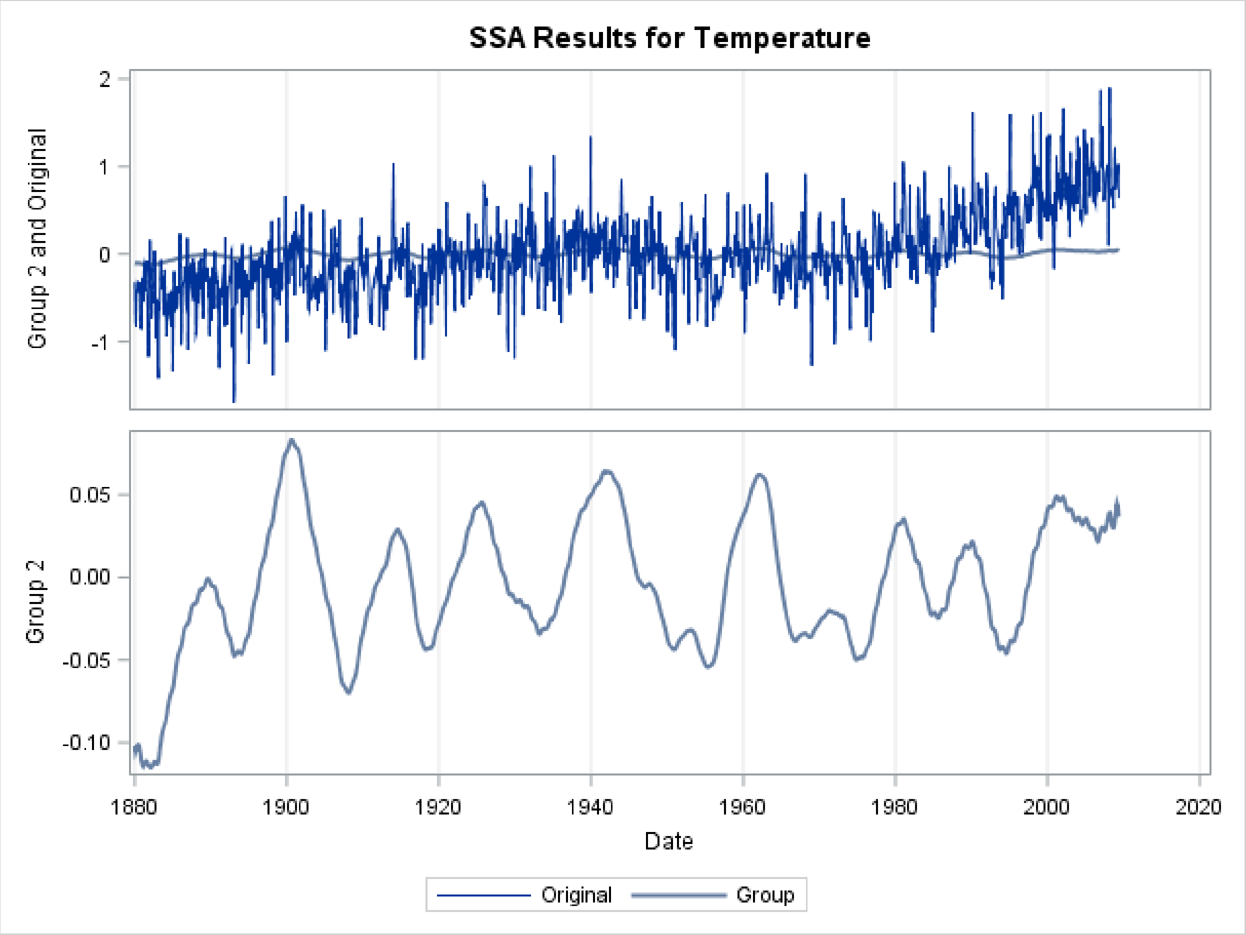

Figure 5. Second Spectral Group of the Temperature Anomaly

Figure 6 illustrates the spectral density plot for the second group. From this plot, there appears to be an approximately 22-year cycle (SEASONALITY=264), possibly related to the Hale solar cycle.

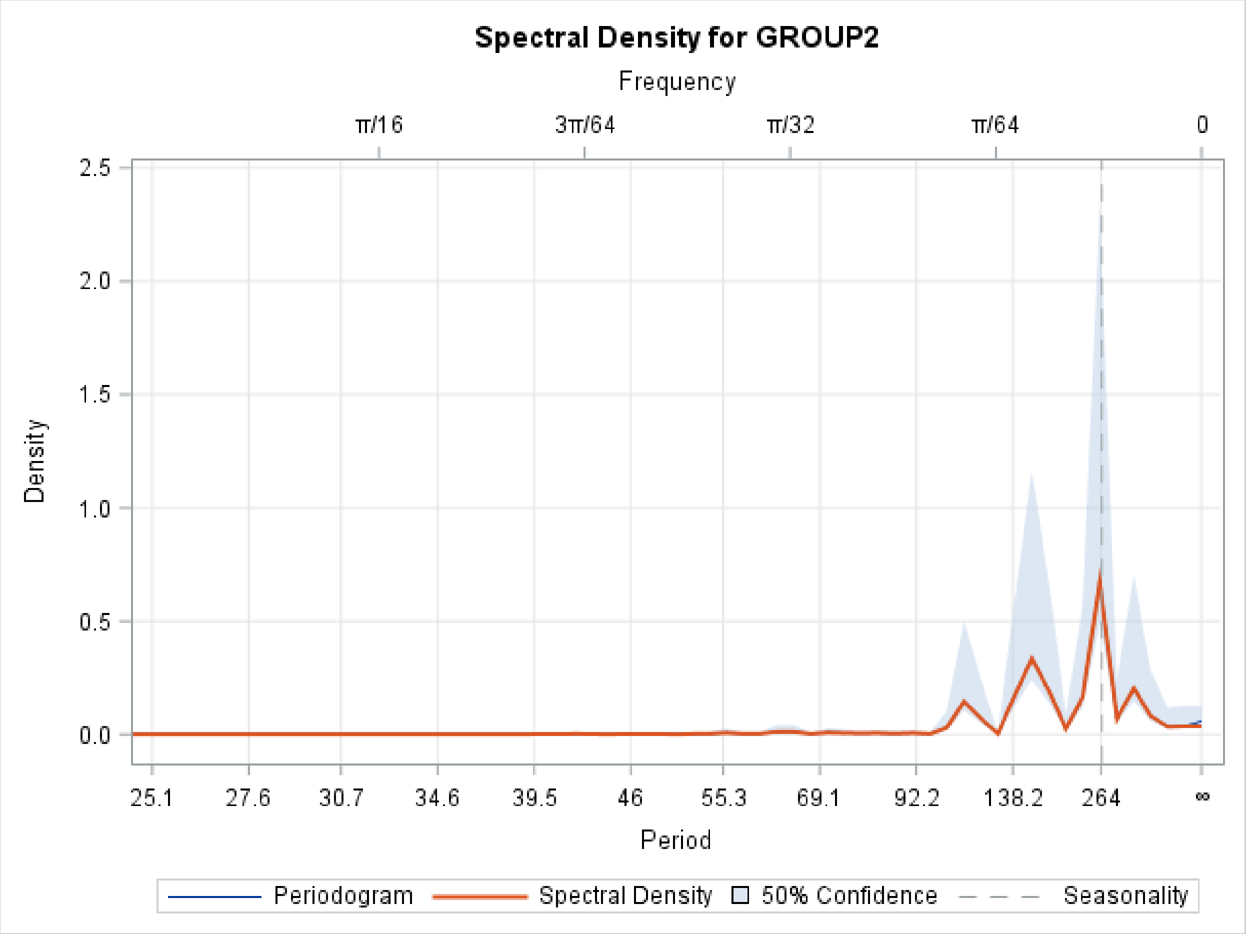

Figure 6. Spectral Density of the Second Spectral Group

Figure 7 illustrates the third group. As you can see from these graphs, the third group represents the dominant short-term cycle in the temperature anomaly series. It appears that the variation is small for the reference decade of the 1960s.

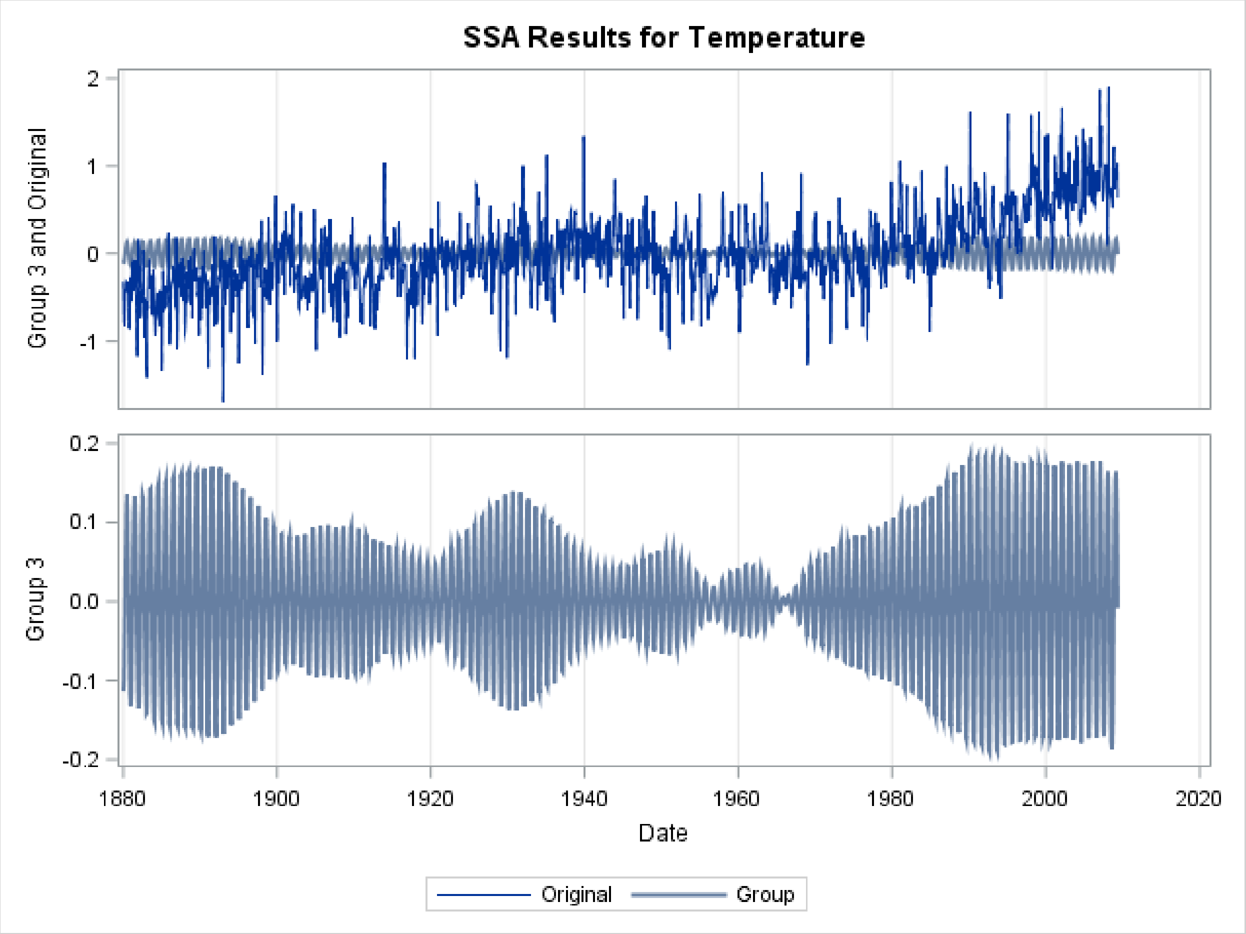

Figure 7. Third Spectral Group of the Temperature Anomaly

Figure 8 illustrates the spectral density plot for the third group. From this plot, there appears to be a monthly cycle (SEASONALITY=12). Although the time series was adjusted for monthly seasonality, there still appears to be a small remnant.

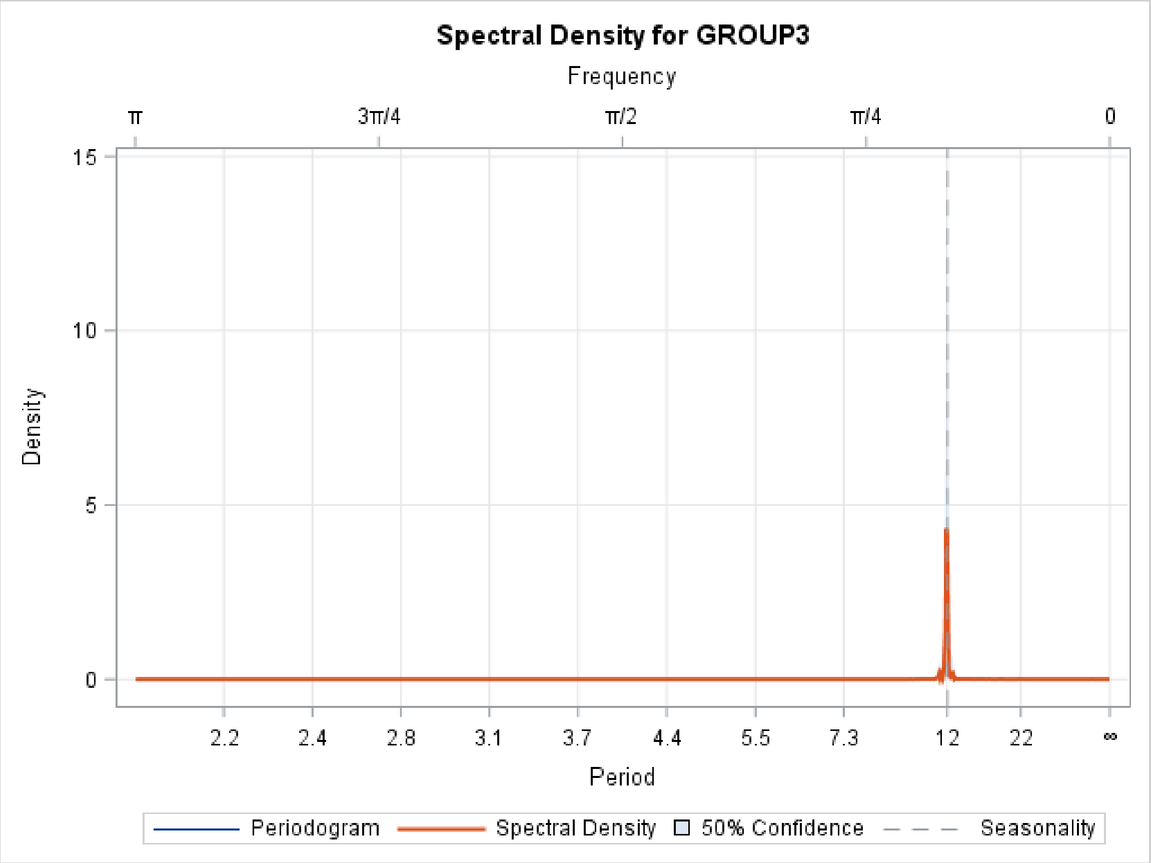

Figure 8. Spectral Density of the Third Spectral Group

Figure 9 illustrates the fourth group. As you can see from these graphs, the fourth group represents the dominant medium-term cycle in the temperature anomaly series.

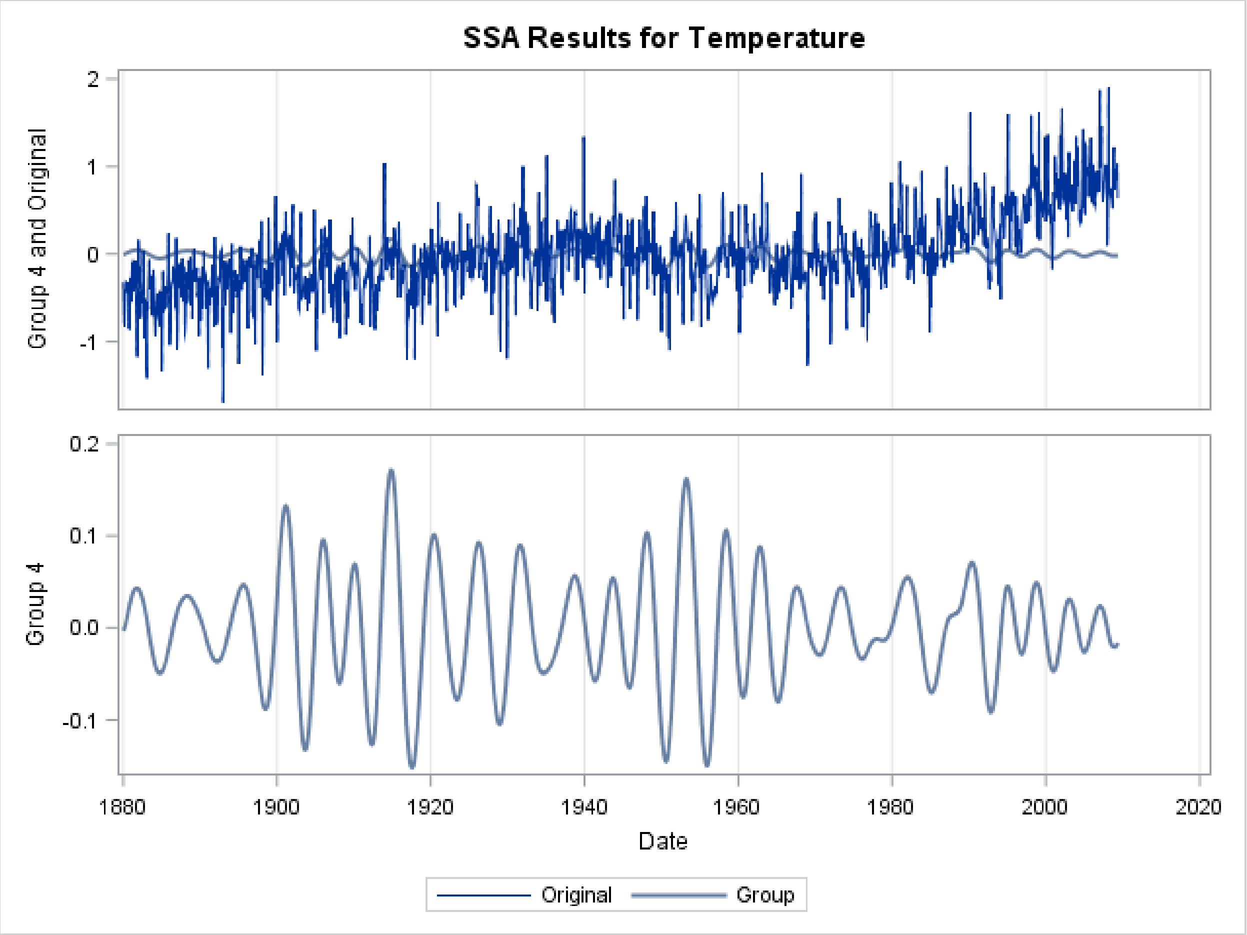

Figure 9. Fourth Spectral Group of the Temperature Anomaly

Figure 10 illustrates the spectral density plot for the fourth group. From this plot, there appears to be an approximately five-year cycle (SEASONALITY=60), possibly related to the El Niño and La Niña cycle.

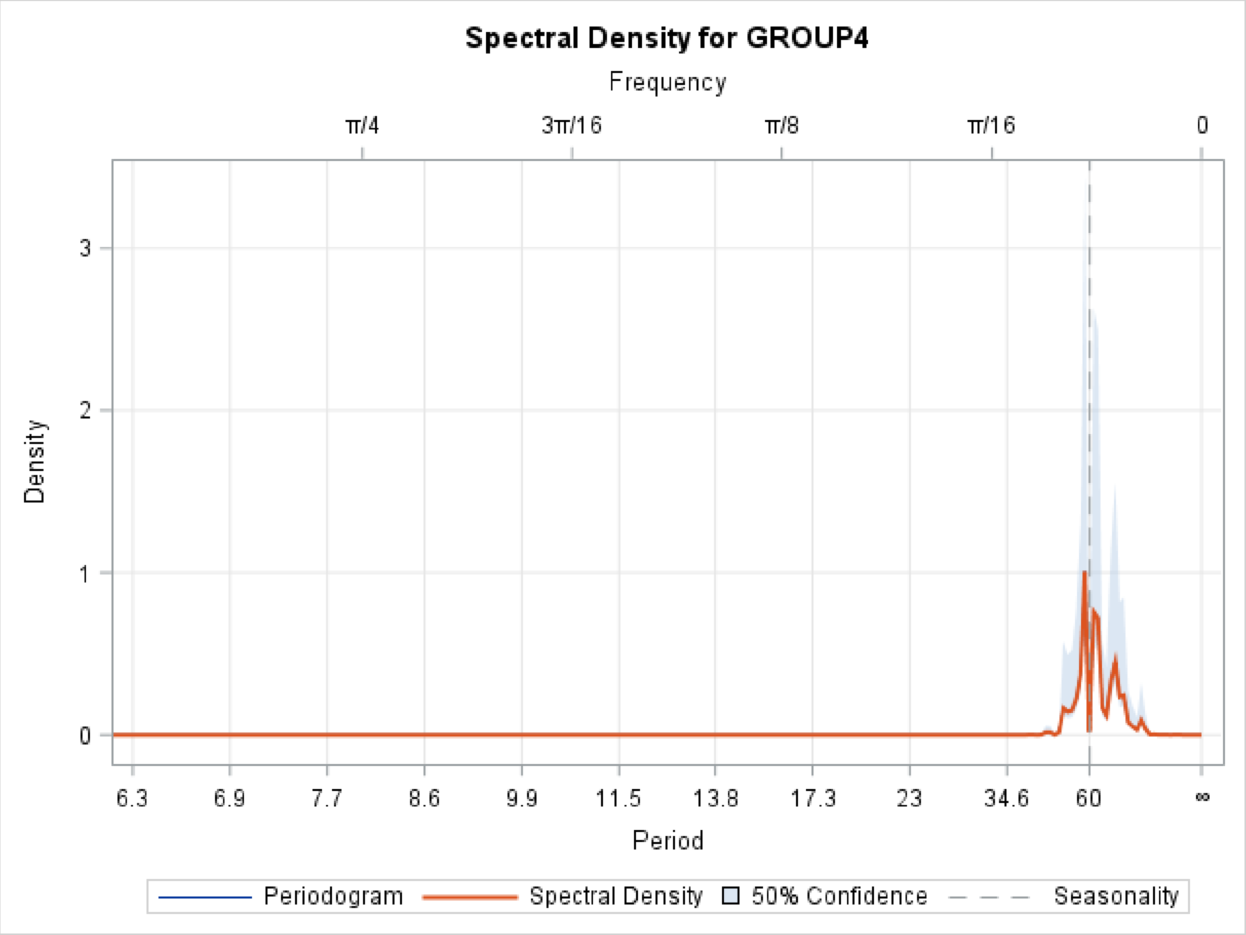

Figure 10. Spectral Density of the Fourth Spectral Group

As you can see from the preceding analysis, this long series is effectively decomposed into spectral groups. Figure 11 illustrates all four spectral groupings. No model assumptions are made other than the window length (LENGTH= option) and spectral groupings (GROUP= option). This analysis demonstrates the value of singular spectrum analysis in finding patterns (especially cyclical patterns) in long series.

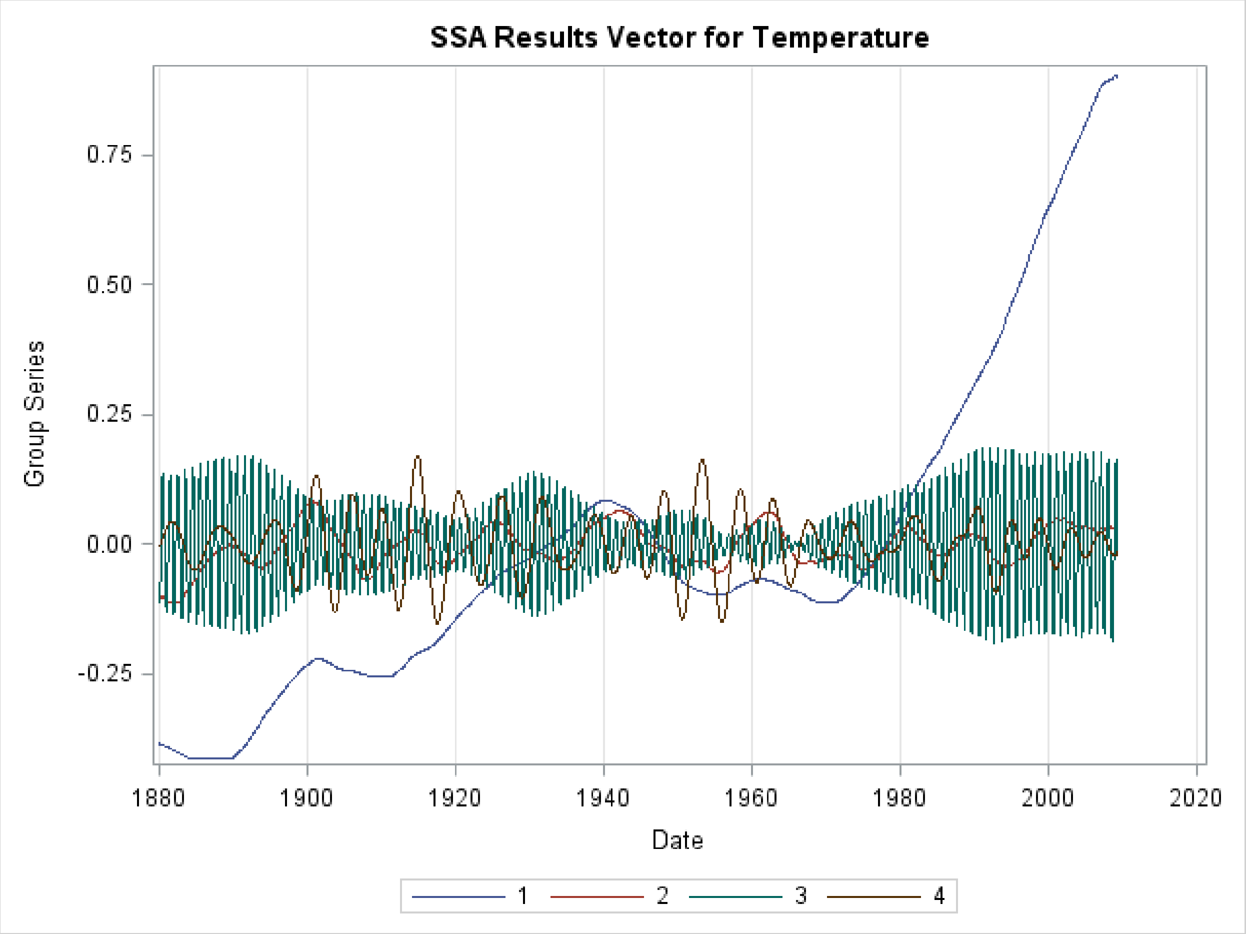

Figure 11. SSA Results for the Temperature Anomaly

Figure 12 illustrates the summation of all four spectral groupings in addition to the original series. This plot demonstrates the spectral components of the dominant singular values.

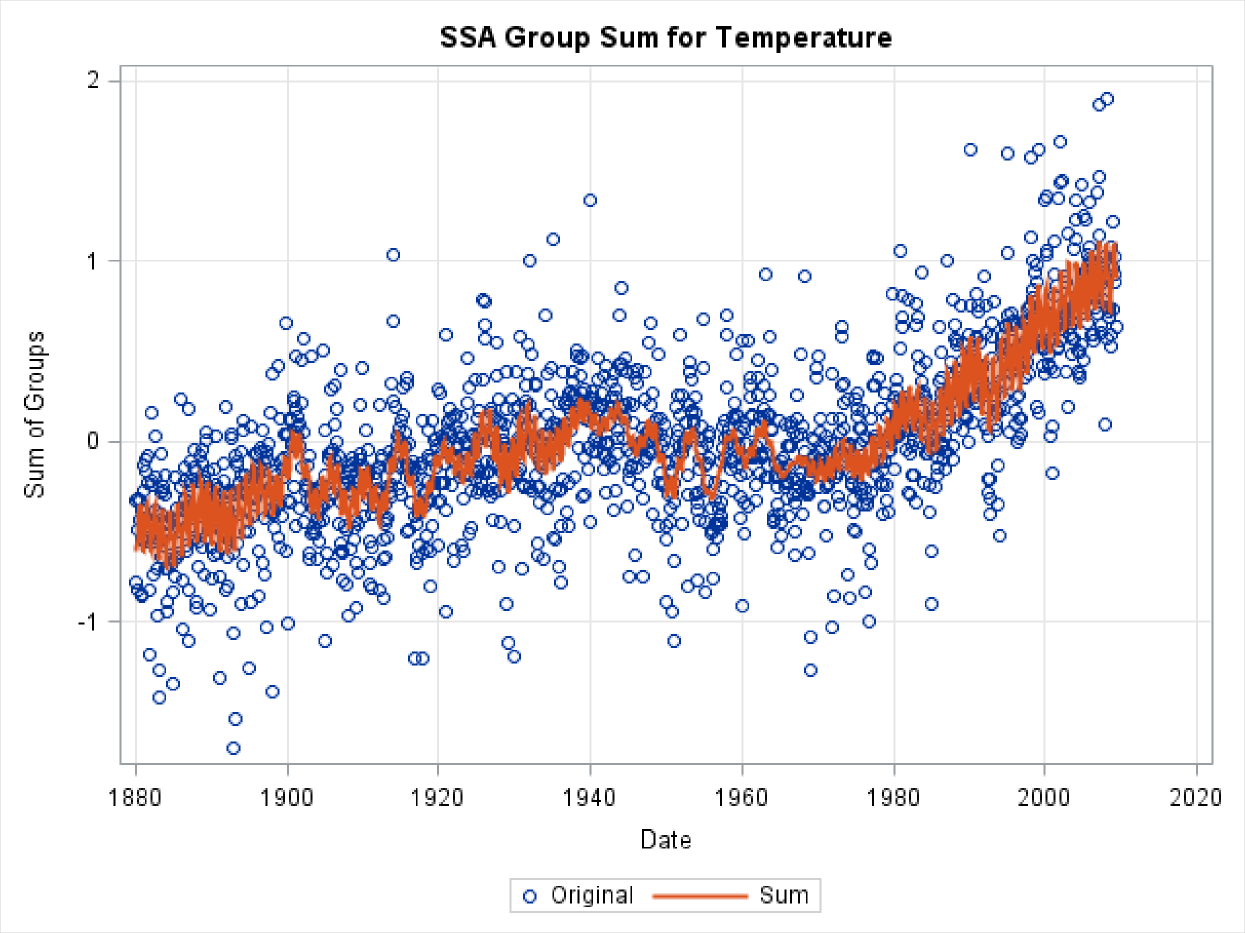

Figure 12. SSA Summation of Groupings for the Temperature Anomaly

The preceding analysis decomposed the time series into additive components. You can decompose the time series into multiplicative components by taking the log transform of the (positive-valued) time series.

AUTOMATIC SPECTRAL GROUPING EXAMPLE

To automatically obtain the results illustrated in the manual spectral grouping example, the GROUPS=AUTO(*number*) option in the TIMESERIES procedure is used to analyze time series as follows:

```
proc timeseries data=noaa out=_NULL_ plot=(series cycles ssa) OUTSSA=SSA;
   ssa / length=120 GROUPS=AUTO(4) THRESHOLD=50;
   id date interval=month;
   var temperature;
run;
```

The generated results, which are shown in Figure 12, are identical to those in the manual spectral grouping example. An additional heat map is generated to illustrate the *w*-correlation analysis. The heat map is based on the correlation between the window indices. Because the LENGTH=120 option was specified, the analysis is confined to 120 window indices. The reddish boxes indicate spectral window indices that have high correlation. Because the GROUP=AUTO(4) option was specified, the first four spectral groupings are used in the remaining analysis.

You can specify the PRINT=SSA option to print the automatic spectral groupings and the PLOT=SSA options to display the *w*-correlations matrix as illustrated in Figure 13.

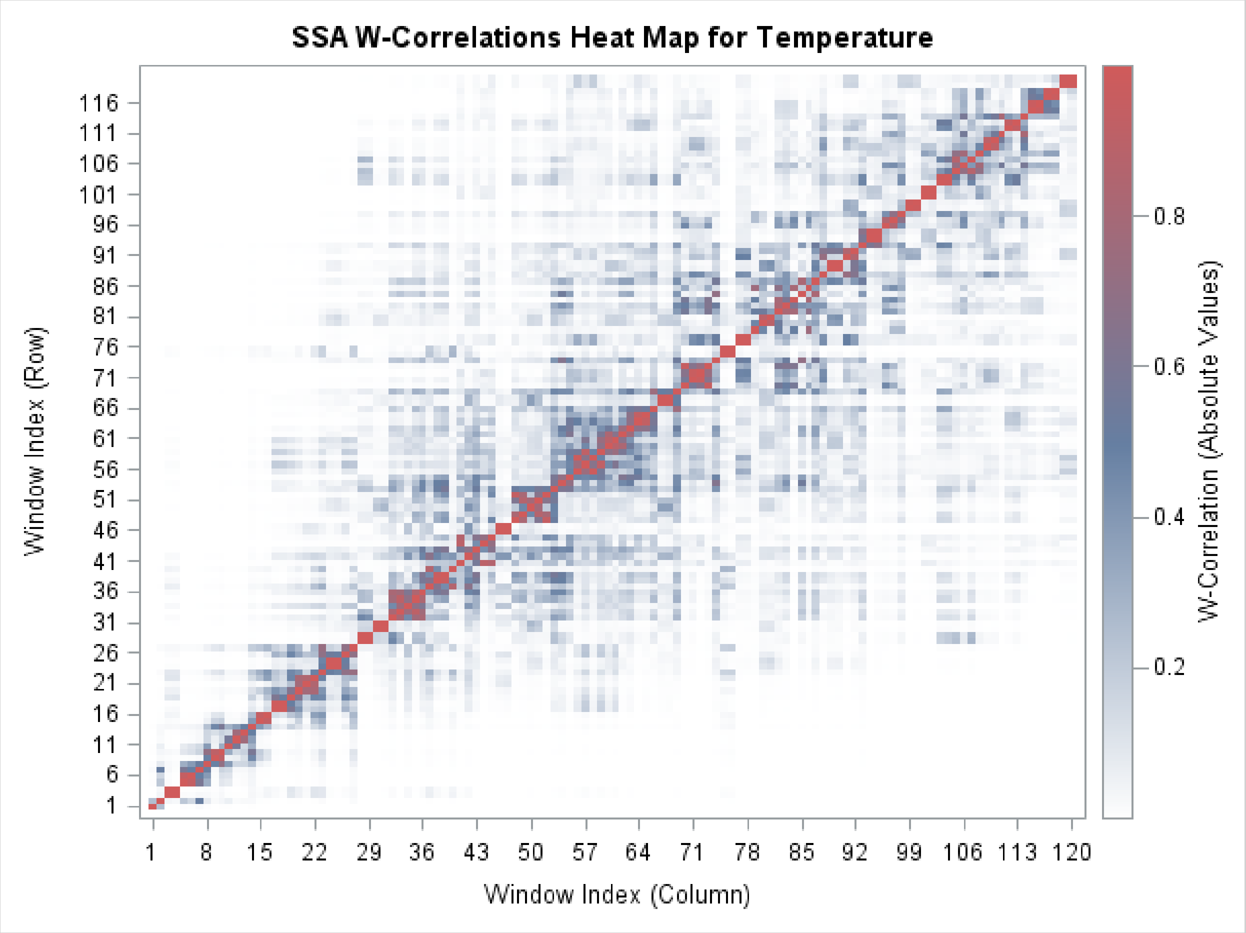

Figure 13. SSA *w*-Correlations Analysis

CONCLUSION

Singular spectrum analysis (SSA) is a very powerful tool for detecting patterns in long time series with few model assumptions. SSA effectively decomposes time series into spectral groupings. These spectral groupings can be individually analyzed using time series analysis techniques such as forecasting and state-space component analysis. This paper uses temperature records to illustrate how SAS/ETS software can be used to perform SSA.

Other cyclical time series can use this technique—for example, in load forecasting (electric, gas, and water consumption), service centers (manpower, call centers, and customer support), and telecommunications (phone service, data centers, and web servers). Geographic analysis shows how SSA can be used to determine localized trends for resource allocation—for example, in new utility construction, new service locations, new telecommunication infrastructure, and others.

REFERENCES

Elsner, J. B., and Tsonis, A. A. 1996. *Singular Spectral Analysis*. *A New Tool in Time Series Analysis*. New York: Plenum Press.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. 2001. *Analysis of Time Series Structure: SSA and Related Techniques*. Boca Raton, FL: Chapman and Hall/CRC.

Hassani, H. 2007. “Singular Spectrum Analysis: Methodology and Comparison.” *Journal of Data Science* 5: 239–257.

Leonard, M. J., Elsheimer, D. B., and Kessler, M. 2010. “Introduction to Singular Spectrum Analysis with SAS/ETS Software.” *Proceedings of the SAS Global Forum 2010 Conference*. Cary NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Michael Leonard
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-6967
Michael.Leonard@sas.com

Bruce Elsheimer
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-5959
Bruce.Elsheimer@sas.com