Using Single Spectrum Analysis for time series data analysis

Introduction:  
 A common trend in time-dependent data series is a general upward/downward trend paired with seasonal effects. This is commonly seen in data sets such as car accidents (Scott, 1986), energy usage (Chou & Tran, 2018), business revenues (Golinelli & Parigi, 2007), and many others. Because of the combination of two trends, as well as the inevitable noise that comes with any real-world date set, it is not possible to determine the contribution of the separate effects. Single Spectrum Analysis (SSA) is a novel technique that breaks down the original data series into components representing the linear trend, periodic/seasonal fluctuations and noise, allowing them to be analyzed independently of each other. This is useful for determining the magnitude and frequency of seasonal trends, or upward/downward trends of a data set independent of seasonal effects, where any of the above may not be apparent from looking at the raw data.

The SSA method used in this report consists of five major operations. The raw data is first “embedded” into a matrix whose dimensions are determined experimentally by the user. The second step performs the Singular Value Decomposition (SVD) function to the newly created matrix, breaking it down into smaller matrices sorted by contribution. The third step groups the matrices together by their associated “Singular Values” as found during the SVD and are then diagonally averaged to revert the matrices back into a list of values. Finally, the values are graphed to give a visual representation of the trends each group represents, and the autocorrelation coefficients are calculated to determine frequency of any periodic effects.

# Methodology:

## Embedding:

Embedding is the process of taking a list of time dependent values (Y) and transforming it into a matrix of “L-Lagged Vectors”. Each column consists of Y values , where L is an experimentally determined value representing the “lag”, resulting in a matrix where . The resulting matrix is what is known as a Hankel Matrix, meaning that all values of Xi,j for (example: X2,2=X3,1=X1,3).

## Singular Value Decomposition (SVD):

Singular Value Decomposition allows for any matrix X to be represented in the form . We define as the eigenvalues associated with in descending order, and U as a orthonormal matrix with rows being the orthonormal eigenvectors of . Orthonormal matrices have two unique properties useful in the SVD process, being

1. The magnitude of any given row or column is 1
2. The result of

If we define V as , we can demonstrate that is true by performing the operations

Note that because of how V is defined, V will also be an orthonormal matrix, except with dimensions . It can be useful to think of SVD in geometric terms, where U and V act as “rotational vectors” and λ as a scalar transformation, as shown in **Figure 1.**

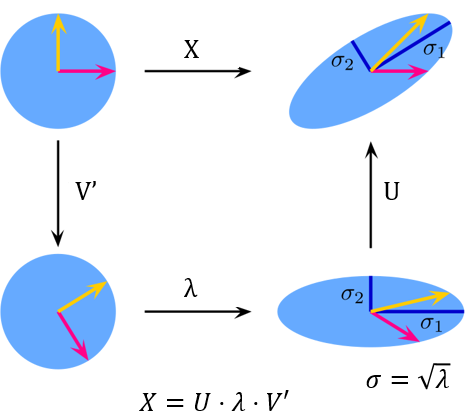


Figure : Graphical Representation of Singular Value Decomposition. Original image by Georg Johann (2010)

## Grouping:

As previously stated, λ is the only component of the equation that affects the magnitude of . By leveraging this fact, we can use λ to group similar matrices based on their contribution to the original matrix. Multiple methods exist to determine which λ values to group together (Hassani, 2007). For this project, we used a graphical approach based on the natural log of each λ value making up 99.5% of the total contribution ().

Diagonal Averaging:  
 Now that we have the grouped matrices, we can reverse the embedding process to return a list of values. The embedded matrix contained the same values on the anti-diagonals (also referred to as the secondary diagonals), however this is no longer the case as the regrouped matrices will have small variances in the values. To extract the grouped values, we take the average value of each anti-diagonal, so that

for i the range and

## Analysis:

We can now analyze the diagonally averaged Y-series to identify trends in the original data set. To first confirm that our new Y-series are representative of the original data set, we first find the individual error terms as well as the mean error. While the mean error is the main indicator that the new Y-series is representative, the error of each term is important to ensure that the new series is not consistently under/over-shooting the original. We expect the error to change between positive and negative randomly, so a consistent sign trend may indicate there is important information that has been cast aside as “error” in the grouping process.

Now that we have confirmed the accuracy of our new series, we can identify trends using the autocorrelation function

This function will return a value between , where values close to 1 indicate strong similarity between and , values close to -1 indicates is close to the additive inverse of , and values close to 0 indicating little if any similarity (Getis, 2009). For the linear trend, we expect values close to 1, and values oscillating between for periodic/seasonal trends. These oscillations can also be used to determine the frequency of any periodic changes without the use of time and computation intensive methods such as Discrete Fourier transformations.

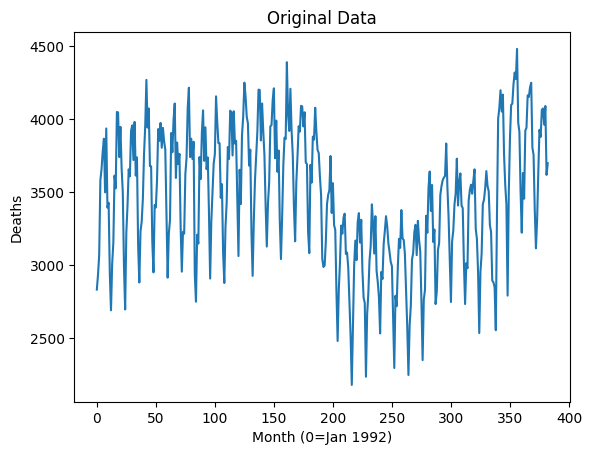


Figure : Graph of Original Data Set of Car Crash Fatalities 1992-2013

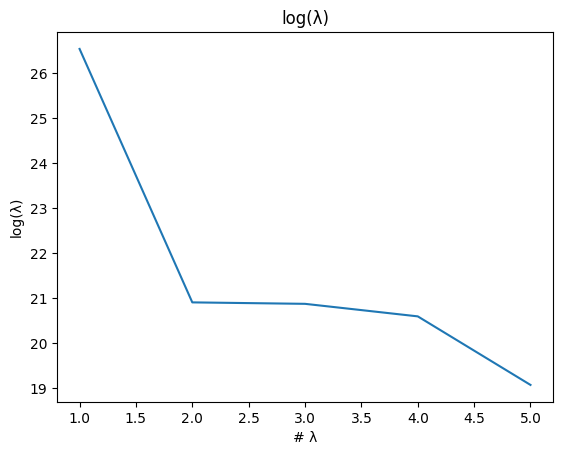
Application:

To demonstrate an application of Singular Spectrum Analysis, we developed a program to perform SSA on a well-known data set of all car crash fatalities in the United States from 1992-2013 by month **(Figure 2**). This data set was chosen because the existence of periodic trends in car fatalities has been known for years by traffic engineers (National Safety Council, 2025) and verified using older analysis techniques such as ARIMA (Dutta, 2020). Our initial hypothesis was that there would be a general downward trend combined with a strong periodic trend with a frequency of 12 (one year).

An initial L value of 192 was chosen based on results from literature showing an optimal value of L to be for computational purposes (Hassani, 2007). However, this value resulted in over-decomposition, where Y-series representing a single trend are decomposed into 2 or more series, thereby becoming far harder to analyze. Note that the opposite effect can happen, where an under-decomposed series containing two or more trends, defeating the purpose of SSA. After performing the analysis with 8 other L values (75, 100, 125, 150 ), a value of L=100 proved ideal for this data set.

With L decided, we then formed the Hankel matrix where the values of columns . The Singular Value Decomposition was performed and the contribution of each was calculated. It was determined that the first five were sufficient to represent %99.5 of the original Y-series, so the values were extracted and graphed (**Figure 3**). A clear similarity developed between and , with showing a strong distinction. This led us to declare Group 1 as the matrix , with Group 2 containing whose values follow the pattern of .

*Figure 3: Graph of the first five values of λ*

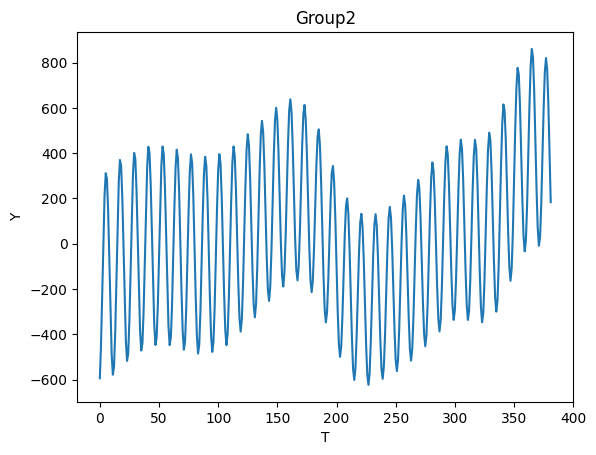


The groups were diagonally averaged and the resulting Y-series graphed, and the autocorrelation function performed for the first 30 lag values. The results confirmed our initial hypothesis that there would be a strong periodic component located in Group 2 (**Figure 4**). The lag graph (**Figure 5**) also indicates a period of 12, as lags of have a correlation coefficient (CC) close to 1, and lags have CC close to -1. Note that the first value at T=1 begins in a “valley”, meaning that January and the surrounding winter months have the least accidents in comparison with the rest of the year, while the summer months have the most.

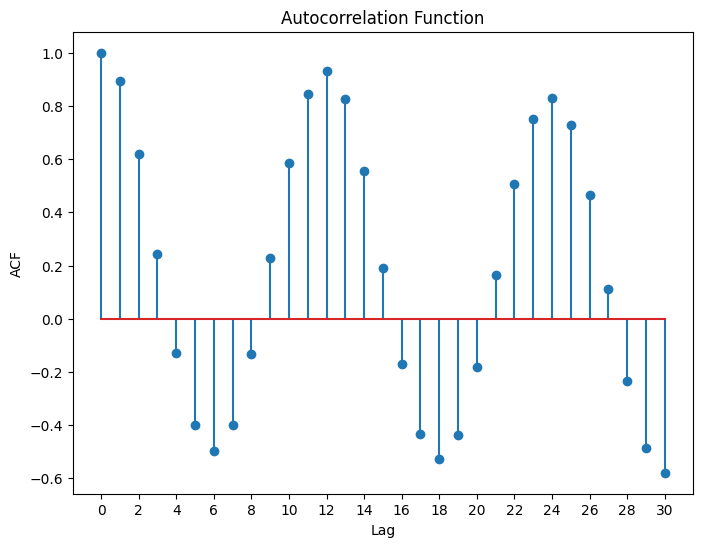
Group 1 surprised us by having a convex shape (**Figure 6**), compared to our initial hypothesis that it would steadily decrease, but the CCs (**Figure 7**) confirmed that the observed trend was accurate. This indicates there was a major shift around 2012-2013 that decreased road safety, though the specifics of what caused this are outside both the scope of this report and the expertise of its authors.

To ensure no meaningful data was lost as error, we calculated the CCs of the error terms themselves (Seabold, 2010). The result (**Figure 8**) showed small and inconsistent CC values, confirming that the error had no trend. A histogram of the error terms also confirmed that the error followed a normal distribution (**Figure 9**). The histogram is consistent with a normal distribution, albeit with a slight positive skew and a negative tail. We also calculated the mean error to be only -4.64, meaning we can confidently say that the error has no useful information inside of it.

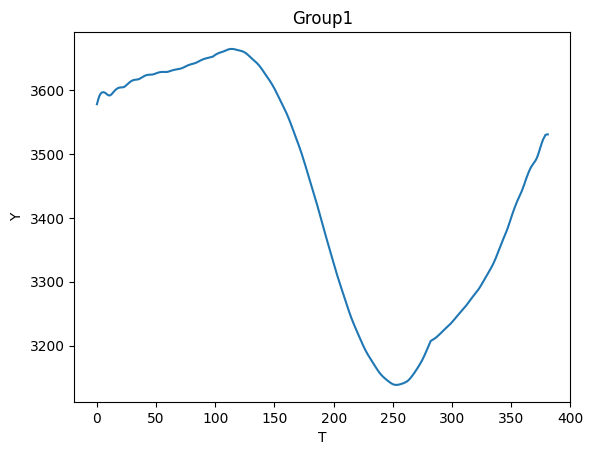
*Figure 4: Graph of the Y-Series values of Group 2*



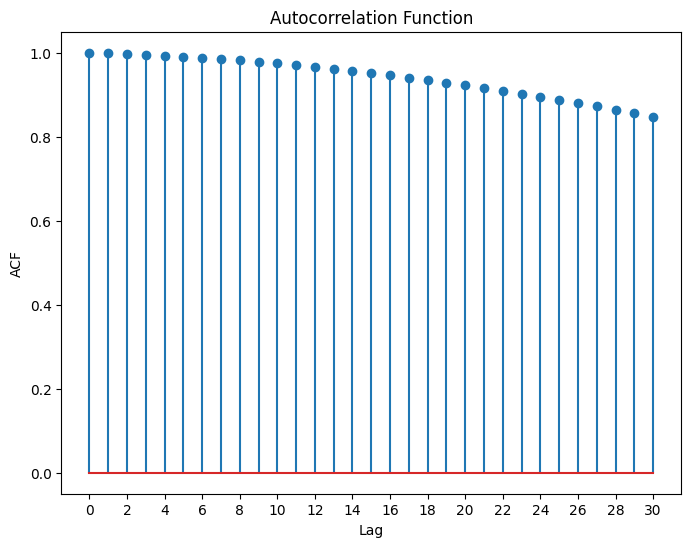
*Figure 5: Graph of the first 30 correlation coefficient of Group 2*



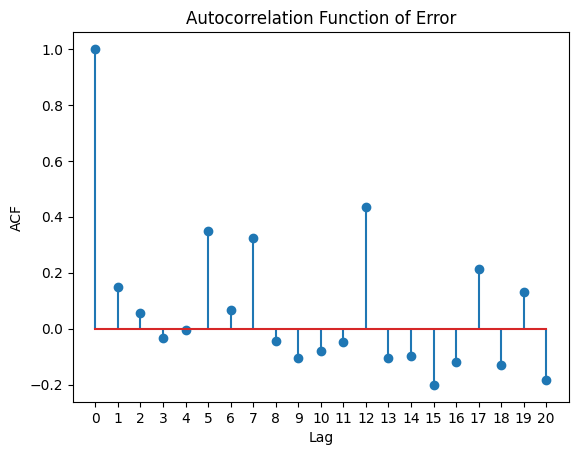
*Figure 6: Graph of the Y-Series values of Group 1*



*Figure 7: Graph of the first 30 correlation coefficient of Group 1*

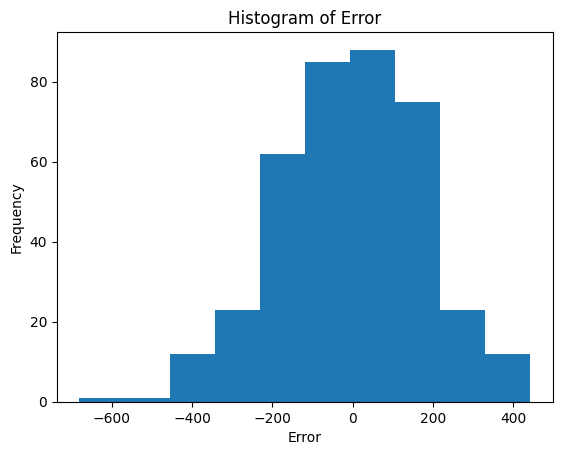


*Figure 8: Graph of the first 30 correlation coefficient of the error terms*



*Figure 8: Graph of the first 20 correlation coefficient of the error terms*

*Figure 8: Histogram of the error terms*



# Discussion:

The SSA method has some unique advantages when compared with other time series data analysis methods. The primary advantage is that the results are easily interpretable without a particularly strong mathematical background. While we may be able to analyze the existence of the trends, often the discussion of the “why” behind them is best left to those with a different expertise. For this reason, it is useful to have the data in an easily accessible form so that others can come to conclusions about the causes and effects of the trends.

SSA does have drawbacks, however, chiefly that initial value of L must be experimentally for any new data set, and for each new L value the grouping must be redone. While this does not take much time with smaller data sets (10s-100s of points) where a run though of the code takes only seconds, it becomes much more cumbersome with larger data sets (1000s-10,000s) where a run can take minutes. However, this is not a problem unique to SSA, as traditional models also require experimentally determined values.

# Conclusion:

We have demonstrated the effectiveness of the Single Spectrum Analysis technique for time-series data analysis. SSA allows for a complex data set to be broken down into its major components, making it far easier to analyze. It also has unique benefits not found in other analysis methods, such as the results being easily understood by non-mathematicians.

# References

Chou, J.-S., & Tran, D.-S. (2018). Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy*, {709-726.

Dutta, B. a. (2020). Application of Arima model for forecasting road accident deaths in India. *International Journal of Agricultural and Statistical Sciences*, 607-615.

Getis, A. (2009). Spatial autocorrelation. In A. Getis, *Handbook of applied spatial analysis: Software tools, methods and applications* (pp. 255-278). Springer.

Golinelli, R., & Parigi, G. (2007). The use of monthly indicators to forecast quarterly GDP in the short run: an application to the G7 countries. *Journal of Forecasting*, 77-94.

Harris, C. R. (2020). Array programming with NumPy. *Nature*, 357-362.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *Cardiff University and Central Bank of the Islamic Republic of Iran*.

Huhtanen, M. (2003). Orthogonal polyanalytic polynomials and normal matrices. *Mathematics of Computation*, 355-373.

National Safety Council. (2025, June 25). *Crashes by Month*. Retrieved from NSC Injury Facts: https://injuryfacts.nsc.org/motor-vehicle/overview/crashes-by-month/

Scott, P. (1986). Modelling time-series of British road accident data. *Accident Analysis & Prevention*, 109-117.

Seabold, S. a. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference.*