STUDENT SUMMER INTERNSHIP TECHNICAL REPORT

Time Series Decomposition and Reconstruction Using Single Spectrum Analysis in the Hanford 200 West Area.

DOE-FIU SCIENCE & TECHNOLOGY
WORKFORCE DEVELOPMENT PROGRAM

Date submitted:

December 16, 2022

Principal Investigators:

Rohan Shanbhag (DOE Fellow Student)
Florida International University

Mark Rockhold, Ph.D., Xuehang Song, Ph.D. (Mentor)
Pacific Northwest National Laboratory

Ravi Gudavalli Ph.D. (Program Manager)
Florida International University

Leonel Lagos Ph.D., PMP® (Program Director)
Florida International University

Submitted to:

U.S. Department of Energy
Office of Environmental Management
Under Cooperative Agreement # DE-EM0005213

FIU | Applied Research Center
FLORIDA INTERNATIONAL UNIVERSITY

EXECUTIVE SUMMARY

This research work has been supported by the DOE-FIU Science & Technology Workforce Development Initiative, an innovative program developed by the U.S. Department of Energy's Office of Environmental Management (DOE-EM) and Florida International University's Applied Research Center (FIU-ARC). During the spring of 2022, a DOE Fellow intern, Rohan Shanbhag, spent 8 weeks doing a summer internship at Pacific Northwest National Laboratory under the supervision and guidance of Dr. Xuehang Song and Dr. Mark Rockhold. The intern's project was initiated on June 29, 2022 and continued through August 20, 2022 with the objective of performing data analysis on the time series data of the 200 West Area.

Hanford Facility 200-Area (USDOE) is a 79-square-mile site that is situated 17 miles to the north-northwest of Richland, Washington. The site is one of the four original EPA National Priorities List (NPL) locations inside the Hanford site, which is under DOE management. Former chemical processing units and waste management facilities can be found at the 200 Area NPL site, which is in the Central Plateau of the Hanford Site. Massive amounts of carbon tetrachloride (CCl_4) were released into the ground during processing operations. Processing, finishing, and maintaining radioactive materials, including plutonium, were additional site activities. About one billion cubic yards of radioactive, mixed, and toxic solid and diluted liquid waste were dumped on-site in trenches, ditches, and a landfill. In order to address this problem, the 200 Area was divided into several cleanup Areas, each with multiple operable units (OUs), to address site contamination more effectively. These cleanup Areas are known as the 200 West, 200 East, and 200 North Areas. To fulfill cleanup requirements, the 200 West P&T processes the extracted water to eliminate pollutants of concern (COCs) from the influent water streams. The 200 West P&T's upkeep and improvements are concentrated on achieving and maintaining nominal design capacity to serve the requirements of each groundwater OU while also meeting the needs of remedial optimization to satisfy site regulatory and cleanup objectives. The main objective of this research is to find underlying patterns and behaviors of carbon tetrachloride (CCl_4) concentration data, mass data, and aqueous mass data through SSA (Single Spectrum Analysis) and then use the components identified as inputs for a temporal clustering model. Through the decomposition and reconstruction of the time series data, we can find the components that best represent the original time series and use those as inputs for a temporal clustering model to see how the extraction wells in the 200 West Area are clustered using their temporal components.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Pump-and-treat (P&T) is the main approach used to contain, intercept, and remove groundwater contaminate plumes at many waste sites. Groundwater withdrawal and contaminant concentration data are routinely collected for P&T operations, which can be mined to enhance site characterization activities, improve remediation performance assessment, and optimize remedy system operation. The primary objective of this research was to assist in data analysis of extraction wells carbon tetrachloride (CCl_4) concentration in the 200 West Area at Hanford by researching and finding a method to better analyze and separate the different components of a time series such as noise, periodicities, and trend. The method that was found to best effective to accomplish this is known as Single Spectrum Analysis (SSA). This method allows us to decompose a time series into separate components that can be used to represent the periodicities, noise, and trend. Through this method, once the components are identified for each well, those components can then be used for temporal clustering which can then be used to visualize the clusters that the extraction wells belong to on a grid.

2. RESEARCH DESCRIPTION

Before using any machine learning algorithms or doing any kind of analysis, understanding the data that will be used is extremely important. For the 200 Areas West, there were 2 files provided that contained different sections of information regarding the wells of the Area. The first file contained the names, ids, x and y coordinates, and elevation data of all the injection and extraction wells in the 200 West Area. The second file contained the ids, sampling dates, carbon tetrachloride (CCl_4) concentration (ug/L), CCl_4 mass (kg), and aqueous mass (GPM) data of all 30 extraction wells. Out of the 30 extraction wells in the dataset, 19 were chosen to be used in the analysis since several of the wells had missing values or had inconsistencies in the data in comparison to most wells. Furthermore, each chosen extraction well had 102 records of CCl_4 concentration (ug/L), CCl_4 mass (kg), and aqueous mass (GPM) data ranging from the years 2012 – 2021 sampled monthly. After plotting the original time series for the wells and observing how the series looked for the CCl_4 concentration, CCl_4 mass, and aqueous mass, it was decided that each time series would be treated separately, and the decomposition and reconstruction would be done for each of the 3-time series for all 19 extraction wells. This was mainly due to the observed fluctuating peaks and trends of each of the time series having different visualizations for all 3 types of data.

Once the data identification and preprocessing were complete, the next step was to incorporate this data into the Single Spectrum Analysis (SSA) technique. The main objective of using this technique was to decompose the time series into its elementary components to better visualize the trend, periodic, and noise components. Furthermore, using the right amount of elementary time series components, it can be reconstructed to be compared to the original to see the similarities between them. Through this visualization, the reconstructed time series, original time series, and noise can easily be seen as well for comparison purposes. Once that was accomplished, the elementary components for each extraction well would be saved and then used for a temporal clustering model. For the temporal clustering model, the tslearn package of Python was used since it consists of useful time series clustering algorithms. Within that package, the time series version the K-Means algorithm was chosen to be the model for the temporal clustering. The primary reason this model was chosen was due to the Dynamic Time Warping (DTW) metric being present for this algorithm. The DTW metric is great at collecting time series of similar shapes. In turn, the cluster centroids for each cluster identified in the algorithm are then computed with respect to the DTW. This is essential since each time series constructed from the elementary components for each of the 19 extraction wells would have different shapes and patterns that should be included and not discarded when fitting the clustering model. Lastly, the number of clusters to separate the wells into was determined through a K-Means elbow plot that showed the optimal number of clusters to choose for the k hyperparameter while minimizing the SSE error. SSE stands for the sum of squared error, and it represents the sum of the distances of the samples or records to their closest cluster centroid. Figure 1 below shows a flowchart regarding the main steps involved in the SSA technique. In addition, Figure 2 displays the reconstruction stage of the SSA technique for a time series of a certain extraction well.

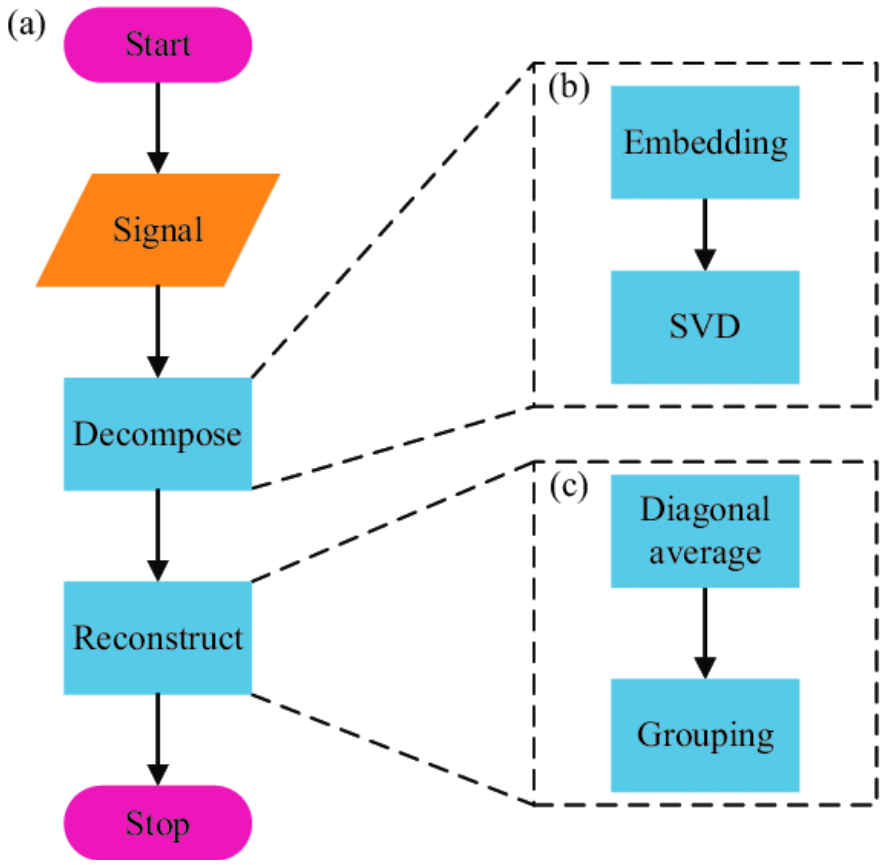

Figure 1. Main steps of the SSA algorithm. (a) is the main procedure of the SSA algorithm, (b) and (c) are sub-procedures involved in the decomposition and reconstruction stages of SSA.

2.1 SSA Algorithm Main Steps

1. Let a time series X of length N exist such that X = (x_1,...,x_N).
2. Determine a window length L such that 2 <= L <= N/2.
3. For the decomposition stage, in the embedding phase, a trajectory matrix is constructed by means of the embedding operator T, which maps a time series to an L×K Hankel matrix (K=N−L+ 1).
4. For the decomposition stage, in the singular value decomposition Phase (SVD) phase, it is given by the formula X = $\sum_{m=1}^{d} \sqrt{\lambda_m}\, U_m V_m^T$ where $\{U_m\}_{m=1}^{d}$ and $\{V_m\}_{m=1}^{d}$ are systems of the left and right singular vectors of X, respectively.
5. For the reconstruction stage, the way of grouping the elements is {1,…,d} = $\bigcup_{j=1}^{c} I_j$
6. Within the grouping phase, the SVD components are grouped using the following formula. X = $X_{I_1} + \cdots + X_{I_c}$, where X = $\sum_{m \in I} \sqrt{\lambda_m}\, U_m V_m^T$.
7. Within the diagonal averaging phase, each matrix X_I is converted to the nearest Hankel matrix $\tilde{X}_I$ by the hankelization process and then $\tilde{X}_I$ is transformed into a time series such that $\tilde{X}^{(I)} = T^{-1}\left(\tilde{X}_I\right)$.
8. The output of the SSA algorithm is the decomposition X = $\tilde{X}^{(I_1)} + \cdots + \tilde{X}^{(I_c)}$, which matches the original time-series X if a correct and proper grouping of the elementary components was performed.

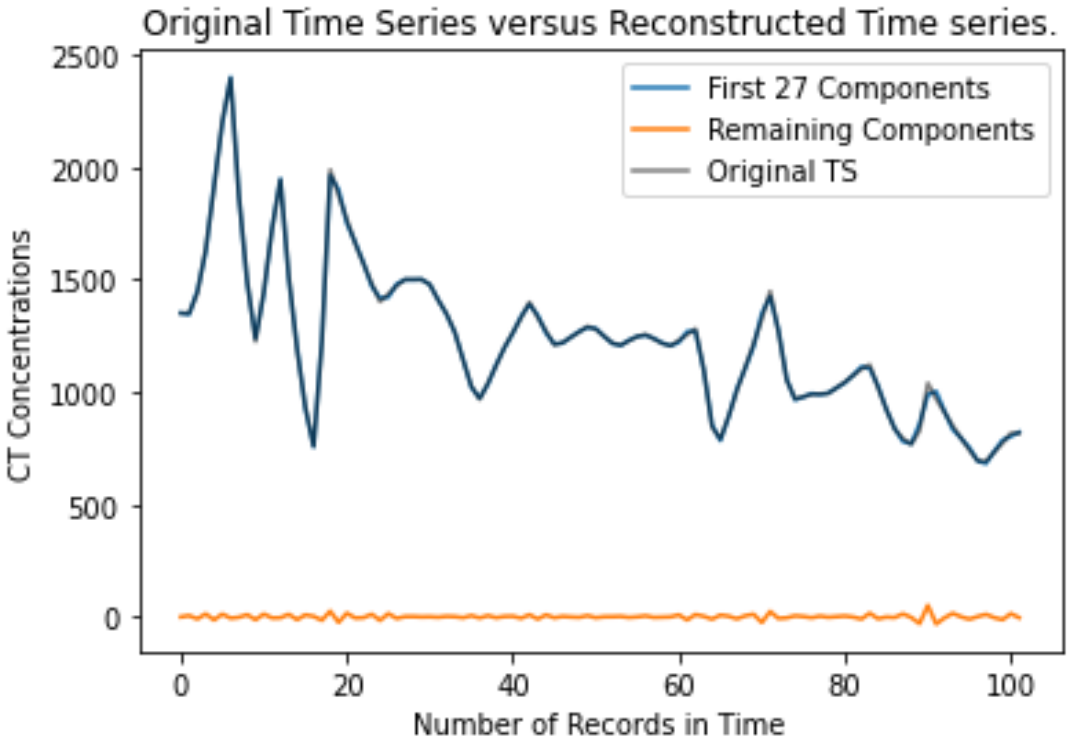

Figure 2. This visualization shows the reconstructed vs original time series for the extraction well 299-W11-90 for CCl4 concentration time series data. The x-axis are the 102 records or samples ranging from 2012 to 2021 and the y-axis are the concentration levels in ug/L.

Through this visualization, we can see that the first 27 out of the first 40 components were the components that were needed to reconstruct his time series to best resemble the original. The orange line consists of the remaining 13 components that make up the noise of the time-series. By looking at the noise components, there are some underlying patterns and behaviors that can be seen such as spikes at certain intervals and stagnant stretches at certain intervals as well.

3. RESULTS AND ANALYSIS

Once the elementary time-series elements were identified for each of the 19 extraction wells, the components were saved to a joblib file to be used as input for the temporal clustering. Two approaches were then determined for the clustering. The first approach was to use the component F_0 from the SSA that represented the trend for all 19 extraction wells as input for the clustering model. The second approach was to sum all time-series components from the SSA that represented the periodicities of the original data for each extraction well used and use that as input for the clustering model. This process was then repeated for the CCl_4 mass data and the aqueous mass data as well. The results of the modeling were the 19 extraction wells identified in the clusters they belonged to based on the temporal clustering for the CCl_4 concentration data, CCl_4 mass data, and CCl_4 aqueous data. Furthermore, the results for the temporal clustering of the concentration data were then overlayed onto a 200 West Area concentration plume map to see where the extraction well clusters resided in comparison to the concentration plume levels of the area. Two metrics were used to evaluate the clustering performance of the time-series K Means clustering. One was the silhouette score and the other was the Davies–Bouldin index (DBI). A silhouette score of 1 means that the clusters are very dense and nicely separated, a score of 0 means that clusters are overlapping, and a score of less than 0 means that data belonging to clusters may be incorrect. The Davies–Bouldin index (DBI) is, a metric for evaluating clustering algorithms, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset. Usually, the lower the DB index value, the better the clustering is. Table 1 shows the clustering results for all three CCL_4 datasets when clustered based on the trend component or the periodic components. Figure 3 displays how the periodic time series components for the CCl_4 concentration data were clustered. Lastly, Figure 4 and Figure 5 are visualizations that show the clusters determined by modeling with periodic components as input and trend components as input, respectively.

Table 1. Results of temporal clustering between using the trend component (F0) only as input versus using the periodic components of the time series as input. The results show that overall, across all 3 data sets for CCl4, clustering based on the trend components performed better than clustering based on the periodic components.

Temporal Clustering Results using Time Series K Means with DTW metric.	*CTET Concentration (ug/L) Time Series Data*	*CTET Mass (kg) Time Series Data*	*Aqueous Mass (GPM) Time Series Data*
Trend component only (F_0)	**Silhouette Score: 0.815 DBI index: 0.513**	**Silhouette Score: 0.79 DBI index: 0.7**	**Silhouette Score: 0.73 DBI index: 0.351**
Periodic Components only (F_1 till F_n summed together)	**Silhouette Score: 0.233 DBI index: 1.699**	**Silhouette Score: 0.058 DBI index: 1.974**	**Silhouette Score: 0.167 DBI index: 1.387**

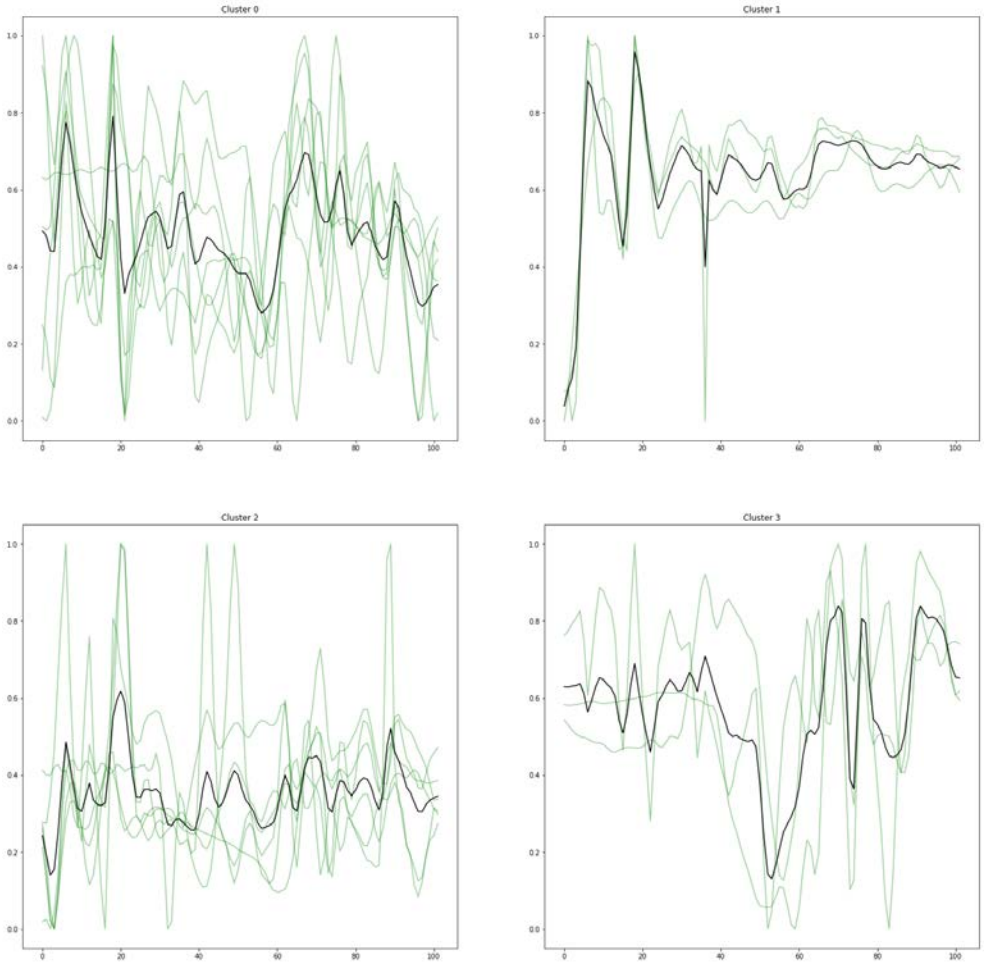

Figure 3. Four clusters shown on how the periodic time series components for the CCl4 concentration data were clustered using the time-series version of the K Means model. The green time-series are the other periodic time-series and the one in black is the average time-series that the clustering was based upon.

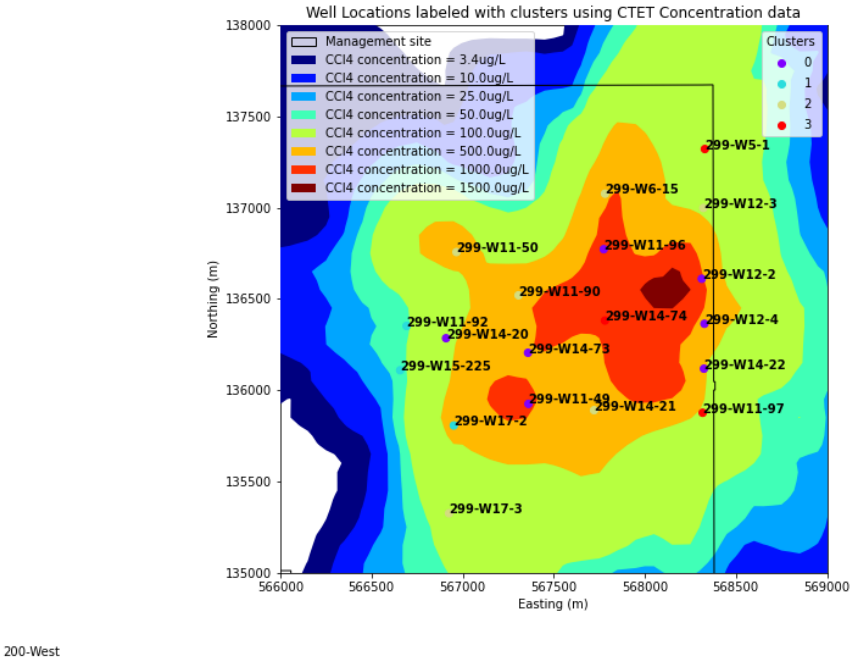

Figure 4. Clusters determined by modeling with periodic components as input, plotted with CCl4 concentration plumes. This is representative of the Ringold E of the 200 West Area. The black line that has the edges of a square represent the 200 West Area well boundary.

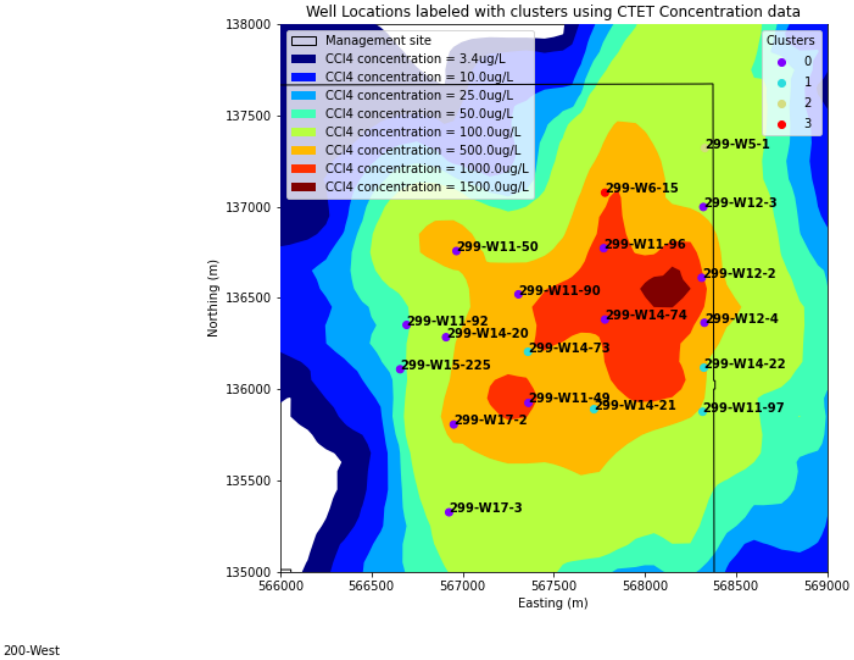

Figure 5. Clusters determined by modeling with trend components as input, plotted with CCl4 concentration plumes. This is representative of the Ringold E of the 200 West Area. The black line with the edges of a square represents the 200 West Area well boundary.

4. CONCLUSION

Overall, Single Spectrum Analysis was effective in deconstructing and reconstructing the original time series data for all 19 chosen extraction wells. Furthermore, the extracted components were able to be used for temporal clustering to compare clustering the trends versus the periodicities. The results demonstrated that clustering based on the trend components was more successful and effective in creating more separable and distinct clusters in comparison to clustering using the periodic components. One of the primary reasons for this is that the periodicities among the 19 wells extracted from the SSA are very similar and thus have a lot of overlap within the clusters.

Several ways to improve this modeling would be to try different machine learning clustering algorithms instead of just the time series K-Means version for result comparison and evaluation purposes. Some of these could include K-Shape or hierarchical clustering algorithms. Furthermore, trying other types of machine learning implementations after decomposing the elementary time series components using SSA such as a classification or regression problem could be another approach as well. Throughout this internship, I learned new skills and gained a lot of knowledge. My data preprocessing skills in Python improved through the use of joblib files and ensuring that the data had no missing values, errors, and was in the right format to be used for clustering. I also learned a lot about the site background and history of the 200 West Area. In addition, I learned how to understand and visualize the different components of a time series through SSA such as the trend, periodicity, and noise. Lastly, understanding how to use a W-correlation matrix in relation to determining which time series components contained noise was essential and also another valuable skill that I learned.

5. REFERENCES

1. Cao, H., Song, Y., Li, Y., Li, R., Shi, H., Yu, J., Hu, M., & Wang, C. (2018). Reduction of moving target time-of-flight measurement uncertainty in femtosecond laser ranging by singular spectrum analysis based filtering. *Applied Sciences (Switzerland)*, *8*(9). https://doi.org/10.3390/app8091625

2. Golyandina, N. (2020). Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. In *Wiley Interdisciplinary Reviews: Computational Statistics* (Vol. 12, Issue 4). Wiley-Blackwell. https://doi.org/10.1002/wics.1487

3. Marques, C. A. F., Ferreira, J. A., Rocha, A., Castanheira, J. M., Melo-Gonçalves, P., Vaz, N., & Dias, J. M. (2006). Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth*, *31*(18), 1172–1179. https://doi.org/10.1016/j.pce.2006.02.061

APPENDIX

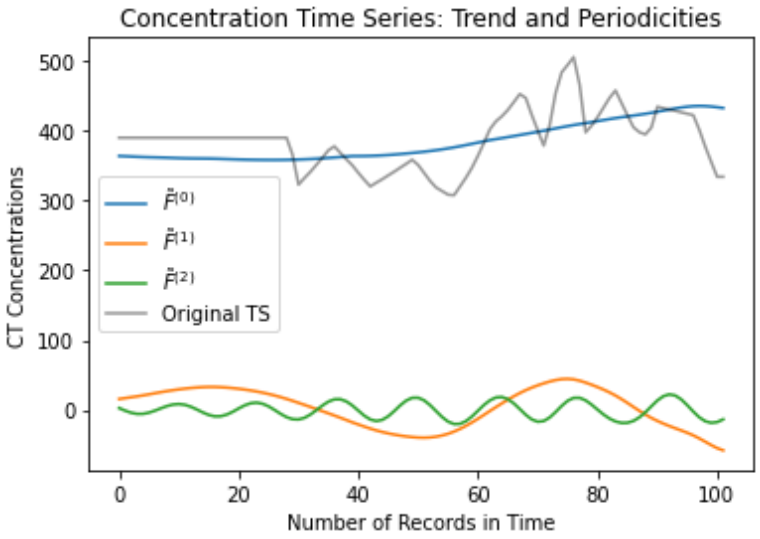

Figure A-1. This visualization shows the separation of the trend and several periodic components using SSA for extraction well 299-W14-22.

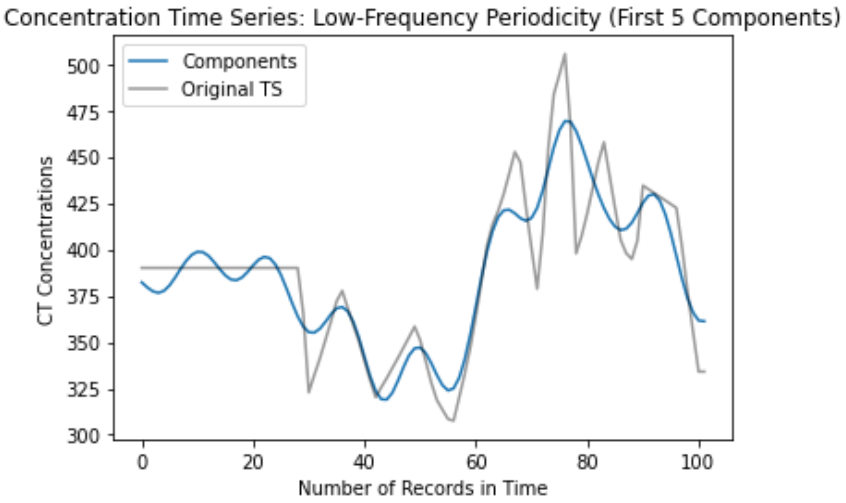

Figure A-2. This visualization shows the first five periodic components in comparison to the original time series for extraction well 299-W14-22.

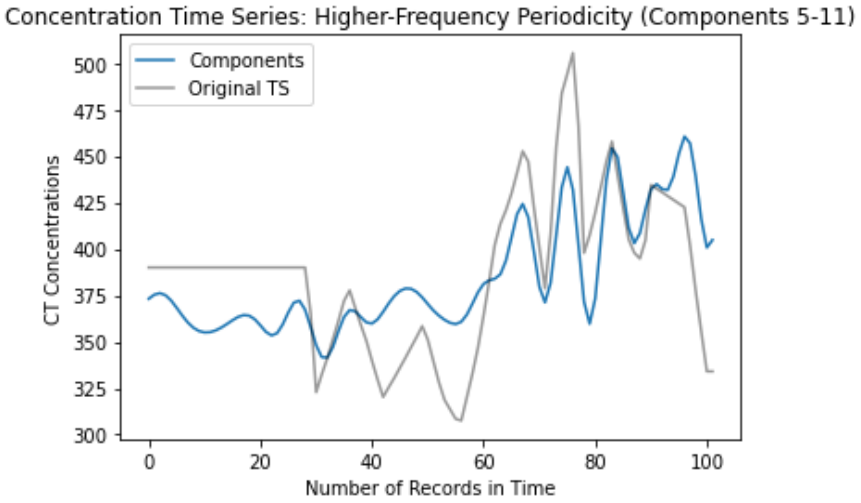

Figure A-3. This visualization shows the fifth till eleventh periodic components in comparison to the original time series for extraction well 299-W14-22.

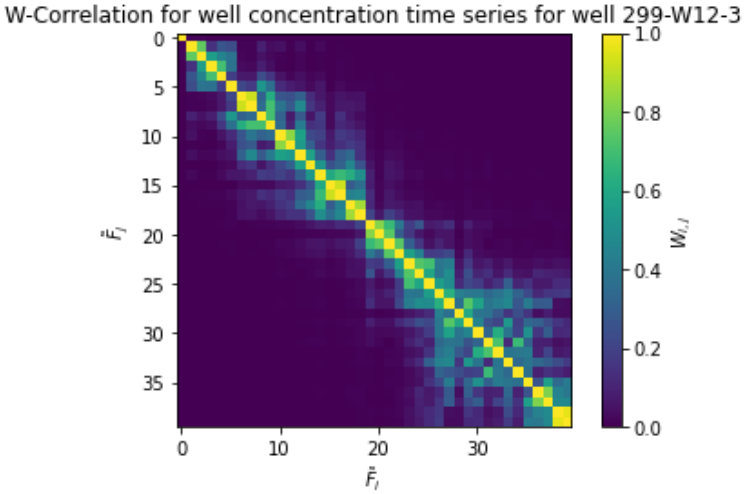

Figure A-4. This visualization shows an example of a W-correlation matrix that determines which time series component groupings to choose to differentiate between the periodicities and the noise. In this instance, the first 19 components would consist of the trend and the periodic components, whereas the remaining components would consist of the noise of the time series.

Figure A-5. These visualizations show all the identified periodic components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 concentration dataset.

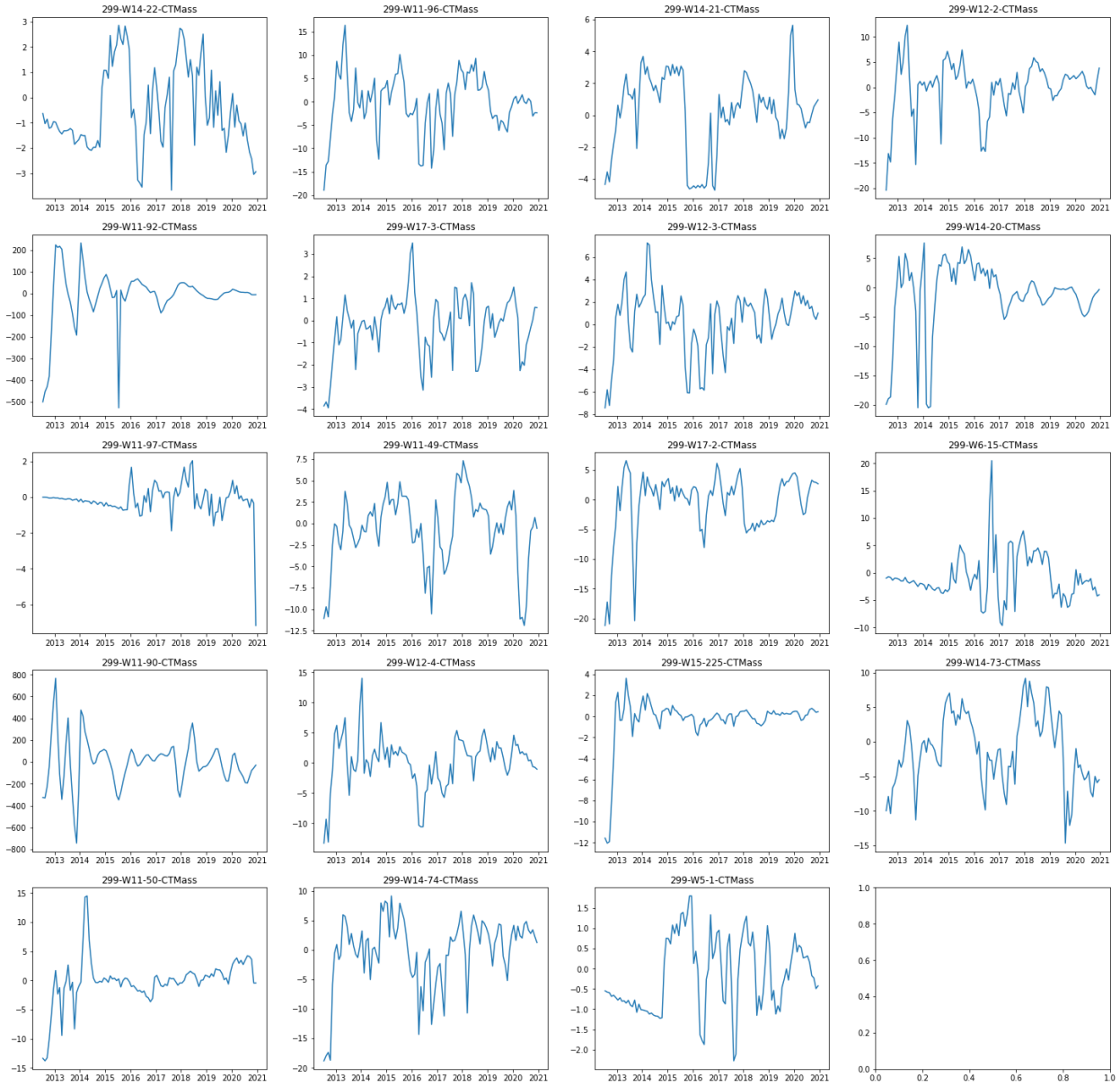

Figure A-6. These visualizations show all the identified periodic components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 mass dataset.

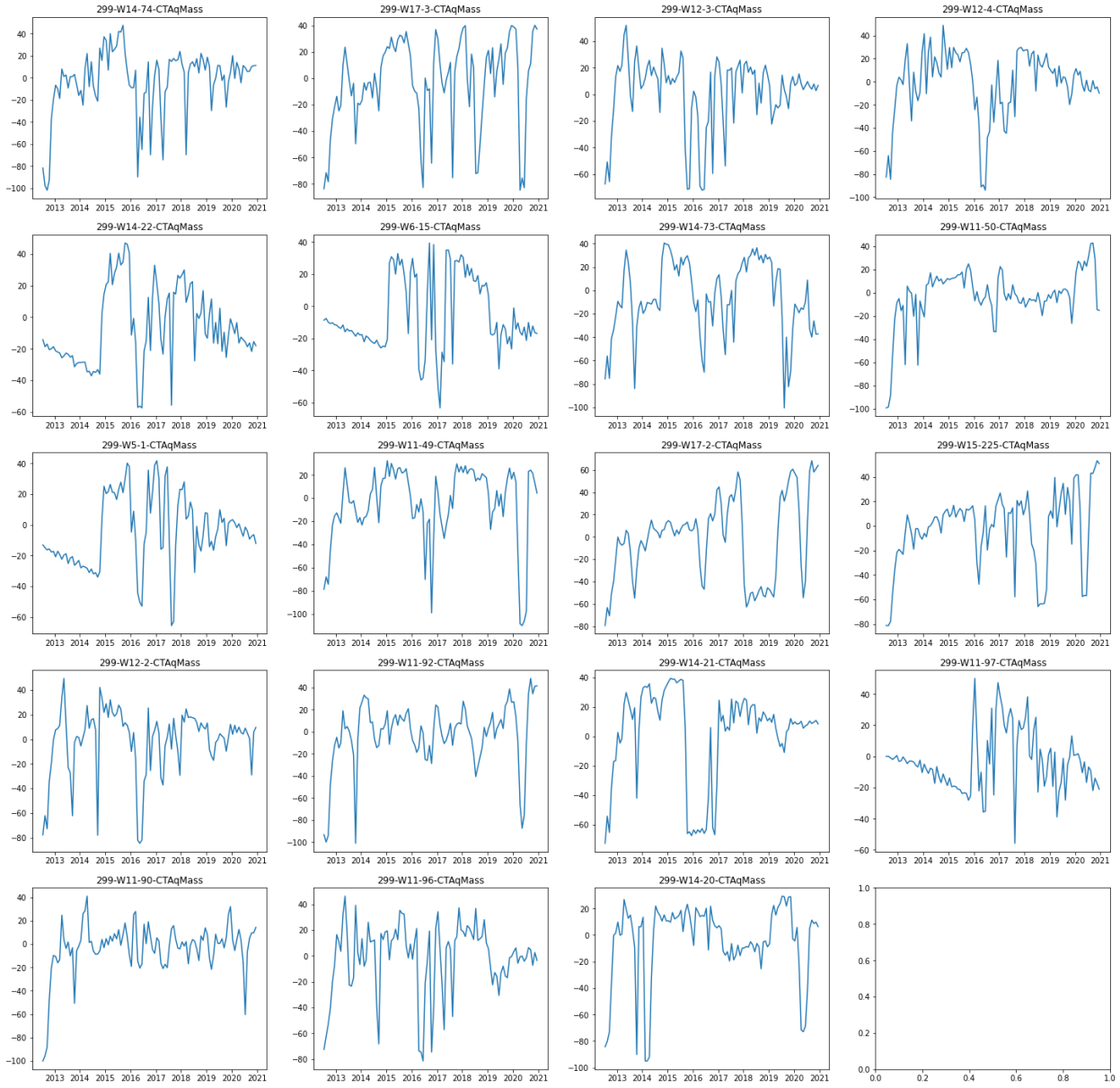

Figure A-7. These visualizations show all the identified periodic components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 aqueous mass dataset.

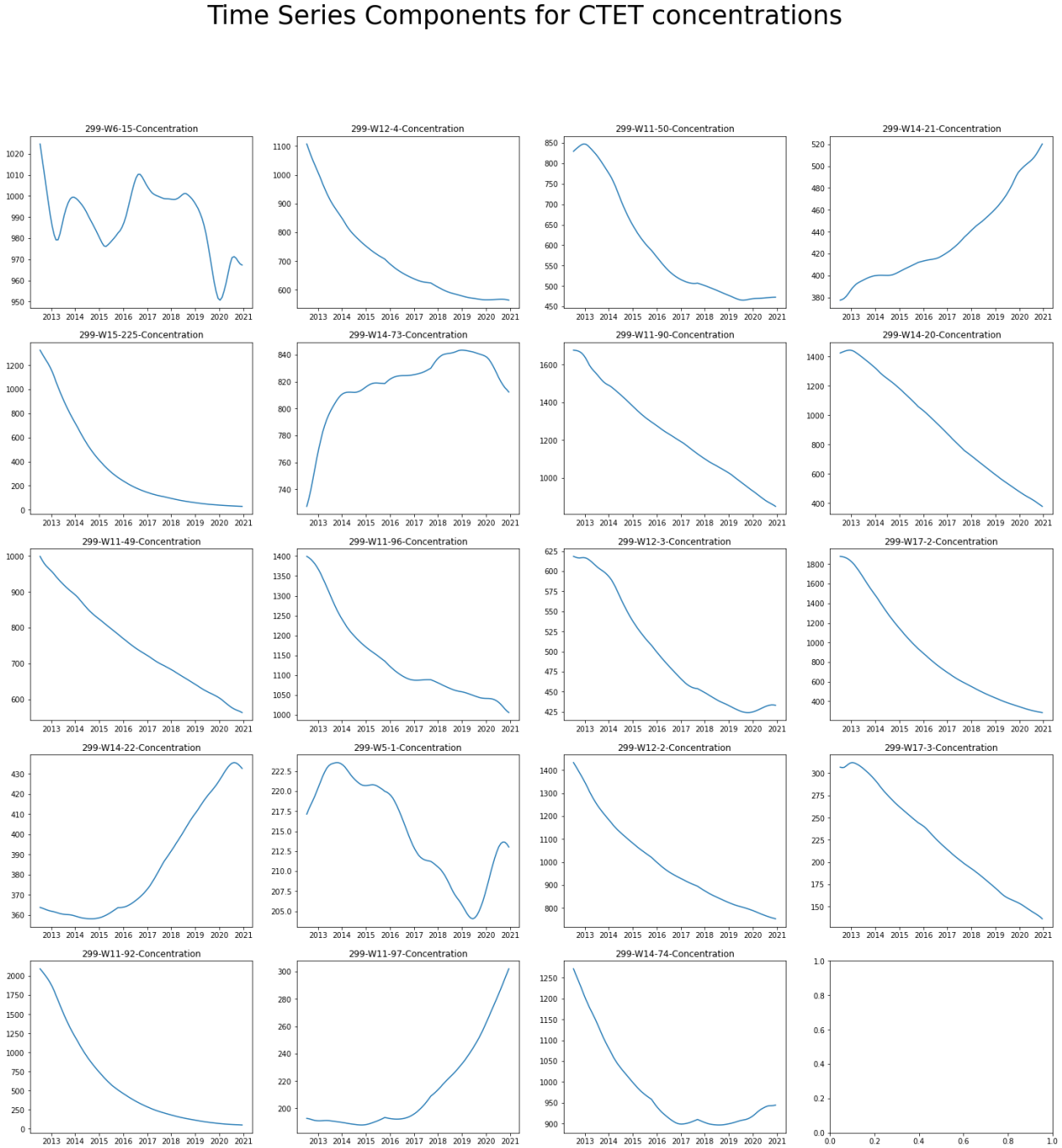

Figure A-8. These visualizations show all the identified trend components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 concentration dataset.

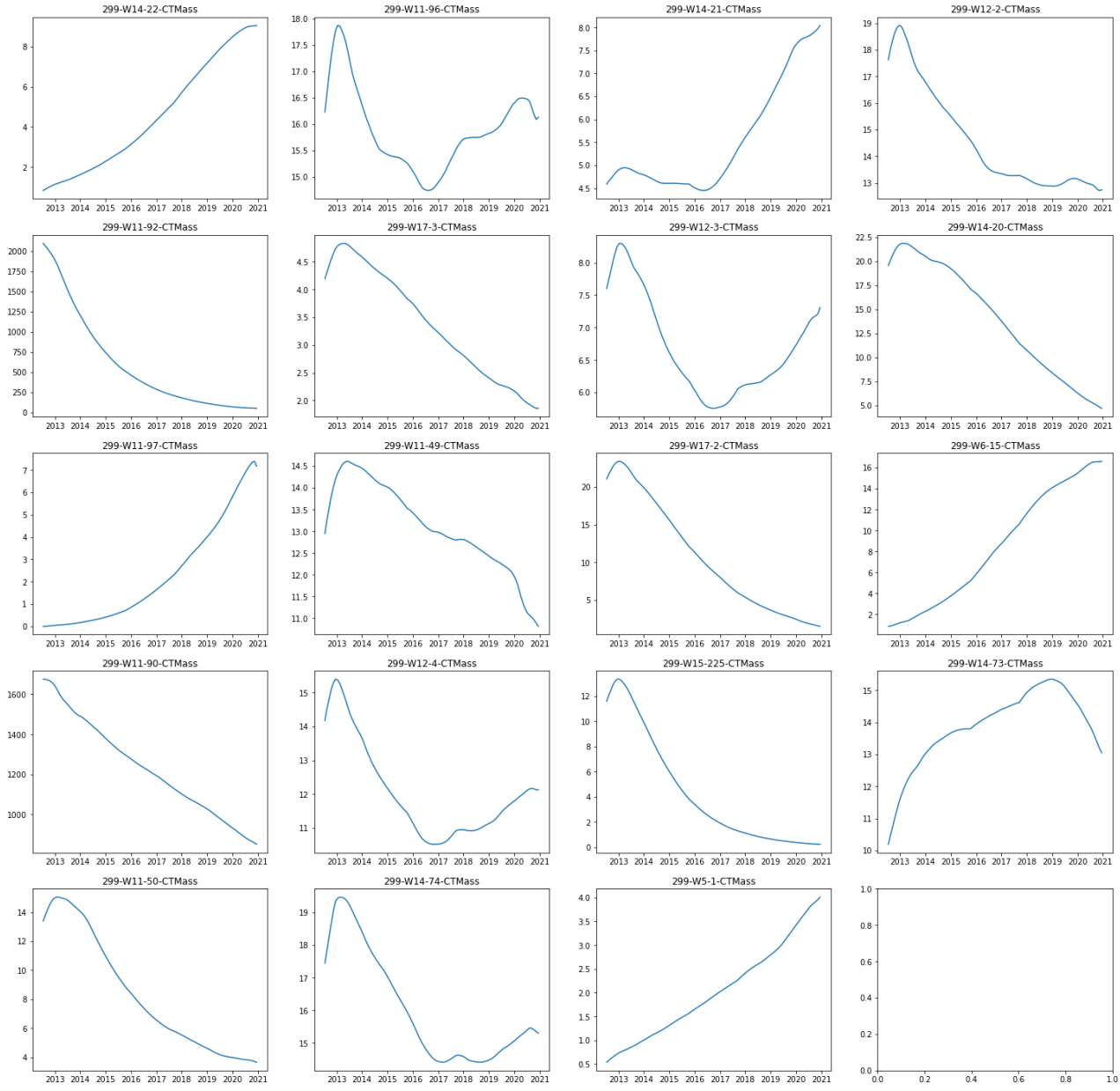

Figure A-9. These visualizations show all the identified trend components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 mass dataset.

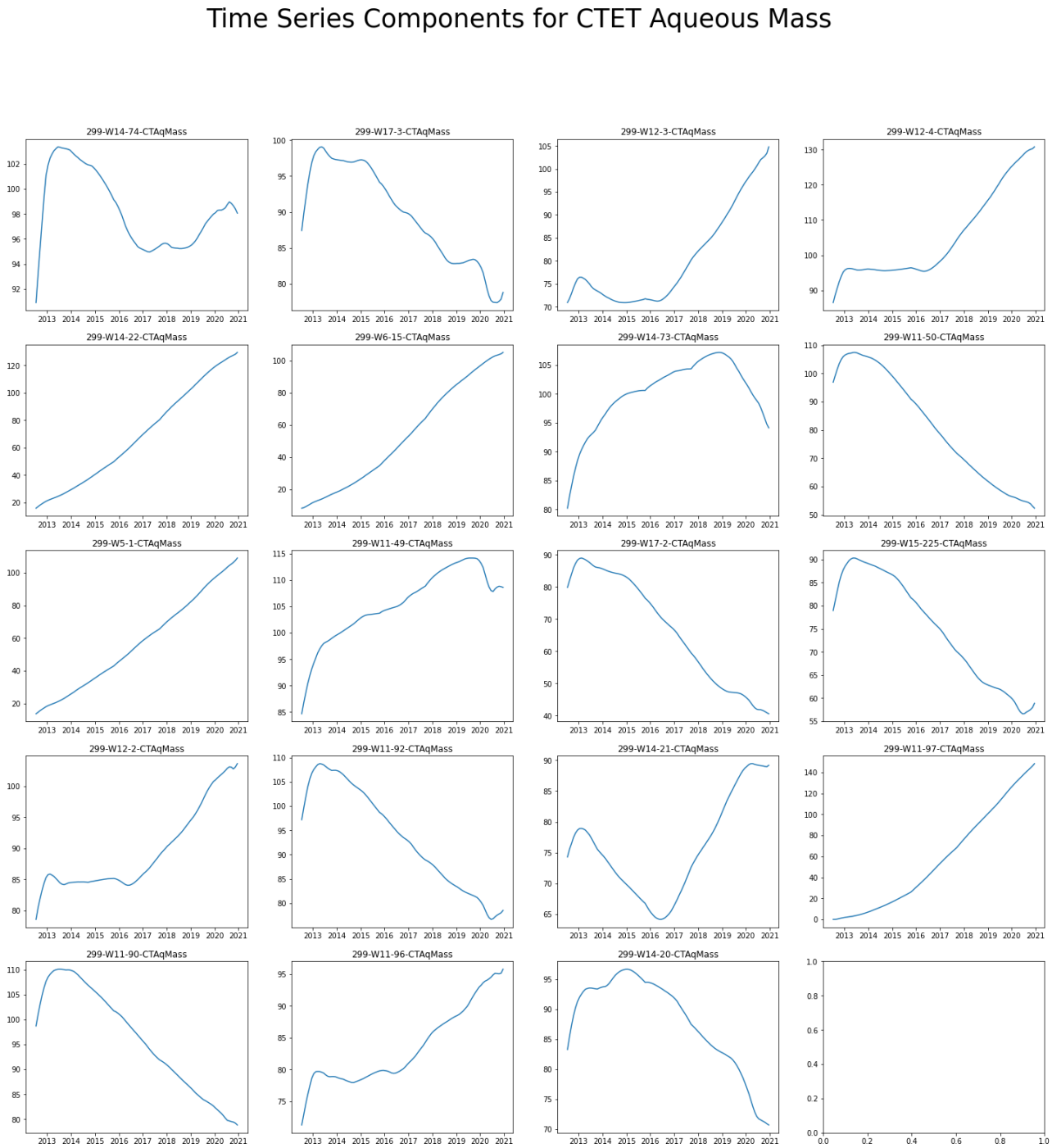

Figure A-10. These visualizations show all the identified trend components for all 19 extraction wells used for data analysis of the 200 West Areas for the CCl4 aqueous mass dataset.