Identifying Seasonal Coherence in Global Lake Surface Water Temperature

Boyang Sun*

Department of Statistics,University of Glasgow, Glasgow, United Kingdom

*Corresponding author: Byron0613@hotmail.com

Abstract. This project examines seasonal water temperature patterns in thirty large lakes worldwide from 2003 to 2012. Analyzing and identifying seasonal patterns can help understand lake water temperature changes better and predict future trends. The seasonal pattern of different lakes is decomposed by a generalized additive model (GAM) and Seasonal and Trend decomposition using Loess (STL), and 30 lakes are clustered and classified according to the results, which shows that different lakes have the same seasonal pattern of water temperature. The reason for the classification is explained using latitude, longitude, and elevation. This study contributes to a deeper understanding of the seasonal patterns of lakes of the same type while distinguishing the differences in latitude, longitude, and elevation of different types of lakes.

Keywords: Seasonal Pattern; Cluster Analysis; Time Series; Generalized Additive Model; Generalized Linear Model.

1. Introduction

1.1. Research Background

Seasonal patterns cover cyclical changes in data that occur regularly throughout the year. The importance of studying seasonal patterns is to reveal and explain cyclical changes in data and help people better understand the trends, fluctuations, and influencing factors of data. By studying seasonal patterns, future trends can be predicted more accurately, helping to make more informed decisions.

As complex ecosystems, lakes contain rich climate and environmental information in their seasonal water temperature patterns. The seasonal pattern of lake water temperature changes cyclically with the seasons. This change is closely related to sunlight and geographical location and profoundly impacts the lake ecosystem.

1.2. Research Questions

To estimate the average seasonal pattern at each of the lakes.

To identify coherent groups of lakes regarding their seasonal patterns of LSWT and Investigate the statistically optimal number of groups required.

Using the location and elevation data provided to informally explore any drivers of the differences in these groups.

2. Literature Review

This section provides an overview of existing studies on seasonal patterns, cluster analysis, generalized additive model (GAM), temperature classification, and generalized linear model (GLM). It highlights key findings, debates, and gaps in the literature, laying the groundwork for current research contributions.

2.1. Seasonal Pattern

STL is a seasonal and trend decomposition method proposed by Cleveland in 1990. After a period of development, many scholars use STL in statistics. Craig A. Stow (2015) used STL to study the seasonal pattern of nitrogen content in rivers and guided production based on the conclusions, which solved the problem of eutrophication in rivers. British statisticians Trevor Hastie and Robert Tibshirani proposed GAM in 1986. GAM extends the linear regression model by allowing predictor variables to model the non-linear response variable. D. de Brogniez (2014) applied GAM to model non-linear data, further proving the significance of GAM in statistics. Cluster analysis can be used to reveal structure and patterns within data.

2.2. Cluster Analysis

This study used two clustering methods, Hierarchical Agglomerative Clustering (HAC) and K-means, to classify the data.S. Ward (1963) proposed a classification method of mutually exclusive samples in his research, which is also considered the origin of the HAC model. HAC solved the problem of large sample classification under limited computing power at that time. K-means clustering is a statistical analysis method that classifies samples according to the distance between samples. Divide data points into groups, each with data points closer to its interior. K-means was further studied by Charles Elkan (2003). Robert Gens (2012) combined K-means with other clustering methods to further improve the accuracy of the results.

John Nelde (1972) proposed the generalized linear model (GLM) for the first time and introduced the concept, construction, and application of GLM in detail. Annette J. Dobson (2018) released the fourth edition of the book on GLM, in which the origin and application of GLM are stated in detail, and the coincidence of Bayesian statistics and generalized linear models is supplemented in the last part.Many scholars have discussed the effect of latitude on temperature. Gabriele C. Hegerl (2009) found that seasonal patterns strongly correlate with latitude. However, Keith Montgomery (2007) stated that latitude is not the only factor affecting temperature. Catherine M (2015) used the time series of lake water temperature and combined it with the geographical location of the lake for analysis, and finally demonstrated that the lake temperature will be affected by many random factors besides geographical location factors.

3. Analysis of the data

3.1. Data processing

The box plot is an effective way to observe multiple variables, and the results obtained are shown in Figure 1.

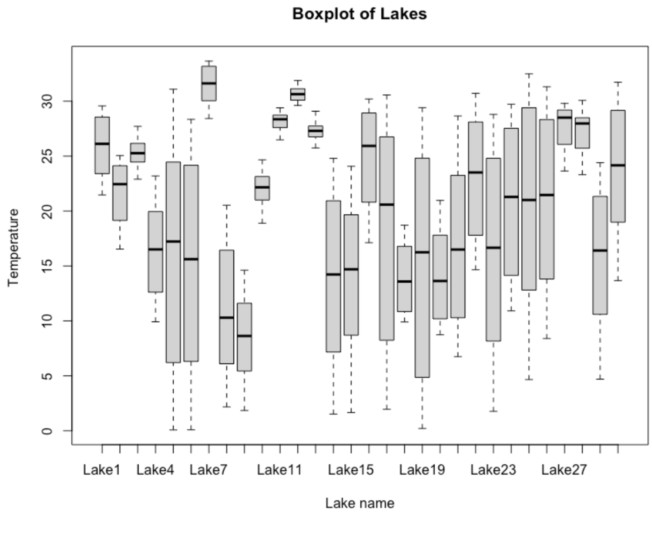

Fig. 1 Boxplot of lakes

The temperature characteristics of different lakes are very different. For example, the average temperature of Lake 1 is about 26 degrees. In comparison, the average temperature of Lake 4 is only 16 degrees, so there is a big difference in the average temperature of the lakes.

Before analyzing a time series, the author must consider whether the time series is stationary. In this project, the ADF test detects whether the time series is stationary. Because the p-value is significantly less than 0.05, it can be concluded that the time series is stationary.

In order to get the average seasonal pattern, the seasonal pattern of the water temperature of each lake is first extracted. Because time series mainly consists of three parts: trend, seasonal pattern, and error, the author must separate the seasonal pattern from the sequence data. Because the conclusion of time series stability has just been obtained, and the water temperature of the same lake in different years is the same. Based on the above judgments, this dataset can extract seasonal patterns.

The result of STL to decompose the time series is shown in Figure 2, which shows that the seasonal pattern repeats continuously within nine years. The time series is fitted with a generalized additive model to get the average seasonal pattern.

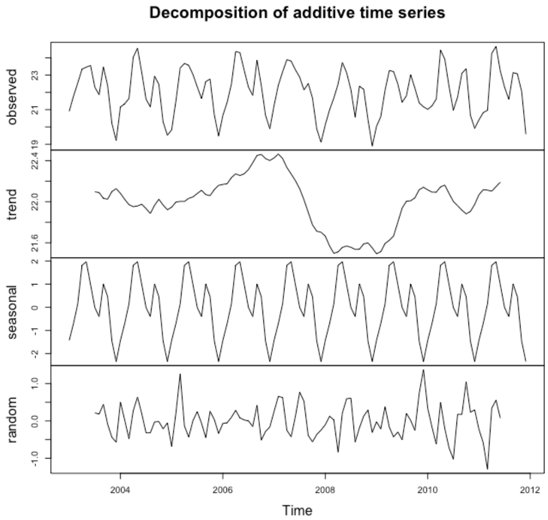

Fig. 2 Decomposition of additive time series

By fitting the time series of each lake with GAM, the seasonal pattern parameters of each lake can be obtained. A line chart can be drawn according to the seasonal pattern parameters, and finally, a line chart of the seasonal pattern containing 30 lakes is obtained in Figure 3.

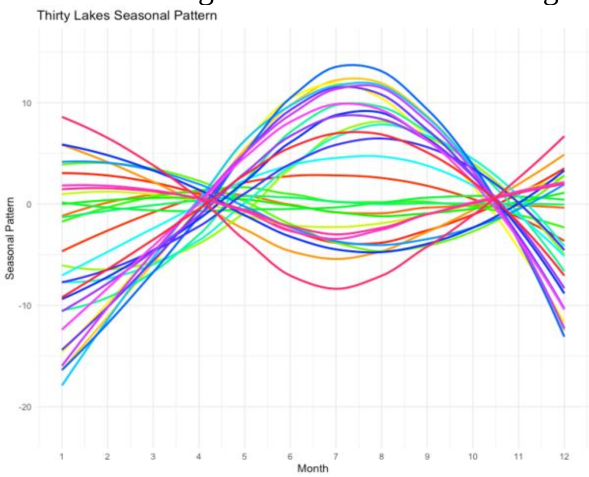

Fig. 3 Thirty lakes seasonal pattern

3.2. Classification

The author used K-means clustering and Hierarchical Agglomerative Clustering(HAC) in the classification process. Because K-means is a clustering method which need to set clusters upfront, the author needs to choose the number of clusters first before classification. From Figure 4 below, the average silhouette width reaches the inflection point when there are two clusters, so it can be determined that the two classes are relatively suitable classification methods.

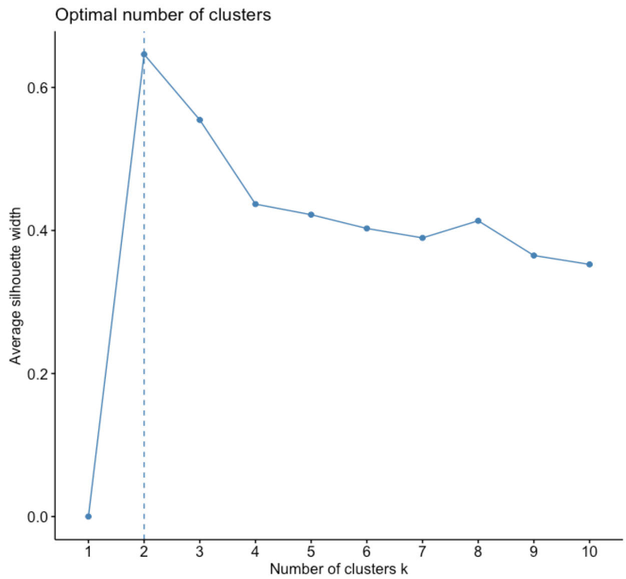

Fig. 4 Optimal number of clusters

3.3. K-means clustering

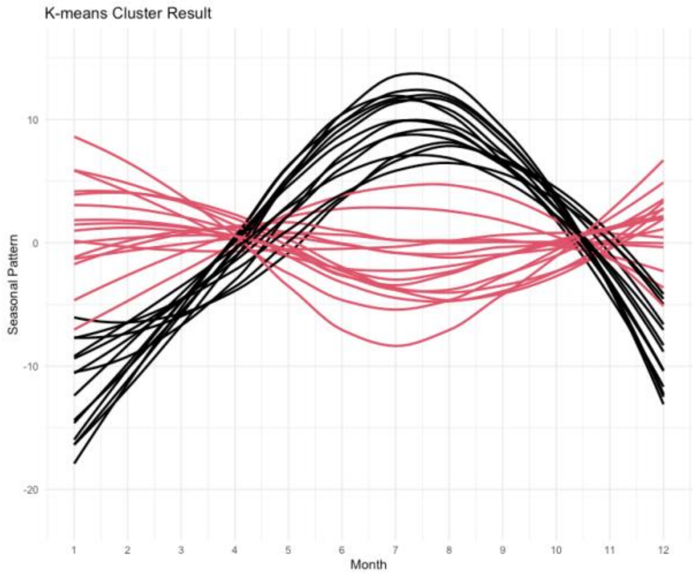

Fig. 5 K-means cluster result

As shown in Figure 5, the 30 lakes were evenly divided into two categories. According to the K-means results, dividing the 30 lakes into two categories can explain 79.6% of the variance. According to the previous analysis, Lake 7 and Lake 27 are classified into Group 2, and Lake 6 and Lake 16 are classified into Group 1. The results obtained by K-means clustering are consistent with the previous analysis.

3.4. HAC

In order to verify the accuracy of K-means clustering and to use the dendrogram to more intuitively display the variables contained in different categories, The author used HAC to retest the results. The result obtained is shown in the figure below.

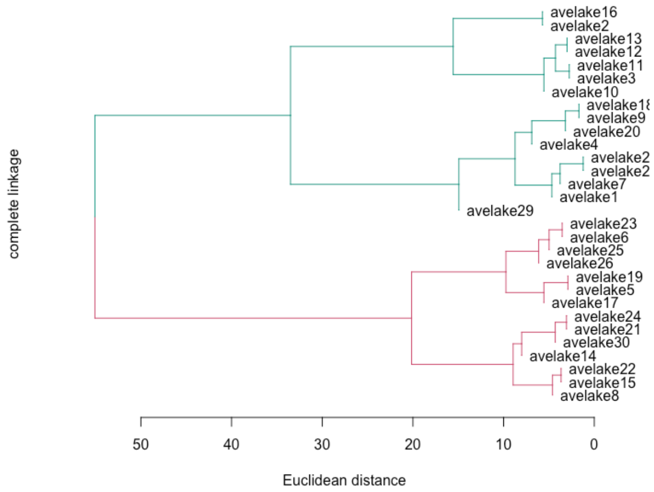

Fig. 6 HAC result

From the figure, except for the difference in the Lake 16 classification, K-means and HAC are entirely consistent in other classification results. Because HAC uses the Euclidean distance, only the distance characteristics between temperatures are considered, and the weight of seasonal pattern characteristics is relatively tiny. Therefore, the results obtained by using K-means clustering are selected here, and Lake16 is divided into the second group.

3.5. Exploring the Motivation of Classification

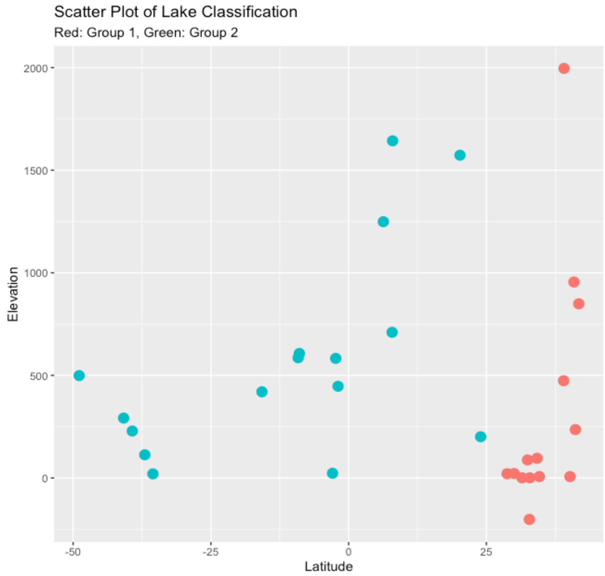

Fig. 7 Scatter plot of lake classification

After incorporating the classification categories into the latitude, longitude, and elevation, the author can find that all lakes with latitude greater than or equal to 23.96°N are classified into the second group. In contrast, all lakes with latitude less than or equal to 20.21°N are classified into group 2.

3.6. Main findings

In this project, through data analysis, it can be found that the seasonal pattern of lake water temperature is obvious. Further, through cluster analysis, two categories can be obtained. Both latitude and elevation are geographic factors that can affect lake classification. Lakes with higher latitudes fall into one category in the northern hemisphere due to the same seasonal pattern. From 10°S to 10°N, there will be multiple peaks and troughs in the lake water temperature.

4. Conclusions

4.1. Summary

This classification divides all lakes into two categories. All lakes in group 1 are located in the northern hemisphere, and the latitude is higher than 23°N. Group 2 contains tropical lakes and lakes in the southern hemisphere. In group 2, these lakes can be further classified according to whether there are multiple inflection points in a year. The main reason for the distinction between group 1 and group 2 is that the seasonal characteristics of the northern hemisphere are summer in July and August and winter in January and February. In group 2, tropical lakes have multiple inflection points in temperature changes within a year due to sufficient sunlight.

4.2. Further research

This study is mainly based on seasonal patterns, so the trend in the time series is deliberately eliminated in the analysis of the article. As one of the essential components in time series, the trend itself also has much content worth studying, such as long-term temperature increase or decrease, drivers of trends, and predictive modeling of lake water temperature using trends. Removing the trend part may lead to some data waste. Interesting conclusions can be drawn in the future by focusing on analyzing trends.

References

[1] Amidon, E.L., Sharpnack, D.A. and Russell, R.M. (1971) The development of an Earth Resources Information System using aerial photographs and digital computers. Berkeley: University of California.

[2] Bonifazi, F. et al. (2013) 'Glutathione transferase-A2 S112T polymorphis predicts survival, transplant-related mortality, busulfan and bilirubin blood levels after allogeneic stem cell transplantation', Haematologica, 99(1), pp. 172–179. doi:10.3324/haematol.2013.089888.

[3] Bruggen, V. et al. (2018) 'Structural validity of the World Assumption Scale', Journal of Traumatic Stress, 31(6), pp. 816–825. doi:10.1002/jts.22348.

[4] Canova, F. and Hansen, B.E. (1995) 'Are seasonal patterns constant over time? A test for seasonal stability', Journal of Business & Economic Statistics, 13(3), p. 237. doi:10.2307/1392184.

[5] Courault, D., Seguin, B. and Olioso, A. (2005) 'Review on estimation of evapotranspiration from remote sensing data: From empirical to numerical modeling approaches', Irrigation and Drainage Systems, 19(3–4), pp. 223–249. doi:10.1007/s10795-005-5186-0.

[6] Courault, D., Seguin, B. and Olioso, A. (2005) 'Review on estimation of evapotranspiration from remote sensing data: From empirical to numerical modeling approaches', Irrigation and Drainage Systems, 19(3–4), pp. 223–249. doi:10.1007/s10795-005-5186-0.

[7] de Brogniez, D. et al. (2014) 'A map of the topsoil organic carbon content of Europe generated by a generalized additive model', European Journal of Soil Science, 66(1), pp. 121–134. doi:10.1111/ejss.12193.

[8] Dunteman, G.H. and Ho, M.-H.R. (2006) An introduction to generalized linear models. Thousand Oaks, CA: Sage Publications.

[9] Efron, B. (1988) 'Logistic regression, survival analysis, and the kaplan-meier curve', Journal of the American Statistical Association, 83(402), pp. 414–425. doi:10.1080/01621459.1988.10478612.

[10] Gens, R. and Domingos, P. (2012) Discriminative learning of sum-product networks. Available at: https://homes.cs.washington.edu/~pedrod/papers/nips12.pdf (Accessed: 22 August 2023).

[11] Hamerly, G. and Elkan, C. (2003) Learning the K in K-means | proceedings of the 16th international ... Available at: https://dl.acm.org/doi/10.5555/2981345.2981381 (Accessed: 21 August 2023).

[12] Hastie, T., Friedman, J. and Tisbshirani, R. (2017) The elements of Statistical Learning: Data Mining, Inference, and prediction. New York: Springer.

[13] Hastie, T.J. and Tibshirani, R.J. (2017) 'Generalized additive models', Generalized Additive Models, pp. 136–173. doi:10.1201/9780203753781-6.

[14] Huber, P.J. (1987) 'Experiences with three-dimensional scatterplots', Journal of the American Statistical Association, 82(398), pp. 448–453. doi:10.1080/01621459.1987.10478447.

[15] Hubert, L.J. (2004) 'Hierarchical Cluster Analysis', Encyclopedia of Statistical Sciences [Preprint]. doi:10.1002/0471667196.ess0947.

[16] Johnson, V.E. and Albert, J.H. (1999) 'Regression models for ordinal data', Ordinal Data Modeling, pp. 126–157. doi:10.1007/0-387-22702-4_4.

[17] Križanić, S. (2020) 'Educational data mining using cluster analysis and decision tree technique: A case study', International Journal of Engineering Business Management, 12, p. 184797902090867. doi:10.1177/1847979020908675.

[18] Lang, S. et al. (2010) 'Earth observation (eo)-basedex postassessment of Internally Displaced Person (IDP) camp evolution and population dynamics in Zam Zam, Darfur', International Journal of Remote Sensing, 31(21), pp. 5709–5731. doi:10.1080/01431161.2010.496803.

[19] Lausch, A. et al. (2016) 'Understanding Forest Health with remote sensing -part I—a review of spectral traits, processes and remote-sensing characteristics', Remote Sensing, 8(12), p. 1029. doi:10.3390/rs8121029.

[20] Mann, M.E. et al. (2000) 'Global temperature patterns in past centuries: An interactive presentation', Earth Interactions, 4(4), pp. 1–1. doi:10.1175/1087-3562(2000)004<0001:gtpipc>2.3.co;2.

[21] Montgomery, K. (2006) 'Variation in temperature with altitude and latitude', Journal of Geography, 105(3), pp. 133–135. doi:10.1080/00221340608978675.

[22] Ozesmi, S.L. and Bauer, M.E. (2014) Satellite remote sensing of wetlands - wetlands ecology and management, SpringerLink. Available at: https://link.springer.com/article/10.1023/A:1020908432489

[23] Russell, K.G. (2018) 'Generalized linear models', Design of Experiments for Generalized Linear Models, pp. 1–20. doi:10.1201/9780429057489-1.

[24] Stow, C.A. et al. (2015) 'Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie', Environmental Science & Technology, 49(6), pp. 3392–3400. doi:10.1021/es5062648.

[25] Stroud, I. and Xirouchakis, P.C. (2000) 'STL and extensions', Advances in Engineering Software, 31(2), pp. 83–95. doi:10.1016/s0965-9978(99)00046-0.

[26] Ward, J.H. (1963) 'Hierarchical grouping to optimize an objective function', Journal of the American Statistical Association, 58(301), pp. 236–244. doi:10.1080/01621459.1963.10500845.

[27] Zhang, H. (1998) 'Bayesian Cart Model Search: Comment', Journal of the American Statistical Association, 93(443), p. 948. doi:10.2307/2669833.