# Discovering relationships among economic variables using machine learning techniques

**Y. Guo, J. Li, T. Lo, Z. Zhu, G. Lee and P. Toscas**

*Data61, CSIRO*
*Email: Geoff.Lee@csiro.au*

**Abstract:** Understanding relationships among economic variables is crucial when establishing an economic multi-variate forecast model, such models are useful tools for generating indicators of overall future market trend and for aiding policy and investment decision making. Numerous researchers have made significant contributions to the traditional study of economic variable relationships over the decades.

Traditional studies have relied on theoretical economic models to investigate the relationships among economic variables. However, these theoretical models and assumptions may not hold in practice. Additionally, the use of historical data to make predictions about future trends may be subject to uncertainty and other factors that cannot be captured by such econometric models alone. While econometric models have been valuable in many areas of economic research, machine learning techniques offer a complementary approach that can provide new insights and uncover relationships among economic variables that may not be apparent using traditional techniques. Some research papers have explored the limitations of theoretical models by applying advanced machine learning algorithms to the economic and finance area. For instance, clustering algorithms have been extensively utilized in unsupervised learning to identify similarities and group observations during economic events. Although some papers present the potential for using clustering methods to solve a range of different problems in economics, most of them apply the methodology to analyse stock market or economic performance rather than interrogating the relationships between economic variables more broadly.

In this paper, we propose a machine learning clustering method to group behavior traces of financial variables and identify relationships between variables. The algorithm is based on k-means clustering with either Dynamic Time Warping (DTW) or Complexity-Invariant Distance (CID). By analyzing the clustering results, we construct comparable variables and investigate the relationships among economic variables. The economic variables include agricultural-related variables (beef price), energy-related variables (gas), and the Australian Consumer Price Index (CPI). In addition to the clustering methodology, we have developed a set of visualisation tools that not only demonstrate the clustering model outputs but also provide insight to aid decision-making. With information readily available, decision makers can make informed decisions and gain valuable insights into the relationships between financial variables, helping them to improve their business strategies and outcomes. By enabling more informed decision-making, our methodology and tools can be particularly useful for those operating in financial markets, where understanding the relationships between variables is crucial for making effective decisions.

***Keywords:*** *Machine learning, clustering, economic variable*

## 1. INTRODUCTION

Understanding relationships among economic variables is crucial when establishing an economic multi-variate forecast model; such models are useful tools for generating insights when analyzing different future scenarios in formulating economic policies and for investment decision making. Numerous researchers have made significant contributions to the traditional study of economic variable relationships over the decades. Engle et al. (1987) introduced a methodology for modeling long-run relationships among economic variables with cointegration and error correction models. Johansen (1995) extended this work to provide likelihood-based inference for cointegrated vector autoregressive models. Stock et al. (1993) proposed a simple estimator of cointegrating vectors in higher-order integrated systems. Fama et al. (1988) research on dividend yields and expected stock returns is a classic in the field of asset pricing, highlighting the importance of dividend yields as predictors of future stock returns. Granger et al. (1974) warns against interpreting correlation as causation in econometrics and propose a method to detect spurious regressions based on testing for stationarity and cointegration of the variables.

The relationships among economic variables constantly change with time, new theoretical economic models are regularly proposed to investigate these relationships. However, these theoretical models and assumptions may not hold in practice. Additionally, the use of historical data to make predictions about future trends may be subject to uncertainty and other factors that cannot be captured by econometric models alone. While econometric models have proven valuable in many areas of economic research, machine learning techniques offer a complementary approach that can provide new insights and uncover relationships among economic variables that may not be apparent using traditional techniques.

Some research papers have explored the limitations of theoretical models by applying advanced machine learning (ML) algorithms to economic and finance studies. For instance, clustering algorithms have been extensively utilized in unsupervised learning to identify similarities and group observations during economic events. The most common clustering algorithm for time series datasets is k-means clustering by MacQueen (1967). For example, Focardi et al. (2001) examined the clustering of economic and financial time series to explore the existence of stable correlation conditions. Lai et al. (2010) proposed a novel two-level clustering method for time series data analysis. Wang et al. (2020) presented a regional intelligent economic decision support system constructed using a fuzzy C-mean clustering algorithm. Marti et al. (2021) provided a comprehensive review of two decades of correlations, hierarchies, networks, and clustering in financial markets. Overall, these papers employed a range of clustering algorithms to investigate the relationships between economic variables and their impacts on financial markets.

Although these papers present the potential of using clustering methods to solve a range of different problems in economics, most of them apply the methodology to analyse stock market or economic performance rather than interrogating the relationships between economic variables more broadly. We propose a machine learning clustering method to group behavior traces of financial variables and identify relationships between variables. The algorithm is based on k-means clustering with either Dynamic Time Warping (DTW) or Complexity-Invariant Distance (CID) by Batista et. al. (2014). By analyzing the clustering results, we construct comparable variables and investigate the relationships among economic variables. The economic variables include agricultural-related variables (beef price), energy-related variables (gas), and the Australian Consumer Price Index (CPI). We also developed a web visualisation toolbox that enables users to choose different economic variables and study their relationships.

The remainder of the paper is organized as follows: section 2 presents the methodological details; in section 3 a case study is presented and the methodology is evaluated; and we end with a conclusion.

## 2. METHODOLOGY

There are generally three different ways to cluster time-series; shape-based, feature-based and model-based (Aghabozorgi, 2015). In the shape-based approach, shapes of two time-series are matched as well as possible, by a non-linear stretching and contracting of the time axes. The choice of a proper distance approach depends on the characteristic of time-series, representation method, and on the objective of clustering time-series. Traditional shape-based similarity measures include Dynamic Time Warping (DTW) (Petitjean et al. 2011, Izakian et al. 2015), and Longest Common Sub-Sequence (LCSS) (Ding et al. 2008). They can be efficiently computed, but are sensitive to noise, scale and time shifts. Batista et. al. (2014) proposed a Complexity-Invariant Distance (CID), which is based on the use of a measure of complexity difference between two time series as a correction factor for a standard distance measure between two time series. CID is reported to be significantly better than DTW on published results (Bagnall et al. 2017).

Once a similarity measure is computed between the time series, clustering needs to be performed with a suitable algorithm. For the commonly used k-means algorithm, the results of clustering depend heavily on the positions of the initial cluster centers, resulting in local minimum; and it can only handle linearly separable clusters. Likas et. al. (2003) proposed a global k-means algorithm, which is not dependent on the initial conditions, and the solution is near optimal. In Dhillon et al. (2004), a global kernel k-means algorithm is proposed as an extension of the global k-means algorithm, where data from the input space is mapped to a high-dimensional feature space by a kernel function that minimizes errors in the clustering feature space. The clusters generated by kernel k-means are phase dependent, hence two time series that differ in phase rather than in shape can also be separated. We implement the shape-based approach, CID, using the Fast Global Alignment kernel (GAK) (Cuturi 2011) as the core of the global kernel k-means algorithm. The approach can clearly separate the time series according to shapes and characteristics.

CID is based on the use of a measure of complexity difference between two time series as a correction factor for a standard distance measure between two time series. Let *X* and *Y* be two time series variables. A simple complexity estimation can be defined as:

$$CE_X = \sqrt{\sum_{i=1}^{n-1} (x_i - x_{i+1})^2}\,.$$

By using a Euclidean distance, the Complexity-Invariant Distance measure is computed as follows:

$$CID_{(X,Y)} = CF_{(X,Y)} \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\,,$$

where $CF_{(X,Y)}$ is the correction factor defined by

$$CF_{(X,Y)} = \frac{max(CE_X, CE_Y)}{min(CE_X, CE_Y)}\,.$$

Once the CID distance between time series has been computed, we perform Fast Global Alignment kernel (GAK) at the core of a global kernel k-means algorithm to perform time series clustering. Typical kernel functions include the polynomial kernel, the Gaussian radial basis function and the sigmoid kernel function. The main idea of global kernel k-means is that a near-optimal solution with *k* clusters can be obtained by starting with a near-optimal solution with *k-1* clusters and initializing the *k*th cluster appropriately based on a local search. During the local search, *N* initialisations are tried, where *N* is the size of the data set. The *k-1* clusters are always initialized to the *k-1*-clustering problem solution, while the *k*th cluster for each initialisation includes a single point of the data set. The solution with the lowest clustering error is kept as the solution with *k* clusters. Since the optimal solution for the *1*-clustering problem is known, the above procedure can be applied iteratively to find a near-optimal solution for different numbers of clusters. More details of the GAK approach are described in Cuturi 2011. We then apply the Elbow method to find the optimal number of clusters *K* for the time series.

## 3. IMPLEMENTATION OF CLUSTERING ALGORITHM

### 3.1. Datasets summary

The proposed clustering algorithm in Section 2 can be used to construct comparable economic variables and analyse the relationships among them. These variables can be broad, such as price inflation, wage inflation, long-term interest rate, short-term interest rate, cash, and unemployment rate amongst others. The list of variables can also extend beyond the financial area to include other domains, such as energy or agriculture, with variables like electricity price and wheat price. To explore the methods across different domains, we will examine the relationship among agricultural-related variables (beef price), energy-related variables (gas), and economic variables (CPI rate). The CPI data for Australia is quarterly data from the Australian Bureau of Statistics (2023). The beef price is the monthly data from Bloomberg, covering the period from January 1993 to December 2022, while the gas price is the daily trading price from the Henry Hub Natural Gas Spot Price, spanning January 1992 to September 2021. Therefore, the matching period for these three variables is quarterly from January 1993 to September 2021, as shown in Figure 1. The x-axis represents the timeline while the y-axis represents the values of the three variables: beef price (in Australian Dollars), gas price (in Australian Dollars), and CPI rate (as a growth ratio of CPI).

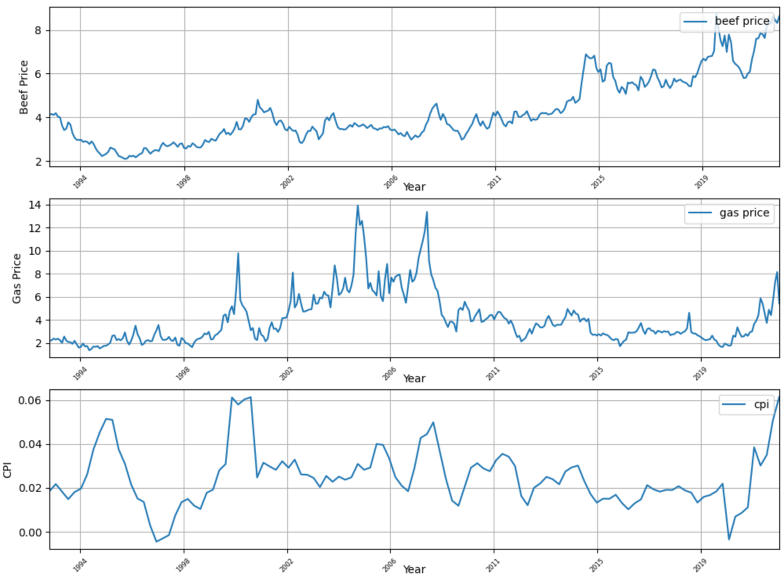


**Figure 1.** Three variable datasets from 1993 to 2021, including beef price, gas price, and CPI growth rate. The x-axis represents the timeline while the y-axis represents the values of three variables: beef price (in Australian Dollars), gas price (in Australian Dollars), and CPI rate (as the rate of change of the CPI).

### 3.2. Clustering algorithm results

The purpose of ML-based clustering is to group time series with similar shapes and characteristics, as described in Section 2. Two more parameters need to be optimized to achieve the best clustering and relationship results. One parameter is the optimal number of clusters, which is found using the commonly used Elbow method. The other parameter is the lag between variables, which is necessary to consider when comparing financial variables, as the market needs time to react and the government needs time to collect and calculate statistical data. We calculated the correlation matrix between variables with lags from 0 to 3 and chose the lag with the highest correlation as the most suitable choice. Figure 2 shows the flowchart of the cluster number and lag optimisation process, with beef and CPI as examples.

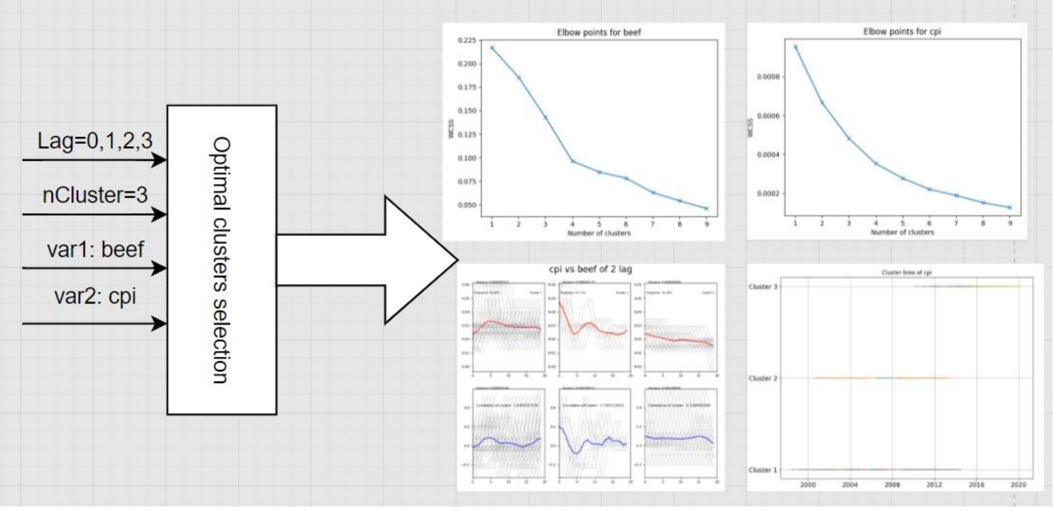


**Figure 2**. Flowchart of the cluster number and lag optimisation process, with beef and CPI as examples.

We segmented the variable data into 5-year windows shifting from 1993 to 2021. Since the data sample rate is uniform at quarterly intervals for all variables, each time series $T$ has a length of $n$=20. With the optimized parameters, we first cluster one variable (called the explanatory variable) into k clusters. Within each cluster, there are many 5-year time series $t$ (grey lines) with similar shapes or characteristics. We then use the optimal lag value to find the corresponding time series of the other variables (called the response variables). Hence, the response variables are also grouped into $k$ clusters. We can observe not only the shape of all these grouped time series but also summarize their statistical features, such as the average curve shape (red or blue lines) and the spread period of each cluster, as shown in the small windows in Figure 2.

Two clustering results are presented in Figures 3 and 4 as examples. In Figure 3, the explanatory variable was CPI, and the response variable was beef price. Each grey line represents a 5-year time series. The optimal cluster number is *k*=3, and the optimal lag value is 2, which indicates that beef prices are influenced by CPI with a delay of six months (as the sample rate is quarterly or every three months). The top plots show the three groups of CPI time series, while the corresponding beef price curves are shown in the bottom plots. The red line indicates the average curve of CPI. The blue line represents the average curve of beef. The average curve of CPI (red line) shows clear patterns or characteristics: cluster one represents a stable period with slow growth in the first five points, cluster two shows a downward trend followed by a stable trend, while cluster three represents a slow declining period. Six months later, the beef price follows very similar patterns caused by CPI.

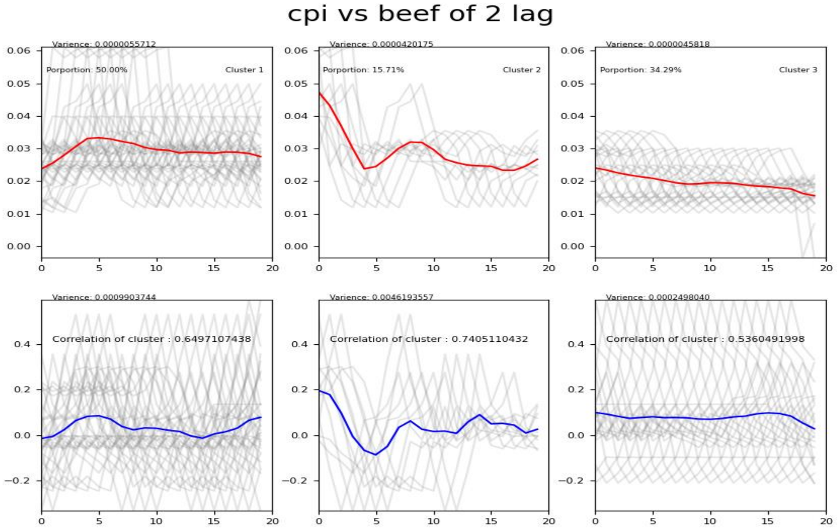


**Figure 3**. CPI as the explanatory variable, and beef price annual change rate as the response variable. Each grey line stands for one 5-year time series *t*. Top row: CPI is clustered into three groups. The y-axis represents the value of the CPI rate. Bottom row: the corresponding beef price growth rate curves with lag=2 (six months delay). The y-axis represents the value of the beef price growth rate. Red line for the average curve of CPI rate. Blue line for the average curve of beef price growth rate.

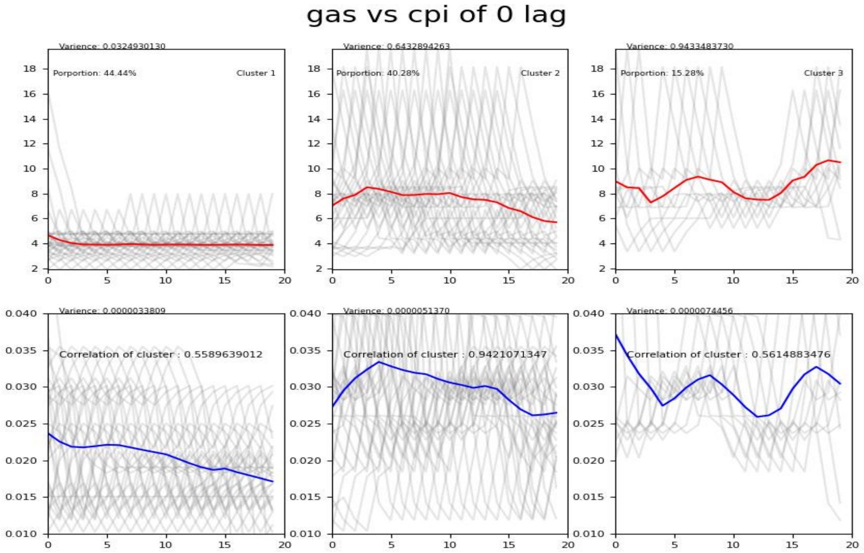


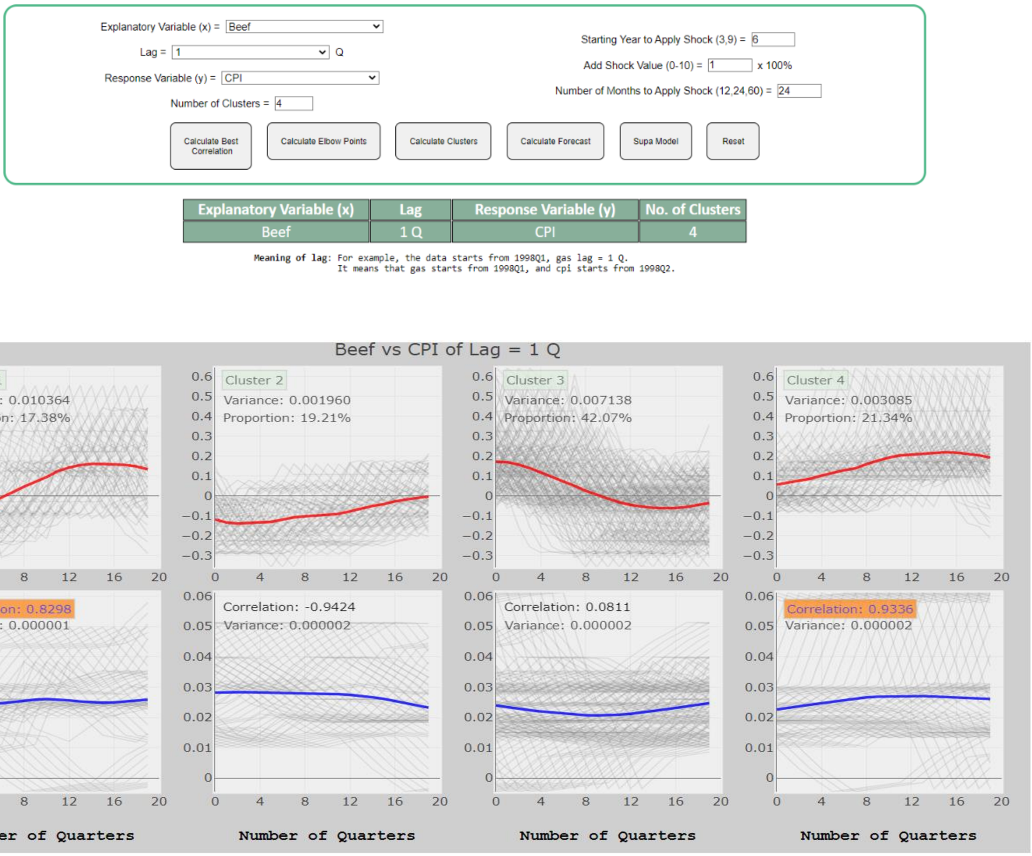
**Figure 4**. Gas price as the explanatory variable, and CPI rate as the response variable. Top row: Gas price clustered into three groups. The y-axis represents the value of the gas price (in Australian Dollars). Bottom row: the corresponding CPI with lag=0 (no delay). The y-axis represents the value of the CPI rate (as the

change rate of CPI). Red line for the average curve of gas price. Blue line for the average curve of CPI growth rate.

Similarly, Figure 4 illustrates the relationship between gas price (explanatory variable) and CPI (response variable). Again, the optimal cluster number is *k*=3, and the optimal lag value is 0, indicating that CPI is almost instantly affected by gas prices. The average curve of CPI exhibits clear patterns or characteristics, with cluster one representing a stable and low period, cluster two showing a slight increase then a downward trend, and cluster three representing an unstable waving period. Almost simultaneously, the CPI curve follows similar patterns caused by gas prices. These findings indicate that gas prices are both a crucial and quick driver of CPI.

### 3.3. Development of web visualisations

In addition to the clustering methodology presented in the paper, the outputs from such a clustering algorithm should be presented graphically in an intuitive way for easy interpretation to gain insights.



Explanatory Variable (x) = Beef

Lag = 1 Q

Response Variable (y) = CPI

Number of Clusters = 4

Starting Year to Apply Shock (3,9) = 6

Add Shock Value (0-10) = 1 x 100%

Number of Months to Apply Shock (12,24,60) = 24

| Explanatory Variable (x) | Lag | Response Variable (y) | No. of Clusters |
|---|---|---|---|
| Beef | 1 Q | CPI | 4 |

**Meaning of lag**: For example, the data starts from 1998Q1, gas lag = 1 Q.
It means that gas starts from 1998Q1, and cpi starts from 1998Q2.

**Figure 5**. Web-based visualisations of the developed clustering methodology. Top: variables that users can set or calculate. The y-axis represents the value of the beef price growth rate (as a ratio). Bottom: clustering results based on the user's choices. The y-axis represents the value of the CPI rate (as a ratio).

Figure 5 displays the web interface that has been implemented in the paper. The top part of the web UI is the input area that users can specify. Firstly, users can select an explanatory variable x (such as Beef or Gas) and then a response variable y (such as CPI) of interest. The "Calculate Elbow Points" function will show the graphs of the WCSS (Within-Cluster Sum of Square) values against the number of clusters, which varies from 1 to 9, for both x and y. This helps users find the optimal number of clusters, *K*. The "Calculate Best Correlation" function will calculate all the correlations between the selected explanatory variable x and the response variable y in each cluster (cluster 1 to cluster *K*) and for each lag (lag 0 to lag 3). The lag with the maximum of all the correlations will be selected as the best lag, L. Users can either input the above calculated values or any pre-set reasonable values into the "Lag" (*L*) and "Number of Clusters" (*K*) in the UI. The "Calculate Clusters" function will display the corresponding 5-year window graphs and the timing graph for each cluster of the selected values of *L* and *K*. The functions related to "Calculate Forecast" and "Super Model" are beyond the scope of this paper. The bottom plot in Figure 5 is just one example. Once the variables in the top part of the UI are chosen, detailed results can be shown in many plots below. This tool can help decision makers explore and analyse the clustering results based on their specific interests and needs. With this information readily available, decision makers can make informed decisions and gain valuable insights into the relationships between financial variables, ultimately helping them to improve their business strategies and outcomes.

## 4. CONCLUSION

In this paper, we have presented a novel clustering methodology for analyzing financial time series data and identifying the relationships between economic variables. Our approach has allowed us to gain valuable insights into how different variables are correlated and how they can impact one another over time, such as the effect of gas prices on CPI rate. We have also developed a more intuitive web-based visualisation tool that allows users to explore and identify relationships among economic variables further by easily setting the number of clusters or lag periods. The methodology and visualisation tool can also allow input of new economic variables, so users are able to explore relationships among economic variables of their own interest. As a generic tool for gaining insight about various time series, our methodology and tool can be particularly useful for those operating in financial markets, where understanding the relationships between economic variables is crucial for making effective decisions.

## REFERENCES

Cuturi, M., 2011. Fast global alignment kernels. Proceedings of the 28th International Conference on Machine Learning (ICML-11), 929-936.

Dhillon, I.S., Guan, Y., and Kulis, B., 2004. Kernel k-means: spectral clustering and normalized cuts. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 551-556.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E., 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 1(2), 1542-1552.

Engle, R.F., and Granger, C.W., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica, 55(2), 251-276.

Fama, E.F., and French, K.R., 1988. Dividend yields and expected stock returns. Journal of Financial Economics, 22(1), 3-25.

Focardi, S.M., and Fabozzi, F.J., 2001. Clustering economic and financial time series: Exploring the existence of stable correlation conditions. Discussion Paper.

Granger, C.W., and Newbold, P., 1974. Spurious regressions in econometrics. Journal of Econometrics, 2(2), 111-120.

Gustavo, E.A.P.A., Batista, A., Keogh, E.J., Tataw, O.M., Vinícius, M., de Souza, A., 2014. CID: An efficient complexity-invariant distance for time series. Data Mining and Knowledge Discovery 28(3), 634-669.

Izakian, H., Pedrycz, W., and Jamal, I., 2015. Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, 235-244.

Johansen, S., 1995. Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press.

Lai, C.P., Chung, P.C., and Tseng, V.S., 2010. A novel two-level clustering method for time series data analysis. Expert Systems with Applications, 37(9), 6319-6326.

Likas, A., Vlassis, N., and Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognition, 36(2), 451-461.

MacQueen, J.B., 1967. Classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1(14), 281-297.

Marti, G., Nielsen, F., Bińkowski, M., and Donnat, P., 2021. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. Progress in Information Geometry: Theory and Applications, 245-274.

Petitjean, F., Ketterlin, A., and Gançarski, P., 2011. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition, 44(3), 678-693.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, Teh Ying Wah. (2015). Time-series clustering – A decade review, Information Systems, Volume 53, 16-38.

Stock, J.H., Watson, M.W., 1993. A simple estimator of cointegrating vectors in higher order integrated systems. Econometrica 61(4), 783-820.

Wang, J., Yu, C., Zhang, J., 2020. Constructing the regional intelligent economic decision support system based on fuzzy C-mean clustering algorithm. Soft Computing 24, 7989-7997.