materials proceedings

MDPI

Proceeding Paper

Assessing Behavior Similarity of Mineral Raw Material Prices through a Feature-Based Clustering Approach †

Konstantinos Oikonomou * and Dimitris Damigos

School of Mining and Metallurgical Engineering, National Technical University of Athens, GR-15773 Athens, Greece; damigos@metal.ntua.gr

* Correspondence: ekonomouk@gmail.com

† Presented at International Conference on Raw Materials and Circular Economy, Athens, Greece, 5–9 September 2021.

Abstract: Mineral raw materials prices have been shown to be affected by macroeconomic factors such as aggregate demand and commodity-specific factors (e.g., supply shocks). In addition, it has been shown that certain mineral raw material prices co-move, meaning that they behave similarly during expansion and contraction phases of the international business cycles. In order to assess the behavior similarity of the prices of different mineral raw materials, we propose a method that utilizes extracted features of time series price data and unsupervised learning techniques to create clusters of price movements having similar long-term behavior.

Keywords: mineral raw material prices; time series; feature extraction; machine learning; principal component analysis; k-means clustering

check for updates

Citation: Oikonomou, K.; Damigos, D. Assessing Behavior Similarity of Mineral Raw Material Prices through a Feature-Based Clustering Approach. *Mater. Proc.* **2021**, *5*, 37. https://doi.org/10.3390/materproc2021005037

Academic Editor: Anthimos Xenidis

Published: 30 November 2021

1. Introduction

Mineral raw materials and their products play crucial roles in international trade due to their importance in sectors such as manufacturing, construction, transportation, etc. Theoretical and practical research efforts suggest that their prices are affected by macroeconomic factors such as aggregate demand and commodity-specific factors such as shocks in supply due to unforeseen events [1]. It has also been shown that commodity prices co-move, meaning that some mineral raw materials prices show some degree of similarity in their behavior [2].

Price data of mineral raw materials are typically studied in the form of time series, that is, data gathered and displayed over time. Time series data exhibit certain properties that can be described by certain measures. For instance, simple statistical measures such as mean and standard deviation can be used to describe the central tendency and the dispersion of the time series, respectively. Nanopoulos et al. [3] used skewness and kurtosis to describe the shape of the time series distribution. More advanced measures have been used to describe other properties. Wang et al. [4] used the Lyapunov exponent to quantify the chaotic behavior of a time series. Therefore, a collection of features can be considered as a partial representation of a time series itself.

Clustering in the context of unsupervised learning refers to the process of grouping together similar entities when prior information regarding classification is unknown. There are two main methods in clustering different time series. One involves directly measuring the similarity of different time series using some sort of metric such as Euclidian distance, and the other utilizes extracted features as an approximation of the time series itself. Feature-based clustering enables the acquisition of information regarding the underlying structure of the clustered time series.

The purpose of this study is to determine the long-term price behavior similarity of mineral raw material prices by utilizing unsupervised machine learning methods such as dimensionality reduction and clustering. This approach aims to provide insights into the common underlying dynamics and possible long-term dependencies of the mineral raw material prices.

2. Materials and Methods

2.1. Data

We obtained the time series data of monthly average prices from the World Bank and the Federal Reserve of St. Louis databases for selected mineral commodities. We included mineral raw materials that cover a wide range of applications—namely, crude oil, coal (Australian), aluminum, iron ore, copper, lead, tin, nickel, zinc, gold, platinum, silver, and uranium, and for which comprehensive time-series data were available for the period January 1990 till March 2021 (all values in nominal US dollars). Since the scaling, as well as the units of the commodities, are different, and this can affect the final clustering, and also because we were interested in clustering time series with similar behavior rather than similar value amplitudes, we rescaled each time series by applying min–max normalization, such that each time series ranges from 0 to 1 (Figure 1).

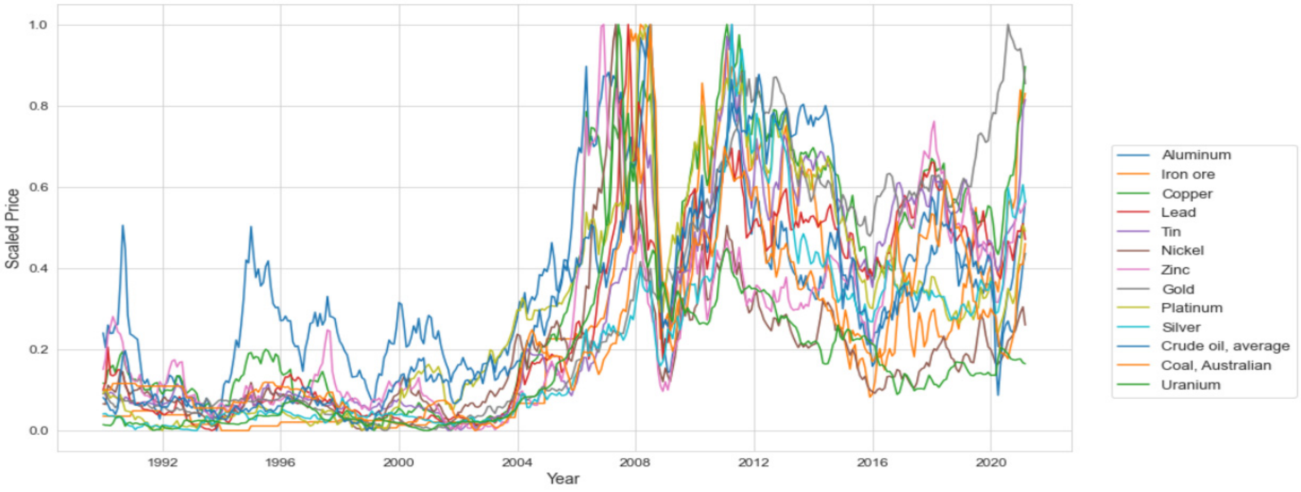


Figure 1. Min–max scaled prices for the different mineral commodities.

2.2. Feature Extraction and Selection

The process of feature extraction was performed using the *tsfresh* package available in the Python programming language. This module enables users to automatically extract hundreds of features for multiple time series. In this study, we chose to include measures that describe the central tendency, dispersion, and shape, as well as other measures such as the relative location of maximum and minimum values, the length of the longest period above and below mean value, the count of values above and below mean, and the free parameters (slope and intercept) of the linear least-squares fitting of the price data versus time. In total, 21 features were used to describe each time series. The description of the selected features can be found in Table S1 of the Supplementary Materials.

2.3. Multivariate Analysis

A common approach when dealing with high dimensional datasets comprising of possibly intercorrelated variables (in this study, 21 features) is to reduce the dimensionality. Principal component analysis (PCA) is a multivariate technique that transforms the initial variables into a new set of orthogonal variables, called principal components, which are linear combinations of the original variables.

Specifically, given a standardized $m \times n$ matrix $\mathbf{A}$, where m is the number of observations, and n is the number of variables, we calculate the variance–covariance matrix $\mathbf{C_A}$ as

$$\mathbf{C_A} = \frac{1}{(n-1)}\mathbf{A^T A}$$

$\mathbf{C_A}$ is a square symmetric matrix that can be decomposed into:

$$\mathbf{C_A} = \mathbf{V \Lambda V^T}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of sorted eigenvalues in descending order, and $\mathbf{V}$ is the matrix where columns are the corresponding orthogonal unit length eigenvectors.

The eigenvectors are called principal components (PCs) and are the new variables. Interpretation is carried out by examining the initial variable loadings on each PC, which are the correlations of the initial variables with the PCs.

PCs are obtained in such a way that the first component explains the largest proportion of the total variation, the second component explains the second largest proportion of the total variation, etc. Thus, by using the first few components, the dimensions of the dataset can be reduced while retaining the largest proportion of the total variance of the dataset.

2.4. K-Means Algorithm and Elbow Method

K-means is a popular clustering algorithm that has been used in many scientific areas [5,6]. It is an iterative algorithm that uses centroids (which can be considered as cluster prototypes) to partition observations in a multidimensional space. The K-means algorithm aims to choose centroid coordinates that minimize the sum of squares criterion, that is, the sum of squared error of all data points from their associated cluster centroid. Formally, it aims to minimize the objective function for all clusters K.

$$E = \sum_{i=0}^{K} \sum_{p \,\in\, Ci} dist\;(\boldsymbol{p}, \boldsymbol{c_i})^{\,2}$$

where E is the sum of the squared error of all data points, $\boldsymbol{p}$ is the point in space representing a given object, $\boldsymbol{c_i}$ is the centroid of the cluster, K is the number of clusters, and $dist(\boldsymbol{x}, \boldsymbol{y})$ is the Euclidian distance between the two points [7].

After randomly initializing the centroids, partitioning is carried out in three steps, as follows:

Step 1. Assign each data point to its closest cluster centroid;
Step 2. Compute new centroids as the mean value of all data points assigned to the previous cluster;
Step 3. Continue until centroids' positions remain unchanged.

K-means requires the number of clusters to be first specified. Many variants to assess the appropriate number of clusters have been proposed [8,9]. In this study, we used a rather heuristic approach, also informally known as the "elbow" method, that involves fitting different numbers of clusters into the dataset and selecting the most appropriate based on the cutoff point of some metric. In this study, the distortion score was employed, that is, the sum of squared error from each point to its corresponding centroid. The idea behind this method is that as the number of clusters increases, the clusters become rapidly refined, and at a certain K number of clusters, the refinement reaches a plateau. The selection of K was performed visually by plotting the values of K versus the distortion score.

3. Results and Discussion

As mentioned, 21 measures that approximate each price time series were utilized, and these measures were used to group together time series with similar features but also separate dissimilar ones. This section presents the results of the analysis.

3.1. Co-Movement of Prices and PCA

The most common measurement of co-movement between two variables is the Pearson correlation coefficient [10]. Table S2 of the Supplementary Materials presents the Pearson correlation between min–max scaled prices for the examined mineral raw material prices. Mineral commodities show different degrees of co-movement with others even though all commodities co-move.

Further, Tables S3 and S4 present the results of the PCA analysis on the 21 extracted features. Three major PCs were identified accounting for more than 80% of the variance of the original dataset. Specifically, the explained variance for the first three components was 47.8%, 23.4%, and 11.4%, respectively. In order to interpret each PC, we first associated the contribution of the original variables to them by examining their loadings. Identifying the process implied by each PC is not always simple and should be accomplished with caution [11].

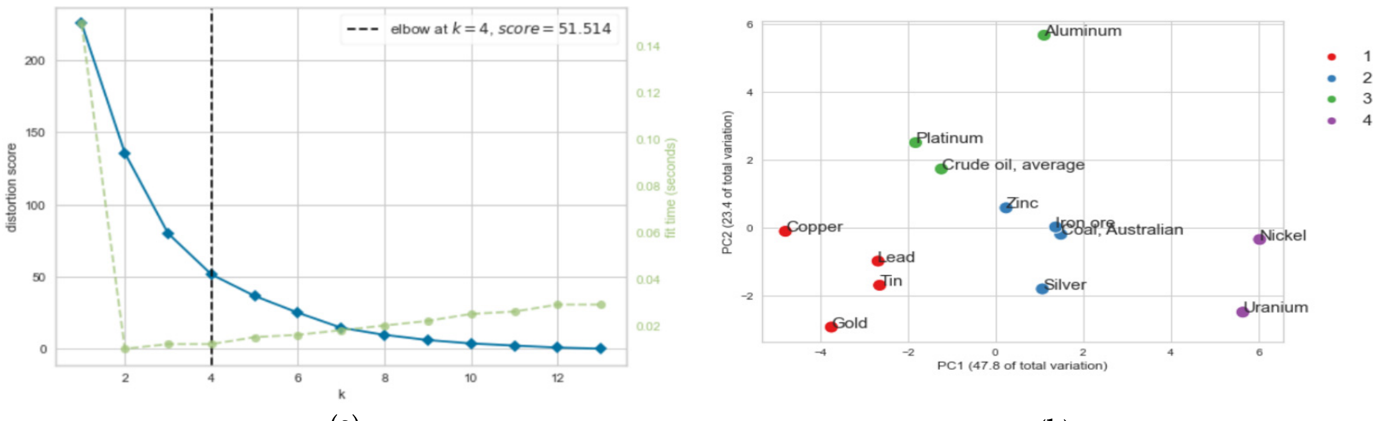
As seen in Table S3, PC1 is positively associated with skewness and kurtosis, as well as the ratio of values exceeding higher thresholds of standard deviations, and negatively associated with the slope of the linear trend and the standard deviation. Nickel price distribution scores high in PC1, meaning that its price is rather skewed, exhibiting relatively low variation, and is prone to contain outliers [12]. Copper price exhibits low skewness and kurtosis and has higher levels of the steepness of linear trend.

The second PC is positively associated with the median value and the intercept of the linear trend and negatively associated with the length of the longest period above and below the mean value. Therefore, the aluminum price has a higher central tendency and initial value, compared with the rest of the commodities, and the gold price for the examined period displayed the longest period both below and above its mean value.

The third PC is positively associated with the ratio beyond 1.5 sigma standard deviations and the count of values below mean and negatively associated with the count of values above mean.

3.2. Clustering

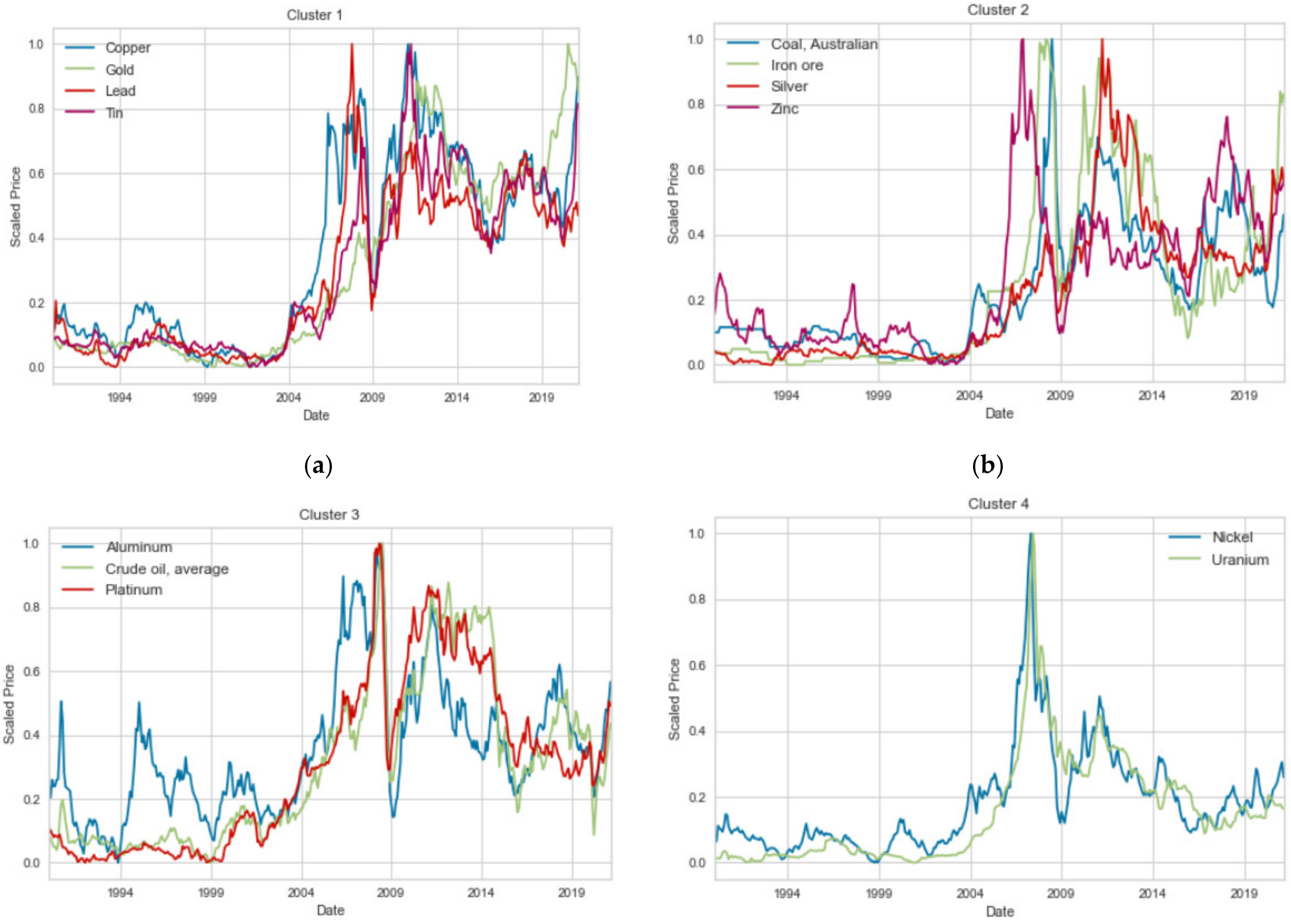
Figure 2a shows the results of the elbow method. The optimal number of clusters was identified to be four, having a distortion score of 51.51. The final obtained clusters can be seen in Figure 2b, where each commodity is plotted against the first two PCs. Figure 3 presents the final clusters of the min–max scaled time series, where each time series contained in each cluster is plotted together. Cluster 1 contains the prices of copper, lead, tin, and gold that score low in PC1. The prices of these commodities exhibit lower kurtosis and skewness, as well as a higher level of the steepness of their linear trend. Zinc, iron ore, coal, and silver prices constitute Cluster 2. The clustering in Cluster 2 is rather incoherent, possibly due to the fact that information regarding the time series properties is not well described by the selected features. Cluster 3 contains the prices of platinum, crude oil, and aluminum. This cluster scores high in PC2, meaning that the central tendency of the price distribution of the included raw materials is higher, compared with the rest of the commodities. The final Cluster 4 contains the prices of nickel and uranium. Nickel and uranium prices score high in PC1, which corresponds to time series of low variation, higher skewness, and being prone to contain outliers.



(**a**)

(**b**)

Figure 2. Distortion score for different numbers of k (**a**) and K-means clusters plotted against the first two PCs (**b**).



(**a**)

(**b**)

(**c**)

(**d**)

Figure 3. Final clusters of the min–max scaled time series: (**a**) Cluster 1 (copper, gold, lead, tin); (**b**) Cluster 2 (coal, iron ore, silver, zinc); (**c**) Cluster 3 (aluminum, crude oil, platinum); (**d**) Cluster 4 (nickel, uranium).

4. Conclusions

The extraction of 21 features from the co-moving commodity time series enabled us to identify four price movement clusters that exhibit similar long-term behavior. According to the findings, copper, lead, tin, and gold prices belonging to Cluster 1 exhibit low kurtosis and skewness and have higher levels of the steepness of linear trend. Aluminum, platinum, and crude oil prices making up Cluster 3 have a higher central tendency. Nickel and uranium prices constitute Cluster 4, which corresponds to higher skewness, higher kurtosis, lower variance, and outlier-containing distributions. Finally, Cluster 2 contains the prices of zinc, iron ore, coal, and silver. Nevertheless, as this is the first study attempting to determine the long-term price behavior similarity of mineral raw material prices by utilizing unsupervised machine learning methods, further research is needed before definitive conclusions can be drawn. A new study could be carried out to clarify whether the clustering is due to purely statistical properties of the time series, or whether there are macroeconomic and other underlying factors that influence each cluster. Additionally, the study could be expanded by introducing additional features toward describing the time series and test whether new clusters are created or by including non-mineral commodity prices and other indicators into the analysis.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/materproc2021005037/s1.

Author Contributions: Conceptualization, K.O. and D.D.; methodology, K.O.; software, K.O.; writing—original draft preparation, K.O.; writing—review and editing, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All commodity price data except Uranium were obtained from the "Pink Sheet" provided by the World Bank available at https://www.worldbank.org/en/research/commodity-markets. Uranium price data are available at the Federal Reserve Bank of St. Louis website at https://fred.stlouisfed.org/series/PURANUSDM. Data were accessed on April 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pindyck, R.S.; Rotemberg, J.J. The Excess Co-Movement of Commodity Prices. *Econ. J.* **1990**, *100*, 1173–1189. [CrossRef]
2. Chiaie, S.D.; Ferrara, L.; Giannone, D. *Common Factors of Commodity Prices*; Banque de France: Paris, France, 2017.
3. Nanopoulos, A.; Alcock, R.; Manolopoulos, Y. Feature-Based Classification of Time-Series Data. In *Information Processing and Technology*; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2001; pp. 49–61. ISBN 1-59033-116-8.
4. Wang, X.; Smith, K.; Hyndman, R. Characteristic-Based Clustering for Time Series Data. *Data Min. Knowl. Discov.* **2006**, *13*, 335–364. [CrossRef]
5. Carvalho, M.J.; Melo-Gonçalves, P.; Teixeira, J.C.; Rocha, A. Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation. *Phys. Chem. Earth Parts ABC* **2016**, *94*, 22–28. [CrossRef]
6. Javadi, S.; Hashemy, S.M.; Mohammadi, K.; Howard, K.W.F.; Neshat, A. Classification of aquifer vulnerability using K-means cluster analysis. *J. Hydrol.* **2017**, *549*, 27–37. [CrossRef]
7. Han, J.; Kamber, M.; Pei, J. 10-Cluster Analysis: Basic Concepts and Methods. In *Data Mining*, 3rd ed.; Han, J., Kamber, M., Pei, J., Eds.; The Morgan Kaufmann Series in Data Management Systems; Morgan Kaufmann: Boston, MA, USA, 2012; pp. 443–495. ISBN 978-0-12-381479-1.
8. Goutte, C.; Hansen, L.K.; Liptrot, M.G.; Rostrup, E. Feature-space clustering for fMRI meta-analysis. *Hum. Brain Mapp.* **2001**, *13*, 165–183. [CrossRef] [PubMed]
9. Pelleg, D.; Moore, A.W. X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2000; pp. 727–734.
10. Rua, A. Measuring comovement in the time–frequency space. *J. Macroecon.* **2010**, *32*, 685–691. [CrossRef]
11. Brown, C.E. Principal Components. In *Applied Multivariate Statistics in Geohydrology and Related Sciences*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 103–111. ISBN 978-3-642-80328-4.
12. Westfall, P.H. Kurtosis as Peakedness, 1905–2014. *R.I.P. Am. Stat.* **2014**, *68*, 191–195. [CrossRef] [PubMed]