Functional cluster regression for commodities and the representatives of stock indexes

Gabriela Hlásková[1], Tomáš Mrkvička[2]

DOI: 10.32725/978-80-7394-976-1.33

Abstract: While using external variables as potential predictors, one might be challenged by numerous possible variables, which while used once-on-time might devalue the predictive ability of individual ones. Thus, the pre-selection of relevant possible predictors should be used. For purpose of risk prediction of exchange rate changes, many external variables (time series) are available, thus most commonly traded ones were selected in the final number of 32 variables. Only a fraction of those should be in fact used while proceeding with the computation. Thus, out of many possible methods of selection of finding the relevant variables, the functional cluster analysis would be used. In this paper, we describe a case study of the functional cluster analysis application on time series as one of the possible methods of explanatory variable selection for the exchange rates.

Keywords: predictors, commodities, functional cluster analysis, stock indexes exchange rate
JEL Classification: F31, F32, F37, F47, F39

1 Introduction

For the needs of regression methods based on the use of external variables - potential predictors, their choice is a crucial moment that can be responsible for the success or failure of the calculated model. The authors use various economic and commodity variables for their prediction models of exchange rates. For example, Lee-Leea Hui-Boon examined the influence of the money supply, real GDP, short-term and long-term interest rates, CPI, and current account balance on the volatility of the exchange rate in the short and long term for the currencies of Southeast Asian countries through the GARCH model, VAR, and VECM models. (Lee-Leea Hui-Boon, 2007) The same models, i.e. VAR and VECM, were used by Englam et al. (2010) to detect the effect of oil price, foreign exchange reserves, and interest rate on exchange rate volatility in the short and long run. Using the multivariate GARCH model in Latin American economies, Grydaki and Fontas described the effect of money supply inflation, and the openness of the economy when taking into account the exchange rate regime (Grydaki and Fontas, 2011). Stančík focused on the volatility of the exchange rate, as an indicator of the stability of the economy (necessary for the adoption of the EURO), primarily on the influence of the openness of the economy, the exchange rate regime, and the "news" factor, while he used the TARCH model to model this influence. (Stančík, 2007) Pearson's correlation analysis was used by Mirchandani to investigate the effect of CPI, interest rate, GDP, foreign direct investment, and current account balance on the volatility of the Indian Rupee. (Mirchandani 2013) Hasaan et al. (2017) using ARCH, ARDL, and Granger causality models demonstrated a significant effect of interest rates and net foreign assets on exchange rate volatility, while the effect of economic openness, oil price, and fiscal policy has no significant effect on exchange rate volatility in Nigeria. The scope of the research of these works indicates a strong diversity of the usability of external variables when there is no established approach to their use. Although the monitored indicators have a significant meaning, only in a certain region or a certain period, others can provide additional information of a global nature. The identification of such variables is an important prerequisite for the success of the model.

For the selection of currency exchange rate predictors, the development of which we follow in a broader context, we have at our disposal a really wide set of potential variables, from commodity markets, and their aggregate index values, to the values of stocks and stock indexes, the use of crypto-assets, macroeconomic statistical data, etc. Only a fraction of these variables (whose time series we use), however, is realistically usable in the following methods, although a number of variables will bear the same or very similar characteristics of historical development, trend, volatility, or correlation.

[1] University of South Bohemia in České Budějovice, Faculty of Economics, Studentská 13,370 05 České Budějovice, Czech Republic, hlaskg00@ef.jcu.cz.
[2] University of South Bohemia in České Budějovice, Faculty of Economics, Studentská 13, 370 05 České Budějovice, Czech Republic, mrkvicka@ef.jcu.cz.

For the pre-selection of variables, we can use various methods such as stepwise backward regression, optimal AIC or BIC criterium, or the mentioned cluster analysis that is intended to identify homogeneous subgroups among observations.

In this contribution, we want to describe the usage of the last-named one and the assessment of its suitability for variable selection. The selection of these variables will subsequently enter as a selection of predictors for predicting changes in exchange rates using Value at Risk methods, extended by quantile regression methods.

2 Methods

Description of the Data

There are 32 pre-selected variables available for our experiment. Those variables (time series of length from December 2008 till February 2022) are provided by Trading Economics. The data are on daily basis, we declare to use daily opening value for each variable. Representatives of both, commodities and stock indexes were chosen for purposes of the functional cluster modelling. Fourteen commodities, two commodity indexes, fourteen stock indexes, and two energetics indexes were particularly chosen. Thus, our goal is to obtain groups of variables that behave similarly with respect to time evolution. Functional clustering analysis enables us to combine functional sources of different kinds. However, it is not in the scope of this paper to describe the method of clustering analysis precisely. A more in-depth functional clustering method is given by Dai et al. (2021) however, we choose this method primarily because it allows the combining of several different views of a function so that these views are equally represented in the distance construct. It, therefore, allows combining several different distances (e.g. volatility, trend, etc.) into one composite distance in which all components are represented equally.

Description of the Method and evaluation

While selecting external variables for the needs of regression models, we chose functional cluster analysis. Alternative methods could also be used for a similar selection, for example, factor analysis and PCA methods, but these create latent variables that are poorly interpreted, so for our needs, it is more appropriate to use individual variables, therefore we will use cluster analysis.

Providing the functional clustering method, the Myllymäki & Mrkvička's (2017, 2020) GET package (Global Envelops) of R studio is used. The *fclustering* function of the GET package is used together with the smoothed curve, the number of clusters given in chapter 3.1, and the "St" type parameter, which specifies the functional measure used to compute the functional distances.

According to the variety of uses of the cluster method, it is necessary to set suitable parameters and an evaluation metric for the suitability of various combinations of these parameters. Such a metric will become the Average Silhouette Width (ASW). (Batool & Hennig, 2021)

The average Silhouette Width of clustering φ is

$$\bar{S}(\varphi, \mathrm{d}) = \frac{1}{n}\sum_{i=1}^{n} S_i\,(\varphi, \mathrm{d}) \tag{1}$$

where the S_i is the silhouette value of datapoint *(i)*

$$S_i = \frac{b_{(i)} - a_{(i)}}{\max(a_{(i)}, b_{(i)})} \tag{2}$$

where *a(i)* is the average distance of x_i to points in the cluster to which it was assigned,

and *b(i)* is the average distance of x_i to the points in the nearest cluster to which it was not assigned.

S_i takes on the values [-1;1]. If the data are appropriately clustered, the S_i value is close to 1, otherwise, the Silhouette value close to -1 indicates that data would be preferable if clustered in a neighboring cluster. Variables on the boundary of two natural clusters are represented with S_i close to zero.

Since clusters are intended to be homogeneous and well-separated, thus larger values indicate better clustering quality. Therefore, the optimal clustering (if different values are compared) is considered to be the one where the ASW values are maximized.

The variety of parameter settings in our case will mainly be in the setting of the "smoothing" parameter and the expected number of clusters. The first one represents the number of days taking into account the trend of the model. Due to the reflection of the exchange rates on the mentioned variables and the observed goal of the medium-term behavior of the explained and explanatory variables, we move to the limit of 2-4 weeks, i.e. the tested values move to the range of 2, 3 and 4 working weeks). The second variable parameter is the expected number of determined clusters. Due to the number of tested variables (32 variables) and the ideal number of variables entering the regression model (10 variables), we set the number of clusters to be tested at 6-10.

3 Research results

In the results chapter, the first part offers the results of the selection of the "smoothing" parameter for the length of the chosen trend period and the number of selected and monitored clusters. The second part is focused on the clusters themselves and their separate performance.

3.1 Setting parameters of cluster analysis - smoothing and number of clusters

The performance results of the cluster analysis for various combinations of these parameters are offered in the following Table 1. It is noticeable that a higher number of clusters and at the same time a longer period of the monitored trend also offer a higher ASW value and thus also the accuracy of the cluster analysis. At the border of the 9th and 10th clusters, the accuracy value breaks down, and we, therefore, evaluate the number of 9 clusters and the length of the trend of the monitored period as an ideal combination of four working weeks.

Table 1: ASW results for combinations of smoothing parameter and number of clusters

Smoothing par./ number of clusters	6	7	8	9	10
2 weeks	0,1541796	0,1589222	0,1870644	0,2083123	0,2040698
3 weeks	0,1568741	0,167891	0,1876303	0,2089254	0,2039036
4 weeks	0,1586368	0,1691971	0,1882855	**0,2089787**	0,2038783

3.2 Results of functional cluster analysis

Using the set parameters to determine the observed trend and the defined number of clusters, the performance of the cluster analysis itself offers us the following results.

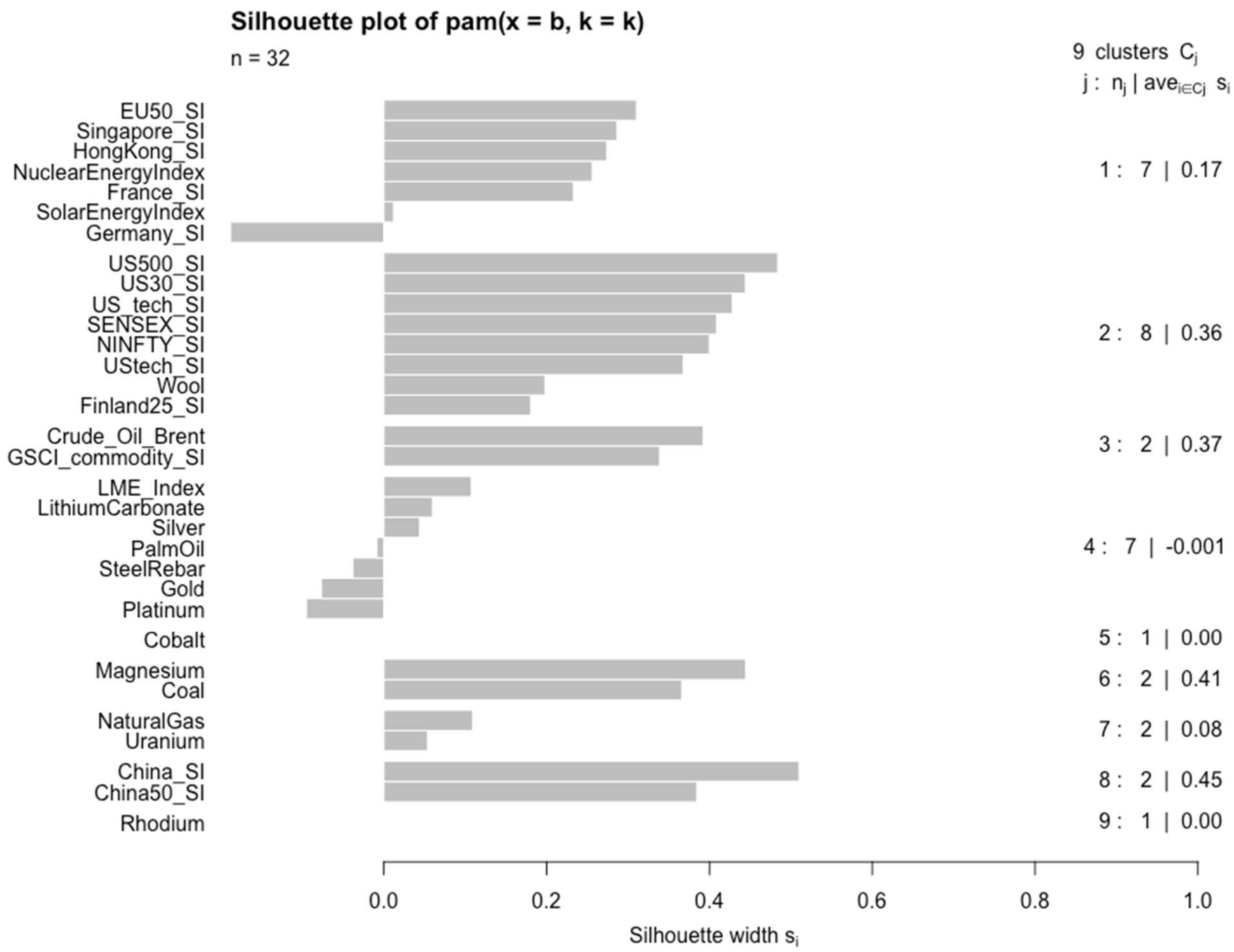
The monitored 32 variables are divided into 9 clusters. This distribution can be seen in the following Table 2. In Figure 1 we also see a detailed distribution and ASW value for each of the observed variables as the same as the closest neighbor of the variable.

Table 2: Distribution of variables through clusters

Cluster	Name variable	Neigbor	Cluster	Name variable	Neigbor
1	EU 50_SI	2	4	LME_Index	1
	Singapore_SI	2		LithiumCarbonate	6
	HongKong_SI	2		Silver	3
	NuclearEnergyIndex	2		PalmOil	1
	France_SI	2		SteelRebar	6
	SolarEnergyIndex	2		Gold	1
	Germany_SI	2		Platinum	3
2	US500_SI	1	5	Cobalt	2
	US30_SI	1	6	Magnesium	4
	US_tech_SI	1		Coal	4
	SENSEX_SI	1	7	NaturalGas	6
	NINGTY_SI	1		Uranium	4
	UStech_SI	1	8	China_SI	1
	Wool	1		China50_SI	1
	Finland25_SI	1	9	Rhodium	2
3	Crude_Oil_Brent	4			
	GSCI_commodity_SI	1			

Source: Own processing

Figure 1: Silhouette plot of the clusters



Source: Own processing

As mentioned in the previous chapter S_i takes on values [-1;1]. The higher the values, the greater the "encapsulation" of the function inside the cluster. Otherwise, the lower (negative) value of S_i, the less it belongs to the given cluster and the greater its probability of leaving the cluster.

The first cluster represents variables with a gradual but noticeable increasing trend combined with specifically three noticeable contortions in the trend curves. Its representatives are EURO STOXX 50[3], Singapore and Hong Kong Stock Indexes, France and German Stock indexes, and both Nuclear and Solar energetical indexes.

The second cluster is represented by an even stronger increase in trend lines, as its members are US stock indexes (US 500, US 30, and both US tech indexes), both Indian stock indexes (SENSEX and NIFTY), and surprisingly the prices of Wool commodity and Finland stock index (HEX 25). However, as seen in Figure 12 and in Table 2 both last named are not the strongest members of the second cluster.

The third cluster, however, contains two variables that perform strongly similarly. The prices of Crude oil (traded in the BRENT system) and the GSCI commodity index[4]. Both variables declare strong dependence and correlation, as the prices of crude oil might be a significant part of the GSCI commodity index itself.

The fourth group might be called the commodity group, as it contains the LME Index[5], Lithium Carbonate, Silver, Steel Rebar, Gold, and Platinum are the metal representatives supplemented with the Palm Oil variable. The cluster itself

[3] It includes the 50 most important companies of the Eurozone. Those are so-called blue-chip companies considered leaders in their respective sectors.
[4] Goldman Sachs Commodity Index actually contains twenty-four different commodities over various sectors (energy, industrial commodities, precious metals, agriculture, or livestock products).
[5] London Metal index (six different metals aluminum, copper, lead, nickel, tin, and zinc).

is represented by a very strong rise at the beginning of the period (2008), followed by a downward trend and gaining strength exceeding historical values at the end of the observed period.

The fifth cluster is defined by just one variable, which is Cobalt. The cluster is characterized by a strong peak followed by a long period of gradual decline. Towards the end of the observed period, there is again a significantly increasing behavior of the trend curve.

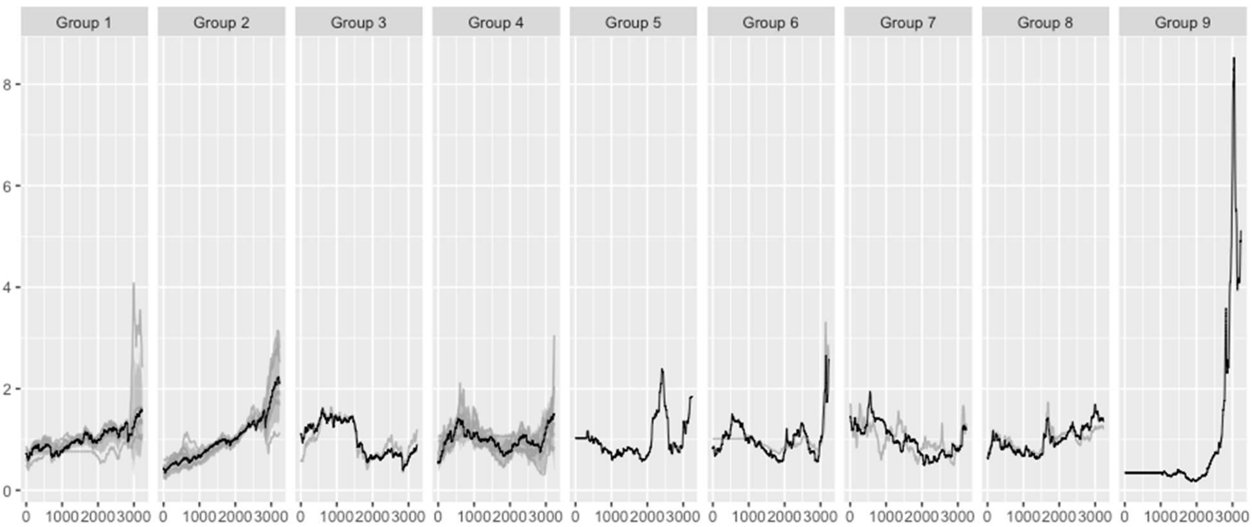
The sixth cluster is defined by both the prices of Magnesium and Coal. Thus, the cluster is specific with major peaks and an exceptionally high increase towards the end of the period, in the order of multiples of past values.

Both members of the seventh cluster are major commodities of the energetical sector – Uranium and Natural Gas. As is seen in both Table 2 and Figure 2, the trend is a long-term decline, hitting its bottom towards the end of the trend period, followed by a rapid rise.

The Eight group is represented by the two Chinese stock indexes (SHANGHAI 50 and China Shanghai Composite Stock Market Index). Both variables perform exactly the same and slightly follow another Chinese representative in the first cluster, as this one is the closest neighbor (as seen in Table 2).

Finally, the ninth cluster is again represented by a single variable – the Rhodium. Its extraordinary raising trend is out of measurements of another cluster. The value of the variable follows usability in industrial use.

Figure 2: Trend development of clusters and ranked variables



Source: Own processing

Clusters defined in this way, or their representatives cover the portfolio of monitored variables and as individual attributes can enter further analyses, for example, the aforementioned quantile regression, which is the next phase of the author's research. For the purpose of further use, the strongest representative from each cluster would be used, i.e. the variable with the highest Silhouette value, i.e. the so-called medoid.

4 Conclusions

The search for significant links between exchange rates and external variables - potential predictions - is part of the fundamental analysis of exchange rates. However, unlike other predictable quantities, the number of variables potentially acting as predictors in the case of exchange rates is considerable, and their mutual correlation or substitutability is hypothetically quite fundamental. The cluster analysis method is one of the possible approaches to the selection of mutually replaceable variables and the final selection of the most relevant groups, which can then enter the final prediction models.

The above-described application of the functional cluster analysis aimed to identify separate clusters from a group of 32 randomly selected representatives of commodities and stock markets groups and then identify their most typical representatives. Cluster analyses are then run on the basis of the trend behavior of sub-variables.

As described in chapter 3. Research results, the ideal distribution represents 9 separate clusters whose representatives show very similar trend behavior. In other words, the representatives of these clusters will perform more or less ambivalently, so they are mutually replaceable. It is certainly promising information that the offered clusters also connect variables cooperating on a "technical level", e.g. different stock market indices listed by the same stock exchanges listed in the same cluster, or the cluster of significantly linked precious metal commodities and commodity indexes, etc. This fact then confirms the credibility of cluster analysis even for supporters of technical analysis. Out of these clusters, it is possible to identify the so-called medoids, i.s. the strongest representatives of the given cluster, and use these representatives for the needs of further analysis.

The logical next steps would be the application of cluster analysis to a larger portfolio of variables and a test of the quality and benefit of the predictive ability when applied directly to currency rate prediction models.

Acknowledgement

This paper was prepared in the project of the Grant Agency of the University of South Bohemia under the number

GA JU 016/2021/S.

References

Batool, F., & Hennig, C. (2021). Clustering with the Average Silhouette Width. *Computational Statistics & Data Analysis*, *158*. DOI: 10.1016/j.csda.2021.107190

Dai, W., Athanasiadis, S., & Mrkvička, T. (2022). A New Functional Clustering Method with Combined Dissimilarity Sources and Graphical Interpretation. In R. López-Ruiz (Ed.), *Computational Statistics and Applications*. IntechOpen. DOI: 10.5772/intechopen.100124

Englama, A., Duke, O. O., Ogundipe, T. S., & Ismail, F. U. (2010). Oil price and Ex- change Rate Volatility in Nigeria: An Empirical Investigation. *CBN Economic and Financial Review*, *48*(3), 31-48.

Grydaki, M., & Fontas, S. (2011). What Explains Nominal Exchange Rate Volatility? Evidence from the Latin American Countries. *Discussion Paper*, *2010*(10).

Hassan, A., Abubakar, M. 'ilu, & Dantama, Y. U. (2017). Determinants of Exchange Rate Volatility: New Estimates from Nigeria. *Eastern Journal Of Economics And Finance*, *3*(1), 1-12. DOI: 10.20448/809.3.1.1.12

Lee-Lee, C., & Hui-Boon, T. (2007). Macroeconomic factors of exchange rate volatility [Online]. *Studies In Economics And Finance*, *24*(4), 266-285. DOI: 10.1108/10867370710831828

Myllymäki, M., & Mrkvička, T. (2020). GET: Global envelopes in R. *arXiv:1911.06583 [stat.ME]*.

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., & Hahn, U. (2017). Global envelope tests for spatial processes [Online]. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)*, *79*(2), 381-404. DOI: 10.1111/rssb.12172

Mirchandani, A. (2013). Analysis of Macroeconomic Determinants of Exchange Rate Volatility in India. *International Journal Of Economics And Financial Issues*, *3*(1), 172-179.

Stančík, J. (2007). Determinants of Exchange-Rate Volatility: The Case ofthe New EU Members. *Czech Journal Of Economics And Finance (Finance A Uver)*, (57), 414-432.

Trading Economics. (2022). Retrieved September 22, 2022, from https://tradingeconomics.com/