4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15080

The demand side of information provision: Using multivariate time series clustering to construct multinational uncertainty proxies

Florian Schütze
Department of Socioeconomics, University of Hamburg, Germany

Abstract

Information demand in the modern world is met to a huge extent by information supply from the search engine Google. Humans use the search engine to gather information which shall help to reduce perceived personal uncertainty about a specific subject. Google Trends is providing insights into this information demand in a timely manner and for a variety of different countries. In this paper, multinational Google Trends data and unsupervised learning techniques are used to construct meaningful country clusters resembling the economic, geographic and political relationships of the considered countries. Additionally, these clusters are stable over time. Under the assumption that an increase in Google search requests reflect elevated uncertainty, the cluster information is used to construct economic and political uncertainty time series for 43 different countries. This uncertainty index Granger causes quarterly GDP growth in more countries compared to an existing multinational uncertainty index proofing its usefulness in the field of forecasting. Furthermore, the new index is available up to a daily frequency and can be applied to additional countries and regions.

Keywords: *Google Trends; Economic Uncertainty; Unsupervised machine learning; Forecasting; Clustering.*

1. Introduction

Economic and political uncertainty can be inferred in diverse ways. For example, by measuring the volatility of macroeconomic variables (Bloom, 2009; Jurado et al., 2015), the dispersion among forecasters (Bachmann et al., 2013) or counting the occurrence of uncertainty related keywords, like the Economic Policy Uncertainty index by Baker et al. (2016) or the World Uncertainty Index by Ahir et al. (2022). A different strand of uncertainty measurement lies in Google Trends data. In contrast to the previous methods, Google allows to measure uncertainty among the general population instead of measuring uncertainty in journalist or forecasters. These Google Trends uncertainty measurements have an influence on real economic variables like investment, consumption, industrial production, and stock market returns (Bontempi et al., 2021; Castelnuovo and Tran, 2017). The underlying assumption is that people feeling uncertain about a subject turn to Google and search for the subject to reduce said uncertainty. Therefore, a higher search request reflects elevated uncertainty.

The mentioned studies used Google Trends keywords, which are prone to language selection. In this paper, the approach uses topics instead of keywords. While the English keyword "economy" only reflects search request which contains the English word "economy", the Google Trend topic "economy" covers keywords like "economic" or "economical" and terms in different language, for example the German word "Wirtschaft". This makes this approach very applicable in a multinational context. Kupfer and Zorn (2019) demonstrated that uncertainty proxies constructed using Google Trends topics have an influence on economic activities in European countries.

The contribution of this paper is as follows: firstly, giving insights in the diverse demand side of information provision across the globe with a cluster analysis. Secondly, showing that these clusters are stable over time. And thirdly, that these insights into the demand side of information can be used to form a timely uncertainty index, which outperforms the uncertainty index by Ahir et al. (2022) when it comes to forecast performance.

The rest of the paper has the following structure: The next chapter gives an insight to the used data and the construction approach of the country clustering. In the third chapter the country clusters of the second chapter are used to form country specific uncertainty measurements to compare them to the uncertainty proxy by Ahir et al. (2022). The last chapter concludes.

2. Multivariate time series clustering

In this section the data collection and clustering approach is explained and subsequently the result of the clustering is shown.

2.1. Data

The data for the approach of this paper stems from Google Trends. Google Trends allows for various search requests, for example the coverage of keywords or topics for different country, for different regions and for a certain time span, up to daily data. The main advantage of topics compared to keywords lies in its robustness against word selection and in its applicability in a multilanguage framework.

The complete set of data contains 109 Google Trends topics for 43 different countries spanning monthly from 01/2004 (the earliest date possible with Google Trends) until 02/2022. The 109 Google Trends topics are based on 184 uncertainty keywords by Bontempi et al. (2021) and by Baker et al. (2016) which are available only in English and Italian. For example, two of the keywords are “taxation” and “taxed”. Both were inserted in the Google Trends interface and the primary suggestion by Google for the underlying topic “tax”, therefore leading to the topic “tax” being among the final 109 different Google Trends topics. This procedure was then repeated for all uncertainty keywords. The R package "gTrendsR" was used to download all topics for all countries. Additionally, the 43 countries are chosen because the complete set of 109 topics exists for each country. The names of all used countries can be seen in figure 1 in the next chapter.

2.2. Cluster construction and optimal number of clusters

A hierarchical clustering procedure was applied to the Google Trends data to obtain country clusters of similar information demand. It is assumed that the entire world can be seen as one major information demand cluster with subclusters regarding to economic, political, geographical and/or historical ties. Hierarchical clustering is used in economics for example when it comes to clustering of countries with similar tax burden (Simkova, 2015).

The similarities between the different time series were identified by using Dynamic Time Warping. In contrast to the Euclidean distance, which compares pairs of datapoints directly, Dynamic Time Warping calculates the smallest distance between all datapoints. Therefore, it allows for possible "leads" and "lags" in the data, which could be important because there might exist “Google search spillover effects” from one country to another. While dynamic time warping stems from the area of speech recognition it is slowly also applied in economic, for example to predict recessions (Raihan, 2017).

The clusters are calculated by using agglomerative Ward´s method (Miyamoto et al., 2015), starting with single clusters for each of the 43 countries. These single country clusters are then merged based on minimum within-cluster variance gain leading in the end to a single cluster containing all countries. Therefore, this approach minimizes the intra-cluster variance.

All time series are Z-Score normalized before used in the clustering process. For the clustering procedure the R-package “dtwclust” was used.

While clustering will always result in different cluster sizes it is paramount to identify the optimal number of clusters which fits the data best. For this purpose, two internal evaluation metrics are used, primarily because in contrast to an external evaluation metric no assumption about cluster size and distribution must be made. The first metric is the Silhouette index. It ranges from -1 to 1, measuring the standardized averaged distance from all points within a cluster A to the next cluster B. Here, zero means a poor fit with a lot of overlapping clusters, whereas a value of one means a perfect fit with no overlapping clusters. Therefore, a higher Silhouette index reflects a better fitting clustering. The second metric is the Davies-Bouldin index. This index measures the ratio of the within cluster separation to the between cluster separation. A lower index can reflect to things: Firstly, that the within cluster separation is better, meaning that the data is more compact within a cluster. Secondly, that the separation between clusters is better, meaning that the cluster are not overlapping.

2.3. Clustering results

In figure 1 the cluster dendrogram of the whole dataset can be seen. The nearer the countries are clustered together the bigger the similarities in information search behaviors using Google among these countries. Potential clusters can be formed wherever the dendrogram splits into subclusters.

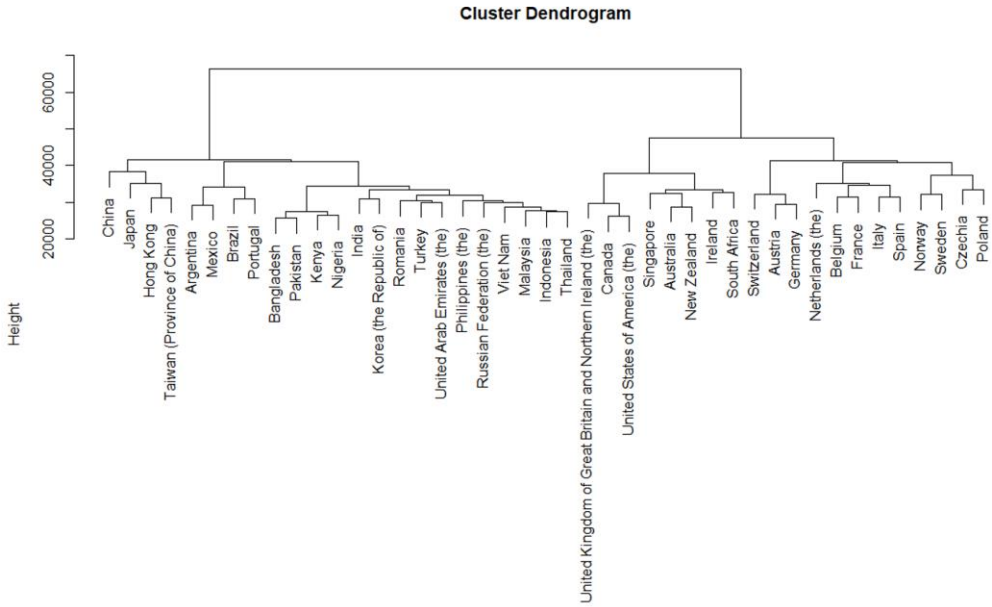

Figure 1: Clustering of Countries based on Google Trends topics data; complete set. Source: own Calculation

Starting at the top there exists a twofold split, resulting in two subclusters. The left-hand side cluster contains (mostly) the emerging economies of the world while the right-hand side

consists only of emerged economies. This alone demonstrates the real-world application of the clustering approach, because the two major branches of the dendrogram are not based on randomness but on similar economic structures resulting in similar information seeking behavior with regards to Google internet searches.

When going to a lower cluster region (or sub-clusters), the splitting is either according to geographical or economical/political reasons. Within the emerging economies cluster on the left side there exists an Asian sub-cluster including China, Japan, Hong Kong and Taiwan. The sub-clustering bordering the Asian one to the right can be interpreted as a South American cluster. The last emerging economies sub-cluster is more based on economic ties instead of geographic ones. In the case of the emerged economies cluster the split is again based on either geographical or economical/political affiliation. The left-hand side split can be interpreted as an "Anglosphere" sub-cluster consisting of mainly English-speaking countries. The right-hand side contains purely (central) European countries.

While it is possible to cluster the countries according to each branching in the dendrogram the optimal fitting number of clusters is based on objective internal evaluation metrics mentioned above. The optimal cluster number is five and was chosen based on the combination of a high silhouette and a low Davies-Boulding index (Silhouette 0.0459, Davies-Boulding index 1.6261) compared to a lower or higher number of clusters.

Table 1. The resulting five clusters with their respective countries

Cluster	**Countries**				
Cluster 1 “Emerging Economies”	Argentina	Bangladesh	Brazil	India	Indonesia
	Korea	Nigeria	Portugal	Thailand	Viet nam
	Malaysia	Pakistan	Romania	United Arab Emirates	Kenya
	Mexico	Phillippines	Russia		Turkey
Cluster 2 “Anglosphere”	Australia	Canada	Ireland	New Zealand	Singapore
	South Africa	United Kingdom	United States		
Cluster 3 “German speaking”	Austria	Germany	Switzerland		
Cluster 4 “Europe”	Belgium	Czechia	France	Italy	Netherlands
	Poland	Spain	Sweden	Norway	
Cluster 5 “Asia”	China	Hong Kong	Japan	Taiwan	

Source: Own calculation.

In table 1 the membership of countries regarding the five clusters are shown. The first cluster can be interpreted as a cluster of mostly emerging countries. The second cluster is an "Anglosphere" cluster. The third and fourth clusters are a "German speaking" and a "European" cluster, respectively. The last cluster is an "Asian" cluster. To sum up, all clusters reflect the connectedness of different countries, either due to geographical, political or economic links or a mixture out of these.

To validate if the cluster results are stable over time, the whole time span (01/2004-02/2022) was cut in half and the clustering has been applied to both subsamples. For the first 9 years the optimal number of clusters is four and the major difference compared to the complete set is the "German speaking" cluster merges with the "European" cluster. When looking at the last nine years the optimal number of clusters is back to five and the "German speaking" cluster is expanded by Czechia and Poland, two non-German speaking countries but with a close distance to Germany. The other clusters stay for the most part the same.

To sum this chapter up the approach using a hierarchical clustering procedure on multinational Google Trends topics data leads to meaningful country clusters being in line with political and geographical proximities. Furthermore, these clusters are stable over time except for the "German" clusters showing up only in the second half of the time span. With these results at hand the next step will be to research if the Google Trend queries of countries within certain subclusters can be used in an economic application context, i.e. as an uncertainty measurement of said countries.

3. Construction of country specific uncertainty indices

This chapter describes how uncertainty proxies using Google Trends topics are constructed and how they perform again an already existing uncertainty proxy.

3.1. Construction based on topic clusters

To construct uncertainty measurements for each country the next step is to identify the optimal number of topic clusters within a respective country. For this task, the 109 Google Trend topics are averaged over all countries within the corresponding cluster. The routine described in the previous chapter is then applied to this new dataset to calculate how many optimal topic clusters exists within each of the five country clusters.

In table 2 the optimal number of topic clusters is stated for all five country clusters. The optimal number is two except for the case of the "Emerging Countries" cluster where the optimal number is four, since this cluster is more diverse than the others when it comes to geography or political affiliation. When looking at the content of the two subclusters for each country, one cluster leans more to theme "economy" while the other cluster is more driven by "politics". For example, in the "Anglosphere" case the first cluster consists of 79 topics ("Tax", "Trade War", "Income tax" etc.) while the second cluster has 30 topics ("Economy", "Business", "Central Bank" etc.).

Table 2. Optimal number of topic clusters within the country clusters

	No. Of optimal topic subclusters	Distribution of topics	Silhouette	Davies-Bouldin index
Cluster 1 (”Emerging Countries”)	4	16-50-22-21	0.19	1.3
Cluster 2 (”Anglosphere”)	2	79-30	0.19	0.9
Cluster 3 (”German speaking”)	2	78-31	0.19	0.8
Cluster 4 (”Europe”)	2	55-54	0.21	1
Cluster 5 (”Asia”)	2	93-16	0.17	1

Source: Own calculation.

3.2. Comparison to the World Uncertainty Index

Published by Ahir et al. (2022) there exists a World Uncertainty Index (WUI) for 143 different countries using the Economist Intelligence Unit reports. The index is constructed counting the word "uncertainty" in the respective report for a certain country and a given time. This uncertainty measurement exists on a quarterly base which is a major disadvantage when it comes to the timely identification of elevated uncertainty regarding time.

To compare the GCUs with the WUI the frequency of the GCU must be adapted, because the GCUs exist on a monthly frequency. For comparing the Google Trends data to the WUI the months for the respective quarters where averaged, for example the first quarter of 2004 consist of the average of the first three month of the year 2004 to keep the informational content as high as possible.

An uncertainty measurement is identified as superior regarding forecast performance if it can significantly (α=0.05**)** Granger causes quarterly GDP growths in more countries. This was evaluated in a VAR approach using the Toda and Yamamoto (1995) procedure for Granger causality. The data for the quarterly GDP growth stems from OECD (2022), is available for 31 countries and the time span is from 01/2004 to 4/2021. All time series were seasonal adjusted and made stationary prior to the procedure. The optimal lag length for each national VAR was evaluated using the AIC.

Table 3. How often does … Granger causes the national quarterly GDP (α=0.05)

	WUI	GCU1	GCU2	GCU3	GCU4
Cluster 1 (”Emerging Countries”)	2	2	0	7	2
Cluster 2 (”Anglosphere”)	4	3	0	-	-
Cluster 3 (”German speaking”)	1	2	0	-	-
Cluster 4 (”Europe”)	3	3	1	-	-
Cluster 5 (”Asia”)	2	2	0	-	-
Total	12	12	1	7	2

Source: Own calculation.

In table 3 the results of the Granger causality procedure are shown. For all 31 considered countries the WUI Granger cause the quarterly GDP twelve times. This is comparable to the GCU performance being also twelve for the first cluster. When using the third cluster for the "Emerging Countries" instead of the first cluster the GCU Granger causes the GDP in 17 countries, which is a distinctly better result compared to the WUI.

To sum this chapter up, it was shown that the constructed multinational Google Cluster Uncertainty indices do Granger cause quarterly national GDP growth in different countries, implying valuable forecast characteristics. The constructed indices perform better doing so in comparison to an existing uncertainty index, the WUI. Furthermore, while the WUI is only available on a quarterly base, the GCU is also available on a monthly base (and even up to a daily base) making it more useful in short term forecasting.

4. Conclusion

In this paper multinational Google Trends search queries were used to show that meaningful country clusters can be formed. These clusters are stable over time and overlap with economic, geographic or political affiliation. These clusters can be used to identify relevant topics in different countries leading to a deeper understanding of the distribution of information demand around the world.

Topics within the country clusters were then clustered to construct Google Cluster Uncertainty indices for 31 different countries. On average, these indices perform better than an already existing uncertainty measurement regarding forecast ability of GDP growth.

The main advantage of the used procedure lies in its applicability and real time availability. Until now, only 43 countries are considered, but it can be applied to almost all countries when there is an a-priori decision to which cluster a new country belongs. Then, there is no need for the complete set of 109 different topics, but only for the topics in the cluster which are needed to construct the GCU. Furthermore, unlike already existing keyword-based Google

Trends uncertainty indices, which are prone to language selection, the usage of topics offers an easy multinational application.

Furthermore, the Google Trends data can be obtained even on a daily base and for subregions within countries, opening a variety of application for future research and practical forecasting considerations.

References

Ahir, H., Bloom, N., & Furceri, D. (2022). The world uncertainty index. NBER Working Paper 29763.

Bachmann, R., Elstner, S., & Sims, E. R. (2013). Uncertainty and economic activity: Evidence from business survey data. American Economic Journal: Macroeconomics, 5 (2), 217–249.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The Quarterly Journal of Economics, 131 (4), 1593–1636.

Bloom, N. (2009). The impact of uncertainty shocks. Econometrica, 77 (3), 623–685.

Bontempi, M. E., Frigeri, M., Golinelli, R., & Squadrani, M. (2021). Eurq : A new web search–based uncertainty index. Economica, 88 (352), 969–1015.

Castelnuovo, E. & Tran, T. D. (2017). Google it up! a google trends-based uncertainty index for the United States and Australia. Economics Letters, 161, 149–153.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. American Economic Review, 105 (3), 1177–1216.

Kupfer, A. & Zorn, J. (2019). A language-independent measurement of economic policy uncertainty in eastern European countries. Emerging Markets Finance and Trade, 4 (1), 1–15.

Miyamoto, S., Abe, R., Endo, Y., & Takeshita, J.-i. (2015). Ward method of hierarchical clustering for non-euclidean similarity measures. In 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), (pp. 60–63). IEEE.

OECD (2022). Quarterly GDP (indicator). 10.1787/b86d1fc8-en (Accessed on 19 March 2022.

Raihan, T. (2017). Predicting us recessions: A dynamic time warping exercise in economics. SSRN Electronic Journal.

Simkova, N. (2015). The hierarchical clustering of tax burden in the eu27. Journal of Competitiveness, 7 (3), 95–109.

Toda, H. Y. & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66 (1-2), 225–250.