UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTY OF ECONOMICS
AND BUSINESS

DEGREE IN ECONOMICS

BACHELOR’S THESIS

TITLE: **Development of a Synthetic Leading Composite Indicator for the U.S. Economy**

AUTHOR: Diana Shilovskaya

TUTOR: Marcos Bujosa Brun

ACADEMIC YEAR: 2023-2024

June 2024

INDEX

ABSTRACT

The objective of this paper is to develop a synthetic composite leading indicator for the U.S economy. Firstly, an explorative analysis is performed, which is aimed at identifying those monthly time series that frequently lead the economic cycle of the United States. Secondly, for each time series selected in the explorative analysis, the trend is estimated using the method of Linear Dynamic Harmonic Regression. Lastly, principal component analysis is employed to construct a linear combination of the estimated trends from the selected time series. This linear combination forms the leading synthetic indicator that aims to improve economic forecasting and provide valuable insights for informed decision making.

INTRODUCTION

In today's complex global economic landscape, the ability to predict economic trends and patterns with accuracy is of utmost importance for policymakers, investors and businesses alike. Leading indicators play a vital role in this endeavour, as they offer insights into the future of economic activity before official data on GDP growth rates, unemployment and inflation are released. Understanding and constructing robust leading indicators is essential for informed decision making and risk management in private and public sectors both.

The objective of this thesis is to construct a composite leading indicator for the economy of the United States of America, capable of predicting economic expansions and recessions. The motivation behind this project comes from a desire to delve deep into time series and econometric analysis. It is also the goal of this thesis to apply and understand the methodology of Linear Dynamic Harmonic Regression, elaborated by Young et al. (1999) and Bujosa et al. (2007), that allows for the dissection of time series data into its three major components: the trend, the seasonal component and the irregular component. Specifically, the component of interest for this project is the trend, uncovered for each economic dataset and then used to develop a composite synthetic leading indicator for the U.S economy. The extracted trends will be aggregated into one composite economic trend, leveraging their respective weights. For this approach to be effective, the datasets selected must exhibit shared behaviours.

2.1 Importance and Criticism of Leading Indicators

Leading indicators are vital instruments for economic performance forecasts, offering crucial insights into future trends before they reveal themselves. In comparison with lagging indicators that simply confirm past occurrences, leading indicators are capable of providing early signs of potential changes in economic activity. By capturing miniature changes in economic behaviour, such as shifts in private consumption patterns or changes in business investment, these indicators aspire to avert economic catastrophes.

Furthermore, leading indicators play an important role in policy making and implementation, as Central Banks and government agencies around the world rely on these indicators to execute informed monetary and fiscal policy decisions that affect interest rates, taxation patterns and spending.

However, it is no secret that the reliability of leading economic indicators in forecasting and preventing significant economic crises, such as the 2008 recession, has been criticised throughout the 20th and 21st centuries. Paul Samuelson, a Noble laureate in economics (1970) has once stated, “In economics, the majority is always wrong”. What he really referred to is that any sort of economic forecasting is inherently uncertain and subject to error due to the complex nature and dynamics of the economy. Samuelson later explored the challenges of accurately predicting economic phenomena by stating, “The stock market has forecast nine of the last five recessions”, explaining that modern forecast strategies often overpredict recessions by focusing on the short-term fluctuations and noise of the time series, leading to false alarms and a more pessimistic view of the economy.

Another well-known economist, John Maynard Keynes, had also criticised the reliability and use of leading economic indicators by stating that “It is better to be roughly right than precisely wrong”. Keynes argued that when modelling and forecasting economic data, perfect results are unachievable and even unpractical. According to him, there is no universal correct forecasting algorithm that would produce perfect results. Instead, there are different statistical procedures that stem different results, and the challenge presented is to know which procedure best answers the questions that are being asked. Keynes highlighted the importance of flexibility in economic analysis, which is especially justified in the face of economic uncertainty.

Even though the examples above do not refute the use of leading indicators, they call attention to the inherent uncertainties and limitations that are connected with economic forecasting. It is also important to take into account behavioural biases and the complexity of market dynamics that significantly challenge the effectiveness of leading indicators. Human decision making is often irrational and influenced by herd mentality which introduces unpredictability and lack of reason into economic forecasting.

For example, the 2008 crisis was complex to predict due to speculative bubbles and the upturn of the housing market, which led the american society to feel a false sense of economic stability. Moreover, external shocks, such as natural disasters, geopolitical tensions and global health crises often interfere with economic trends and diminish the effectiveness of leading indicators.

EXPLORATORY ANALYSIS FOR VARIABLE SUITABILITY

3.1 Recessions and Expansions: the key for economic health diagnostics

In this study, the accurate development of a leading composite synthetic indicator depended on the adequate election of economic variables that would comprise this indicator. Therefore, the selected variables needed to provide useful insights into consumer confidence, overall economic health and potential future of economic trends. Thus, in selecting these datasets, it was necessary to choose those variables that could signal a recession or expansion before the National Bureau of Economic Research (NBER) would officially announce such economic phases. If a potential variable can signal a downturn of the economy before the Business Cycle Dating Committee of NBER identifies the starting point of a recession, it may be useful to observe the evolution of this variable in order to uncover the future changes in economic patterns.

According to (Leamer, Macroeconomic Patterns and Stories, 2008) a recession can be defined as a market failure which causes persistent and substantial increase in unwanted idleness. A recession is followed by a recovery in which idleness returns to normal levels. The **National Bureau of Economic Research** provides a more technical definition of a recession, defining a recession as a period between a peak of economic activity and its following trough, characterised by a significant decline of economic activity that is spread across various sectors and lasts more than a few months. These two definitions mention a crucial common idea: a recession signals a downturn in economic activity that results in the idleness of productive assets. Idleness of productive assets is not only unwanted but costly because capital rental costs are fixed and must be paid even when the asset is not working. For instance, the manufacturing sector is greatly affected by recessions due to reduced revenues while the fixed costs remain unchanged.

Contrary to a recession, the **Nation Bureau of Economic Research** defines an expansion as the phase of the business cycle when the economy moves from a trough to a peak. This is a period in which economic activity rises and is usually characterised by an increase in industrial production, consumer spending on durable goods and overall rise in economic activity across many sectors. It is possible to single out various economic variables that may prematurely signal the onset of an economic expansion. For example, the rise of consumer confidence is often reflected in increased spending on big purchases, such as housing and automobiles. When households feel optimistic about the future of their finances, it is more likely that they invest in new housing and vehicles. Optimistic consumer behaviour enhances construction and manufacturing sectors, further driving economic growth. Subsequently, firms may hire new staff to keep up with the rising demand, resulting in higher disposable incomes and reinforcing the cycle of rising spending, investment and production.

3.2 Identification of Key Leading Economic Variables

In the search for economic variables that could potentially signal a recession and an expansion prior to its official beginning according to the NBER, there are four key datasets selected in this study that are considered adequate for their inclusion in the synthetic composite leading indicator for the U.S economy. A sample of 56 years of monthly observations was taken from each one of these four datasets, spanning from January 1967 to December 2023. Natural logarithms were applied to the original variables in order to compress the scale of the data, which was later useful for the computation of the leading indicator.

It was important to use non seasonally adjusted raw data because the selected variables needed to contain unmanipulated data. **The reason for this is that pre-handling seasonality would have a negative impact on the study. For the construction of the leading composite synthetic indicator, it is more prudent to address seasonality simultaneously with the estimation of the trends for each of the datasets. The trends of the variables are the key unobservable components for this study. However, if a trend of a seasonally adjusted series is estimated, it would exhibit undesirable properties that would be the result of a chain of filters applied that are harmful to the study.**

The first dataset contains the number of monthly, new privately owned housing units started, in thousands of units, for the United States of America. This data is published by the Federal Reserve Bank of St Louis (FRED), and is originally sourced from the United States Census Bureau and the U.S Department of Housing and Urban Development. This dataset provides insights into the number of new residential construction units that have begun in a specific month. Data on housing starts is especially helpful in determining the overall economic health of a country because housing starts are typically high when individuals are optimistic about their future. The decision to invest into a new home involves substantial financial commitment that would reflect the speculation for long term financial stability. As mentioned earlier, the housing sector greatly influences other industries, such as construction, manufacturing (orders of machinery, raw materials), and services. A rise in housing starts leads to reduced economic idleness and promises prosperity in diverse economic sectors.

Historically, a decline in housing starts has often preceded economic recessions, while a rise in the variable indicated economic recoveries or expansions. For instance, during the boom of the mid 2000s, the housing market in the U.S experienced noticeable growth with a surge in new housing construction. This boom was produced by various financial factors, such as low interest rates on mortgages and substantial speculative investments in real estate.

However, signals of instability in the housing market appeared as early as 2006. Although the construction of new housing units continued, its rate of growth began to diminish all around the U.S. This downturn in the trend of new housing units started was a sign of a weaker consumer demand which served as a precursor of the Great Recession (2007-2009).

The second dataset used for the construction of the leading synthetic composite indicator was a dataset on privately - owned housing units authorised in permit - issuing places. This dataset presents monthly observations of housing permits issued around the U.S and measured in thousands of units. This data has also been sourced from the U.S Census Bureau and the U.S Department of Housing and Urban Development and published in the FRED database. Housing permits represent the number of new housing units that have received the necessary authorizations to begin construction.

This variable is closely related to new housing units started as it serves as a precursor for the initiation of real estate development. A rise in housing permits would indicate an increase in upcoming construction activity, providing an optimistic view of the future economic conditions. Moreover, the number of new housing permits issued may also reflect the changes in the regulations and government policies in different regions, and these changes, when beneficial to the housing market, may anticipate the future health of the real estate sector.

Economic growth often results in more subsidies and government transfers being available, which can stimulate the housing market through tax incentives destined to boost construction activity. Favourable conditions in the real estate and construction sectors would be reflected in the rise of housing permits issued, leading the economy towards a sustained expansion. Consequently, a rise in housing permits issued would boost infrastructure development in urban areas, which would further attract real estate investors and developers.

The third dataset that forms part of the composite leading indicator developed in this study is a dataset on the retail sales of domestic automobiles for the U.S economy. Sourced from the U.S Bureau of Economic Analysis and stored in the FRED database, this dataset provides monthly, not seasonally adjusted data on the thousands of units of domestic automobiles sold. Motor vehicle sales, in particular domestic autos, is another crucial economic variable that can anticipate a surge in economic growth.

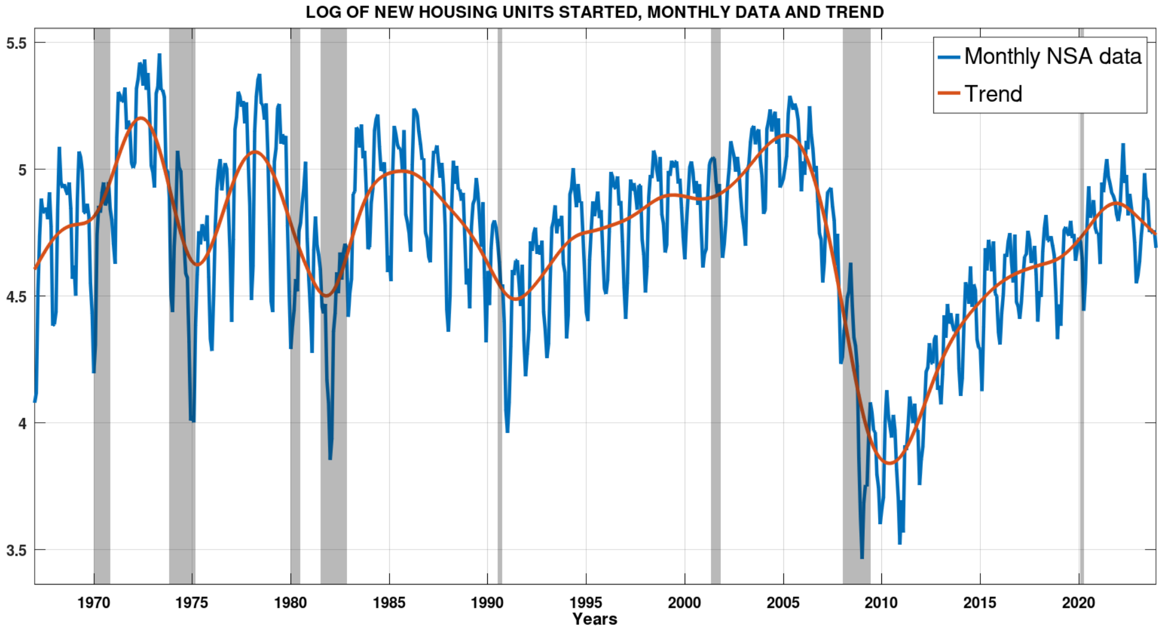
Automobile sales are a considerable part of consumer spending as they reflect the financial health and confidence of households. The purchase of a new car is a long term investment that is costly to maintain, therefore when a decision to make such a purchase is taken, it necessarily reflects the faith of an individual in their own future income stability. If an individual is particularly confident about their own future financial prospects, they may opt for purchasing a new vehicle rather than fixing an old one. Therefore, an increase in vehicle sales can indicate the future improvement of economic conditions of households.

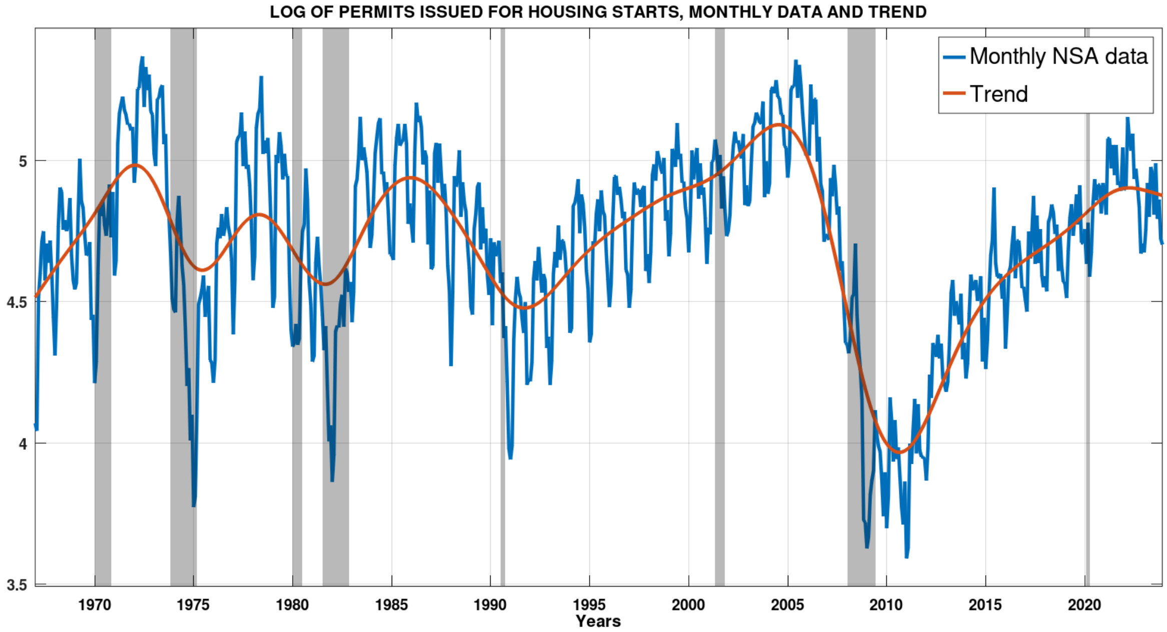
The automotive industry forms a considerable part of the manufacturing sector and is a major employer for workers with both high value added and low value added skills. High auto sales pushed by a positive shift in consumer demand would result in a boost of industrial activity and therefore signal higher economic growth. Moreover, auto sales nowadays are increasingly financed through credit, so a rise in vehicle purchases can reflect favourable credit conditions and consumer confidence in the ability to repay the loans.

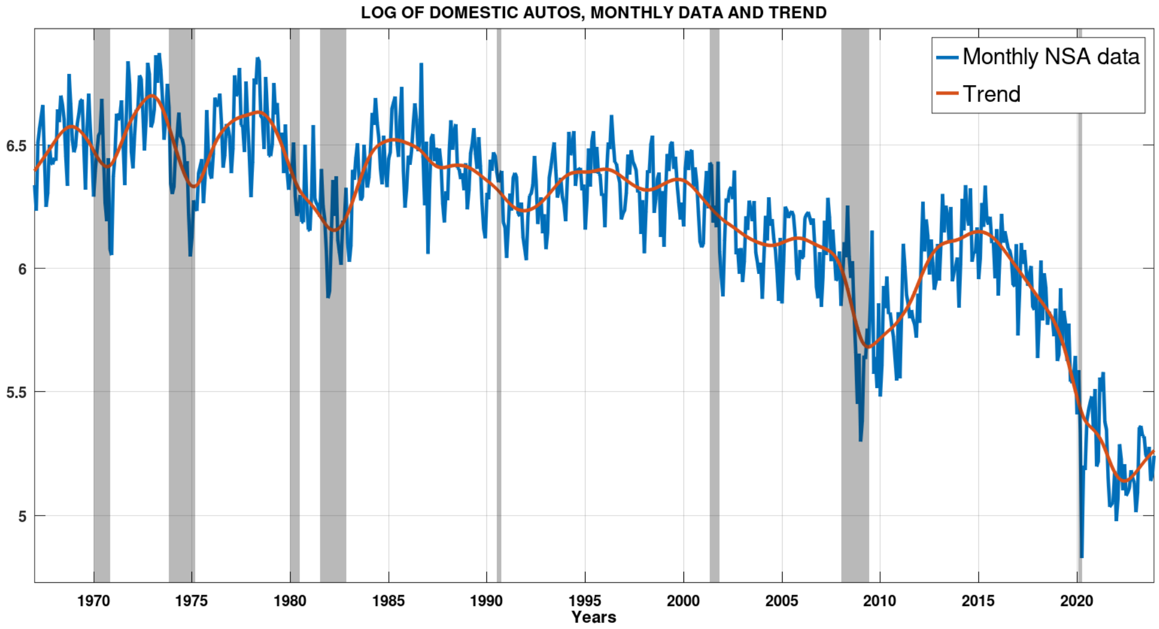
The fourth dataset used in the construction of a composite leading indicator was a dataset (not seasonally adjusted data) on the monthly retail sales of heavy weight trucks, measured in thousands of units. This dataset was sourced from the U.S Bureau of Economic Analysis and published on the FRED database. Sales of heavyweight trucks can act as a precursor for the future of economic growth for a number of reasons. Heavy weight trucks are primarily used for commercial purposes, including transportation and logistics. This means that a spike in heavy weight truck sales can reflect the optimistic expectations about the future economic activity. A positive shift in consumer demand can push businesses to invest in heavy weight trucks because the volume of their operations is expected to increase.

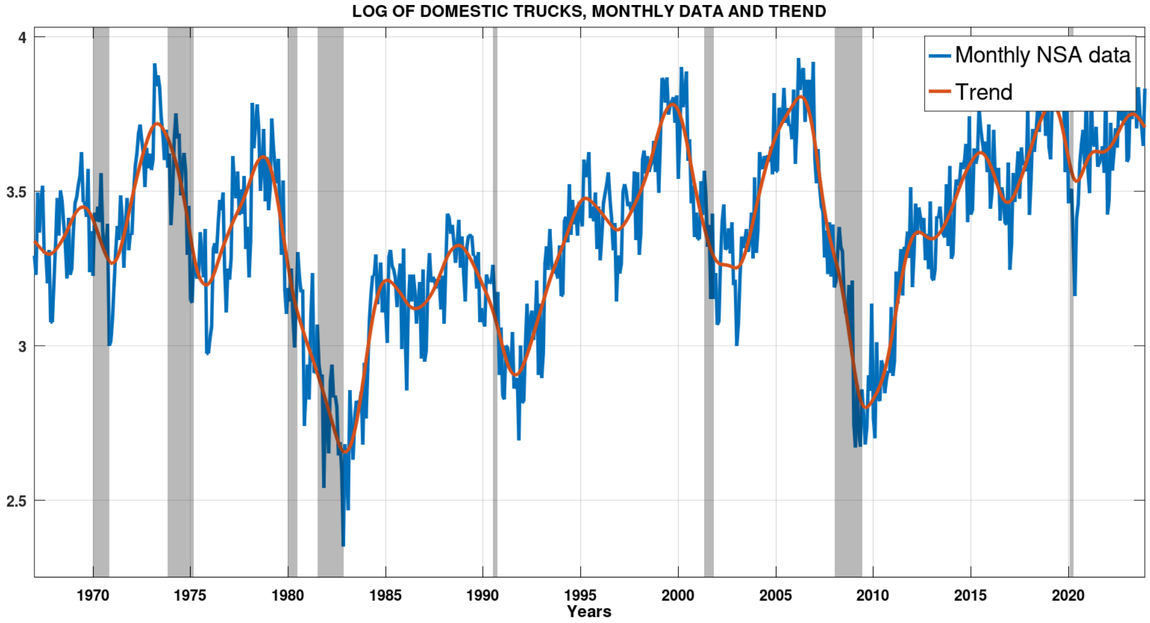
In the late 1990s, heavy weight truck sales were rising which was a reflection of the strength of the U.S economy and overall business confidence. However, in the year 2000 the sales of heavy weight trucks began to decline significantly before the onset of a recession in March 2001. This decrease in sales was an indication that businesses were scaling back on investment in capital goods, anticipating a downturn for the economy.

In the figures below, the shaded regions indicate recession periods for the U.S economy, announced by the NBER's Business Cycle Dating Committee. The trend of each time series is depicted along with the log transformed series.









3.3 Exploratory analysis for Variable Suitability

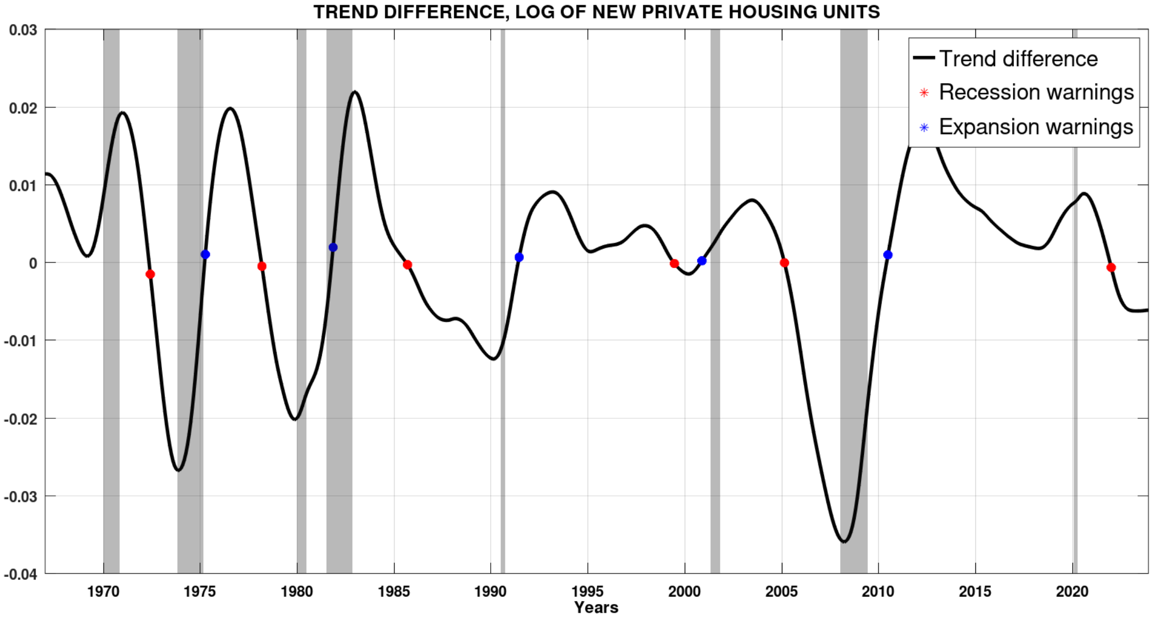
It was necessary to test the suitability of these four datasets in order to justify their inclusion in the synthetic composite leading indicator. The first step in exploring the suitability of each dataset was to estimate the trend of each variable. The trend captures the overall long-term evolution of a variable over time, reflecting persistent increases or decreases in economic activity. The trend is able to reveal the underlying pattern in time series data by smoothing out the noise, or the short - term fluctuations in the variable. Thus, the trend shows sustained changes in the dataset over time, revealing the big picture of the evolution of the variable. Once the trend for each of the datasets was estimated through the method of LDHR (this algorithm is explained below), it was necessary to take the first regular difference of each trend in order to reveal the rate of change of each variable over time. By working with trend differences, it was possible to effectively identify turning points in economic data. Taking the first difference of the trend enables a clear visualisation of changes occurring in the data over time, enhancing the ability to identify shifts present in data patterns, which were not so apparent when observing the trend of the data alone.

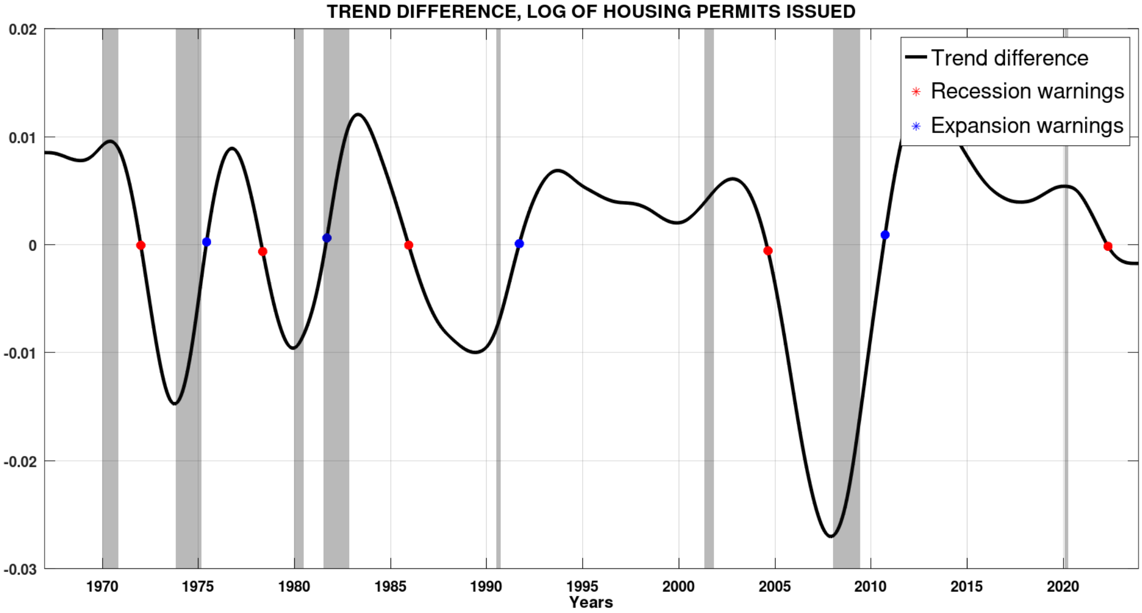
Once the trend difference for each dataset was computed, it was necessary to determine the dates indicating recessions and expansions within each dataset. These dates would have to precede the dates of expansions and recession announced by the NBER for a dataset to be considered suitable for its inclusion in the synthetic leading indicator.

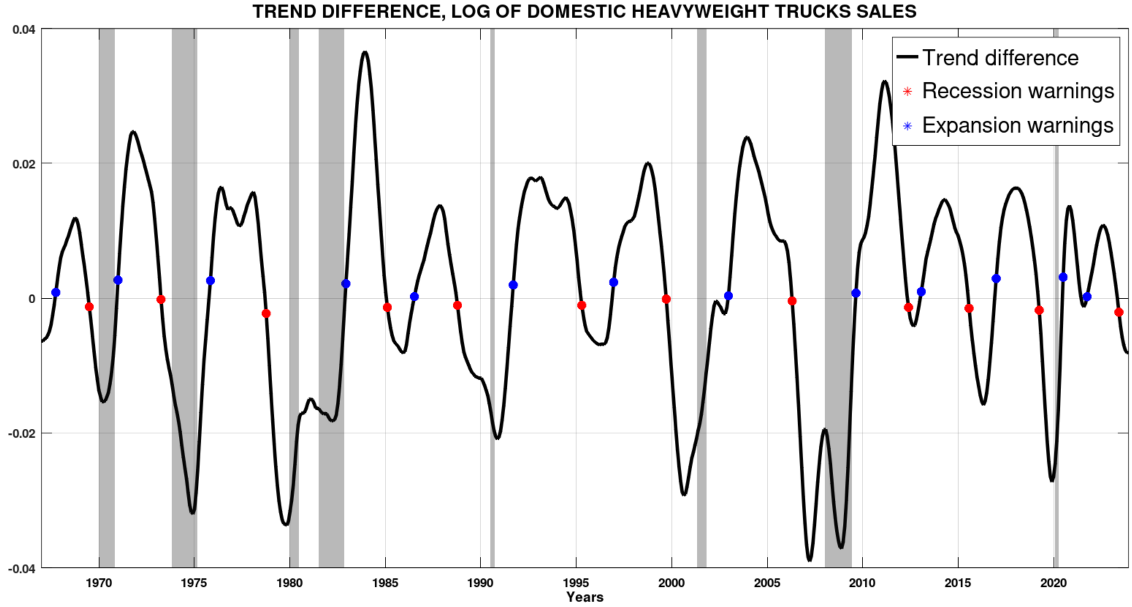
To identify recessions, the focus was on selecting those data points (within each trend difference) with values smaller than zero, and which are preceded by a positive value and followed by six consecutive negative values. Therefore, a recession was identified when a variable exhibited negative growth for six consecutive months, preceded by positive growth. The first negative data in this sequence of consecutive negative growth points would therefore signal the onset of a recession.

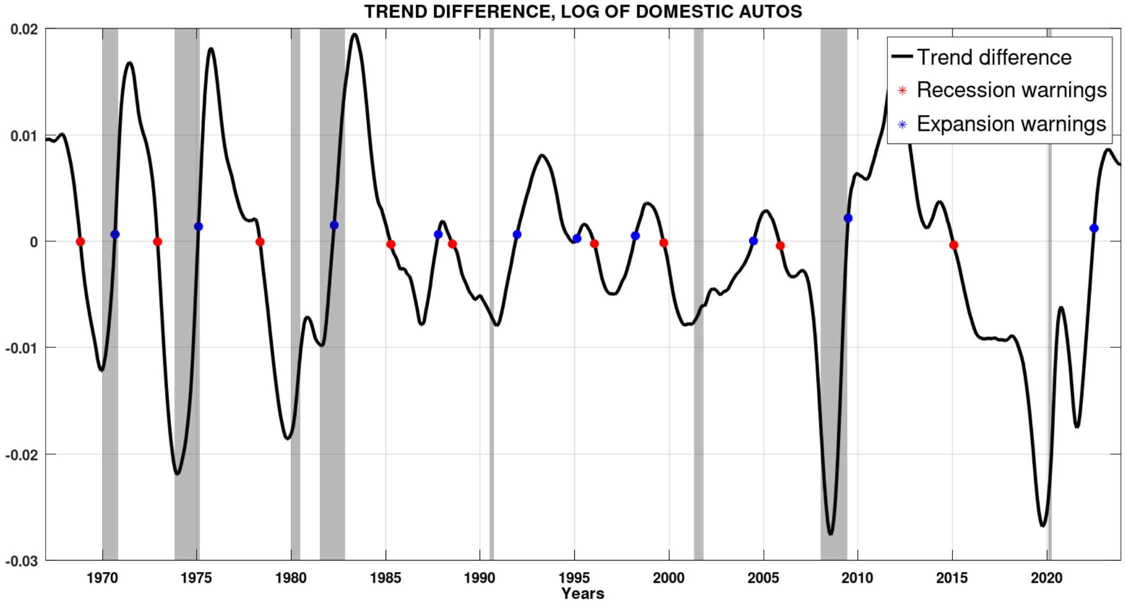
Similarly, to identify the beginning of an expansion period, the procedure involved selecting each positive data point which was preceded by a negative data point and followed by 6 consecutive positive values. Thus, an expansion was identified when a variable exhibited positive growth for 6 consecutive months, and the onset of an expansion was signalled by the first positive value in this sequence of positive economic growth. **(García-Ferrer, A. and Bujosa-Brun, M. (2000). Forecasting OECD industrial turning points using unobserved components models with business survey data. International Journal of Forecasting, 16(2), 207–227. ISSN 0169-2070) URL http://www.sciencedirect.com/science/article/pii/S0169207099000497**

Consequently, by plotting the trend difference of each dataset, shading the regions that correspond to recession periods identified by the NBER, and marking the selected values that signal oncoming recessions and expansions, it was possible to visually identify whether the selected datasets acted as leading variables for the U.S economy.









Each of the four variables selected for the computation of the leading synthetic composite indicator were able to signal the onset of a recession before the U.S economy officially entered a recessive period according to the NBER's Business Cycle Dating Committee.

THEORETICAL FRAMEWORK FOR INDICATOR CONSTRUCTION

4.1 Introduction to Linear Dynamic Harmonic Regression

The methodology of Dynamic Harmonic Regression is employed in this study because it has proven to be useful for working with non stationary time series data. The Linear Dynamic Harmonic Regression algorithm consists in the identification and the estimation of the unobserved components that form the time series, assigning a sensible model for each of them. This estimation algorithm is based on a spectral approach that decomposes the series into several DHR components whose variances are concentrated around certain frequencies.

Considering a univariate time series and applying the LDHR algorithm, the series can be decomposed into the trend, the seasonal, or a periodic component, and the irregular component. The variance of each of those components can be decomposed at different frequencies and this decomposition is known as the spectrum of the series. The trend of the series is the component of most interest for this study, and its variance, which is infinite in the theoretical model, is mainly located at the zero frequency. **The trend is also known as a low frequency component; the objective of this paper is to estimate it for each time series chosen and then utilise a weighted aggregation of trends in order to construct a synthetic composite indicator.**

Before exploring how LDHR is applied in this specific study, it is essential to first discuss how this algorithm would work on a theoretical model. There are two types of models of interest: mean non-stationary and mean stationary models. A mean non-stationary model is a time series model where the average value of the series is non-constant over time, which would imply that the average value of the series may increase or decrease substantially over time. A mean stationary time series is the opposite: the average value of the series is unchanged over time, and even though data points fluctuate around this mean, the average value remains constant along different time periods. It is crucial to mention that stationary time series are much easier to analyse and work with because they exhibit a well defined variance, autocovariance and mean, which are the statistical properties necessary to build a model over which the time series can be interpolated and extrapolated.

However, it must be noted that the time series subject to analysis and identification in this particular study are non stationary: this type of data is most commonly encountered in economics due to the dynamic and non-constant growth of most business and financial variables over time. GDP components, inflation rates, unemployment rates and stock prices manifest systematic patterns with non - constant mean and variance over time, affected by complex consumer behaviour choices and external shocks. This is why non-stationary time series are challenging to model and forecast.

4.2 Theoretical framework of non stationary and stationary time series data

In a theoretical framework with a mean stationary time series, LDHR is set in motion with the concept of the autocovariance generating function.
A time series inherently displays a degree of correlation between data points across time. This correlation can be quantified by calculating covariances between a data point at time 't' and previous data points at 't-1', 't-2', 't-3', and so forth. The computed autocovariances can be recorded into a sequence of numbers which forms the autocovariance generating function, which is actually a vector with it´s first element being the variance of the time series.

A visual representation of the autocovariance generating function can be presented in a correlogram, which is a depiction of autocovariances that are uniformly graphed in bar format, and each bar is commonly referred to as a lag. This is a widely used tool in ARMA model specification where the speed with which these lags decrease with time or the statistical significance of each lag is crucial in identifying whether the model has significant MA (moving average) or/and AR (autoregressive) components.
However, in order to specify a reasonable model for a specific time series, it is also necessary to look at the Partial Autocorrelation Function, which computes the correlation between two observations while holding constant the effect of the remaining lags of the time series. Therefore, PACF is an exceptionally useful tool for identifying an appropriate lag order of a time series model.

A stationary time series that may be identified with an ARMA (Autoregressive Moving Average) model may be expressed in polynomial form, where time series data multiplied by the autoregressive (AR) polynomial are equal to a white noise process

multiplied by a moving average (MA) polynomial. For a series to be mean stationary, the autoregressive polynomial has to be invertible (it has to possess roots outside of the unit circle), which means it would be possible to multiply this polynomial by its own **summable** inverse and the result would yield '1'. This result is known as a unique stationary solution for an infinite stationary sequence of time series data. **Reference.** It is also important to mention that there is an infinite number of inverses for this polynomial, but only one of those inverses is summable.
To expand, every inverse of this AR polynomial is an infinite sequence of numbers, but only one of those inverses has an infinite sequence of numbers such that these numbers are summable to a **finite** quantity (in literature, this is known as an MA (∞) Wold Representation of a Stationary Process). Therefore, in order to adequately express the time series in polynomial form, the autoregressive polynomial must be multiplied by a summabe inverse and the product of the multiplication would yield '1'.
Assuming that the above requirements are satisfied, a theoretical ARMA model is the one for which the original data is equal to an MA polynomial multiplied by the inverse of an AR polynomial, multiplied by white noise process.

To visualise the frequency content of a theoretical ARMA model, the autocovariance generating function must be passed through the Fourier Transform in order to yield a spectrum of a stationary time series. The spectrum of a univariate time series can be defined as a representation of the frequency content of the variance of this series. The analysis and the characteristics of the spectrum will reveal the distribution of the variance frequencies in the data.
It is possible to depict the spectrum between 0 and pi, and the area underneath, or the integral of the spectrum function is the variance of the time series. **The regions where the spectrum function exhibits large values indicate the frequencies that significantly contribute to the variance of the series. The frequency of interest for this study is the zero frequency, at which the spectrum function possesses a maximum.**

Although it is possible to transform the autocovariance generating function into the spectrum of a stationary time series, non - stationary stochastic processes present a huge problem inside this theoretical framework. This is because the autocovariance generating function is not defined for a non-stationary stochastic process. Since non-stationary series possess no explicitly defined mean and have infinite variance, it is not possible to compute the autocovariance generating function because it does not actually exist. Therefore, there is a level of abstraction regarding the statistical methods applied to analyse and model non stationary data. There is no unique stationary solution, nor is there a **summable** inverse of an AR polynomial which would multiply this polynomial and yield '1'. So, whenever the AR polynomial has roots on the unit circle, the spectrum is not defined.

However, for a non-stationary time series, it is possible to define a certain 'pseudo covariance function', which can be passed through an extended Fourier Transform in order to yield the pseudospectrum of the time series. **(M. Bujosa, A. Bujosa y A. García-Ferrer (2015). Mathematical Framework for Pseudo-spectra of Linear Stochastic Difference Equations. — IEEE Transactions on Signal Processing. Volume 63 , Num. 24, pp. 6498-6509. DOI:10.1109/TSP. 2015.2469640).**
The pseudospectrum is a function that describes the distribution of variance over frequency for a non stationary stochastic process. The area underneath, or the integral of this function, is the variance of the time series, just like in the mean stationary case, but with non-stationary data this variance is infinite for some frequencies. **Reference.** Specifically, the representation of the pseudospectrum is similar to the one of the stationary time series spectrum, except that each root of the unit circle of non - stationary data results in a peak with infinite variance for some frequencies represented in the pseudospectrum.

4.3 Application of LDHR to time series framework

The goal of Linear Dynamic Harmonic Regression within time series modelling is to adjust the "pseudospectrum" of a theoretical DHR model to the periodogram of the time series analysed. The periodogram of the data is an estimation of the spectrum of a time series.

The objective of the LDHR algorithm is to find a sufficiently large autoregressive polynomial to fit to the data, and to analyse its polynomial roots. Each one of those polynomial roots corresponds to a certain frequency, and the frequencies of interest within this study are the zero frequency and the seasonal frequency with

its harmonics. After identifying these polynomial roots, the task of the LDHR algorithm is to specify a DHR model for each one of the components of the time series. Each one of those time series components correspond to a spectral peak associated with the polynomial roots of the autoregressive polynomial.

Once the models for the DHR components of the series are specified, the algorithm must adjust the "pseudospectrums" of the models to the estimated spectrum of the series.

The challenge of this algorithm, however, lies in minimising the distance between the 'real' pseudospectrum of the time series and the estimated spectrum of the data. This minimization is impossible to achieve when the pseudospectrum has spectral peaks where the variance is infinite. The area underneath those poles is infinite which means that the distance to minimise is not defined.

Linear Dynamic Harmonic Regression addresses this challenge by using a certain function that is able to transform the poles of infinite variance of the pseudospectrum into points with zero variance. In other words, the minimization problem is multiplied by a certain function that allows for a finite minimal distance between the pseudospectrum and the estimated spectrum. LDHR uses an Ordinary Least Square algorithm in order to minimise this distance.

PRACTICAL DEVELOPMENT OF SYNTHETIC INDICATOR

5.1 The erroneous application of Moving Average filter for trend identification

Before using the Linear Dynamic Harmonic Regression method to extract the smoothed trend for each of the chosen datasets, a more common technique known as the Moving Average smoothing has been applied to this study. Moving Average smoothing is a well known algorithm used in time series analysis to reduce noise and reveal the underlying patterns within the data. **This technique involves calculating the mean of a subset of neighbouring data points and substituting each data point in this subset with this calculated mean.**

Within the Octave framework, the ‘movmean’ function had been used to smoothen each one the datasets by replacing the original data with a set of moving average values, which revealed the trend of each series.

However, the moving average filter has not proven to be useful in this study because of the way that the filter has manipulated the data. Once the Moving Average filter was applied, it showed a trend that exhibited significant noise, and it was desired to eliminate such irregularities during the smoothing process. These irregularities were enhanced when the first difference of the trend was taken. Consequently, in order to achieve a smoothed and different trend, another moving average filter was applied to the data, further distorting the true pattern of the series. The reason why the Moving Average filter had revealed such undesirable results is because Moving Average filters manifest a poor frequency response in time series data smoothing. Time series data exhibit various frequency components, including the low-frequency trend and the high-frequency noise. Seasonal components display a range of frequencies between the highest and the lowest of them. The Moving Average filter aims to diminish the high frequency noise in the data and preserve the low frequency components, such as the trend, or any other gradual variations in time series data. However, the issue is that this filter uniformly weakens all frequencies, resulting in insufficient reduction of high frequencies during the smoothing process. Consequently, the trend pattern does not appear as smooth as desired.

This issue is aggravated when applying the first difference to the trend estimated with the moving average filter. It is necessary to apply the first difference to the trend of the data because the objective of this study is to highlight the cyclical behaviour of the trend of the variables so that it is possible to compare them to the U.S business cycles defined by the NBER committee.

Taking the first difference is a widely used technique to put emphasis on short-term fluctuations to better capture the rate of change of the data. That said, the first difference filter amplifies high-frequencies once again, and since the Moving Average filter does not reduce the high frequencies sufficiently, once they are amplified, the differenced trend exhibits noisy and unclear behaviour.

In this particular case study, the Moving Average filter had also exhibited poor performance when it came to handling anomalies in the data, and this was particularly harmful because these anomalies provided valuable insights into the underlying patterns of the data. The filter smoothed out important spikes and dips, discarding vital information about the behaviour of the time series. Furthermore, the ‘movmean’ function in Octave operated by sliding a window of values across the data in order to compute the average for each window. This meant that the initial and the final observations of the sample were gravely affected by the smoothing process. At the beginning of the sample, only succeeding values were averaged, while at the end of the sample, only preceding values were averaged. This asymmetric averaging distorted the edges of the sample, resulting in the higher effectiveness of the filter in the middle of the sample but providing inaccuracies on the edges.

In contrast, the Linear Dynamic Harmonic Regression algorithm had proven to effectively reduce high frequencies of the original data to zero in the smoothed trend. Consequently, when the first difference was applied to the trend, higher frequencies were amplified, but since these high frequencies had been eliminated successfully in the previous stage, the outcome was a smoothed differenced trend. This is because the annulled high noise frequencies remained unaffected during the amplification process of applying the first difference.

However, the LDHR approach cannot avoid the problem of only averaging with past and future data points at the extremes of the sample, which generates greater uncertainty in the estimations at the extremes of the sample, as well as introduces a temporal lag. Specifically, at the beginning of the sample, only the future data points are considered, while at the end of the sample only past data is used. This problem is unavoidable regardless of the smoothing procedure used.

5.2 Practical Application of Linear Dynamic Harmonic Regression

After the selection of time series datasets from the Federal Reserve of Economic Data (FRED), the primary tool employed in this study was the GNU Octave scientific programming language. The objective was to apply the LDHR algorithm (through custom made Octave functions) to each, monthly, not seasonally adjusted dataset, and estimate each one of the DHR components for every dataset. Consequently, the estimated trend of each time series was taken and used for the construction of the leading synthetic indicator.

Initially, monthly time series data that has not been seasonally adjusted, has been selected. It was vital for this study to employ raw data that has not been modified or filtered using other statistical techniques since the objective was to use the LDHR algorithm only to transform and dissect raw original data.

Data used in this analysis was taken from mean non - stationary, infinite stochastic processes. However, the samples extracted from these time series are composed of a finite set of observations. Therefore, these samples have a defined mean and finite variance, and so it is possible to define the autocovariance generating function and the corresponding spectrum that can be represented on the periodogram. The spectral peaks in this study would take very large values (however, they are not infinite), and these peaks are located at frequencies with a significant contribution to the variance of the time series.

For each dataset, the practical application of the LDHR algorithm has been set in motion with the ‘autodhr’ custom made function. This is a key function that identifies and estimates a model for each DHR component of the series. For the correct execution of this function, one of the inputs used was a vector that contains the periodicities of the harmonics corresponding to seasonality of the series. This vector instructed the ‘autodhr’ program to seek models with a trend and a seasonal component with the corresponding harmonics.

Specifically, for monthly observations, the seasonal component follows a harmonic pattern with oscillations at periods of 12, 6, 4, 3, 2.4, and 2. The spectral peaks for the seasonal component of monthly data are accordingly found at frequencies such as '2pi/12', '2pi/6', '2pi/4', '2pi/3', '2pi/2.4' and '2pi/2'.

Another input for the ‘autodhr’ function was a matrix that indicated the modulus of the roots of the autoregressive process of the dataset, which would help model the amplitude of the oscillations for each model component. This matrix was key to capturing the dynamic behaviour of the time series components.

The ‘autodhr’ program processed the data and had identified and estimated the most sensible model for each DHR component of each dataset.

The next step is to estimate the variance of innovations of the stochastic process that would model the oscillations of each of the DHR components. In the theoretical LDHR framework, the variance of innovations of the DHR components are the coefficients from the OLS estimation that was used to minimise the distance between the pseudospectrum of the data and the estimated spectrum.

The final step to extract DHR components for each time series involved applying the ‘dhrfilt’ function. This program decomposed the data into its trend, its seasonal and its irregular component. This custom instruction had used the Kalman Filter that was able to accurately extract the necessary components from the series. Although not part of the Linear Dynamic Harmonic regression algorithm, the ‘dhrfilt’ function used the parameters obtained by the ‘autodhr’ function.

5.3 The theoretical basis of Principal Component Analysis

Once the trends of the four time series chosen for the construction of the leading indicator have been obtained through LDHR, Principal Component Analysis (PCA) was employed to compute the synthetic composite indicator for the US economy for the years between 1967 and 2023.

PCA is a widely used statistical technique that allows to simplify large datasets while retaining the most joint variability of the data. This methodology transforms the original variables into a new set of variables called principal components, identifying which variables account for the most variability in the dataset. To perform PCA on a dataset, the first step is to compute the variance - covariance matrix of the variables. This matrix contains the information about the variation of each variable from its mean value, as well as how much the variables vary together, or the covariance between them. The variance indicates the spread or dispersion of each variable, while the covariances measure the degree to which two variables change together. The transformation of the original dataset into a matrix of covariances is crucial as it reveals shared behaviours between the variables.

Once the covariance matrix of the dataset is defined, PCA proceeds by calculating the eigenvalues and eigenvectors of this matrix. **An eigenvector of a matrix can be defined as a vector that, when the matrix is multiplied by it, yields this vector being scaled by a certain constant, known as the eigenvalue.** The covariance matrix is symmetric, which is a fundamental property to perform PCA: the matrix is diagonalizable and therefore has real, non-zero eigenvalues and corresponding eigenvectors **(linear algebra theorem).** Thus, the symmetry of the covariance matrix is vital because it ensures that the eigenvalues and eigenvectors can effectively represent the directions and magnitudes of the variability in the data.

The eigenvector with the largest eigenvalue is a linear combination of the original dataset that displays the largest variability of the data. Essentially, when the covariance matrix is multiplied by this eigenvector, the result captures the largest variability of the dataset. Therefore, the objective of Principal Component Analysis is to reduce the dimensionality of the data and preserve the maximum variance of this data at the same time.

PCA is especially useful for the manipulation of economic data because economic datasets often contain a large number of variables that are correlated between each other. PCA can reduce the dimensionality of the data as well as uncover underlying patterns in the datasets. This can help identify the behaviour of complex economic phenomena.

5.4 Practical implementation of Principle Component Analysis

The initial step in the computation of the composite synthetic indicator involved the aggregation of the four trend variables obtained through LDHR into a single matrix, representing the dataset to analyse. Each row of this matrix corresponded to one of the trends considered. Specifically, the first row contains the trend values for the log of monthly new privately owned housing units started, the second row corresponds to the trend values from the log of heavy weight trucks sales dataset, the third row contains the trend values for the log of domestic autos sold, and the fourth row corresponds to the trend values for the log of new housing permits issued in the U.S.

To sum up, the dataset was recorded in a matrix of four rows and as many columns as there were observations, with each column representing the four trend values for a particular month for a sample of years observed (1967 to 2023). However, this original data needed to be transformed into a symmetric covariance matrix which would facilitate the detection of eigenvalues and eigenvectors of the dataset.

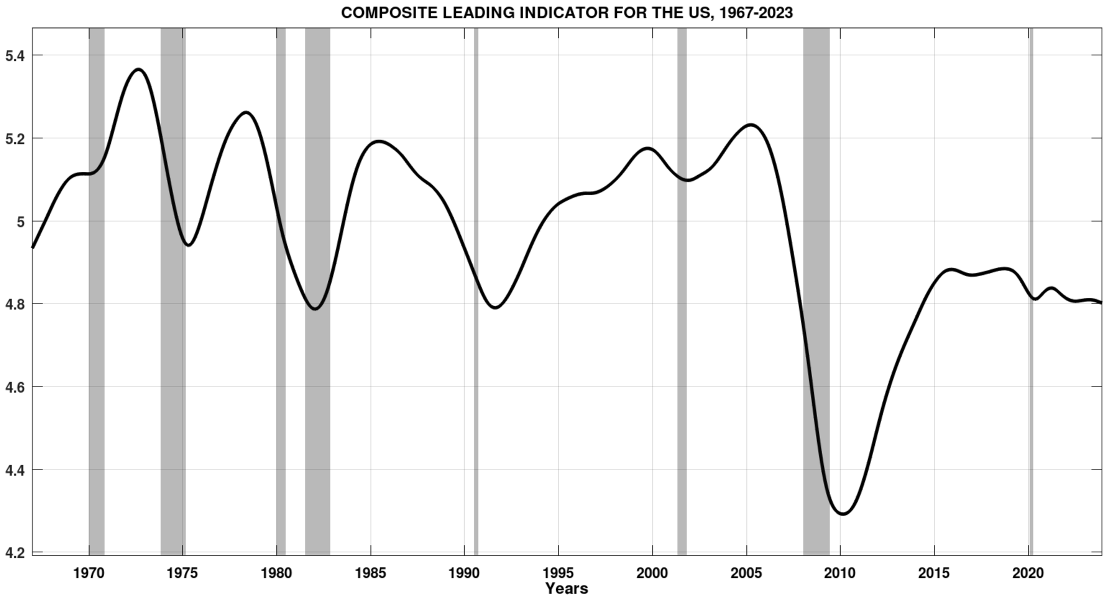
Therefore, the second step to complete the PCA in this study was to calculate the variance - covariance matrix from the matrix of original data. A time lag of 5 months was introduced for the computation of this covariance matrix in order to measure the autocovariance of each trend (or the variance of each trend over time). The covariances between trends were also intertemporal measurements between the present value of one trend variable and the past value of another trend variable from 5 months ago. A covariance matrix of such design captured the intertemporal correlations between trends. **(Peña, D. and Poncela, P. (2006). Nonstationary dynamic factor analysis. Journal of Statistical Planning and Inference, 136(4), 1237–1257.)**
URL http://www.sciencedirect.com/science/article/pii/S0378375804003659

The third step of this analysis was to compute the eigenvalues and the eigenvectors of this covariance matrix, which were calculated with the use of built-in functions in the GNU Octave programming language. There were two outputs for this computation. The first one was a four - by - four matrix where each column was an eigenvector for the dataset. The second output was a diagonal matrix where each entry on the main diagonal corresponded to an eigenvalue related to the eigenvector from the eigenvector matrix. The largest eigenvalue in this diagonal matrix was the first entry of this matrix, and this eigenvalue corresponded to the first column, or the first eigenvector of the eigenvector matrix.

Thus, the eigenvector with the largest eigenvalue contains the coefficients of the linear combination of the four trends that capture the direction of the maximum joint variability of the original dataset. However, in order to obtain a synthetic composite indicator which is a weighted aggregate of the four trends chosen, it was necessary to normalise, or reweight the elements of the eigenvector so that their sum yields ‘1’. This was done by dividing each element of the eigenvector by the sum of all its elements. The result yielded a vector of weights, indicating the weighted composition of the synthetic leading indicator.

Specifically, the composite leading indicator constructed for the U.S economy in this study is **34%** composed from the **housing units started dataset**, **12%** composed from the **heavy weight trucks dataset**, **28%** composed from the **domestic autos sold dataset** and **26%** composed from the **housing permits issued dataset.**

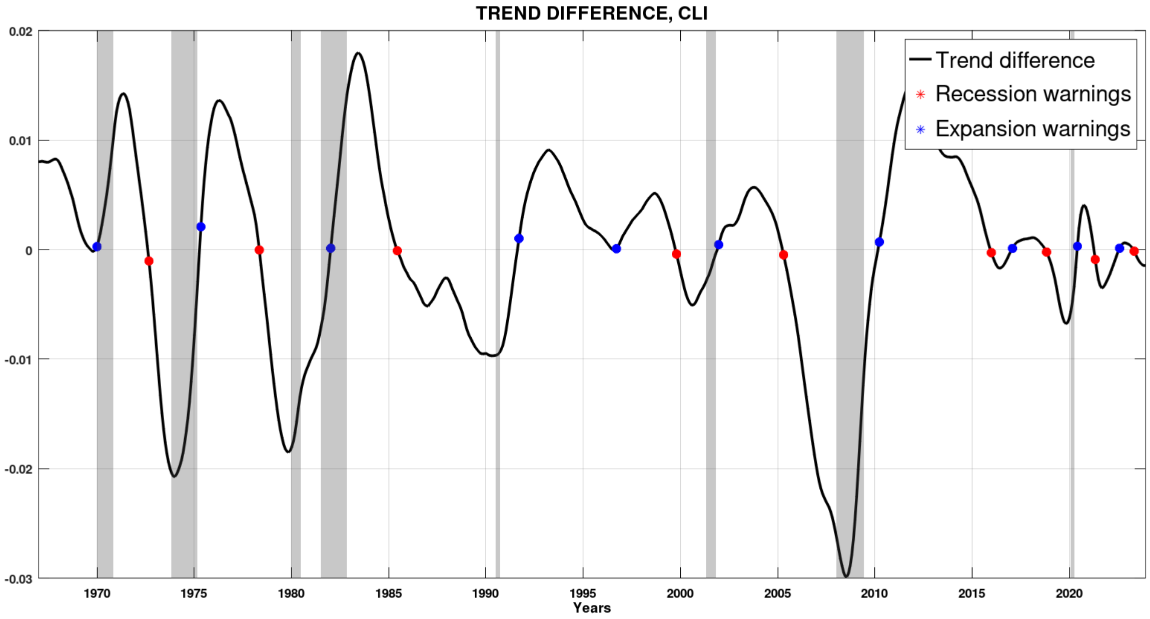
The composite leading indicator is depicted in the figure below. Shaded regions indicate NBER - dated recessions.



EMPIRICAL RESULTS

In order to test whether the synthetic indicator depicted above predicts the upcoming economic conditions, it was necessary to conduct the same explorative analysis that was used to identify those economic variables that lead the U.S. business cycles. Therefore, the first regular difference of the composite indicator was taken to capture the cyclical behaviour of the indicator. Then, it was necessary to identify the dates of recession and expansion warnings according to the composite indicator. For expansion warnings, positive data points within the indicator were selected that were preceded by a negative data point and followed by six consecutive positive values. For recession warnings, negative values within the indicator that were preceded by a positive value and followed by six consecutive negative values were selected.

According to the depiction of the first regular difference of the composite synthetic indicator, dates of recession warnings are preceeding the NBER's Business Cycle Dating Committee's recession announcement dates. This is an adequate result because the synthetic indicator was able to predict six of the 8 recessions that were reported by the NBER between 1967 and 2023.



CONCLUSIONS