Vincenzo Brigandì
Email: brigandi.vincenzo@yahoo.com
Mobile: +39 3484467743

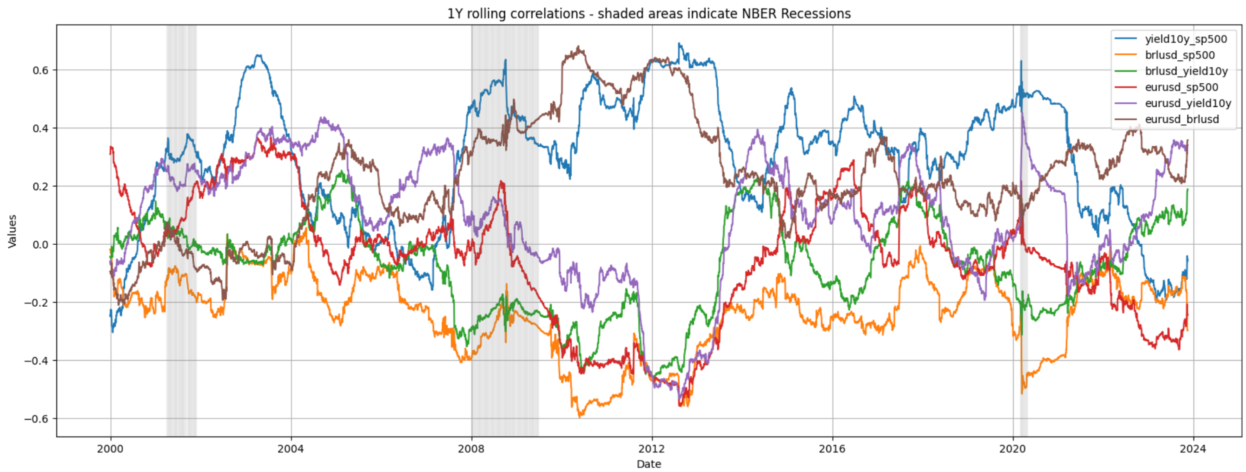
[redacted] Macro Quant Analyst Business Case

Descriptive analysis

The objective of this analysis is to ascertain whether four series, specifically the US 10Y yield (henceforth *yield10y*), S&P500 Total Return (*sp500*), EUR/USD (*eurusd*), and BRL/USD (*brlusd*), exhibit movement patterns during periods of robust economic growth versus economic slowdowns. Data were gathered from St. Louis Fred and Yahoo Finance, covering the period from December 31, 1999, to November 17, 2023, resulting in a total of 6265 daily observations. Data points prior to December 31, 1999, were omitted to maintain a consistent number of data points for all four series in the dataset.

Initially, it is noteworthy to calculate the historical correlation of the daily percentage change in these series to discern if there is a consistent pattern of positive or negative correlation among them. Furthermore, it is of interest to examine how these correlations, as well as the first and second moments of the series, evolve during periods of economic growth or recession. The primary definition of market regimes utilized in this analysis is based on the NBER Recession Indicators for the United States. This time series is an interpretation of the US Business Cycle Expansions and Contractions data provided by The National Bureau of Economic Research (NBER). Our time series consists of dummy variables representing periods of expansion (indicator is 0) and recession (indicator is 1). We employ this indicator as it is widely accepted and originates from a well-known and reputable organization.

Throughout our analysis period, three recessions, as defined by NBER, occurred: the early-2000 recession, the Great Financial Crisis, and the Covid-19 recession. Correlations exhibit considerable volatility across this period, fluctuating from negative to positive values. Generally, it appears that *sp500* and *yield10y* are more positively correlated during contractions, while *brlusd* and *sp500* are more negatively correlated.

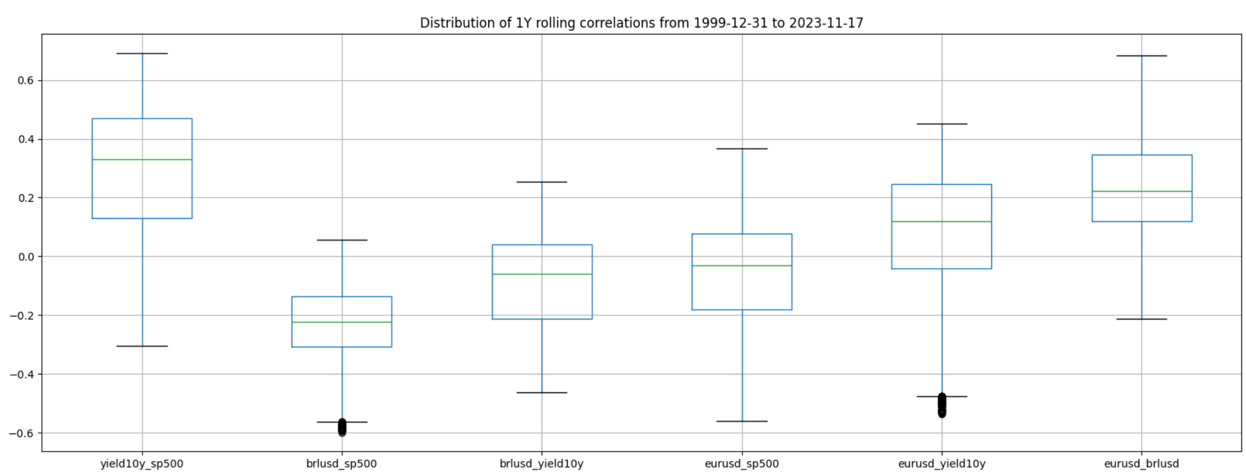


To corroborate the intuition, we compute the median, 25% and 75% percentiles of correlation between the series in case of a normal economic growth or recession and indeed correlation of *sp500* and *yield10y* are effectively higher during recessions. This can be explainable considering that the stock market normally underperforms during recessions and interest rates drop because of the anticipation of a monetary policy ease by Central Banks to support the economy.

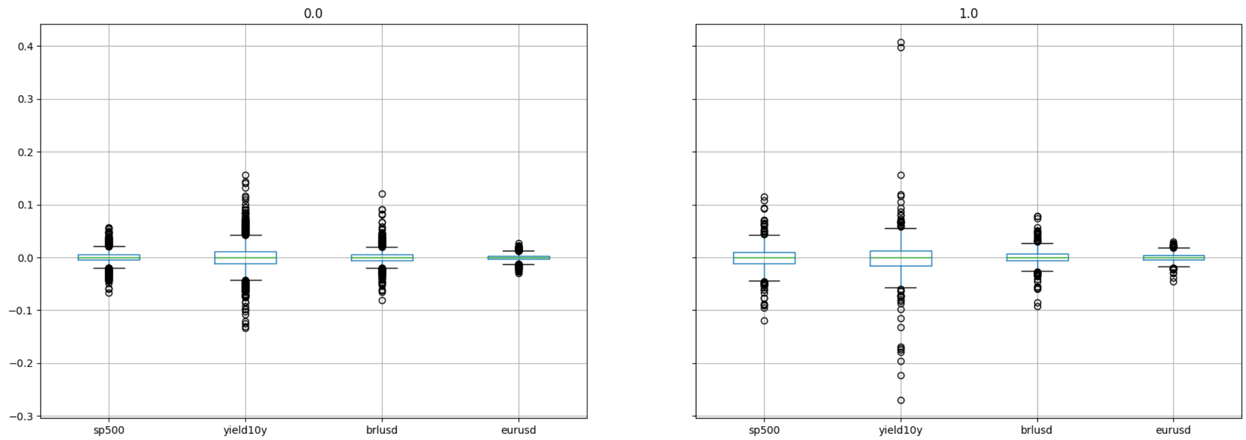
nber_recession	0.0			1.0		
percentile	0.25	0.50	0.75	0.25	0.50	0.75
yield10y_sp500	0.113692	0.314493	0.465104	0.348786	0.437835	0.517808
brlusd_sp500	-0.306877	-0.217359	-0.134526	-0.308049	-0.258622	-0.173083
brlusd_yield10y	-0.189452	-0.053288	0.052091	-0.242061	-0.218613	-0.010322
eurusd_sp500	-0.214491	-0.037806	0.063841	-0.051033	0.037511	0.090046
eurusd_yield10y	-0.059299	0.119263	0.250770	0.042623	0.108744	0.205995
eurusd_brlusd	0.122174	0.216720	0.326487	0.044407	0.353926	0.389857

Correlations 25%, 50%, 75% percentiles in case of Recession (1) or Economic growth (0)

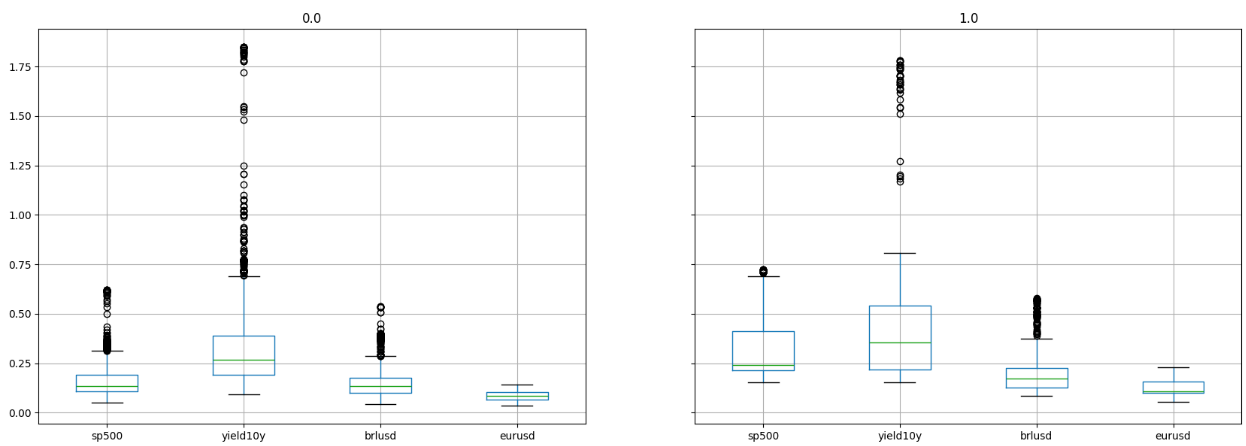
However, in general the correlations between our series are quite dynamic, ranging from positive to negative values, but over time the series are quite uncorrelated except for few outliers.



It is worth exploring the average %change (also called daily returns for asset like stocks, fx, but not for yields, that is why we prefer the notation of %change) for our series as well the 3-months annualized volatility of %change between the two states.



Average %change during periods of economic growth (flag 0) and recessions (flag 1)



3-months annualized volatility of %change during periods of economic growth (flag 0) and recessions (flag 1)

What is clearly visible is that daily %change is on average centered in 0 for both the two states; during recessions one may observe more outliers and volatility increase particularly for equity and rates, and less for the forex series.

In summary, the initial phase of our analysis indicates that during recessions, there is a notable increase in volatilities, while daily percentage changes do not significantly differ from periods of economic growth. Additionally, correlations among our four series do not exhibit a distinct pattern; instead, they appear to be randomly cyclical and less predictable.

Unsupervised Clustering for regime identification

Thus far, we have utilized an external measure provided by the NBER to identify economic regimes. However, this approach may appear overly stringent, as it only delineates periods of recession and non-recession. Presently, our objective is to adopt a more sophisticated methodology that categorizes the economic landscape into three potential states (ideally: recession, mild growth, and strong growth). This task poses a challenge, as determining the key variables and establishing appropriate thresholds is not straightforward. Additionally, the complexity of the economy, involving households, firms, the government, and the Central Bank, makes it impossible to identify a single indicator as a reliable barometer without oversimplifying the intricate nature of the system. Therefore, our proposed approach involves employing a variety of widely recognized indicators, crucial components of the macro picture, and assessing them collectively on a monthly basis to gauge the state of the economy.

The indicators employed are:

- Non-farm payrolls (labour market indicator)
- Initial jobless claim (labour market indicator)
- Personal consumption expenditure (FED's inflation preferred measure)
- Industrial production (proxy for GDP)
- PPI (producer price index, another measure for inflation)
- CPI less food and energy (inflation less volatile items, captures stickier inflation)
- Oil price (proxy for cost of energy)
- Real 10y rates (real cost of long-term borrowing)
- M2 (proxy for amount of money in the system)
- Housing starts (captures the housing sector which has a cascade effect on other sectors and is considered a leading indicator)
- New orders of non-defense capital goods excluding aircrafts (proxy for economic activity)

Real GDP was not employed due to its quarterly frequency and its nature as a predominantly lagging indicator. Additionally, the breakeven inflation rate was excluded as well because the series has a relatively recent inception.

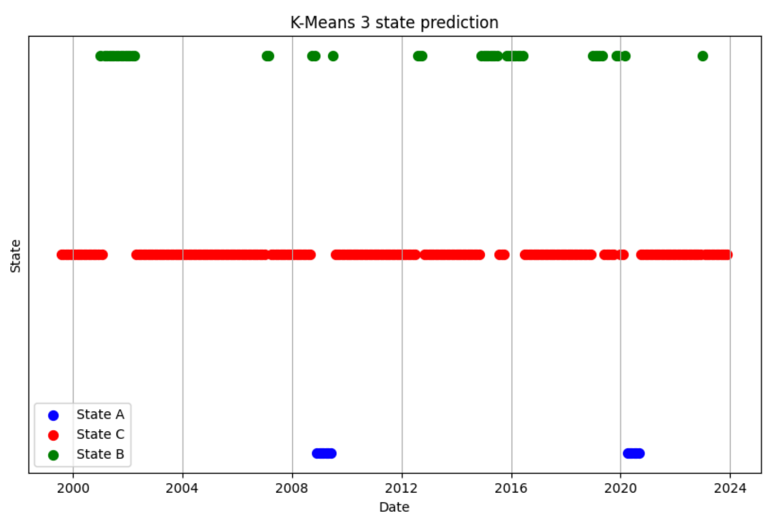
While these series serve as imperfect proxies for assessing the state of the economy — considering numerous others that could be relevant — they offer a reasonable approximation without succumbing to the curse of dimensionality problem. Our data collection involves a monthly frequency, with the series undergoing percentage changes to eliminate trends. Subsequently, a 6-month geometric mean is applied for smoothing, which proves more effective than an arithmetic mean when handling percentage changes. Standardization is then carried out by subtracting the mean and dividing by the standard deviation.

The challenge lies in blending these series together. Given the difficulty in establishing optimal rules and thresholds for each variable, a task prone to subjectivity, we have opted for a 3-class K-means approach. This unsupervised method clusters observations into three states, denoted as A, B, and C. Essentially, the algorithm groups months with similar characteristics in terms of the selected variables, with the aim that months within the same class share a corresponding economic state.

	f1	f2	...	f10	f11	State
m1						A
m2						B
m3						A
m4						C
...						C
...						C
...						B
...						B

Representation of how the model works (m1 is month 1, f1 is feature 1)

To achieve this, we utilized a K-means algorithm with 3 clusters and a maximum of 1000 iterations in training. To enhance the robustness of the model, for each month (i), we executed the same model with 100 different seeds. Subsequently, we classified the month (i) by determining the mode of the clusters obtained from these 100 different seed runs. A clustering ensemble is a popular method when dealing with unsupervised algorithms aimed at combining multiple models to produce a better result in terms of consistency and quality. The ensemble methods used is a naïve one and is based on the frequencies of occurrences of the three different states: for each random seed a state is labeled as infrequent (A), moderately frequent (B), or highly frequent (C) and then the mode is computed across all random seeds. This is the outcome of the analysis:

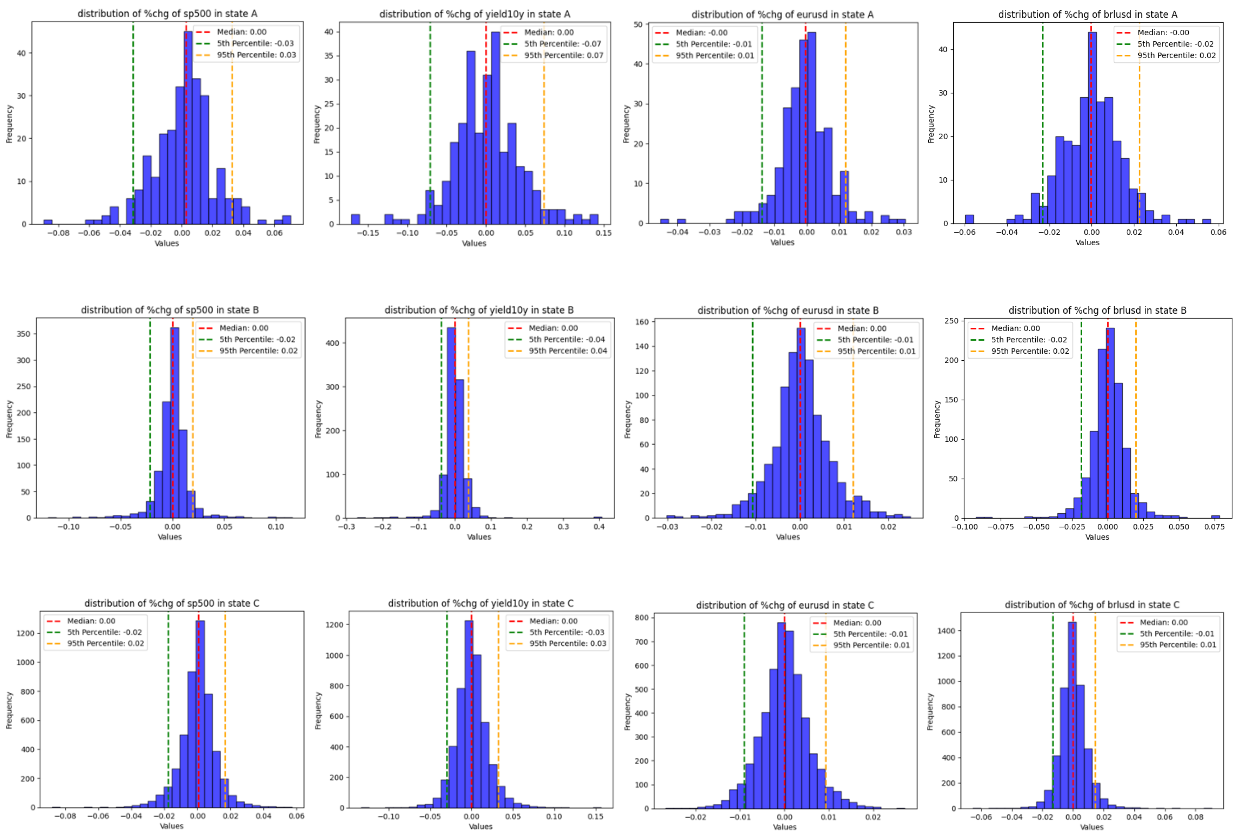


K-Means unsupervised state prediction. Each month is labeled as belonging to state A, B, or C

State C is the most frequent and represent the average state of the economy, state A captures recessions corresponding with the Great financial crisis and the Covid-19 contraction and state B is a third state of economic that often overlaps with periods of economic growth. Being this approach unsupervised it is not always straightforward to map the states with standard classification of economic regimes, however the results of the clustering seem reasonable.

After categorizing the economic states, it's valuable to investigate the behavior of the four analyzed series within these distinct states. In state A, both the *sp500* and *yield10y* series show wider tails in terms of percentage change, characterized by larger 5% and 95% percentiles. This effect is more pronounced for *yield10y*, where the range expands from approximately +/-3% in state C to around +/-7% in state A.

The medians remain constant across all states for each series, and the percentiles for the two foreign exchange (fx) series do not vary.



Distribution of %change during the 3 K-means macro regimes

The impact of the state is evident in the annualized volatility as well. Across all series, the volatility increases in state A or B compared to state C. Notably, this increase is particularly significant for *sp500* and *yield10y*, where it more than doubles.

	sp500	yield10y	brlusd	eurusd
km_state				
A	0.322735	0.711154	0.235046	0.136621
B	0.260295	0.525666	0.207321	0.105165
C	0.169500	0.314000	0.146330	0.087861

Annualized volatilities of %change series in the 3 k-mean macro regimes

In summary of this segment of the analysis, it appears that the unsupervised model effectively identifies distinct macro states within the economy. This identification is subsequently reflected in the volatilities and the fatter tails of the distributions in the daily returns for our series. This impact is more pronounced in the cases of *sp500* and *yield10y*, that are the asset classes normally most impacted in a recession, while the distributions of *brlusd* and *eurusd* are comparatively less affected.

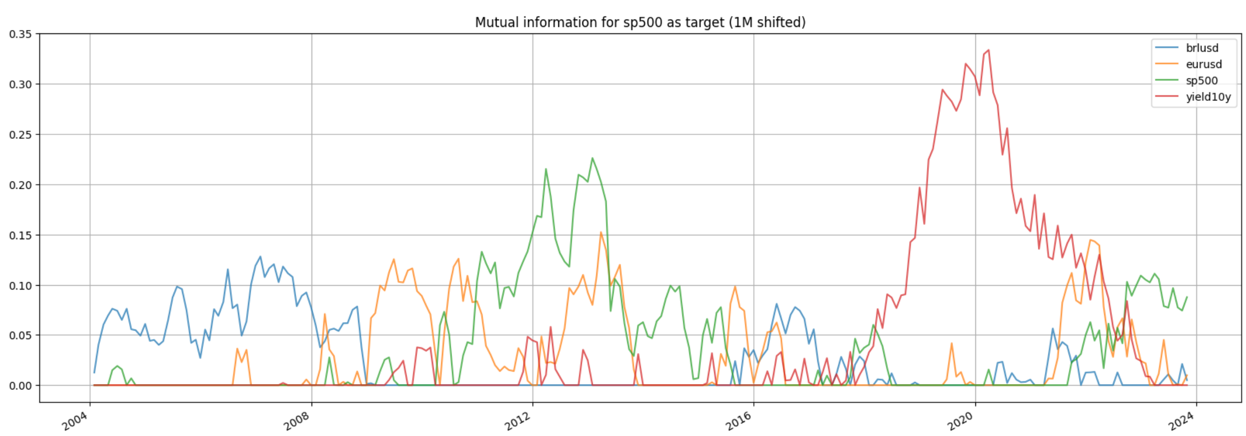
Asset leading indicator property

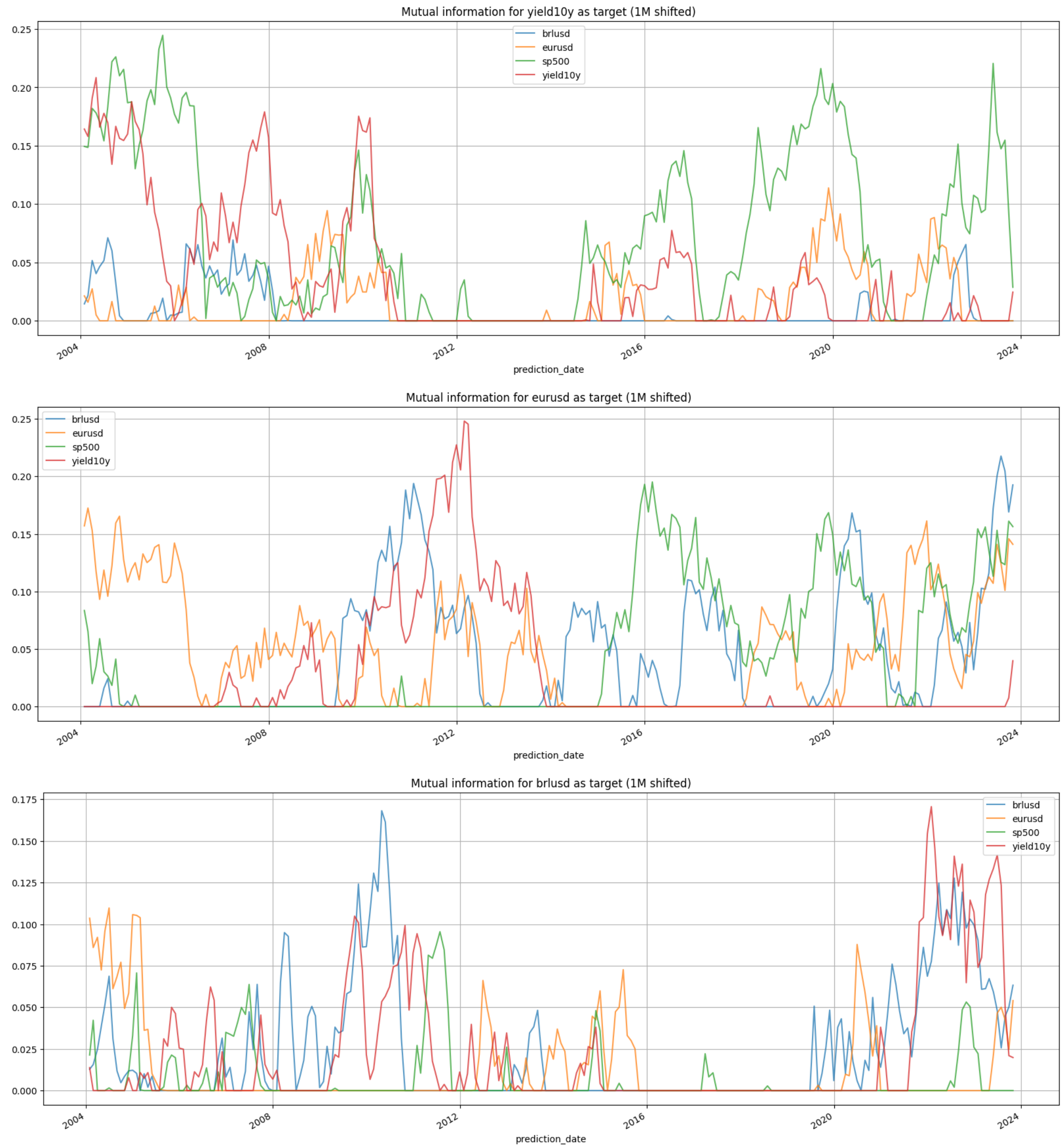
We aim to investigate whether any of the four %change series (*sp500*, *yield10y*, *brlusd*, *eurusd*) holds leading information over another. While one possible approach involves predicting a series using another, forecasting returns is typically challenging, both in terms of predicting the sign and magnitude. Consequently, employing such an approach would likely yield poor out-of-sample performance and a low R2, especially when utilizing a mono-factorial linear model.

To address this question, we have chosen to utilize pairwise mutual information, a metric that measures the mutual dependence between two random variables. Specifically, it gauges the amount of information acquired about one random variable by observing another. This technique is commonly employed to assess whether a relationship exists between two variables, not restricted to linear relationships, as seen in correlations. A mutual information value of 0 or close to 0 indicates no relationship, while a positive value signifies an existing relationship.

For this analysis, we computed the 1-month %change of the series as our set of features. We opted for a monthly timeframe to filter out noise in daily data, providing a stronger signal on the direction and magnitude of the change. The target was derived by shifting our series backward by 1 timestamp, allowing us to calculate mutual information between a feature at time t and a target variable at time t+1 i.e. we compute $MI(target_{t+1}, feature_t)$.

We performed this exercise using a rolling window spanning 5 years of monthly data, and the results are depicted in the plots below.





Pairwise Mutual Information of each target variable vs. the 1-month %change of *sp500*, *yield10y*, *eurusd*, *brlusd*

As illustrated in the plots, the connections among variables demonstrate dynamic patterns over time, marked by periods of strong interdependence and intervals where no apparent relationships are evident.

Regarding the *sp500*, notable variables include the lagged version of *sp500* itself during the years 2010 to 2014. *yield10y* appears to exert significant influence from 2017 to 2021, with diminished significance in other periods. Exchange rates play a minor role, mainly between 2000 and 2010.

In the context of *yield10y*, *sp500* exhibits a significant relationship, with peaks and troughs in the entire period, except for the years spanning from 2010 to 2014. The lagged version of *yield10y* itself seems to have a notable relation, especially in the years before 2010.
eurusd is primarily associated with *brlusd*, *yield10y*, and its lagged values in the period 2000-2014. After 2016, *yield10y* loses its influence, while *sp500* becomes relevant alongside *brlusd*.
For *brlusd*, all variables appear to have a relationship with the target up to 2016. Subsequently, there is no relationship with any of them until 2020 when *brlusd* and *yield10y* emerge as strong predictors.

It is crucial to note that achieving a high mutual information score alone does not guarantee precise inference of future feature values. To achieve this, it is important to identify a broad set of features that can explain the target when others lose significance, especially for series that may cease to impact during certain periods. The model must be well calibrated, and factors such as prediction horizon and target selection significantly impact forecasts. While mutual information helps identify relationships, leveraging this information is challenging due to potential non-linear relationships and changes in correlation signs over time. Nonetheless, this exercise is valuable for detecting relations and, more importantly, uncovering spurious relationships between variables that may not be apparent through simple charts or Pearson correlations.