Instituto BME

A logo with blue and grey letters

Description automatically generated

Master of Science Thesis

Optimizing Global Portfolio Weights using Artificial Intelligence and Macroeconomic Analysis

Master in Artificial Intelligence and Quantum Computing Applied to Financial Markets, 11th edition (mIA-X)

Lukasz Tadeusz Harazin

July 2024

Table of Contents

[1. Introduction 5](#_Toc166683198)

[2. Background 6](#_Toc166683199)

[2.1 Basic portfolio creation with global diversification 6](#_Toc166683202)

[2.2 Global portfolio with index funds and ETFs 6](#_Toc166683203)

[2.3 MSCI ACWI 7](#_Toc166683204)

[2.4 Modern Portfolio Theory 8](#_Toc166683205)

[2.4.1 Efficient Frontier and Optimal Portfolios 8](#_Toc166683206)

[2.4.2 Solving Optimal Portfolio using Quadratic Programming 9](#_Toc166683207)

[2.4.3 Markowitz’s Curse 9](#_Toc166683208)

[2.5 Black-Litterman Model 9](#_Toc166683209)

[2.6 Hierarchical Risk Parity 10](#_Toc166683210)

[2.7 Covariance Matrix Denoising 10](#_Toc166683211)

[2.8 Nested Clustered Optimization 11](#_Toc166683212)

[3. Methodology 12](#_Toc166683213)

[3.1 Data sources 12](#_Toc166683215)

[3.1.1 Organisation for Economic Co-operation and Development (OECD) 12](#_Toc166683216)

[3.1.2 World Bank (WB) 12](#_Toc166683217)

[3.1.3 Bank of International Settlements (BIS) 13](#_Toc166683218)

[3.1.4 Yahoo Finance 13](#_Toc166683219)

[3.1.5 Investing.com 13](#_Toc166683220)

[3.1.6 Morgan Stanley Capital International (MSCI) 14](#_Toc166683221)

[3.1.7 Other sources 14](#_Toc166683222)

[3.2 Macroeconomic Data 14](#_Toc166683223)

[3.2.1 Full list of economic indicators 15](#_Toc166683224)

[3.2.2 Key Economic Indicators 17](#_Toc166683225)

[3.2.3 Additional Economic Indicators 18](#_Toc166683226)

[3.3 Benchmark 19](#_Toc166683227)

[3.3.1 MSCI ACWI inclusions and exclusions 20](#_Toc166683228)

[3.3.2 MSCI ACWI Country Weights Changes 21](#_Toc166683229)

[3.3.3 Countries selection 23](#_Toc166683230)

[3.4 Investment Instruments 23](#_Toc166683231)

[3.4.1 Selected ETF Data 23](#_Toc166683232)

[3.4.2 Synthetic ETF Prices 24](#_Toc166683233)

[3.5 Data Collection and Storage 25](#_Toc166683234)

[3.6 Data Preparation 27](#_Toc166683235)

[3.6.1 Filling missing values 27](#_Toc166683236)

[3.6.2 Calculating available indicators data 29](#_Toc166683237)

[3.6.3 Formatting input data 30](#_Toc166683238)

[3.6.4 Splitting training data 31](#_Toc166683239)

[4. Presentation of work 32](#_Toc166683240)

[5. Conclusions 33](#_Toc166683241)

[6. References 34](#_Toc166683242)

Table of Figures

Figure 1Countries included in MSCI AWCI index as of March 2024. Source: MSCI 7

Figure 2 Efficient frontier with Capital Market Line 8

Figure 3 MSCI ACWI Complete Geographic Breakdown (Source www.msci.com/acwi) 20

Figure 4 Initial additions to MSCI Emerging Markets Index 21

Figure 5 Synthetic ETFs using different normalization and rescaling methods. 25

Figure 6 Synthetic ETF prices using reverse returns. 25

Figure 7 Cloud infrastructure diagram for data collection and data storage. 26

# Introduction

The pursuit of optimal portfolio diversification and risk-adjusted returns has long been a cornerstone of investment management. Traditional methods often rely on Modern Portfolio Theory (MPT) and historical data analysis. However, these approaches can be limited by static assumptions and the inherent difficulty of accurately forecasting future market behaviour.

This thesis explores the potential of Artificial Intelligence (AI) and macroeconomic analysis to create a more dynamic and data-driven approach to global portfolio weight optimization. By leveraging the power of AI techniques to identify complex patterns and relationships within vast datasets, I aim to develop an investment algorithm that can adapt to evolving market conditions and generate superior portfolio management.

This research will delve into various AI techniques, such as machine learning methods of Random Forest, XG Boost and Learning to Rate as well as different neural network models. It will validate their applicability in the context of portfolio optimization. Moreover, it will examine the role of macroeconomic analysis in shaping investment strategies and how it can be integrated with AI for portfolio management.

The main motivation for this thesis is to evaluate an alternative to passive investing, via index funds and Exchange-Traded Funds (ETFs), which has been rapidly gaining market share for both individual and institutional investors. A desired result for this investment algorithm would be to achieve both superior absolute returns and risk-adjusted returns in comparison to MSCI All Country World Index (ACWI), taken as a benchmark.

This thesis will try to offer a fresh perspective on global portfolio management in the age of AI. It is hoped to provide new insights and practical applications for the world of finance, helping to navigate the complexities of the global investment landscape.

# Background



## Basic portfolio creation with global diversification

Creating a globally diversified portfolio is a strategic approach to investing that can help mitigate risk and potentially enhance returns. The process begins by defining investment goals and constraints. A sound investment plan outlines the investor’s objectives and any significant constraints. Most investment objectives can be viewed in the context of a required rate of return (RRR). This is the return a portfolio would need to generate to bridge the gap between an investor’s current assets, any future cash flows, and the investment goals.

The next step is broad strategic allocation among the primary asset classes such as equities, fixed income, and cash. The allocation should be based on the investor's risk tolerance, investment horizon, and financial goals. This is followed by sub-asset allocation within classes, such as domestic and non-domestic securities or large-, mid- or small-capitalisation equities.

The investor then decides on the allocation to indexed or actively managed funds or both. Depending on the investment strategy, the investor can choose to invest in indexed or actively managed funds, or a combination of both.

The final step in the process is rebalancing. This is the process of realigning the weightings of the portfolio of assets. Rebalancing involves periodically buying or selling assets in a portfolio to maintain an original or desired level of asset allocation. (Vanguard UK Professional, 2024)

The goal of global diversification is not necessarily to boost performance but rather to spread risk across a variety of investments from different countries and sectors. This can potentially smooth out returns over time and help protect the portfolio from the volatility of any single investment or market.

## Global portfolio with index funds and ETFs

In recent years there has been a growing trend of passive investing through index funds and global ETFs for both individual and institutional investors. Index funds and Exchange-Traded Funds (ETFs) based on the MSCI World and MSCI All Country World Index (ACWI) have become key instruments for achieving global diversification in an investment portfolio.

The MSCI World Index tracks large and mid-cap representation across 23 developed markets countries, covering approximately 85% of the free float-adjusted market capitalization in each country. The MSCI ACWI goes a step further by including both developed and emerging markets, covering large and mid-cap representation across 23 developed and 24 emerging markets countries. This allows investors to gain exposure to a wide range of global equities through a single investment.

Moreover, these indices prioritize diversification by incorporating a broad range of sectors. The top sectors within the MSCI World and MSCI ACWI are information technology, financial services, and healthcare. This sector diversification can help to mitigate risk and potentially enhance returns.

Therefore, investing in index funds and ETFs based on the MSCI World and MSCI ACWI indices can be an effective strategy for achieving global diversification. It allows investors to spread their investments across different countries and sectors, thereby reducing the risk associated with any single investment or market.

## MSCI ACWI

The MSCI All Country World Index (ACWI) is a stock index designed to provide a broad measure of equity-market performance throughout the world. It is maintained by Morgan Stanley Capital International (MSCI) and comprises the stocks of nearly 3,000 companies from 23 developed countries and 24 emerging markets (Mitchell, 2024)

The MSCI ACWI captures large and mid-cap representation across these countries, covering approximately 85% of the global investable equity opportunity set. The index is used as a benchmark for global equity funds and as a guide to asset allocation. Approximately $4.6 trillion in assets are benchmarked to the index as of December 31, 2023.

Along with MSCI World and FTSE All World Index it is the most selected index for global portfolios. As of March 2024, it covers following markets:

A table with different countries/regions

Description automatically generated

Figure 1Countries included in MSCI AWCI index as of March 2024. Source: MSCI

## Modern Portfolio Theory

Traditional way to optimize portfolio has been Modern Portfolio Theory (MPT), also known as mean-variance analysis. It is a fundamental framework used by investors and financial professionals to construct portfolios that aim to maximize expected returns while managing risk. Developed by American economist Harry Markowitz in 1952, MPT revolutionized portfolio management by introducing rigorous mathematical principles.

MPT recognizes that diversification is essential for managing risk. By combining assets with different risk profiles, investors can achieve better risk-adjusted returns. The central idea is that owning a mix of assets—such as stocks, bonds, and real estate—reduces the impact of adverse events affecting any single investment.

During a portfolio construction MPT evaluates investments within the context of an entire portfolio rather than in isolation. The expected return of a portfolio is a weighted average of the expected returns of its individual assets. The key is to find an optimal combination of assets that balances risk and return. Moreover, an important assumption is that investors are risk-averse, meaning they prefer less risk for a given level of return. Utility theory quantifies an investor's preferences and trade-offs between risk and reward.

The variance of return or its transformation, the standard deviation, is used as a measure of risk, because it is tractable when assets are combined into portfolios.

### Efficient Frontier and Optimal Portfolios

The efficient frontier is a curve that represents the best possible trade-offs between risk and return. Portfolios lying on the efficient frontier offer the highest expected return for a given level of risk. The optimal portfolio is the point on the efficient frontier that aligns with an investor's risk tolerance. Visual representation is presented in the figure below (Lindquist, 2022)

A diagram of a financial market

Description automatically generated with medium confidence

Figure 2 Efficient frontier with Capital Market Line

Capital Market Line (CML) extends the efficient frontier by incorporating a risk-free asset (such as government bonds). The risk-free asset allows investors to create a risk-free portfolio by combining it with risky assets. The tangent point between the CML and the efficient frontier identifies the optimal risky portfolio.

### Solving Optimal Portfolio using Quadratic Programming

Quadratic optimization is a type of mathematical optimization that deals with quadratic functions. In the context of Modern Portfolio Theory (MPT), it plays a crucial role in determining the optimal allocation of assets in a portfolio.

Optimization can be done by minimizing variance to specific level of expected return or by maximizing expected return for a given variance level. Quadratic programming allows us to add various restrictions when it comes to number or assets, allowed weights or even minimal rebalancing when transitioning from one period to the next.

### Markowitz’s Curse

In his book ‘Advances in Financial Machine Learning’ Marcos Lopez de Prado points out various drawbacks of Modern Portfolio Theory with the main one being its instability. (Prado M. L., Advances in Financial Machine Learning, 2018) With more correlated instruments, we need more diversification and yet we might receive unstable results. In that case, the benefits of diversification are often offset by estimation errors. As a possible solution to that issue, the author proposes his own method known as Hierarchical Risk Parity, described later on.

## Black-Litterman Model

The Black-Litterman model is a mathematical model for portfolio allocation developed by Fischer Black and Robert Litterman at Goldman Sachs in 1990. It seeks to overcome problems that institutional investors have encountered in applying modern portfolio theory (MPT) in practice.

The model begins with modern portfolio theory and adds investor views of expected returns. The MPT model relies solely on historical market data and assumes that past expected returns will continue into the future. The Black-Litterman model enhances portfolio construction by incorporating investor expectations and adjusting the allocation based on those views.

One of the limitations of MPT is that it assumes that past expected returns will continue in the future. The Black-Litterman model incorporates observed market data along with an investor's projections of future expected returns based on models like the capital asset pricing model (CAPM) or other. The model essentially modifies the default MPT allocation by considering expectations of future performance.

The Black-Litterman model was created to develop "a quantitative and disciplined approach to structuring international bond portfolios in a manner consistent with the portfolio manager’s unique view of markets". It continues to be used and is recognized as a key mathematical tool in the investment world. The goal of the Black-Litterman model is to improve MPT-generated asset allocations by incorporating opinions on future outlook.

However, that model must be used with great care as it can introduce bias when investor’s views are inaccurate or based on false assumption. In that sense it doesn’t promise or guarantee optimal portfolio when working on poor assumptions. (Kenton, 2023)

## Hierarchical Risk Parity

Hierarchical Risk Parity (HRP) is a portfolio optimization strategy introduced by Marcos Lopez de Prado which addresses three major concerns of quadratic optimizers and Markowitz’s Critical Line Algorithm (CLA). It applies modern graph theory and machine learning techniques to build a diversified portfolio based on the information retrieved from the covariance matrix. As author proves in his paper (Prado M. L., Building Diversified Portfolios that Outperform Out-of-Sample, 2016), it delivers less risky portfolios out-of-sample compared to other traditional risk-parity methods.

The first step in the HRP method is Hierarchical Tree Clustering. This involves breaking down the assets into hierarchical clusters. Hierarchical clustering algorithms uncover a hierarchical structure of the considered investment universe, resulting in a tree-based representation.

Following the clustering of securities, the covariance matrix is reordered in a process known as Quasi-Diagonalization. This reorganization places similar stocks close to each other, ensuring that similar assets are grouped together.

The final step in the HRP method is Recursive Bisection. This involves assigning weights to each asset in the portfolio. The allocation is done in such a way that equities only have to compete with similar equities for spots in the portfolio.

Hierarchical Risk Parity is a robust method of portfolio construction, allowing user to introduce constraints and restructure the tree structure without compromising the algorithm’s search.

## Covariance Matrix Denoising

In the context of financial applications, covariance matrices are often numerically ill-conditioned, because of few independent observations used to calculate large number of features. That leads to classical portfolio optimization solutions often underperforming the naïve allocation out-of-sample. Matrix denoising, introduced by Marcos Lopez de Prado in Machine Learning for Asset Managers (Prado M. M., 2020), is used to correct the covariance matrix and reduce its instability.

This technique is based on linear algebra, particularly eigenvalues and eigenvectors. In practice, it aims at removing market components present in the correlation matrix, used in portfolio optimization. That magnifies more subtle signals hiding under the general market trends. According to the author, matrix denoising is a highly recommended step before any optimization operations.

Matrix denoising has several practical implications in portfolio management. It can significantly improve investors’ returns by providing more accurate risk assessments, enabling robust portfolio construction, and enhancing simulation and prediction capabilities.

## Nested Clustered Optimization

Nested Clustered Optimization (NCO) is another portfolio optimization method developed by Marcos Lopez de Prado (Prado M. M., 2020). It applies hierarchy logic and k-means clustering to the Mean-Variance Optimization problem, which suffers from instability when the covariance matrix is large. The algorithm consists of several steps:

* Denoising the Covariance Matrix: The covariance matrix of the outcomes is taken as an input.
* Choosing the Optimal Number of Clusters: The optimal number of clusters is determined using k-means algorithm or one of its variants.
* Clustering the Correlation Matrix: The covariance matrix is clustered into subsets of highly correlated variables.
* Finding Intra-Cluster Weights: Computes optimal intracluster allocations by using for example minimum variance allocation.
* Finding Inter-Cluster Weights: The optimal weights allocation for the reduced covariance matrix is computed by minimizing portfolio volatility.
* Computing Final Allocations: The final allocations are computed as a dot-product of the allocations between the clusters and inside the clusters.

The NCO algorithm can also maximize the Sharpe ratio by taking the expected returns as input. The purpose of the NCO algorithm is to introduce hierarchy logic to the portfolio optimization problem, making it more robust and efficient. By containing the numerical instability at each level of the tree, it reduces instability propagation to parent cluster or the rest of the correlation matrix. Moreover, it can be utilized in combination with other methods like Black-Litterman or constrained optimization.

# Methodology



## Data sources

In the course of conducting research for my master’s thesis, I have exclusively utilized open databases accessible via the internet. While these sources may not match the quality of paid data providers such as the Bloomberg Terminal, which is considered the gold standard, they nonetheless provide adequate data coverage to facilitate a detailed macroeconomic analysis.

There are various international organizations collecting broad range of macroeconomic statistics which are publicly available. Main organizations with open datasets are World Bank (WB), Organisation for Economic Co-operation and Development (OECD), International Monetary Fund (IMF), United Nations (UN) and Bank for International Settlements (BIS).

Additionally, there are multiple financial websites that offer a variety of resources for investors and traders. The most popular portals providing market data and economic indicators are Investing.com, Yahoo Finance, Trading Economics and FX Empire.

After extensive analysis of all available data sources, I have selected those with the highest data coverage for all countries selected for global portfolio. In the subsequent sections, I will provide a detailed description of all the data sources utilized in my algorithm.

### Organisation for Economic Co-operation and Development (OECD)

The Organisation for Economic Co-operation and Development (OECD) compiles a comprehensive collection of economic, social, and environmental datasets encompassing member countries and selected non-member economies. These datasets include:

* National accounts, trade, labour, and productivity statistics
* Science and technology innovation indicators
* Entrepreneurship and formation data
* Environmental-economic accounting information
* Development resource flows and official development assistance (ODA) data

In my macroeconomic analysis, I have used a wide range of monthly and quarterly economic indicators. It has the best data coverage for most countries included in MSCI ACWI index. Data can be downloaded using csv exports or developer’s API. In my algorithm, I have managed to use multiple indicators from OECD.Stat website reaching back to 1999.

### World Bank (WB)

World Bank Data is a comprehensive resource for global development data. It offers data by country or indicator, databases, pre-formatted tables, reports, and other resources. It includes tools for data analysis and visualization, access to microdata, international debt statistics, and development indicators. It also provides information on World Bank’s finances and lending projects.

For my analysis, it includes even more countries than OECD. Unfortunately, it provides mainly yearly data with significant delay. That limits its use in terms of monthly portfolio rebalancing. Data can be exported using csv files or third-party python libraries providing an easy access to all indicators formatted automatically to pandas’ data frames.

### Bank of International Settlements (BIS)

The BIS Data Portal provides global financial statistics. It includes data on international banking activity, debt securities, credit to non-financial sectors, credit-to-GDP gaps, debt service ratios, global liquidity indicators, exchange-traded derivatives, OTC derivatives, triennial survey data, and residential property prices. All data can be downloaded in CSV or SDMX format. Moreover, databases can be accessed via REST API providing data in XML format.

In my work I have used its data for central bank rates. I haven’t found any other place with such a complete daily data for interest rates for almost all central banks in both developed and emerging economies. Data is updated weekly what provided very precise and up to date information.

### Yahoo Finance

Yahoo Finance is a financial website that provides investors with:

* Market Data: Real-time stock quotes, news, and analysis.
* Investment Tools: Portfolio management tools and research resources.
* Financial News: Up-to-date headlines and insights on global markets.

Data can be easily downloaded using open-source Python library which reads real time market data from publicly available APIs. I have used this library to read historical prices for ETFs, stock indices and currency rates.

### Investing.com

Investing.com is a leading financial website and platform that offers a variety of tools and resources for investors and traders. It offers:

* Real-time market data: Stocks, bonds, commodities, currencies, futures, options across 70+ exchanges.
* Analysis & news: Articles on market trends, company performance, and economic events.
* Investment tools: Economic calendar, earnings calendar, technical analysis tools, portfolio tracker.

In my work, its economic calendar has provided a great value. While it has slightly worse data coverage than OECD dataset, it does provide report date time for all indicators. That is immensely important for proper back testing because we know exactly what data was available at a certain point in time. While other databases only hold final revision for given indicators like GDP values, with exact investing calendar we can read all values for subsequent revisions and know on which days they were released. Economic calendar cannot be easily downloaded, so I have used techniques of web scrapping using Selenium library. I have managed to download over 320 thousand data points for 50 countries in MSCI ACWI index dating back to 1999.

Apart from economic calendar, I have also downloaded csv data for missing currency rates and selected stock indices that aren’t available on Yahoo Finance. Moreover, it has complete historical data for Manufacturing PMI indicators, hard to find anywhere else, which are very useful to calculate economic cycles.

### Morgan Stanley Capital International (MSCI)

Morgan Stanley Capital International maintains a family of stock market indexes which are widely followed by investors around the world. MSCI indexes are widely used benchmarks for global stock markets. They track different segments (like developed, emerging or country-specific) and by market cap (where bigger companies influence more). This allows investors to see how their portfolios perform compared to a specific market segment.

In my algorithm, I have used MSCI indexes to both obtain MSCI ACWI benchmark and also fill missing returns on selected exchange traded funds where Yahoo Finance doesn’t have complete data.

### Other sources

Apart from data sources listed above, I have explored other datasets and web portals:

* **The International Monetary Fund (IMF)** Data provides comprehensive economic, financial, and socio-demographic statistics. It includes data on direct investment, climate transition, greenhouse gas emissions, world economic outlook, international finance, global financial stability, fiscal monitor, and exchange rates. It covers many similar indicators to OECD. However, it has worse historical data coverage and more missing values for old data. In the end I opted to use OECD datasets instead.
* **Trading Economics** is a platform that provides similar features to Investing.com but is better protected against automatic web scrapping and offers paid subscriptions to access its data.
* **FX Empire** is another data provider resembling Investing.com and Trading Economics. Unfortunately, its interactive website makes it very difficult to download data using web scrapping.
* **FRED** (Federal Reserve Economic Data) is a trusted source for economic data since 1991. It provides access to over 824,000 US and international time series. It does provide an excellent coverage for US economy, but lacks international indicators, referencing data from OECD and other public database.
* **EBS Statistics** provide comprehensive data that supports all aspects of the ECB’s work, including monetary policy, financial stability, and banking supervision. It covers European economies in great details but covers few international indicators required to optimise global portfolio.

## Macroeconomic Data

To perform macroeconomic analysis, we need to collect a wide range of economic indicators. In my research I have downloaded 70 indicators and divided them into 9 distinct categories of related measures. That has made it easier to get an overall view and select the best indicators for an investment algorithm.

Each of these categories provides a different perspective on the health of the economy and can be used together to get a comprehensive understanding of economic conditions. Below I provide a brief description of each category:

1. ***Stock Market***: This refers to the collection of markets and exchanges data. Moreover, provides an overview of stock market for each country with its relevant metrics.
2. ***GDP***: It refers to various measures of Gross Domestic Product when it comes to total value, growth rate or value per capita.
3. ***Labour***: Labour market indicators provide an overview of the economic health of the employment sector. They include metrics such as the unemployment rate, employment rate and population.
4. ***Prices***: This provides various measures of price changes over time. For example, an indicator like CPI examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care.
5. ***Money***: This refers to various indicators like money supply, interest rates, and credit availability. These indicators can influence spending and investment activities in an economy.
6. ***Trade***: Trade indicators include metrics related to imports, exports, trade balance, and terms of trade. They provide insights into a country's competitiveness, the demand for its goods and services, and its economic ties with other countries.
7. ***Government***: Government indicators include government spending, budget deficits, and public debt. These indicators can show how government policy is affecting the economy.
8. ***Business***: Business indicators include measures of business confidence, industrial production, and manufacturing output. They provide insights into the health of the business sector and can be leading indicators for the overall economy.
9. ***Customer***: This refers to measures of consumer confidence, which provide an indication of consumers' attitudes about the health of the economy and their willingness to spend.

### Full list of economic indicators

Full list of economic indicators downloaded for analysis is included in the table below. There are 5 columns to describe each indicator:

* Indicator – full name of the indicator.
* Source – data source from which the indicator has been downloaded.
* Freq – frequency of the data. Can be D-daily, M-monthly, Q-quarterly, or Y-yearly.
* Measure – additional description of applied measure.
* Data coverage – a percentage of available data for 50 countries for years 1999 to 2024.

Table 1 Full list of macroeconomic indicators

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Indicator** | **Source** | **Freq** | **Measure** | **Data Coverage** |
| **1. Stock Market** | | | | |
| Stock Indices | Yahoo Finance, Investing.com | D | Index |  |
| Currency Rates | Yahoo Finance, Investing.com | D | Exchange rates |  |
| ETFs in USD | Yahoo Finance | D | US Dollars |  |
| ETFs in EUR | Yahoo Finance | D | EUR |  |
| MSCI Indices | MSCI | D | Index |  |
| Stock Market Cap | World Bank | Y | Current US Dollars | 76.2 |
| Stock Market Cap Pct of GDP | World Bank | Y | Pct of GDP | 76.2 |
| Listed Domestic Companies Total | World Bank | Y | Count | 80.5 |
| Stocks Traded Total Value | World Bank | Y | Current US Dollars | 78.2 |
| Stocks Traded Total Value Pct of GDP | World Bank | Y | Pct of GDP | 78.2 |
| **2. GDP** | | | | |
| GDP Annual Growth Rate | OECD | Q | Growth YoY | 73.9 |
| GDP Growth Rate | OECD | Q | Growth QoQ | 73.1 |
| GDP Per Capita | OECD | Q | Per Head, US dollars (2015) | 60.1 |
| GDP Current Prices US Dollars | OECD | Y | Current US Dollar | 77 |
| GDP Current Prices PPP | OECD | Y | Current US Dollar | 77 |
| GDP | World Bank | Y | Current US Dollar | 98 |
| GDP Per Capita | World Bank | Y | Current US Dollar | 98 |
| GDP (QoQ) | Investing.com | Q | Growth QoQ | 36.2 |
| GDP (YoY) | Investing.com | Q | Growth YoY | 39.6 |
| **3. Labour** | | | | |
| Unemployment Rate | OECD | M | Level | 53.6 |
| Unemployment Rate | OECD | Q | Level | 58.2 |
| Population | World Bank | Y | Count | 98 |
| Unemployment Rate | Investing.com | M | Level | 53.8 |
| **4. Prices** | | | | |
| Inflation Rate | OECD | M | Growth YoY | 70.6 |
| Inflation Rate | OECD | Q | Growth YoY | 74.2 |
| Inflation Rate MoM | OECD | M | Growth MoM | 70 |
| CPI | OECD | M | Index 2015 = 100 | 69.5 |
| PPI Manufacture of food products | OECD | M | Index 2015 = 100 | 36.8 |
| PPI Manufacturing | OECD | M | Index 2015 = 100 | 47.7 |
| CPI (MoM) - Inflation Rate MoM | Investing.com | M | Growth MoM | 61.8 |
| CPI (YoY) - Inflation Rate | Investing.com | M | Growth YoY | 80 |
| PPI (MoM) | Investing.com | M | Growth MoM | 26.2 |
| PPI (YoY) | Investing.com | M | Growth YoY | 35.6 |
| **5. Money** | | | | |
| Central Bank Rates | BIS | D | Level |  |
| Overnight Interbank Rate | OECD | M | Level | 73.7 |
| Short Term Interest Rate | OECD | M | Level | 68.4 |
| Long Term Interest Rate | OECD | M | Level | 66.4 |
| Narrow Money M1 | OECD | M | National Currency | 49.5 |
| Broad Money M3 | OECD | M | National Currency | 49.5 |
| **6. Trade** | | | | |
| Current Account to GDP | OECD | Q | Pct of GDP | 70 |
| Export of goods and services | OECD | Q | Growth YoY | 70.3 |
| Import of goods and services | OECD | Q | Growth YoY | 70.3 |
| Export - Value (goods) | OECD | M | US Dollars, monthly level | 74.2 |
| Import - Value (goods) | OECD | M | US Dollars, monthly level | 74.4 |
| Net Trade - Value (goods) | OECD | M | US Dollars, monthly level | 74.2 |
| Current Account Pct of GDP | World Bank | Y | Pct of GDP | 91.4 |
| Current Account | World Bank | Y | Current US Dollars | 93.3 |
| Trade Balance | Investing.com | M | Local Currency, monthly level | 64.7 |
| **7. Government** | | | | |
| Government Debt to GDP | OECD | Y | Pct of GDP | 54.8 |
| Total Government Expenditure | OECD | Y | Current prices, local currency | 61.9 |
| Total Government Revenue | OECD | Y | Current prices, local currency | 61.9 |
| Government Budget | OECD | Y | Pct of GDP | 77 |
| Government Expense | World Bank | Y | Pct of GDP | 77.8 |
| **8. Business** | | | | |
| OECD Business Confidence Indicator | OECD | M | Index, Amplitude adjusted | 72.7 |
| OECD Composite Leading Indicators | OECD | M | Index, Amplitude adjusted | 34 |
| Industrial Production | OECD | M | Growth YoY | 60.6 |
| Total Manufacturing | OECD | M | Index 2015=100, s.a. | 66.7 |
| Total Industry ex Construction | OECD | M | Index 2015=100, s.a. | 60.3 |
| Total Construction | OECD | M | Index 2015=100, s.a. | 44.7 |
| Changes in Inventories | OECD | Q | Current prices, local currency | 62.3 |
| Manufacturing PMI | Investing.com | M | Index | 31.9 |
| Services PMI | Investing.com | M | Index | 13.8 |
| Industrial Production | Investing.com | M | Growth YoY | 42.3 |
| **9. Consumer** | | | | |
| OECD Consumer Confidence Indicator | OECD | M | Index, Amplitude adjusted | 66.1 |
| Private Consumption | OECD | Q | Growth YoY | 67.7 |
| Total Retail Sales Value | OECD | M | Index 2015=100, s.a. | 52.7 |
| Passenger Car Registration | OECD | M | Index 2015=100, s.a. | 48.4 |
| Permits Issued (Residential Buildings) | OECD | M | Index 2015=100, s.a. | 42.6 |
| Retail Sales | Investing.com | M | Growth YoY | 46.8 |
| Consumer Confidence | Investing.com | M | Index | 33.9 |

### Key Economic Indicators

Out of all indicators I have selected 6 most relevant measures to include in all models. Each of these indicators provides a different perspective on the health of the economy and can be used together to get a comprehensive understanding of economic conditions. Moreover, they are widely commented in both economic and general press after each release and are closely followed by traders and investors alike. Below are brief descriptions of each economic indicator:

1. ***GDP Annual Growth Rate***: This measures the percentage change in a country's Gross Domestic Product (GDP) over a year. It indicates whether an economy is expanding or contracting, serving as a critical indicator of its performance.
2. ***GDP Growth Rate****:* Like the annual growth rate, the GDP Growth Rate measures economic growth by comparing GDP from one period to the next, in my study over following quarters. It's an indicator of the health of an economy and helps policymakers adjust fiscal and monetary policy to achieve economic objectives.
3. ***Unemployment Rate***: This is the percentage of unemployed individuals in an economy among individuals currently in the labour force. It's calculated as the number of unemployed individuals divided by the total labour force, multiplied by 100. It's a key economic indicator as it signals the ability (or inability) of workers to obtain gainful work and contribute to the productive output of the economy.
4. ***Inflation Rate***: The Inflation Rate is the percentage increase in price for a basket of goods and services for a particular period. It's used to measure the general increase in the cost of goods and services. It's contrasted by deflation, which refers to the appreciation of the currency and leads to decreased prices of commodities.
5. ***Inflation Rate MoM (Month over Month)***: This measures the change in the Inflation Rate from one month to the next. It provides a more granular view of inflation trends and can be useful for identifying short-term changes in the rate of inflation.
6. ***Manufacturing PMI (Purchasing Managers' Index):*** The PMI is an index of the prevailing direction of economic trends in the manufacturing sector. It consists of a diffusion index that summarizes whether market conditions are expanding, staying the same, or contracting as viewed by purchasing managers. It's based on five major survey areas—each of which is weighted equally: New Orders, Inventory Levels, Production, Supplier Deliveries, and Employment.

### Additional Economic Indicators

For a more comprehensive macroeconomic analysis, I have selected additional nine indicators that predominantly exhibit high data coverage and a monthly data frequency.

Firstly, I have excluded yearly indicators. Indicators with an annual frequency tend to offer limited value for an investment algorithm that operates with monthly rebalancing. This is primarily due to the substantial delay, often up to two years, in their release. Consequently, when these are combined with monthly indicators, they frequently retain the same value over a six to eight-month rolling window.

Secondly, indicators with low data coverage, their utility is inherently restricted. The incorporation of such indicators would necessitate a significant amount of estimated or interpolated data, which could potentially introduce considerable noise into the model. Thus, their use is generally not recommended for precision-oriented tasks such as this.

Additional indicators selected for extra analysis are:

1. ***Producer Price Index (PPI):*** This is a measure of the average change over time in the selling prices received by domestic producers for their output. It is a measure of inflation at the wholesale level.
2. ***Central Bank Rate:*** This is the interest rate set by the central bank of a country that commercial banks are charged to borrow money. It is a key tool used by central banks to implement monetary policy and control inflation.
3. ***Short Term Interest Rate:*** This refers to the interest rates on financial instruments that mature in less than one year. They are generally averages of daily rates and are based on three-month money market rates where available.
4. ***Long Term Interest Rate:*** This refers to the interest rates on financial instruments that mature in ten years or more. These rates are mainly determined by the price charged by the lender, the risk from the borrower, and the fall in the capital value.
5. ***Current Account to GDP***: This ratio provides an indication of a country's level of international competitiveness. A positive current account balance indicates that the nation is a net lender to the rest of the world, while a negative current account balance indicates that it is a net borrower.
6. ***Total Manufacturing:*** This refers to the total output of the manufacturing sector in an economy. It includes the production of goods and services across all manufacturing industries.
7. ***Industrial Production:*** This is a measure of the real output of the manufacturing, mining, electric, and gas industries. It is a key economic indicator that measures the level of production in the industrial sector of an economy.
8. ***OECD Consumer Confidence Indicator (CCI):*** This indicator provides an indication of future developments of households’ consumption and saving, based upon answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment, and capability of savings.
9. ***Retail Sales***: Retail sales refer to the total amount of goods and services sold by retailers to consumers within a specific period¹. They are a vital economic indicator as they reflect consumer spending patterns, which account for a significant portion of overall economic activity. As a leading macroeconomic indicator, healthy retail sales figures typically elicit positive movements in equity markets.

## Benchmark

MSCI All Country World Index (ACWI) has been selected as a benchmark for all designed algorithms. A more detailed description of that index has already been provided in chapter 2.3. In this section I would like to focus on the weights of different countries, its inclusions, and exclusions from the index and finally a selection of the countries for AI models.

Exact geographical breakdown of the index by country and region can be found on MSCI ACWI official website (MSCI, 2024). The figure presented below shows percentage weights of all countries and their changes over the past 11 years.

A blue and red circle with text

Description automatically generated

Figure 3 MSCI ACWI Complete Geographic Breakdown (Source www.msci.com/acwi)

### MSCI ACWI inclusions and exclusions

To be able to perform back testing correctly we cannot use current list of countries included in the index because that would give an algorithm unfair advantage using survivorship bias. Instead, we need to investigate all inclusions and exclusions from the index over the years.

All 23 developed economies have been present in the index since 1999. The only change has been an upgrade for Israel, Portugal, and Greece from emerging economies to developed ones.

However, there has been significant changes to emerging markets over the last 25 years. During my research I found an analysis prepared by MSCI for the Norwegian Ministry of Finance from October 2019 provides detailed additions to MSCI Emerging Markets (MSCI, 2019). That paper includes initial additions to MSCI Emerging Markets Index since 1988 when the index was created.

A screenshot of a computer

Description automatically generated

Figure 4 Initial additions to MSCI Emerging Markets Index

Apart from these initial inclusions, there have been additional removals, additions, and reinstallations:

* Sri Lanka was removed in 2001.
* Venezuela was removed in 2006.
* Pakistan and Jordan were removed in 2009.
* Argentina was removed in 2010.
* United Arab Emirates and Qatar were added in 2014.
* Pakistan was re-instated in 2017.
* Argentina was re-instated in 2019.
* Saudi Arabia was added in 2019.
* Kuwait was added in 2020.
* Russia, Argentina, and Pakistan were removed in 2022.

### MSCI ACWI Country Weights Changes

Country weights are changing constantly due to market capitalization changes, addition or removal of constituents, quarterly rebalancing of the index, and changes in free float. In the investment algorithm, the exact weights within the index are not required, but approximate values are used in portfolio optimization used as a target. Their usage will be described in detail in section 3.6.

Historical weights for all countries aren’t publicly available on the MSCI website. However, during my research I have come across a study on emerging market portfolio strategies, where it lists composition of the MSCI ACWI in years 1987-2012. (Roberto Violi, 2018).

Combining weights from the research paper (Roberto Violi, 2018), MSCI report (MSCI, 2019) and MSCI geographical breakdown (MSCI, 2024), I have been able to approximate all country weights in period 1998 to 2024.

Table 2 MSCI ACWI Country Weights Changes 1998 - 2024

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| MSCI World: | 1998 | 2003 | 2008 | 2013 | 2021 | 2024 |
| United States | 46.56 | 53.99 | 41.80 | 47.45 | 58.60 | 62.60 |
| Japan | 11.26 | 8.44 | 8.60 | 7.50 | 6.20 | 6.10 |
| United Kingdom | 10.40 | 11.50 | 9.60 | 7.90 | 4.00 | 3.50 |
| Canada | 2.29 | 2.20 | 3.67 | 4.30 | 2.90 | 2.80 |
| France | 4.10 | 3.10 | 4.20 | 3.10 | 2.70 | 2.70 |
| Switzerland | 3.40 | 3.10 | 4.10 | 3.00 | 2.50 | 2.10 |
| Germany | 4.20 | 2.60 | 3.70 | 2.90 | 2.30 | 2.00 |
| Australia | 1.22 | 1.88 | 2.79 | 3.31 | 1.90 | 1.80 |
| Netherlands | 1.50 | 1.60 | 1.30 | 0.80 | 1.10 | 1.10 |
| Sweden | 1.20 | 1.20 | 1.30 | 1.20 | 1.20 | 0.84 |
| Spain | 1.20 | 1.50 | 1.90 | 0.96 | 0.61 | 0.62 |
| Hong Kong | 1.25 | 0.64 | 1.05 | 1.09 | 0.79 | 0.42 |
| Italy | 1.90 | 1.80 | 1.50 | 0.79 | 0.67 | 0.70 |
| Singapore | 0.41 | 0.33 | 0.48 | 0.71 | 0.32 | 0.34 |
| Denmark | 0.30 | 0.30 | 0.40 | 0.37 | 0.68 | 0.83 |
| Finland | 0.45 | 0.45 | 0.40 | 0.27 | 0.33 | 0.24 |
| Belgium | 0.30 | 0.50 | 0.50 | 0.49 | 0.27 | 0.24 |
| Norway | 0.40 | 0.40 | 0.60 | 0.34 | 0.24 | 0.20 |
| Israel | 0.17 | 0.13 | 0.24 | 0.21 | 0.26 | 0.25 |
| Ireland | 0.10 | 0.10 | 0.20 | 0.19 | 0.18 | 0.08 |
| New Zealand | 0.13 | 0.07 | 0.06 | 0.05 | 0.09 | 0.06 |
| Austria | 0.10 | 0.10 | 0.10 | 0.08 | 0.08 | 0.07 |
| Portugal | 0.10 | 0.10 | 0.10 | 0.06 | 0.05 | 0.05 |
| MSCI EM: |  |  |  |  |  |  |
| China | 0.03 | 0.26 | 1.80 | 2.23 | 3.80 | 2.40 |
| Taiwan | 1.14 | 0.51 | 1.12 | 1.34 | 1.80 | 1.90 |
| India | 0.40 | 0.20 | 0.94 | 0.79 | 1.40 | 2.00 |
| Korea | 0.20 | 0.86 | 1.62 | 1.94 | 1.70 | 1.30 |
| Brazil | 1.02 | 0.27 | 1.51 | 1.64 | 0.65 | 0.54 |
| Russia | 0.37 | 0.19 | 1.15 | 0.76 | 0.38 | 0.00 |
| South Africa | 0.67 | 0.56 | 0.76 | 1.01 | 0.44 | 0.29 |
| Mexico | 0.81 | 0.31 | 0.51 | 0.63 | 0.23 | 0.27 |
| Malaysia | 0.36 | 0.22 | 0.28 | 0.46 | 0.18 | 0.16 |
| Thailand | 0.10 | 0.07 | 0.15 | 0.27 | 0.22 | 0.18 |
| Indonesia | 0.11 | 0.04 | 0.19 | 0.36 | 0.14 | 0.19 |
| Türkiye | 0.20 | 0.05 | 0.19 | 0.21 | 0.05 | 0.10 |
| Poland | 0.03 | 0.05 | 0.19 | 0.17 | 0.10 | 0.11 |
| Chile | 0.25 | 0.06 | 0.13 | 0.24 | 0.06 | 0.05 |
| Argentina | 0.29 | 0.02 | 0.05 | 0.00 | 0.02 | 0.00 |
| Saudi Arabia | 0.00 | 0.00 | 0.00 | 0.00 | 0.36 | 0.44 |
| Greece | 0.17 | 0.00 | 0.00 | 0.00 | 0.03 | 0.05 |
| United Arab Emirates | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.12 |
| Qatar | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.08 |
| Kuwait | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.09 |
| Philippines | 0.09 | 0.02 | 0.06 | 0.12 | 0.07 | 0.07 |
| Peru | 0.08 | 0.02 | 0.07 | 0.08 | 0.02 | 0.03 |
| Hungary | 0.08 | 0.05 | 0.09 | 0.04 | 0.03 | 0.02 |
| Czechia | 0.06 | 0.02 | 0.09 | 0.04 | 0.01 | 0.01 |
| Egypt | 0.00 | 0.01 | 0.09 | 0.04 | 0.01 | 0.01 |
| Colombia | 0.06 | 0.00 | 0.03 | 0.16 | 0.02 | 0.01 |
| Pakistan | 0.05 | 0.01 | 0.02 | 0.00 | 0.01 | 0.00 |
| Jordan | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| Sri Lanka | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Venezuela | 0.10 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| Marocco | 0.00 | 0.01 | 0.03 | 0.01 | 0.00 | 0.00 |

### Countries selection

Over the past 25 years, the MSCI ACWI index has included 54 countries. However, incorporating all these countries into the portfolio would introduce excessive complexity to the model while providing minimal impact for countries with negligible weight.

Upon analysis of available macroeconomic data and historical ETF data, a decision has been made to use a 0.03% cutoff index weight to determine which countries would be included in the algorithm. This approach reduced the number of countries by half, to 27, while still covering over 98% of the total MSCI ACWI index. This effectively excluded all countries that only contribute a fraction of a percentage to the index and were also absent from the OECD database, thus lacking most of the macroeconomic indicators.

The model will include 18 developed economies and 9 emerging markets:

* MSCI World: United States, Japan, United Kingdom, Canada, France, Switzerland, Germany, Australia, Netherlands, Sweden, Spain, Hong Kong, Italy, Singapore, Denmark, Finland, Belgium, Norway
* MSCI EM: China, Taiwan, India, Korea, Brazil, Russia, South Africa, Mexico, Malaysia

## Investment Instruments

There exist various investment instruments that can be utilized to construct a global stock portfolio, representing markets in specific countries. Given the selection of 27 countries for the investment algorithm, exposure to the main stock index is necessary, as purchasing individual stocks for numerous companies is not feasible with monthly rebalancing. There are three primary options to replicate the returns of a stock market index:

1. Derivatives, particularly futures: These are primarily used for risk hedging. However, in the context of investment algorithms, they do not provide the required flexibility due to their fixed settlement day and high unit price (typically over 10,000 euros like for Mini Ibex).
2. Index Funds: These have substantial coverage for major economies, but there is a lack of funds for smaller countries. Furthermore, it can take up to five days to transfer funds from one fund to another during rebalancing, potentially resulting in lost gains due to being out of the market. Nonetheless, in certain countries, such as Spain and Poland, they offer tax advantages for retail investors, eliminating the need to pay capital gains tax when transferring money from one fund to another.
3. Exchange-Traded Funds (ETFs): These funds are traded on stock exchanges. They offer high liquidity and instant market orders, similar to regular stocks. They have been selected as the preferred option for the investment algorithms in this thesis.

### Selected ETF Data

In order to conduct the most reliable back-testing possible, the decision has been made to utilize historical data spanning from 1999 to the end of 2023. This encompasses a period of 25 years and nearly three full economic cycles. The inclusion of historical data from various stages of the cycle in both the training and testing phases is crucial to ensure the development of a reliable AI model that is not biased towards any specific part of the cycle.

Upon investigation of available Exchange-Traded Funds (ETFs), it was observed that the US markets offer the longest price history. This is due to the fact that ETFs emerged in the United States in the early 1990s, while in Europe, the first ETFs were introduced in 2005 and only became widely available after 2008. To obtain market data from 1999, ETFs from the Nasdaq stock exchanges have been selected, with BlackRock offering the broadest range for the selected countries and the highest liquidity.

Below I present a list of ETFs downloaded from Yahoo Finance and used in this research.

Table 3 Selected Exchange Trades Funds (ETFs)

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Symbol | Full Name | Year Created |
| United States | SPY | SPDR S&P 500 ETF Trust (SPY) | 1999 |
| Japan | EWJ | iShares MSCI Japan ETF | 1999 |
| United Kingdom | EWU | iShares MSCI United Kingdom ETF | 1999 |
| Canada | EWC | iShares MSCI Canada ETF | 1999 |
| France | EWQ | iShares MSCI France ETF | 1999 |
| Switzerland | EWL | iShares MSCI Switzerland ETF | 1999 |
| Germany | EWG | iShares MSCI Germany ETF | 1999 |
| Australia | EWA | iShares MSCI Australia ETF | 1999 |
| Netherlands | EWN | iShares MSCI Netherlands ETF | 1999 |
| Sweden | EWD | iShares MSCI Sweden ETF | 1999 |
| Spain | EWP | iShares MSCI Spain ETF | 1999 |
| Hong Kong | EWH | iShares MSCI Hong Kong ETF | 1999 |
| Italy | EWI | iShares MSCI Italy ETF | 1999 |
| Singapore | EWS | iShares MSCI Singapore ETF | 1999 |
| Denmark | EDEN | iShares MSCI Denmark ETF | 2012 |
| Finland | EFNL | iShares MSCI Finland ETF | 2012 |
| Belgium | EWK | iShares MSCI Belgium ETF | 1999 |
| Norway | NORW | iShares MSCI Norway ETF | 2012 |
| China | MCHI | iShares MSCI China ETF | 2011 |
| Taiwan | EWT | iShares MSCI Taiwan ETF | 2000 |
| India | INDA | iShares MSCI India ETF | 2012 |
| Korea | EWY | iShares MSCI South Korea ETF | 2000 |
| Brazil | EWZ | iShares MSCI Brazil ETF | 2000 |
| Russia | ERUS | iShares MSCI Russia ETF | 2010 |
| South Africa | EZA | iShares MSCI South Africa ETF | 2003 |
| Mexico | EWW | iShares MSCI Mexico ETF | 1999 |
| Malaysia | EWM | iShares MSCI Malaysia ETF | 1999 |

### Synthetic ETF Prices

Unfortunately, not all ETFs have full historic prices since 1999. Countries like Denmark, Finland or India has prices starting from 2012 when these funds were launched. To be able to use these instruments in my algorithm I had to recreate ETFs for missing periods to have full history for all selected countries. These missing periods can by recreated using MSCI indexes or main stock indexes from given country with currency exchange rate applied.

My first approach has been to extrapolate index data using either min-max and Z-score normalization and rescaling. However, generated prices have been distorted, not representing correctly returns from underlying indexes.

A graph showing different colored lines

Description automatically generated

Figure 5 Synthetic ETFs using different normalization and rescaling methods.

In some cases, like in the figure above, prices were reaching impossible, negative values. To prevent that from happening and more importantly achieve the same relative returns, I have used reverse returns from indexes. That would start from the first date available ETFs and going backwards till start of 1999, applying cumulative product for reverse returns multiplying by the first price. This method has proven much more reliable, resulting in exact representation of underlying indexes but converted to real ETF prices.

A graph showing a line of blue and orange lines

Description automatically generated

Figure 6 Synthetic ETF prices using reverse returns.

Using calculated synthetic ETFs together with the actual ones from Nasdaq stock exchange I have been able to calculate returns for all individual countries over the period of 25 years.

## Data Collection and Storage

In this research, due to large amounts of data for economic indicators, cloud computing solutions have been selected to help with the process of data collection and storage. They offer many advantages over traditional on-site solutions. Cloud computing enables rapid deployment, allowing developers to spin up or retire instances in seconds, thereby accelerating development processes. The inherent scalability and flexibility of cloud computing dynamically allocate resources based on workload, ensuring applications can handle sudden demand spikes effortlessly. In terms of data security, the risk of data loss due to hardware failures is mitigated by networked backups. Furthermore, the pay-as-you-go model of cloud computing ensures cost-effectiveness by charging only for the actual compute resources consumed during execution. Lastly, the accessibility of cloud-based applications and data from virtually any internet-connected device enhances the ease of.

In this project, I have selected solutions from Amazon Web Services (AWS). All market data and macroeconomic indicators are stored in Microsoft SQL Server, managed within the Amazon Relational Database Service.

Keeping all the data up to date is handled by six Lambda functions that periodically retrieve the latest data from all data sources via API calls or techniques of web scrapping. These functions, which are both scalable and cost-effective, operating only couple seconds every morning for data updated daily and weekly for less frequently updated indicators. This might prove beneficial for extending back-testing over time and for production algorithm deployment.

The final significant component of the cloud infrastructure is the Elastic Container Registry, utilized for updating and storing Docker images used to run Lambda functions. The use of containers has been particularly advantageous for web scraping tasks of the economic calendar from the Investing.com website, as it facilitated the installation of additional chrome drivers and the selenium library, while being isolated from other functions and system updates.

The figure below shows the general cloud architecture for data collection and data storage. It does not include all Event Briges, used to trigger remaining 5 Lambda functions as that would decrease the visibility of the diagram.

A diagram of software development

Description automatically generated

Figure 7 Cloud infrastructure diagram for data collection and data storage.

Source code for all Lambda functions is versioned in GitHub Repository. Deployment process is implemented using Terraform, an infrastructure as code tool, what allows to provision and destroy an entire cloud infrastructure automatically by executing a single bash script. That can be used in times when a given data source is not needed or must be updated due to API changes.

Different Lambda functions read various data sources:

* Economic Calendar Lambda – uses web scrapping techniques with Selenium to read all the indicators from the previous day. While Investing.com page has some protections against automatic selection of data range from the calendar, it does allow to select yesterday values. For that reason, the initial web scrapping for data all the way to 1999, was performed in semi-automatic way by selecting manually date range of several months, waiting for an entire period to load and then running a python function to read all presented indicators.
* Investing Stock Indices Lambda – uses web scrapping techniques with Selenium to read 8 stock indices from Investing.com web page that are not available in Yahoo Finance. They were used for synthetic ETFs data.
* BIS Lambda – reads central bank rates from Bank of International Settlements API. At the time of writing these rates were updated at the end of each week.
* Yahoo Finance Lambda – reads ETFs prices, stock indices values and currency rates from API using yfinance library. Can be extended to new instruments by simply adding new symbols to the database which will be included automatically in the next update.
* OECD Lambda – reads all indicators from official API and saves values to the database if any new records are available.
* World Bank Lambda – reads worlds bank yearly indicators from API using third-party wbgapi library.

## Data Preparation

Data preparation plays a pivotal role in the success of machine learning models. It involves several critical tasks, including data cleaning, feature engineering, and handling imbalanced datasets. By ensuring that the data is well-structured and relevant, we enhance the accuracy and reliability of our models. Properly prepared data serves as the foundation for robust and effective machine learning solutions.

In this research there are several steps required to prepare downloaded indicators into correct input data ready for ML training. First, missing values have to be filled while introducing minimal noise. Next, correct values must be selected for each rebalance date. That’s because we cannot use indicators released in later date in relation to a given training step. Finally, indicators must be normalized, and additionally they can be factorized to reduce dimensionality.

### Filling missing values

One of the principal problems with open data sources is much higher percentage of missing values in comparison to paid providers. In my research I have managed to combine values from Investing.com and OECD to reduce number of NAs for most key indicators available in both sources. The best example is an inflation rate where both sources have missing values for different countries and combining them has resulted in almost complete dataset.

Table 4 Missing values in data sources for Inflation Rate YoY

|  |  |  |  |
| --- | --- | --- | --- |
|  | Inflation Rate (Investing) | Inflation Rate (OECD) | Inflation Rate (Combined) |
| United States | 0 | 0 | 0 |
| Japan | 6 | 30 | 0 |
| United Kingdom | 1 | 0 | 0 |
| Canada | 2 | 0 | 0 |
| France | 230 | 0 | 0 |
| Switzerland | 0 | 0 | 0 |
| Germany | 0 | 0 | 0 |
| Australia | 257 | 200 | 157 |
| Netherlands | 0 | 0 | 0 |
| Sweden | 6 | 0 | 0 |
| Hong Kong | 1 | 300 | 1 |
| Spain | 0 | 0 | 0 |
| Italy | 2 | 0 | 0 |
| Singapore | 1 | 300 | 1 |
| Denmark | 0 | 0 | 0 |
| Finland | 144 | 0 | 0 |
| Belgium | 2 | 0 | 0 |
| Norway | 0 | 0 | 0 |
| China | 0 | 0 | 0 |
| Taiwan | 0 | 300 | 0 |
| India | 160 | 0 | 0 |
| Korea | 3 | 0 | 0 |
| Brazil | 0 | 0 | 0 |
| Russia | 1 | 21 | 0 |
| South Africa | 1 | 0 | 0 |
| Mexico | 0 | 0 | 0 |
| Malaysia | 1 | 300 | 1 |

Unfortunately for other key indicators there are still many missing values. Below is the summary of missing values count for key macroeconomic indicators.

Table 5 Missing values for Key Macroeconomic Indicators

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | GDP Annual Growth Rate (Combined) | GDP Growth Rate (Combined) | Unemployment Rate (Combined) | Inflation Rate (Combined) | Inflation Rate MoM (Combined) | Manufacturing PMI (Investing) |
| United States | 0 | 0 | 0 | 0 | 0 | 160 |
| Japan | 0 | 0 | 0 | 0 | 5 | 109 |
| United Kingdom | 0 | 0 | 0 | 0 | 0 | 109 |
| Canada | 0 | 0 | 0 | 0 | 0 | 263 |
| France | 0 | 0 | 0 | 0 | 0 | 115 |
| Switzerland | 3 | 3 | 0 | 0 | 0 | 0 |
| Germany | 0 | 0 | 0 | 0 | 0 | 109 |
| Australia | 3 | 3 | 0 | 157 | 300 | 239 |
| Netherlands | 0 | 0 | 0 | 0 | 0 | 300 |
| Sweden | 0 | 0 | 0 | 0 | 0 | 0 |
| Hong Kong | 177 | 129 | 0 | 1 | 292 | 161 |
| Spain | 0 | 0 | 0 | 0 | 0 | 157 |
| Italy | 0 | 0 | 0 | 0 | 0 | 161 |
| Singapore | 141 | 180 | 239 | 1 | 292 | 300 |
| Denmark | 0 | 0 | 0 | 0 | 0 | 290 |
| Finland | 0 | 3 | 0 | 0 | 0 | 300 |
| Belgium | 0 | 0 | 0 | 0 | 0 | 300 |
| Norway | 3 | 0 | 0 | 0 | 0 | 63 |
| China | 0 | 144 | 238 | 0 | 0 | 75 |
| Taiwan | 129 | 300 | 2 | 0 | 0 | 161 |
| India | 3 | 3 | 300 | 0 | 0 | 159 |
| Korea | 0 | 0 | 0 | 0 | 0 | 167 |
| Brazil | 3 | 3 | 55 | 0 | 0 | 161 |
| Russia | 63 | 78 | 3 | 0 | 0 | 162 |
| South Africa | 3 | 3 | 253 | 0 | 0 | 174 |
| Mexico | 0 | 0 | 0 | 0 | 0 | 161 |
| Malaysia | 144 | 300 | 190 | 1 | 293 | 300 |

There are various strategies to fill missing values. The simplest one is to fill with 0 or -1 values. Other would be to use mean or median values. Finally for more advanced methods we can use predictive models or interpolation.

For selected indicators, there are certain countries not included in the OECD, like Singapore, Hong Kong, Taiwan and Malaysia, which tend to have more missing values. In this case, for the most accurate approximation, I have decided to use mean values of the 5 most correlated countries. That should give much more precise approximation than taking a mean value of all countries.

Correlation values are calculated on the ETFs returns for the last 12 months for a given date with missing values. For example, when filling missing values for May 2002, we can only use correlations calculated on returns till that date. We cannot use single calculation of correlations based on the entire 25-year period to fill all missing values. That’s because in many instances we would be using information from future dates. Looking forward is one of the most common mistakes made when back testing investment algorithms that distorts results and achieve unrealistic performances.

### Calculating available indicators data

When working with macroeconomic indicators we must be very careful about release dates and consecutive revisions. We cannot calculate portfolio weights using indicators released in future dates or using revisions to original values that were updated later. We can only use indicator values available at the time of rebalancing.

In my research I have examined release dates for all selected indicators and used values only after the last revisions for all countries. That introduces certain delay for some indicators, having to wait for all countries to release their data, but it assures I only use available data.

Selected 15 indicators have different release schedule that must be considered:

* Quarterly indicators - GDP Annual Growth Rate, GDP Growth Rate and Current Account to GDP have the most delay of all indicators. They have up to 3 releases, with 2 provisional readings and one final value. For example, for the first quarter, we have releases at the end of April, at the end of May and the final one at the end of June. That way we can only use data from the first quarter starting from July, having the final values. Moreover, these quarterly values need to be converted into monthly format, using ffill method, to be compatible with other monthly indicators.
* Most monthly indicators like Unemployment Rate, Inflation Rate, OECD Business and Consumer Confidence Indicators, Short- and Long-Term Interest Rates provided by OECD, Total Manufacturing and Producer Price Index are released during the following month after the given month ends. As the rebalancing is done at the beginning of each month, we need to use 2-month delay. For example, in March we can use values from January, but not from February that haven’t been released at that time.
* Manufacturing PMI has very little delay, being released at the beginning of the following month. At the 4th or 5th, we already have data for all countries from previous month. This indicator is one of the best predictors for economic cycle and its release dates have been selected for algorithm rebalancing.
* Central Bank Rates provided by Bank of International Settlement has daily values with minimal delay. In the algorithm, the most recent values can be used.
* Industrial Production and Retail Sales are taking more time to collect all relevant data. Their values are released up to 50 days after given month ends. That gives 3-month delay, one month more than most other monthly indicators. In this case we can use data from January only starting from April.

### Formatting input data

Having filled missing values and calculated available indicators for the given date, the next step is to format input data for machine learning models and neural networks. First of all, all indicators must be normalized to be on the same scale. Normalizing the data can lead to improved model performance by reducing the impact of outliers and extreme values. Moreover, it fosters stability in the optimization process, promoting faster convergence of the models during training.

Normalized indicators can be passed separately, concatenating all indicators for selected countries. However, with this approach input data will have significant size. When using all 15 selected indicators for 27 countries multiplied by 6 months of data, it results in 405 columns by 6 rows. In flatten version, required by machine learning models or simple deep neural network has 2430 inputs. With monthly rebalancing there are relatively few data points to train having in total 300 months for 25 years minus period reserved for testing. Having that many inputs with so few training examples, the model might not be able to learn all complex relationships from macroeconomic analysis.

One solution to excessive input size is dimensionality reduction which aims to reduce the number of input features while preserving essential information. Reduced dimensions lead to faster training times without significant information loss. The easiest way to do that is to take either equal or fixed weights applied to all indicators to create a new single composite indicator with shape of 27x6 instead of 405x6.

Another, more complex method is to use Principal Component Analysis (PCA). PCA transforms the original features into a new set of uncorrelated features (principal components). It captures the most significant variance in the data, allowing us to represent it in a lower-dimensional space.

In my research I have examined different input formatting, from separate indicators, simple complex indicator to various number principal components. More details on the testing model and achieved results will be presented in the next chapter.

### Splitting training data

Consistent with established practices, the input dataset has been partitioned into an 80% training set and a 20% test set. Additionally, for neural networks, an additional 10% has been reserved from the training data for validation purposes. Furthermore, the initial 12 months from 1999 have been excluded from the training set to account for the requisite initial knowledge needed for the first rebalancing.

In summary, total 25-year period is divided into following date ranges:

* 01/1999 – 12/1999 – reserved for initial training step
* 01/2000 – 09/2016 – training data
* 10/2016 – 02/2019 – validation data
* 03/2019 – 12/2023 – test data

With broad economic periods being selected for both training and testing, more realistic models can be trained, not skewed toward any particular market conditions. Moreover, testing models through different stages of economic cycles will help to show the performance of the investment algorithm in different scenarios.

# Presentation of work

TODO:

* Describe all models and results they achieved.

# Conclusions

TODO:

* Summary of results
* Recommendations: generalize conclusions to appropriate design decisions, practices and/or procedures
* Implications for further study
* Future Work

# References

Kenton, W. (2023). *Black-Litterman Model: Definition, Basics, and Example*. Retrieved from Investopedia: https://www.investopedia.com/terms/b/black-litterman\_model.asp

Lindquist, W. R. (2022). Modern Portfolio Theory. In W. R. Lindquist, *Advanced REIT Portfolio Optimization.* vol 30. Springer, Cham.

Mitchell, C. (2024). *MSCI All Country World Index (ACWI): Definition and Countries*. Retrieved from Investopedia: https://www.investopedia.com/terms/m/msci-acwi.asp

MSCI. (2019). *Selected geographical issues in the global listed equity market, Analysis for the Norwegian Ministry of Finance.*

MSCI. (2024). *ACWI IMI's Complete Geographic Breakdown*. Retrieved from https://www.msci.com/research-and-insights/visualizing-investment-data/acwi-imi-complete-geographic-breakdown

Prado, M. L. (2016). Building Diversified Portfolios that Outperform Out-of-Sample. *Journal of Portfolio Management*, 31.

Prado, M. L. (2018). *Advances in Financial Machine Learning.*

Prado, M. M. (2020). *Machine Learning for Asset Managers (Elements in Quantitative Finance).* Cambridge University Press.

Roberto Violi, E. C. (2018). *Emerging market portfolio strategies, investment performance, transaction cost and liquidity risk.* ResearchGate.

Vanguard UK Professional. (2024). *The power of portfolio diversification*. Retrieved from Vanguard: https://www.vanguard.co.uk/professional/vanguard-365/investment-knowledge/principles-of-investing/the-power-of-diversification