Uncovering Market Structures Through Correlation Matrix Ordering Techniques

Author: Juan José Pellicer Querol.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Josep Perelló Palou.

Abstract: The objective of this project is to characterise market behaviour and analyse potential correlations using Random Matrix Theory (RMT), applying different criteria to reorder the correlation matrix heatmap. In addition, a basic strategy is explored for constructing an investment portfolio based on the eigenvectors of the correlation matrix, and a Minimum Spanning Tree (MST) is also constructed to visualise the market structure in a simplified topological form. The results show that the distribution of asset returns is leptokurtic, with heavier tails than a Gaussian distribution. The reordered heatmaps reveal structural insights into market groupings under different criteria and confirms that patterns are not result of random fluctuations.
Keywords: Random Matrix Theory, Econophysics, Correlation Analysis, Market Behaviour, Portfolio construction.
SDGs: 8. Decent Work and Economic Growth.

I. INTRODUCTION

Econophysics is an interdisciplinary field that applies concepts and methods from statistical physics to analyse and model complex phenomena in economics. The relevance of economics and physics arises from the fact that financial markets can be seen as a large system influenced by the actions of interacting agents (investors).[1]

This behaviour is found across all asset classes (indices, commodities, bonds...), but the distribution of price fluctuations deviates significantly from the Brownian model and Gaussianity.[4]

Understanding the correlations between financial assets is crucial for risk management, portfolio construction, and identification of collective market behaviour. Correlation structures help investors diversify, reduce exposure to systemic shocks, and detect patterns that may not be visible through individual asset analysis.

Here, we determine whether the 28 selected assets exhibit significant correlations and explore their underlying structure. To achieve this, we apply Random Matrix Theory (RMT) [2], which enables us to construct a correlation matrix that quantifies the strength of relationships between assets and helps uncover the factors that may be driving these patterns.

By diagonalising the correlation matrix, we extract its eigenvalues and eigenvectors, which reveal dominant collective behaviours and underlying substructures.[7] In addition, we explore the performance of simple portfolios constructed using the leading eigenvectors, assessing their cumulative returns over time. [8]

We also construct a Minimum Spanning Tree (MST) by converting the correlation matrix values into distances. [6], [3], [5] This provides a simplified topological view of the market structure, highlighting essential links between assets.

To deepen the analysis, we generate and compare four heatmaps of the correlation matrix, each ordered using a different method: an arbitrary grouping criterion, a time-shuffled order, the components of the leading eigenvector (market mode) and the structure revealed by the MST. Each approach offers a distinct perspective on market organisation and helps identify clusters and hidden structures within asset correlations.

II. DATA COLLECTION AND PREPROCESSING

The datasets used were obtained from *Investing.com* and contain daily closing prices for 28 assets from October 3, 2005, to December 30, 2024. Only the closing price and date were used, ensuring a consistent reference point across markets in different parts of the world.

To standardise the format, all datasets were converted to European style numerical notation (period for thousands, comma for decimals), addressing variations between source markets. Dates when any market was closed due to holidays were removed, aligning all assets on the same trading days. The final dataframe contains 3,421 rows and 28 columns.

The selected assets fall into three main categories. The first is **stock indices**, which represent the performance of a group of stocks, usually tracking an entire market or a specific sector. The second category includes **commodities**, which are physical goods or raw materials that can be traded in markets. Lastly, **bonds** are debt instruments issued by governments or corporations; they represent loans from investors to issuers and typically offer fixed income returns. [8]

All processing was done in Python 3, using **pandas**, **numpy**, **matplotlib**, **seaborn**, **scipy**, **os**, and **networkx**.

Finally, an Appendix is included, listing all the assets used in the study along with a brief description of each one and a link to its corresponding dataset.

III. DEVELOPING SECTIONS

A. Universal Behaviour of Returns

We begin by analyzing the statistical behaviour of price fluctuations in financial markets with the logarithmic returns for each market i. These are calculated as follows:[4],[8]

$$R_i(t) = \ln\left(\frac{P_i(t+1)}{P_i(t)}\right) \tag{1}$$

Once computed, they are then standardised by subtracting the mean and dividing by the standard deviation.

$$X_i = \frac{R_i(t) - \langle R_i(t)\rangle_t}{\sigma_i} \tag{2}$$

This step ensures that all assets are represented on the same scale, allowing for a clear and visual comparison of their return distributions, regardless of their individual volatility levels.

The individual distributions are then plotted to examine their empirical shapes and to compare the overall behaviour across different assets, as shown in FIG. 1.

In the same figure, the PDFs of the standardized returns deviate from the standard Gaussian. This behaviour is consistent across all asset types (indices, commodities, and bonds). These distributions are more leptokurtic, exhibiting sharper peaks and fatter tails than the Gaussian distribution. [4]

To better visualise the fat tails, we have added two insets corresponding to each tail of the distribution on a log-log scale. The fat-tailed behaviour becomes more clearly visible, indicating a higher probability of extreme events (large gains or losses) than predicted by a normal distribution. These insets also reveal an asymmetry: both tails follow power-law distributions, but with slightly different exponents, approximately $\alpha \approx -2.497$ for the left tail and $\alpha \approx -2.548$ for the right tail. This suggests that extreme losses are marginally more likely than extreme gains, highlighting a small bias in the distribution of logarithmic returns.

B. Markets correlation

We analyse return correlations to uncover hidden market structures. Starting with a heatmap of the correlation matrix using an arbitrary asset order, we then perform a spectral analysis based on its eigenvalues and eigenvectors. From the leading eigenvectors, we construct theoretical portfolios and evaluate their performance. Finally, we reorder the correlation matrix according to the market mode to better highlight asset clusters and collective behaviours.

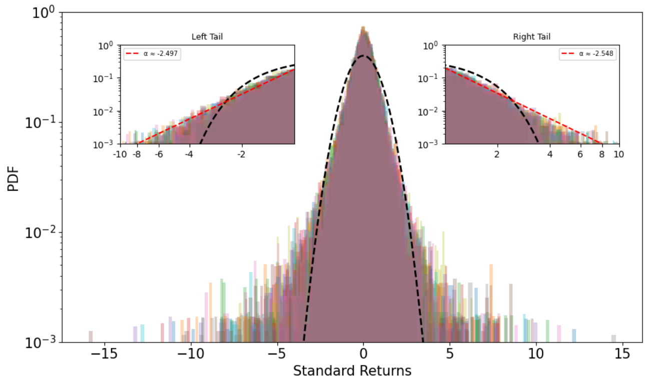


FIG. 1: **Empirical distributions (histograms) vs Gaussian with zoom in the fat tails insets.** Empirical probability density functions (PDFs) of the 28 assets in front of the normal Gaussian on a logarithmic scale. The x-axis shows standardized returns. We include insets of the right and left tails plotted on a log-log scale, revealing power law behaviour with different exponents.

1. *Correlation matrix*

One way to study the internal structure of financial markets is through Random Matrix Theory (RMT). To apply this approach, we construct a correlation matrix using the Pearson correlation coefficient on the logarithmic returns of the 28 selected markets. This matrix is defined by: [7]

$$\rho_{ij} = \frac{\langle R_i(t)R_j(t)\rangle_t - \langle R_i(t)\rangle_t\langle R_j(t)\rangle_t}{\sigma_i\sigma_j} \tag{3}$$

Here $\langle\cdot\rangle_t$ denotes the temporal average and σ_i is the standard deviation of the returns of the asset i.

The order chosen to represent the assets in the correlation matrix heatmap is arbitrarily selected to make certain patterns more visible and improve interpretability. They are grouped by continent and asset type. [6][7]

FIG. 5 heatmap shows the correlation of all assets. We can see that the diagonal of the matrix holds the maximum correlation values, as each asset is perfectly correlated with itself. We also observe distinct correlation blocks, suggesting the presence of market-specific clusters, such as European, Chinese, Asian, North American markets, and precious metals. Furthermore, certain assets show weak or even negative correlations, for example, between precious metals and US 10-year bonds, or between Chinese markets and other equity indices.

2. *Eigenvalues and eigenvectors*

A powerful approach consists in diagonalising the correlation matrix to obtain its eigenvalues and eigenvectors. These form a new orthogonal basis in which the collective dynamics of the market become more transparent.

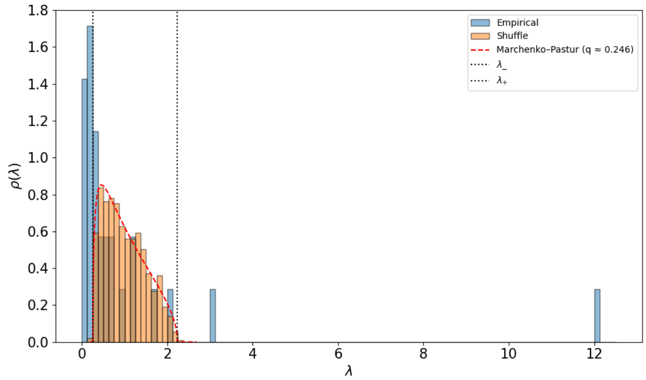


FIG. 2: **Empirical eigenvalue spectrum vs shuffled spectrum of the correlation matrix.** Three regions appear in the empirical histogram: a large group near the lower end, a smaller group at medium range, and one dominant eigenvalue. The shuffled spectrum fits well within the theoretical Marchenko–Pastur distribution 4, highlighting the presence of non-random structure in the empirical data.

The eigenvectors capture correlated modes of asset behaviour, while the eigenvalues quantify the strength of each mode, allowing us to identify dominant market factors and noise.

An interesting thing you can do is to see the eigenvalue distributions of the correlation matrix.

The structure of FIG. 2 suggests the presence of noise in the market data, reflected in the lower part of the eigenvalue spectrum. To confirm this, we apply a methodology similar in spirit to that of Bouchaud and Potters [2], comparing the empirical spectrum with the Marchenko–Pastur distribution. The spectrum is obtained by dividing the full time period into 30 windows of 114 trading days, resulting in a ratio of $q = \frac{28}{114} \approx 0.246$. Within each window, the time series are randomly shuffled to destroy temporal correlations.

In FIG. 8, we present the heatmap of the logarithmic returns after shuffling the time series over the entire study period. As temporal correlations are destroyed, no meaningful structure can be observed.

The eigenvalue spectrum of the resulting shuffled matrices fits well within the bounds predicted by the Marchenko–Pastur distribution, defined as:

$$\lambda_\pm = (1 \pm \sqrt{q})^2\,; \quad \rho(\lambda) = \frac{\sqrt{4\lambda q - (\lambda + q - 1)^2}}{2\pi\lambda q} \tag{4}$$

It is worth noting that some empirical eigenvalues fall outside the theoretical bounds, particularly near zero. This is expected in the very low limit $q \to 0$ (in our total time period $q_T = \frac{28}{3420} \approx 0.0082$), where the distribution underestimates the density near $\lambda \approx 0$. Theoretical bounds prediction for this limit are $\lambda_\pm = 1$.

The mid-range values of the spectrum can explain the correlation of some clusters in the market. Finally, the largest eigenvalue suggests a collective movement of all assets. This eigenvalue is known as the **market mode**. [1],[7]

TABLE I: **Leading eigenvector components.** compilation of the five most significant components from the eigenvectors associated with the three largest eigenvalues.

$\lambda_1 = 12.0480$		$\lambda_2 = 3.0349$		$\lambda_3 = 2.0897$	
Market Mode		China Market		Commodities	
AEX	0.258550	Shangai	0.467841	Gold	-0.561025
CAC 40	0.257636	CHINA A50	0.457892	Silver	-0.531310
FTSE 100	0.253655	SZSE	0.455710	Copper	-0.263112
DAX	0.252566	Hang Seng	0.284068	Oil	-0.224447
OMXS	0.243449	Kospi	0.199065	S&P TSX	-0.174358

From the eigenvalues, we extract the associated eigenvectors, which define the basis of the diagonalized correlation markets. We focus on the three largest and characterise them using the five most significant components per eigenvector. (See TABLE I).

The first eigenvector, associated with the market mode, is dominated by European indices, suggesting strong regional coherence. The second reflects the Chinese market, with an additional Asian presence, also suggesting a regional grouping. The third is mainly composed of commodities and Canada, an economy highly exposed to raw materials. Interestingly, all of its components are negative, indicating a collective behaviour distinct from equity markets.

We can construct a portfolio using the previously calculated eigenvectors (eigenportfolios). The sign of each component is interpreted as a position, positive for the buy position and negative for the sell position. The cumulative return of each portfolio is calculated using the cumulative growth factor ($\eta(t)$): [8]

$$\eta_i(T) = e^{\sum_{t=0}^{T-1} R_i(t)} = \frac{P_i(T)}{P_i(0)} \tag{5}$$

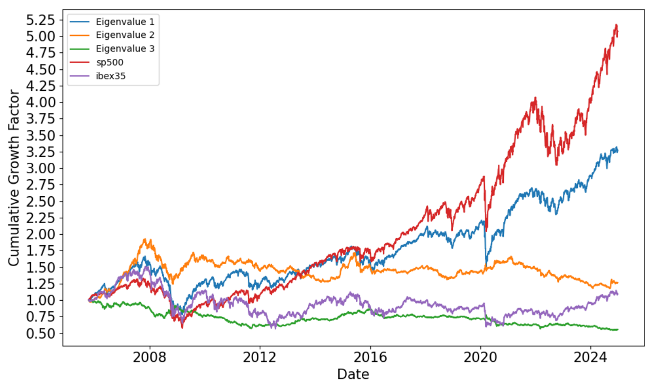


FIG. 3: **Cumulative performance of portfolios based on leading eigenvectors.** Performance over time of the S&P 500, Ibex 35, and the three eigenportfolios constructed from the largest eigenvalues of the correlation matrix.

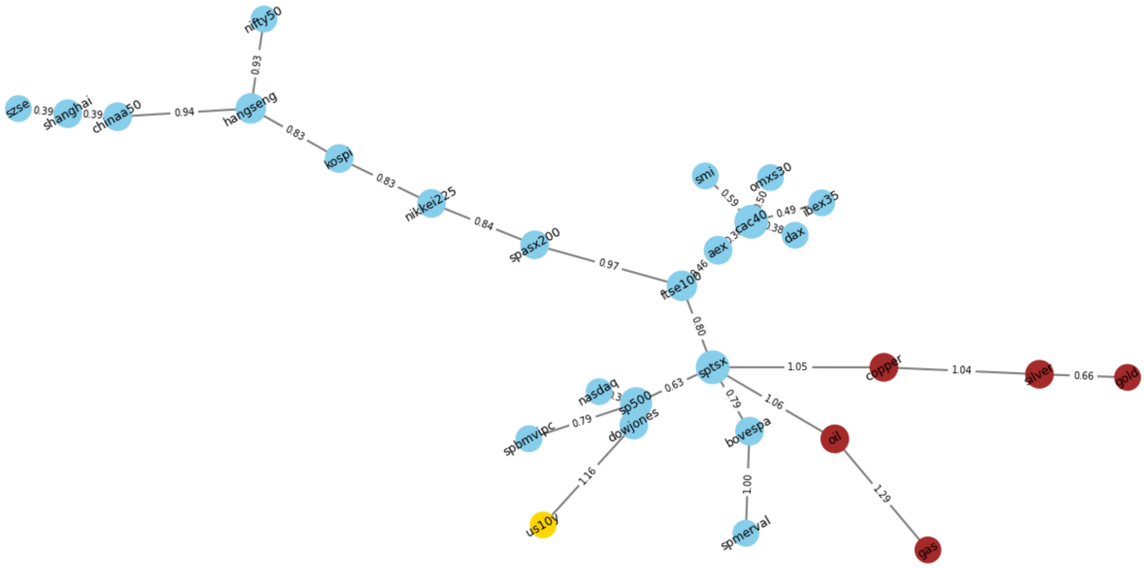


FIG. 4: **Minimum Spanning Tree of the assets. Minimum Spanning Tree of the assets.** Built from the correlation matrix by applying Mantegna's distance transformation (6).

We compare these eigenportfolios in FIG 3 with a benchmark indices, the S&P 500, one of the most widely followed equity indices, and IBEX 35, the Spanish stock index.

The market mode portfolio shows relatively good performance, multiplying the initial capital by approximately ($\times 3.27$) during the study period. Eigenportfolio 2 achieves a modest gain ($\times 1.26$), while eigenportfolio 3 suffers a significant loss ($\times 0.55$), resulting in approximately half the initial capital. Despite the solid performance of the market mode, clearly outperforming Ibex 35, the S&P 500 surpasses all portfolios with a total gain of $\times 5.07$. This example is basic; more advanced techniques could improve the optimization of the eigenportfolio.

3. Correlation Matrix ordered by Market Mode

The asset can be reordered according to the values of the components in the eigenvectors associated with the market mode. The ordering follows the components of the highest values (highlighted in Table I) to the lowest (Gas: 0.0381).

In FIG. 6 a gradual fading of correlation is visible across the heatmap. This reflects each asset correlates with the dynamics of the market mode: Western markets exhibit strong mutual correlation, while Chinese markets, commodities, and bonds appear less connected to the dominant market behaviour.

C. Minimum Spanning Tree

Finally, we construct a Minimum Spanning Tree (MST) from the correlation matrix (3). To do this we need to transform the Pearson correlation coefficient ρ_{ij} is transformed into a distance metric d_{ij} using the equation proposed by Mantegna: [6],[3],[5],[7]

$$d_{ij} = \sqrt{2(1-\rho_{ij})} \tag{6}$$

We reorder the correlation matrix based on its structure criterion. Starting from the most isolated node (SZSE), we traverse the tree by always following the shortest available branch at each bifurcation, recursively continuing until all assets are covered.

FIG. 7 shows similar regional groupings as in previous heatmaps. Moreover, the structure reveals how correlations between markets are connected: strong correlations concentrate in the centre, dominated by Western markets, and gradually weaken toward the upper-left corner, where Chinese markets appear, linked through other Asian markets. In the opposite direction, the structure leads to the commodity markets, which appear to be more distant and less correlated with the rest.

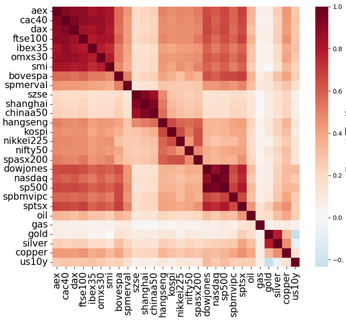


FIG. 5: **Correlation Matrix of logarithmic returns.** Correlation matrix of logarithmic returns for the 28 assets (1), computed using equation (3) ordered for visually identify main clusters.

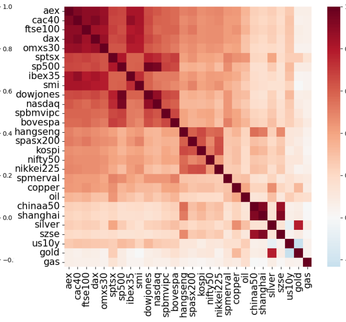


FIG. 6: **Heatmap Ordered by Market Mode Eigenvector.** Correlation matrix of logarithmic returns, ordered by the market mode eigenvector. (TABLE I)

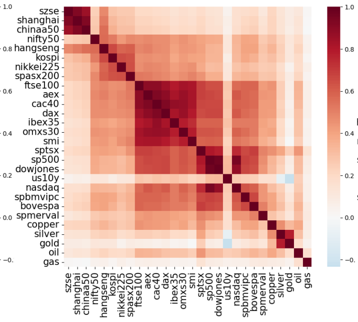


FIG. 7: **Heatmap Ordered by Minimum Spanning Tree.** Correlation matrix of logarithmic returns for the 28 assets, ordered using the Minimum Spanning Tree. (FIG. 4)

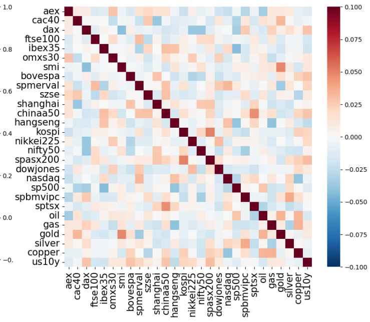


FIG. 8: **Heatmap of shuffled correlation matrix.** Correlation matrix of logarithmic returns after randomly shuffling the time series of each asset independently.

IV. CONCLUSIONS

In the project, we analyse how different ordering criteria of the correlation matrix can reveal underlying structures in financial markets. Four heatmaps were generated based on an arbitrary grouping, the market mode eigenvector, the Minimum Spanning Tree, and a final one based on a single temporal shuffle of the dataset (see Appendix B for more resolution). The first three heatmaps show some clustering according to regional markets. The arbitrarily ordered heatmap one provides a general overview, but lacks any other meaningful structure. The market mode ordering reveals how each asset aligns with the dominant collective behaviour, producing a smooth gradient. The Minimum Spanning Tree reordering reflects the topological organisation of the market, identifying which assets serve as links between regions. In contrast, the shuffled heatmap, where the temporal structure has been destroyed, shows no meaningful pattern, confirming that the structures seen in the other heatmaps arise from genuine market correlations.

To complement this, we analysed the statistical behaviour of logarithmic returns, which exhibit a leptokurtic probability distribution. The decomposition of the eigenvalues of the correlation matrix revealed three distinct regions: a zone dominated by noise (consistent with the Marchenko–Pastur distribution), a mid-range cluster region and a dominant market mode. Finally, we evaluated a simple portfolio strategy based on the leading eigenvectors. The results were mixed: the portfolio aligned with the market mode delivered strong returns, while those based on the other eigenvectors performed significantly worse.

Acknowledgments

I express my sincere gratitude to my advisor, Josep Perelló Palou, for his guidance and support throughout the development of this project. I also thank my family for their unconditional encouragement and my friends and classmates, both from the faculty and the residence, for their companionship and support during these years of studying physics.

[1] Mantegna, R. N., & Stanley, H. E. (1999). *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge University Press.

[2] Bouchaud, J. P., & Potters, M. (2015). Financial applications of random matrix theory: a short review.

[3] Cheng, H. J. Z., Rea, W., & Rea, A. (2015). A comparison of three network portfolio methods: Evidence from the Dow Jones. *arXiv preprint* arXiv:1512.01905.

[4] Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance, 1*(2), 223–236.

[5] Cheng, H. J. Z., Rea, W., & Rea, A. (2015). An application of correlation clustering to portfolio diversification. *arXiv preprint* arXiv:1511.07945.

[6] Mantegna, R. N. (1999). Hierarchical structure in financial markets. The European Physical Journal B-Condensed Matter and Complex Systems, 11, 193-197.

[7] Mantegna, R. N. (2023). Noise and information in economic and financial systems [Video]. *YouTube.* `https://www.youtube.com/watch?v=zUI7_VTloxA`

[8] Bouchaud, J.-P., & Potters, M. (2003). *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management.* Cambridge University Press.

Descobriment d'estructures de mercat mitjançant tècniques d'ordenació de la matriu de correlació

Author: Juan José Pellicer Querol.
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Josep Perelló Palou.

Resum: L'objectiu d'aquest projecte és caracteritzar el comportament del mercat i analitzar les possibles correlacions utilitzant la Teoria de Matrius Aleatòries (RMT), aplicant diferents criteris per reordenar el mapa de calor de la matriu de correlació. A més, s'explora una estratègia bàsica per construir un portafoli d'inversió a partir dels autovectors de la matriu de correlació, i també es construeix un Arbre Generador Mínim (MST) per visualitzar l'estructura del mercat en una forma topològica simplificada. Els resultats mostren que la distribució dels rendiments dels actius és leptocúrtica, amb cues més pesades que una distribució gaussiana. Els mapes de calor reordenats revelen informació estructural sobre els agrupaments de mercats segons diferents criteris i confirmen que aquests patrons no són resultat de fluctuacions aleatòries.
Paraules clau: Teoria de Matrius Aleatòries, Econofísica, Anàlisi de correlacions, Comportamiento del mercado, Contrucció de carteres.
ODSs: 8. Treball digne i creixient econòmic.

Objectius de Desenvolupament Sostenible (ODSs o SDGs)

1. Fi de la es desigualtats		10. Reducció de les desigualtats	
2. Fam zero		11. Ciutats i comunitats sostenibles	
3. Salut i benestar		12. Consum i producció responsables	
4. Educació de qualitat		13. Acció climàtica	
5. Igualtat de gènere		14. Vida submarina	
6. Aigua neta i sanejament		15. Vida terrestre	
7. Energia neta i sostenible		16. Pau, justícia i institucions sòlides	
8. Treball digne i creixement econòmic	X	17. Aliança pels objectius	
9. Indústria, innovació, infraestructures			

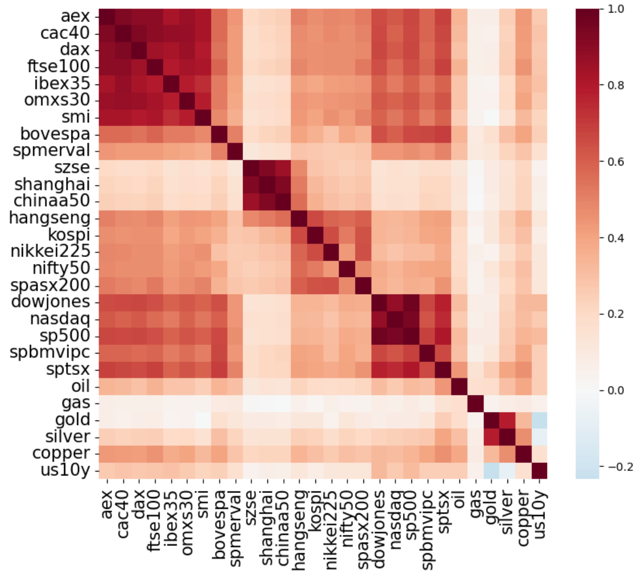
El contingut d'aquest TFG està relacionat amb l'ODS 8, en particular amb els objectius 8.1 i 8.3, ja que una bona estratègia d'inversió pot contribuir a mantenir o fins i tot augmentar el PIB. Un exemple clar és el fons sobirà de Noruega, que no només protegeix l'economia nacional davant les oscil·lacions del mercat, sinó que també té com a finalitat garantir el benestar de les generacions futures. A més, aquest treball també es vincula amb l'objectiu 8.10, ja que l'estudi de les correlacions i del comportament col·lectiu dels mercats pot ajudar les entitats financeres a dissenyar estratègies més inclusives. Aquest coneixement permetria oferir serveis més accessibles, adaptats al perfil de cada persona, valorant els beneficis i riscos estadístics per a poder assessorar i monitorar millor les seves inversions.

Appendix A: ASSETS

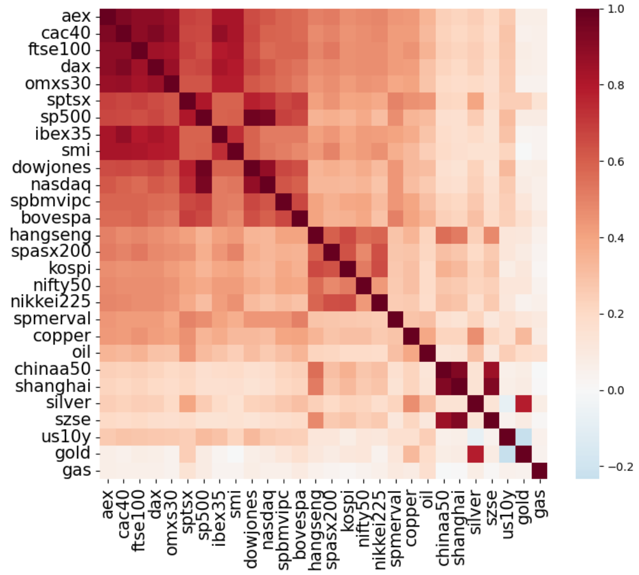
TABLE II: A compilation of the assets used in the study, with a brief description of each. The dataset spans from October 3, 2005 to December 30, 2024.

ASSETS TABLE		
Asset	Description	Source
AEX	Index of the Amsterdam Stock Exchange (Netherlands).	https://es.investing.com/indices/netherlands-25
Bovespa	Main index of São Paulo Stock Exchange (Brazil).	https://es.investing.com/indices/bovespa
Cac 40	Index of the Paris Stock Exchange (France).	https://es.investing.com/indices/france-40
China A50	Index of large-cap companies from Shanghai and Shenzhen Stock Exchanges (China).	https://es.investing.com/indices/ftse-china-a50
DAX	Index of the Frankfurt Stock Exchange (Germany).	https://es.investing.com/indices/germany-30
Dow Jones	Industrial index of the New York Stock Exchange (USA).	https://es.investing.com/indices/us-30
FTSE 100	Index of the London Stock Exchange (UK).	https://es.investing.com/indices/uk-100
Hang Seng	Index of the Hong Kong Stock Exchange (Hong-Kong).	https://es.investing.com/indices/hang-sen-40
IBEX 35	Index of the Madrid Stock Exchange (Spain).	https://es.investing.com/indices/spain-35
KOSPI	Index of the Korea Exchange (South Korea).	https://es.investing.com/indices/kospi
NASDAQ	Technology index of the New York Stock Exchange (USA).	https://es.investing.com/indices/nasdaq-composite
Nifty 50	Index of the National Stock Exchange of India (India)	https://es.investing.com/indices/s-p-cnx-nifty
Nikkei 225	Index of the Tokyo Stock Exchange (Japan).	https://es.investing.com/indices/japan-ni225
OMXS 30	Index of the Stockholm Stock Exchange (Sweden).	https://es.investing.com/indices/omx-stockholm-30
Shanghai	Composite index of the Shanghai Stock Exchange (China).	https://es.investing.com/indices/shanghai-composite
SMI	Index of the Zurich Stock Exchange (Switzerland).	https://es.investing.com/indices/switzerland-20
S&P 500	Index of the New York Stock Exchange (USA).	https://es.investing.com/indices/us-spx-500
S&P ASX 200	Index of the Sydney Stock Exchange (Australia).	https://es.investing.com/indices/aus-200
S&P BMV IPC	Index of the Mexico City Stock Exchange (Mexico).	https://es.investing.com/indices/ipc
S&P TSX	Index of the Toronto Stock Exchange (Canada).	https://es.investing.com/indices/s-p-tsx-composite
SZSE	Index of the Shenzhen Stock Exchange (China).	https://es.investing.com/indices/szse-component
Copper	Copper price.	https://es.investing.com/commodities/copper
Oil	Crude oil price.	https://es.investing.com/commodities/crude-oil
Gas	Natural gas price.	https://es.investing.com/commodities/natural-gas
Silver	Silver price.	https://es.investing.com/currencies/xag-usd
Gold	Gold price.	https://es.investing.com/currencies/xau-usd
US 10 Years	Yield of the 10 year U.S Treasury bond (USA).	https://es.investing.com/rates-bonds/u.s.-10-year-bond-yield

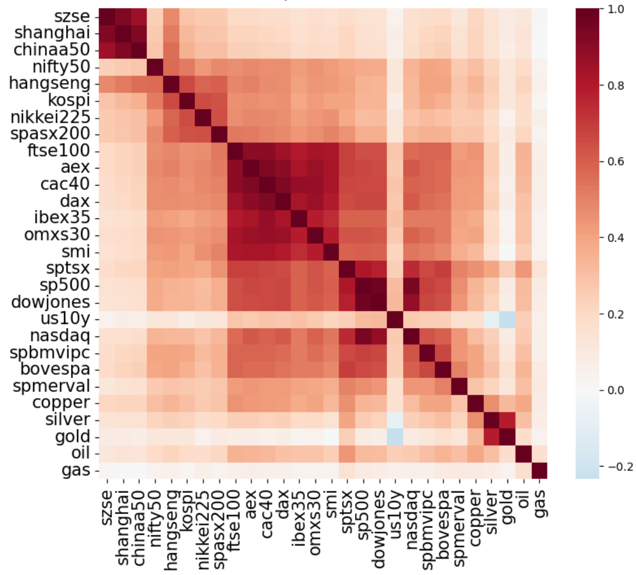
Appendix B: Correlation Matrix Heatmaps



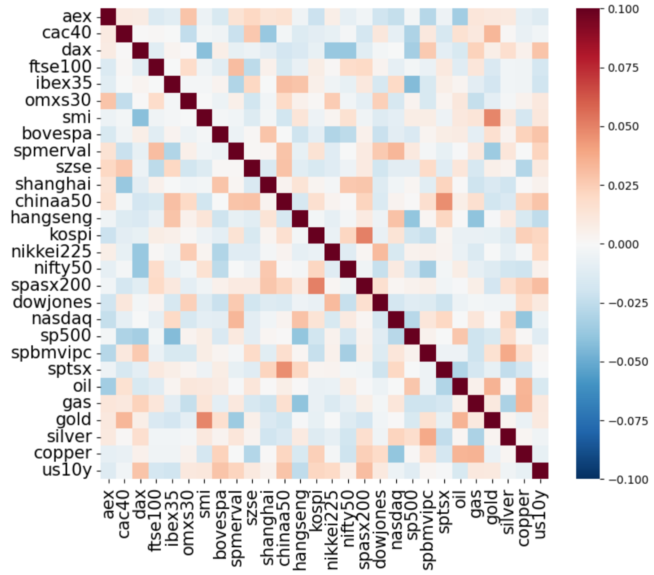
Correlation Matrix of logarithmic returns. Correlation matrix of logarithmic returns for the 28 assets (1), computed using equation (3) ordered for visually identify main clusters.



Heatmap Ordered by Market Mode Eigenvector. Correlation matrix of logarithmic returns, ordered by the market mode eigenvector. (TABLE I)



Heatmap Ordered by Minimum Spanning Tree. Correlation matrix of logarithmic returns for the 28 assets, ordered using the Minimum Spanning Tree. (FIG. 4)



Heatmap of shuffled correlation matrix. Correlation matrix of logarithmic returns after randomly shuffling the time series of each asset independently.