Review of Global Industry Classification

László Nagy,
Department of Finance
Budapest University of Technology and Economics
Magyar tudósok körútja 2, Budapest H-1117, Hungary
E-mail: nagyl@finance.bme.hu

Mihály Ormos
Department of Economics
J. Selye University
Bratislavská cesta 3322, SK-94501 Komárno, Slovakia
E-mail: ormosm@ujs.sk

KEYWORDS
Cluster analysis, market classification, GICS, machine learning

ABSTRACT

This paper introduces the financial market implied industry classification standard. Besides current industry classifications we propose a spectral clustering based quantitative methodology. The main drawback of current standards come from their qualitative classification techniques which can be eliminated in this purely mathematical concept. Calculating the market implied clusters and compering them with global industry classification explores that market implied classification provides better statistical results. However it also turns out that different clustering techniques provide similar classifications, moreover, both methods determine "Real Estate" as a cluster.

INTRODUCTION

Practitioners are using Global Industry Classification Standards (GICS) to keep track of sector based moves. It is widely accepted that firms with similar business activities have similar macro and micro factor sensitivities. Thus, building portfolios with names from the same industry lets us to eliminate specific risk. Sector based bundling gives the opportunity to tailor the macro factor exposure, understand the contamination of macro level shocks and determine regulatory standards. Moreover, it also helps to identify firms which under or outperform the industry hence it helps to separate systematic and non-systematic risks.

Asset managers use industry classification standards for their asset allocation and risk management strategies. Moreover, industry classification is also important for regulators, governments and labour economists to have a deeper understanding of the given state of the economy and be able to implement policies.

GICS is the mainly used classification. The methodology was developed by Morgan Stanley Capital International Inc. (MSCI) and Standard and Poor's (S&P). The categorization combines quantitative and qualitative techniques (MSCI 2015.) to obtain a market oriented economically thoughtful classification standard.

In this article we introduce a spectral clustering based (Nagy and Ormos 2016.) purely quantitative model to unveil the Financial Market Implied Classification (FMIC). Using daily closing prices, we show that the normalized modularity cut (Bolla 2011.) implied clusters and GICS are highly comparable. Moreover, standard Capital Asset Pricing Model based regressions give further evidences for using spectral clustering. The results show that GICS can be looked on as an approximation of the spectral clustering based classification. In addition, the clusters support the latest review of GICS in which Real Estate was added as distinct asset class.

The article structured as follows: Section 2 is a brief overview of industry classification standards. In Section 3 we introduce our spectral clustering based concept. In Section 4 we present the market implied clusters, compare them with groups given by GICS and carry out regression analysis to investigate the risk explanation power of FMIC. Section 5. summarizes the article.

INDUSTRY CLASSIFICATION STANDARDS

Investors are using different classification techniques from long ago to identify industries and characterize their specific behaviour. In 19th century all the market participants had their own classifications which were mainly used to evaluate risk, understand the macro factor sensitivities and build investment strategies (Vose 1916). However, there was no officially accepted framework. Hence, it was unmanageable to aggregate sector level sensitivities and implement economic policies.

Standard Industrial Classification

After the Great Depression in 1937, the US Central Statistical Board (Kolesnikoff 1940.) unified the different industry classifications and established the Standard Industry Classification. SIC was the main

Proceedings 32nd European Conference on Modelling and Simulation ©ECMS Lars Nolle, Alexandra Burger, Christoph Tholen, Jens Werner, Jens Wellhausen (Editors)
ISBN: 978-0-9932440-6-3/ ISBN: 978-0-9932440-7-0 (CD)

guideline of the Federal Government, banks and investors during the 20th century, in addition, the U.S. Securities and Exchange Commission (SEC) still uses it to its industry classification. The system had to be revised several times, in 1958, 1963, 1967, 1972, 1977 and 1987, but it still had several limitations. The main drawbacks of SIC are that it is designed for the economy of early 20th century, it is hard to identify new groups and follow the changes of global economy.

North American Industrial Classification System

NAICS is and industry classification standard which was designed to response to the increasing criticism of SIC (Executive Office of the President Office of Management and Budget. 2017). In 1980s the rapid changes of world economy forced the Office of Management and Budget (OMB) to overhaul SIC. Thus in 1997 the Economic Classification Policy Committee (ECPC) in Nacional the cooperation with Mexico's Instituto de Estadística, Geografía e Informática (now the Instituto Nacional de Estadística y Geografía, INEGI) and Statistics Canada suggested a new industry classification system which supplanted SIC. They agreed in that NAICS should be reviewed in every five years to reflect economic changes. The current form of NAICS defines 20 sectors and 1,057 industries. The classification standard is used for various administrative, regulatory, and taxation purposes.

Global Industry Classification Standards

Besides the SIC and NAICS in 1999 Morgan Stanley Capital International Inc. and Standard and Poor's created the widely used Global Industry Classification Standard. The primary goal of S&P was to enhance its business with introducing sector indices of S&P 500 index. In order to achieve this ambition an adequate industry categorization rule was needed. The classification has a market oriented nature, which incorporates quantitative and qualitative techniques. As GICS is used in the sub-index decomposition of S&P 500 thus it should follow all the changes of the market, hence it has to be reviewed at least annually. At first GICS introduced 10 sectors, 23 industry groups, 59 industries and 123 sub-industries. The classification was revised several times, currently it stands from 11 sectors, 24 industry groups, 68 industries and 157 sub-industries. Each sector, namely Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Telecommunication Services and Utilities represents an economically understandable market segment which is the key of the popularity of the classification. GICS is the most famous actively used standard which is assiduously reviewed by the financial market. Because, almost all U. S. market participants benchmark their positions against the performance of GICS sector indices. Moreover, the most liquid products in the world are the Standard & Poor's Depositary Receipts (SPDR) ETFs which tracks the S&P 500 index and sub-indices. Thus, the behaviour of GICS must be in line with the market because like a stock it is continuously reviewed by the market and like earning season at least annually revised by MSCI and S&P.

In this article we propose a purely quantitative technique (FMIC) to analyse the connections between financial market and GICS and study the efficiency of Global Industry Classification Standards.

SPECTRAL CLUSTERING

The original concept of classifying the market, defining sub-groups, distinguishing sectors raises the fundamental question: Who can judge the market?

If we do not want to make any a priori assumptions then we have to look at the data and dismiss other subjective classification guidelines. Mathematically it is possible to represent a datasets as a graph, hence, we can construct an abstract network of stocks. The most straightforward method would be representing stocks with nodes, connection strengths with weights. Thus, we can define the network with $\mathrm{G}(V, W)$ graph where V represents the set of stocks and W contains the connection information. It is widely used that the adjacency matrix W represents the graph, thus all the structure information is embedded in the matrix. Note that if we normalize the sum of each row to one, then we get the transition matrix of the random walk on the graph (Luxburg 2007.);

$$P = D^{-1}W \tag{1}$$

where D represents the diagonal matrix of the sums of rows. Studying stopping times of random walks on graphs sheds some light on the structure of the graph, because, if it takes a long time to reach a subgraph of the graph from a given node then it would mean that the node and the subgraph are well separated. Moreover, the largest eigenvalue of the submatrix of P which belongs to the subgraph controls the distribution of the stopping time;

$$\mathrm{Prob}(\tau \geq n) = 1 - \pi_0 \sum_{i=1}^{n} Q^{i-1} R \cdot 1 \tag{2}$$

where τ represents the stopping time, π_0 the initial distribution, Q the submatrix of transient points, R defines the transition probabilities from transient to recurrent points.

Using symmetricity and spectral theorem

$$\mathrm{Prob}(\tau \geq n) = 1 - \pi_0 V \sum_{i=1}^{n} \Lambda^{i-1} V^T R \cdot 1 \tag{3}$$

Equation 3. shows that the spectrum of adjacency matrix, stopping times and clustering properties are strongly connected.

In addition, the problem can be looked at from an analytical point of view. Fourier analysis is a widely used tool in pattern recognition theory (Shi and Malik 2000). Considering an arbitrary real valued function on the vertexes and defining the below incidence matrix led

some colour to the connections between Laplace operator and graph theory (Chung 1997.);

$$B_{ev} = \begin{cases} 1 & \text{if v is the initial vertex of e} \\ -1 & \text{if v is the terminal vertex of e} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

then $B^T B$ would be exactly the negative discrete Laplace operator because:

$$B^T B f(v_i) = \sum_{v_i \sim v_j} f(v_i) - f(v_j)$$

Notice that if we subtract the adjacency matrix from the diagonal matrix of row-sums then we get the same operator. Thus Laplace matrix can be defined as follows;

$$L = D - W \quad (5)$$

If we think of the adjacency matrix as a noisy matrix and would like to maximize the information content, then we get the modularity matrix;

$$M = W - dd^T \quad (6)$$

where d represents the vector of row-sums. Normalizing the matrix gives as the normalized modularity matrix which helps us to cluster tightly connected networks (Bolla 2011.).

$$M_D = D^{-\frac{1}{2}}(W - dd^T)D^{-1/2} \quad (7)$$

Equation 5. shows that L is the negative discrete Laplace operator hence its eigenvectors would be sine and cosine functions. Analogously to noise filtering techniques we can calculate a Fourier approximation. However, we would like to optimize the normalized modularity cut thus we have to use the normalized modularity matrix. The optimal representation of the original vertices are the rows of that matrix which contains its columns the eigenvectors of in absolute sense largest eigenvalues of the normalized modularity matrix (Bolla 2013.);

$$\left(D^{-\frac{1}{2}}u_1, \dots, D^{-\frac{1}{2}}u_k\right)$$

where $u_1, \dots, u_2$ are the corresponding eigenvectors of $|\lambda_1(M_D)| \geq \cdots \geq |\lambda_k(M_D)| \gg |\lambda_{k+1}(M_D)|$.
To unveil the market implied classification we should calculate the similarity matrix, then determine the normalized modularity, identify the spectral gap in the spectrum of the normalized modularity matrix and finally clustering the optimal representation with k-mean algorithm.
The only hurdle is the calibration of edge weights. Note that a spectral clustering method is effective if and only if in absolute sense decreasing sequence of eigenvalues goes to zero and gap appears in the spectrum. Otherwise, spectral based methods cannot give appropriate classification. Hence, we should test different similarity measures thus we have to calculate the spectrum of different similarity measures implied normalized modularity and Laplace matrices and choose that which provides the best spectral properties. Ormos and Nagy showed tightly connected financial dataset can be analysed with Gaussian square distances.

$$W_{ij} = e^{-||s^i - s^j||^2} \quad (8)$$

Combining Fourier analysis, pattern recognition techniques, random walk and graph theory provides us the optimal cluster property, thus spectral clustering gives us the opportunity to find an approximation of the market implied classification.

FINANCIAL MARKET IMPLIED CLASSIFICATION

Spectral clustering can be used to unveil the hidden market structure of stock indices. The purely quantitative approach needs only closing prices so that the market segments are formed by stock prices. The fundamental assumption behind the model is that at least the weak form of market efficiency holds on a daily scale. This assumption allows intra-day inefficiencies, but accepts that the daily closing auction forces the market into the equilibrium.

Data

The current study identifies FMIC based on stock splits and dividends adjusted daily closing prices between 01/01/2007 and 01/03/2017 of current S&P 500 constituents. The data is provided by Yahoo! Finance.

FMIC and GICS

Calculating the spectrum of normalized Laplace and modularity matrices of Gaussian based similarities shed some light on the structure of the network. Bolla showed that if a graph is dense then the normalized Laplacian matrix cannot be used because the norm of eigenvalues slowly converge to zero. Normalized modularity matrix, however, provides better spectrum properties.

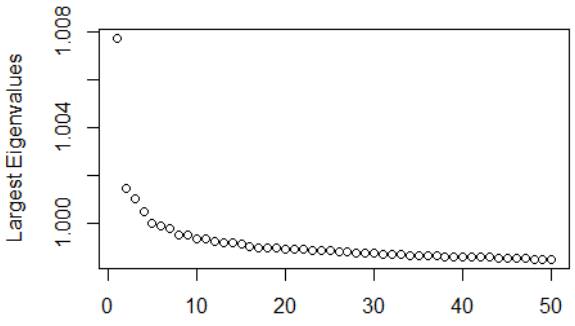

Figure 1: Spectrum of Gaussian based normalized Laplace matrix

Calculating the spectrum of normalized Laplace matrix displays that the eigenvalues converge slowly to zero thus Fourier based approximation techniques cannot be used (Figure 1.). It also suggests that the network structure is not scarce, different clusters should be connected. Hence, we should optimize the information theory based Newman Girvan cut thus we have to calculate the spectrum of normalized modularity matrix.

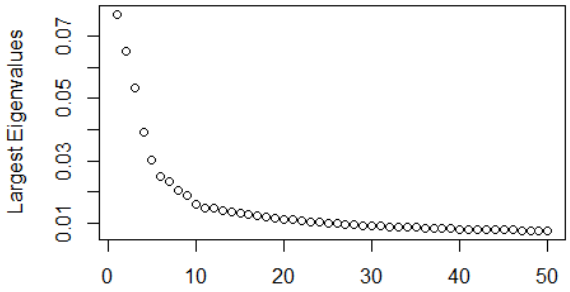

Figure 2: Spectrum of Gaussian based normalized modularity matrix

The spectrum of normalized modularity matrix provides appropriate spectral properties. Figure 2. shows that it has several large eigenvalues and the decreasing sequence of them converges to zero. Nevertheless, the normalized Laplacian cannot be used for clustering. All these implies that the equity index network is dense, most of the stocks are connected.
Identifying spectral gaps highlights that the optimal number of market implied clusters would be 5, 7, 9 or 12.
Note that GICS distinguishes 11 sectors. If we calculate FIMC with 11 clusters then we get controversial results.

Table 1. shows that the 7th cluster dominates the graph, most of the nodes are put into that cluster despite of the cluster size penalty. Moreover, GICS and FIMC 11 classifications are fundamentally different.

Table 1: Frequency table of GICS and FMIC 11

GICS/FMIC 11											
	1	2	3	4	5	6	7	8	9	10	11
D	10	3	10	5	4	8	21	10	7	4	4
S	4	3	0	2	3	2	7	4	3	4	4
E	1	0	4	1	3	2	13	2	5	4	0
F	9	3	2	3	8	5	9	6	9	0	10
H	8	4	0	3	4	1	17	13	6	1	2
I	6	4	5	3	4	4	24	2	8	4	2
IT	8	2	9	4	7	2	15	4	10	1	6
M	3	2	2	5	0	2	6	1	4	0	0
R	4	4	1	3	0	3	8	1	3	2	1
T	0	0	0	0	0	0	2	0	1	2	0
U	2	1	5	1	2	1	10	1	2	1	2

Notes: This table is the contingency table of GICS and FMIC 11; D denotes Consumer Discretionary, C: Consumer Staples, E: Energy, F: Financials, H: Health Care, I: Industrials, IT: Information Technology, M: Materials, R: Real Estate, T: Telecommunications Services and U: Utilities.

However, spectral clustering based methodology proposes to use 12 clusters. If we calculate FMIC 12 then we could see that GICS is in line with market implied classification, see Table 2.

Table 2: Frequency table of GICS and FMIC 12

GICS/FMIC12												
	1	2	3	4	5	6	7	8	9	10	11	12
C	0	8	33	3	0	0	0	0	0	42	0	0
S	0	0	5	1	0	0	0	0	0	4	0	26
E	31	2	2	0	0	0	0	0	0	0	0	0
F	0	5	9	0	0	50	0	0	0	0	0	0
H	0	0	5	0	0	0	0	39	0	0	16	0
I	0	5	55	0	0	0	0	0	6	0	0	0
IT	0	6	15	47	0	0	0	0	0	0	0	0
M	4	1	19	0	0	0	0	0	0	1	0	0
R	0	1	5	1	23	0	0	0	0	0	0	0
T	0	1	4	0	0	0	0	0	0	0	0	0
U	1	0	1	0	0	0	26	0	0	0	0	0

Notes: This table is the contingency table of GICS and FMIC 12

Clusters 1, 4, 5, 6, 7 and 12 cover Energy, Information Technology, Real Estate, Financials, Utilities and Consumer Staples respectively.
The first cluster is dominated by Energy companies. It also contains four Utilities and NRG Energy. These firms are closely related to energy business. Thus, we could say that GICS Energy sector can be quantitatively supported.
Information Technology names are put into Cluster 4. Checking non Information Technology names in Cluster 4 we see that the most of them are highly technological. The three Consumer Discretionary companies are Amazon Inc., Expedia Inc. and Garmin Ltd. that all carry out strongly technology based business. Equinix specialized in data and cloud business thus putting it to a technological cluster is also in line with economic thinking. Monster Beverage could be surprising, however, if we compare it with its peers (Table 3.) we could see that Monster Beverage differs from them.

Table 3: Peer review of Monster Beverage Corporation (Reuters)

Company Name	EV/ Revenue	Price/Book Value Per Share	Price / Revenue	Dividend / Share Yield %
Monster	7.97	7.9	8.19	0.00%
Pepsi	2.95	12.98	2.6	2.70%
Dr Pepper	3.17	8.11	2.52	2.50%
Starbucks	3.85	14.43	3.78	1.70%
Cott	1.18	2.15	0.5	1.80%

Notes: This table contains the peer review of of Monster Beverage Corporation.

Cluster 5 gives mathematical evidence to handle Real Estate as a different sector. Moreover, Financials, Utilities and Consumer Staples are also separated in Cluster 6, 7 and 12 respectively.
Clusters 8, 9, 10 and 11 incorporates Health Care, Industrials and Consumer Discretionary firms. Table 2. shows that these clusters split GICS sectors into two parts. Scrutinizing the results we could see that names are divided along GICS sub industries (Appendix 1).
Cluster 9 is the smallest cluster which encompasses 6 Industrial firms. Studying the sub industry classification we see that 5 companies out of 6 are Airlines and the outlier is Fortive Corporation which has large exposure to aviation business.
In conclusion, we can say that FIMC 12 is in line with GICS sectors and sub-industries, however it gives us a purely quantitative technique to identify clusters.

The spirit of CAPM

Understanding systematic risk is essential part of asset allocation because prices can be scrutinized only within an equilibrium model. Researchers, investors, regulators and banks are seeking models which could distinguish systematic and non-systematic risk.
Following a purely market oriented prospect leads to the Capital Asset Pricing Model (CAPM) which incorporates the systematic risk into the market portfolio. Several empirical studies concluded that CAPM can be used as a benchmark, but has to be made it more precise.
Analysing risk and reward in different frameworks explores different aspects of risk. Note that while standard deviation counts all the moves, β takes into account only that moves which can be explained linearly with the fluctuation of market portfolio. However, investors are sensitive to losses, filtering out therefore gains leads to Expected Downside Risk (Ormos and Timotity 2016.), in addition, various information theory based measures (Ormos and Zibriczky 2014.) can be defined which are strongly connected to log-optimal portfolio theory (Urban and Ormos 2013.).

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta \cdot SRM + \varepsilon \qquad (9)$$

All the linear models explain different aspects of systematic risk (SRM) but they can be compared with regression statistics. Correspondingly, if we add GICS and FMIC 12 to the regressions then we could compare them.

$$\begin{cases} \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 SRM + \beta_2 GICS + \epsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 SRM + \beta_2 FMIC12 + \theta \end{cases} \qquad (10)$$

Table 4. shows that the FMIC 12 outperforms GICS, except when risk is characterized by entropy and semi-variance.

Table 4: Regression statistics of GICS and FMIC 12

	p-value risk	R2	p-value GICS	R2	p-value FMIC 12	R2
Var	0.000	0.110	0.000	0.173	0.000	0.185
Sig	0.000	0.104	0.000	0.169	0.000	0.179
Semi-var	0.151	0.004	0.000	0.111	0.000	0.106
CAPM	0.000	0.187	0.002	0.232	0.002	0.234
EDR	0.000	0.143	0.000	0.210	0.000	0.227
H	0.000	0.026	0.000	0.122	0.000	0.117

Notes: Table 4. contains the regression statistics of different systematic risk models with GICS and FIMC 12.

Notice that semi-variance is not significant and entropy could explain only 2.6% of total variance. Hence, we could say that risk measures with high R^2 values can be used with cluster-variable like FIMC or GICS. Otherwise, there is no need adding cluster-variable to regression, because linear terms are not explained, thus the risk measure captures the same non-linear effect. If we would like to compare the behaviour of entropy (**H**) with FIMC and GICS we have to look the following models:

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = r^f + \beta_1 \cdot SRM + \beta_2 \cdot \mathbf{H} + \varepsilon$$

Calculating the regressions we get Table 5.

Table 5: Comparison of risk measures

Entropy	p-value of entropy	R2
Variance	0.164	0.114
Sigma	0.161	0.107
Semi-variances	0.000	0.029
CAPM Beta	0.085	0.191
EDR	0.090	0.148

Notes: This table presents the regression statistics of variance based systematic risk models with entropy.

The results show that entropy works differently, because using it as a cluster-variable adds only little explanatory power. However FIMC and GICS split the date such that the cluster-wise risk is linear in the main risk factor.

Figure 3. highlights that spectral clustering identifies concave clusters. Thus, adding the industry classification to the regressions filters out non-linear effects.

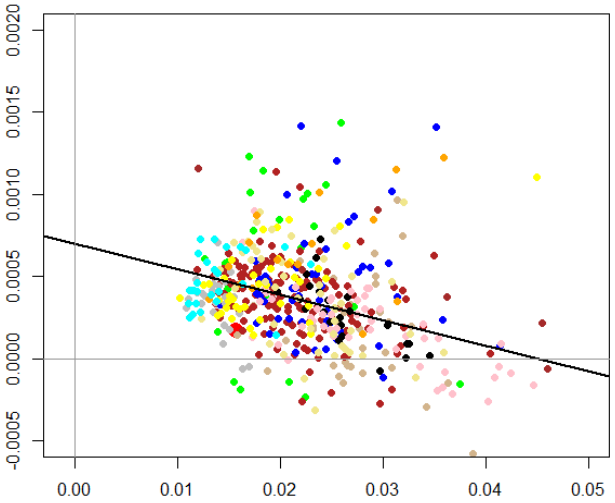

Figure 3: Standard deviation, mean return plot of FMIC 12

Nevertheless, the outcomes imply negative connection between risk and reward which is controversial with economic sense.
Analysing the market implied risk free rates (the interception of the regressions) we see that entropy is the only risk metric whose regression estimation (Table 6.) is in line with the market conventions.

Table 6: Estimated interceptions and p-values

Risk measure	Interception	p-value	Implied 1-year rate
Variance	0.001	0.000	0.207
Sigma	0.001	0.000	0.291
Semi-variances	0.000	0.000	0.148
CAPM Beta	0.001	0.000	0.283
EDR	0.001	0.000	0.341
Entropy	0.000	0.255	0.032

Note: Table 6. summarizes the regression statistics of the constant term.

The source of the problem could be the extreme low interest rate environment. If we calculate regressions without the constant term,

$$\begin{cases} \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i(SRM)^i + \beta_{GICS}GICS + \varepsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i(SRM)^i + \beta_{FMIC\ 12}FMIC\ 12 + \theta \end{cases} \quad (11)$$

we get positive connection between risk and reward.

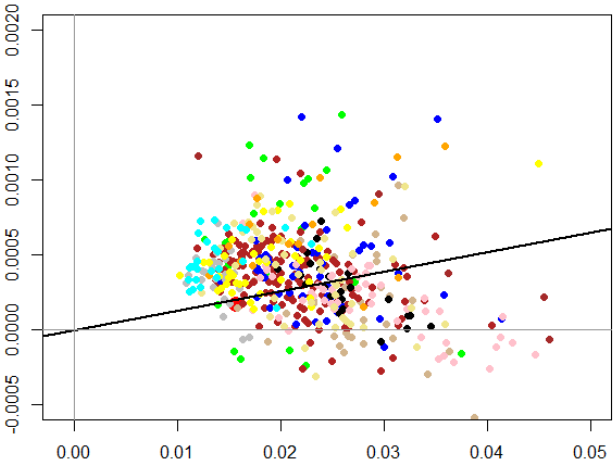

Figure 4: Standard deviation, mean return plot of FMIC 12 with zero interception

Setting the constant coefficient zero and calculating the regressions gives Table 7. which is in line with the 2007-2017 market conditions. Moreover, comparing the position of S&P 500 in Figure 3. and 4. gives further evidence to use regressions without constant, because, the benchmark portfolio (red dote) is on the regression line in Figure 4, meanwhile in Figure 3. is far away from the model expectations.

Table 7: Regression statistics of GICS and FMIC 12 with zero interception

	p-value risk	R2	p-value GICS	R2	p-value FMIC 12	R2
Var	0.000	0.203	0.000	0.595	0.000	0.601
Sigm	0.000	0.374	0.000	0.593	0.000	0.598
Semi-var	0.000	0.253	0.000	0.565	0.000	0.562
CAPM	0.000	0.277	0.000	0.624	0.000	0.625
EDR	0.000	0.375	0.000	0.613	0.000	0.622
H	0.000	0.522	0.000	0.571	0.000	0.568

Notes: This table contains the regression statistics of Equation 12.

It also can be seen that different linear based risk factors explain 20%-52% of the total variance, but adding cluster-variables the explanatory power of the model jumps to 60%. This means that cluster specific and linear risks explains 60% of the fluctuation.

If we generalize the baseline linear model we could specify the following regression;

$$\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \sum_{i=1}^{10}\beta_i \cdot (SRM)^i + \varepsilon \qquad (12)$$

Risk / R2 of n-th order	1st	2nd	3rd	4th	5th	10th
Var	0.203	0.365	0.495	0.537	0.567	0.582
Sig	0.374	0.547	0.568	0.568	0.569	0.581
Semi-var	0.253	0.348	0.411	0.454	0.477	0.518
CAPM	0.277	0.559	0.574	0.575	0.591	0.626
EDR	0.375	0.587	0.593	0.594	0.595	0.598
H	0.522	0.524	0.526	0.527	0.532	0.535

Expanding the linear model with higher order terms could shed more light on non-linear dependencies, see Table 8.

Table 8: Estimated R2 statistics of polynomial

Notes: Table 8. highlights the non-linear connections between returns and systematic risk factors.

The results are in line with our expectations, because, adding higher order terms of linearly inspired risk metrics to the regression increases the explanatory power, while entropy shows different behavior.
Generalizing Equation (11) we could get the following models;

$$\begin{cases}\mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \beta_1 SRM + \beta_2 GICS + \varepsilon \\ \mathbf{E}\left(\ln\frac{S_t}{S_{t-1}}\right) = \beta_1 SRM + \beta_2 FMIC\ 12 + \theta\end{cases} \qquad (13)$$

Equation (13) lets us to distinguish cluster and non-cluster specific higher order connections. The results (Table 9.) show that polynomial terms explain similar effects like GICS and FMIC 12.

Table 9: Estimated R2 statistics of polynomial regressions with FMIC 12

Risk/ R2 of n-th order model with FMIC 12						
	1st	2nd	3rd	4th	5th	10th
Var	0.601	0.602	0.605	0.606	0.609	0.616
Sig	0.598	0.601	0.604	0.605	0.608	0.614
Semi-var	0.562	0.563	0.564	0.565	0.566	0.569
CAPM	0.625	0.636	0.637	0.640	0.640	0.652
EDR	0.622	0.634	0.634	0.634	0.635	0.636
H	0.568	0.569	0.569	0.573	0.573	0.577

Notes: This table summarizes the non-linear filter behaviour of the cluster variable (FMIC 12).

Tables 8, 9. and 10. show that higher order terms do not increase the explanatory power of the regressions. Thus, the remaining 30% part of the variance cannot be explained by liner-based risk factors and clusters.

Table 10: Estimated R2 statistics of polynomial regressions with GICS

Risk measures/ R2 of n-th order model with GICS						
	1st	2nd	3rd	4th	5th	10th
Var	0.595	0.596	0.597	0.598	0.600	0.607
Sig	0.593	0.595	0.597	0.597	0.599	0.606
Semi-var	0.565	0.566	0.566	0.566	0.567	0.570
CAPM	0.624	0.631	0.632	0.634	0.634	0.646
EDR	0.613	0.619	0.620	0.621	0.622	0.624
H	0.571	0.571	0.572	0.574	0.574	0.578

Notes: Table 10. describes the non-linear filter behaviour of GICS.

Analysing returns with linear and polynomial regressions show that FMIC 12 outperforms GICS, however, cluster variables explain similar non-linear connections.

Optimal number of clusters

Analysing the structure of FIMC 12, GICS industries and subindustries the natural question arise; how many clusters do we need?
In spectral clustering and signal analysis there are different analytical and simulation based techniques. The most widely used methodology is spectral gap analysis which is a Fourier based technique. The goal is to approximate the norm of the object with as few elements as possible.
Calculating the spectrum of normalized modularity matrix (Figure 2) shows that the GICS and FIMC 12 could be too elaborated and 5 clusters would be enough to explain non-linear cluster specific effects.
Comparing FMIC 12 with FMIC 5 it shows that FIMC 12 is a more detailed subdivision of the stock market graph. Because, FIMC 5 bundles the FIMC 12 clusters into bigger groups (Table 11.).
However, Real Estate remains a single cluster and Consumer Staples with Utilities are put into FIMC 5 Cluster 3. Other FIMC 5 clusters are also dominated with FIMC 12 clusters.

Table 11: Frequency table of FMIC 12 and FMIC 5

FMIC 12/FMIC 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	0	36	0	0	0
Cluster 2	4	13	0	0	12
Cluster 3	141	7	2	0	3
Cluster 4	46	5	0	0	1
Cluster 5	0	0	0	23	0

Cluster 6	3	0	0	0	47
Cluster 7	0	0	26	0	0
Cluster 8	5	33	1	0	0
Cluster 9	6	0	0	0	0
Cluster 10	43	4	0	0	0
Cluster 11	0	16	0	0	0
Cluster 12	3	7	16	0	0

Notes: This table is the contingency table of FIMC 12 and FMIC 5.

Analysing FMIC 5 we saw that it is closely related to GICS and FMIC 12. Regression statistics (Table 12.) are in line with Table 7, relative entropy and semi-variance show different behaviour and CAPM has the strongest explanatory power.

Table 12: Regression statistics of FMIC 5 with zero interception

Risk	p-value risk	R2	p-value GICS	R2	p-value FMIC 5	R2
Var	0.000	0.203	0.000	0.595	0.000	0.578
Sig	0.000	0.374	0.000	0.593	0.000	0.579
Semi-var	0.000	0.253	0.000	0.565	0.000	0.536
CAPM	0.000	0.277	0.000	0.624	0.000	0.614
EDR	0.000	0.375	0.000	0.613	0.000	0.604
H	0.000	0.522	0.000	0.571	0.000	0.543

Notes: Table 12 sheds some light on the optimal, lower dimensional market classification.

Regression and frequency statistics support the theoretically proposed five cluster model which incorporates roughly the same information like FIMC 12 and GICS.

Conclusion

Spectral clustering is an adequate technique to unveil the embedded market structure, filter out non-linear effects and make CAPM more precise. The purely quantitative method gives us the opportunity to categorize firms based on their stock market returns, in addition, it lends some colour to Global Industry Classification Standards.

ACKNOWLEDGEMENTS

Mihály Ormos acknowledges the support of the János Bolyai Research Scholarship of the Hungarian Academy.

REFERENCES

A. Urban and M. Ormos. 2013 “Performance analysis of log-optimal portfolio strategies with transaction costs” *Quantitative Finance*. 13: (10) pp. 1587-1597.

D. Timotity. And M. Ormos 2016. “Generalized asset pricing: Expected Downside Risk-based equilibrium modeling” *Economic Modelling* 52:pp. 967-980.

D. Zibriczky. and M. Ormos. 2014. “Entropy-Based Financial Asset Pricing” *PLoS* ONE 9(12): e115742.

Fan R. K. Chung. 1997. “Spectral Graph Theory” *American Mathematical Society*

J. Shi and J. Malik. 2000. “Normalized cuts and image segmentation” *Pattern Analysis and Machine Intelligence* IEEE, Transactions, N.J., 888-905.

L. Nagy and M. Ormos. 2016. “Friendship of Stock Indices” *30th Conference on Modelling and Simulation*

M. Bolla. 2011. “Penalized version of Newman-Girvan modularity and their relation to normalized cuts and k-means clustering” *Physical review E*, Vol. 84, 016108

M. Bolla. 2013. “Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables” *Wiley*

MSCI. 2015. “S&P Dow Jones Indices and MSCI announce Further Revision to the Global Industry Classification Standards (GICS®) Structure in 2016” *Working Paper*.

Executive Office of the President Office of Management and Budget. 2017. “North American Industry Classification System” *Working Paper*.

E. N, Vose. 1916 “Seventy-five years of the Mercantile agency, R.G. Dun & co., 1841-1916” *Brooklyn, N.Y., Priv. Print. at the Printing House of R.G. Dun & Co.*

U. von Luxburg. 2007. “Tutorial on Spectral Clustering” *Statistics and Computing* Vol. 17, 395-416.

V. S. Kolesnikoff. 1940. “Standard Classification of Industries in the United States” *Journal of the American Statistical Association*, Vol. 35, No. 209, pp. 65-73

AUTHOR BIOGRAPHIES

László Nagy is a PhD student at the Department of Finance, Institute of Business at the School of Economic and Social Sciences, Budapest University of Technology and Economics. His main area of research is financial risk measures and asset pricing

Mihály Ormos is a Professor of Finance at Eötvös Loránd University and at J. Selye University. His area of research is financial economics especially asset pricing, risk measures, risk perception and behavioral finance. He serves as one of the contributing editors at Eastern European Economics published by Taylor and Francis. His teaching activities concentrate on financial economics, investments and accounting.