COURSE – CS6242 DATA AND VISUAL ANALYTICS

Final Report
for
Enhanced Chilean Mutual Fund Data Explorer

By,
Pedro Pablo Ramirez,
Christopher J Santiago,
Collin G Kruger,
Shannon R Flynn,
Nagasree Chelamalla.

Introduction

In Chile, limited tools and static reports exist to help Chilean retail investors (average citizens who do not work for an investment firm) understand their investment options[1, 2]. We recognize this community as underserved and aim to help retail investors make well-informed decisions by serving intuitive data visualizations, informed by a combination of machine learning methods and investment theory. Our solution coalesces multi-platform data sets to hasten, and make possible, insights that are difficult to achieve with the status quo. Our initial research indicates that the results of this project will be highly impactful, as mutual funds are one of the primary investment avenues in Chile[3], representing nearly 20% of total assets under management (AUM) in Chile (see Table 1 in the Appendix), and an initial collection of traders found our software to be significantly more useful than what they use day to day.

TABLE 1: Assets Under Management

Institution	AUM	Market Share
Mutual Funds	62.500	18%
Pension Funds	200.000	58%
Insurance Companies	60.000	17%
Private Funds	25.000	7%

***MMUSD**

Data obtain from the Chilean Financial Regulator (Financial Market Commission "CMF") https://estadisticas.aafm.cl/

Problem Definition

Currently, retail investors must collect data, aggregate, clean, then apply common or novel analysis techniques to their data in the hopes to glean insights on what mutual funds to purchase. Collecting and cleaning the data is currently a manual process performed by manually downloading, then manually aggregating and cleaning with tools such as Excel. Analysis is then performed by extracting only a few days of the data. Excel exports only support up to 10 days of daily information. The data must then be processed to obtain price returns (month-to-date, year-to-date, etc.), and changes in asset under management. Due to the time required to go through that process, few (if any) retail traders take the time to do daily exports. Given this room for improvement, and in our case, we chose to reasonably automate the download, cleaning, analysis, and visualization of Mutual Fund historical and present data.

Survey

In industry internationally we see a range of state-of-the-art practices in finance, from "robo-advisors" that employ Modern Portfolio Theory (MPT, a way of statistically maximizing reward vs. risk) for the masses, to natural language processing (NLP) sentiment analysis with neural networks (NN) deployed by hedge funds with unfathomable resources. Chilean investors, however, are limited to price monitoring and investment consulting services. In Chile, much of the reasoning behind specific asset allocations are hidden behind "black boxes" that lack online, consumer-led interaction[1]. Further, investment companies often display bias in their recommendations, pushing their own funds onto investors. Overall, Chilean retail investors lack access to open and transparent investment analytics.

MPT provides a method for finding an optimal combination of assets that maximizes an investor's return, given their specific risk tolerance[7]. MPT can fall short, especially during financial crises, if it over-weights an investor's exposure to correlated risk assets. This happens if the covariance matrix (a mathematical way of defining similarity, in our case how similar mutual funds behave) is mis-specified, which can arise when assets are arbitrarily grouped based on investment styles or descriptions, vice underlying risk and return

profiles. We use semi-supervised clustering methods[6,11] to better classify Chilean mutual funds and improve MPT optimization.

Proposed Method

Our project had four primary steps: 1) consolidate disparate sources of Chilean financial data; 2) apply unsupervised machine learning (ML) techniques to identify funds with abnormal risk/return profiles compared to their stated objective[4] (AAFM category, a financial classification); 3) apply semi-supervised ML techniques[5] to re-classify funds based on their historic performance[6], vice their stated objectives, for portfolio optimization via Modern Portfolio Theory[7]; and 4) create a user interface (UI) to visualize distributions and key statistics of Chilean mutual funds, any identified abnormalities, and correlations of newly-clustered funds. Compared to the very manual process retail investors go through today, our software brings their experience into the modern age.

There are several unsupervised ML techniques[4,8] that, paired with intuitive visualizations, could help Chilean investors identify mutual funds that behaved abnormally compared to their stated classification. Originally, we hypothesized Autoencoders and Principle Component Analysis (dimensionality reduction techniques) would yield good results for finding outliers in mutual fund classification; however, at least for the nature of the data that we have, this was not the case due to the final useful data scale being smaller than anticipated (we started with ~4,000,000 records and have cleaned and transformed it down to 312 records).[9] We discovered a fairly recent technique of clustering called t-SNE, which became our resolution to the Autoencoder and PCA problem. We used t-SNE to project 72 periods of mutual fund returns into a two-dimensional space. This created a self-organizing map, with similar funds residing near each other in space. We calculated pairwise distances from our t-SNE embedding to identify funds whose Euclidean distance was far from their respective category mean, allowing us to identify anomalous or outlying funds, as compared to their respective categories. Along the way to this resolution, we experimented with several other dimensionality reduction embedding techniques including Multi-Dimensional Analysis[10], Isomap, and Uniform Manifold Approximation and Projection (UMAP)[20]. Ultimately, we decided that t-SNE best embedded our data for visualization in a two-dimensional space.

The t-SNE technique provides dimensionality reduction and projection onto a self-organizing map, but the algorithm does not output labels for each observation, which are needed to re-classify mutual funds based on their underlying performance metrics. We used the k-means algorithm to cluster the mutual funds based on each fund's average annual return, average annual risk (standard deviation) and the t-SNE embedding. K-means is a "semi-supervised" method that required use of several metrics to determine the optimal number of clusters[21, 22, 23]; our analysis indicated 10 clusters to be appropriate. We used these labels to group the mutual funds, construct a new covariance matrix and an efficient frontier using Modern Portfolio Theory.

The efficient frontier was built by simulating around six thousand portfolios. Further, each portfolio had a different weight of its corresponding assets that was iterated in the simulation. In this case, the assets were the mutual funds categories and reduce categories performed by the space reduction technique. Later, we grouped the return series of each portfolio to build a covariance and variance matrix to multiply weights, covariance and returns. Finally, we calculated the return and standard deviation of returns to generate points on a plot. Points plotted in this efficiency frontier are superior to samples outside the frontier. These points maximize the returns asked by certain risk premia. It is a frontier (area) instead of just a border, because an investor can choose an acceptable range of risk.

To provide an interactive UI, we used Tableau as our starting point. Tableau is beneficial because its dashboards and packaged workbooks can be accessed anywhere by anyone utilizing open and closed

available platforms, such as Tableau Reader, Tableau Server, Tableau Public, and embedded workbooks in personal websites. For this reason, we decided Tableau would be an excellent resource to the people of Chile for improving investment fund decision making.

The UI created in Tableau consists of four major dashboards for interaction:

1. Chilean Mutual Funds Anomaly Detection (Figure 1): This dashboard enables investors to quickly view a self-organized map of the available funds, observe outliers and fund misclassifications, and compare funds to their proper classifications. Filters are included, such as filters for asset categories and fund names, enable quickly scoping data. Lastly, a contextual filter exists accessed by clicking a bar to jump to a filtered dashboard for funds that exist within that AAFM category.
2. AAFM Category Fund Summary (Figure 2): This is the destination dashboard from the bar filter above. It is a table that shows the funds that are in the selected AAFM category, whether the fund is marked as an anomaly or not, average annual return, and average annual risk of each fund. This can be used by a user to gather additional information about available funds in an AAFM category they may be interested in.
3. Similar Chilean Mutual Funds (Figure 3): This dashboard enables users to view funds that have been classified as similar to a selected fund from a table. Fund similarity is designated solely as whether funds fall into the same AAFM Category. The information that is displayed includes a line chart of time series data of historical return information of the selected fund, scatter plots to show how other funds compare to the chosen fund for risk/return, and various other features. Filters are available to allow users to choose which fund's data they would like to see.
4. Modern Portfolio Theory (MPT) (Figure 4): This dashboard is used to display the Modern Portfolio Theory options for AAFM categories and K-means clustering categories. Users are able to quickly visualize how different portfolio options compare to each other, and they are colored on a user-entered Sharpe ratio (a financial performance metric). Each portfolio category distribution is shown in a tooltip. The ideal portfolio, determined by the highest Sharpe ratio, for each category set is shown as a red star to allow the user to quickly identify the best option for their investment.

Figure 1: Anomaly Detection Dashboard

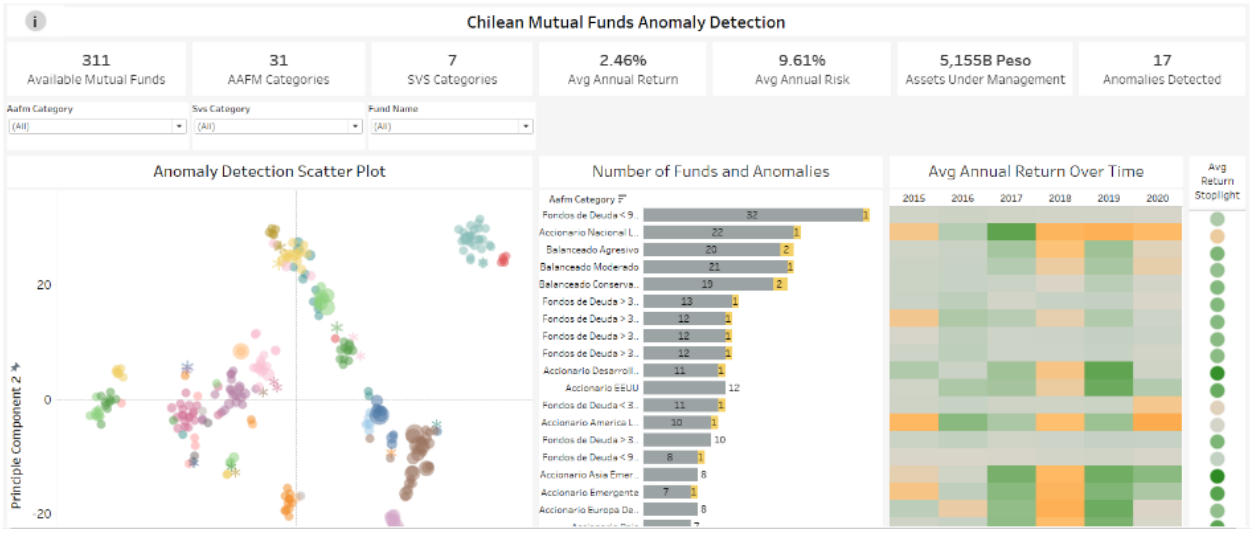


Figure 2: AAFM Category Fund Summary Dashboard

AAFM Category Fund Summary

Aafm Category	Fund Name	Anomaly	Annual Return	Annual Risk
Balanceado	ACTIVA C	no	3.82%	6.84%
Moderado	ACTIVO EQUILIBRADO	no	5.44%	6.91%
	ACTIVO MODERADO	no	5.27%	5.49%
	BALANCEADO MODERADO	no	2.83%	6.87%
	BANCOESTADO PERFIL TRADICIONAL C	no	5.23%	5.73%
	BICE EST AGRESIVA	no	4.47%	8.48%
	BICE EST BALANCEADA	no	5.13%	6.89%
	BTG PACTUAL GESTION ACTIVA	no	4.83%	7.04%
	ESTRATEGIA EQUILIBRADA	no	2.83%	6.70%
	ESTRATEGIA MODERADA	yes	-28.24%	37.22%
	FMCODBAL	no	3.40%	7.39%
	ITAU GESTIONADO MODERADO	no	5.54%	7.67%

Figure 3: Similar Chilean Mutual Fund Dashboard

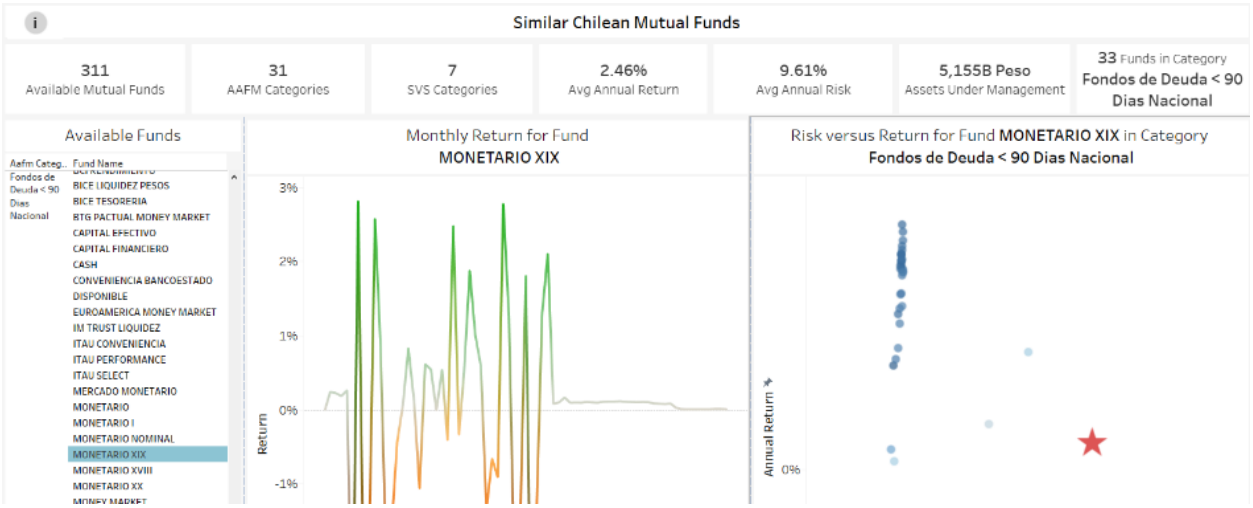
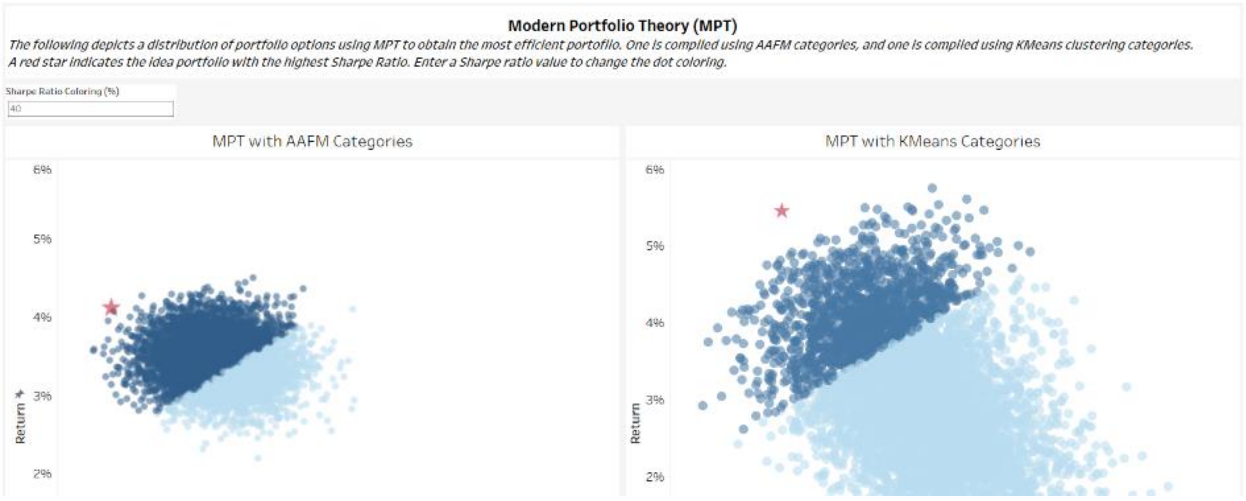


Figure 4: Modern Portfolio Theory Dashboard



Bugs are always a present danger in software development. Given that our software is ambitious and that we are writing it outside of our daily jobs, there is significant risk that we implement an algorithm incorrectly, transform data incorrectly, or other similar errors. Studies in software development have found that, typically, there are "about 15 – 50 errors [occur] per 1,000 lines of delivered code[12]." This means throughout our process we've had multiple eyes on our code, helping reduce the possibility of errors. However, given this project had a very aggressive deliverable timeline, we recommend that our initial users should not use our software as their only source of information—perhaps using it for simulated trades, only, until they feel comfortable with the data and software reliability. In our opinion, the possibility of greatly benefiting the people of Chile outweighs the risks.

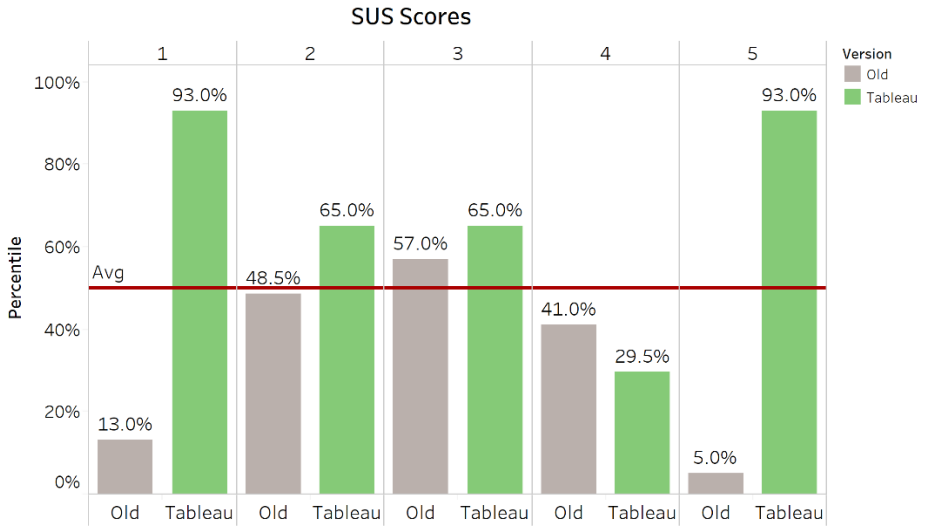
Design of experiments / evaluation

Given timeline and geographic limitations, a full long term user study was not feasible. Therefore, to get a rough understanding of usage/impact, we could observe web traffic data, including geographic information, to understand how many people are utilizing the tool and where they are located. Once the semester has concluded, we are handing our software over to our Chilean teammate to open source, expand on its capabilities, extend where appropriate, embed it in a public facing website, and evangelize its usefulness to the Chilean retail investment community through his work at Itau Chile Bank.

In lieu of a full user study, usability analysis was conducted to gauge the ease of use of the new tool compared to a privately available investment bank platform. Our Chilean teammate recruited participants from his work to use their current tool, our new tool, and then complete a System Usability Scale (SUS) industry standard survey. The SUS is a survey used often in the industry to obtain a subjective measurement of the usability of a website, online tool, or other interactive system. There are 10 questions measured on a likert scale, and scores range from 0 to 100. Each score is then converted into a percentile ranking to determine how the tool compares to all other tools measured using the SUS. This survey enabled us to make a reasonable estimation of whether or not our tool in fact improved on the status quo. We did this with the knowledge that we implicitly assert that the tools available to investment bank platforms are better than manually scraping data, and transforming it with Excel, the status quo for retail investors.

Figure 5 shows a summary of the results from the usability study. The results suggest that the new tool in Tableau is an improvement over the current privately available tool. Overall, the average SUS score for the new Tableau tool is 69% and the average for the current tool is 33%. This means that our Tableau tool scored higher on usability than 69% of all systems/tools tested, and the current tool is only better than 33% of all systems/tools. It is important to note that a new tool, such as our Tableau tool, typically does not score above a 50% on the SUS. This would indicate that our tool is headed in the right direction and should be seriously considered as an alternative to the current tool they have available.

Figure 5: SUS Usability Scores by Participant



It is important to note that our sample size was small due to time constraints and availability of participants due to the COVID environment we are currently in. However, we believe that the results we obtained are representative of how the larger population will feel, and therefore we are confident in our results. The participants were all main users of the current tool, and they noted that the ability to have an interactive tool for their analysis would greatly improve their decision-making abilities. As a follow on to this project, additional participants should be utilized for a larger usability study with a larger sample size. It is also important to note that there was a language barrier with this project; our Tableau tool and SUS questionnaire are both in English, and the participants are primarily Spanish speaking. Our Chilean teammate interpreted all the information for them, but this language barrier could have potentially influenced some of the results, such as participant 4, who rated the old tool, which is in Spanish, as better than the new tool.

Conclusion/Discussion

In conclusion we find that our software is more usable than what is privately available to an investment bank. From this we make the claim that our software will improve the lives of retail investors who do not have access to the same resources as an investment bank.

During our work with Chilean traders, many areas of improvement were uncovered. For instance, one of the traders noted that though the tool was much easier to use, it did need some deeper information for it to be usable to him as part of his daily analysis. Also, there was a large consensus that having more of our analysis charts be interactive would be helpful. These we see as excellent work items for continued development later this year.

Lastly, due to time and immediate data availability constraints, there were several areas of research that we did not include in this project: 1) how political parties affect the selection of mutual funds[13]; 2) how predicting the performance of constituent stocks can be used to predict mutual fund performance[14]; 3) application of evolution-inspired algorithms to make predictions[15]; and 4) using support vector machines (SVM) to identify types of assets and groups for building a covariance matrix[16]. These ideas would be solid options for follow-on work, given more time and data.

All team members contributed similar amounts of work to this project.

References:

1. Informes de Recomendacion (2020). Sura. SURA Asset Management. https://inversiones.sura.cl/nosotros/Paginas/informes-de-recomendacion.aspx
2. IPSA, Chile (2021). Btg Pactual. CME Group. https://www.mercadosenlinea.cl/www/chile/resume.html
3. Ahumada, L., Alvarez, N., & Diego, S. (2011). Valorización de Fondos Mutuos Monetarios y su Impacto sobre Estabilidad Financiera. Central Bank of Chile.
4. Hastie, T., Tibshirani, R. & Friedman, J. (2009). *Unsupervised learning.* The elements of statistical learning (pp. 501-528). New York, New York: Springer Science+Business Media, LLC.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An introduction to statistical learning with applications in R. New York, N.Y: Springer.
6. Mehta, D., Desai, D. & Pradeep, J. (2020). *Machine learning fund categorizations.* ArXiv.
7. Markowitz, H. (1952). Portfolio selection. Journal of Finance, vol 7(1), (pp. 77-91).
8. Dixon, R., Halperin, I. & Bilokon, P. (2020). *Autoencoders*. Machine learning in finance: From theory to practice (pp. 266-271). Cham, Switzerland: Springer Nature Switzerland AG.
9. Kim, M., et al. (2000). Mutual Fund Objective Misclassification. Journal of Economics and Business (pp. 309–323.).
10. Aflalo, Y., Dubrovina, A. & Kimmel, R. (2016). Spectral Generalized Multi-dimensional Scaling. Int J Comput Vis 118, 380–392. https://doi.org/10.1007/s11263-016-0883-8
11. Pattarin, F., et al. (2004). Clustering Financial Time Series: an Application to Mutual Funds Style Analysis. *Computational Statistics & Data Analysis (*pp. 353–372).
12. McConnell, S. (2004). Code Complete, Second Edition. Microsoft Press, USA.
13. Bubb, R. & Catan, E. (2020). The Party Structure of Mutual Funds. European Corporate Governance Institute - Law Working Paper 560/2020. http://dx.doi.org/10.2139/ssrn.3124039
14. Li, B. & Rossi, A. G. (2020). Selecting Mutual Funds from the Stocks They Hold: A Machine Learning Approach http://dx.doi.org/10.2139/ssrn.3737667
15. Kyong Joo Oh, Tae Yoon Kim, Sungky Min (2005). Using genetic algorithm to support portfolio optimization for index fund management. Expert Systems with Applications, Volume 28, Issue 2, Pages 371-379, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2004.10.014
16. Guglietta, J. (2018). *Support vector machine-based global tactical asset allocation.* Big data and machine learning in quantitative investment (pp. 211-224). John Wiley & Sons, Ltd.
17. Tufte, E. R. (2001). The visual display of quantitative information. Graphic Press.
18. Geraldi, J., & Arlt, M. (2015). *Visuals Matter! Designing and using effective visual representations to support project and portfolio decisions.* Project Management Institute.
19. van der Maaten, L. & Hinton, G. (2008). Visiualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov):2579-2605.
20. Leland McInnes, John Healy, James Melville (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/abs/1802.03426
21. T. Calinski and J. Harabasz, 1974. "A dendrite method for cluster analysis". Communications in Statistics
22. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53-65.
23. Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227