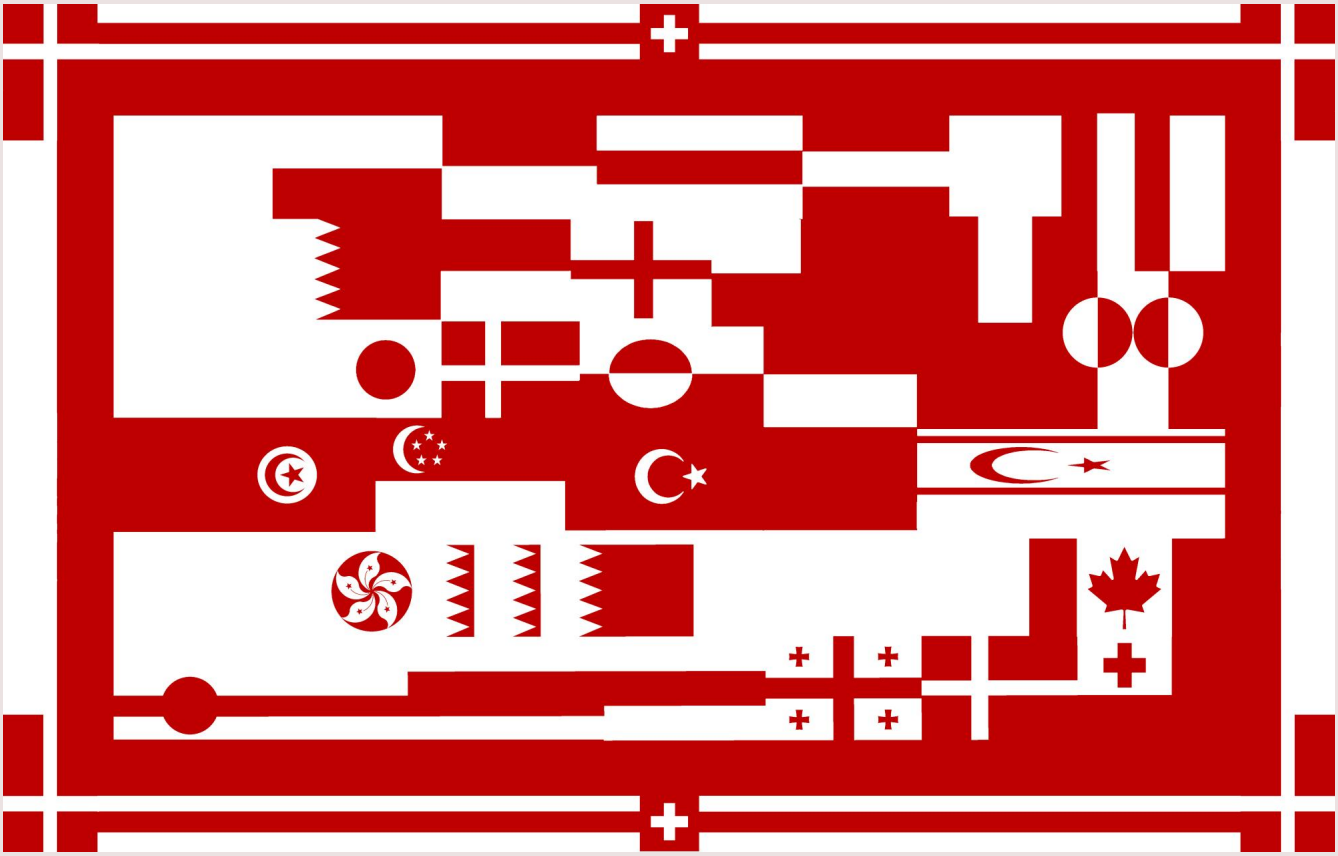
Annals of Computer Science and Information Systems
Volume 29

Recent Advances in Business Analytics

Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics

November 11-12, 2021
Olomouc, Czech Republic



Jan Stoklasa, Pasi Luukka and Maria Ganzha (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 29

ISSN: 2300-5963

Recent Advances in Business Analytics

Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics

Jan Stoklasa, Pasi Luukka and Maria Ganzha (eds.)

POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

Annals of Computer Science and Information Systems, Volume 29
Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics

USB: ISBN 978-83-962423-6-5
WEB: ISBN 978-83-962423-7-2
ISSN: 2300-5963
DOI: 10.15439/978-83-962423-7-2

Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover art: Flag
Jana Waleria Denisiuk,
Elbląg, Poland

Also in this series:

Volume 28: Proceedings of the 2021 International Conference on Research in Management & Technovation, **ISBN WEB: 978-83-962423-4-1, ISBN USB: 978-83-962423-5-8**
Volume 27: Proceedings of the Sixth International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-962423-2-7, ISBN USB: 978-83-962423-3-4**
Volume 26: Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-959183-9-1, ISBN USB: 978-83-962423-0-3**
Volume 25: Proceedings of the 16th Conference on Computer Science and Intelligence Systems, **ISBN Web 978-83-959183-6-0, ISBN USB 978-83-959183-7-7, ISBN ART 978-83-959183-8-4**
Volume 24: Proceedings of the International Conference on Research in Management & Technovation 2020, **ISBN WEB: 978-83-959183-5-3, ISBN USB: 978-83-959183-4-6**
Volume 23: Communication Papers of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-959183-2-2, ISBN USB: 978-83-959183-3-9**
Volume 22: Position Papers of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-959183-0-8, ISBN USB: 978-83-959183-1-5**
Volume 21: Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-955416-7-4, ISBN USB 978-83-955416-8-1, ISBN ART 978-83-955416-9-8**
Volume 20: Communication Papers of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-955416-3-6, ISBN USB: 978-83-955416-4-3**
Volume 19: Position Papers of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-955416-1-2, ISBN USB: 978-83-955416-2-9**
Volume 18: Proceedings of the 2019 Federated Conference on Computer Science and Information Systems, **ISBN Web 978-83-952357-8-8, ISBN USB 978-83-952357-9-5, ISBN ART 978-83-955416-0-5**
Volume 17: Communication Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-952357-0-2, ISBN USB: 978-83-952357-1-9**
Volume 16: Position Papers of the 2018 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-949419-8-7, ISBN USB: 978-83-949419-9-4**

DEAR readers. Business analytics as a field is becoming more and more relevant for practice, companies are trying to utilize the data they have and obtain more relevant data in order to manage and optimize their processes as well as to reach their goals in all relevant areas including economic, societal, ecological, and sustainability-related goals. The demand for university education that would combine information science, operations research and management science, statistics and insights into economics, business and industry is still growing. Business analytics in all its forms and application areas is also becoming a strong and wide field of research. This development is being mirrored also in the growing number of conferences and research seminars that are being organized on various topics related to business analytics.

This year was the first that brought the North-European Society for Adaptive and Intelligent Systems (NSAIS) and the international conference on Knowledge in Economics and Management (KNOWCON) together to organize a joint KNOWCON-NSAIS workshop on business analytics. This event took place on November 11-12, 2021 in Olomouc, Czech Republic in the historical buildings of Palacký University Olomouc. It brought together researchers and practitioners from the field of business analytics as well as students of business analytics, economics and related subjects, and provided an inspiring place for scientific discussion of new ideas, problems to be solved and methods that are being developed by the researchers in the field. The topics discussed in the workshop and the related social events covered the use of analytics in various areas of business and finance as well as the development of new instruments and models for business and data analytics, for the processing of social science and business data, operations research, intelligent systems, machine learning and soft-computing methods, their development, analyses and use in the business and financial setting.

This issue of ACSIS presents the selected full papers the contents of which were presented and thoroughly discussed at the KNOWCON-NSAIS workshop. We are very happy that we can share with you these contributions that range from the development of machine learning methods and their application through econometric analyses to solve business problems to multiple-criteria decision-making methods dealing with uncertainty to address business, managerial and social science decision-making and evaluation problems. All the papers of this issue went through a rigorous review process by at least two independent reviewers and the assessment by the KNOWCON-NSAIS scientific committee members.

We would like to thank Palacký University Olomouc, Faculty of Arts, Department of Economic and Managerial Studies, mainly to associate professor Pavla Slavíčková, for organizing the whole event and to NSAIS for the cooperation on the organization of the workshop. We would also like to extend our thanks to all the researchers, scholars, practitioners and students that took part in the discussions and the presentations of the current research results, and also in the review process, for maintaining a high scientific quality of the discussions and for creating a very pleasant and inspiring atmosphere to share ideas and open problems and to find innovative solutions. Last but not least we would like to thank ACSIS and its editorial board for their support and kind collaboration on the creation of this issue and for the opportunity to share the recent advances in the field of business analytics represented by the selected full papers published in this issue with a wide audience of readers.

Jan Stoklasa, Pasi Luukka and Maria Ganzha, editors of this issue.

Annals of Computer Science and Information Systems, Volume 29

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics

November 11–12, 2021. Olomouc, Czech Republic

TABLE OF CONTENTS

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 1–6
DOI: 10.15439/2021B8
ISSN 2300-5963 ACSIS, Vol. 29

Material demand forecasting with classical and fuzzy time series models

Sergey Zakrytnoy
Sievo Oy, Mikonkatu 15 A,
00100 Helsinki, Finland,
Email: sergey.zakrytnoy@sievo.com

Pasi Luukka
School of Business and
Management, LUT University,
Yliopistonkatu 34,
53851 Lappeenranta, Finland
Email: pasi.luukka@lut.fi

Jan Stoklasa
School of Business and
Management, LUT University,
Yliopistonkatu 34,
53851 Lappeenranta, Finland, and
Palacky University Olomouc,
Faculty of Arts, Department of
Economic and Managerial Studies
Email: jan.stoklasa@lut.fi
Email: jan.stoklasa@upol.cz

***Abstract*—Direct material budgeting is an essential part of financial planning processes. It often implies the need to predict quantities and prices of hundreds of thousands of materials to be purchased by an enterprise in the upcoming fiscal period. Distortion effects in demand projections and overall uncertainty cause the enterprises to rely on internal data to build their forecasts.**

In this paper we are dealing with material demand forecasting and evaluate the feasibility of fuzzy time series forecasting models as compared to classical forecasting models. Relevant methods are shortlisted based on existing practice described in academic research. Three datasets from industry are used to evaluate the predictive performance of the shortlisted methods. Our findings show an improvement in prediction accuracy of up to 47% compared to naïve approach. Fuzzy time series models are reported to be the most reliable forecasting method for the analyzed intermittent time series in all three datasets.

I. Introduction

MODERN digitalization technologies and computational methods provide new levers for business decision-support impacting financial performance of an enterprise. Many of those levers are to be found (either to originate or to be applied) in supply chain management. Chopra and Meindl [1] state that the objective of a supply chain is the maximization of the overall generated value, where value is defined as the difference between sales revenue and total incurred costs throughout the chain of decision-making units. With shortened delivery timelines, those units are looking to introduce supply chain forecasting (SCF) models in order to meet customer's demand with the highest possible efficiency in terms of accuracy of the forecast and the work effort required for its generation. While this paper analyzes material forecasting from requirements planning perspective, downstream demand forecasting has recently been outlined as a symmetrically important business challenge with major impact on profitability of an enterprise [2]. To find a suitable approach to upstream SCF, the physiology of a supply chain should be considered from three different perspectives: length, depth and time.

When trying to quantify and forecast upstream demand propagation, it is important to recognize the complexity of the supply chain and different factors that may influence or distort the projections. Lee et al. [3] defined Bullwhip Effect as the amplification of demand variance that takes place as the value proceeds through the chain nodes. Main reasons for this are operational inefficiencies and external factors that affect the deviation between expected and realized demand quantities.

It was noted by Chopra and Meindl [1] that one way to handle incomplete information, its distortion effect on demand projections and operational inefficiency of manufacturers, would be the development of collaborative concepts where information is shared between supply chain entities. The main concepts that were proposed are collaborative forecasting and replenishment (CFAR) systems where interchange of decision-support models and strategies to facilitate forecasting processes is suggested [4]. Other concepts that have emerged include Collaborative planning, forecasting and replenishment (CPFR), Vendor Management Inventory (VMI) and other information systems [5]. While unintuitive, it was shown that collaborative supply chain forecasting can yield negative dynamics in the performance, widening the Bullwhip Effect and burdening the procurement function [6], [7].

Since collaborative forecasting mechanisms prove to be ineffective both in terms of accuracy and incurred workload, it is becoming increasingly relevant to explore possibilities for the autonomous forecasting of demand. This research is based on anonymized historical purchasing data from several industry partners operating globally. In terms of the length of a supply chain, the dataset provides full visibility to the first-tier suppliers of different products, while lacking an extended view to adjacent nodes of the supply chain, which represents a typical setup for developing SCF process as a business application. The main objective of this research is to evaluate and compare the performance of different forecasting

This work was supported by Sievo Oy. The authors would also like to acknowledge the support of this publication by LUT research platform AMBI- Analytics-based management for business and manufacturing industry.

methods in the SCF domain and, if possible, to identify methods of choice for material demand forecasting.

II. METHODOLOGICAL BACKGROUND

Time series forecasting, i.e. prediction of future or missing entries in a series of numerical values indexed in time order [8], is a broad research domain which is historically relevant for multiple application areas, incl. natural sciences, industrial engineering, economy, business and many others. Time series forecasting can be divided into three main methodological types [9]. These are 1) Explanatory models where the dependent variable is represented as a function of external factors (regressors or independent variables) and a causal relationship is assumed (or at least the ability to compute the values of the regressand from the values of the regressors) for modelling by fitting the function to existing data 2) Autoregressive models where the forecast is generated based on historical values of time series without external variables 3) Mixed models, which contain explanatory and dynamic components, that include dynamic regressions, transfer function models, linear systems, vector alternatives of the above mentioned models, machine learning models etc. In this paper, autoregressive time series models are used due to a lack of numerical data points available in an independent enterprise SCF concerning additional explanatory variables. We selected three different model types to be fitted to the data and also considered a naïve benchmark model to be able to compare the performance of the selected models with a reference model. Given the type of the time series being forecasted, only models capable of reflecting seasonality in the time series were considered.

A. Naïve benchmark

The naïve forecasting method is the basic estimation technique in which time series value from the last period is taken as the forecast for the next one, without attempting to adjust it or establish causal factors. It is represented as

$$y_{t+1} = y_t \tag{1}$$

where y_t is time series in question and it is assumed that at time t we need to make a forecast of the value of the time series for times $t+m$, $m \in \mathbb{N}, m > 0$. In other words we assume that the historical values of the time series being forecasted are known including the current value of the series, but no information is available after time t. Predicting the last known value, that is $y_{t+m} = y_t$ for all $m > 0$, is one of the most commonly used benchmark methods due to its simplicity.

B. Holt-Winters exponential smoothing

The exponential smoothing models were proposed as forecast generators through weighted average of previous observations while weights decrease exponentially over time periods (more historical values influence the forecast less than more recent ones).

In Holt-Winters (HW) seasonal method [10]–[12] the time series are decomposed, and the series estimation formula is split into three equations: level, trend and seasonality. All of them consider different smoothing coefficients and comprise a system of simultaneous equations as follows:

$$\begin{cases} S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \\ b_t = \gamma(S_t - S_{t-1}) + (1-\gamma) b_{t-1} \\ I_t = \beta \frac{y_t}{S_t} + (1-\beta) I_{t-L} \\ y_{t+m} = (S_t + m b_t) I_{t-L+m} \end{cases} \tag{2}$$

where y_t is observation of the series, S_t is the smoothed observation, b_t is the trend factor, I_t is the seasonal index, y_{t+m} is the forecast at m periods ahead; α, β and γ are smoothing parameters that are estimated so as to minimize the fitting error. The baseline value for trend can be computed as

$$b_0 = \frac{1}{L}\left(\frac{y_{L+1} - y_1}{L} + \frac{y_{L+2} - y_2}{L} + \cdots + \frac{y_{L+L} - y_L}{L}\right) \tag{3}$$

where L is the length of the season, y_t are observation series, while the initial season factor is calculated as

$$I_0 = \frac{\sum_{p=t}^{N} \frac{y_{t+pL}}{A_p}}{N} \tag{4}$$

where t is the time period, N is the number of complete seasons we have the data for, y_t are observation series and $A_p = \frac{\sum_{i=1}^{L} y_i}{L}, p = 1,2,\ldots,N$.

C. Seasonal Autoregressive Moving Average

Autoregressive Moving Average (ARMA) model family consist of autoregressive (AR) and stochastic (MA) components [13]. Autoregressive components reflect the dynamic structure of the series describing its linear relation to order p while the moving average component is a linear combination of q lags of the error term. Alongside with exponential moving average models, they are commonly used in SCF for benchmarking purposes [2].

ARMA models are formulated as follows and they require the time series to be weakly stationary.

$$y_t = C + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{5}$$

where y_t is the estimated series, C is the constant term, φ_i is the coefficient for the autoregressive component of order i, θ_j is the coefficient for the moving average component of order j, and ε_t is the error term.

Seasonal autoregressive integrated moving average (SARIMA) model is an extension of the traditional integrated ARMA activating the pattern recognition potential through a set of new parameters: seasonal autoregressive component (P), seasonal integration (D) and seasonal moving average (Q). These parameters are combined (the order of seasonal integration being set to $D = 0$) in the following equation:

$$y_t = C + \sum_{i=1}^{p} \varphi_i y_{t-i} + \sum_{k=1}^{P} \gamma_k y_{t-kL} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{r=1}^{Q} \mu_r \varepsilon_{t-rL} \quad (6)$$

where, in addition to the terms from (5), we introduce γ_k and μ_r as seasonal parameters to be estimated with the length of seasonal period L.

D. Fuzzy time series model

Fuzzy time series (FTS) is a concept from the fuzzy data analysis domain, which is based on the fundamental concept of a fuzzy set, introduced by Zadeh [14]. A fuzzy set is a flexible way to model uncertainty through assigning a gradual membership value $\mu_A(x) \in [0,1], x \in U$ to a specified set A for every element x of a universe of discourse U, instead of quantifying phenomena with a single crisp value from the set $\{0,1\}$.

In 1993 Song and Chissom [15] introduced fuzzy time series $F(t)$ on the subset of real numbers $Y(t)$ (t = 0,1,2,…). A fuzzy time series $F(t)$ is a collection of fuzzy sets $A_t(t = 1,2,\dots)$ with membership functions $\mu_{A_t}(x)$ $(t = 1,2,\dots,n \in Y_t)$. The real time series can be transformed into their fuzzy representation with the appropriate membership function, universe of discourse, and assigning membership degree values for real numbers in question.

The fuzzy time series forecasting models rely on the notion of fuzzy logical relationships (FLR). If A_i and A_j denote the fuzzy sets that form part of fuzzy time series $F(t)$, the logical relationship can be expressed with notation $A_i \rightarrow A_j$ (FTS model of order 1) or $[A_i, A_k] \rightarrow A_j$ (high-order FTS model with 2 lags). In the examples above, A_i and $[A_i, A_k]$ are called left-hand side (LHS) of an FLR, while A_j is its right-hand side (RHS).

The FLRs observed from historical data can be organized into fuzzy logical relationship groups (FLRGs). They comprise the knowledge- or rule base that is further inferred to generate forecast for future or missing values.

A simple FTS model generates forecast based on the following algorithm; let $F(t) = A_i$, then

- if $A_i \rightarrow \emptyset$, that is if there is no rule in the FLRG with A_i as LHS, then $F(t+1) = A_i$ and the defuzzified forecast $Y(t+1)$ is the midpoint of A_i, if defuzzification is needed;
- if $A_i \rightarrow A_j \in FLRG$, then $F(t+1) = A_j$, $Y(t+1)$ being the midpoint of A_j;
- if $A_i \rightarrow A_{j_1}, A_{j_2}, \dots, A_{j_k} \in FLRG$, there is no single fuzzy representation of $F(t+1)$, there are more possible fuzzy-set outputs, and the defuzzified value, if needed, is derived directly as the arithmetic average of the midpoints of $A_{j_1}, A_{j_2}, \dots, A_{j_k}$.

Weighted FTS (WFTS) is a model type that handles the scenario of $A_i \rightarrow A_{j_1}, A_{j_2}, \dots, A_{j_k}$ in a different way. The defuzzification of the forecast is then calculated as

$$Y(t+1) = \sum_{j \in RHS} w_j c_j \quad (7)$$

with

$$w_j = \frac{\#A_j}{\#RHS} \; \forall A_j \in RHS \quad (8)$$

where $\#A_j$ is the number of occurrences of A_j in FLRs with the same LHS and $\#RHS$ is the total number of temporal patterns within that FLRG and c_j is j^{th} midpoint [16].

Probabilistic Weighted FTS (PWFTS) incorporate information about membership degrees of the LHSs of the FLRs. The knowledge base for PWFTS is given as

$$\begin{matrix} \pi_1 A_1 \rightarrow w_{11} A_1, \dots, w_{1i} A_i \\ \dots \\ \pi_i A_i \rightarrow w_{i1} A_1, \dots, w_{ii} A_i \end{matrix} \quad (9)$$

where each weight π_i is the normalized sum of all LHS values of membership functions where the LHS is fuzzy set A_i [17]. Thus, π_i can be interpreted as the empirical a priori probability of having A_i as an LHS. The weight w_{ij} is the normalized sum of all RHS memberships where LHS is A_i and RHS is A_j, which can be understood as a conditional probability $P(F(t+1) = A_j | F(t) = A_i)$.

The forecasting procedure in PWFTS starts with the computation of probability distribution

$$P\big(Y(t)\big|Y(t-1)\big) = \sum_{A_j \in \tilde{A}} \frac{P\big(Y(t)\big|A_j\big) * \sum_{i=1}^{k} P\big(Y(t+1)\big|A_i, A_j\big)}{\sum_{i=1}^{k} P\big(Y(t)\big|A_i\big)} = \sum_{A_j \in \tilde{A}} \frac{\pi_j \frac{\mu_{A_j}(Y(t))}{Z_{A_j}} * \sum_{i=1}^{k} w_{ij} \frac{\mu_{A_i}(Y(t+1))}{Z_{A_i}}}{\sum_{i=1}^{k} \pi_i \frac{\mu_{A_i}(Y(t))}{Z_{A_i}}} \quad (10)$$

where, in addition to previous notations, $\mu_A(Y)$ is degree of membership of continuous value Y to a fuzzy set A, Z_A is the total area under membership function of A, and $\tilde{A}$ is the set of all fuzzy sets considered on the given universe, for example $\tilde{A}$ can be a set of the fuzzy-set meanings of the linguistic values of a linguistic variable used to describe the values of the time series to be forecasted. The point forecast is then produced by

$$Y(t+1) = \sum_{A_j \in \tilde{A}} \frac{P\big(Y(t)\big|A_j\big) * E[A_j]}{\sum_{A_j \in \tilde{A}} P(Y(t)|A_j)} \quad (11)$$

with $E\big[A_j\big] = \sum_{i \in A_j^{RHS}} w_{ij} * mp_i$, mp denoting a midpoint of a fuzzy set A_j.

FTS represent a real alternative to the traditional econometrics methods since fuzzification of original time series makes the stationarity requirement redundant and reduces the allowed value domain to a finite number of fuzzy sets (or linguistic values the meanings of which are

represented as fuzzy sets) which works as an embedded normalization technique.

III. Data extraction and preprocessing

Extraction and preprocessing of the data included such subtasks as 1) selection of appropriate time period, 2) temporal aggregation, 3) scoping (reducing dimensionality – due to computational reasons) the list of time series included in the analysis and 4) handling outliers.

For the time period selection, three main criteria were considered: availability of data on codified direct purchases, potential to reveal annual seasonality and relevance for the business. Based on those criteria, the timeframe for the transactional dataset was set to January 2016 – November 2020, the latter being the most recent reported period in the source data. The time step was one calendar month. Cross-sectional aggregation was performed on a product-location level, which means that each time series represents monthly values of purchased quantities of a product by a given operating unit.

Pareto principle also known as the "80-20" rule was considered to narrow down the focus of the quantitative research. The dominating share of spend originated from a relatively low number of biggest purchase items, hence the scope of research could be limited to comply with limitations of available computing resource. Depending on the industry partner, from 1.34% to 3.70% of available time series were such that they added up to 90% of cumulative spend, and thus were included to the research scope. Further filtering of the data is described assuming that 100% of original time series represent the reduced number.

The underlying products of purchased quantity time series need to remain relevant to the business. We therefore only included the time series that contained non-zero values of quantity and spend in the last 12 months of the recorded period. Across the three partner datasets, 77.31-90.41% of time series fulfilled the requirements.

In order to ensure availability of sufficient training data, we removed the series where the period between the earliest and the most recent observation was under 3 years. In the study, 18.30-53.96% of the time series have enough observations. Combined with the previous filtering criterion, there is an overall acceptance rate of 17.99-50.99% of the initial number of time series for the subsequent analysis and experiments.

IV. Design of experiments

In this part, we describe the experiments conducted to evaluate the quality of the selected time series forecasting methods.

Performance measurement

For performance measure Root Mean Squared Error (RMSE) was selected. RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{pred_i}-y_i\right)^2}{n}} \tag{12}$$

where y_{pred_i} is the forecasted value of the series for time i,and y_i is the corresponding original value for all the investigated values of i.

Training and testing data

An appropriate representation of data is essential in a quantitative study. Time series data are commonly split over temporal indices to ensure original order.

First, we specify the forecast horizon i.e. the number of future observations that we want to generate as a model output. The business needs dictate that budgetary revisions are performed on a quarterly basis; thus the forecast horizon is specified to 3 months.

In order to avoid potential bias related to seasonality or coincidence in externalities, multiple testing windows are included in the analysis as per availability and volume of original data. An expanded rolling window approach is adopted, meaning a gradual increase in the number of observations in the training dataset, shifting the index of the testing period in a way that provides additional dimension to the analysis of results by revealing the sensitivity of algorithms to the amount of training data. The resulting cuts of the original time series range between 36 and 58 observations in length; we characterize the amount of training data as "low" if it represents a time period of less than 4 full years, and "high" otherwise.

All things considered, the experiment for each series is carried out in the following steps:

1. Identify the first and last period with non-zero normalized quantity values and remove leading and lagging null observations;

TABLE I.
COMPARISON OF PERFORMANCE OF DIFFERENT METHODS

	HW RMSE		**SARIMA RMSE**		**FTS RMSE**		**Naive RMSE**	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Company A	0.163	0.075	0.178	0.076	0.121	0.055	0.199	0.101
Company B	0.165	0.084	0.179	0.087	0.125	0.069	0.219	0.123
Company C	0.210	0.107	0.226	0.109	0.156	0.088	0.295	0.153

2. Split the resulting series into $n_{windows}$ expanding windows, starting with the first 33+3=36 months of data (33 observations for training and 3 for testing purposes) and incrementing the index of last observation included in the sample by $[(i_{max} - 36)/n_{windows}]$ where i_{max} is the largest integer index of the series (starting with 1, equal to number of observations) and $n_{windows}$ is the target number of windows per series;
3. Run all configurations of each model family (Exponential Smoothing, SARIMA or FTS) on each of the windows and store the results in such a format that it would include full information regarding the tested series, values of hyperparameters and observed RMSE. Apply the naïve forecast for benchmarking purposes.

V. Results

In Table I, the methods are evaluated based on average RMSE error term and its standard deviation across time series. There is a visible improvement in prediction accuracy of all three methods compared to the naïve solution.

The improvement is further validated with a visualization of RMSE in form of histograms (Figure 1), one per each dataset.

If we compare individual performance of the models on time series level, we see that FTS would be the best choice in 7947 cases representing almost 60% of the total count, followed by HW and Naïve benchmark with 2642 and 1624 respectively while SARIMA would only be optimal in 1099 cases. Looking at detailed specifications of the respective models, it is notable that in majority of cases SARIMA becomes a simple arithmetic average, serving as additional benchmark solution.

FTS shows a higher and more stable prediction accuracy which may be explained by its ability to handle the intermittency by fuzzification of original series whereas other methods operate on a continuous scale. Zero values alternating with non-zero ones are translated into a discrete number of fuzzy sets, which reduces the noise in identifying sequential patterns.

VI. Conclusion

In scope of this research, we have tested such time series forecasting methods as Holt-Winters exponential smoothing, SARIMA and Fuzzy Time Series forecasting models, on three independent datasets containing historical direct material purchasing data of industry partners. The results reveal that using the Fuzzy Time Series approach, there is a potential to reveal hidden intrinsic and seasonal patterns and achieve a substantial improvement in accuracy compared to simple statistical forecast.

Fuzzy Time Series models showed the best performance in all datasets because of their ability to reduce the noise caused by intermittency of the original series. Holt-Winters is a viable alternative showing stable improvement to the error

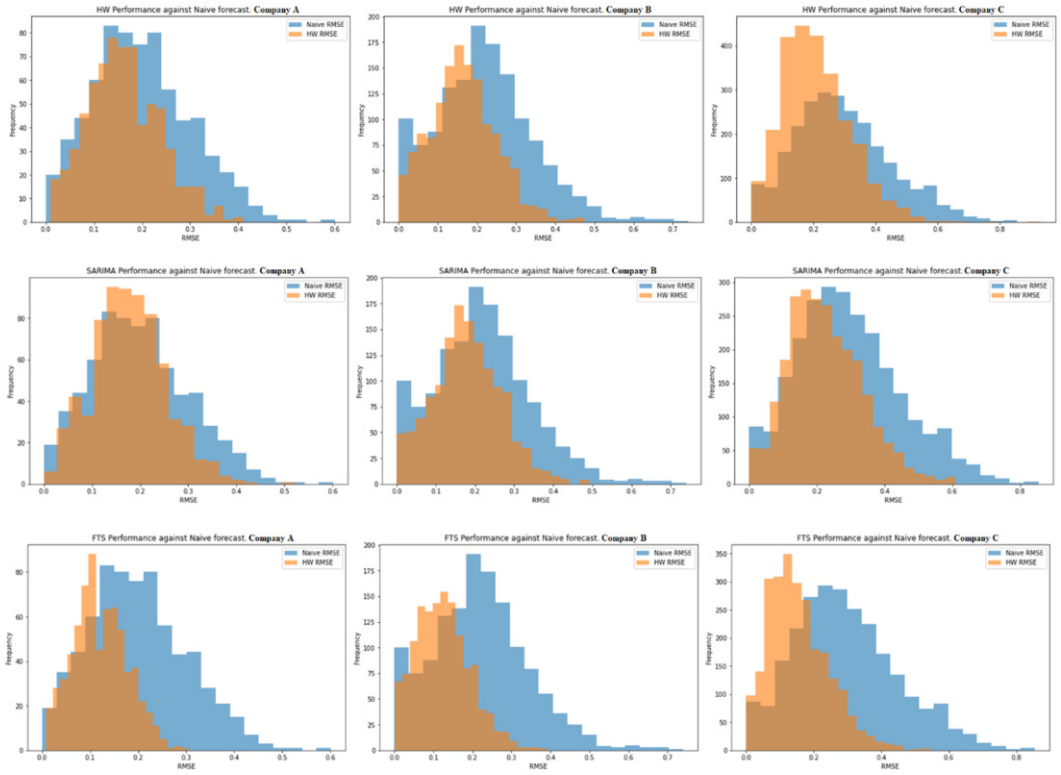


Figure 1. RMSE histogram of HW (row 1), SARIMA (row 2) and FTS (row 3) against benchmark

metrics whereas SARIMA is not recommended in this case due to insufficient amount of training data and its intermittent nature which results in a notable underperformance. Fuzzy Time Series and Holt-Winters can be used to generate automatic forecasts of direct material purchases when the amount of historical data is sufficient.

Acknowledgment

The paper represents a processed summary of the research performed in scope of a Master thesis [18].

References

[1] G Chopra, S., Meindl, P. Supply chain management: strategy, planning and operation.5th edition. USA New Jersey: Pearson, 2012
[2] Ganzha, M., Maciaszek, L., Paprzycki, M. & Ślęzak, D. Impact of time series clustering on fuel sales prediction results. Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems, ACSIS, Vol. 26, pages 13–21. 2021
[3] Lee, H. L., Padmanabhan, V., & Whang, S. Information distortion in a supply chain: the bullwhip effect. Management Science, 43, 546–558, 1997
[4] Raghunathan, Srinivasan. Interorganizational Collaborative Forecasting and Replenishment Systems and Supply Chain Implications. Decision Sciences. 30. 1053 - 1071, 2007
[5] Synthetos, A. A., Kholidasari, I., & Naim, M. The effects of integrating management judgement into OUT levels: in or out of context? European Journal of Operational Research, 2015
[6] Thonemann, U.W. Improving supply-chain performance by sharing advance demand information. European Journal of Operational Research 142-1, 81–107, 2002
[7] Heikkilä, J., From supply to demand chain management: Efficiency and customer satisfaction. Journal of Operations Management 20 (6), 747–767, 2002
[8] Box, G. & Jenkins, G., Time Series Analysis: forecasting and control, Oakland, California: Holden-Day, 1976
[9] Hyndman, R.J. & Athanasopoulos, G. Forecasting: principles and practice, OTexts: Melbourne, Australia, 2nd Edition, 2018
[10] Brown, R. G. Statistical forecasting for inventory control. McGraw/Hill, 1959
[11] Holt, C. E. Forecasting seasonals and trends by exponentially weighted averages (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA, 1957
[12] Winters, P. R. Forecasting sales by exponentially weighted moving averages. Management Science, 6(3), 324–342, 1960
[13] Mills, T. Time Series Techniques for Economists. Cambridge University Press, 1990
[14] Zadeh, L. Fuzzy sets. Information and Control, 8, 338–353, 1965
[15] Song, O., Chissom, B. Fuzzy time series and its model. Fuzzy Sets and Systems. 54. 269–277. 1993
[16] Ortiz-Arroyo, D., Poulsen, J. R. A Weighted Fuzzy Time Series Forecasting Model. Indian Journal of Science and Technology, 11(27), 1-11. 2018
[17] Silva, P. Scalable Models for Probabilistic Forecasting with Fuzzy Time Series, Thesis for: Ph.D. 2019
[18] Zakrytnoy, S. Comparative study of classic and fuzzy time series models for direct materials demand forecasting. Thesis for: MSc. 2021

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 7–17
DOI: 10.15439/2021B7
ISSN 2300-5963 ACSIS, Vol. 29

The voter's guide to the galaxy—a multiple-criteria fuzzy decision-support tool for voters and a fresh take on election survey methods

Sofia Maria Panzeri
Universitá degli Studi di Milano,
Via Festa del Perdono 7, 20122
Milano, Italy, and
Palacky University Olomouc, Faculty of Arts, Department of Economic and Managerial studies, Email:
sofiamaria.panzeri@studenti.unimi.it

Jana Stoklasová
School of Business and Management, LUT University,
Yliopistonkatu 34,
53851 Lappeenranta, Finland
Email: jana.stoklasova@lut.fi

Jan Stoklasa
School of Business and Management, LUT University,
Yliopistonkatu 34,
53851 Lappeenranta, Finland, and
Palacky University Olomouc,
Faculty of Arts, Department of Economic and Managerial Studies
Email: jan.stoklasa@lut.fi
Email: jan.stoklasa@upol.cz

***Abstract*—This paper suggests a multiple-criteria decision-support tool for voters, that compares the attitudes of the voters with the declared attitudes of the political parties in several sets of relevant issues. The model intends to identify parties that seem to provide the best fit with the voter attitude-wise. The data input methodology uses discrete 5-point Likert-type scales. We investigate the effect of the inclusion of weights of different sets of issues, of the numerical anchors of the values of the Likert-type scales and also of the potential presence of extremity/leniency effect on the suggestion of the "most compatible" political party suggestion. We also propose a simple fuzzy-rule based evaluation tool to identify serious incompatibilities or desired compatibilities in the attitudes of the voter and the party to the relevant issues. This tool introduces (un)acceptability thresholds for the differences in attitudes between the parties and the respondents and provides lists of parties to vote for or to avoid voting for accompanied by the strengths of these suggestions. The tool is shown to have several desirable features including lower sensitivity to small differences in the attitudes, respondents' ability to express their preferences and also preventing the compensation of unacceptable differences in some categories of important issues by high compatibility in the other categories.**

I. Introduction

THE issue of elections is a topical one since the very beginning of democracy. It is becoming even more complex nowadays with increased (cyber) security concerns [1]. Choosing one's representatives or at least the political party that reflects one's values well enough is, however, a difficult problem to face. The choice of the most appropriate representing party would be a difficult one even if the voters had full information concerning the program, values, intentions, and goals of the parties/individuals to be chosen. In many cases, however, the assumption of full information is unachievable. In these cases, one might decide based on a sample of key issues and the similarity/difference of his/her attitudes towards these and the attitudes of the political parties. Ballot and voter decision-support systems are being discussed and proposed to help voters get oriented in the vast amount of information available and to facilitate information management [2]. However, these are very information extensive and require a sufficient knowledge of their users and advanced knowledge of human-computer interaction. As such their introduction in practice might be difficult. Still a simpler and less advanced voter decision-support can be beneficial and constitute a step towards the desired integrated voter-support information systems.

This is exactly the point of departure of this paper. We assume that a voter intends to choose as rationally as possible. Rational choice in this case is operationalized as the act of choosing the party that expresses its opinions or attitudes to the key issues (or sets/groupings thereof) in the way most similar to one's own attitudes. Our question is how to decide what is "most similar" in this context – particularly in such a way that would be applicable for political and social sciences research and also for election surveys as well as for actual voter decision support. This means that we will be relying on the information available in the program statements of the parties and on expert assessment thereof, when needed. We will also be using simple tools for data input, namely Liker (type) scales [3].

In line with the finding of Rogowski [4] we assume that voters tend to vote for those parties that have similar (general) ideological orientations. This means that we can afford to focus on several key issues that overall capture the attitude (or ideology) of the party and of the voter to be supported by the proposed system. Several tools for the assessment of agreement of one's attitudes with those of others that deal with the closeness of the attitudes (including their uncertainty) in the semantic space are already available [5,6]. Even though various types of consensus have recently been proposed for these methods [7] and the attitudes can be represented with the corresponding uncertainty stemming from the data input method as well as from the nature of the decision-makers and the concepts being assessed, these

This work was supported by LUT research platform AMBI- Analytics-based management for business and manufacturing industry.

methods require a more complex data input method than would be desirable in the context of voter decision-support with multiple key areas being considered.

In this paper we therefore suggest a method that is simpler in terms of data input, but still allows for the assessment of compatibility of one's own values with those declared by the parties in line with [4] and provides valid decision-support. Obviously, when the goal of simplicity of obtaining input data is set, there are drawbacks to be expected in the process of the analysis of the data. In this case we will discuss the effect of the calibration of the numerical values of discrete Likert scales with linguistic labels [8,9] and also the possibility of getting more insights or more real-life representation of the preferences, attitudes or values using the tools of fuzzy set theory [10]. We are well aware that some election surveys and popular voter "calculators" providing fast and popular "compatibility" suggestions to voters use a similar approach, but these are frequently using just a binary scale and do not offer any customizability. Our approach strives to allow for the reflection of different strengths of support/opposition concerning a specific issue and thus on different magnitudes of differences in the attitudes to the selected crucial issues.

II. Problem definition and context

In this paper we assume the perspective of 5 young potential Italian voters represented by actual respondents, and we set the goal of identifying the most "fitting" party to vote for based on the compatibility (difference) of the attitudes of the respondents and declared attitudes of the parties.

The set of parties consists of eight Italian political parties:

- Movimento 5 Stelle ("anti-establishment")
- Lega (right wing)
- Partito Democratico (centre-left)
- Forza Italia (centre-right)
- Fratelli d'Italia (far right wing)
- Italia Viva (centre-left)
- Liberi e Uguali (centre-left)
- +Europa e Azione (centre-left)

The above listed parties serve as real-life examples of parties, are selected so that they represent different declared attitudes to the selected crucial issues and at the same time allow the assessment of the reasonability of the provided voter decision-support and its sensitivity to the calibration of the used scales. We are not assuming a position of support/opposition with respect to any of these parties. The summary labels of the parties provided in the brackets are intended as "guides to the understanding of the overall philosophy/ideology of the party", they have been assigned by the authors of the paper and might constitute a large simplification of the actual goals and attitudes of the party. Nevertheless, we think that since this represents an example setting for the proposed method, the labels can provide the reader a better ability to assess the results of the decision support suggested in this paper.

A. The important issues used to assess the compatibility between the respondents' attitudes and the attitudes of the political parties

The crucial issues to be considered were compiled by the authors and in partial cooperation with the respondents with the aim to cover the most important areas as considered by the respondents. This is well in line with the idea of the use of the proposed framework as a voter decision-support tool. On the other hand, if an overall "attitude compatibility study" were to be conducted, then the list of the important issues can be compiled by the researcher in accordance with the needs and goals of the study. The crucial issues are grouped into 6 main categories. This allows for a detailed issue-by-issue attitude-compatibility analysis but also for a more complex (potentially repeated) assessment of the attitudes towards the overall issue categories. The considered issues and their categories are the following:

C1. SOCIAL ISSUES

C1,1 Are you for or against ABORTION?
C1,2 Are you for or against EUTHANASIA?
C1,3 Are you for or against the DEATH PENALTY?
C1,4 Are you for or against LGBTQIA+ ADOPTION RIGHTS?
C1,5 Are you for or against SAME SEX MARRIAGE?

C2. FOREIGN POLICY ISSUES

C2,1 Are you for or against ITALY'S WITHDRAW FROM THE EU?
C2,2 Are you for or against the GONVERNMENT INFLUENCING FOREIGN ELECTIONS?
C2,3 Are you for or against the UNITED STATES of EUROPE?
C2,4 Are you for or against an INCREASE in MANDATORY MILITARY SPENDING?
C2,5 Are you for or against the creation of an EU ARMY?

C3. IMMIGRATION ISSUES

C3,1 Are you for or against a TEMPORARY IMMIGRATION BAN?
C3,2 Are you for or against DEPORTING CRIMINAL IMMIGRANTS? (violent crimes)
C3,3 Are you for or against BANNING MUSLIMS IMMIGRANTS FROM ENTERING THE COUNTRY?
C3,4 Are you for or against an EU IMPOSED QUOTA OF MIGRANTS PER COUNTRY?
C3,5 Are you for or against IMMIGRANTS taking a CITIZENSHIP TEST?

C4. HEALTHCARE ISSUES

C4,1 Are you for or against the ISSUE of VACCINE PASSPORTS?
C4,2 Are you for or against an INCREASE in FUNDING for MENTAL HEALTH?
C4,3 Are you for or against the PRIVATIZATION of HOSPITALS?

TABLE I.

ASSESSMENT OF THE ATTITUDES OF THE PARTIES TO THE J-TH ISSUE IN THE I-TH CATEGORY (REPRESENTED BY THE VALUE IN THE I-TH ROW AND J-TH COLUMN IN EACH RESPECTIVE MATRIX) UNDER DIFFERENT SETUPS OF THE LIKERT SCALE. SETUPS I AND II ARE STANDARD 5-POINT EQUIDISTANT LIKERT SCALE CODINGS, SETUP III IS 5-POINT NON-EQUIDISTANT BUT SYMMETRICAL LIKERT SCALE CODING AND SETUP IV IS A 3-POINT SCALE ANALOGY TO STOKLASA ET. AL. [9]. EXAMPLE OF A RESULT OF EXPERT ASSESSMENT.

SETUP I and II

MOVIMENTO 5 STELLE

2	1	5	2	2
4	5	5	4	4
4	1	5	1	1
2	2	5	2	1
1	4	3	5	1
5	5	2	3	1

LEGA

2	3	1	5	4
4	5	5	1	4
1	1	1	1	1
2	2	1	5	5
1	4	1	1	4
2	5	5	4	1

PARTITO DEMOCRATICO

1	1	5	1	1
5	5	1	5	1
5	4	5	1	5
1	1	5	3	1
1	2	4	5	1
5	4	1	3	1

FORZA ITALIA

2	5	4	5	3
4	5	1	2	1
3	1	4	1	1
1	2	1	5	3
1	5	1	5	4
1	5	3	5	4

FRATELLI D'ITALIA

5	5	3	5	5
4	5	5	2	5
1	1	1	1	1
4	2	5	5	5
1	5	2	1	3
5	5	5	5	1

ITALIA VIVA

1	2	5	1	2
5	5	1	3	2
3	3	4	4	2
2	1	4	3	2
1	2	2	4	4
4	2	3	4	4

LIBERI E UGUALI

1	1	5	1	1
5	5	1	4	1
5	5	5	1	5
1	1	5	1	1
1	1	5	5	1
5	5	1	2	1

+EUROPA and AZIONE

1	1	5	1	1
5	5	1	5	1
5	5	5	1	5
1	1	5	1	1
1	3	4	5	1
5	4	1	2	1

SETUP III

MOVIMENTO 5 STELLE

2.5	1	5	2.5	2.5
3.5	5	5	3.5	3.5
3.5	1	5	1	1
2.5	2.5	5	2.5	1
1	3.5	3	5	1
5	5	2.5	3	1

LEGA

2.5	3	1	5	3.5
3.5	5	5	1	3.5
1	1	1	1	1
2.5	2.5	1	5	5
1	3.5	1	1	3.5
2.5	5	5	3.5	1

PARTITO DEMOCRATICO

1	1	5	1	1
5	5	1	5	1
5	3.5	5	1	5
1	1	5	3	1
1	2.5	3.5	5	1
5	3.5	1	3	1

FORZA ITALIA

2.5	5	3.5	5	3
3.5	5	1	2.5	1
3	1	3.5	1	1
1	2.5	1	5	3
1	5	1	5	3.5
1	5	3	5	3.5

FRATELLI D'ITALIA

5	5	3	5	5
3.5	5	5	2.5	5
1	1	1	1	1
3.5	2.5	5	5	5
1	5	2.5	1	3
5	5	5	5	1

ITALIA VIVA

1	2.5	5	1	2.5
5	5	1	3	2.5
3	3	3.5	3.5	2.5
2.5	1	3.5	3	2.5
1	2.5	2.5	3.5	3.5
3.5	2.5	3	3.5	3.5

LIBERI E UGUALI

1	1	5	1	1
5	5	1	3.5	1
5	5	5	1	5
1	1	5	1	1
1	1	5	5	1
5	5	1	2.5	1

+EUROPA and AZIONE

1	1	5	1	1
5	5	1	5	1
5	5	5	1	5
1	1	5	1	1
1	3	3.5	5	1
5	3.5	1	2.5	1

SETUP IV

MOVIMENTO 5 STELLE

1	1	5	1	1
5	5	5	5	5
5	1	5	1	1
1	1	5	1	1
1	5	3	5	1
5	5	1	3	1

LEGA

1	3	1	5	5
5	5	5	1	5
1	1	1	1	1
1	1	1	5	5
1	5	1	1	5
1	5	5	5	1

PARTITO DEMOCRATICO

1	1	5	1	1
5	5	1	5	1
5	5	5	1	5
1	1	5	3	1
1	1	5	5	1
5	5	1	3	1

FORZA ITALIA

1	5	5	5	3
5	5	1	1	1
3	1	5	1	1
1	1	1	5	3
1	5	1	5	5
1	5	3	5	5

FRATELLI D'ITALIA

5	5	3	5	5
5	5	5	1	5
1	1	1	1	1
5	1	5	5	5
1	5	1	1	3
5	5	5	5	1

ITALIA VIVA

1	1	5	1	1
5	5	1	3	1
3	3	5	5	1
1	1	5	3	1
1	1	1	5	5
5	1	3	5	5

LIBERI E UGUALI

1	1	5	1	1
5	5	1	5	1
5	5	5	1	5
1	1	5	1	1
1	1	5	5	1
5	5	1	1	1

+EUROPA and AZIONE

1	1	5	1	1
5	5	1	5	1
5	5	5	1	5
1	1	5	1	1
1	3	5	5	1
5	5	1	1	1

C4,4 Are you for or against the institution of SAFE HAVENS?
C4,5 Are you for or against the LEGALIZATION of MARIJUANA?

C5. ECONOMIC ISSUES

C5,1 Are you for or against the SAME SALARY for MEN and WOMEN for the SAME JOB?
C5,2 Are you for or against RAISING TAXES on the RICH?
C5,3 Are you for or against CUTS to PUBLIC SPENDING in order to REDUCE NATIONAL DEBT?
C5,4 Are you for or against an INCREASE on TARIFFS on PRODUCTS IMPORTED into the country?
C5,5 Are you for or against FEWER RESTRICTIONS on CURRENT WELFARE BENEFITS?

C6. CRIMINAL ISSUES

C6,1 Are you for or against the PRIVATIZATION of PRISONS?
C6,2 Are you for or against the RELEASE from JAIL of NON-VIOLENT PRISONERS? (to reduce overcrowding)
C6,3 Are you for or against CONVICTED CRIMINALS having the RIGHT TO VOTE?
C6,4 Are you for or against the DEFUNDING of the POLICE?
C6,5 Are you for or against passing laws which PROTECT WHISTLEBLOWERS?

In order to allow for some expression of the strength of support or opposition of a specific issue, Likert scales are used to obtain the assessment of the attitudes of the parties and also of the individual respondents (i.e. potential voters). For the purpose of this paper we first adopt a 5-point Likert scale with linguistic values “strongly for”, “slightly for”, “neutral”, “slightly against” and “strongly against”.

B. Different configurations of the Likert scales used in the decision support tool

To be able to perform calculations, we need to assign numerical values to the linguistic values of the scales. This step potentially introduces several methodological issues (see e.g. [9-13] for a more detailed discussion of some of them). In this research we are focusing on the reasonability of performing calculations with the numerical values of the scales [11] that is connected with the (non)equidistance of the used numerical meanings of the linguistic values of the scale [9] and the differences in the perception of the relative distances between the linguistic values as compared to the perceived distances of their numerical meanings. We also reflect the potential ambiguity of the linguistic terms and their different interpretation by different individuals that might result in a different numerical value being the appropriate meaning of the linguistic term for different individuals [10]. Last but not least we consider the effect of leniency/central tendency [12,13] and apply an analogy to the 3-bin histogram based solution proposed in [9]. For this reason, we propose the use of the following configurations of the numerical meanings of the linguistic values of the Likert scales:

- Setups I and II assign integer values to the linguistic values. In other words, “strongly for”, “slightly for”, “neutral”, “slightly against” and “strongly against” are represented by 1, 2, 3, 4 and 5 respectively. This setup uses the usual approach to Likert scales and considers the linguistic values equidistant meaning-wise. As the equidistance of the perceived meaning of the linguistic terms cannot be guaranteed, this assignment of numerical meanings to the linguistic values is questioned by some authors as the source of serious limitations for the subsequent reliable processing of these values. Nevertheless, this configuration is being used frequently in practice and as such it provides a good benchmark for the other proposed setups. Setups I and II use the same Likert scale configuration, Setup I considers all the categories of issues equally important, whereas Setup II assigns different weights to different categories.
- Setup III assumes a “calibration” of the numerical meanings of the linguistic terms of the Likers scale has been performed and as a result of it the values “slightly for” and “slightly against” are semantically closer to the “neutral” term than to their respective “strongly for” and “strongly against” counterparts. The numerical values used as meanings of the linguistic terms are 1, 2.5, 3, 3.5 and 5 for “strongly for”, “slightly for”, “neutral”, “slightly against” and “strongly against” respectively. The use of this setup does not assume that the calibration of the values proposed here is valid universally. It is meant to show what could be the results of appropriate calibration of the meanings of the linguistic values (as stressed in [14,15] for example) compared to the use of the standard use of Likert scales represented by Setup I. Note that the proposed calibration at least preserves the symmetry of the scale with respect to the mean value, which is required by Likert [3].
- Setup IV offers a possible solution to the presence of central tendency or leniency bias, that is to the tendency of some people to avoid extreme values of the scales or to prefer using these values respectively. In essence it can be argued that “strongly for” can be representing the same strength of support for one respondent as “slightly for” for another respondent. If this is the case, then assigning different numerical meanings (or treating these answers as different, even though they might represent an identical strength of support) can be incorrect. Stoklasa *et al.* suggest in [9] the use of aggregated +/0/- classes. This means that all positive answers are grouped in one class (denoted +), all negative answers in another one (denoted -) and those that can be considered neutral in a third one (denoted 0). This can be achieved by representing “strongly for”, “slightly for”, “neutral”, “slightly

against" and "strongly against" by 1, 1, 3, 5 and 5 values respectively.

C. Expert assessment of the political parties' attitudes towards the important issues

Table I summarizes the assessment of the attitudes of the selected eight political parties to the chosen thirty important issues by an expert evaluator. The first column of matrices represents the numerical values corresponding with the standard configuration of the 5-point Likert scale (Setup I and II), the middle column of matrices represents the same linguistic assessments but transformed into numerical ones using a recalibrated scale (Setup III), the right column represents the same linguistic assessments as the previous two columns of matrices, just coded using the +/0/- configuration (Setup IV). In each matrix the element in the position (i, j) represents the evaluation of the issue $C_{i,j}$, that is the evaluation of the *j*-th issue in the *i*-th category.

The presented values are an example of a possible assessment by an expert and do not need to fully reflect the actual attitudes declared by the parties. In a real-life setting the task of obtaining the assessments depicted in Table I could be performed by a group of domain experts. We will, however, consider them as representative of the parties' program declarations for the purpose of the calculations.

D. Attitudes of the respondents

The attitudes of the five respondents towards the thirty important issues were obtained using the 5-point Likert scales too and coded in accordance with the three above discussed Likert scale codings. Table II summarizes the results of this process. The respondents (we will call them Melania, Anna, Marco, Carlo and Sara) come from similar cultural and social backgrounds and for this reason the results of decision support for them can be expected to be similar to one another. Based on the expert assessment by the authors, we can conclude that most of the respondents are "close" to center-left parties and inclined to vote for parties closer to the left side of the spectrum.

TABLE II.

ASSESSMENT OF THE ATTITUDES OF THE RESPONDENTS TO THE J-TH ISSUE IN THE I-TH CATEGORY (REPRESENTED BY THE VALUE IN THE I-TH ROW AND J-TH COLUMN IN EACH RESPECTIVE MATRIX) UNDER DIFFERENT SETUPS OF THE LIKERT SCALE. SETUPS I AND II ARE STANDARD 5-POINT EQUIDISTANT LIKERT SCALE CODINGS, SETUP III IS 5-POINT NON-EQUIDISTANT BUT SYMMETRICAL LIKERT SCALE CODING AND SETUP IV IS A 3-POINT SCALE ANALOGY TO STOKLASA ET. AL. [9].

SETUP I and II

Melania

1	1	5	1	1
5	5	2	3	2
5	4	5	4	2
1	1	5	1	1
1	4	2	4	3
5	4	1	4	1

Anna

2	1	2	4	2
4	4	2	4	4
2	2	5	2	2
2	2	4	2	1
1	2	2	2	4
5	2	2	4	2

Marco

4	4	5	4	2
5	5	1	3	1
3	1	3	1	2
2	2	4	4	1
1	2	1	4	2
5	4	1	4	4

Carlo

2	1	5	2	1
5	5	5	3	3
5	5	5	3	2
3	1	1	1	1
1	3	2	3	1
5	1	1	5	3

Sara

1	1	4	1	1
5	5	2	4	2
5	1	5	1	1
1	1	4	2	1
1	1	3	4	3
3	4	2	5	1

SETUP III

Melania

1	1	5	1	1
5	5	2.5	3	2.5
5	3.5	5	3.5	2.5
1	1	5	1	1
1	3.5	2.5	3.5	3
5	3.5	1	3.5	1

Anna

2.5	1	2.5	3.5	2.5
3.5	3.5	2.5	3.5	3.5
2.5	2.5	5	2.5	2.5
2.5	2.5	3.5	2.5	1
1	2.5	2.5	2.5	3.5
5	2.5	2.5	3.5	2.5

Marco

3.5	3.5	5	3.5	2.5
5	5	1	3	1
3	1	3	1	2.5
2.5	2.5	3.5	3.5	1
1	2.5	1	3.5	2.5
5	3.5	1	3.5	3.5

Carlo

2.5	1	5	2.5	1
5	5	5	3	3
5	5	5	3	2.5
3	1	1	1	1
1	3	2.5	3	1
5	1	1	5	3

Sara

1	1	3.5	1	1
5	5	2.5	3.5	2.5
5	1	5	1	1
1	1	3.5	2.5	1
1	1	3	3.5	3
3	3.5	2.5	5	1

SETUP IV

Melania

1	1	5	1	1
5	5	1	3	1
5	5	5	5	1
1	1	5	1	1
1	5	1	5	3
5	5	1	5	1

Anna

1	1	1	5	1
5	5	1	5	5
1	1	5	1	1
1	1	5	1	1
1	1	1	1	5
5	1	1	5	1

Marco

5	5	5	5	1
5	5	1	3	1
3	1	3	1	1
1	1	5	5	1
1	1	1	5	1
5	5	1	5	5

Carlo

1	1	5	1	1
5	5	5	3	3
5	5	5	3	1
3	1	1	1	1
1	3	1	3	1
5	1	1	5	3

Sara

1	1	5	1	1
5	5	1	5	1
5	1	5	1	1
1	1	5	1	1
1	1	3	5	3
3	5	1	5	1

III. Different Setups of the Decision Support Tool and the Suggestions Generated by Them

Before we introduce the different setups of the proposed decision support tool to be compared, we first need to introduce the in-model notation. Let us consider a set of eight political parties $P = \{P_1, \dots, P_8\}$ and a set of five respondents $R = \{R_1, \dots, R_5\}$. We also consider a set of six categories of important issues $C = \{C_1, \dots, C_6\}$ each of which can be subdivided into five important issues relevant for the specific category, in other words $C_i = \{C_{i,1}, \dots, C_{i,5}\}$ for all $i = 1, \dots, 6$. The assessment of the attitude of each party and respondent to each of the important issues is captured by the matrix $A^r = \{a^r_{i,j}\}$, where $i = 1, \dots, 6, j = 1, \dots, 5,$ and $r \in P$ for the parties or $r \in R$ for the respondents. The categories $C_1, \dots, C_6$ can be assigned normalized weights $w_1, \dots, w_6$ such that $w_i \geq 0$ for all $i = 1, \dots, 6$ and $\sum_{i=1}^{6} w_i = 1$. It is even possible to assign respondent-specific weighting vectors. The vector of the possible numerical values representing the linguistic values of the 5-point Likert scale $L =$ ("strongly for", "slightly for", "neutral", "slightly against" and "strongly against") is denoted as $N = (n_1, \dots, n_5)$. Therefore $a^r_{i,j} \in N$ for any i, j, and r.

A. Setup I

In this setup the standard coding of the Likert-scale linguistic values by subsequent integer (equidistant) values is used. This means that $N^I = (1,2,3,4,5)$. We also assume that all the categories $C_1, \dots, C_6$ are considered equally important. We therefore do not need to define category weights for this purpose. As all the categories contain the same amount of important issues, we can define the difference between the attitudes expressed by a respondent $u \in R$ and attitudes of a party $v \in P$ simply as

$$d^I(A^u, A^v) = \sum_{i=1}^{6} \sum_{j=1}^{5} \left| a^u_{i,j} - a^v_{i,j} \right|, \qquad (1)$$

where $a^u_{i,j}, a^v_{i,j} \in \{1,2,3,4,5\}$. This way the difference in attitudes is defined as the sum of differences in the numerical meanings of the linguistic terms of the Likert scales used to assess the thirty important issues. Given the fact that all the assessments use the same 5-point scale, the maximum possible difference can be defined as $d^I_{\max} = 30(\max(N^I) - \min(N^I)) = 30(5 - 1) = 120$. As such an absolute-type measure of compatibility of the attitudes of respondent $u \in R$ with the attitudes expressed by a party $v \in P$ can be defined as

$$c^I(A^u, A^v) = 1 - \frac{d^I(A^u, A^v)}{d^I_{\max}}. \qquad (2)$$

Clearly $c^I(A^u, A^v) \in [0,1]$ for any $u \in R$ and $v \in P$. $c^I(A^u, A^v) = 1$ then means absolute (100%) compatibility, or in other words, zero difference in attitudes expressed by the respondent and the party. The suggested party to vote for should then be such that maximizes the compatibility value for the given respondent. Under this approach we get the compatibilities summarized in Table III.

TABLE III.
SUMMARY OF THE COMPATIBILITY OF THE PREFERENCES EXPRESSED BY THE FIVE RESPONDENTS TO THE PREFERENCES EXPRESSED BY THE POLITICAL PARTIES UNDER SETUP I. FOR EACH RESPONDENT THE HIGHEST COMPATIBILITIES ARE DENOTED IN GREEN, THE LOWEST COMPATIBILITIES IN RED.

Setup I	Melania	Anna	Marco	Carlo	Sara
MOVIMENTO 5 STELLE	77%	76%	71%	73%	79%
LEGA	47%	62%	57%	48%	52%
PARTITO DEMOCRATICO	83%	66%	71%	70%	81%
FORZA ITALIA	57%	61%	74%	53%	64%
FRATELLI D'ITALIA	47%	58%	60%	46%	50%
ITALIA VIVA	81%	77%	78%	72%	77%
LIBERI E UGUALI	82%	62%	66%	70%	79%
+EUROPA and AZIONE	84%	63%	67%	72%	78%

B. Setup II

This setup uses again the standard coding of the Likert-scale linguistic values, that is $N^{II} = (1,2,3,4,5)$. We however allow the respondents to reflect the perceived importances of the categories of issues in the form of respondent-specific weights. This allows for the customizability of the decision support by reflecting the relative importance of each category of issues.

We still assume that the issues within one category represent "repeated measurements" of the attitude towards the overall category and as such are considered equally important within a single category. This assumption can also be relaxed, but it would require us to obtain 30 weights from each respondent, which is not feasible in reality. We also assume that the normalized weights of categories can be specified reliably by all the respondents in the form $w^u = (w^u_1, w^u_2, \dots, w^u_6)$ for all $u \in R$. More specifically we have:

- $w^{Melania} = (0.33, 0.27, 0.20, 0.07, 0.07, 0.07)$,
- $w^{Anna} = (0.36, 0.21, 0.14, 0.14, 0.07, 0.07)$,
- $w^{Marco} = (0.25, 0.25, 0.13, 0.13, 0.13, 0.13)$,
- $w^{Carlo} = (0.37, 0.19, 0.19, 0.15, 0.07, 0.04)$,
- $w^{Sara} = (0.26, 0.21, 0.21, 0.16, 0.11, 0.05)$.

The (weighted) difference of the attitudes of any respondent $u \in R$ from the attitudes expressed by the party $v \in P$ can be calculated as

$$d^{II}(A^u, A^v) = \sum_{i=1}^{6} \left(w^u_i \cdot \sum_{j=1}^{5} \left| a^u_{i,j} - a^v_{i,j} \right| \right). \qquad (3)$$

Given the normalized weights and the same number of issues in every category being assessed by the same Likert scales, we can define the maximum possible difference as $d^{II}_{\max} = \sum_{i=1}^{6} w^u_i \cdot 5 \cdot (\max(N^{II}) - \min(N^{II})) = 20 \sum_{i=1}^{6} w^u_i = 20$ for any $u \in R$. The compatibility of the attitudes of respondent $u \in R$ to the attitudes declared by a party $v \in P$ can thus be calculated as

$$c^{II}(A^u, A^v) = 1 - \frac{d^{II}(A^u, A^v)}{d^{II}_{\max}}. \qquad (4)$$

Again $c^{II}(A^u, A^v) \in [0,1]$ for any $u \in R$ and $v \in P$ and any vector of normalized weights w^u. Under this approach we get the compatibilities summarized in Table IV.

TABLE IV.
SUMMARY OF THE COMPATIBILITY OF THE PREFERENCES EXPRESSED BY THE FIVE RESPONDENTS TO THE PREFERENCES EXPRESSED BY THE POLITICAL PARTIES UNDER SETUP II. FOR EACH RESPONDENT THE HIGHEST COMPATIBILITIES ARE DENOTED IN GREEN, THE LOWEST COMPATIBILITIES IN RED.

Setup II	Melania	Anna	Marco	Carlo	Sara
MOVIMENTO 5 STELLE	81%	77%	71%	71%	79%
LEGA	46%	63%	57%	46%	48%
PARTITO DEMOCRATICO	85%	66%	74%	69%	84%
FORZA ITALIA	55%	59%	75%	54%	60%
FRATELLI D'ITALIA	43%	53%	59%	38%	44%
ITALIA VIVA	83%	74%	81%	71%	81%
LIBERI E UGUALI	84%	64%	70%	72%	84%
+EUROPA and AZIONE	88%	65%	70%	75%	82%

TABLE V.
SUMMARY OF THE COMPATIBILITY OF THE PREFERENCES EXPRESSED BY THE FIVE RESPONDENTS TO THE PREFERENCES EXPRESSED BY THE POLITICAL PARTIES UNDER SETUP III. FOR EACH RESPONDENT THE HIGHEST COMPATIBILITIES ARE DENOTED IN GREEN, THE LOWEST COMPATIBILITIES IN RED.

Setup III	Melania	Anna	Marco	Carlo	Sara
MOVIMENTO 5 STELLE	75%	79%	74%	72%	76%
LEGA	47%	63%	61%	49%	55%
PARTITO DEMOCRATICO	85%	67%	72%	71%	79%
FORZA ITALIA	53%	63%	74%	52%	64%
FRATELLI D'ITALIA	48%	59%	59%	46%	52%
ITALIA VIVA	80%	79%	81%	72%	77%
LIBERI E UGUALI	82%	62%	66%	71%	76%
+EUROPA and AZIONE	85%	64%	68%	74%	76%

We can see that while the parties with the lowest compatibility scores are almost the same as in Setup I, the suggested parties to vote for are different for Anna, Carlo and Sara. For all three respondents the new suggestion or the additional suggestion introduced by Setup II is such that its compatibility value in Setup I was close to the highest one in Setup I. Still, the introduction of weights changes the suggestions.

C. Setup III

The previous two setups used the standard coding of the linguistic values of Likert scales. From the linguistic modelling perspective, it is, however, highly unlikely that the meanings of "strongly for" and "slightly for" would have the same distance as "slightly for" and "neutral", for example. In this setup we therefore propose a different vector of numerical meanings of the linguistic terms – one that considers the "slight…" labels to be closer to the "neutral" labels than they are to the extreme labels. We define the meaning vector N^{III} in such a way that the values remain symmetrically distributed with respect to the middle-value meaning. This way we get $N^{III} = (1,2.5,3,3.5,5)$. Even though the calibration of the meanings of the linguistic terms should be ideally done separately for each expert, we propose here a single calibration for all that at least removes the most obvious discrepancies between the linguistic values and their numerical meanings. We do not claim this is the best or optimal modification of the meaning vector – it is simply one possible modification and its effect on the final recommendations is being studied in this paper. For simplicity we do not consider the individual category weights in this setup. The difference between the attitudes of the respondent u and attitudes of a party v can be defined using (1), just $a_{i,j}^{u}, a_{i,j}^{v} \in \{1,2.5,3,3.5,5\}$ in this case. Because $\max(N^{III}) = \max(N^{I})$ and $\min(N^{III}) = \min(N^{I})$, the maximum possible difference between the attitudes of a respondent and a party can again be expressed by $d_{\max}^{I}$ and (2) can be used to calculate the compatibility of the attitudes of respondent $u \in R$ with the attitudes expressed by a party $v \in P$. This way we get the compatibilities summarized in Table V.

The different assignment of numerical meanings to the linguistic values of the Likert scale here results in just one suggestion of a party to vote for that is being changed (Carlo), but for two respondents (Melania and Anna) the new setup introduces a second party suggestion that is equally compatible as the one suggested in Setup I. Also, for Marco the suggestion of the least compatible party is different than in Setup I.

D. Setup IV

In this case we investigate what happens with the suggestion provided by our decision-support model, if we remove the potential leniency/central-tendency effects in the data by grouping together the answers provided in the positive direction ("strongly for" and "slightly for"), the answers provided in the neutral direction ("neutral"), and the answers provided in the negative direction ("slightly against" and "strongly against"). This can be achieved by defining the vector of numerical meanings of the linguistic values of the Likert scale, for example, as $N^{IV} = (1,1,3,5,5)$.

The difference between the attitudes of the respondent u and attitudes of a party v can again be defined using (1), just $a_{i,j}^{u}, a_{i,j}^{v} \in \{1,3,5\}$ in this case. Because of the choice of N^{IV} we again have $\max(N^{IV}) = \max(N^{I})$ and $\min(N^{IV}) = \min(N^{I})$, the maximum possible difference between the attitudes of a respondent and a party can again be expressed by $d_{\max}^{I}$ and (2) can be used directly to calculate the compatibility of the attitudes. Note that the actual numerical values used in the vector N^{IV} do not matter as long as the three numerical values assigned are ordered and the minimum has the same distance from the middle value as the maximum does. In other words (2) gives in this case the same result for any alternative definition of the vector $N^{IV} = (b-a, b-a, b, b+a, b+a)$ for any two real numbers a and b, $a > 0$.

The compatibilities of the attitudes of the respondents and the parties under this setup are summarized in Table VI.

This Setup also changes the initial suggestions of the parties to vote for with respect to Setup I – Melania is suggested a second option, Sara two additional (equally compatible) options and Marco is suggested a different party. In terms of least compatible parties Anna is now left with two such parties while Melania with only one.

TABLE VI.
SUMMARY OF THE COMPATIBILITY OF THE PREFERENCES EXPRESSED BY THE FIVE RESPONDENTS TO THE PREFERENCES EXPRESSED BY THE POLITICAL PARTIES UNDER SETUP IV. FOR EACH RESPONDENT THE HIGHEST COMPATIBILITIES ARE DENOTED IN GREEN, THE LOWEST COMPATIBILITIES IN RED.

Setup IV	Melania	Anna	Marco	Carlo	Sara
MOVIMENTO 5 STELLE	80%	70%	65%	75%	85%
LEGA	45%	62%	50%	47%	47%
PARTITO DEMOCRATICO	80%	63%	68%	68%	85%
FORZA ITALIA	63%	57%	75%	55%	65%
FRATELLI D'ITALIA	47%	57%	62%	45%	45%
ITALIA VIVA	82%	72%	73%	73%	77%
LIBERI E UGUALI	80%	63%	65%	68%	85%
+EUROPA and AZIONE	82%	62%	63%	70%	73%

E. A linguistic fuzzy modelling interface for decision support

It is clear from the comparison of results presented in Tables III to VI that even though the differences in the ordering of parties with respect to their compatibility with the respondents (in terms of attitudes to the important topics and their categories) can be found for all the respondents across the four setups, the relative differences in the compatibility values are rather small. Moreover, in many cases a difference of one or a just few percentage points determines which party will be suggested as the most compatible one. As such the models (setups) can be considered sensitive even to a single answer - note that if the total maximum difference is 120 for the non-weighted models, then a difference of two levels of the linguistic assessment (numerically a difference of 2) can already result in a 1% difference in compatibility.

This does not mean that the models would not be useful. It might, however, be a good idea to accompany the suggestion of a "most compatible" and "least compatible" party by a piece of information on a completely different level of granularity. We therefore propose here an additional linguistic fuzzy modelling based tool, that helps the respondents answer a more general question – "Should I consider voting for a given party?" This can be considered a question on *sufficient compatibility of the attitudes* of the respondent and the given party. In other words, this question does not ask for the ordering of the parties in terms of their compatibility. It is more of an absolute-type evaluation question aiming to identify which parties are "compatible enough". As a consequence, the answer to this question does not need to distinguish between the parties that are considered compatible enough and might not offer their ordering. We also add the opposite perspective and ask "Should I avoid voting for this party?" – with similar reasoning this is a question looking for too large a difference in the attitudes of the respondent and the party in order for the party to still constitute a reasonable choice for the respondent.

To get the necessary answers to these questions we will focus on the categories of issues and define an "acceptable difference in attitudes" $ADA_u^{C_i}$ of a respondent $u \in R$ in the category of important issues C_i as a trapezoidal fuzzy set on the universe [0,20], where $20 = d_{max}^{C_i}$ is the largest possible total difference in the numerical values of the assessments of attitudes towards the respective five important issues in category C_i, $i = 1, \dots, 6$. In other words the membership function of the fuzzy set $ADA_u^{C_i}$ (denoted for simplicity $ADA_u^{C_i}(x)$) maps [0,20] into [0,1] such that for any $x \in [0,20]$ the value $ADA_u^{C_i}(x)$ represents the extent of acceptability of that particular size of difference (1 meaning fully acceptable and 0 meaning 0% acceptable). For simplicity we will use trapezoidal-shaped membership functions that can be fully characterized by 4 characteristic values $k, l, m, n \in [0,20]$ such that $k \le l \le m \le n$, $ADA_u^{C_i}(x)$=0 for all $x \in [0,k] \cup [n,20]$, $ADA_u^{C_i}(x)$=1 for all $x \in [l,m]$ and $ADA_u^{C_i}(x)$ is linear between k and l and also between m and n. In this case we write $ADA_u^{C_i} \sim (k,l,m,n)$. We will also use the minimum triangular norm to represent the intersection of fuzzy sets (and thus the logical conjunction of their linguistic meanings) and the maximum triangular co-nom to represent the union of fuzzy sets (and thus the logical disjunction of their linguistic meanings). See [16,17] for more details on fuzzy set theory.

Let $ADA_u^{C_i} \sim (k_u^i, l_u^i, m_u^i, n_u^i)$ be a trapezoidal fuzzy number representing the acceptable values of the difference between the attitudes of the respondent and a given party with respect to category C_i defined by (valid for) the respondent $u \in R$, $i = 1, \dots, 6$. In this case it is reasonable to expect that $k_u^i = l_u^i = 0$. Let $UDA_u^{C_i} \sim (K_u^i, L_u^i, M_u^i, N_u^i)$ be a trapezoidal fuzzy number representing the unacceptable values of the difference in attitudes of the respondent and a given party with respect to category C_i defined by (valid for) the respondent u. Let (5) define the numerical value of a difference in attitudes in category C_i between the respondent $u \in R$ and a party $v \in P$. In this case we would expect that $M_u^i = N_u^i = d_{\max}^I$ for all $i = 1, \dots, 6$. Then the overall strength supporting the claim "*Respondent u should consider voting for party v.*" can be calculated as $support_u^v \in [0,1]$ using (6).

$$d^{C_i}(A^u, A^v) = \sum_{j=1}^{5} \left| a_{i,j}^u - a_{i,j}^v \right| \quad (5)$$

$$support_u^v = \\ = \min\{ADA_u^{C_1}\left(d^{C_1}(A^u, A^v)\right), \dots, ADA_u^{C_6}\left(d^{C_6}(A^u, A^v)\right)\} \quad (6)$$

Formula (6) represents the requirement of the distances in attitudes being acceptable in all the categories at the same time. The overall strength supporting the claim "*Respondent u should avoid voting for party v.*" can be calculated as $avoid_u^v \in [0,1]$ using (7):

$$avoid_u^v = \\ = \max\{UDA_u^{C_1}\left(d^{C_1}(A^u, A^v)\right), \dots, UDA_u^{C_6}\left(d^{C_6}(A^u, A^v)\right)\}. \quad (7)$$

This way (7) represents the idea that if at least one of the categories is such that the difference in preferences there is unacceptable, then one should avoid voting for that party.

Note that the definitions of the fuzzy numbers $ADA_u^{C_i}$ and $UDA_u^{C_i}$ substitute the need for the definitions of weights of the categories and directly reflect the requirements on the strength of compatibility of the respondent with the party in terms of the attitudes towards a given category of criteria. For

example, the closer the interval $[m_u^i, n_u^i]$ is to the left side of the $[0,20]$ interval, the more important the compatibility (agreement) in this category is for the respondent.

Let us now see what kind of decision support such an approach can provide. If Melania's definitions of the fuzzy numbers representing the acceptable values of the differences in the categories are:

- $ADA_{Melania}^{C_1} \sim (0,0,4,10)$,
- $ADA_{Melania}^{C_2} \sim (0,0,6,12)$,
- $ADA_{Melania}^{C_3} \sim (0,0,6,12)$,
- $ADA_{Melania}^{C_4} \sim (0,0,8,14)$,
- $ADA_{Melania}^{C_5} \sim (0,0,8,14)$,
- $ADA_{Melania}^{C_6} \sim (0,0,8,14)$,

and the fuzzy numbers representing unacceptable values of differences in the categories are:

- $UDA_{Melania}^{C_1} \sim (10,16,20,20)$,
- $UDA_{Melania}^{C_2} \sim (12,16,20,20)$,
- $UDA_{Melania}^{C_3} \sim (12,16,20,20)$,
- $UDA_{Melania}^{C_4} \sim (14,18,20,20)$,
- $UDA_{Melania}^{C_5} \sim (14,18,20,20)$,
- $UDA_{Melania}^{C_6} \sim (14,18,20,20)$,

then the following decision-support would be provided. Decision support is formulated linguistically as an answer to the original question, the number in bracket represents the $support_u^v$ value:

- Consider voting for Movimento 5 Stelle (100%),
- Consider voting for +Europa e Azione (100%).
- Consider voting for Italia Viva (100%),
- Consider voting for Partito Democratico (83%),
- Consider voting for Liberi e Uguali (50%),

and also (now values in brackets represent the $avoid_u^v$ values):

- Avoid voting for Fratelli d'Italia (100%),
- Avoid voting for Lega (66%), and
- Avoid voting for Forza Italia (33%).

We can see that those parties with overall high compatibilities in Setups I-IV are suggested for consideration while those that were scored frequently as incompatible are suggested to be avoided. This approach does not provide a clear answer to the question whom to support, but it seems to be able to summarize the situation reasonably well, to provide linguistic outputs and to cover the information obtained through the use of setups I-IV. It is definitely an approach to consider at least as an additional source of information for informed choice in the election situation. Note that linguistic summaries are being applied more and more often recently [18,19] because of their easy understandability by laymen.

IV. Discussion

Setups I-IV investigate different possible uses of a 5-point Likert scale for the assessment of attitudes of parties and respondents towards a given set of thirty important issues grouped into six categories. The important take-away message of these approaches is the fact that there were differences in the suggestion of the most compatible and least compatible party at least for one of the respondents between each pair of setups. This implies that the choice of the setup has to be done correctly – mainly the coding of the linguistic values and the need for the reflection of perceived importances of the issues (or categories thereof) by the respondents, potentially the need for countermeasures to the central tendency/leniency issue need to be well thought through. As such the choice of the setup to provide correct voters' decision support and also correct research data for political sciences is not trivial. All the four setups considered in this paper define the suggestion of a party to vote for based on the maximum compatibility score, that might have a very close runner up with only a slightly lower compatibility score, which is then, however, discarded. The setups thus seem to be very sensitive to the precision of the actual values provided by the respondents and the experts assessing the programs of the parties.

One can also argue that the very goal of the decision support is not specified well in the four setups. How reasonable is it to look for the most compatible party (in terms of attitudes to the important issues)? First of all, we need to understand that this goal calls for a relative-type evaluation. This type of evaluation is, by definition, dependent on the set of available parties and also on the actual values of compatibilities. If the set of parties does not contain all the parties to be considered, the relative-type decision support can be biased. There is also one more potential, and well known, issue connected with relative type evaluation and decision support based on such evaluation – the inability of the model to assess whether the best choice that is to be presented to the decision-maker as a decision-support is "good enough" to be accepted. The most compatible party might still not be a party to vote for, if its compatibility with one's attitudes is low. Unfortunately, if the goal is formulated in terms of finding the "most compatible" party, then the answer we are getting is formally correct, even though it might not be practically correct or relevant. One should therefore at least make sure that the set of parties is complete. Then it might be justifiable to accept that the most compatible party is the one to vote for, as there is no better one available. Still, it seems that not knowing whether the suggested party is "compatible enough" with the voter to choose that party in reality can seriously bias the research based on such data.

We have therefore suggested a linguistic fuzzy modelling based approach to the assessment of the available data. The proposed linguistic fuzzy modelling based tool allows for the definition of an acceptable magnitude of difference in attitudes and also for the definition of an unacceptable (too high) magnitude of difference in preferences. Having done so the respondent (or the researcher conducting a research on voter preferences or assumed choices) can be provided with a list of parties that are "sufficiently close" in terms of the attitudes to the categories of important issues to be considered for selection and also with a list of parties that "differ in their attitudes too much" to be voted for. Both that with the

measure of support or strength of that suggestion represented by the $support_u^v$ and $avoid_u^v$ values respectively.

Clearly the linguistic approach provides a different perspective on the voters' potential choices than the usual models aiming on suggesting the most fitting solution. Instead of a single party with the best compatibility (which can be the best by a very slim margin) it lists all the parties that are "acceptable enough". The other side of the universe is also covered by listing those parties that are "too different in their attitudes towards the important issues" to be voted for.

It is also interesting to note that the linguistic fuzzy modelling perspective offers more customizability to the decision support. First, the membership functions of the fuzzy numbers that represent acceptable and unacceptable values of the differences are defined in an absolute way. This means that the respondent can define the thresholds for (un)acceptable values of differences prior to the very task of data input, independently of the already available data. The information is provided in the units of the distance of the attitudes (expressed numerically) and as such these units can be expected to be well understood by the respondent/voter. It might also mean that the (un)acceptability thresholds required here (i.e. the definitions of the membership functions of the fuzzy numbers that represent (un)acceptable magnitudes of differences in attitudes) are much easier to define than abstract unitless weights of the categories. And as we have discussed previously, the fuzzy numbers defined for this purpose essentially perform a similar function as the weights would perform for example in the Setup II.

Second, the aggregation using the min t-norm and the max t-conorm respectively does not allow for compensation, unlike the approaches relying on (weighted) sums or (weighted) arithmetic means. In other words, the model is built in such a way that a clear incompatibility in a single category of important issues (the difference in attitudes is acceptable with the strength of 0) results in the strength of the suggestion to vote for that party to be 0% regardless of the compatibility in the other categories of important issues. Analogously the suggestion not to vote for a party is provided with the strength of 100% if the difference in attitudes is considered 100% unacceptable in a single category of important issues regardless of the compatibility in the other categories. This is a rather strict approach, that can still be customized by the choice of different t-norms and t-conorms. In any case it reflects a risk-averse approach to the evaluation of the compatibility of the attitudes of respondents and parties and as such it provides a good benchmark to the other models. Alternatively, one could also consider representing the attitudes of the parties and the respondents including the perceived relevance of the issues and apply the tools of interval-valued semantic differential [20] and see which parties are the closest in the semantic space in terms of the n-dimensional representation of their attitudes. This however remains out of the scope of this particular paper.

V. Conclusion

The paper investigated the possibility of constructing a voter decision-support using surveys with Likert-type answers. Based on the studied Setups we can conclude that the use of calibrated and non-calibrated Likert scales can result in a different suggestion (decision-support being provided). The inclusion of weights of categories of important issues also influences the results of the decision support. Lastly switching to the +/0/- understanding of the values of the Likert scales also changes the decision support provided by the models. This all implies that if a calibration of the Likert scale is needed (i.e. if its linguistic terms cannot be considered equidistant) or if the categories have different perceived importance for the choice to be made by the voter, then these need to be reflected in the model. Otherwise, the decision-support can be biased. We have also proposed a linguistic fuzzy modelling interface for the evaluation of the data obtained through Likert scales that allows for the reflection of voters' preferences and priorities by the definitions of (un)acceptable values of differences. These definitions have the benefit of being absolute-type (i.e. they directly specify what values are acceptable and what values are not, independently of the actual available data), no standardization is needed), they can be expressed directly in the units of the magnitude of difference and also introduce "(un)acceptability thresholds" above or below which the parties are no longer discriminated and considered "too different" or "compatible enough" to be either discarded or considered viable choices. This reduces the potentially undesirable sensitivity of the model to minor changes in the answers provided by the Likers scales. Even though the definition of the fuzzy-number representations of the (un)acceptable magnitudes of differences in attitudes might be slightly more demanding for the voter, we still strongly recommend this approach to be able to obtain a less sensitive and well understandable decision support in the voting process. And also, to obtain a fresh and novel type of data for the pre-election surveys and analytics.

VI. Acknowledgement

The authors would like to express their thanks to Douglas Adams [21] for the inspiration for the title of the paper.

References

[1] S. R. Muller, and C. E. Thomas, "Election Infrastructure Security: Grants and Reimbursement to the States for Usage of their National Guards in State Active Duty Status to Provide Cybersecurity for Federal Elections," in *Proceedings of the 2020 International Conference on Research in Management & Technovation,* Shivani Agarwal, Darrell Norman Burrell, Vijender Kumar Solanki (eds). ACSIS, Vol. 24, pp. 73–78, 2020, doi: 10.15439/2020KM7

[2] S. P. Robertson, "Voter-centered design: Toward a Voter Decision Support System," *ACM Trans. Comput. Interact*., vol. 12, no. 2, pp. 263–292, 2005, doi: 10.1145/1067860.1067866.

[3] R. Likert, "A technique for the measurement of attitudes," *Arch. Psychol*., vol. 22, no. 140, pp. 5–55, 1932.

[4] J. C. Rogowski, "Voter Decision-Making with Polarized Choices," *Br. J. Polit. Sci*., vol. 48, no. 1, pp. 1–22, 2018, doi: 10.1017/S0007123415000630.

[5] J. Stoklasa, T. Talášek, and J. Stoklasová, "Semantic differential for the twenty-first century: scale relevance and uncertainty entering the semantic space," Qual. Quant., vol. 53, no. January 2019, pp. 435–448, May 2019, doi: 10.1007/s11135-018-0762-1.

[6] J. Stoklasa, T. Talášek, and J. Stoklasová, "Reflecting emotional aspects and uncertainty in multi-expert evaluation: one step closer to a soft design-alternative evaluation methodology," in *Advances in Systematic Creativity: Creating and Managing Innovations*, L. Chechurin and M. Collan, Eds. Palgrave Macmillan, 2019, pp. 299–322.

[7] J. Stoklasová, T. Talášek, and J. Stoklasa, "Attitude-based multi-expert evaluation of design," in *Intelligent Systems and Applications in Business and Finance*, P. Luukka and J. Stoklasa, Eds. Springer, (in press).

[8] G. Hoang, J. Stoklasa, and T. Talášek, "First steps towards a lossless representation of questionnaire data and its aggregation in social science and marketing research," in *Proceedings of the international scientific conference Knowledge for Market Use 2018*, 2018, pp. 112–118.

[9] J. Stoklasa, T. Talášek, J. Kubátová, and K. Seitlová, "Likert scales in group multiple-criteria evaluation," *J. Mult. Log. Soft Comput*., vol. 29, no. 5, pp. 425–440, 2017.

[10] J. Stoklasa, T. Talášek, and P. Luukka, "Fuzzified Likert scales in group multiple-criteria evaluation," in *Soft Computing Applications for Group Decision-making and Consensus Modeling*, vol. 357, M. Collan and J. Kacprzyk, Eds. Springer International Publishing AG, 2018, pp. 165–185.

[11] G. Norman, "Likert scales, levels of measurement and the 'laws' of statistics," *Adv. Heal. Sci. Educ.*, vol. 15, no. 5, pp. 625–632, 2010.

[12] A. Furnham, "Response Bias, Social Desirability Dissimulation," *Pers. Individ. Dif.*, vol. 7, no. 3, pp. 385–400, 1986.

[13] A. Furnham and M. Henderson, "The good, the bad and the mad: Response bias in self-report measures," *Pers. Individ. Dif.*, vol. 3, no. 3, pp. 311–320, 1982.

[14] J. Stoklasa and T. Talášek, "On the use of linguistic labels in AHP: calibration, consistency and related issues," in *Proceedings of the 34th International Conference on Mathematical Methods in Economics*, 2016, pp. 785–790.

[15] J. Stoklasa, *Linguistic models for decision support*. Lappeenranta: Lappeenranta University of Technology, 2014.

[16] G. J. Klir and B. Yuan, F*uzzy Sets and Fuzzy Logic: Theory and Aplications*. New Jersey: Prentice Hall, 1995.

[17] D. Dubois and H. Prade, Eds., *Fundamentals of Fuzzy Sets*. Massachusetts: Kluwer Academic Publishers, 2000.

[18] J. Stoklasa, T. Talášek, and J. Stoklasová, "Executive summaries of uncertain values close to the gain/loss threshold – linguistic modelling perspective," *Expert Syst. Appl.*, vol. 145, p. 113108, 2020, 10.1016/j.eswa.2019.113108

[19] Ł. Sosnowski, and T. Penza, "Generating Fuzzy Linguistic Summaries for Menstrual Cycles," *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 21, pp. 119–128, 2020, doi: 10.15439/2020F202

[20] J. Stoklasová, "Interval-valued semantic differential in multiple criteria and multi-expert evaluation context: possible benefits and application areas" *Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics*, Jan Stoklasa, Pasi Luukka and Maria Ganzha (eds). ACSIS, Vol. 29, pp. 53–61, 2021, doi: 10.15439/2021B3

[21] D. Adams, *The Hitchhiker's Guide to the Galaxy*. New York: Harmony Books, 1980.

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 19–28
DOI: 10.15439/2021B1
ISSN 2300-5963 ACSIS, Vol. 29

Volatility Risk Premium and European Equity Index Returns

Antti Ihalainen
Varma Mutual Pension Insurance Company, P.o.Box: 1, 00098 Varma, Finland
Email: antti.ihalainen@varma.fi

Sheraz Ahmed
School of Business and Management, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland
Email: sheraz.ahmed@lut.fi

Eero Pätäri
School of Business and Management, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland
Email: eero.patari@lut.fi

***Abstract*—For most of the time, equity index option implied volatilities exceed the corresponding realized volatilities. The resulting volatility risk premium seems to be directly linked with the equity risk premium, which motivates to study whether this investor risk aversion-related premium has explanatory power on the future stock index returns. Based on several linear regression models, this study shows that volatility risk premiums can explain a non-trivial fraction of the aggregate stock returns in Europe. Furthermore, both local and global risks are found to be systematically priced. Our findings confirm the consistency and deterministic power of volatility risk premium in the European equity markets. Additionally, the evidence supports the hypothesis that the global volatility risk and equity market premium are inter-linked.**

I. Introduction

VOLATILITY risk premium (VRP) represents the compensation for investing in risky securities instead of risk-free assets. It is essential to understand how investors deal with the uncertainty and variance of future returns not only in risk management, asset allocation, and pricing purposes, but also in attempts to understand the behavior of financial assets in general [11]. This study aims to explore the components that drive the equity risk premium, and the expected equity index returns, and to identify the risks that are ultimately being compensated for investors.

There is broad evidence of volatility risk premium and its significant explanatory power over expected stock returns in the U.S. market. [3] examined volatility risk premium using statistical properties of delta hedged option portfolios constructed from S&P 500 index options and concluded that negative volatility risk premium and mean delta-hedged gains share the same sign. [11] presented the existence of a systematic variance risk factor in the U.S. stock market as evidenced by highly negative risk premium. [8] found similar results derived from the squared VIX index and realized variance measures calculated using intraday data. They provided empirical evidence that stock market returns are predictable from the difference between model-free implied variance and realized variance and concluded that a strong positive relationship exists between the variance risk premium and following equity index returns in the U.S.

[12] studied higher moments estimated from the S&P 500 index option data and found highly negative and economically significant market skewness risk premium related to the cross-section of stock returns. Focusing on the higher moments of the probability distribution, [20] introduced a concept of a synthetically built skew swap to explore the relationship between the option implied skew and realized skew. They showed that skew risk premium (SRP) can explain almost half of the implied skew in index option prices, implying that common risk factors drive both variance and SRP.

Comprehensive study of volatility risk premiums in the European stock market has not been implemented at broad aggregate level. [16] studied moment risk premia in Europe using portfolio sorting techniques to obtain volatility risk from Euro Stoxx 50 index options and reflected it to a cross-section of STOXX Europe 600 index constituent returns. Evidence of negative variance risk premium and positive skewness premium was found amongst the individual stocks. Their findings were robust to the inclusion of other risk factors such as size, book-to-market, and momentum.

In contrast to the majority of existing studies on the U.S. stock market returns, this study focuses solely on European stock markets. Although [16] provide valuable insights of individual European stocks, there is a lack of evidence on aggregate stock market returns. A comprehensive study of a broad set of European indexes is needed to distinguish whether investors require compensation for the volatility risk, whether these premiums show predictive power on expected returns in the short term, and in addition, whether global variance risk premium exhibits significant predictive power on future European equity index returns.

The primary contribution of this study is to provide new empirical evidence on the predictive power of option-implied information on subsequent aggregate equity returns. The aim is to provide new evidence from understudied European stock markets and gain a better understanding of the risks and their pricing in expected stock returns. Since the information content of option prices seems to be superior to the historical measures, especially for short-term horizons, this paper focuses on one-month (21 business days) and three-month (63 business days) equity index excess returns. The examined equity indexes are Euro Stoxx 50 (European-wide), DAX index (Germany), FTSE 100 index (United Kingdom), SMI index (Switzerland), and STOXX Europe 600 (European-wide). Model-free volatility indexes are used to capture information in option prices. Option-implicit information is then used to explain the subsequent returns of these stock indexes. The studied period spans from the beginning of 2007 until the end of October 2017. This period

This work was carried out during the postgraduate studies of Antti Ihalainen at LUT University. Finland.

is selected in order to include the regimes of high and low stock market returns and volatility during the most recent times. The 2007-2009 financial crisis, EU sovereign debt crisis 2009-2012, and the period of growth from 2012 to 2017 are all included in the sample period. Special attention is paid to *ex-ante* volatility premiums (forward-looking) in explaining the future equity index returns. The distinction between *ex-ante* and *ex-post* premiums (future and historical, respectively) is important because the expected returns of the financial assets and option prices are determined on the basis of past, present, and future information of the underlying assets' volatility at any given point of time according to the notion of strong-form market efficiency.

Potential existence of global volatility risks is examined by using the information in VIX index. All the implied volatility indexes being examined are calculated in a similar model-free manner and they utilize a broad set of out-of-the-money (OTM) call and put options expiring in 30 days, providing risk-neutral and model-free expectations of second and third moments of risk-neutral probability distributions (RNPDs).

According to the results, volatility risk premiums explain a non-trivial fraction of the equity index return in Europe and both local and global volatility risk are systematically priced into the European equity index returns. The findings of the explanatory power of volatility risk premiums on aggregate stock market returns are consistent with the previous evidence reported on the U.S. market.

This remainder of the paper is structured as follows: previous literature on volatility risk premium and option-implied information is summarized in section II. Section III describes in detail the data and methodology. Section IV presents empirical results of the univariate and multivariate regression analyses. Finally, section V concludes and discusses the limitations and suggestions for future research.

II. Literature Review

A. Volatility and risk aversion

As stock market volatility seems to be harmful to most investors, they demand compensation. It is well-established that market volatility of equity returns varies over time. While time-varying volatility changes the expectations of future returns or risk-return tradeoff, rational investors whose utility increases as a function of wealth require compensation for being exposed to the changes in market volatility. Yet the relationship between market volatility and stock returns has proven to be ambiguous.

[1] studied the pricing of aggregate volatility risk in the cross-section of equity returns and found that stocks with high sensitivities to innovations in market volatility have low average returns. They used changes in implied volatility index VIX as a proxy of changes in market volatility and made a reservation regarding the use of the VIX, noting that it incorporates both stochastic volatility and the stochastic volatility risk premium. According to their research, aggregate volatility may be a priced factor, partly because assets with high sensitivities to volatility risk hedge against the risk of substantial market declines.

In lieu of risk aversion, [4] noted that out-of-the-money options became remarkably expensive during the year prior to the market crash of October 1987. His interpretation was that conditional expectations in jumps in asset prices revealed significant time variation. According to the [5], the volatility smile should be a flat line, because only one volatility parameter rules the underlying stochastic process based on which all options are priced. [18] showed that the observed RNPDs describing investor expectations for equity indexes across the globe are mostly left-skewed and leptokurtic. The corresponding distributions of realized returns are somewhat lognormally distributed, implying that investors are pricing some non-occurring risks in asset prices. These downward sloping volatility smirks and negatively skewed risk-neutral densities representing the "crash-o-phobia" phenomenon, meaning that investors, who consider a market crash as a risk, buy OTM put options to cover their positions and to put a floor on their maximum losses.

Building on the notion of investor risk aversion and fears of a crash, [24] studied the perceived *ex-ante* risks by using S&P 500 index options. They made a distinction between diffusion risk and jump risk, the former referring to the quadratic variation of the realized price process, and the latter to the anticipated risk of large price movements. Their findings showed that the premium embedded in option prices is, on average, 40% higher than the premium required to compensate for the realized stock returns and support the risk aversion-explanation for the equity premium puzzle. [13] studied volatility and jump risk. Their result showed strong evidence of a priced jump risk, and stocks with high sensitivities to jump and volatility risk had low expected returns. Investors' risk aversion was revealed through the jump and volatility premiums. Implied volatilities can be high due to high volatility expectations, high risk aversion, or a combination of these, therefore using the implied volatilities as an indicator of general risk aversion is somewhat fallacious. Nevertheless, implied volatilities provide a valuable tool for revealing the risk-neutral expectations of investors when combined with corresponding realized volatility information. This leads us to the use of the volatility risk premiums instead of pure volatility estimates in predicting expected returns.

B. Options-implied information

Option prices reflect the market's common assessment of the probability distribution of the underlying asset prices on the date of expiry, and this assessment is adjusted to include the degree of investors' risk tolerance. As the option-implicit factors provide the market's forward-looking risk-neutral approximation of the expected prices of an underlying, they provide RNPDs that cannot be derived from historical prices.

The superiority of option implied volatilities over the backward-looking volatility estimates is widely documented by [6], [14], [19], and [22], among many others. Traditionally, some specific option pricing models have been used to extract option implied information from option prices, but this kind of approach have some shortcomings: Probably the most important of these is that the implementation of a certain model for the purposes of implied volatility estimation

is always a combined test of the option-implicit information content and the option model itself.

By using a continuous set of options with strike prices from zero to infinity, [9] showed that it is possible to form the entire risk-neutral probability distribution. [10] extended the work of [9] by deriving implied volatility from a set of current option prices without the use of any specific option model. The suggested model-free approach does not assume a constant volatility or suffer from the inconsistencies of traditional models. [19] showed that the calculation of VIX, which is the most well-known model-free implied volatility (MFIVI) index, is essentially consistent with the theoretical framework of [10].

Following [26], VIX is calculated on the basis of near- and next-term put and call options with more than 23 days but less than 37 days to expiry. Once each week, the index options used to calculate the VIX are rolled to new maturities, making the previous next-term options (more than 30 days until expiry) now near-term options (30 or fewer days until expiry). Both standard monthly options expiring on the 3rd Friday of each months and weekly options expiring every Friday are employed in the calculations. In order to make the time-to-expiry calculations more straightforward, monthly options are deemed to expire at the open of trading on the S&P 500 settlement day (i.e., on the 3rd Friday of the month), whereas for the weekly options, the expiry is assumed to at the close of trading (i.e., 4:00 p.m. ET).

The risk-free interest rates used in the calculations are yields of the U.S. Treasury bills maturing closest to the corresponding S&P 500 index option, implying that the used risk-free rates may vary between near- and next-term options. The options included in the VIX index calculation of are out-of-the-money put and calls and centered around an at-the-money strike price. Only the options quoted with non-zero bid prices are used in the calculations. Finally, the put and call prices for the same strike price are averaged to produce a single value. After the options included in the VIX calculation are identified, the variance is first calculated as follows:

$$\sigma^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{rT} Q(K_i) - \frac{1}{T}\left[\frac{F}{K_0}\right]^2 \quad (1)$$

where $\sigma = VIX/100$, T is time to expiration, K_0 is the first strike price below the forward index level F, K_i is the strike price of the i^{th} OTM option (call if $K_i > K_0$, put if $K_i < K_0$, and both if $K_i = K_0$), r is the risk-free interest rate, and $Q(K_i)$ is the average of bid-ask spread for each option with strike K_i. ΔK is defined by halving the difference between the strikes on both sides of K_i:

$$\Delta K_i = \frac{K_{i+1} - K_{i-1}}{2} \quad (2)$$

The formula presented in Equation 2 is then applied for both near- and next-term options by using times to expirations T_1 and T_2, respectively. The resulting If_1^2 (for T_1 near-term options) and If_2^2 (for T_2 next-term options) are then averaged over 30 days. The VIX index value is obtained by taking the square root of the 30-day weighted average of σ_1^2 and $\sigma_2^2\sigma_2^2$, and multiplying it by 100:

$$VIX = 100 \times \sqrt{\left\{T_1\sigma_1^2\left[\frac{N_{T_2} - N_{30}}{N_{T_2} - N_{T_1}}\right] + T_2\sigma_2^2\left[\frac{N_{30} - N_{T_1}}{N_{T_2} - N_{T_1}}\right]\right\} \times \frac{N_{365}}{N_{30}}} \quad (3)$$

[25] conducted a comprehensive global review of all available implied volatility indexes, concluding that the European MFIVIs, namely VSTOXX, V1X-NEW, VSMI, and VFTSE index calculation methodologies follow closely the one introduced by CBOE VIX. The methodology for the calculation of the VIX's European equivalents involves a summation over a band of OTM option prices. The intuition behind the use of option implied information is relatively simple: a cross-section of option prices (and implied volatilities) for the same underlying asset and the same maturity reveals the RNPD, which then reveals an estimate of the future state and its pricing at the maturity of the options' cross-section. Specific to Eurex-based indexes VSTOXX, V1X-NEW, and VSMI is that they are calculated based on eight expiry months and a sub-index is calculated for each option expiry. Linear interpolation is then used to calculate the main indexes from the sub-indexes (E.g., [27] describes the calculation and interpolation scheme of VSTOXX in detail).

C. *Volatility risk premium*

The academic literature has documented a consistently positive spread between implied and realized volatilities. Figure I shows one-month volatility spread of Euro Stoxx 50 index over the studied period. On average, volatility selling over the Euro Stoxx 50 index has been profitable over the ten-year period.

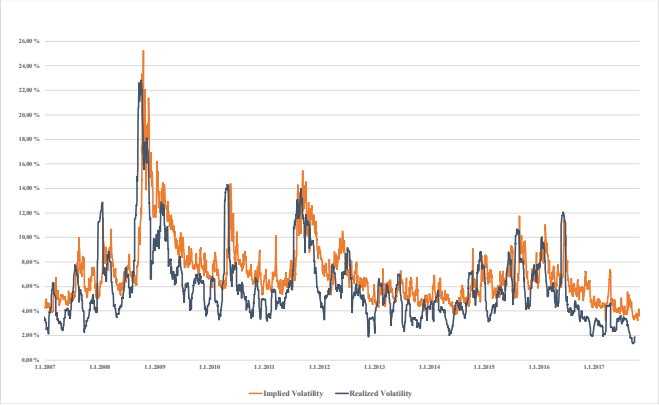


Figure I: Daily volatility spread of Euro Stoxx 50 index from Jan 2007 to Oct 2017

This difference between risk-neutral and realized volatility is proven to have predictive power for equity returns on both individual and aggregate level. The second moment of a return distribution, quantified by variance or volatility, seems to exhibit significant explanatory power on following equity returns. Using the S&P 500 and S&P 100 index options, [3] present the VRP in a non-parametric way by analyzing delta hedged option positions. They show that the implication of the volatility risk premium is that the profits on delta-neutral option strategy are non-zero and are determined mutually by the volatility risk premium and option vega. Moreover, the volatility risk premium and delta

hedged gains seem to have the same negative sign. Negative VRP implies an equilibrium, where equity-index options act as a hedge to the market portfolio. Investors are willing to pay a premium to hold options in their portfolio for hedging purposes, which makes options' price higher than it would be when volatility is not priced.

The intuition behind the existence of VRP is again relatively simple: if investors do not want to be exposed to the variation in prices and therefore in expected returns, they require being compensated for it. [2] showed that implied return distribution of the S&P 100 index was much more volatile than its physical equivalent. They concluded that "rational risk-averse investors are sensitive to extreme loss states and willing to counteract these exposures by buying protection." Investors need to hedge against extreme losses drives up the option implied probability of occurrence relative to the actual probability of occurrence, causing the volatility spread to widen.

The risk-neutral expectation of variance can also be interpreted as variance swap rate, following the methodology introduced by [11]. The fixed leg of the swap is the option implied variance, and the floating leg represents the realized variance. The spread between the risk-neutral and physical values unveils the variance risk premium. Variance swap rate represents the market's risk-neutral expected value of the realized variance and is synthesized by a linear combination of option prices. Their findings prove the existence of the common and stochastic risk factor, that the Fama-French factors cannot explain. This negative premium indicates that investors regard rises in market volatility as an unfavorable shock and are willing to pay a large premium against market volatility increases. Writing variance swaps is therefore on average profitable, since the fixed swap rate is prone to exceed the floating rate.

The evidence of the existence of return impacts of variance risk premium has been established both on an aggregate market level and an individual stock level. [17] focused on the cross-section of large-cap stock returns and found that an individual stock's expected return increases with its variance risk premium. They used a model-free approach and found that the top VRP quintile stock returns outperform the stocks in the lowest. Low VRP stocks seem to be serving as useful hedges against systematic and therefore also have lower expected returns. Investors seem to have preferences about equity volatility at both individual and aggregate levels.

[8] proved the existence of a significant risk-return relationship and found that the variance risk premium is most effective in forecasting equity index returns in quarterly to six-month horizons, even though the results hold for shorter one-month and longer annual periods as well. Their results hold when other, more common equity index return predictor variables are included in multiple regressions. The predictive power of P/E ratios becomes more effective and significant when combined with the variance risk premium.

[15] argued that the variance risk premium is closely linked to the uncertainty of economic fundamentals. They found a strong statistically significant relationship between the variance risk premium and aggregate stock market returns, and their findings support the superior short-term predictive power of VRPs. They concluded that the variance risk premium is an extremely useful tool in measuring the market's perceptions of uncertainty and the risks of influential shocks to the economy. Not only does the VRP provide a measure for uncertainty perceptions, but it is also a useful tool in understanding what preferences are able to map the risk onto asset prices. VRP can be seen to provide a vehicle to capture investor risk aversion and its pricing in equity markets.

In this paper, the volatility risk premium is defined to be the difference between the model-free implied volatility (*IV*) and the corresponding realized volatility (*RV*) estimated as standard deviation of each equity index. The formula of *RV* is presented in Equation 4.

$$RV = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}, \tag{4}$$

where x refers to daily logarithmic returns and $\overline{xx}$ is the corresponding average return calculated over n trading days.

III. Data and methodology

The European equity indexes employed in the empirical tests are Euro Stoxx 50, DAX, FTSE 100, SMI, STOXX Europe 600, and S&P 500. Descriptive statistics of all the used equity indexes are presented in Table I. These equity indexes also have dedicated model-free one-month implied volatility indexes. The option-implied volatility is derived from the corresponding options underlying the volatility index of the respective stock index. The risk premiums (implied minus realized volatilities) are used to test their prediction power on future returns of the underlying indexes.

Table I
Descriptive statistics of monthly excess returns of the equity indexes.

	Euro Stoxx 50	*DAX*	*FTSE 100*	*SMI*	*STOXX Europe 600*	*S&P 500*
Mean	-0.001837	0.004397	0.000475	-0.000566	-0.00035	0.003647
Median	0.005037	0.012246	0.006023	0.005383	0.008328	0.008333
Maximum	0.135916	0.153838	0.080398	0.095632	0.125243	0.101949
Minimum	-0.162803	-0.214371	-0.143831	-0.121759	-0.146255	-0.188121
Standard Deviation	0.052744	0.055848	0.040235	0.037745	0.044108	0.043425
Skewness	-0.599381	-0.830805	-0.659834	-0.525199	-0.690271	-0.987473
Kurtosis	3.63626	4.978845	3.855028	3.468004	4.325541	5.511011
N	130	130	130	130	130	130

The table shows one-month logarithmic excess returns of the risk-free rates for all equity indexes. The risk-free rates used in excess return calculations are three-month Euribor for Euro Stoxx 50, DAX, FTSE 100, SMI, and STOXX Europe 600 and three-month Libor for S&P 500.

The existence of the global risks is analyzed by using volatility and skew risk premiums (SRP) embedded in the U.S markets explaining the European stock market returns with S&P 500 equity index (SPX) and CBOE's model-free volatility index. The US market volatilities are used to test the relationship between global sources of risk and local European stock index returns. This is an important step to see the extent of global risk premiums affecting the aggregate European stock returns.

Each index prices are downloaded from Refinitiv Eikon in their base currency, and the returns are reported in

percentages. The risk-free rate used in the calculation of the excess returns of European (US) indexes is three-month EURIBOR (LIBOR). Risk-free rates are modified considering the day count convention and conversion into continuously compounded rates. All the employed variables and their abbreviations are presented in Appendix I.

The calculation methodology of VRP and SRP closely follows the approach of [7] and [8]. Annualized implied volatilities obtained from MFIVIs are translated to monthly (quarterly) volatilities simply by dividing the index levels by $\sqrt{12}$ ($\sqrt{4}$). This approach has a clear advantage from the viewpoint of forecasting. One-month VRP at time t is obtained by using the implied volatility observed at t for the time period $t + 21$ and subtracting the realized volatility that is calculated using the returns of the preceding month. One-month ex-post volatility and skew risk premiums (EPVRPs and EPSRPs, respectively) for each time interval (t) are obtained by subtracting the *ex-post* observed realized volatility of time period $t + 21$ from the implied volatility observed at time t. Autoregressive conditional heteroscedasticity effects and autocorrelation effects are tested in post estimation purposes by conducting ARCH and the Breusch-Godfrey tests. As evidence of both heteroscedasticity and autocorrelation is found in OLS standard errors, the regressions are run by adjusting the standard errors by the Newey-West [21] procedure, which simultaneously controls for the biases stemming from heteroscedasticity and autocorrelation, therefore providing a better estimation accuracy.

The forecasts are based on linear regressions of the excess returns of European equity indexes. Both univariate and multivariate regression models are used to determine the relationship between risk premium and returns. All the conducted regressions with volatility and skew risk premium variables, as well as control variables, can be formally expressed in the general form of regression equation presented in Equation 5.

$$ER = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \cdots + \beta_k\alpha_k + \varepsilon, \qquad (5)$$

where ER = daily excess return (over and above the risk-free rate), β_0 = intercept term, $\beta_{1\ldots k}$ = regression slopes of risk premiums or control variables represented by $\alpha_{1\ldots k}$, respectively.

The observed period for implied volatilities, realized volatilities, and closing prices of the stock indexes spanned from 28.11.2006 to 31.10.2017. This analysis directly employs the model-free implied volatilities provided by Refinitiv Eikon. The Term spread (TERM) is the difference between 10-year and 3-month government liability yields, explicitly 10-year Bund yield and 3-month German government liability BD deposit. Default spread (DEFT) is defined as the Difference between Moody's Baa Corporate bond yield and 10-year Treasury yield. The main explanatory variables are *ex-ante* volatility risk premiums (VRP) and *ex-post* volatility risk premiums (EPVRP) of the Euro Stoxx 50, DAX, FTSE 100, SMI index, and S&P 500 index.

IV. Results

The results show that the European volatility risk premium is able to explain the subsequent equity index returns for next one month. Table II shows that the *ex-post* observable volatility risk premium is significantly and substantially related to the European equity index returns. On average, one percentage point increase in observed volatility risk premium leads to 0.81%–1.64% increase in monthly returns, highly significant t-statistics of the volatility risk premium coefficients altering correspondingly from 4.05 all the way to 9.88. An increase in the volatility risk premium can potentially result from an increase in implied volatility, decrease in realized volatility, or as a result of occurrence of both.

TABLE II
MONTHLY REGRESSION RESULTS FOR VRP VARIABLES

One-month returns				
	Euro Stoxx 50	*DAX*	*FTSE 100*	*SMI*
VRP coefficient	0.0195	0.0392	0.0894	0.0095
	(-0.06)	(0.11)	(0.39)	(0.04)
Constant	-0.0021	0.0032	-0.0004	-0.0007
	(-0.29)	(0.44)	(-0.09)	(-0.13)
Adj. R Squared (%)	-0.77	-0.77	-0.66	-0.77
EPVRP coefficient	1.4576	1.6483	1.2434	0.8169
	(9.88)***	(7.45)***	(6.95)***	(4.05)***
Constant	-0.0195	-0.0121	-0.0184	-0.0085
	(-5.27)***	(-2.86)***	(-4.01)***	(-2.74)***
Adj. R Squared (%)	40.56	37.82	35.85	21.26
N	129	130	129	130

*p < 0.1; **p < 0.05; ***p < 0.01

The table shows univariate OLS-regression results with the Newey-West standard errors for one-month (21 business days) excess returns of the European equity indexes. These excess returns are explained by the volatility risk premium variables of each index. Corresponding NW-based t-statistics are presented in parentheses.

The largest return-impact is for the DAX index. One percentage point increase in *ex-post* volatility risk premium increasing the one-month excess returns by 1.65%. The results for all of the examined indexes are significant at the 1% level. Volatility risk is clearly priced in the aggregate equity markets, and the volatility risk premium provides consistent explanatory power on subsequent equity index returns.

The strong explanatory power might be due to the EPVRP's relationship to the equity risk premium, or it can result from the observation that realized volatilities are prone to be higher in the downward markets, and lower in upward markets. It is likely that both of these explanations are right and that the substantial return impact of *ex-post* observed volatility risk premium is a joint result of the connection of the volatility risk premium to the equity risk premium and volatility's connections to market trends.

Since the *ex-post* measure does not provide forecasting value in a real decision-making context, special interest lies on *ex-ante* volatility risk premium. VRP does not deliver statistically significant forecasting results for European equity index returns in one-month periods, but for Euro Stoxx 50, DAX, and SMI index, the VRP slopes become significant in quarterly periods (see Table III). One percentage increase in three-month VRP leads on average to a 1.35% increase in quarterly returns of Euro Stoxx 50 index, a 1.28% increase of quarterly SMI index returns, and 1.14% increase in quarterly returns of the DAX index. Local VRP seems to exhibit significant predictive power over following equity index returns when tested in isolation.

TABLE III
QUARTERLY REGRESSION RESULTS FOR VRP VARIABLES

Three-month returns				
	Euro Stoxx 50	*DAX*	*FTSE 100*	*SMI*
VRP coefficient	1.3472	1.1390	0.6153	1.2781
	(2.41)**	(1.77)*	(1.59)	(2.47)**
Constant	-0.0262	-0.0004	-0.0061	-0.0166
	(-1.73)*	(-0.02)	(-0.55)	(-1.11)
Adj. R Squared (%)	8.06	4.10	1.23	12.95
EPVRP coefficient	0.1263	0.0212	0.2108	-0.3446
	(0.46)	(0.05)	(0.49)	(-1.14)
Constant	-0.0079	0.0124	-0.0014	0.0021
	(-0.48)	(0.70)***	(0.09)	(0.14)
Adj. R Squared (%)	-2.27	0.24	-1.81	-0.37
N	43	43	43	43

*p < 0.1; **p < 0.05; ***p < 0.01

The table shows univariate OLS-regression results with the Newey-West standard errors for three-month excess (63 business days) returns of the European equity indexes. These excess returns are explained by volatility risk premium variables of each index. Corresponding NW-based t-statistics are presented in parentheses.

The findings support the hypothesis that the volatility risk premium would have explanatory power over short-term European equity index returns. The hypothesis of the return-forecasting nature of the European volatility risk premiums is also supported. Own-country-based volatility risk premiums explain a significant fraction of the European equity index returns and provide predictive power for return forecasting purposes.

Univariate models with pure S&P 500 volatility premiums are similar to the local evidence. Table IV shows that S&P 500-based *ex-post* volatility risk premium explains a substantial fraction of the broader European equity index (STOXX Europe 600) returns on one-month periods. The corresponding coefficients of other local market indexes are also significant at the 1% level. One percentage point increase in the global *ex-post* volatility risk premium increases the STOXX Europe 600 index one-month logarithmic excess returns at 1.36%, on average.

TABLE IV
MONTHLY REGRESSION RESULTS FOR SPX VARIABLES.

One-month returns					
	STOXX Europe 600	*Euro Stoxx 50*	*DAX*	*FTSE 100*	*SMI*
SPX VRP coefficient	0.4611	0.4416	0.3946	0.2802	0.3793
	(1.46)*	(1.41)*	(1.09)	(1.22)	(1.94)**
Constant	-0.0060	-0.0072	-0.0004	-0.0030	-0.0052
	(-0.85)	(-0.99)	(-0.06)	(-0.59)	(-1.06)
Adj. R Squared (%)	2.30	1.12	0.63	0.59	2.06
SPX EPVRP coefficient	1.3763	1.5901	1.6996	1.1391	0.8377
	(11.41)***	(10.12)***	(7.61)***	(11.17)***	(7.56)***
Constant	-0.0162	-0.0201	-0.0152	-0.0126	-0.0102
	(-5.39)***	(-5.52)***	(-3.45)***	(-4.75)***	(-3.35)***
Adj. R Squared (%)	39.62	36.93	37.65	32.47	19.66
N	129	130	129	129	130

*p < 0,1; **p < 0,05; ***p < 0,01

The table shows univariate OLS-regression results with the Newey-West standard errors for one-month (21 business days) excess returns of the European equity indexes. These excess returns are explained by the *ex-ante* (SPXVRP) and ex-post (SPXEPVRP) volatility risk premiums of S&P 500 index. Corresponding NW-based t-statistics are presented in parentheses.

By contrast, the results of the forecasting power of S&P 500-based *ex-ante* volatility risk premium are ambiguous. These results were confirmed by cutting the observed period in shorter sub-periods, calculating quarterly returns with monthly data, as well as by using the actual three-month implied volatilities without any major improvement in significance. Although the evidence for the predictive power of SPX VRP over European equity indexes is weak, the *ex-post* measure of volatility risk premium implies the existence of positive cross-market relation between innovations in S&P 500 volatility risk premium and European equity index returns. The results show that the ex-post S&P 500 volatility risk premium is consistently positively related to the corresponding European equity index returns, explaining significantly the short-term future return variability of the FTSE100 and SMI indices.

TABLE V
QUARTERLY REGRESSION RESULTS FOR SPX VARIABLES.

Three-month returns					
	STOXX Europe 600	*Euro Stoxx 50*	*DAX*	*FTSE 100*	*SMI*
SPX VRP coefficient	-0.1590	-0.4095	-0.4011	-0.3788	0.0596
	(-0.23)	(-0.55)	(-0.50)	(-0.71)	(0.11)
Constant	0.0016	0.0017	0.0202	0.0082	-0.0030
	(0.06)	(0.06)	(0.73)	(0.42)	(-0.14)
Adj. R Squared (%)	-2.29	-1.60	-1.77	-1.14	-2.40
SPX EPVRP coefficient	0.4031	0.3379	0.3615	0.6031	0.0970
	(1.06)	(1.06)	(0.78)	(2.28)**	(0.35)**
Constant	-0.0097	-0.0130	0.0052	-0.0113	-0.0039
	(-0.63)	(-0.78)	(-0.30)	(-1.06)	(-0.30)
Adj. R Squared (%)	0.17	-1.14	-1.17	5.32	-2.24
N	43	43	43	43	43

*p < 0,1; **p < 0,05; ***p < 0,01

The table shows univariate OLS-regression results with the Newey-West standard errors for three-month (63 business days) excess returns of the European equity indexes. These excess returns are explained by the *ex-ante* (SPXVRP) and ex-post (SPXEPVRP) volatility risk premiums of S&P 500 index. Corresponding NW-based t-statistics are presented in parentheses.

Local and global sources of volatility risk premiums consistently explain a non-trivial part of the European excess returns. The predictive power of the local VRP is stronger at quarterly return periods than monthly returns, whereas the global SPX VRP does not show any sign of *ex-ante* predictability of European returns (see Table V). This means that on quarterly basis, the local VRP seems to be more consistently predicting the subsequent European equity index returns than the global SPX VRP.

The main empirical results from the local part of the study are robust to the inclusion of traditional explanatory variables. Results from the multivariate controlled regressions shown in Table VI indicate that the monthly impact of local *ex-post* volatility risk premium (*EPVRP*) remain highly similar to the univariate results in all four indices. The local *ex-ante* volatility risk premium displayed significant forecasting power over subsequent equity index returns for Euro Stoxx 50 and DAX but not for FTSE100 and SMI on monthly basis. While corresponding univariate monthly regressions results (see table 2) showed no significant relationships between expected returns and *ex-ante* volatility risk premium (*VRP*) in all cases.

The local sources of volatility risk premium displayed significant forecasting power over subsequent equity index returns on quarterly horizons in isolation and remained relatively robust to the inclusion of control variables. The obtained coefficients from the controlled regressions de-

TABLE VI
MONTHLY MULTIVARIATE REGRESSION FOR VRP AND CONTROL VARIABLES

One-month returns				
	Euro Stoxx 50	*DAX*	*FTSE 100*	*SMI*
VRP	-0.5113	-0.6245	-0.1296	-0.3579
	(-2.46)**	(-2.09)**	(-0.64)	(-1.97)
IV	-1.5688	-1.6693	-1.1587	-1.6951
	(-4.91)***	(-5.74)***	(-4.12)***	(-4.50)***
ln(DIV)	0.0396	-0.0090	-0.1173	-0.0379
	(1.23)	(-0.20)	(-1.72)*	(-1.78)*
ln(PE)	0.0279	0.0391	0.01500	0.0059
	(1.30)	(2.81)***	(1.66)*	(0.48)*
DEFT	0.0181	0.0203	0.0312	0.0241
	(1.42)	(1.30)	(2.85)***	(2.31)**
TERM	0.0117	0.0056	0.0079	0.0062
	(2.11)	(1.02)	(1.39)	(1.36)
Constant	-0.0683	-0.0433	0.0816	0.0440
	(-0.83)	(-0.93)	(1.29)	(1.33)
Prob (F-statistic)	0.00	0.00	0.00	0.00
Adj. R Squared (%)	25.72	29.13	23.34	32.58
EPVRP	1.2543	1.2969	0.9983	0.5195
	(7.06)***	(5.08)***	(4.85)***	(2.90)***
IV	-0.4932	-0.6979	-0.6606	-1.1235
	(-1.56)	(-2.05)**	(-2.70)***	(-2.77)***
ln(DIV)	0.0079	-0.0540	-0.0959	-0.0254
	(0.32)	(-1.56)	(-1.89)*	(-1.37)
ln(PE)	0.0379	0.0209	0.0199	-0.0050
	(2.39)**	(1.99)**	(2.97)***	(-0.40)
DEFT	0.0041	0.0153	0.0215	0.0139
	(0.40)	(1.44)	(2.41)**	(1.48)
TERM	0.0007	0.0034	0.0012	0.0029
	(0.17)	(1.00)	(0.33)	(0.71)
Constant	-0.1067	-0.0068	0.0360	0.0563
	(-1.89)*	(-0.18)	(0.77)	(1.44)
Prob (F-statistic)	0.00	0.00	0.00	0.00
Adj. R Squared (%)	47.36	44.17	42.94	36.88
N	130	129	129	130

*p < 0.1; **p < 0.05; ***p < 0.01

The table presents multivariate OLS linear regression results with the Newey-West standard errors explaining one-month logarithmic excess returns of the European equity indexes. The independent variables are *ex-ante* volatility risk premium (VRP), *ex-post* volatility risk premium (EPVRP), monthly implied volatility (IV), log-dividend yield (ln(DIV)), log-price-to-earning-ratio (ln(PE)), default spread (DEFT), and term spread (TERM). Corresponding NW-based t-statistics are presented in parentheses.

TABLE VII
QUARTERLY MULTIVARIATE REGRESSION FOR VRP AND CONTROL VARIABLES

Three-month returns				
	Euro Stoxx 50	*DAX*	*FTSE 100*	*SMI*
VRP	0.7329	0.6525	0.3490	0.9170
	(1.97)**	(1.53)*	(0.88)	(2.18)**
IV	-1.6820	-1.6469	-1.4174	-1.8906
	(-4.73)***	(-3.53)***	(-2.87)***	(-4.05)***
ln(DIV)	0.0696	0.0208	-0.1977	-0.0429
	(0.92)	(0.16)	(-1.48)	(-1.18)
ln(PE)	0.0971	0.1118	0.0369	0.0492
	(1.66)	(3.02)***	(1.75)*	(1.65)
DEFT	0.0277	-0.0046	0.0486	0.0169
	(-1.59)	(-0.15)	(1.58)	(0.97)
TERM	0.0229	0.0071	0.0147	0.0177
	(2.14)**	(0.56)	(0.82)	(1.49)
Constant	-0.2541	-0.1409	0.1388	-0.0055
	(-1.10)	(-1.36)	(1.12)	(-0.08)
Prob (F-statistic)	0.00	0.00	0.00	0.00
Adj. R Squared (%)	48.55	46.92	39.46	58.18
EPVRP	0.1834	-0.1816	0.2067	-0.2138
	(0.40)	(-0.44)	(0.92)	(-0.53)
IV	-1.7017	-1.5196	-1.3851	-1.8538
	(-4.03)***	(-3.08)***	(-2.91)***	(-4.36)***
ln(DIV)	0.0663	0.0165	-0.1937	-0.0401
	(0.83)	(0.12)	(-1.56)	(-1.26)
ln(PE)	0.0953	0.1095	0.0341	0.0451
	(-1.56)	(2.75)***	(1.91)*	(1.35)
DEFT	0.0237	-0.0140	0.0439	0.0170
	(-1.22)	(-0.38)	(1.63)	(1.35)
TERM	0.0259	0.0120	0.0157	0.0255
	(2.09)**	(0.83)	(0.91)	(1.55)
Constant	-0.2261	-0.1119	0.1520	0.0063
	(-0.95)	(-0.98)	(1.20)	(0.07)
Prob (F-statistic)	0.00	0.00	0.00	0.00
Adj. R Squared (%)	45.83	45.12	39.06	51.59
N	43	43	43	43

*p < 0,1; **p < 0,05; ***p < 0,01

The table presents multivariate OLS linear regression results with the Newey-West standard errors explaining three-month logarithmic excess returns of the European equity indexes. The independent variables are *ex-ante* volatility risk premium (VRP), *ex-post* volatility risk premium (EPVRP), quarterly implied volatility (IV), log-dividend yield (ln(DIV)), log-price-to-earning-ratio (ln(PE)), default spread (DEFT), and term spread (TERM). Corresponding NW-based t-statistics are presented in parentheses.

creased, but the results remained consistent and statistically significant (see Table VII). Decrease in predictive return-impact of VRP might be subject to minor fading when other return-predictors are added into the same model. Similarly, Appendix II & III demonstrates that return-impacts of S&P 500 *ex-post* measure remained sufficiently unaffected when tested along with other controlling variables. The empirical findings of volatility risk premiums remain robust to the controlled effects for both, local and global measures.

V. CONCLUSIONS

The results show that volatility risk premiums are able to explain a non-trivial fraction of the equity index return and that volatility risk is systematically priced into the European equity index returns locally and globally. Our findings of the explanatory power of volatility risk premiums on aggregate stock market returns are in line with the results from the U.S. markets, for example, by [3] and [11].

The negative sign of variance premium means that variance buyers are willing to accept negative returns to hedge against the volatility risk. As the volatility risk premium, on average, is positive for all the examined indexes, volatility selling over European equity indexes has been consistently profitable over the sample period. By contrast, buying volatility would have been unprofitable. Our results show that this negative premium is related to the equity risk premium and explains a relatively large portion of European equity index excess returns.

All European equity index monthly returns are positively related to the *ex-post* volatility risk premium so that one percentage point increase in EPVRP has on average resulted in increase of 1.28% *p.m.* for equity index excess returns. For quarterly return predictions, ex-ante risk premiums are better than ex-post volatility risk premiums. On average, one percentage point increase in quarterly volatility risk pre-

mium leads to 1.25% *per quarter* increase in subsequent quarterly European equity index excess returns.

The S&P 500-based quarterly *ex-ante* and *ex-post* volatility risk premiums did not show a significant forecast ability of subsequent European equity index excess returns. However, the corresponding monthly volatility risk premiums are found to be significantly related to all the European equity index returns. On average, one percentage point increase in the S&P 500 volatility risk premium leads to 1.36% *p.m.* increase in excess returns of European stock indexes in univariate settings. The S&P 500-based monthly measure of the *ex-post* volatility risk premium provides slightly stronger (by 8 basis points) predictive power for European equity index returns than the corresponding local measures. These findings are consistent with [23], who showed an increasing inter-market linkage between US and developed European markets during post-financial crisis period. Overall, ex-post volatility risk premiums show better explanatory power (Adj. R Squared) than ex-ante variants.

This study contributes to the existing literature in two ways. Firstly, the finding of the positive relationship between the volatility risk premium and European equity index excess returns is significant, since this is the first time the phenomenon is addressed at a market-wide level in the European stock markets. Our results are in line with the previous findings from the U.S. markets and complement the existing literature in this respect. Investors demand compensation for bearing volatility risk, and a part of equity risk premium can be explained by this risk-aversion-related information implicit in option prices. The local European forward-looking volatility risk premium provides forecasting power on subsequent quarterly equity index excess returns.

Secondly, the finding of globally priced volatility and skewness preferences is important in understanding risks that are priced in equity index returns. The risks in aggregate stock market volatility and skewness of the S&P 500 index seem to be important in Europe as well. Particularly, the risk-neutral implied volatility that is captured by the S&P 500 index options and further by the VIX index, provides a useful tool when assessing the risks embedded globally in the equity markets. Risk aversion captured by the S&P 500-based volatility risk premium exhibits a substantial return-explanatory power across markets and displays as a useful measure for understanding the risks that are relevant in the aggregate European equity market.

To better understand driving forces of the positive relation between the volatility risk premium and expected equity index returns, it might be worthwhile to extend this study to analyze European equity market's exposure to jump risks. Forming a specific measure of tail risk would enrich the knowledge of risk-return tradeoffs in the European equity markets.

References

[1] Ang, A., Hodrick, R., Xing, Y., Zhang, X., 2006. The Cross-Section of Volatility and Expected Returns, The Journal of Finance 61, 259–299.

[2] Bakshi, G. Madan, D., 2006. A Theory of Volatility Spreads. Management Science 52, 1–12.

[3] Bakshi, G., Kapadia, N., 2003. Delta Hedged Gains and the Negative Market Volatility Risk Premium. The Review of Financial Studies 16, 527–566.

[4] Bates, D., 1991. The Crash of '87: Was It Expected? The Evidence from Option Markets. The Journal of Finance 46, 1009–1044.

[5] Black, F., Scholes, M., 1973. The Pricing of Options and Corporate Liabilities. Journal of Political Economy 81, 637–654.

[6] Blair, B., Poon, S., Taylor, S., 2001. Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High Frequency Index Returns. Journal of Econometrics 105, 5–27.

[7] Bollerslev, T., Gibson, M., Zhou, H., 2011. Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. Journal of Econometrics 160, 235–245.

[8] Bollerslev, T., Tauchen, G., Zhou, H., 2009. Expected Stock Returns and Variance Risk Premia. Review of Financial Studies 22, 4463–4492.

[9] Breeden, D., Litzenberger, R., 1978. Prices of Stage-Contingent Claims Implicit in Option Prices. The Journal of Business 51, 621–651.

[10] Britten-Jones, M., Neuberger, A., 2000. Option Prices, Implied Price Processes, and Stochastic Volatility. The Journal of Finance 55, 839–866.

[11] Carr, P., Wu, L., 2008. Variance Risk Premiums. Review of Financial Studies 22, 1311–1341.

[12] Chang, B., Christoffersen, P., Jacobs, K., 2013. Market Skewness Risk and the Cross-Section of Stock Returns. Journal of Financial Economics 107, 46–68.

[13] Cremers, M., Halling, M., Weinbaum, D., 2015. Aggregate Jump and Volatility Risk in the Cross-Section of Stock Returns. The Journal of Finance 70, 577–614.

[14] Day, T., Lewis, C., 1992. Stock Market Volatility and the Information Content of Stock Index Options. Journal of Econometrics 52, 267–287.

[15] Drechsler, I., Yaron, A., 2010. Tails, Fears and Risk Premia. The Review of Financial Studies 24, 1–45.

[16] Elyasiani, E., Gambarelli, L., Muzzioli S., 2016. Moment Risk Premia and Cross-Section of Stock Returns. DEMB Working Paper Series 103.

[17] Han, B., Zhou, Y., 2012. Variance Risk Premium and Cross-Section of Stock Returns. Working Paper. University of Texas Austin.

[18] Jackwerth, J., 2004. Option-Implied Risk-Neutral Distributions and Risk Aversion. The Research Foundation of AIMR.

[19] Jiang, G., Tian, Y., 2005. The Model-Free Implied Volatility and its Information Content. The Review of Financial Studies 18, 1305–1342.

[20] Kozhan, R., Neuberger, A., Schneider, P., 2013. The Skew Risk Premium in the Equity Index Market. The Review of Financial Studies 26, 2174–2203.

[21] Newey, W., West, K., 1987. Hypothesis Testing with Efficient Methods of Moments Estimation. International Economic Review 28, 777–787.

[22] Poon, S., 2005. A Practical Guide to Forecasting Financial Market Volatility. Chichester, England: John Wiley & Sons.

[23] Pätäri, E., Ahmed, S., John, E. & Karell, V. 2019. The changing role of emerging and frontier markets in global portfolio diversification, Cogent Economics & Finance, 7:1,

[24] Santa-Clara, P., Yan, S., 2010. Crashes, Volatility, and the Equity Premium: Lessons from S&P 500 Options. Review of Economics and Statistics 92, 435–451.

[25] Siriopolous, C., Fassas, A., 2009. Implied Volatility Indexes – A Review. Working paper. University of Patras.

White papers:

[26] CBOE VIX White Paper (2017). https://cdn.cboe.com/resources/vix/vixwhite.pdf

[27] STOXX Strategy Index Guide (2021) https://www.stoxx.com/document/Indices/Common/Indexguide/stoxx_strategy_guide.pdf

Appendices

Appendix I
Variables and Abbreviations Used in the Empirical Section

Class	*Variable*	*Abbreviation*
Equity indexes	*Euro Stoxx 50 index*	*ESTOXX*
	DAX index	*DAX*
	FTSE 100 index	*FTSE*
	SMI index	*SMI*
	STOXX Europe 600 index	*STOXX*
	S&P 500 index	*SPX*
Volatility- and skew indexes	*Euro Stoxx 50 implied volatility index*	*VSTOXX*
	DAX implied volatility index	*VIX-NEW*
	FTSE 100 implied volatility index	*VFTSE*
	SMI implied volatility index	*VSMI*
	S&P 500 implied volatility index	*VIX*
	S&P 500 implied skew index	*SKEW*
Moment risk premiums	*Ex ante volatility risk premium*	*VRP*
	Ex post volatility risk premium	*EPVRP*
	Ex ante S&P 500 volatility risk premium	*SPX VRP*
	Ex post S&P 500 volatility risk premium	*SPX EPVRP*
	Ex ante skew risk premium	*SRP*
	Ex post skewrisk premium	*EPSRP*
Control variables	*Implied volatility*	*IV*
	S&P 500 implied volatility	*SPX IV*
	S&P 500 implied skewness	*IS*
	Dividend yield	*DIV*
	Price-to-earnings-ratio	*PE*
	Default spread	*DEFT*
	Term spread	*TERM*

Appendix II
monthly Multivariate Regression for S&P 500 Volatility Risk Premiums and Control Variables

One-month returns

	STOXX Europe 600	*Euro Stoxx 50*	*DAX*	*FTSE 100*	*SMI*
SPX VRP	-0.1362	-0.1061	-0.2200	-0.1900	-0.1198
	(-0.54)	(-0.39)	(-0.69)	(-0.76)	(-0.66)
SPX IV	-0.8509	-1.0112	-1.0721	-0.9379	-1.0149
	(-3.44)***	(-3.05)***	(-4.04)***	(-3.63)***	(-2.86)***
ln(DIV)	-0.0255	0.0277	0.0001	-0.1132	-0.0287
	(-0.54)	(0.86)	(0.00)	(-1.73)*	(-1.12)
ln(PE)	0.0194	0.0236	0.0232	0.0144	0.0008
	(1.39)	(1.08)	(1.67)*	(1.76)*	(0.07)
DEFT	0.0117	0.0090	0.0090	0.0296	0.0166
	(0.92)	(0.81)	(0.59)	(2.81)***	(1.26)
TERM	0.0116	0.0089	0.0080	0.0099	0.0122
	(2.47)	(1.57)	(1.25)	(1.80)*	(2.64)***
Constant	-0.0174	-0.0711	0.0293	0.0704	0.0304
	(-0.34)	(-0.87)	(-0.61)	(1.17)	(0.98)
Prob (F-statistic)	0.00	0.00	0.00	0.00	0.00
Adj. R Squared (%)	25.15	17.55	17.02	19.98	23.00
SPX EPVRP	1.1141	1.4370	1.5498	0.9412	0.5195
	(5.32)***	(6.44)***	(5.07)***	(7.06)***	(1.44)***
SPX IV	-0.5132	-0.4132	-0.5165	-0.5628	-1.1235
	(-2.45)**	(1.59)	(2.15)**	(2.81)***	(2.77)***
ln(DIV)	-0.0437	0.0077	-0.0526	-0.0850	-0.0254
	(-1.28)	(0.28)	(-1.48)	(-1.62)	(-1.37)
ln(PE)	0.0159	0.0349	0.0131	0.0155	-0.0050
	(1.40)	(2.07)**	(1.25)	(2.12)**	(-0.39)
DEFT	0.0111	0.0049	0.0145	0.0215	0.0139
	(1.30)	(0.61)	(1.37)	(2.50)**	(1.48)
TERM	0.0027	-0.0034	-0.0006	0.0014	0.0029
	(0.89)	(-1.00)	(-0.16)	(0.40)	(0.71)
Constant	-0.0083	-0.1088	-0.0009	0.0273	0.0563
	(-0.20)	(-1.73)*	(-0.02)	(0.61)	(1.44)
Prob (F-statistic)	0.00	0.00	0.00	0.00	0.00
Adj. R Squared (%)	46.49	42.20	41.88	37.85	36.88
N	130	129	130	129	130

*p < 0.1; **p < 0.05; ***p < 0.01

Appendix III
quarterly Multivariate Regression for S&P 500 Volatility Risk Premiums and Control Variables

Three-month returns

	STOXX Europe 600	*Euro Stoxx 50*	*DAX*	*FTSE 100*	*SMI*
SPX VRP	-1.1846	-1.2448	-1.2704	-1.3778	-0.7284
	(-3.16)***	(-2.49)**	(-2.42)**	(-4.33)***	(1.27)***
SPX IV	-0.6786	-1.0593	-1.1081	-0.9183	-0.9267
	(-1.49)	(-1.42)	(-1.44)	(-2.89)***	(-3.48)***
ln(DIV)	-0.1551	0.0530	0.0753	-0.2306	-0.0457
	(-1.46)	(0.56)	(-0.87)	(-2.85)***	(-1.55)
ln(PE)	0.0788	0.0588	0.0690	0.0360	0.0202
	(1.68)	(0.85)	(-0.96)	(2.04)	(0.51)
DEFT	0.0098	-0.0019	0.0107	0.0400	0.0045
	(0.45)	(-0.06)	(-0.40)	(2.21)**	(0.22)
TERM	0.0446	0.0408	0.0305	0.0384	0.0429
	(4.60)***	(3.19)***	(2.66)**	(3.42)***	(3.60)
Constant	-0.0106	-0.1264	-0.2004	0.1717	0.0421
	(-0.08)	(-0.47)	(-0.73)	(1.77)*	(0.41)
Prob (F-statistic)	0.00	0.00	0.00	0.00	0.00
Adj. R Squared (%)	58.94	44.73	46.85	53.26	44.36
SPX EPVRP	-0.1985	-0.0920	-0.1937	0.2447	-0.3977
	(-0.71)	(-0.25)	(-0.54)	(1.48)	(-1.88)*
SPX IV	-0.7897	-1.1786	-1.3094	-1.0735	-1.0164
	(-2.03)*	(-1.79)***	(-2.09)**	(-3.04)***	(-3.27)***
ln(DIV)	-0.1044	0.0783	-0.0273	-0.1418	-0.0281
	(0.90)	(0.89)	(-0.23)	(-1.29)	(-0.87)
ln(PE)	0.0956	0.0813	0.0960	0.0270	0.0380
	(1.97)*	(1.17)	(2.26)**	(1.70)*	(1.09)
DEFT	0.0053	0.0023	0.0008	0.0326	0.0012
	(0.25)	(0.09)	(0.02)	(1.45)	(0.07)
TERM	0.0332	0.0283	0.0248	0.0220	0.0430
	(3.23)***	(2.23)*	(1.86)*	(1.72)*	(2.56)**
Constant	-0.1098	-0.2274	-0.1169	0.1009	-0.0202
	(-0.77)	(-0.87)	(-0.95)	(0.95)	(-0.23)
Prob (F-statistic)	0.00	0.00	0.00	0.00	0.00
Adj. R Squared (%)	52.77	38.70	41.94	40.46	43.72
N	43	43	43	43	43

*p < 0,1; **p < 0,05; ***p < 0,01

KNOWCON
Knowledge on Economics and Management
International Scientific Conference

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 29–35

DOI: 10.15439/2021B5
ISSN 2300-5963 ACSIS, Vol. 29

Using the generalized fuzzy k-nearest neighbor classifier for biomass feedstocks classification

Mahinda Mailagaha Kumbure, Pasi Luukka
LUT University
Yliopistonkatu 34, 53850 Lappeenranta, Finland
Email: {mahinda.mailagaha.kumbure, pasi.luukka}@lut.fi

***Abstract*—This paper proposes a novel framework based on a recently introduced classifier called multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) and the Minkowski distance to classify biomass feedstocks into property-based classes. The proposed approach uses k nearest neighbors from each class to compute class-representative multi-local power mean vectors and the Minkowski distance instead of the Euclidean distance to fit the most suitable distance metric based on the properties of the data in finding the nearest neighbors to the new data point. We evaluate the performance of the proposed approach using three biomass datasets collected from several articles published in reputable journals and the Phyllis 2 biomass database. Input features of the biomass samples include their characteristics from the proximate analysis and ultimate analysis. In the developed framework, we interpret the biomass feedstocks classification as a five-class problem, and the classification performance of the proposed approach is benchmarked with the results obtained from classical k-nearest neighbor-, fuzzy k-nearest neighbor- and support vector machine classifiers. Experimental results show that the proposed approach outperforms the benchmarks and verify its effectiveness as a suitable tool for biomass classification problems. It is also evident from the results that the features from both ultimate and proximate analyses can offer a better classification of biomass feedstocks than the features considered from each of those analyses separately.**

***Index Terms*—Biomass feedstocks, Fuzzy k-nearest neighbor, Machine learning, Minkowski distance, Proximate properties, Ultimate properties**

I. Introduction

BIOMASS is a biological material obtained from living organisms such as animals and plants. Biomass feedstocks are diverse, usually derived from agricultural residues, forest products waste, food waste, green waste, municipal solid waste, and other waste [1]. Due to its organic nature and abundant supply, biomass is considered as an essential renewable energy source [2] and has received much attention in the world [1]. Biomass is typically used to derive various energy products, for example, biogas, bioethanol, biodiesel, and solid fuel [3]. Following oil, coal, and natural gas, biomass has been the fourth largest energy source globally to date [4].

Primary concerns regarding biomass investigations include enhancing and extending the general understanding of the biomass properties and compositions, and also using this knowledge for achieving sustainable development in energy generation [5]. In the study of biomass, in general, two different types of analyses: proximate analysis and ultimate analysis, are used to determine the nature of biomass in terms of the chemical compounds [6]. The proximate analysis is applied to measure the compositions of volatile matter, moisture, ash and fixed carbon in the biomass. On the basis of ash and moisture content, ultimate analysis yields the amount of carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S) [6]. These properties and their classification corresponding to the various biomass materials are considered more important when they are selected as energy feedstocks [5]. The energy conversion process has also encouraged the studies for biomass feedstocks classification considering their properties such as proximate properties, thermal properties, chemical properties, to mention few [7].

Artificial intelligence, particularly machine learning (ML), has been extensively used to analyze various types of data classification and prediction problems effectively. However, applying ML-based techniques in biomass analysis is still a new development [8]. In the literature, a few studies have focused on the potential of some ML techniques for biomass classification and related research. Tao et al. [9] used a principal component analysis (PCA) based approach to attribute the biomass properties within five groups. Wang et al. [10] also applied the PCA to find the most influential features of biomass for the decision-making process in bioenergy production. Olatunji et al. [5] attempted to grade the biomass feedstocks based on their proximate properties using k-nearest neighbor (KNN) method. The best performance they found with the KNN model [11] was around 70% in the training and validation. A recent study by [8] examined the effectiveness of several ML techniques, including Random Forest, KNN, Gaussian Naïve Bayes, and Decision Tree models to predict and differentiate biomass types based on the Pyrolysis molecular beam mass spectrometry (py-MBMS) analyses. They showed that the KNN classifier generally performed the best compared to others. The present work introduces a novel ML-based approach for biomass classification by interpreting the classification task as a five-class problem.

Our proposed approach is based on the multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) method that is an enhanced version of the KNN classifier, which was recently introduced in [12]. This new KNN method is chosen as it has showed more robust to outliers and random variables than original ones according to [12]. This technique can perform well in situations where clear imbalances in class distributions of the data are found [12]. In this study, we generalize the

performance of the MLPM-FKNN classifier using k nearest neighbors from each class to compute class-representative multi-local power mean vectors. In addition to that, we also introduce the Minkowski distance for the k nearest neighbor search in the learning part instead of the Euclidean distance to fit the most suitable distance metric according to the data properties in finding the nearest neighbors for the unclassified data point from each class. Since the Minkowski distance is a generalized distance of the Euclidean and Manhattan distances, its utilization allows greater flexibility for obtaining more relevant neighboring points close to the unclassified data point.

To examine the classification performance of the proposed approach, we use three biomass datasets collected from several articles [7], [13], [14], [15], [16] and the Phyllis 2 biomass database [17]. Four well-known performance measures are used to assess the performance of the proposed method, and the observed results are benchmarked with three state-of-art techniques such as the KNN, fuzzy k-nearest neighbor (FKNN) [18], and support vector machine (SVM) [19] classifiers. From the wide variety of machine learning techniques [20], [21], these were chosen since they are similar to proposed method and easily available. In summary, the main contributions of this paper include (i) proposing a generalized MLPM-FKNN classifier with Minkowski distance for biomass classification, (ii) using chemical compound features derived from ultimate analysis for biomass classification, and empirically examining whether they have a great influence on the classification of biomass, (iii) applying biomass data from Phyllis 2 data repository for classification purpose, and (iv) comparing the classification performance of the proposed intelligent model with the performance of several well-known ML techniques.

II. Preliminaries

This section briefly presents the preliminaries of relevant k-nearest neighbor classifier variants, the Power mean operator, and the Minkowski distance measure. In addition, the Minkowski distance-based generalized MLPM-FKNN classifier is introduced.

A. KNN and FKNN Classifiers

The KNN classifier [11] is a simple, effective, and robust supervised machine learning technique. Due to its high accuracy and capability in the pattern classification, the KNN classifier has been widely used in many real-world applications (for examples, see [22], [23]). It begins with calculating the Euclidean distances from the query sample (i.e., unclassified data point) to the training instances. Then, a set of k nearest neighbors is identified for the query sample from the sorted training instances in ascending order according to the Euclidean distances measured. Finally, the query sample is assigned to the class represented by the majority of the nearest neighbors. However, the KNN method intuitively suffers from some weaknesses. For instance, it gives equal importance to all nearest neighbors neglecting the fact that different instances have different impacts on the classification of the query sample [24]. Moreover, it does not take into account the strength of the class membership for the query sample [25]. To deal with these issues, the FKNN model [18] has been introduced as an enhancement of the original algorithm.

In the FKNN, the set of k nearest neighbors of the query sample (Q) is searched first as in the KNN classifier. After that, a membership degree for each class is measured for the query sample using weighted distances from k nearest neighbors to the query sample. Lastly, it classifies the query sample into the class with the highest membership degree among all classes. To compute the class memberships (u_i for class i) for Q, the formula used can be defined as follows:

$$u_i(Q) = \frac{\sum_{j=1}^{k} u_{ij}(1/\|Q - X_j\|^{2/(r-1)})}{\sum_{j=1}^{k}(1/\|Q - X_j\|^{2/(r-1)})} \tag{1}$$

where, $r \in (1, +\infty)$ is a fuzzy strength parameter and u_{ij} is the membership degree of the j^{th} nearest neighbor X_j from the i^{th} class. Also,

To compute u_{ij}, there are two main approaches: one is through the crisp membership, and the other is through the fuzzy membership [18]. In this study, we use the crisp labeling approach where the full membership is assigned to the known class and zero memberships to all other classes.

B. Power Mean and Minkowski Distance

Power mean (also called generalized mean) is a function of means. If $\{x_1, x_2, .., x_m\}$ is a set of real numbers and p is a real parameter, then power mean (M_p) is defined as:

$$M_p = (\frac{1}{m}\sum_{l=1}^{m} x_l^p)^{1/p} \text{ for } p \neq 0 \tag{2}$$

When $p \to 0$, $M_p \to \prod_{i=1}^{m} X_i^{1/m}$. With the power mean function, different types of means can be generated including well-known harmonic mean ($p = -1$), arithmetic mean ($p = 1$), and quadratic mean ($p = 2$). Additionally, M_p approaches to geometric mean when $p \to 0$.

The Minkowski distance (also referred to as L_p norm) between two data points $X = \{x_1, x_2, .., x_m\}$ and $Y = \{y_1, y_2, .., y_m\}$ in m-dimensional space is defined as follows:

$$d_{Md}(X, Y) = \big(\sum_{l=1}^{m} |x_l - y_l|^q\big)^{1/q} \text{ for } q \geq 1 \tag{3}$$

The Minkowski distance represents a class of distance functions that are formed by the parameter q. For instance, by setting $q = 1$, we obtain the Manhattan distance (also called City block distance). Similarly, the Euclidean distance is observed in the case of $q = 2$.

C. Modified MLPM-FKNN Classifier

The concept of the multi-local power mean fuzzy k-nearest neighbor (MLPM-FKNN) classifier is easy to understand. It has been developed by introducing a local-mean computation into the learning part of the FKNN method. The local mean

vectors are calculated for each class in the set of nearest neighbors by using the power mean function. These vectors are called multi-local power mean vectors. In this way, the MLPM-FKNN method creates "representative vectors" for each class to perceive the class information for query sample instead of comparing it directly to the k-nearest neighbors. Also, changing the power mean parameter allows us to find its best possible options, which will enhance the classification accuracy [12].

In this study, we deploy a generalized version of this method. The Minkowski distance function is applied according to the study by [5] instead of the Euclidean distance to measure the distances from the query sample to the training instances. The purpose of using Minkowski distance here is to generate greater flexibility for obtaining more relevant neighbors close to the query sample since it has an optimizable parameter to adjust the function to the data set available. A formal definition of the developed method can be presented as follows.

Let $\{X_j, c_j\}_{j=1}^{n}$ be a training set with n instances, where $X_j = \{x_j^1, x_j^2, .., x_j^m\}$ is an input instance j from m-dimensional feature space, and its output class label is $c_j \in C$ ($C = \{\omega_1, \omega_2, .., \omega_T\}$: the set of class labels and T is the number of classes). For a given query sample $Q = \{q^1, q^2, .., q^m\}$, the goal is to fit the classifier from the training set in order to predict the class ω^* for Q. The steps of the generalized MLPM-FKNN classifier in this study can be presented as follows:

(i) Group the training data $\{X_j, c_j\}_{j=1}^{n}$ into each class ω_i. The resulting class subsets can be denoted as $\{X_j, \omega_i\}_{j=1}^{n_i}$ for $i = 1, 2, .., T$. Here n_i is the number of instances in class ω_i.

(ii) Find the sets of k nearest neighbors of Q from each class ω_i. In this case, the Minkowski distances are calculated from the training instances in $\{X_j, \omega_i\}_{j=1}^{n_i}$ to Q and the set of k nearest neighbors are identified from the reordered training instances according to the increasing distances.

(iii) For each set of k nearest neighbors $\{X_j^{nn}\}_{j=1}^{k}$ from each class ω_i (nn means nearest neighbor), power mean vectors M_i ($i = 1, 2, .., T$) are measured and which are called multi-local power mean vectors.

$$M_i = \Big(\frac{1}{k}\sum_{j=1}^{k}(X_j^{nn})^p\Big)^{1/p} \text{ for } p \neq 0 \tag{4}$$

(iv) Compute the Minkowski distances from Q to $M_i = \{\tilde{m}_1^i, \ldots, \tilde{m}_m^i\}$ for $i = 1, 2, .., T$ such as:

$$d_{Md}(Q, M_i) = \Big(\sum_{l=1}^{m}|q^l - \tilde{m}_i^l|^q\Big)^{1/q} \tag{5}$$

(v) Compute the memberships to $\{\omega_i\}_{i=1}^{T}$ according to Eq. (1) using the distances from Step (iv) and the crisp approach for calculating u_{ij} (i.e., $u_{ij} = 1$ for the known class and $u_{ij} = 0$ for other classes).

Algorithm 1 Updated MLPM-FKNN classifier

Input: $\{X_j, c_j\}_{j=1}^{n}$, k, p, q, Q
Output: ω^*
START
for $i \leftarrow 1$ to T **do**
  for $j \leftarrow 1$ to n_i **do**
    Compute $d_{Md}(Q, X_j) \leftarrow \Big(\sum_{l=1}^{m}|q^l - x_j^l|^q\Big)^{1/q}$
  end for
  Sort $\{d_{Md}(Q, X_j)\}_{j=1}^{n_j}$ in ascending order
  if ($n_j < k$) **then**
    $k \leftarrow n_j$
  end if
  Find $\{X_j^{nn}\}_{j=1}^{k}$
  Find $M_i \leftarrow (\frac{1}{k}\sum_{j=1}^{k}(X_j^{nn})^p)^{1/p}$
end for
for $i \leftarrow 1$ to T **do**
  Compute $d_{Md}(Q, M_i) \leftarrow \Big(\sum_{l=1}^{m}|q^l - \tilde{m}_i^l|^q\Big)^{1/q}$
  Compute $u_i(Q) \leftarrow \frac{\sum_{j=1}^{T} u_{ij}(1/d_{Md}(Q,M_i)^{2/(r-1)})}{\sum_{j=1}^{T}(1/d_{Md}(Q,M_i)^{2/(r-1)})}$
end for
return ω^* such that

$$\omega^* = \arg\max_{\omega_i} u_i(Q)$$

(vi) Classify Q into the class ω^* that has the highest membership degree. In other words:

$$\omega^* = \arg\max_{\omega_i} u_i(Q) \tag{6}$$

This method generates class-representative power mean vectors using k nearest neighbors obtained from each class subset instead of the entire training dataset. This distinguishes the proposed method from the original MLPM-FKNN algorithm. Moreover, utilizing the Minkowski distance metric to measure the distances from the query sample to the training instances allows the classifier to choose the most suitable distance metric based on the properties of the data. In the developed framework, we also examine the performance of the updated MLPM-FKNN classifier based on the Euclidean distance, which is denoted as MLPM-FKNN (E). At the same time, the Minkowski distance-based generalized approach is shown as MLPM-FKNN (M).

III. Data and Experimental Setting

A. Data Description

In this study, we used three datasets of biomass feedstocks, two of them were generated from several articles [7], [13], [14], [15], [16] published in respective journals, and other one was collected from the Phyllis 2 biomass data repository [17]. Information and the properties of each of the datasets are summarized in Table I. It is noteworthy to mention that these datasets are based on experimental outcomes of the proximate and ultimate analyses of biomass produced by previous studies. We attempt to use them for classification purposes in this study.

TABLE I: Properties of the data used

Data	Source	# Instances	# Features	# Classes
Dataset 1	[13], [14]	212	4	5
Dataset 2	[7], [15], [16]	135	5	5
Dataset 3	[17]	344	9	5

In these datasets, we included five classes of biomass feedstocks considering the property-based definitions in [5]. In particular, class 1 contained energy grasses and their parts (fiber materials, leaves), and class 2 comprised fruit residues and relevant sources (shells, seeds, pit). For class 3, materials from wood, wood chips, chips-barks, pruning were considered, while food crop residues (straws, stalks, dust, husk, hull, cob) were set for class 4. Class 5 included other waste materials such as milling industry waste, refuse, and municipal solid waste. Fig. 1 illustrates the percentages of the classes included in each dataset.

According to Fig. 1, it is clear that there are imbalances of the classes in each of datasets. Among them, class 3 is the most frequent class in all datasets, even though it does not account for over 50% of each dataset. In contrast, class 4 and 5 in dataset 1, and class 2 and 5 in dataset 2, and class 5 in dataset 3 are associated with a small number of biomass samples. The features considered in dataset 1 were fixed carbon (FC), volatile matter (VM), ash, and higher heating value (HHV) that had been extracted from the proximate analysis. In dataset 2, the features were the chemical properties of the biomass substances from the ultimate analysis, such as carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulfur (S). For dataset 3, all feature types included in both dataset 1 and dataset 2 were considered. In all cases, we assumed that these features have significant influences on the class variable. Notice that dataset 2 and dataset 3 have not been used earlier for classification purposes, and this paper is the first one showing classification results for them. In particular, we utilize biomass data instances for the Phyllis 2 database for machine learning-based classification.

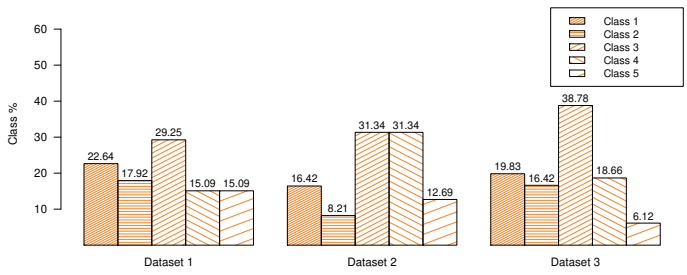


Fig. 1: Distribution (%) of each class in dataset 1, dataset 2, and dataset 3.

B. Testing methodology

The proposed framework for biomass classification has two main phases: i) training and validation and ii) testing. In the training and validation step, the model was developed by optimizing values for parameters k (number of nearest neighbors), p (power mean parameter), and q (Minkowski distance parameter). A grid search technique was deployed to optimize the model parameters. The performance of the classification models with optimal parameters were evaluated in the testing phase. To compare the performance of the generalized MLPM-FKNN classifier, we applied three well-known methods, namely k-nearest neighbor (KNN) [11], fuzzy k-nearest neighbor (FKNN) [18] and support vector machine (SVM) [19] classifiers. In addition to them, the MLPM-FKNN classifier based on the Euclidean distance [i.e., MLPM-FKNN (E)] was also applied, and the results were compared.

The analysis started with normalizing all features in the data into the unit interval. Next, datasets were randomly split into 60% for training, 20% for validation and 20% for testing. Stratified random sampling method was applied to ensure that all instances have the same proportions of units representing the different classes present as the whole data set. The holdout technique [26] was adopted for cross-validation, where the training and validation datasets were randomly generated 20 times. In the parameter settings, the number of nearest neighbors k was selected from $\{1, 2, ..., 15\}$ for all nearest neighbor methods. The value for p in power mean was chosen from the range $\{1, 1.1, .., 5\}$. The values from $\{1, 1.5, ..., 5\}$ were selected for the parameter q of the Minkowski distance. The fuzzy strength parameter $r = 2$ was kept, as in [12], [25] for MLPM-FKNN (M), MLPM-FKNN (E), and FKNN classifiers. Radial basis function kernel was used with the SVM model. To measure the classification performance, accuracy was used as the primary evaluation metric. Additional performance measures such as sensitivity and specificity were also measured as displaying classification results with accuracy alone is often not enough to adequately emphasize the effectiveness of the applied method [12]. The formulas used for sensitivity and specificity, especially to multi-class problems can be found from [25]. Additionally, the standard deviation (STD) of the accuracies was also computed. Based on the resulting confusion matrixes, we further examined the results of each classifier in the classification of biomass samples into each class.

IV. Results and Discussion

This section first presents the results from the training & validation phase of our methodology. Then the classification results in the test phase are presented.

A. Classification results with the training and validation data

We collected the accuracy, sensitivity, and specificity values in each run during the training and validation and averaged them for all repetitions from the holdout process. When the mean accuracy reached the maximum, the optimal values for the parameters (p, q and k) were observed. Table II

TABLE II: Classification performance with the validation data

Model	Measure	Dataset 1	Dataset 2	Dataset 3
MLPM-FKNN (Minkowski)	Accuracy	**0.5000**	**0.6217**	**0.7815**
	Sensitivity	0.4775	0.5208	0.7435
	Specificity	0.8697	0.8973	0.9447
	STD	0.0707	0.0761	0.0722
	Op. k, p, q	{9, 1.7, 1}	{2, 5, 3}	{3, 1, 1.5}
MLPM-FKNN (Euclidean)	Accuracy	0.4824	0.6152	0.7667
	Sensitivity	0.4558	0.5252	0.7175
	Specificity	0.8619	0.8968	0.9410
	STD	0.0676	0.0737	0.0636
	Op. k, p	{15, 2}	{2, 4.1}	{3, 1.4}
KNN	Accuracy	0.4588	0.5804	0.7370
	Sensitivity	0.4402	0.5402	0.6557
	Specificity	0.8582	0.8892	0.9317
	STD	0.0736	0.1183	0.0546
	Op. k	7	3	5
FKNN	Accuracy	0.4471	0.5804	0.7704
	Sensitivity	0.4313	0.5173	0.6839
	Specificity	0.8550	0.8866	0.9398
	STD	0.0676	0.0928	0.0500
	Op. k	15	11	6
SVM	Accuracy	0.4029	0.5348	0.7704
	Sensitivity	0.3600	0.3848	0.7056
	Specificity	0.8413	0.8684	0.9423
	STD	0.0312	0.0632	0.0211

summarizes those maximum performance measures and corresponding parameter values ("Op.") obtained with the proposed approach and the benchmarks with each dataset. To assess the reliability of the achieved mean accuracy value, its standard deviation ("STD") is also reported.

According to Table II results, we can see that the MLPM-FKNN (M) classifier achieves better results than the benchmarks in the training & validation for all datasets. It also has a reasonable standard deviation of accuracy and explicit support from mean sensitivity and specificity values. Moreover, used classifiers give outstanding performance with dataset 3 among all datasets while the proposed approach performs the best, achieving an accuracy of 78.15%. It is also apparent that the mean accuracy of all classifiers with dataset 2 is comparatively high compared with dataset 1, even though the sample size of dataset 2 is relatively small. This implies that the chemical properties of the biomass from the ultimate analysis offer great support than the proximate properties for their classifications, and having features from both analyses may provide even better results. Moreover, despite the influence of the class imbalance (as shown in Fig. 1) and the class overlapping issues [27], having a small number of instances in dataset 1, and dataset 2 might also have caused all classifiers to yield a relatively low performance.

Looking at the optimal values of the model parameters, a low value of k has yielded better results for MLPM-FKNN (M) than for the KNN and FKNN methods, which is surprising. This indicates that when the class-representative power mean vectors are computed using the k nearest neighbor from each class, it does not necessarily need to have more instances to make local power mean vectors more robust (and representative). It also can be seen that $p \in \{1.7, 5, 1\}$ and $q \in \{1, 3, 5\}$ have produced the maximum accuracy with the proposed MLPM-FKNN (M) approach for all datasets. Turning into the distance measure in the MLPM-FKNN classifier, the Minkowski distance-based approach has achieved slightly better accuracy than the Euclidean distance-based approach in all cases considered, which signifies the effectiveness of using Minkowski distance in the proposed method for biomass feedstock classification.

To visually inspect the impact of the different values of k and p on the classification performance of the proposed MLPM-FKNN (M) approach, Fig. 2 illustrates the mean accuracies during the training and validation with all datasets when q at its optimum.

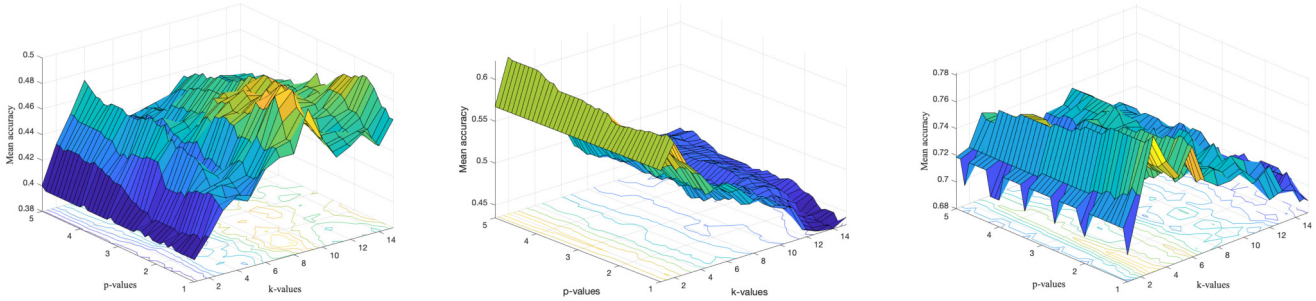
B. Classification performance with the test data

The classification results of each classifier with the test data instances are presented in Table III. In the testing step, we evaluated the performance of the trained models with the test data instances using the training instances that were stored during the holdout validation. As a result, the mean values of the performance measures are reported.

The results with the test data instances show that the proposed MLPM-FKNN (M) approach has a high classification accuracy compared to the benchmarks. In particular, it has a good accuracy of 70.88% with dataset 3, acceptable performance with dataset 2, and somewhat low accuracy of 42.62% with dataset 1. Along with them, other performance measures also remain reasonable, while the specificity is always higher

TABLE III: Classification performance with the test data

Model	Measure	Dataset 1	Dataset 2	Dataset 3
MLPM-FKNN (Minkowski)	Accuracy	**0.4262**	**0.5320**	**0.7088**
	Sensitivity	0.3850	0.4788	0.6935
	Specificity	0.8490	0.8736	0.9248
	STD	0.0508	0.0552	0.0161
MLPM-FKNN (Euclidean)	Accuracy	0.3786	0.5180	0.7059
	Sensitivity	0.3305	0.4721	0.6932
	Specificity	0.8349	0.8702	0.9242
	STD	0.0208	0.0527	0.0180
KNN	Accuracy	0.4000	0.4780	0.5853
	Sensitivity	0.3463	0.4538	0.5719
	Specificity	0.8442	0.8634	0.8931
	STD	0.0369	0.0458	0.0192
FKNN	Accuracy	0.3952	0.4760	0.6265
	Sensitivity	0.3508	0.3942	0.6232
	Specificity	0.8396	0.8570	0.9035
	STD	0.0256	0.0428	0.0305
SVM	Accuracy	0.4143	0.4960	0.6912
	Sensitivity	0.3605	0.3871	0.6721
	Specificity	0.8453	0.8591	0.9200
	STD	0.0392	0.0398	0.0180

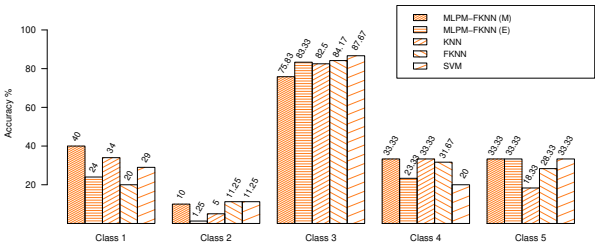


(a) Dataset 1 ($q = 1$) (b) Dataset 2 ($q = 3$) (c) Dataset 3 ($q = 1.5$)

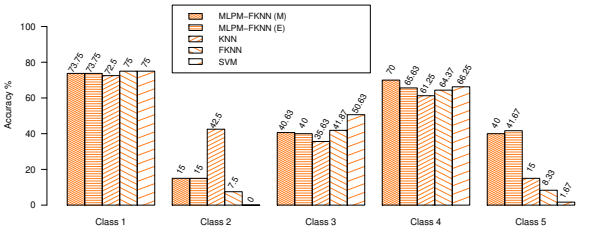
Fig. 2: Classification performance of the MLPM-FKNN (M) model with different p and k values for dataset 1, dataset 2, and dataset 3.

than the sensitivity. By looking at the others, even though the test performance of the KNN, FKNN and SVM models have comparable and generally good performance with dataset 3, they have relatively low performance with dataset 1 and dataset 2. Furthermore, it is apparent that for all methods used, the SDT is considerably lower for the test data (especially data set 3) than for the training and validation data.

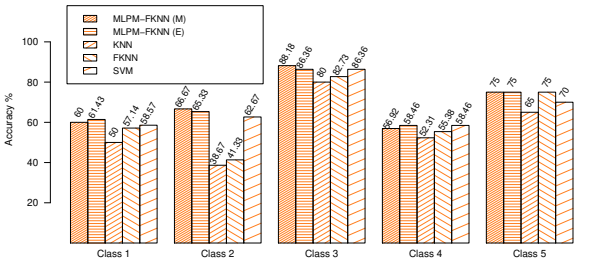
Fig. 3 shows the mean classification accuracy (measured from the confusion matrices) of each model for each class during the testing. It is apparent from the figure that all classifiers yielded good classifications on class 3 (that includes the wood-based energy crops) in dataset 1, whereas the SVM model performed the best. In dataset 2, class 1 (that includes energy grasses and their parts) and class 4 (that includes food crop residues-based biomass samples) have offered good and reasonable performance with all classifiers. In contrast, the classification performance of all methods in other classes of dataset 1 and dataset 2 appear to be poor—it is even worst for some cases, for instance, with class 2 in dataset 1. This might be because these classes are represented by a small number of biomass samples in the data. On the contrary, the classes (for example, class 3 in dataset 1) that are largely represented in the data have offered better classification. This indicates that the classification performance of these classes can be improved by introducing more data with approximately the same number of instances from all classes. It is also apparently supported by the results on dataset 3, where one can observe that the biomass samples in all classes generally produced good classification performance with all methods. This finding indicates that more biomass samples with relevant features from the proximate and ultimate analyses contribute to better results in their classification. Overall, it is evident from the result on dataset 3 that even though all the classifiers have comparable good performance, the MLPM-FKNN classifiers appear to be performing well for all classes classifications, whereas the KNN method performs the least.



(a) Dataset 1



(b) Dataset 2



(c) Dataset 3

Fig. 3: Comparison of classification performance of each model for each class with test data.

V. Conclusion

This paper presents a novel approach based on the MLPM-FKNN classifier and Minkowski distance for biomass feed-

stocks classification. An essential characteristic of this approach is that the generalization through power means and the Minkowski distance allows testing of different parameter values and enables a better fit of the method, consequently improving classification accuracy. We interpreted the biomass feedstocks classification as a five-class problem. Input features of the biomass samples included their characteristics from the proximate analysis and ultimate analysis. The experimental classification results clearly show that the proposed approach can achieve better performance than the benchmarks and can potentially produce an efficient classification that can benefit categorization of biomass sources for generating energy. The experimental results also validate the usefulness of the proposed MLPM-FKNN (M) method for multi-class imbalance real-world problems. Besides, it is evident from the results that the features from both ultimate and proximate analyses can offer a better classification of biomass feedstocks than the features considered from each of those analyses separately.

Future research possibilities include, for example, testing the classification performance of the proposed approach with more extensive biomass data that adequately comprises all classes specified in this study. Additional data will enhance the accuracy and the classification performance for wider range of biomass types and characteristics, in general.

REFERENCES

[1] A. Demirbas, "Biomass feedstocks," *In: Biofuels; Green Energy and Technology,* Springer, 2009, pp. 45-85.

[2] S. Gent and M. Twedt and C. Gerometta and E. Almberg, "Chapter Two - Introduction to Feedstocks," in *Theoretical and Applied Aspects of Biomass Torrefaction,* Butterworth-Heinemann, 2017, pp. 17-39

[3] A. A. Adeleke and J. K. Odusote and P. P. Ikubanni and O. A. Lasode, and M. Malathi, and D. Paswan,"The ignitability, fuel ratio and ash fusion temperatures of torrefied woody biomass," *Heliyon,* vol. 6, 2020, pp. e03582.

[4] A.A. Adeleke and P.P. Ikubanni and T.A. Orhadahwe and C.T. Christopher and J.M. Akano and O.O. Agboola and S.O. Adegoke and A.O. Balogun and R.A. Ibikunle, "Sustainability of multifaceted usage of biomass: A review," *Heliyon,* vol. 7, 2021, pp. e08025.

[5] O. O. Olatunji and S. Akinlabi and N. Madushele, "Property-based biomass feedstock grading using k-nearest neighbor technique," *Energy,* vol. 190, 2020, pp. 116346.

[6] P. Basu, Chapter 2 - Biomass Characteristics. *Biomass Gasification and Pyrolysis,* 2010, pp. 27-63.

[7] A. A. Khan and W. D. Jong and P. J. Jansens and H. Spliethoff, "Biomass combustion in fluidized bed boilers: Potential problems and remedies," *Fuel Process,* vol. 90, 2009, pp. 21-50.

[8] A. Nag and A. Gerritsen and C. Doeppke and A. E. Harman-Ware, "Machine Learning-Based Classification of Lignocellulosic Biomass from Pyrolysis-Molecular Beam Mass Spectrometry Data," *Int. J. Mol. Sci.,* vol. 22, 2021, pp. 4107.

[9] G. Tao and T. A. Lestander and P. Geladi and S. Xiong, "Biomass properties in association with plant species and assortments I: a synthesis based on literature data of energy properties," *Renew. Sustain. Energy Rev.,* vol. 16, 2012, pp. 3481-3506.

[10] M. Wang et al., "To distinguish the primary characteristics of agro-waste biomass by the principal component analysis: An investigation in East China," *Waste Manage.,* vol. 90, 2019, pp. 100-120.

[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory,* vol. 13, 1967, pp. 21-27.

[12] M. M. Kumbure and P. Luukka and M. Collan, "An enhancement of fuzzy k-nearest neighbor classifier using multi-local power means," *Proc. 11th Conf. European Society for Fuzzy Logic and Technology (EUSFLAT),* Atlantis Press. 2019, pp. 83-90.

[13] J. Parikh and S. A. Channiwala and G. K. Ghosal, "A correlation for calculating HHV from proximate analysis of solid fuels," *Fuel,* vol. 84, 2005, pp. 487-494.

[14] D. R. Nhuchhen and P. A. Salam, "Estimation of higher heating value of biomass from proximate analysis: A new approach," *Fuel,* vol. 99, 2012, pp. 55-63.

[15] S. V. Vassilev and D. Baxter and L. K. Andersen and C. G. Vassileva, "An overview of the chemical composition of biomass," *Fuel* vol. 89, 2010, pp. 913-933.

[16] M. Sajdak and O. Piotrowski, "C&RT model application in classification of biomass for energy production and environmental protection," *Cent. Eur. J. Chem.,* vol. 11, 2013, pp. 259-270

[17] Energy Research Centre of the Netherlands. *Phyllis 2: database for biomass and waste,* [Online]. Available: https://phyllis.nl/Browse/Standard/ECN-Phyllis#eucalyptus. [Accessed: July 31, 2021].

[18] J. M. Keller and M. R. Gray and J. A. Givens, "A Fuzzy K-Nearest Neighbor Algorithm," *EEE Trans. Syst. Man Cybern. Syst. ,* vol. 15, 1985, pp. 580-585.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.,* vol. 20, 1995, pp. 273-297.

[20] B. Salami and K. Haataja and P. Toivanen, "State-of-the-Art Techniques in Artificial Intelligence for Continual Learning: A Review" *Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems,* M. Ganzha, and L. Maciaszek, M. Paprzycki and D. Ślęzak, Eds. ACSIS, vol. 26, 2021, pp. 23-32.

[21] P. Gepner, "Machine Learning and High-Performance Computing Hybrid Systems, a New Way of Performance Acceleration in Engineering and Scientific Applications" *Proceedings of the 16th Conference on Computer Science and Intelligence Systems,* M. Ganzha, and L. Maciaszek, M. Paprzycki and D. Ślęzak, Eds. ACSIS, vol. 212, 2021, pp. 27-36.

[22] A. Coluccia and A. Fascista and G. Ricci, "A k-nearest neighbors approach to the design of radar detectors," *Signal Process.,* vol. 174, 2020, pp. 107609.

[23] R. Arian and A. Hariri and A. Mehridehnavi and A. Fassihi and F. Ghasemi, "Protein kinase inhibitors' classification using K-Nearest neighbor algorithm," *Comput. Biol. Chem.,* vol. 86, 2020, pp. 107269.

[24] S. Wua et al., "Evolving fuzzy k-nearest neighbors using an enhanced sine cosine algorithm: Case study of lupus nephritis," *Comput. Biol. Med.,* vol. 135, 2021, pp. 104582.

[25] M. M. Kumbure and P. Luukka and M. Collan, "A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean," *Pattern Recognit. Lett.,* vol. 140, 2020, pp. 172-178.

[26] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat Surv.,* vol. 4, 2010, pp. 40–79.

[27] P. Vuttipittayamongkol and E. Elyan and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowl.-Based Syst.,* vol. 212, 2021, pp. 106631

KNOWCON
Knowledge on Economics and Management
International Scientific Conference

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 37–44

DOI: 10.15439/2021B4
ISSN 2300-5963 ACSIS, Vol. 29

Image based classification of shipments using transfer learning

Markus Leppioja
markus.leppioja@gmail.com
Loihde Analytics
Valtakatu 49, 53100 Lappeenranta, Finland

Pasi Luukka
pasi.luukka@lut.fi
LUT University
School of Business & Management
Yliopistonkatu 34, 53850
Lappeenranta, Finland

Christoph Lohrmann
christoph.lohrmann@lut.fi
LUT University
School of Business & Management
Yliopistonkatu 34, 53850
Lappeenranta, Finland

***Abstract*—This paper focuses on recognizing different postal shipment types from images taken by the sorting machine. Greyscale images obtained from sorting machines are used to build a classifier using transfer learning to recognize seven different classes of shipments. Three convolutional neural networks (VGG16, GoogLeNet and ResNet50), that were pretrained using the ImageNet dataset, were used as feature extractors and the extracted features were subsequently supplied to a neural network classifier. VGG16 demonstrated the best performance for six out of the seven classes and achieved an overall mean accuracy of 95.69% on the independent test set. The model accomplished F1 scores exceeding 90% for five out of seven classes, only having a lower recall for the aggregated class "Other" and shipments from abroad. The results of this study highlight the potential of transfer learning for computer vision in the context of shipment classification.**

I. Introduction

THE objective in this study is to build a classifier to effectively recognize different shipment types from images taken by a sorting machine. Data for the shipment type classification problem is obtained from a company operating in the field of postal and logistics services. Different types of shipments arrive from several sources to the company's networks. These shipments pass through a sorting process which divides the shipments based on the location of the destination. However, the sorting machine is not capable of recognising the type of each shipment and the number of shipments of each type, which are both of interest to the company. Especially the recognition of consumer-to-consumer letters is pivotal since there are no preannouncements related to this shipment type whereas some larger customers make preannouncements about future shipments to ensure their smooth processing. Thus, being able to recognize the type of shipments, especially the "Consumer Letter" type, but also all other types, is the main aim of this work. The problem presents itself as a computer vision problem where an image is taken by a sorting machine and a classifier needs to be built to recognize which shipment type is present in the image. From this information, the quantities for all types of shipments can be inferred, thus addressing both objectives for the case company.

For this type of problem deep learning and convolutional neural networks (CNN) have proven to be useful. Nowadays the databases that CNNs are trained on are so large that at least low-level features extracted in the first convolutional blocks are useful in almost any computer vision application. Thus, the features extracted from such pretrained models are commonly used, whereas training a new CNN from scratch is rare [1]. The advantage of using a pretrained CNN is that it is computationally less complex, and less data is needed to fit a new classifier than for fully training a CNN model. Limitations on the computational complexity are also the reason for the application of a pretrained CNN in this study.

Pretrained convolutional neural networks are usable in many different fields. For example, Pardamean et al. [2] had a small size mammogram dataset and used transfer learning of a convolutional neural network pretrained on chest X-ray data to overcome this problem. The best model was able to achieve a 90.38% accuracy. Sun and Qian [3] worked on a Chinese herbal medicine recognition task from images using a pretrained convolutional neural network VGG16. They managed to achieve an average precision of 71% which these authors considered promising. Reddy and Juliet [4] used transfer learning with the objective to classify malarial infected cells and improve the malaria diagnostics accuracy with the pretrained convolutional neural network ResNet50. They reported to have obtained an accuracy of 95.4%. In the study of Chmielinska's and Jakubowski [5] the problem was to develop a detector for driver fatigue symptoms based on facial images. Driver fatigue is considered one of the main causes for car accidents. In this case the authors used a pretrained convolutional neural network called AlexNet. Their results indicate that it is possible to use transfer learning for the detection of driver fatigue symptoms. The best class had an error rate of less than 2%. Abu Mallouh et al. [6] worked on classifying peoples' age range from images. They managed to show that pretrained CNNs can be used for this problem. Their model outperformed the previous state of the art solution by 12%. Sert and Boyacı [7] worked on a free-hand sketch recognition problem. They deployed three pretrained convolutional neural networks for feature extraction: AlexNet, VGG16 and GN-Triplet [8]. A support vector machine was used as a classifier. The model which was able to achieve the best accuracy of 97.91% used a combination of AlexNet and GN-Triplet together with PCA. Fu and Aldrich [9] used convolutional neural networks for analysing a froth flotation process from images. In their study AlexNet performed the best and managed to outperform the previous best solutions. Shao et al. [10] worked on a machine fault diagnostic problem. They selected the VGG16 pretrained convolutional neural network for their study. The best performing, finetuned VGG16 model's accuracy was reported to be almost 100%. The recognition of plant species was the sub-

ject in the research problem covered by Ghazi et al. [11]. They used three different pretrained convolutional neural networks: VGG16, AlexNet and GoogLeNet. The best performing model with accuracy of 80.18% was achieved with a combination of VGG16 and GoogLeNet. Data augmentation and finetuning the number of iterations was considered the most important factors influencing the results. Tree species identification from wooden boards was the subject in the study by Shustrov et al. [12]. They used the four convolutional neural network architectures AlexNet, VGG16, GoogLeNet and ResNet to address this problem. The highest accuracy of 94.7 % was obtained with GoogLeNet. Besides this, Camargo et al. [13] used the pretrained convolutional neural network AlexNet to classify sunspots and were able to achieve an accuracy of 91.70%. Finally, Zhao et al. [14] built a classifier for land-use with a transfer learning technique and spatial resolution images available for the land-use.

The results show that transfer learning based on pretrained convolutional neural networks was successfully applied in many different fields and contexts. It is thus also selected for the machine vision problem in this study.

II. Convolutional Neural Networks

Fully connected neural networks connect each neuron in a layer with all neurons in the subsequent layer [15]. Since the weight of each of these connections represents a parameter to be learned during model training, fully connected neural networks tend to have a large number of parameters that need to be trained [1]. This problem is amplified when there are many neurons in each layer and / or there are many layers in the network - which is not uncommon in deep learning problems. The key idea behind convolutional neural networks (CNN) is to create a solution in a way that reduces the number of parameters compared to fully connected neural networks. This allows to train deeper networks with less parameters [16], [29], [30].

One of the first convolutional architectures was LenNet-5, which was applied to identify hand-written numbers [17]. Since LeNet-5, convolutional neural networks have evolved in terms of the number of layers and the use of different activation functions.

Convolutional neural networks are combinations of convolutional and pooling layers. The last layers are usually fully connected ones. The network can be defined through the number of filters, stride lengths, the number of convolution pooling combinations and the fully connected layers. Fig. 1 represents such a simple network [18].

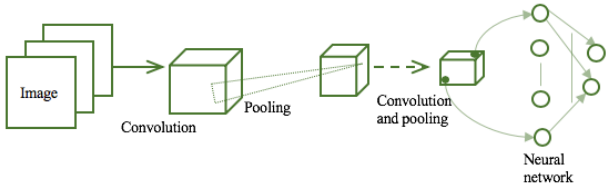


Fig 1. A simple convolutional neural network, reproduced from Rebala et al. [18]

The key aspect of convolutional neural networks is an operation called “convolution”. Convolution is a dot product operation between grid-structured inputs and a grid-structured set of weights which is drawn from different spatial localities in the input volume. It is useful when there is a high level of spatial locality in the data, for instance, in case of image data.

The goal of the pooling layer is to reduce the dimensionality of feature maps. Hence, the pooling can be called “down sampling”. In a pooling operation, the maximum (or sometimes the average) of a small grid region is returned [1]. The pooling is applied to every feature map separately, whereas a convolution operation uses all feature maps simultaneously [1], [16]. This is the reason why the pooling operation doesn’t change the number of feature maps – the depth stays the same [1]. Nevertheless, the dimensionality of the feature maps reduces spatially [16].

The convolutional neural network works in a similar way as a regular feed-forward neural network. The difference is that the operations in the layers are spatially organized with sparse connections. The ReLU activation typically follows the convolutional operation hence it is not usually shown independently when illustrating convolutional neural networks. Compared to other common activation functions, ReLU is advantageous in terms of speed and accuracy [1].

Convolutional neural networks allow translation invariance [19]. This means, for instance, in images that an object is the same object no matter where it is located in the image [19]. This is related to weight (or parameter) sharing - a particular shape should be processed the same way regardless of its spatial location [1]. There has been a great advancement in the field of image classification in the 2010s due to the development of the ImageNet database [20]. It contains over 14 million images with a large number of sub-categories [21]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a competition where participants use the ImageNet database in different tasks. ILSVRC has been arranged from 2010 to 2017 yearly and many state-of-the-art CNN architectures have participated and won the challenge.

A. VGG

Visual Geometry Group’s (VGG) convolutional neural network placed second in the ILSVRCs image classification task. Simonyan and Zisserman [22] present different versions of their model in their article, for instance VGG16 and VGG19. The architecture of VGG16 is shown in Fig. 2.

There are 16 weight layers in VGG16, out of which there are 13 convolutional weight layers. In between each two to three convolutional layers is a max-pooling layer. Moreover, the three last layers are fully connected. The ReLU activation function is selected in the convolutional part and in the first two fully connected layers, while the softmax activation function is used in the last layer which provides the class probabilities (outputs). The core idea is to use 3x3 filters instead of the widely used 5x5 or 7x7 filters. In particular, a 3x3 filter is used three times in a row. The advantage of this approach is that the decision function is more discriminative. Another advantage is that there are less parameters in this approach compared to the versions with 5x5 or 7x7 fil-

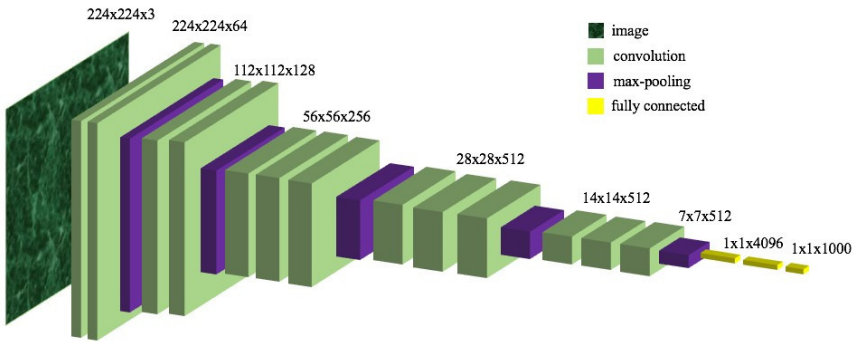


Fig 2. Illustration of the VGG16 architecture

ters, reducing the overfitting problem. There are altogether 138 million parameters in the VGG16 model [22].

B. GoogLeNet

GoogLeNet is a convolutional neural network architecture and the winner of the ILSVRC 2014 challenge in image classification [23]. To reduce the dimensionality and the computation load, GoogleLeNet heavily relies on 1x1 convolutions. The inception module is displayed in Fig. 3. The idea of inception modules is to extract features using 1x1, 3x3, 5x5 convolutions and 3x3 max-pooling and then combine them together [24].

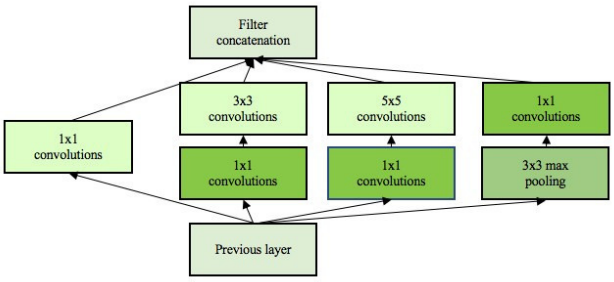


Fig 3. Inception module, reproduced from Szegedy et al. [24]

GoogLeNet is a deep CNN, containing 27 layers - counting both weight layers (22) and pooling (5) layers. All the convolutions are using the ReLU activation, also the convolutions inside the inception modules.

GoogLeNet uses one average pooling layer instead of a fully connected layer after the convolutional layers and, thus, reduces overfitting. There is also one dropout layer after the average pooling layer. The last layer is fully connected, and it uses a softmax activation. On top of the original GoogLeNet model, some of the authors have introduced modifications called InceptionV2 and InceptionV3. The goal of these modifications is to scale up the network and add regularization in as computationally efficient ways as possible [24].

C. ResNet

ResNet is a CNN architecture and the winner of the ILSVRC 2015 image classification task. The winning model contained 152 trainable layers. It is the deepest model ever presented in the ILSVRC. However, it is noteworthy that the complexity of ResNet-152 is still lower than VGG's CNN [25]. Deep convolutional neural networks suffer from the vanishing/exploding gradient problem. This increases the error in a very deep CNN. The solution to the stated problem is shortcut connections as shown in Fig. 4. The shortcut connection can skip one or more layers and the outputs are added to the outputs of the stacked layer. This reduces the vanishing/exploding gradient problem and allows to build deeper networks. Basic identity shortcut connections do not add parameters or complexity to the model. Identity shortcuts can be used when the input and output have the same dimensions [25].

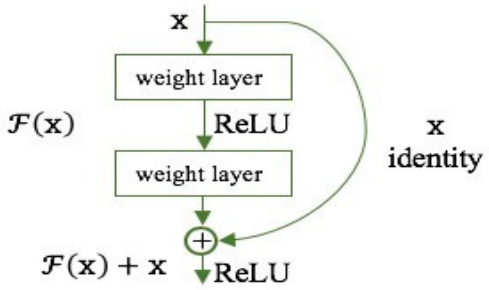


Fig 4. The identity shortcut connection, reproduced from [25]

A bottleneck design is used for the deep ResNet models. In particular, 1x1 convolutions are added to the start and the end of the network. This approach is the same kind as in GoogLeNet. Convolutional layers use the ReLU activation. After the convolutional part, the average pooling and one fully connected layer are used [25].

D. Transfer learning and finetuning

Deep convolutional neural networks contain a large number of parameters. The large number of parameters ensures their ability to learn complex tasks. However, it also means that a considerable amount of data is needed to fully train such models adequately. Having said this, for many applications a large amount of labelled data might not be available [1], [26]. If there is not sufficient training data, the model will suffer from overfitting and won't generalize well [26]. Data availability is the key reason why the technique called transfer learning has been developed.

Pattanayak [26] (p. 211) describes transfer learning as follows: "Transfer learning in a broad sense refers to storing knowledge gained while solving a problem using that

knowledge for a different problem in a similar domain." Aggarwal [1] points out that using features extracted from public data sources, such as ImageNet, can be viewed as transfer learning. This is beneficial for image data since features extracted from a certain dataset are reusable across data sources [1]. For a new problem, less data is needed because low-level features were already extracted previously from another data domain. The reason for this is that when images are processed through many layers of convolutions, the initial layers learned to detect universal features such as shapes and edges [1], [26].

The simplest way to implement transfer learning is to remove the original output layer of an existing, trained model and replace it with the one suitable to the new problem [19]. Another option is to remove the topmost layers of the original network and use the output as features (inputs) in a new machine learning model [19]. The new machine learning model can be, for instance, a support vector machine, a random forest or a neural network [19]. There is also the possibility to freeze certain layers of the pretrained model and then retrain the model [19]. This means that weights of the frozen layers are not updated during the training [19]. Retraining some of the layers is often referred to as 'finetuning' [1], [27].

III. Shipment Type Classification Problem and Results.

A. Dataset and transfer learning strategy

The dataset used in this study contains images of shipments from sorting machines with the shipment ID and shipment type. The shipment type was classified manually (by hand). The size of the dataset is 25'979 shipments with 13 different shipment types. The rarest shipment types were grouped together to the class "Other", so that this classification problem eventually contained only seven different classes (Table I).

TABLE I.
Classes and their frequency of occurrence in the dataset.

Class number	Class name	Number of samples
0	Image not found	333
1	Consumer letter	1'677
2	Commercial shipment	3'938
3	Shipment from abroad	1'125
4	Corporate letter	16'265
5	Magazine	1'956
6	Other	685

To build the classifier, a technique of transfer learning is used. Three pretrained convolutional neural networks, VGG16, GoogLeNet and ResNet50 are used. All three models were selected since they are commonly used for image classification in the literature and, additionally, have demonstrated their ability to perform well on challenging image classification problems e.g., in the ILSVRC. The top layers of these models are removed, and all the other layers remain frozen. The pretrained models are used as feature extractors. On top of these models, a simple three layered fully connected neural network classifier is utilized. The first layer is a dense layer, which takes the features as an input. It contains 256 nodes and uses the ReLU activation. The next layer is a dropout layer. This is used to avoid overfitting and to add regularization. A dropout ratio of 0.5 is used. The final layer is the output layer, which makes the actual prediction. The activation function softmax is used in this layer. The output is a probability distribution. The class which has the highest probability is the one that the model predicts. Different classes are evaluated in terms of F1 score, precision as well as recall and based on the results, there is a variability of the model's performance on the different classes.

Based on the scientific literature, three different CNN architectures were selected for the application to this problem. These were the 16-layer VGG model (VGG16), GoogLeNet (InceptionV3) and the 50-layer ResNet (ResNet50). The strategy was to apply transfer learning to these models, which had been pretrained on the ImageNet dataset, and to compare these models' performance to find the best one for classifying the shipment types from images.

The data is divided into training and testing sets with a 90-10% split (holdout method). Additionally, a 10-fold cross validation is performed for the training data and the results are averaged over the 10 folds of the cross validation. The batch size is set to 25 and the number of epochs to 100.

B. Results from the models

The classification accuracies, cross-entropy losses and standard deviations of the validation results are displayed in Table II. The highest accuracy of 95.11% was obtained with the VGG16. However, the other two models were also capable of achieving an accuracy of over 90 %. The lowest categorical cross-entropy loss was obtained by ResNet50 and the highest by VGG16. The sample standard deviation of VGG16's loss was relatively high compared to the other models. The results indicate that VGG16 might be suffering from some degree of overfitting. This was supported by the observation that the loss value varied much between the folds, compared to GoogLeNet and ResNet50. However, the model is clearly performing best in terms of accuracy.

Additionally, F_1 scores and their sample standard deviations are presented for each class and each model in Table III. It is noteworthy that all models tend to perform poorer on the classes "Other" and "Shipment from abroad" and also have clearly higher sample standard deviations. Overall, VGG16 produces the best F_1 scores in six out of seven classes.

TABLE II.
Accuracies, categorical cross-entropy losses and their sample standard deviation (validation results).

	VGG16	GoogLeNet	ResNet50
Accuracy	**95.11%** (+-0.6265%)	91.87% (+-0.5622%)	93.24% (+-0.4261%)
Categorical cross-entropy loss	1.121 (+-0.2988)	0.3621 (+-0.0337)	0.3773 (+-0.0341)

In Table IV the precision and recall values together with their standard deviations are reported for VGG16. Since F_1 scores are based only on precision and recall, the results for

TABLE III.
F1 scores and sample standard deviations of each class, the highest F1 score for each class is in bold.

Method	VGG16 F_1 score	GoogLeNet F_1 score	ResNet50 F_1 score
Image not found (0)	89.47% (+-4.457%)	**89.51%** (+-3.483%)	87.58% (+-4.640%)
Consumer letter (1)	**92.89%** (+-1.234%)	87.74% (+-1.414%)	90.10% (+-1.913%)
Commercial shipment (2)	**92.75%** (+-1.739%)	86.74% (+-1.785%)	89.54% (+-1.081%)
Shipment from abroad (3)	**83.85%** (+-3.750%)	72.17% (+-4.434%)	77.09% (+-4.103%)
Corporate letter (4)	**97.78%** (+-0.3716%)	95.96% (+-0.3776%)	96.70% (+-0.2292%)
Magazine (5)	**91.27%** (+-1.721%)	88.43% (+-1.875%)	89.82% (+-1.534%)
Other (6)	**79.48%** (+-5.286%)	68.49% (+-5.723%)	73.21% (+-6.879%)

TABLE IV.
Precision, recall and their sample standard deviations for VGG16 (validation results).

VGG16	Precision	Recall
Image not found (0)	97.36% (+-3.445%)	83.06% (+-6.969%)
Consumer letter (1)	93.45% (+-1.972%)	92.38% (+-1.361%)
Commercial shipment (2)	92.94% (+-1.866%)	92.59% (+-2.086%)
Shipment from abroad (3)	91.66% (+-4.002%)	77.52% (+-5.812%)
Corporate letter (4)	96.89% (+-0.5723%)	98.69% (+-0.3279%)
Magazine (5)	90.09% (+-1.827%)	92.52% (+-2.484%)
Other (6)	86.95% (+-5.098%)	73.51% (+-7.310%)

GoogleLeNet and ResNet50 are lower for most of the classes also in terms of these two metrics and can be found in Table VII and Table VIII in the appendix.

According to Table IV, all precision values for VGG16 are relatively high. Two of the lowest precision values, which are also characterized by high sample standard deviations, are linked to the "Other" and "Shipment from abroad" classes. For instance, a precision value of over 90 % was achieved for all classes, except for the class "Other".

A similar situation is encountered for the recall of VGG16, where the "Other" and "Shipment from abroad" classes both show values below 80% - the lowest recalls of all classes. On the "Consumer letter" class, which is one of the classes of the highest interest for the case company, the model is overall performing well: the precision value is 93.45% and the recall value is 92.38%.

C. *Test set results*

Applying VGG16 on the test set, an accuracy of 95.69% and a categorical cross-entropy loss value of 0.9176 were achieved, which are close to the average results obtained on the validation sets. These results indicate that VGG16 is indeed performing well and has the ability to generalize its performance for shipment classification. The test set's F_1 score, precision and recall are presented in Table V. The F_1 score is higher than 90% for five out of seven classes and is still above 80% for the "Shipment from abroad" and "Other" classes. When compared to the validation results, it is apparent that for the "Consumer letter" class the F_1 score, precision and recall are a bit lower in the test set results.

The recall values of the "Shipment from abroad" and "Other" class are comparably low. The low recall values indicate that the classifier is not able to identify these classes very well from the samples and many of the samples that actually belong to these classes are falsely assigned to one of the other classes. One reason for the low recall value is that the class "Other" consists of several smaller classes which were combined to one (13 classes originally of which seven were aggregated into this class). This of course also entails that samples in this class are more dissimilar among each other than in other classes. The results indicate that this clearly has an effect on the recall (and precision) for this class. Another reason for the low recall in this class can be the low sample size. Overall, there were only 685 samples in this class which is the second smallest of all classes. The fact that the class "Image not found", which has the smallest sample size but is not aggregated, has a considerably lower recall than all other classes (other than "Other" and "Shipment from abroad") reinforces this reasoning.

TABLE V.
F1 score, precision and recall for each class of the test set.

VGG16	F_1 score	Precision	Recall
Image not found (0)	90.14%	96.97%	84.21%
Consumer letter (1)	91.93%	92.25%	91.61%
Commercial shipment (2)	94.01%	93.42%	94.62%
Shipment from abroad (3)	83.81%	92.63%	76.52%
Corporate letter (4)	98.02%	96.98%	99.09%
Magazine (5)	93.99%	94.24%	93.75%
Other (6)	80.65%	90.91%	72.46%

For the class "Shipment from abroad" the comparably low performance values can be explained by the fact that it – even though it was not aggregated from classes - also contains different types of shipments, which are all coming from abroad. These shipments can vary considerably, and it seems that the classifier has some difficulty in finding the similarities between shipments belonging to this class (see Table VI). Table VI highlights that the class "Shipment from abroad" is most often misclassified into the classes "Consumer letter" and "Corporate letter".

The reason for this can be that shipments coming from abroad are often letter type shipments – making it hard to differentiate the "Shipment from abroad" class from these other two classes and, to some smaller degree, vice versa.

A noticeable misclassification error can also be detected between the classes "Commercial shipment" and "Magazine". This appears plausible since some magazines have commercial contents on the back cover. Besides this, "Commercial shipment" is a relatively heterogeneous class since it contains different kinds of shipments. Overall, the test set indicates that the classifier is performing well for the shipment type classification. Moreover, the confusion matrix and the misclassification errors are consistent with those obtained during the validation (see Table IX in the appendix).

IV. Conclusions

In this study, pretrained convolutional neural networks were applied for a shipment type recognition problem. The convolutional neural networks were pretrained using the ImageNet dataset and a transfer learning strategy that is suitable for shipment type classification was developed. In particular, three different models were selected for the application to this particular problem: VGG16, GoogLeNet and ResNet50.

These models were used as feature extractors and the extracted features were subsequently supplied to the classifier. The classifier developed for this purpose was a simple neural network. The dataset available for this study contained images of shipments taken by sorting machines and differentiates seven classes of shipments. The highest mean accuracy of 95.11% was obtained with VGG16 selected as the feature extractor on the validation data. ResNet50 achieved a mean accuracy of 93.24% and GoogLeNet of 91.87%. For the validation data sets VGG16 performed overall the best and produced the best results in every class except one. From the business perspective, the most important class to recognise in this study was "Consumer letter". The model demonstrated on this class its second-best performance of all classes in terms of the F_1 score (92.89%) and precision (93.45%) and a comparably high recall (92.38%). On the independent test set, VGG16 obtained an accuracy of 95.69%, which is almost identical to the mean accuracy obtained on the validation data sets. Moreover, given that the majority class accounts for only 62.61% of the data, this result seems overall very promising. The F_1 score for the "Consumer letter" class in the test set was with 91.93% also comparable to that obtained during the validation. Overall, the confusion matrix also indicated that the misclassification error is largely based on plausible misclassifications that are linked to same classes being similar to each other and/or heterogenous within (e.g., "Shipments from Abroad" with "Consumer Letter" and "Corporate Letter").

It is noteworthy that there was more variability for the categorical cross-entropy loss and accuracy for the cross validated results of VGG16 in terms of the sample standard deviations than for the other models. It should be kept in mind, that the trained classifier with the VGG16 model possesses considerably more parameters than the other two models due to the larger output vector of VGG16. Because of this, there is a larger possibility to run into overfitting problems with VGG16. When for a dataset of given size, the number of parameters is larger, there is a greater chance to tune also the less useful parameters' values as part of the final model. However, given the consistently high and similar results of VGG16 for cross-validation and the test set, this is likely neither a major concern nor critical. The training of all three models was relatively fast – which is one of the main advantages of the transfer learning approach. Unsurprisingly, VGG16 took the longest to train since it has more parameters than the two other classifiers. However, training a full model from scratch would have taken considerably longer. ˇ

TABLE VI.
Confusion matrix of VGG16 of the test set.

True labels \ Predictions	VGG 16 (n = 2598)	Image not found (0)	Consumer letter (1)	Commercial shipment (2)	Shipment from abroad (3)	Corporate letter (4)	Magazine (5)	Other (6)
	Image not found (0)	32	0	3	0	3	0	0
	Consumer letter (1)	0	131	2	4	5	0	1
	Commercial shipment (2)	0	0	369	0	13	8	0
	Shipment from abroad (3)	0	9	4	88	14	0	0
	Corporate letter (4)	0	1	7	2	1636	1	4
	Magazine (5)	0	0	10	1	1	180	0
	Other (6)	1	1	0	0	15	2	50

Acknowledgment

The authors acknowledge that this paper is based on Markus Leppioja's master's thesis titled "Shipment type classification from images" [28]. The authors would like to thank Artur Vuorimaa for his help in editing the text.

References

[1] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Cham: Springer International Publishing, 2018.

[2] B. Pardamean, T. W. Cenggoro, R. Rahutomo, A. Budiarto, and E. K. Karuppiah, "Transfer Learning from Chest X-Ray Pre-trained Convolutional Neural Network for Learning Mammogram Data,"

Procedia Comput. Sci., vol. 135, pp. 400–407, 2018, doi: https://doi.org/10.1016/j.procs.2018.08.190.
[3] X. Sun and H. Qian, "Chinese Herbal Medicine Image Recognition and Retrieval by Convolutional Neural Network," *PLoS One*, vol. 11, no. 6, pp. 1–19, 2016, doi: 10.1371/journal.pone.0156327.
[4] A. S. B. Reddy and D. S. Juliet, "Transfer Learning with ResNet-50 for Malaria Cell-Image Classification," in *2019 International Conference on Communication and Signal Processing (ICCSP)*, 2019, pp. 945–949, doi: 10.1109/ICCSP.2019.8697909.
[5] J. Chmielinska and J. Jakubowski, "Detection of driver fatigue symptoms using transfer learning," *Bull. Polish Acad. Sci.*, vol. 66, no. 6, pp. 869–874, 2018, doi: 10.24425/bpas.2018.125934.
[6] A. Abu Mallouh, Z. Qawaqneh, and B. D. Barkana, "Utilizing CNNs and transfer learning of pre-trained models for age range classification from unconstrained face images," *Image Vis. Comput.*, vol. 88, pp. 41–51, 2019, doi: https://doi.org/10.1016/j.imavis.2019.05.001.
[7] M. Sert and E. Boyaci, "Sketch Recognition Using Transfer Learning," *Multimed. Tools Appl.*, vol. 78, no. 12, pp. 17095–17112, 2019, doi: 10.1007/s11042-018-7067-1.
[8] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The Sketchy Database: Learning to Retrieve Badly Drawn Bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, 2016, doi: 10.1145/2897824.2925954.
[9] Y. Fu and C. Aldrich, "Froth image analysis by use of transfer learning and convolutional neural networks," *Miner. Eng.*, vol. 115, pp. 68–78, 2018, doi: https://doi.org/10.1016/j.mineng.2017.10.005.
[10] S. Shao, S. McAleer, R. Yan, and P. Baldi, "Highly Accurate Machine Fault Diagnosis Using Deep Transfer Learning," *IEEE Trans. Ind. Informatics*, vol. 15, no. 4, pp. 2446–2455, 2019, doi:10.1109/TII.2018.2864759.
[11] M. Mehdipour Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing*, vol. 235, pp. 228–235, 2017, doi: https://doi.org/10.1016/j.neucom.2017.01.018.
[12] D. Shustrov, T. Eerola, L. Lensu, H. Kälviäinen, and H. Haario, "Fine-Grained Wood Species Identification Using Convolutional Neural Networks," in *Image Analysis*, 2019, pp. 67–77.
[13] T. O. Camargo *et al.*, "Detecting a predefined solar spot group with a pretrained convolutional neural network," in *2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, 2019, pp. 1–6, doi: 10.1109/ColCACI.2019.8781990.
[14] B. Zhao, B. Huang, and Y. Zhong, "Transfer Learning With Fully Pretrained Deep Convolution Networks for Land-Use Classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1436–1440, 2017, doi: 10.1109/LGRS.2017.2691013.
[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer ScienceBusiness Media, 2006.
[16] H. H. Aghdam and E. J. Heravi, *Guide to convolutional neural networks. A practical application to traffic-sign detection and classification*. Springer International Publishing, 2017.
[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.
[18] G. Rebala, A. Ravi, and S. Churiwala, *An introduction to machine learning*. Springer International Publishing, 2019.
[19] M. Salvaris, D. Dean, and W. H. Tok, *Deep learning with Azure. Building and deploying artificial intelligence solutions on the Microsoft AI platform*. Apress, 2018.
[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[21] ImageNet, "Summary and Statistics," 2020. http://image-net.org/about-stats (accessed Feb. 29, 2020).
[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *3rd International Conference on Learning Representations, (ICLR)*, 2015, pp. 1–14, [Online]. Available: http://arxiv.org/abs/1409.1556.
[23] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.
[24] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[26] S. Pattanayak, *Pro deep learning with TensorFlow. A mathematical approach to advanced artificial intelligence in Python*. Apress, 2017.
[27] L. Mou and Z. Jin, *Tree-Based Convolutional Neural Networks: Principles and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2018.
[28] M. Leppioja, Shipment type classification from images, Master's thesis, LUT University, 2020
[29] K. Danilchenko and M. Segal, An efficient connected swarm deployment via deep learning, Annals of Computer Science and Information Systems, 25, 2021, pp. 1-7.
[30] A. M. Nguyen and H.S. Nguyen, Rotation Invariance in Graph Convolutional Networks, Annals of Computer Science and Information Systems, 25, 2021, pp. 81-90.

APPENDIX

TABLE VII.
Precision, recall and their sample standard deviations for GoogLeNet.(validation results)

GoogLeNet	Precision	Recall
Image not found (0)	97.40% (+-3.799%)	83.09% (+-6.022%)
Consumer letter (1)	87.97% (+-3.000%)	87.62% (+-2.434%)
Commercial shipment (2)	88.03% (+-3.011%)	85.62% (+-3.264%)
Shipment from abroad (3)	85.03% (+-5.892%)	63.27% (+-7.008%)
Corporate letter (4)	94.46% (+-0.8449%)	97.52% (+-0.5742%)
Magazine (5)	86.74% (+-3.205%)	90.36% (+-3.545%)
Other (6)	80.94% (+-3.453%)	59.93% (+-8.844%)

TABLE VIII.
Precision, recall and their sample standard deviations for ResNet50 (validation set results).

ResNet50	Precision	Recall
Image not found (0)	95.11% (+-3.213%)	81.66% (+-8.461%)
Consumer letter (1)	91.16% (+-1.892%)	89.11% (+-2.856%)
Commercial shipment (2)	90.57% (+-1.658%)	88.58% (+-2.186%)
Shipment from abroad (3)	84.28% (+-4.302%)	71.19% (+-5.331%)
Corporate letter (4)	95.45% (+-0.4993%)	97.98% (+-0.5020%)
Magazine (5)	88.44% (+-3.589%)	91.44% (+-3.084%)
Other (6)	85.50% (+-7.973%)	64.76% (+-9.042%)

TABLE IX.
Confusion matrix of VGG16 (Average validation performance).

		Predictions						
True labels	**VGG 16** (n = 2338.1)	**Image not found (0)**	**Consumer letter (1)**	**Commercial shipment (2)**	**Shipment from abroad (3)**	**Corporate letter (4)**	**Magazine (5)**	**Other (6)**
	Image not found (0)	24.50 (+-2.07)	0.30 (+-0.48)	1.40 (+-1.17)	0.00 (+-0.00)	2.40 (+-1.84)	0.30 (+-0.48)	0.60 (+-0.84)
	Consumer letter (1)	0.00 (+-0.00)	141.70 (+-1.89)	1.50 (+-1.51)	2.80 (+-1.81)	7.20 (+-1.93)	0.10 (+-0.32)	0.10 (+-0.2)
	Commercial shipment (2)	0.10 (+-0.32)	0.00 (+-0.00)	328.50 (+-7.32)	0.70 (+-0.68)	14.30 (+-6.06)	11.00 (+-2.75)	0.20 (+-0.42)
	Shipment from abroad (3)	0.00 (+-0.00)	7.00 (+-2.21)	3.90 (+-2.47)	78.30 (+-5.87)	9.20 (+-3.65)	1.70 (+-1.42)	0.90 (+-0.74)
	Corporate letter (4)	0.20 (+-0.63)	1.80 (+-0.92)	8.40 (+-2.37)	3.00 (+-2.16)	1442.30 (+-4.62)	2.10 (+-1.79)	3.60 (+-2.01)
	Magazine (5)	0.10 (+-0.32)	0.00 (+-0.00)	9.30 (+-3.30)	0.30 (+-0.68)	2.00 (+-1.25)	163.20 (+-4.34)	1.50 (+-1.51)
	Other (6)	0.30 (+-0.48)	0.90 (+-0.99)	0.50 (+-0.71)	0.50 (+-0.71)	11.30 (+-4.11)	2.80 (+-1.81)	45.30 (+-4.72)

KNOWCON
Knowledge on Economics and Management
International Scientific Conference

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 45–51

DOI: 10.15439/2021B6
ISSN 2300-5963 ACSIS, Vol. 29

Similarity based TOPSIS with linguistic-quantifier based aggregation using OWA

Pasi Luukka
School of Business and Management, LUT University,
Yliopistonkatu 34, 53851 Lappeenranta, Finland
Email: pasi.luukka@lut.fi

Jan Stoklasa
School of Business and Management, LUT University
Yliopistonkatu 34, 53851 Lappeenranta, Finland, and
Palacky University Olomouc, Faculty of Arts
Department of Economic and Managerial Studies
Email: jan.stoklasa@lut.fi
Email: jan.stoklasa@upol.cz

***Abstract*—In this paper we present similarity based TOPSIS with OWA operators. The motivation behind this new method is the fact that in many real world problems it is more important to consider the amount of criteria that a particular alternative is able to satisfy instead of simply concentrating on the importance of particular criteria. Here with OWA operators we can tackle this problem together with multi-criteria decision making method called TOPSIS by aggregating alternatives' similarities towards positive ideal solution and negative ideal solution and aggregating these similarities using OWA. The use of linguistic quantifiers represented by OWA weights generated by a selected RIM quantifier allows for the reflection of decision-maker's attitude to risk in the calculation of the similarities of the alternative with positive and negative ideal solutions.**

I. Introduction

THE name TOPSIS is shortening from the Technique for Order Preference by Similarity to Ideal Solution. This tool belongs to multi-criteria decision making methods which are of increasing importance [1], [2]. TOPSIS is based on the idea of forming two ideal solutions (best possible case called the *(positive) ideal solution* and denoted PIS and worst possible case called the *negative ideal solution* and denoted NIS), both relative to the set of available alternatives, and comparing the current alternative to these two. Unlike the name of the method suggests originally [3] this was done by computing the distances of each alternative to both ideal solutions and then forming the so called *relative closeness to the ideal solution* from these distances. The relative closeness to the ideal solution is originally defined in [3] in such a way that its value is equal to 1 for alternatives identical with PIS and 0 for alternatives identical with NIS. This way the relative closeness to the ideal solution takes into account the minimization of the distance of an alternative from PIS and the maximization of its distance from NIS and introduces a specific tradeoff between the two distances. The distances from PIS and NIS are calculated as Euclidean distances and as such do not reflect any behavioral or personality traits of the decision-maker. The weights of criteria are already reflected in the vectors representing all the alternatives, PIS and NIS in the calculation of relative closeness to the ideal solution.

In similarity based TOPSIS [4] similarity is used to compare the alternatives with both ideal solutions. Later a generalized version of the similarity based TOPSIS was developed [5], where the aggregation of similarities was done using Bonferroni mean [6].

TOPSIS has not been examined much in connection with OWA operators [7] and to our knowledge similarity based TOPSIS variants with OWA aggregation do not exist in earlier literature. Chen et al. [8] examined OWA operator together with standard TOPSIS and used OWA in both internal and external aggregation. Wang et al. [9] developed OWA-TOPSIS approach in intuitionistic fuzzy environment. There OWA was used to aggregate preference and source and to calculate the distance; overall six different types of information aggregation processes are analysed in the paper. Liu et al. [10] used OWA operators to create additive reciprocal matrices to be used as ideal solutions for TOPSIS. Also Yusoff et al.[11] applied Minkowski OWA distance to aggregate distances to positive and negative ideal solutions. However none of these OWA TOPSIS combinations consider aggregating the information on differences of values representing the alternatives under separate criteria into an overall distance or similarity with respect to positive and negative ideal solutions separately by posing different (possibly linguistic) requirements for the distance from or similarity to PIS and NIS.

Intuitively approaching the distance/similarity to PIS and NIS in a different way seems reasonable since if we want to pose a requirement as 'most' of the criteria should have highly similar values for the positive ideal solution and the alternative in question for the alternative to be considered similar to PIS or to be desirable, it is highly unlikely that we want to do the same with this alternative and negative ideal solution and still call it desirable; on the other hand we might require only 'a few' criteria having highly similar values for the NIS and the alternative in question to consider the alternative similar to NIS. Also notable is that if you apply same linguistic weights derived from OWA to aggregate both distances (to PIS and NIS) eventhough relative closeness values differ actual rankings usually does not show statistically

The research was supported by the Finnish Strategic Research Council, grant number 313396 / MFG40 - Manufacturing 4.0, and by LUT research platform AMBI- Analytics-based management for business and manufacturing industry.

significant differences [8].

Here we introduce similarity based TOPSIS with OWA operators with two different motivations behind this. One motivation is that with OWA operator we are able to make linguistic quantifications like 'at most half' 'almost all' or 'at least two'. All this can be done without expressing preference on which of the criteria are required to satisfy these needs. Besides this it is unlikely that we want to have high similarity on e.g. 'most' criteria to be met for NIS even though for PIS this would clearly be desirable. One could, for example, expect, that a careful decision-maker would require 'at least a few' high similarities with NIS across all the criteria to consider the alternative in question *similar to NIS*, but the same decision-maker would require 'most' of the values of criteria to be highly similar with PIS for the alternative to be considered *similar to PIS*. On the other hand a overly optimistic (i.e. less careful or more risk-taking) decision maker might consider 'a few' highly similar values of criteria between the alternative in question and PIS to be sufficient to consider it *similar to PIS*, while he/she would require 'almost all' the criteria to have similar values to those of NIS to consider the alternative in question as *similar to NIS*. Other linguistic quantifications that define an alternative similar to PIS and one similar to NIS can be also considered depending on the purpose of the model, the problem being solved and also on the characteristics, preferences and risk attitude of the decision-maker. Customizability in this matter is definitely reasonable and can lead to better fitting decision support using TOPSIS. Hence in this paper we introduce two sets of linguistic weights for OWA separately for PIS and NIS. In context of supplier evaluation this kind of requirement is at least as important as simple weighting of criterions of their importance.

II. Preliminaries

Yager [7] defined ordered weighted averaging (OWA) operator as follows.

Definition 1: An ordered weighted averaging (OWA) operator of dimension n is a mapping $F : \mathbb{R}^n \longrightarrow \mathbb{R}$, that has an associated weighting vector such that $w_i \in [0, 1]$, $1 \leq i \leq n$, $\sum_{i=1}^{n} w_i = 1$

$$F(a_1, ..., a_n) = \sum_{j=1}^{n} w_j b_j = w_1 b_1 + ... + w_n b_n, \qquad (1)$$

where b_j is the j-th largest element of the collection of objects $a_1, a_2, ..., a_n$.

In our research we are interested in linguistic quantification of weights, or to be more precise of the quantification of the amount of criteria that need to be fulfilled/satisfied sufficiently. With linguistic quantification we mean terms like 'at least some' of the criteria, 'almost all' criteria etc. For this purpose quantifier guided aggregation with OWA operators was established in [12], [13]. One field of quantifiers is called RIM quantifier [12] which is defined as follows.

Definition 2: A fuzzy subset Q of the unit interval is called a Regular Increasing Monotone (RIM) quantifier, if it satisfies the following conditions

1) $Q(0) = 0$,
2) $Q(1) = 1$,
3) $Q(x) \geq Q(y)$, if $x > y$.

The RIM quantifiers can be used to express terms like 'all', 'most', 'many' and 'at least k', where k is an integer number. Often used quantifier is $Q(x) = x^{\alpha}, \alpha \geq 0$ where the weights are calculated as follows

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), \quad i = 1, ..., n. \qquad (2)$$

By using this with proper selection of α we are able to model different types of linguistic terms as described in Table I which is following Yager's example.

Note that the linguistic quantifiers listed in Table I and represented by vectors of OWA weights in fact all have the 'at least' interpretation, in other words we are defining the quantifiers 'at least one', 'at least few', 'at least some', 'at least many', 'at least most' and 'at least all' in Table I. Even though Yager does not directly specify so in [7]. The reason for this might be that the linguistic labels without the 'at least' part are easier to understand and thus to be used by decision-makers, and also for example 'at least all' is identical with 'all'. We, however, consider it important to point out that the linguistic quantifiers defined in Table I *do not* represent 'just one', 'just few', 'just some' and so on. We need to stress that the use of these quantifiers (and the respective OWA weight vectors) does not guarantee that only the specified amount of criteria will be satisfied sufficiently. It is possible that, for example, all the criteria will be satisfied to a high degree even if we use the 'few' quantifier. We therefore strongly suggest to keep the 'at least ...' meaning of the quantifiers in mind when using them.

TABLE I
WEIGHTS WITH DIFFERENT LINGUISTIC QUANTIFIERS

Weight	At least one	Few	Some	Many	Most	All
α	$\alpha \to 0$	0.1	0.5	2	10	$\alpha \to \infty$
w_1	1	0.8513	0.4472	0.04	0	0
w_2	0	0.0611	0.1852	0.12	0.0001	0
w_3	0	0.0378	0.1421	0.20	0.0059	0
w_4	0	0.0277	0.1198	0.28	0.1013	0
w_5	0	0.0221	0.1056	0.36	0.8926	1

III. Method

To apply similarity based TOPSIS with OWA operator we require a specification of the decision matrix for a set of alternatives over a set of criteria. Given a set of m alternatives $A = \{a_i | i = 1, 2, \ldots, m\}$, a set of n criteria $C = \{c_j | j = 1, 2, \ldots, n\}$ and a set of weigths $W = \{w_j | j = 1, 2, \ldots, n\}$, $w_j > 0$, $\sum_{j=1}^{n} w_j = 1$, where w_j denotes the weight of the criterion c_j, let $X = \{x_{ij} | i = 1, 2, \ldots, m, j = 1, 2, \ldots, n\}$ denote the decision matrix where x_{ij} is the performance measure of the alternative a_i with respect to the criterion c_j. Besides this we will also be using two sets of OWA

operator weights $W^+ = \{w_j^+ | j = 1, 2, \ldots, n\}$ used in the context of the similarity of an alternative to PIS and $W^- = \{w_j^- | j = 1, 2, \ldots, n\}$ used in the context of the similarity of an alternative to NIS. Given the decision matrix, the similarity based TOPSIS with OWA involves following steps.

1. Normalize the decision matrix into a unit interval.

$$z_{ij} = \frac{x_{ij} - min_i(x_{ij})}{max_i(x_{ij}) - min_i(x_{ij})}, \quad (3)$$

$$i = 1, 2, \cdots, m, j = 1, 2, \cdots, n$$

2. Compute the weighted normalized decision matrix $V = [v_{ij}]$:

$$v_{ij} = z_{ij} w_j, i = 1, \cdots, m, j = 1, \cdots, n \quad (4)$$

3. Determine the positive and negative ideal solutions

$$V^+ = [v_1^+, \ldots, v_m^+] \text{ such that}$$
$$v_j^+ = \begin{cases} \max_i v_{ij} \text{ if } j \in B \\ \min_i v_{ij} \text{ if } j \in C \end{cases} \quad (5)$$

$$V^- = [v_1^-, \ldots, v_m^-] \text{ such that}$$
$$v_j^- = \begin{cases} \min_i v_{ij} \text{ if } j \in B \\ \max_i v_{ij} \text{ if } j \in C \end{cases} \quad (6)$$

where B is the set of indices of benefit-type criteria, and C is the set of indices of cost-type criteria. Note that given the normalization (3) the definitions of V^+ and V^- can be simplified, as we know that for any $j = 1, \ldots, n$ it holds that:

- $\max_i v_{ij} = 1 \cdot w_j$
- $\min_i v_{ij} = 0$

This means that it is sufficient to know the orientation of the criteria (cost/benefit type) in order to be able to define the PIS and NIS in this alternative of TOPSIS. For example if we consider five criteria $c_1, \ldots, c_5$ such that $1, 2, 5 \in B$ and $3, 4 \in C$ then under (3) we automatically get $V^+ = [w_1, w_2, 0, 0, w_5]$ and $V^- = [0, 0, w_3, w_4, 0]$ regardless of the actual performance values of the alternatives.

4. Compute OWA operator weights by using suitable quantifier for the linguistic requirement of aggregation. Since linguistic requirement for similarity toward positive ideal solution is clearly different than to negative ideal solution we can derive two sets of weights with different linguistic requirements for an alternative to be considered similar to PIS or to NIS.

$$w_j^+ = Q_1\left(\frac{j}{n}\right) - Q_1\left(\frac{j-1}{n}\right), \quad j = 1, \cdots, n \quad (7)$$

$$w_j^- = Q_2\left(\frac{j}{n}\right) - Q_2\left(\frac{j-1}{n}\right), \quad j = 1, \cdots, n \quad (8)$$

where Q_1 and Q_2 denotes RIM functions for different linguistic requirements.

5. Compute similarity vector for each alternative a_i w.r.t. positive ideal solution (i.e. the vectors $[s_{i1}^+, \ldots, s_{in}^+]$, $i = 1, \ldots, m$) and negative ideal solution (i.e. the vectors $[s_{i1}^-, \ldots, s_{in}^-]$, $i = 1, \ldots, m$):

$$s_{ij}^+ = \sqrt[p]{1 - |(v_{ij})^p - (v_j^+)^p|}, i = 1, \cdots, m, j = 1, \cdots, n \quad (9)$$

$$s_{ij}^- = \sqrt[p]{1 - |(v_{ij})^p - (v_j^-)^p|}, i = 1, \cdots, m, j = 1, \cdots, n \quad (10)$$

Here $\hat{s}_i^+ = [s_{i1}^+, s_{i2}^+, \cdots, s_{in}^+]$ denotes the similarity vector of the alternative a_i with the positive ideal solution and $\hat{s}_i^- = [s_{i1}^-, s_{i2}^-, \cdots, s_{in}^-]$ denotes the similarity vector of the alternative a_i with the negative ideal solution.

Theorem 1: Under the normalization (3), if $p = 1$ then

$$s_{ij}^+ + s_{ij}^- = 2 - w_j \text{ for any } i = 1, \ldots, m \text{ and } j = 1, \ldots, n. \quad (11)$$

Proof 1: Since (3) normalizes the values in the decision-matrix into a unit interval, either $v_j^+ = 1 \cdot w_j \ \wedge v_j^- = 0$ or $v_j^+ = 0 \ \wedge v_j^- = 1 \cdot w_j$ for any $j = 1, \ldots, n$. Either way we get

$$s_{ij}^+ + s_{ij}^- = 1 - |v_{ij} - 0| + 1 - |v_{ij} - w_j| = \\ = 1 - v_{ij} + 1 - (w_j - v_{ij}) = 2 - w_j,$$

because $v_{i,j} \in [0, w_j]$ for any $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

6. Compute the similarity of each alternative w.r.t. positive ideal solution and negative ideal solution by aggregating the respective similarity vector using the OWA operator. This aggregation can reflect the requirements on how many of the criteria need to have high similarity for the alternative and PIS (or NIS) for the alternative to be considered 'similar to PIS' (or 'similar to NIS'). These requirements can be expressed using the linguistic quantifiers summarized in Table I.

$$s_i^+ = \sum_{j=1}^{n} w_j^+ b_j^+ \quad (12)$$

where b_j^+ is the j^{th} largest element of $\hat{s}_i^+ = [s_{i1}^+, s_{i2}^+, \cdots, s_{in}^+]$.
Similarly

$$s_i^- = \sum_{j=1}^{n} w_j^- b_j^- \quad (13)$$

where b_j^- is the j^{th} largest element of $\hat{s}_i^- = [s_{i1}^-, s_{i2}^-, \cdots, s_{in}^-]$.

Note that it is the W^+ OWA weights that reflect the requirements for the alternative to be considered similar to PIS in terms of the linguistically quantified (described) minimum number of criteria with respect to which the alternative needs to be similar with PIS for the alternative to be considered '*similar to PIS overall*'. On the other hand the W^- OWA weights reflect the requirements for the alternative to be considered similar to NIS in terms of the linguistically quantified (described) minimum number of criteria with respect to which the alternative needs to be similar with NIS for it to be considered '*similar to NIS overall*'.

7. Compute the relative closeness of the alternative to the positive ideal solution:

$$RC_i = \frac{s_i^+}{s_i^+ + s_i^-}, i = 1, \cdots, n \tag{14}$$

The definition of RC_i by (14) does not guarantee that full similarity with PIS ($s_i^+ = 1$) would imply that $RC_i = 1$ by itself. Also full similarity with NIS ($s_i^- = 1$) does not mean that $RC_i = 0$. However, zero similarity with NIS ($s_i^- = 0$) does imply that $RC_i = 1$ regardless of the actual similarity with PIS. Still increasing the similarity with PIS (s_i^+) increases RC_i while increasing similarity with NIS (s_i^-) decreases the value of RC_i. It can therefore be considered a reasonable value for the ranking of alternatives.

8. Arrange the ranking indexes in a descending order with respect to the values of RC_i to obtain the best alternative.

The above proposed method differs from original TOPSIS in four ways:

1) The normalization is done to unit interval unlike in the original version of TOPSIS. This simplifies the definition of PIS and NIS and makes it independent on the actual performance of the alternatives w.r.t. the criteria. It is sufficient to know the orientation of the criteria to be able to define PIS and NIS[1].
2) The computation of how similar alternatives and ideal vectors (PIS and NIS) are is done using a similarity measure instead of a distance measure. This fully introduces the concept of similarity into a method that has a 'similarity to ideal solution' in its very name.
3) The aggregation of similarity vectors (criteria-wise similarities to PIS and NIS) is done using two different ordered weighted averaging operators. This allows different linguistic quantifications of "how similar an alternative needs to be to PIS or NIS criteria-wise to be considered 'overall similar to PIS or NIS' respectively". This opens doors for the reflection of the risk-preference of the decision-maker and for more detailed specification of the requirements on a 'alternative similar to PIS' and an 'alternative similar to NIS' by the decision-maker. The proposed method uses linguistic quantification for this purpose to facilitate the expression of these requirements for a wide set of decision-makers.
4) The relative closeness computation is adjusted to similarity measures instead of the distance measures that are used in the calculation of the relative closeness to the ideal solution in [3].

[1]In original method normalization is done as $z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$.

In the following example we will illustrate the performance of the method proposed in this paper on a supplier evaluation and selection problem. We will also show how different requirements on the similarity to PIS and NIS (potentially representing the risk-attitude of the decision-maker) can be reflected, and discuss how this can influence the results obtained by the method.

IV. Supplier evaluation

Here we introduce how we can use similarity based TOPSIS with OWA operators in supplier evaluation. Our basic problem is the following. A car manufacturing company wants to select its supplier. Most important criteria which manufacturer selected to focus on are: price, duration of the project, quality, the amount of equipment and distance. From these quality and the amount of equipment are considered to be benefit-type criteria and others are considered to be cost-type criteria. After preliminary selection, five suppliers remain and the decision matrix given in Table II is obtained.

Linguistic assessments for quality and equipment are transformed into numerical scale between $[0, 10]$ resulting in Table III.

The first step is to calculate the normalized decision matrix which can be found in Table IV.

In this problem we consider all the criteria to be equally important leading to the weighting expressed by $W = [1, \ldots, 1]$ to be redundant. Even though the original method requires

TABLE II
ORIGINAL DECISION MATRIX

	Price(c_1)	Time (c_2)	Quality (c_3)	Equipment (c_4)	Distance (c_5)
Supplier 1	80	12	very good	good	260
Supplier 2	75	14	very good	very good	230
Supplier 3	72	13	good	medium	50
Supplier 4	65	15	medium	medium	140
Supplier 5	78	13	very good	medium	180

TABLE III
DECISION MATRIX ON NUMERICAL SCALE

	Price(c_1)	Time (c_2)	Quality (c_3)	Equipment (c_4)	Distance (c_5)
Supplier 1	80	12	9	7	260
Supplier 2	75	14	9	9	230
Supplier 3	72	13	7	5	50
Supplier 4	65	15	5	5	140
Supplier 5	78	13	9	5	180

TABLE IV
NORMALIZED DECISION MATRIX

	Price(c_1)	Time (c_2)	Quality (c_3)	Equipment (c_4)	Distance (c_5)
Supplier 1	1	0	1	0.5	1
Supplier 2	0.67	0.67	1	1	0.86
Supplier 3	0.47	0.33	0.5	0	0
Supplier 4	0	1	0	0	0.43
Supplier 5	0.87	0.33	1	0	0.62

normalized weights, the final ordering of the alternatives (suppliers in this example) will not change if we do not normalize the weights and with the weighting vector $W = [1, 1, 1, 1, 1]$ the calculations will be easier to follow and will thus serve better as an example of the proposed method. This does not limit the applicability of the results presented further. It allows us to see the effects of linguistic quantification in the method more clearly. Given the fact that c_1, c_2 and c_5 are cost-type criteria and c_3 and c_4 are benefit-type criteria, we get the positive ideal solution in the form of $V^+ = [0, 0, 1, 1, 0]$ and the negative ideal solution if the form of $V^- = [1, 1, 0, 0, 1]$.

The similarity vectors of the alternatives to PIS (vectors $\hat{s}_i^+$, $i = 1, \ldots, 5$) and to NIS (vectors $\hat{s}_i^-$, $i = 1, \ldots, 5$) are presented in Tables V and IV.

TABLE V
SIMILARITY VECTORS $\hat{s}_i^+ = [s_{i1}^+, \ldots, s_{i5}^+]$ OF THE ALTERNATIVES TO PIS REPRESENTED BY $V^+ = [0, 0, 1, 1, 0]$

	Price(c_1)	Time (c_2)	Quality (c_3)	Equipment (c_4)	Distance (c_5)
Supplier 1	0.000	1.000	1.000	0.500	0.000
Supplier 2	0.333	0.333	1.000	1.000	0.143
Supplier 3	0.533	0.667	0.500	0.000	1.000
Supplier 4	1.000	0.000	0.000	0.000	0.571
Supplier 5	0.133	0.667	1.000	0.000	0.381

TABLE VI
SIMILARITY VECTORS $\hat{s}_i^- = [s_{i1}^-, \ldots, s_{i5}^-]$ OF THE ALTERNATIVES TO NIS REPRESENTED BY $V^- = [1, 1, 0, 0, 1]$

	Price(c_1)	Time (c_2)	Quality (c_3)	Equipment (c_4)	Distance (c_5)
Supplier 1	1.000	0.000	0.000	0.500	1.000
Supplier 2	0.667	0.667	0.000	0.000	0.857
Supplier 3	0.467	0.333	0.500	1.000	0.000
Supplier 4	1.000	0.000	0.000	1.000	0.429
Supplier 5	0.867	0.333	0.000	1.000	0.619

We used Regular Increasing Monotone (RIM) type of quantifier guided aggregation in order to emphasize importance of getting high ratings from at least some criteria (or to be more precise to express in how many criteria we need to find high similarity with the respective ideal to consider the whole alternative similar to the ideal). We decided to use the exponential function $Q(x) = x^\alpha$ as our monotonic function. This choice here is simply based on the fact that it is the most commonly used quantifier function in the literature. Next we need to set up linguistic requirement for suppliers similarity towards PIS and NIS.

We will be considering three different cases representing different types of decision-makers:

Careful decision-maker

This decision-maker is rather pessimistic. To consider an alternative to be similar to PIS he/she requires high similarity in (at least) 'many' criteria between the given alternative and PIS. Note, that it is not specified in which criteria the similarity needs to be found. On the other hand similarity with NIS in (at least) 'few' criteria is considered enough by this decision-maker to consider the alternative similar to NIS (See Table I for linguistic evaluations). In other words this decision maker requires more strong evidence of high qualities of the given alternative to consider it good (similar to PIS), while some evidence of its badness is enough to consider it bad (similar to NIS). Such a behavior could be considered close to risk avoidance. This setup means, that 'many' will be represented by the following OWA weights for the calculation of the overall similarity to PIS:

$$w^+ = [0.04, 0.12, 0.2, 0.28, 0.36]$$

derived using the equation (7) and the value of $\alpha = 2$. The OWA weights used for the calculation of the overall similarity to NIS are calculated using (8) and the value of $\alpha = 0.1$:

$$w^- = [0.8513, 0.0611, 0.0378, 0.0277, 0.0221].$$

Using these weights we can next calculate similarities of each alternative to positive ideal solution and to negative ideal solution. These can be found in Table VII. Similarly on the fourth column relative closeness values have been computed.

Based on the relative closeness values we get the ordering of suppliers to be $2 \succ 3 \succ 1 \succ 5 \succ 4$ meaning best choice of a supplier would be supplier 2, the second best choice is supplier 3 etc.

TABLE VII
SIMILARITIES OF THE SUPPLIERS TO PIS AND NIS AND THE VALUES OF RELATIVE CLOSENESS OF THE SUPPLIERS TO PIS - THE CASE OF A CAREFUL (RISK-AVOIDING) DECISION-MAKER

Attribute	s^+	s^-	RC
Supplier 1	0.26	0.93	0.22
Supplier 2	0.37	0.80	0.32
Supplier 3	0.37	0.91	0.29
Supplier 4	0.11	0.96	0.10
Supplier 5	0.23	0.94	0.20

Optimistic decision-maker

This decision-maker is much more willing to evaluate an alternative as good (similar to PIS) when its performance is similar with the performance of PIS in (at least) 'few' criteria. On the other hand to consider an alternative to be bad (similar to NIS) it would have to be similar to NIS in (at least) 'many' criteria. This approach can be considered close to risk-seeking. In this case:

$$w^+ = [0.8513, 0.0611, 0.0378, 0.0277, 0.0221],$$

$$w^- = [0.04, 0.12, 0.2, 0.28, 0.36]$$

The respective results can be found in Table VIII. We can see that in this more benevolent approach the relative closeness of all the alternatives to PIS is much larger than in the case of the careful decision-maker. The suggested ordering of suppliers in this case would be $2 \succ 1 \succ 3 \succ 5 \succ 4$. We can see that while the most promising supplier remained the same, the runner up has changed from supplier 3 to

TABLE VIII
SIMILARITIES OF THE SUPPLIERS TO PIS AND NIS AND THE VALUES OF RELATIVE CLOSENESS OF THE SUPPLIERS TO PIS - THE CASE OF AN OPTIMISTIC (RISK-SEEKING) DECISION-MAKER

Attribute	s^+	s^-	RC
Supplier 1	0.93	0.26	0.78
Supplier 2	0.94	0.25	0.79
Supplier 3	0.93	0.29	0.76
Supplier 4	0.89	0.48	0.65
Supplier 5	0.91	0.36	0.72

supplier 1. This is the result of the best performance of supplier 1 in the two of the criteria. This ranking is more focused on the potential of the suppliers.

Ignorant (risk-indifferent) decision-maker

This decision-maker treats the similarity to PIS and to NIS in an identical way - for the alternative to be considered similar to PIS or to NIS its performance has to be similar with the given ideal in (at least) 'many' criteria. This approach is the closest to the original TOPSIS as it calculates the similarity to PIS and NIS in the same way. In this case

$$w^+ = w^- = [0.04, 0.12, 0.2, 0.28, 0.36]$$

TABLE IX
SIMILARITIES OF THE SUPPLIERS TO PIS AND NIS AND THE VALUES OF RELATIVE CLOSENESS OF THE SUPPLIERS TO PIS - THE CASE OF AN IGNORANT (RISK-INDIFFERENT) DECISION-MAKER

Attribute	s^+	s^-	RC
Supplier 1	0.26	0.26	0.50
Supplier 2	0.37	0.25	0.60
Supplier 3	0.37	0.29	0.56
Supplier 4	0.11	0.48	0.18
Supplier 5	0.23	0.36	0.39

The respective results can be found in Table IX. We can see that in this approach that treats both similarities to PIS and NIS in the same way we are getting the same final ordering as with the careful decision-maker. The suggested ordering of suppliers in this case would be $2 \succ 3 \succ 1 \succ 5 \succ 4$. The overall relative closeness to PIS is much larger for all the alternative than with the careful decision-maker, but not as large as with the pessimistic one. This is due to the fact that the similarities of the alternatives to NIS are much smaller than with the careful decision-maker. The reason for this being that the ignorant/indifferent decision-maker requires the performance in 'many' criteria to be similar with the performance of NIS for the alternative to be considered similar to NIS and thus to lower the respective value of RC_i.

Obviously, there are many other possible choices of linguistic quantifications for the definition of overall similarity with PIS and NIS. In this paper we will focus on just these three.

We have, however, examined whether internal aggregations for all the combinations of linguistic quantifiers are different from each other by using Friedman's test. This is inline with [8] who studied aggregation of multiple experts with different OWA weights. The hypothesis in this case is

H_0: The 49 rankings (combinations from $\alpha_1 = 0, 0.1, 0.5, 1, 2, 10, 1000$ and $\alpha_2 = 0, 0.1, 0.5, 1, 2, 10, 1000$) of five alternatives are the same.

H_1: At least two rankings are different

TABLE X
FRIEDMAN TEST RESULT

χ	55.58
df	4
p	$2.4584e^{-11}$

From the Friedman's test results we can conclude that the results are highly significant showing that by posing different linguistic requirements on similarity of supplier w.r.t. positive and negative ideal solutions it is possible to get significantly different ranking orders. These requirements need to reflect the needs and preferences (and potentially also the risk-attitude) of the decision-maker well, as for different linguistically quantified requirements we can get significantly different rankings of the alternatives.

V. CONCLUSIONS

In this paper similarity based TOPSIS with OWA operator is introduced. The advantage of using OWA operator in aggregation of similarities is that we are able to model such linguistic requirements as similarity should be high to at least some (at least half, most) criteria without needing to specify to which criteria. This changes decision making procedure clearly compared to situation where similarities/distances to particular criteria are needed to specify. Often in real world cases analysis requirements as 'at least some', 'at least half', 'most' are more suitable to practical problem at hand than the need to emphasize particular criteria. For this purpose similarity based TOPSIS with OWA operator is designed. Besides this it allows for the expression of preference/needs of the particular decision-maker with respect to what should be considered similar to PIS and NIS. From the presented examples it is clear that the same linguistic requirement may not be suitable for modeling requirements for both similarities toward PIS and NIS. For this purpose we allow the use of two different linguistic quantifiers reflecting the requirements of the decision-maker. We demonstrate the method by applying it to supplier selection problem for car manufacturing company in the context of three different types of decision-makers. Here we managed to show that different ranking orders can be gained which reflect of decision makers attitude towards situation at hand.

REFERENCES

[1] V. Traneva, S. Tranev, D. Mavrov, Interval-Valued Intuitionistic Fuzzy Decision-Making Method using Index Matrices and Application in Outsourcing, Annals of Computer Science and Information systems, 25, pp. 251-254, 2021, http://dx.doi.org/10.15439/2021F77

[2] A. Karczmarczyk, J. Jankowski, J. Watrobski, Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks, Annals of Computer Science and Inofrmation systems, 18, pp. 663-673, 2019, http://dx.doi.org/10.15439/2019F199

[3] C.L. Hwang, K. Yoon, Multiple Attributes Decision Making Methods and Applications, 1980, Springer, Berlin, Heidelberg

[4] P. Luukka, M. Collan, Histogram ranking with generalized similarity-based TOPSIS applied to patent ranking, International Journal of Operational Research 25 (4), 2016, pp. 437-448, http://dx.doi.org/10.1504/IJOR.2016.075290

[5] P. Luukka, M. Collan, Bonferroni mean based similarity based TOPSIS, 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 704-709, 2016, http://dx.doi.org/10.1109/FUZZ-IEEE.2016.7737756

[6] C. Bonferroni, Sulle medie multiple di potenze, Bollettino Matematica Italiana 5, 1950, pp. 267-270.

[7] R. R. Yager, On ordered weighted averaging aggregation operators in multi-criteria decision-making, Systems, Man and Cybernetics, IEEE Transactions, Vol. 18, 1988, pp. 183-190, 1988, http://dx.doi.org/10.1109/21.87068

[8] Y. Chen, K.W. Li, S. Liu, An OWA-TOPSIS method for multiple criteria decision analysis, Expert Systems with Applications, 38, pp. 5205-5211, 2011, http://dx.doi.org/10.1016/j.eswa.2010.10.039

[9] T. Wang, J. Liu, J. Li, C. Niu, An integrating OWA-TOPSIS framework in intuitionistic fuzzy settings for multiple attribute decision making, Computers & Industrial Engineering, 98, pp. 185-194, 2016, http://dx.doi.org/10.1016/j.cie.2016.05.029

[10] F. Liu, Y-.F. Shang, L-.H. Pan, A modified TOPSIS method for obtaining the associated weights of the OWA-type operators, International journal of intelligent systems, 30, pp. 1101-1116, 2015, http://dx.doi.org/10.1002/int.21737

[11] B. Yusoff, J.M. Merigo, D.C. Hornero, Generalized OWA-TOPSIS model based on the concept of majority opinion for group decision making, FIM 2015, AISC 730, pp. 124-139, 2018.

[12] R. R. Yager, Quantifier Guided Aggregation using OWA Operators, International Journal of Intelligent Systems, Vol. 11, pp. 49-73, 1996, http://dx.doi.org/10.1002/(SICI)1098-111X(199601)11:1<49::AID-INT3>3.0.CO;2-Z

[13] Z. S. Xu, An Overview of Methods for Determining OWA Weights, International Journal of Intelligent Systems, Vol. 20, pp. 843-865, 2005, http://dx.doi.org/10.1002/int.20097

KNOWCON
Knowledge on Economics and Management
International Scientific Conference

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 53–61

DOI: 10.15439/2021B3
ISSN 2300-5963 ACSIS, Vol. 29

Interval-valued semantic differential in multiple criteria and multi-expert evaluation context: possible benefits and application areas

Jana Stoklasová
School of Business and Management,
LUT University Yliopistonkatu 34,
53851 Lappeenranta, Finland
Email: jana.stoklasova@lut.fi

***Abstract*—The paper discusses the possibilities of adapting the recently introduced interval-valued semantic differential method to the multiple-criteria decision-making and evaluation context. It focuses on the differences and common ground of the intended use of the original semantic differentiation method and general multiple-criteria evaluation problems. The paper identifies the aspects of the interval-valued modification of the method that can be useful in multiple-criteria evaluation and also aspects that can be beneficial in the multi-expert evaluation setting and also possible limitations stemming from the transition to the multiple-criteria (or multi-expert) evaluation context. Finally the paper suggests potential application areas for the (interval-valued) semantic differential based methods.**

I. Introduction

THE set of methods available for multiple-criteria and multi-expert evaluation problems is large and it is being continuously expanded (see e.g. [1], [2], [3], [4], [5], [6]). The currently available methods include methods for weights determination (see e.g. [7], [8], [9]), methods for the standardization of values of criteria, various methods for the aggregation of values across different criteria [10], [11], [12], [13], methods for preference representation [14], [15], [16] and aggregation [17], [18], [19]. We have specific methods based on pairwise comparisons (see e.g. [20], [21], [22]), methods utilizing ideals in the evaluation process [23], [24], special methods for ordinal data [25], [26], [27], methods equipped to deal with different types of uncertainty [28], [29], [30], [31], [32], [33], [34], methods capable of dealing with linguistic inputs/outputs and to process natural language [35], [36], [37], [38], methods for consensus modeling and analysis [39], [40], [27], [41]. The list is definitely not complete, nor is it reasonably structured. The point we would like to make here is that currently there are many methods available to model and assist with human-like decision making. They focus on different aspects of the evaluation in these problems and are able to reflect many different specific features of the decision-makers, of the alternatives, of the scales used for the evaluation etc. The behavioral perspective has entered the multiple-criteria and multi-expert evaluation and decision-making domain long ago and is still gaining momentum [42], [43], [44], [45], [46], [47], [48], [49], [50]. Most of the methods assume at least some sort of measurability of the features of the alternatives that are being evaluated, or of the values of the criteria that are being used in the process; some circumvent the requirement of measurability by pairwise comparisons, by the use of linguistic assessments, by the use of ordinal values only etc. There are, however, very few methods in the multiple-criteria and multi-expert evaluation field that would be focused or tailored for dealing with intangible criteria.

The research was supported by LUT research platform AMBI - Analytics-based management for business and manufacturing industry.

Even though current research is aiming also on the ability of computers and models to recognize, process, mimic and interpret emotions [51], the efforts to incorporate affects and other less tangible criteria in evaluation models are limited. This might be stemming from the difficulties with measuring or obtaining the information on affect and other less tangible concepts like attitudes, political preferences, religion, values, connotative meaning of words etc. On the other hand there are methods in psychology, anthropology, linguistics and related fields that are designed for the very purpose of capturing non-measurable and intangible concepts. One of these methods, the semantic differential method by Osgood, Suci and Tannenbaum [52] is going to be investigated in this paper. We will describe the main principles of this method, briefly recall its recent interval-valued generalization [53], [54] and identify how the concepts intended for the capturing of intangible characteristics can be applied in the multiple-criteria evaluation and multi-expert evaluation setting. We will particularly focus on those aspect that are crucial in the original definition of this tool and have psychological value (such as partial projectivity, the requirement on the bipolar adjectives scales being non-descriptive for the evaluated alternative/concept, etc.) and their meaningfulness, usefulness or potential drawbacks if transferred directly into the multiple-criteria evaluation setting. Our aim is to identify those features of the semantic differentiation method (and its interval-valued generalization) rooted in its original social-science use, that can be beneficial in multiple-criteria evaluation models.

II. Semantic differential and its main features

Semantic differential (SD) is a method introduced by Osgood, Suci and Tannenbaum in 1957 [52] as a technique for the quantification or representation of connotative meaning of words. Soon enough it found its way to anthropology [55] and obviously also to psychology for the measurement (quantification) of attitudes [56], [57].

The basic tool in the method are bipolar adjective scales that are used for the assessment of the given object/term/concept (the bipolar adjective scales will also be called items in the text for more simplicity). These scales are the basic "measurement" instrument in the method. The scales are assumed to share the same universe, let us say $[a, b] \subset \mathbb{R}$. Some authors suggest that $0 \in [a, b]$, some suggest that $a > 0$, some that $a = -b$, but the actual form of the scale influences mainly the comfort and reliability of the respondent's answer. Let us now assume that the underlying scale is a continuum with extremes a and b representing the opposite poles of the scale. Originally, discrete (7-point) scales were used in [52] but the actual form of the scale was more tailored for that time's methods of data collection and analysis. The transition to continuous scales is of no actual consequence for the design and performance of the semantic differential method. In other words we can also use continuous scales instead of discrete ones and the method works as well.

The method targets the less tangible aspects of the evaluated object/concept, that is, it intends to capture the connotative (individual-specific) meaning of the concept, reflect the individual's experience and specifics. In social psychology the ability to capture not-measurable aspects connected with the assessed concept led to the use of semantic differential in the quantification of attitudes (mainly in the three-factor model of attitudes where attitudes are assumed to have cognitive, conative and affective components - the latter two being difficult to directly measure). It is therefore suggested by Osgood et al. ([52]) to avoid such bipolar-adjective scales that would have actual descriptive power over the evaluated object. Note that this is a very particular requirement for a method that should be considered for multiple-criteria evaluation. There are, however, good psychological reasons behind this requirement. First the use of descriptive items (e.g. sharp-blunt for the description of a knife) and non-descriptive items (e.g. happy-sad for the description of the same knife) together in one assessment tool (inventory or set of bipolar scales) could result in a lower reliability of the non-descriptive items. The respondents might simply wonder whether they understand the evaluation task well as some items have clear connection with the evaluated concept while others do not. Second the use of descriptive items provides a description of the object/concept rather than its actual evaluation. Third the use of non-descriptive scales decreases respondents' ability to provide desirable, "fake-good" or "fake-bad" answers, which decreases the potential deliberate distortion of information by the respondent/evaluator. The whole procedure of using a semantic differential in the assessment of connotative meanings of concepts or the assessment of attitudes of the respondents towards these concepts can be summarized in the following steps (more details can be found in [52]):

1) Generate a set $S = \{s_1, \ldots, s_n\}$ of bipolar-adjective scales. It should contain sufficiently many scales, the meanings of their endpoints should be understandable to the potential evaluators (respondents), enough of these scales should be non-descriptive for the concepts to be evaluated.
2) Carry out pilot run where all these scales are used to assess some concepts by a representative sample of the target population.
3) Carry out a factor analysis (both exploratory and confirmatory versions are suggested) to determine whether the dimensionality of the original set of scales can be reduced to just a few underlying factors. The factors are to be identified ideally in such a way that they could be named and interpreted accordingly (apply factor notation if needed). For example in [52] three factors were identified: Evaluation, Potency and Activity. These factors are expected to represent orthogonal evaluation dimensions. Let us assume that k factors $F_1, \ldots, F_k$ are identified. Then the factor loadings of the scale $s_i \in S$ for factors $F_1, \ldots F_k$ can be denoted $f_{1_{s_i}}, \ldots, f_{k_{s_i}}$ respectively. Note that the factor loadings (and the factors) are therefore domain- and culture-specific. In other words the factor analysis should be performed every time we apply the chosen scales to the evaluation/assessment of concepts in a different context, also when we change the target population. Given the fact that the extreme values (poles) of the scales are described linguistically, every language mutation of the scales should have its own factor analysis performed.
4) Select a subset of the bipolar-adjective scales $Z \subseteq S$, $Z = \{z_1, \ldots, z_m\}$ that would be used for the given application. Usually scales that sufficiently load at least one factor are used, it is also good to use scales that would allow all the factors to be "measured" and also to allow for repeated measurement of each factor.
5) Obtain data from the respondents. In other words let each respondent assess the concept using all m chosen bipolar-adjective scales $z_1, \ldots, z_m$. If the assessment of the concept/object by a respondent X on scale z_i is denoted as x_{z_i}, $i = 1, \ldots, m$, then the object/concept O is represented as a point O^X in the k-dimensional space $[a, b]^k$ with the following coordinates:

$$O^X = \left(\frac{\sum_{i=1}^{m} x_{z_i} \cdot f_{1_{z_i}}}{\sum_{i=1}^{m} |f_{1_{z_i}}|}, \ldots, \frac{\sum_{i=1}^{m} x_{z_i} \cdot f_{k_{z_i}}}{\sum_{i=1}^{m} |f_{k_{z_i}}|} \right) = (x_{F_1}, \ldots, x_{F_k}). \quad (1)$$

In other words the coordinates are the factor-loading weighted average of the answers provided by the respondent. Sometimes only the contribution of the item to the factor with the highest factor loading is reflected in the practical applications of semantic differential.

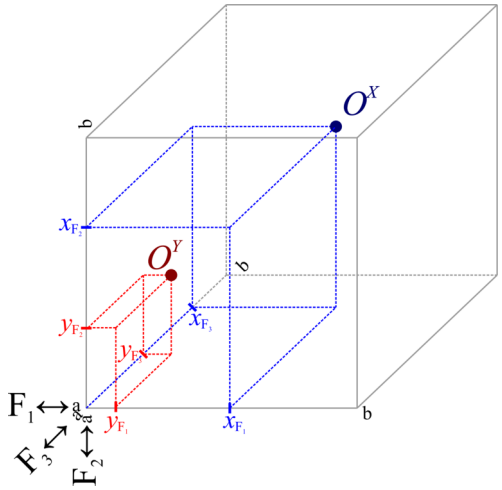


Fig. 1. An example of the output of the standard semantic differential method [52] for two objects/concepts X and Y. Three factors are assumed. The assessment of X is represented by the point $O^X = (x_{F_1}, x_{F_2}, x_{F_3})$ in the three-dimensional space defined by the factors F_1, F_2 and F_3, the assessment of Y is represented by the point $O^Y = (y_{F_1}, y_{F_2}, y_{F_3})$.

The above described procedure allows for the representation of a connotative meaning of a concept or an attitude towards that concept to be represented as a point in an k-dimensional space where each dimension represents one factor (higher level characteristic of the object/concept) that is orthogonal to the other factors. An example of the output of the standard semantic differential method is presented in Fig. 1.

To summarize, the benefits of the method as proposed by Osgood et al. ([52]), assessed from a multiple-criteria evaluation perspective, are the following:

- The factor analysis applied provides a few orthogonal evaluation dimensions to work with. This means that a visualization of the results, that is easy to understand, might be possible.
- The use of non-descriptive bipolar-adjective scales prevents deliberate distortions of the data by respondents.
- The use of bipolar-adjective scales provides a "projective-like" feature of the data collection that in terms allows for the assessment of less tangible criteria/aspects of the concept.
- The fact that more items have non-zero loadings to the same factor means that we have repeated assessment of each factor.
- Data input using the semantic differential scales is rather simple.
- It is possible to define distances in the $[a, b]^k$ space to decide which representations of objects/concepts are close to each other, which are far from each other.
- As long as the factors are defined with appropriate labels and can be seen as consistent characteristics "measured" by multiple items (repeated "measurement"), the coordinates of the concepts in the $[a, b]^k$ space can be interpreted. It is also possible to define "desired" or "undesired" values in this space, that is to define ideals to be used in the evaluation or decision-making.
- There is no need for aggregation across the factors. Aggregation within one factor (1) is understood as repeated measurement of the factor, other aggregation is not necessary. The final representation of the result of semantic differentiation can therefore be understood as virtually lossless. The aggregation across factors, if needed, can also be done, for example, via the definition of the distance from a given ideal in the $[a, b]^k$ space.

It is therefore clear that many features of the semantic differential can be seen as beneficial for the standard multiple-criteria or multi-expert evaluation. On the other hand there are certain clear limitations or drawbacks of the method when considered for practical multiple-criteria evaluation:

- The factor analyses need to be done. As the factors, their number, definition and loadings of the scales can be context and culture dependent, it might take a lot of time to set up the scales and find their factor loadings.
- Also a conversion to other languages and other domains of application requires new factor analyses. The language issues are even more complex than might be apparent at first sight. If the tool is calibrated, for example, for English scales for a given context of application (factor loadings of items are determined with English labels of the endpoints of the bipolar adjective scales) it should still not be directly applied with non-native speakers of English, unless these were present in the original sample used to determine the factors and their loadings.
- The factors are stemming from the factor analysis. They are therefore constructed and might not have clear interpretation. This could limit the interpretability of the results of semantic differential in evaluation applications.
- The issue of concept-scale interaction and lower perceived scale relevance may be present [58]. This means that the respondents might see some scales as inappropriate for the assessment of a given concept and thus the value of the given item provided by them can be arbitrary without the researcher knowing so.
- The method has no means of incorporating uncertainty stemming from lower perceived item relevance for the evaluation of the given object/concept, from the misinterpretation of the meanings of the endpoints of the scales or simply from the inability of the respondent to provide answers using some items because their connection with the assessment might be too value or unclear.
- The single-point in $[a, b]^k$ space might appear much more precise than it should.
- It might not be clear if a "middle" answer means the inability of the respondent to use the given bipolar-adjective scale, or whether his/her assessment is really neutral.

Even though there are clear benefits that speak in favor of the semantic differential being used in multiple-criteria evaluation, there are still some shortcomings that make its use

problematic. Some of these shortcomings can be overcome by generalizing the semantic differential into an interval-valued method as proposed by Stoklasa et al. [54]. The interval-valued methods are being applied in other areas as well [33].

III. Interval-valued generalization of the semantic differential

The generalized semantic differential (GSD) method was introduced in 2019 by Stoklasa, Talášek and Stoklasová with the intention of introducing means for the reflection of uncertainty of the answers provided by the respondents in the form of the x_{z_i} values [54]. GSD assumes that each bipolar adjective scale $z_i \in Z$ is accompanied by another scale r_{z_i} designed to assess the relevance of the scale z_i for the assessment of the given object/concept as perceived by the decision maker. The authors suggest a $[0\%, 100\%]$ universe for each relevance scale r_{z_i} for any $i = 1, \ldots, m$. The term "perceived relevance" can be replaced by any potential source of uncertainty of the values x_{z_i} provided by the respondent/evaluator. The source of uncertainty discussed specifically in [54] is the incompatibility (partial or full) of the bipolar adjective scale with the assessment/evaluation task perceived by the respondent. In other words if the scale is perceived as partially irrelevant by the person who is supposed to use it to assess the given concept, the actual value x_{z_i} is not reliable and should not be considered precise. Due to the partial irrelevance of the scale z_i, the value x_{z_i} might be misspecified by the respondent due to the fact that it was difficult to him/her to established a connection between the evaluated object/concept and the bipolar adjective scale. As such the actually expressed value x_{z_i} is in these cases accompanied by an interval of "also possible values" $I_{z_i} = [x^L_{z_i}, x^R_{z_i}] \subseteq [a, b]$, whose length is proportional to the perceived irrelevance of the scale. Stoklasa et al. [54] suggest the use of Dombi's kappa function [59] to parameterize the calculation of the length of the "interval of also possible values" from the the perceived (ir)relevance of the scale, in other words $\kappa(r_{z_i}) = \left|[x^L_{z_i}, x^R_{z_i}]\right|$. This interval is centered around x_{z_i}, if possible. If this is not possible, then it is shifted so that the shift is minimal, the whole "interval of also possible values of x_{z_i} fits within the $[a, b]$ universe and its length calculated using the kappa function is preserved. The final representation of the output of GSD for an object/concept X is the point O^X in the $[a, b]^k$ space determined from the x_{z_i} values by (1) accompanied by the box of uncertainty B^X (or box of also possible values) surrounding it determined by (2), which is a direct analogy to (1) using interval algebra.

$$B^X = \left(\frac{\sum_{i=1}^{m} I_{z_i} \cdot f_{1_{z_i}}}{\sum_{i=1}^{m} |f_{1_{z_i}}|}, \ldots, \frac{\sum_{i=1}^{m} I_{z_i} \cdot f_{k_{z_i}}}{\sum_{i=1}^{m} |f_{k_{z_i}}|} \right) \quad (2)$$

Interval algebra (see [31, p. 103] for more details) is applied to obtain the final outputs from the generalized semantic differential. This method provides not only the outputs available in the original version of the method - that is the representation of the object/concept X as a point O^X in the k-dimensional Cartesian space - but also a box of uncertainty B^X surrounding the point O^X. See Fig 2 for an example of the outputs of the GSD method. The size of the box of uncertainty is proportional (with respect to the selection of parameters for the kappa function) to the average perceived irrelevance of the bipolar adjective scales used for the assessment of the object/concept. In Fig. 2 it is apparent that the items with high factor loadings for the factor F_1 are perceived as much less relevant for the assessment of X than they are for the assessment of Y.

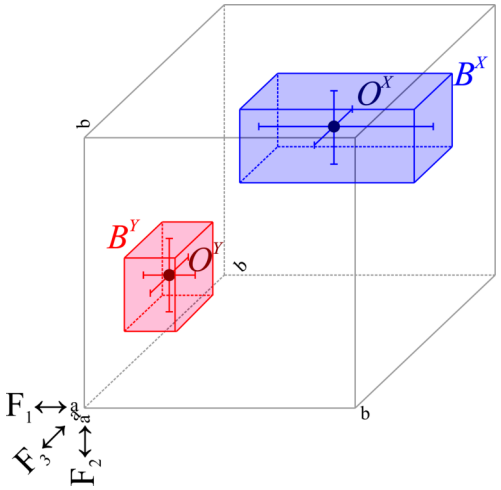


Fig. 2. An example of the output of the generalized semantic differential method [54] for two objects/concepts X and Y. Three factors are assumed. The same two objects X and Y are considered as in Fig. 1 with the same coordinates of O^X and O^Y respectively. Boxes of uncertainty stemming from lower perceived relevance of some scales for the assessment of X and Y are depicted as B^X and B^Y around the O^x and O^Y respectively.

The steps needed to apply GSD are similar to those for SD just with a few minor changes:

1) We again need to have the set $S = \{s_1, \ldots, s_n\}$ of bipolar-adjective scales generated with the same requirements as in SD.
2) We need to administer all those scales to a representative sample of the target population to be able to derive the factors and the factor loadings of the scales (again exploratory and confirmatory factor analysis is recommended). Note that at this point the perceived relevance scales are not used yet. This means that the factors and the factor loadings of the scales are determined independently of the perceived relevance. This also means that if factors and factor loadings are already available for an applicable set of bipolar adjective scales derived for a compatible area of application, these can be used in GSD.
3) We select a subset of the bipolar-adjective scales $Z \subseteq S$, $Z = \{z_1, \ldots, z_m\}$ that will be used for the given application in the same way as for SD. To each of these scales we attach a "perceived relevance" scale r_{z_i}. The r_{z_i} scales, $i = 1, \ldots, m$, are used to capture uncertainty of the x_{z_i} evaluations provided by the respondents. It is possible to label this scale so that it captures different sources of uncertainty of the evaluations as well.

4) We obtain data from the respondents. The values x_{z_i} are assigned "intervals of also possible values" that reflect the irrelevance of the scales and a possible uncertainty of the evaluations stemming from the scale irrelevances (either directly, or through the kappa function).
5) The final representation of the objects/concepts is represented by O^X calculated using equation (1) and by the "box of uncertainty" B^X calculated using the equation (2). Both these representations are depicted graphically (see e.g. Fig. 2).

Allowing the uncertainty in the semantic differentiation process takes care of some of the issues connected with scale-concept interactions. The generalized semantic differential has the same advantages as the original method, plus the ability to reflect uncertainty of the answers provided by the respondents. It can show that the respondent was not very sure about the answers (contributing to particular factors or to all of the factors) by increasing the respective dimension of the "box of uncertainty". As for the disadvantages, the need to perform the factor analyses to get the factors and factor loadings of the bipolar-adjectives scales is still there. Also the factors are defined automatically in the process. All the other limitations or drawbacks listed for SD are mitigated or removed. The method is now slightly more tedious form as it needs to include two sets of scales, meaning a slightly larger workload for the respondents. Also there are more parameters in the GSD to set (the parameters of the kappa function, the framing of the "relevance" scale). Nevertheless, most of the drawbacks listed for the original method can be mitigated by the use of GSD and the method still retains the ability to deal with less tangible criteria. Let us therefore now see, how applicable the method might be in a multiple-criteria or multi-expert evaluation setting.

IV. Generalized Semantic differential and multiple-criteria evaluation

Before we are able to assess the potential benefits of applying GSD in multiple-criteria and multi-expert evaluation, and to suggest the needed modifications of the GSD method for this purpose, we need to define the multiple-criteria evaluation problem first. In multiple-criteria evaluation we assume that we have several objects/alternatives that need to be assessed and assigned a final evaluation of some sort. Usually the expected form of the evaluation is a numerical or vector one, that is in many evaluation methods we are looking for a real-value (or a vector of real values) that would summarize the qualities and the downsides of the alternative sufficiently. We also assume that the k criteria that represent the relevant features of the alternatives are known in advance (usually along with their types, underlying scales and also relative importances). The ultimate goal of the evaluation is then to a) obtain an ordering of the alternatives to be able to decide which ones to select (relative-type evaluation) or b) decide about the acceptability/unacceptability of the alternative (absolute-type evaluation). Let us now have a look at the features of the GSD method and comment on their usefulness or the need for the modification of these aspects for GSD to become a valid multiple-criteria evaluation method.

We need to start with one clear incompatibility between SD or GSD and the multiple-criteria evaluation setting. This is the fact that in GSD (and SD) the factors are defined through factor analysis and thus independent on the user of the evaluation. On the other hand in multiple-criteria evaluation, criteria are usually given and need to be used as defined by the user of the analysis. As we usually expect the k criteria to be independent, we can easily assume that each criterion would be represented by one axis in a k-dimensional Cartesian space. This would mean that if we substitute criteria for factors, we can obtain a method applicable to multiple-criteria evaluation. We can even assume that each criterion is "measured" or assessed repeatedly either through subcriteria, or through different items in a questionnaire or scorecard. Discarding the bipolar adjective scales completely we, however, lose the "projectivity" of the GSD and also the ability to capture less tangible and intangible aspects of the alternatives, as long as we do not have specific items for them in the data input tool (survey, scorecard, etc.). There is always a possibility to keep those bipolar adjective scales that measure the intangible factor(s) that might be relevant for our analysis (e.g. affect) and use externally defined criteria as other dimensions in the final output space. Being able to include the criteria as separate dimensions in the final output space, we can now analyze the other features of the GSD method:

- repeated measurement - semantic differential is built on the idea of repeated measurement of the factors. This is stemming from the fact that factor analysis is applied as a dimensionality reduction technique here. It also means that GSD is ready to process e.g. data from questionnaires where criteria are being assessed by more items. The aggregation of the values provided via different items in a questionnaire or a scorecard can be done either by arithmetic mean or any other feasible aggregation operator, weights can also be reflected, if needed (but weights of items in a questionnaire might not be frequently available).
- ability to capture less tangible aspects/criteria - if bipolar adjectives scale that are not descriptive for the evaluated object are kept in the pool of the items and their respective factor(s) constitute(s) separate dimension(s) in the final output space, then this feature is maintained. On the other hand factor analysis needs to precede the use of the bipolar adjective scales to know which ones are contributing to the desired factor.
- if the intangible criteria/aspects are not important in the evaluation process, then factor analysis might not be needed and the method is much simpler to apply as it does not longer require pre-analysis and an availability of a sample prior to the main analysis/evaluation.
- simple data input procedure - the data input can still be done through questionnaires, where groups of items would contribute to particular criteria. Each item can also

be assigned a scale for the reflection of uncertainty of the answer provided through this scale. Instead of "scale relevance", it might be better to talk about evaluator's confidence with the answer or something similar, though. Afterwards the kappa function can again be used to calibrate the method for the given purpose and to calculate the length of the interval of also possible values based on the confidence with the particular answer.
- graphical outputs - the original SD method was frequently shown to result in three factors. This allowed for a simple three-dimensional graphical representation if the outputs as points in the three-dimensional Cartesian space (called semantic space). For more factors or criteria, or simply for more dimensions of the output space, graphical outputs might not be achievable or easily understood. Still the intuition from three dimensional graphical summaries of the outputs (e.g. those presented in Fig. 2 and Fig. 3) can prove useful in explaining how the methods works in higher dimensions of the output space.
- presentation of results without final aggregation - this is one of the very desirable properties for multiple-criteria but also multi-expert evaluation. The design of the outputs allows for separate treatment of all the criteria representing the dimensions in the output space. The output objects (points or boxes of uncertainty around them) can be defined without an explicit knowledge of the relative importances of the criteria. If the evaluations represent outputs for different experts, the weights of experts do not need to be known either.

If a final ordering of the alternatives is required, one needs to be able to aggregate the information across all the dimensions of the output space. For this we can either introduce weights of criteria, or simply work with the k-dimensional representations of the objects directly and define distances on them. The multiple-criteria evaluation setting has the benefit of being able to define the most desired values of the criteria, and based on them the ideal (potentially non-existent) alternative, or at least the evaluation thereof (see the preference directions in Fig. 3 and the "ideal" evaluation defined based on them). The evaluation task can then be approached by defining distances between the evaluations of the alternatives (up to k-dimensional entities) from the ideal evaluation (also more ideals can be considered like e.g. in TOPSIS). The introduction of uncertainty in the SD represented by GSD then allows for the determination of interval-valued distances (for example shortest distance from the box of uncertainty to the ideal and longest distance from the box of uncertainty to the ideal defining the interval of possible distances - see Fig. 3). Overall GSD has the needed properties to be applied in multiple-criteria evaluation:

- it can handle multiple criteria (including less tangible ones - see the discussion above)
- it is capable of handling uncertainty of the evaluations with respect to the (sub)criteria
- the uncertainty can be assessed using a simple questionnaire-based input procedure. It does not require the respondent to be able to express uncertainty/risk in a complex way and can still derive the intervals of also possible values around the crisp evaluations provided by less certain or less experienced evaluators.
- it is designed for repeated measurement
- for low values of k it provides a graphical interface to present the results to the evaluators
- the k-dimensional representation of the final evaluation of the object does not require aggregation across criteria
- ordering of the alternatives can be obtained applying a suitable distance (interval-valued, if needed) of the k-dimensional representations of the evaluations of the alternatives and the k-dimensional representation of an ideal or desirable alternative (its evaluation). The distances from the least desirable alternative (its evaluation) can also be reflected 'TOPSIS-style'. These distances can reflect also the weights of criteria, or even be based on OWA operators as proposed in the linguistic OWA-TOPSIS [60].

As such GSD-based multiple-criteria evaluation seems to be particularly promising in areas where uncertainty of the evaluations is to be expected and needs to be reflected somehow. The design of the method and the application of the kappa function in combination with a simple assessment of (un)certainty or relevance of the provided evaluation is particularly suitable for those evaluation problems where laymen (in terms of risk/uncertainty representations) are evaluating, and where either less tangible or less usual criteria are being used, or where the alternatives are complex, abstract or novel in some way. The area of design management and design evaluation comes to mind as a first representative [53], [61]. But the applications are much wider and include social sciences and business in general, the evaluation of alternatives with emotional value for the evaluators, the assessment of risk, etc.

What seems to be an even stronger argument speaking in favor of the application of the GSD framework in the multiple-criteria evaluation setting is its capability of serving as a multi-expert evaluation analysis tool. In the multi-expert evaluation problem, we can assume the same that we did in the multiple-criteria evaluation setting, plus the fact that the evaluations are being provided by more evaluators and all their views/evaluations need to be reflected in the final decision to some extent. If we assume k criteria are used and m experts are involved in the evaluation task, then the evaluation of a single alternative can be represented by m k-dimensional objects in the k-dimensional Cartesian evaluation space. Apart from the desirable properties of GSD listed before, we can now consider also:

- the ability to see potential clusters of experts with similar evaluations - their number, distance etc. Obviously clustering techniques can be applied directly to the k-dimensional evaluations to define the clusters, if needed. This can bring understanding concerning the composition of the set of evaluators in terms of their priorities, mutual

agreement or even the number of potential points of view on the evaluation

- the overall evaluation of the alternative does not need to be represented as a single k-dimensional object in the evaluation space calculated as an average of all the expert evaluations (applying some aggregation operator) but instead can be represented by:
  - centroids of clusters of experts with similar background/opinion or simply similar evaluations of the alternative (information summarization such that different groups of experts and their views/evaluations remain visible)
  - by the set of m $k-$dimensional evaluations (no information reduction)
  - by a 'union' of the m $k-$dimensional evaluations constructed e.g. as a minimum k-dimensional evaluation such that all the other evaluations are its subsets in the k-dimensional space (maximally careful but potentially very uncertain summary)
  - by an 'intersection' of the m $k-$dimensional evaluations (if a nonempty intersection exists) - this would represent the 'common ground' or 'full agreement' of the experts in terms of their evaluations
  - by an evaluation of a specified shape (in the k-dimensional space) that is the closest to all the other evaluations (ideal compromise)
  - etc.
- the benefit from the possibility of finding consensus of expert evaluations [61] (intersections of the evaluations within a specified (sub) group of evaluators) and analyzing the compatibility of expert assessments by investigating the intersections of the k-dimensional evaluations either overall or dimension by dimension, or the distances of the expert evaluations from each other, from the centroids of clusters (if available), etc. Stoklasová et al. [61] define various types of consensus of expert evaluations that can be applied in the multiple-criteria multi-expert evaluation setting using the GSD evaluation method.

Overall the method allows for various definitions of the overall evaluation of the alternative based on m expert evaluations including such that lose very little information, it allows for the identification of (non) existence of the consensus of expert evaluations (overall and in terms of specific criteria), and for the identification of various types of consensus proposed in [61]. It can be used not only for the determination of the final group evaluation, but also for the analysis of the group of evaluators based on their evaluations. From the above mentioned points it seems that the GSD evaluation applied in the multi-expert setting can prove to be a useful tool for the evaluation and also for the understanding of the evaluation process.

V. Conclusions

Given the above mentioned analysis of the main features and potential benefits of the use of the GSD in multiple-criteria and multi-expert evaluation, the tool seems to be a reasonable candidate for future research concerning its applicability in this domain. We have managed to identify and stress some possible benefits of the use of this tool including the ability to assess less tangible aspects, the ability to model uncertainty, a convenient way of the presentation of results, simplicity of obtaining inputs etc. We have also analyzed the requirements of the method and there do not seem to be any major drawbacks preventing the applicability of the GSD-based tools in multiple-criteria and multi-expert evaluation problems. We have outlined a possible way to apply the main ideas of GSD in this context. More detailed description of the applicability of the method and practical application studies will be the subject of future research.

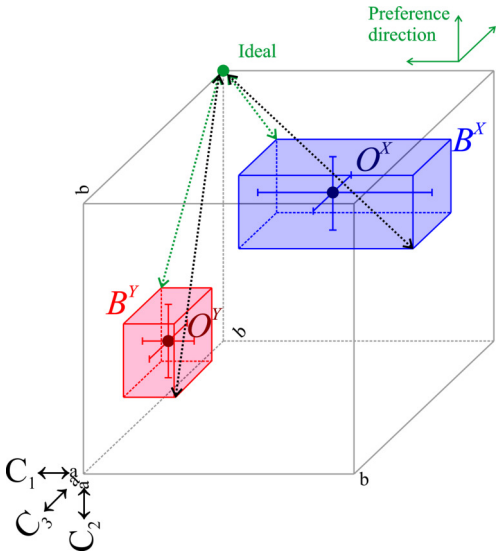


Fig. 3. An example of the output of the generalized semantic differential method [54] for two objects/concepts X and Y in the multiple-criteria evaluation setting. Three criteria C_1, C_2 and C_3 are assumed. The two evaluated objects X and Y are represented by O^X and O^Y and by the "boxes of uncertainty" B^X and B^Y around the O^x and O^Y respectively. The preference direction for all three criteria is shown in green and an ideal evaluation is defined based on this. Green dashed arrows denote the closest Euclidean distances from the ideal to the boxes of uncertainty, black ones the largest distance from the ideal to the points in the boxes of uncertainty.

References

[1] G. Mitra, H. J. Greenberg, F. A. Lootsma, M. J. Rickaert, and H.-J. Zimmermann, Eds., *Mathematical Models for Decision Support*. Berlin Heidelberg New York London Paris Tokyo: Springer-Verlag, 1988, vol. F48. ISBN 978-0387500843

[2] J. Figueira, S. Greco, and M. Ehrgott, Eds., *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005. ISBN 0387230815

[3] M. Doumpos, J. R. Figueira, S. Greco, and C. Zopounidis, Eds., *New Perspectives in Multiple Criteria Decision Making: Innovative Applications and Case Studies*. Springer Nature Switzerland AG, 2019.

[4] G. Silahtaroglu, H. Dinçer, and S. Yüksel, *Data Science and Multiple Criteria Decision Making Approaches in Finance: Applications and Methods*. Cham: Springer Nature Switzerland AG, 2021.

[5] G.-H. Tzeng and K.-Y. Shen, *New Concepts and Trends of Hybrid Multiple Criteria Decision Making*. New York: CRC Press, 2017. ISBN 978-1-4987-7708-7

[6] C. Zopounidis and M. Doumpos, Eds., *Multiple criteria decision making - Applications in Management and Engineering*. Springer International Publishing AG Switzeland, 2017.

[7] M. Metfessel, "A proposal for quantitative reporting of comparative judgements," *Journal of Psychology*, vol. 24, no. 2, pp. 229–235, 1947.

[8] B. F. Hobbs, "A Comparison of Weighting Methods in Power Plant Siting," *Decision Sciences*, vol. 11, no. 4, pp. 725–737, 1980. doi: 10.1111/j.1540-5915.1980.tb01173.x

[9] T. L. Saaty, "How to Make a Decision: The Analytic Hierarchy Process," *European Journal of Operational Research*, vol. 48, no. 1, pp. 9–26, 1990.

[10] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Information sciences*, vol. 3, no. 2, pp. 177–200, 1971.

[11] L. A. Zadeh and J. Kacprzyk, Eds., *Computing with words in information/intelligent systems 2: Applications*. Berlin Heidelberg GmbH: Springer-Verlag, 1999. ISBN 9783790824612

[12] J. Mingers and L. White, "A review of the recent contribution of systems thinking to operational research and management science," *European Journal of Operational Research*, vol. 207, no. 3, pp. 1147–1161, 2010. doi: 10.1016/j.ejor.2009.12.019

[13] P. Holeček, J. Talašová, and I. Müller, "Fuzzy Methods of Multiple-Criteria Evaluation and Their Software Implementation," in *Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition: Advancing Technologies*, V. K. Mago and N. Bhatia, Eds., Hershey, 2012, pp. 388–411. ISBN 9781613504291

[14] Z. Gong, Y. Lin, and T. Yao, *Uncertain Fuzzy Preference Relations and Their Applications*. Heidelberg New York Dordrecht London: Springer-Verlag, 2013. ISBN 9783642284472

[15] A. Sen, "Behaviour and the Concept of Preference," *Economica*, vol. 40, no. 159, pp. 241–259, 1973. doi: 10.2307/2552796

[16] Z. Xu, "A method based on linguistic aggregation operators for group decision making with linguistic preference relations," *Information Sciences*, vol. 166, no. 1, pp. 19–30, oct 2004. doi: 10.1016/j.ins.2003.10.006

[17] E. Herrera-Viedma, J. L. García-Lapresta, J. Kacprzyk, M. Fedrizzi, H. Nurmi, and S. Zadrozny, Eds., *Consensual processes*. Berlin Heidelberg: Springer, 2011. ISBN 9783642133428

[18] S. Greco, B. Bouchon-Meunier, G. Coletti, M. Fedrizzi, B. Matarazzo, and R. R. Yager, Eds., *Advances in Computational Intelligence: 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU 2012, Catania, Italy, July 9-13, 2012, Proceedings, Part IV*. Heidelberg New York Dordrecht London: Springer, 2012.

[19] G. Tzeng and J. Huang, *Multiple Attribute Decision Making: Methods and Applications*. Boca Raton, London, New York: CRC Press, Taylor & Francis group, 2011. ISBN 978-1-4398-6157-8

[20] J. Krejčí and J. Stoklasa, "Aggregation in the analytic hierarchy process: Why weighted geometric mean should be used instead of weighted arithmetic mean," *Expert Systems with Applications*, vol. 114, pp. 97–106, 2018. doi: 10.1016/j.eswa.2018.06.060

[21] V. Jandová, J. Krejčí, J. Stoklasa, and M. Fedrizzi, "Computing interval weights for incomplete pairwise-comparison matrices of large dimension - a weak consistency based approach," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 6, pp. 1714–1728, 2017. doi: 10.1109/TFUZZ.2016.2633364

[22] T. L. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234–281, jun 1977. doi: 10.1016/0022-2496(77)90033-5

[23] T. Talášek and J. Stoklasa, "Ordering of fuzzy quantities with respect to a fuzzy benchmark - how the shape of the fuzzy benchmark and the choice of distance / similarity affect the ordering," in *Proceedings of the 36th International Conference on Mathematical Methods in Economics*, L. Váchová and O. Kratochvíl, Eds. Jindrichuv Hradec: MatfyzPress, 2018. ISBN 978-80-7378-372-3 pp. 573–578.

[24] C. Hwang, Y. Lai, and T. Liu, "A new approach for multiple objective decision making," *Computers & Operations Research*, vol. 20, no. 8, pp. 889–899, 1993. doi: 10.1016/0305-0548(93)90109-V

[25] L. Basile and L. D'Apuzzo, "Transitive matrices, strict preference order and ordinal evaluation operators," *Soft Computing*, vol. 10, no. 10, pp. 933–940, 2006. doi: 10.1007/s00500-005-0020-z

[26] R. R. Yager, "An approach to ordinal decision making," *International Journal of Approximate Reasoning*, vol. 12, no. 3-4, pp. 237–261, 1995.

[27] J. Stoklasa, T. Talášek, and P. Luukka, "Fuzzified Likert scales in group multiple-criteria evaluation," in *Soft Computing Applications for Group Decision-making and Consensus Modeling*, M. Collan and J. Kacprzyk, Eds. Springer International Publishing AG, 2018, vol. 357, pp. 165–185. ISBN 978-3-319-60206-6

[28] G. Bojadziev and M. Bojadziev, *Fuzzy logic for business, finance and management*, 2nd ed. World Scientific, 2007.

[29] J. Talašová, *Fuzzy metody vícekriteriálního hodnocení a rozhodování*. Olomouc: Palacký University in Olomouc, 2003. ISBN 8024406144

[30] H.-J. Zimmermann, *Fuzzy Set Theory and Its Applications*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2001. ISBN 0792374355

[31] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Aplications*. New Jersey: Prentice Hall, 1995. ISBN 0131011715

[32] M. Smithson and J. Verkuilen, *Fuzzy set theory: applications in the social sciences*. Thousand Oaks, London, New Delhi: Sage Publications, 2006. ISBN 076192986X

[33] V. Traneva, S. Tranev, and D. Mavrov, "Interval-Valued Intuitionistic Fuzzy Decision-Making Method using Index Matrices and Application in Outsourcing," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems, ACSIS, Vol. 25*, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., 2021. doi: 10.15439/2021F77 pp. 251–254.

[34] T. T. Yaman, "Pythagorean Fuzzy Analytical Network Process (ANP) and Its Application to Warehouse Location Selection Problem," in *Proceedings of the Federated Conference on Computer Science and Information Systems, ACSIS, Vol. 21*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., 2020. doi: 10.15439/2020F187 pp. 137–140.

[35] F. Herrera and E. Herrera-Viedma, "Linguistic decision analysis: steps for solving decision problems under linguistic information," *Fuzzy Sets and Systems*, vol. 115, no. 1, pp. 67–82, oct 2000. doi: 10.1016/S0165-0114(99)00024-X

[36] J. Stoklasa, *Linguistic models for decision support*. Lappeenranta: Lappeenranta University of Technology, 2014. ISBN 978-952-265-686-2 (PDF)

[37] J. Kacprzyk and S. Zadrozny, "Linguistic Data Summarization: A High Scalability through the Use of Natural Language?" in *Scalable Fuzzy Algorithms for Data Management and Analysis: Methods and Design*, A. Laurent and M.-J. Lesot, Eds. IGI Global, 2010, pp. 214–237.

[38] M. Delgado, J. L. Verdegay, and M. A. Vila, "Linguistic decision-making models," *International Journal of Intelligent Systems*, vol. 7, no. 5, pp. 479–492, 1992. doi: 10.1002/int.4550070507

[39] M. Collan and J. Kacprzyk, Eds., *Soft Computing Applications for Group Decision-making and Consensus Modeling*. Cham: Springer international publishing AG, 2018. ISBN 978-3-319-60206-6

[40] V. Sukač, J. Talašová, and J. Stoklasa, "'Soft' consensus in decision-making model using partial goals method," in *Proceedings of the 34th International Conference on Mathematical Methods in Economics*. Liberec: Technical University of Liberec, 2016. ISBN 978-80-7494-296-9 pp. 791–796.

[41] M. Fedrizzi, M. Fedrizzi, R. A. M. Pereira, and M. Brunelli, "Consensual dynamics in group decision making with triangular fuzzy numbers," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2008. doi: 10.1109/HICSS.2008.100. ISBN 0769530753. ISSN 15301605 pp. 1–9.

[42] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[43] D. Kahneman, "The Semantic Differential and the Structure of Inferences Among Attributes," *The American journal of psychology*, vol. 76, no. 4, pp. 554–567, 1963.

[44] A. Tversky and D. Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, vol. 5, no. 4, pp. 297–323, 1992. doi: Doi 10.1007/Bf00122574

[45] ——, "The framing of decisions and the psychology of choice." pp. 453–458, 1981.

[46] R. P. Hämäläinen, J. Luoma, and E. Saarinen, "On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems," *European Journal of Operational Research*, vol. 228, no. 3, pp. 623–634, aug 2013. doi: 10.1016/j.ejor.2013.02.001

[47] J. Stoklasa, T. Talášek, and J. Stoklasová, "Executive summaries of uncertain values close to the gain/loss threshold - linguistic modelling perspective," *Expert Systems with Applications*, vol. 145, p. 113108, 2020. doi: 10.1016/j.eswa.2019.113108

[48] J. Stoklasa and M. Kozlova, "What is your problem, decision maker? Do we even care anymore?" in *KNOWCON 2020, Knowledge on Economics and Management, Conference Proceedings*, P. Slavíčková

and J. Stoklasa, Eds. Olomouc: Palacký University Olomouc, 2020. doi: 10.5507/ff.20.24457987. ISBN 9788024457987 pp. 208–214.

[49] A. Morreale, J. Stoklasa, and T. Talášek, "Fuzzy Grouping Variables in Economic Analysis. A Pilot Study of a Verification of a Normative Model for R&D Alliances," *Fuzzy Economic Review*, vol. 21, no. 2, pp. 19–46, 2016.

[50] Y. Mitev and L. Kirilov, "Group Decision Support for e-Mail Service Optimization through Information Technology Infrastructure Library Framework," in *Proceedings of the 16th Conference on Computer Science and Intelligence Systems, ACSIS, Vol. 25*, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., 2021. doi: 10.15439/2021F93 pp. 227–230.

[51] R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds., *The Oxford Handbook of Affective Computing*. Oxford University Press, jan 2015. ISBN 9780199942237

[52] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Chicago: University of Illinois Press, 1957. ISBN 0252745396

[53] J. Stoklasa, T. Talášek, and J. Stoklasová, "Reflecting emotional aspects and uncertainty in multi-expert evaluation: one step closer to a soft design-alternative evaluation methodology," in *Advances in Systematic Creativity: Creating and Managing Innovations*, L. Chechurin and M. Collan, Eds. Palgrave Macmillan, 2019, pp. 299–322. ISBN 978-3-319-78074-0

[54] ——, "Semantic differential for the twenty-first century: scale relevance and uncertainty entering the semantic space," *Quality & Quantity*, vol. 53, no. January 2019, pp. 435–448, may 2019. doi: 10.1007/s11135-018-0762-1

[55] C. E. Osgood, "Semantic Differential Technique in the Comparative Study of Cultures," *American Anthropologist*, vol. 66, no. 3, pp. 171–200, 1964. doi: 10.1515/9783110215687.109

[56] O. Friborg, M. Martinussen, and J. H. Rosenvinge, "Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience," *Personality and Individual Differences*, vol. 40, no. 5, pp. 873–884, 2006. doi: 10.1016/j.paid.2005.08.015

[57] N. Kervyn, S. T. Fiske, and V. Y. Yzerbyt, "Integrating the stereotype content model (warmth and competence) and the Osgood semantic differential (evaluation, potency, and activity)," *European Journal of Social Psychology*, vol. 43, no. 7, pp. 673–681, 2013. doi: 10.1002/ejsp.1978

[58] D. R. Heise, "Some methodological issues in semantic differential research." *Psychological Bulletin*, vol. 72, no. 6, pp. 406–422, 1969. doi: 10.1037/h0028448

[59] J. D. Dombi and A. Kertész, "Advanced Scheduling Techniques with the Pliant System for High-Level Grid Brokering," in *Agents and Artificial Intelligence Second International Conference, ICAART 2010*, J. Filipe, A. Fred, and B. Sharp, Eds. Heidelberg New York Dordrecht London: Springer, 2011, pp. 173–185. ISBN 9783642198892

[60] P. Luukka and J. Stoklasa, "Similarity based topsis with linguistic-quantifier based aggregation using owa," in *Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics*, ser. Annals of Computer Science and Information Systems, J. Stoklasa, P. Luukka, and M. Ganzha, Eds., vol. 29. PTI, 2021. doi: 10.15439/2021B6 p. 45–51. [Online]. Available: http://dx.doi.org/10.15439/2021B6

[61] J. Stoklasová, T. Talášek, and J. Stoklasa, "Attitude-based multi-expert evaluation of design," in *Intelligent Systems and Applications in Business and Finance*, P. Luukka and J. Stoklasa, Eds. Springer, 2022, p. (in press).

KNOWCON
Knowledge on Economics and Management
International Scientific Conference

Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics pp. 63–73

DOI: 10.15439/2021B2
ISSN 2300-5963 ACSIS, Vol. 29

Possible drivers of high performance of European mutual ESG funds - an fsQCA view on sustainable investing

Fanni Welling
School of Business and Management, LUT University,
Yliopistonkatu 34, 53851 Lappeenranta, Finland
Email: fanni.welling@student.lut.fi

Jan Stoklasa
School of Business and Management, LUT University
Yliopistonkatu 34, 53851 Lappeenranta, Finland, and
Palacky University Olomouc, Faculty of Arts
Department of Economic and Managerial Studies
Email: jan.stoklasa@lut.fi
Email: jan.stoklasa@upol.cz

***Abstract*—The paper applies the tools of fsQCA and their recent modifications by Stoklasa, Luukka and Talášek to analyze the possible drivers of high performance of European ESG funds. 429 mutual equity growth ESG funds from the European area are being analyzed. We focus mainly on the connection of Morningstar sustainability rating with the performance of the funds during 2018-2021 measured by Jensen's alpha and the Sharpe ratio. Other possible drivers of the success of these funds are also being explored. We identify the prevailing assumed relationships between funds' sustainability and other characteristics with their performance and formulate rules to be investigated using the fsQCA methodology. More specifically the possibility of high performance being associated with a high sustainability rating of the funds is explored in detail. Our results indicate that although the high performance cannot be clearly associated with the high sustainability rating of a fund, high sustainability rating seems to be preventing the low performance of the fund.**

I. Introduction

SUSTAINABILITY and responsibility are not only topical issues in business scientific literature and practice [1], [2], but these concepts are also potentially influencing the investment decision-making of individual investors. In this paper, we discuss three factors that might potentially influence the performance of mutual funds, namely the size of the fund, the length of its managers' tenure and its sustainability rating, show the relationships that have already been identified in the literature between these factors and the performance of the fund. In line with the usual approaches in the literature, the performance of the funds is measured using Jensen's alpha and the Sharpe ratio in this paper.

We then set the goal of validating the existence of the "prevailing" relationships on a chosen sample of 429 European growth funds in the 2018-2021 period. Given the fact that most studies (see the brief literature reviews for each feature further in the text) use statistical methods (regression etc.) to investigate the existence of relationships, it is reasonable to try to verify the (non)existence of the relationships between the chosen features and the fund performance using a different methodology.

We therefore apply the tools of the set-theoretic approach and its fuzzification, that are utilized in the frame of the fuzzy set qualitative comparative analysis (fsQCA) - namely we focus on the concepts of the consistency of the rules representing specific assumed relationships with the data and the coverage of these relationships by the available data [3], [4]. Given the recent advances in the methods for fsQCA focusing on the investigation of consistency and coverage of assumed relationships in the fuzzy context, we also apply the recently introduced fuzzified consistency and coverage measures and their alternatives [5], [6]. Another reason to reach for the set-theoretic methods is the fact that based on the definition of the rules (investigated relationships formulated as IF-THEN rules) we can postulate and verify the existence of non-linear relationships between the features of the funds and their performance. This has proven to be beneficial in the recent studies on strategic decision-making [7], [8].

Even though our focus is mainly on the possible relationship between sustainability (or sustainability ratings) of the funds and their performance, we include the other fund features too to be able to assess the performance of the fsQCA methods on the data. This way we will be able to interpret the results concerning sustainability in the context of fund size and manager tenure as well. Other potentially relevant features such as green approach to HR management [9], corporate social responsibility or company's reputation [1], [2] and others are left out of the scope of this paper.

II. Preliminaries

Let U be a nonempty set. A *fuzzy set* A on U is defined by a mapping $\mu_A : U \to [0, 1]$, where μ_A is called a *membership function* of A (see e.g. [10], [11] for more details). The set of all fuzzy sets on U is denoted $\mathcal{F}(U)$. For simplicity, we can denote a fuzzy set and its membership function by the same

This work was supported by LUT research platform AMBI- Analytics-based management for business and manufacturing industry and partially also by the grant IGA_FF_2021_001 Barriers to the expansion of sustainable consumption.

symbol (that way the membership function of a fuzzy set A will be denoted $A(.)$). Let $A \in \mathcal{F}(U)$, then

- the *kernel* of A is a crisp set $\text{Ker}(A) = \{x \in U \mid A(x) = 1\}$.
- the *support* of A is a crisp set $\text{Supp}(A) = \{x \in U \mid A(x) > 0\}$.
- the *height* of A is $\text{hgt}(A) = \sup\{A(x) \mid x \in U\}$
- the $\alpha-$*cut* of A is a crisp set $A^\alpha = \{x \in U \mid A(x) \geq \alpha\}$ for any $\alpha \in [0, 1]$.

A *negation of a fuzzy* set $A \in \mathcal{F}(U)$ is a fuzzy set $\neg A \in \mathcal{F}(U)$ such that for any $x \in U$ we have $\neg A(x) = 1 - A(x)$. Let A be a fuzzy set on $\mathbb{R}$, such that all the following conditions are met:

1) A is normal that is, $\text{hgt}(A) = 1$,
2) A^α is a closed interval for all $\alpha \in (0, 1]$,
3) $\text{Supp}(A)$ is bounded,

then A is called a *fuzzy number* on $\mathbb{R}$, denoted as $A \in \mathcal{F}_N(\mathbb{R})$. Each fuzzy number $B \in \mathcal{F}_N(\mathbb{R})$ can be represented by a quadruple of characteristic values $B \sim (b_1, b_2, b_3, b_4)$, where $b_1, ..., b_4 \in \mathbb{R}$, $b_1 \leq b_2 \leq b_3 \leq b_4$, and $[b_1, b_4] = \text{Cl}(\text{Supp}(B))$, $[b_2, b_3] = \{x \in \mathbb{R} \mid B(x) = 1\} = \text{Ker}(B)$ and $B(x) = 0$ for all $x \in (-\infty, b_1] \cup [b_4, \infty)$. For a *triangular fuzzy number* we have $b_2 = b_3$ and the membership function is continuous, linear and strictly increasing between the points b_1 and b_2 and continuous, linear and strictly decreasing between b_3 and b_4. For a *trapezoidal fuzzy number* we assume the same, we just allow $b_2 \neq b_3$. If $[b_1, b_4] \subseteq [r, s]$ we call B a *fuzzy number on an interval* $[r, s]$. The set of all fuzzy numbers on an interval $[r, s]$ will be denoted $\mathcal{F}_N([r, s])$. In this paper, we will only consider these two types of fuzzy numbers to represent the linguistically defined values of the features under investigation.

As the main methodology chosen for this paper is the set-theoretic investigation of the consistency of the investigated rules with the data, we will need to introduce the basic (fuzzy) set-theoretic concepts of consistency and coverage as used in the fsQCA [12] and as recently generalized by Stoklasa et. al [5], [6]. We will be employing the revised fuzzification of the consistency and coverage measures [5], [6] as these have already proven useful in practical investigation of real-life relationships in business data [7]. Let us consider a set of observations $U = \{x_1, x_2, \ldots, x_n\}$. Let us consider a feature A and an indicator function $\chi^A : U \to \{0, 1\}$ such that $\chi^A(x_i) = 1$ if and only if x_i has the feature A and $\chi^A(x_i) = 0$ otherwise, for all $i = 1, \ldots, n$. Let us also consider a feature B with an analogous indicator function $\chi^B : U \to \{0, 1\}$. Let us also introduce a negation of the feature B representing the absence of the feature B (denoted B' and meaning "not B"), for which the indicator function is $\chi^{B'} : U \to \{0, 1\}$ such that $\chi^{B'}(x_i) = 1$ if and only if x_i does not have the feature B and $\chi^{B'}(x_i) = 0$ otherwise. In other words, we have $\chi^B(x_i) = 1 - \chi^{B'}(x_i)$ and as long as "possessing a feature" is considered as a crisp (binary) state, we have $\chi^B(x_i), \chi^{B'}(x_i) \in \{0, 1\}$. Now we assume that we need to investigate the assumption that *an observation having a feature A also implies it having the feature B as well*, or $A \Rightarrow B$ for short. Given the set of observations U and given $A \subseteq U, B \subseteq U$, we can assess the consistency [12] of such a crisp assumption with the data (its support by the data) computing the consistency of $A \Rightarrow B$:

$$\text{Consistency}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} \min\{\chi^A(x_i), \chi^B(x_i)\}}{\sum_{i=1}^{n} \chi^A(x_i)} = \frac{\text{Card}(A \cap B)}{\text{Card}(A)}, \quad (1)$$

where $\text{Card}(A)$ represents the cardinality of the set A, i.e. the number of its elements, and $\cap$ is the standard set intersection, i.e. $\chi^{(A\cap B)}(x_i) = \min\{\chi^A(x_i), \chi^B(x_i)\}$. Note, that U is fully consistent with $A \Rightarrow B$ as long as $A \subseteq B$ (which implies that $A \cap B = A$), i.e. in this case $\text{Consistency}(A \Rightarrow B) = 1$ and we can interpret this as the absence of counterexamples to $(A \Rightarrow B)$; obviously we need to assume that $\text{Card}(A) \neq 0$. If the cardinality of A was zero, then there would be no observations that possess the feature A and it would make no sense to try to investigate the compatibility of the assumption $A \Rightarrow B$ with the given dataset. Analogously we can calculate a measure of "universality" of the assumption $A \Rightarrow B$ for the given set of observations U as the coverage of $A \Rightarrow B$ (assuming again that $\text{Card}(B) \neq 0$):

$$\text{Coverage}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} \min\{\chi^A(x_i), \chi^B(x_i)\}}{\sum_{i=1}^{n} \chi^B(x_i)} = \frac{\text{Card}(A \cap B)}{\text{Card}(B)}. \quad (2)$$

Apparently $\text{Coverage}(A \Rightarrow B) = 1$ if and only if $B \subseteq A$. In other words, both measures are based on subsethood. This means that the validity of the assumption that A leads to B is assessed based on the available data - if the set of observations having feature A is a subset of those observations that have the feature B, then having the feature A can be considered a sufficient condition for having the feature B too (see [12] or [5] for more details). If the possession of the feature can be understood in gradual and not binary terms, a fuzzification of the whole approach is necessary. We can still assume that the possession of the feature A by an element of U can be described by its membership to A, we just need to allow $A \in \mathcal{F}(U)$, that is we need to allow for A to be a fuzzy subset of U.

If we now assume that A and B are fuzzy sets ($A, B \in \mathcal{F}(U)$) and $\mu_A : U \to [0, 1]$ and $\mu_B : U \to [0, 1]$ are their respective membership functions, we need to introduce at least the fuzzy-set subsethood, fuzzy-set intersection operation and the notion of a cardinality of a fuzzy set to be able to generalize (1) and (2). The intersection of two fuzzy sets A and B on the same universe U is a fuzzy set $(A \cap B)$ on U with the membership function $\mu_{A\cap B} : U \to [0, 1]$ such that for any $x \in U$ we have $\mu_{A\cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$. A is a fuzzy subset of B (denoted $A \subseteq_F B$) if for all $x \in U$ it holds that $\mu_A(x) \leq \mu_B(x)$. The cardinality of a fuzzy set $A \in \mathcal{F}(U)$ is calculated as $\text{Card}(A) = \sum_{x_i \in U} A(x_i)$ as long as U is a discrete set, and $\text{Card}(A) = \int_{x_i \in U} A(x_i)dx$ as long as U is a continuous universe (e.g. a subinterval of the real axis). The direct fuzzification of (1) and (2) stemming from

the subsethood interpretation of consistency and coverage can be expressed by the following formulas [12] (the fuzzified formulas will be denoted by the subscript F):

$$\text{Consistency}_{F_1}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} \min\left(\mu_A(x_i), \mu_B(x_i)\right)}{\sum_{i=1}^{n} \mu_A(x_i)}, \tag{3}$$

$$\text{Coverage}_{F_1}(A \Rightarrow B) = \frac{\sum_{i=1}^{n} \min\left(\mu_A(x_i), \mu_B(x_i)\right)}{\sum_{i=1}^{n} \mu_B(x_i)}. \tag{4}$$

Stoklasa et al. [5] proposed a different fuzzification of (1) and (2) that deals with the fact that the transition to a gradual possession of a feature ultimately implies that a feature can be partially possessed and partially not possessed (in a nonzero degree) by the same observation at the same time. This results in the ambivalence of evidence in the set-theoretic investigation, as the same observation can now simultaneously support $A \Rightarrow B$ and $A \Rightarrow B'$ to some extent. Stoklasa et al. therefore suggested several alternative fuzzifications of formulas (1) and (2) - namely the F_2 fuzzification [5] represented by formulas (5) and (6) that removes that part of evidence that is ambivalent, F_3 fuzzification [5] that focuses of "pure support" of the investigated relationship by removing ambivalent evidence as well as reducing the evidence by the amount of available "pure" counterevidence represented by formulas (7) and (8). Finally, a modification of (7) and (8) was proposed in [6] that deals with the partial loss of information introduced to F_3 formulas by the use of the maximum operator. These F_4 fuzzifications are represented by formulas (9) and (10); note that the results of these formulas have a slightly different interpretation - for example if $\text{Consistency}_{F_4}(A \Rightarrow B) = 0.5$, then there is the same amount of "pure" evidence as there is counterevidence with regards to the investigated relationship, whereas if $\text{Consistency}_{F_4}(A \Rightarrow B) = 1$, then there is only "pure evidence" in its favor etc. It should be noted that Stoklasa et al. also proposed a completely different approach to the assessment of consistency and coverage of the investigated relationships [5] represented by the degree of (unconditional) support and degree of (unconditional) disproof, that are based on α-cuts of the fuzzy numbers used to represent the investigated values of the variables, namely it takes into account the amount of fulfillment of the outcome of the investigated rule. A more detailed discussion of the degrees of support/disproof is not necessary here, we therefore refer the interested readers to [5] and here we will simply calculate and discuss the values.

To make the description of the methods complete, we need to specify the measures applied to the assessment of the performance of the selected mutual equity growth ESG funds. The first measure applied in this paper is Jensen's alpha [13] which is calculated for a portfolio i using equation (11), where r_i is the return of the portfolio, β_i is the beta coefficient of the portfolio, r_m is the return of the market and r_f is the risk-free rate. From its construction, it is apparent that α_i is a risk-adjusted measure of portfolio performance that represents the excess returns of the portfolio above the expected level (derived through the capital asset pricing model (CAPM)). It is a benefit-type criterion of fund performance and positive values are interpreted as desirable as they represent situations when the portfolio under investigation outperforms the benchmark market portfolio.

$$\alpha_i = r_i - (\beta_i(r_m - r_f)) \tag{11}$$

Another performance measure applied in this paper is the Sharpe ratio [14]. This measure S_i reflects the returns of portfolio i per unit of risk and it is defined using (12), where r_i and r_f have the same interpretation as in Jensen's alpha and δ_i is the standard deviation of the i-th portfolio.

$$S_i = \frac{r_i - r_f}{\delta_i} \tag{12}$$

Unfortunately, Sharpe ratio's interpretability is limited when the information about the actual size of risk is not available or when a reference investment is not available. Higher values of this measure are preferred as they indicate better performance, however one can never be sure whether a high value of the ratio is obtained due to high excess returns, or due to low volatility of the portfolio. Sharpe ratio is therefore used as a secondary performance measure in this analysis.

III. Features of the mutual funds and their relationships with fund performance

In this part, we will briefly summarize the results of previous research on the possible links between the performance of mutual funds and their size, the length of their managers' tenure and their sustainability ratings. We do not claim the literature review in this aspect is complete, we mainly use the presented papers as a basis for the formulation of the assumed relationships to avoid data-mining bias.

A. Relationship between mutual fund size and its performance

Table I lists seven papers that focus on the relationship between the performance of the fund and its size. The analyzed periods do not cover the last 20 years, yet the most recent papers tend to agree on the existence of a negative relationship between the size of the fund and its performance. The only discovered relationships that can be considered positive are dealing with economies of scale and suggest that the larger the funds get, the lower the fees and thus the higher the potential returns for the investors (a simplified interpretation). Most of the research also relies on regression or other statistical methods. Based on the presented summary, we postulate the following potential relationship to be investigated: *If the fund size is large, then the risk-adjusted returns are low.* We will specify the meanings of "large" fund size and "low" risk-adjusted returns in the data section, where the meanings of these linguistic descriptions will be provided in terms of fuzzy numbers. In line with the recommendations by Stoklasa et al. [5], the opposite relationship *If the size of the fund is large, then its risk-adjusted returns are not low* will also be investigated to get a more complete picture.

$$\text{Consistency}_{F_2}(A \Rightarrow B) = \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_A(x_i), \mu_B(x_i), \mu_{B'}(x_i)))}{\sum_{i=1}^{n} \mu_A(x_i)} \tag{5}$$

$$\text{Coverage}_{F_2}(A \Rightarrow B) = \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_A(x_i), \mu_B(x_i), \mu_{A'}(x_i)))}{\sum_{i=1}^{n} \mu_B(x_i)} \tag{6}$$

$$\text{Consistency}_{F_3}(A \Rightarrow B) = \max\left\{0; \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_A(x_i), \mu_{B'}(x_i)))}{\sum_{i=1}^{n} \mu_A(x_i)}\right\} \tag{7}$$

$$\text{Coverage}_{F_3}(A \Rightarrow B) = \max\left\{0; \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_B(x_i), \mu_{A'}(x_i)))}{\sum_{i=1}^{n} \mu_B(x_i)}\right\} \tag{8}$$

$$\text{Consistency}_{F_4}(A \Rightarrow B) = \frac{1}{2}\left(1 + \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_A(x_i), \mu_{B'}(x_i)))}{\sum_{i=1}^{n} \mu_A(x_i)}\right) \tag{9}$$

$$\text{Coverage}_{F_4}(A \Rightarrow B) = \frac{1}{2}\left(1 + \frac{\sum_{i=1}^{n}(\min(\mu_A(x_i), \mu_B(x_i)) - \min(\mu_B(x_i), \mu_{A'}(x_i)))}{\sum_{i=1}^{n} \mu_B(x_i)}\right) \tag{10}$$

TABLE I
SUMMARY OF THE REVIEWED PAPERS DEALING WITH THE RELATIONSHIP OF THE SIZE OF THE FUND AND ITS PERFORMANCE.

Year	Paper, Authors	Declared objective(s)	Period	Data characteristics	Methodology	Results	Assumed effect[a]
2009	Chan, Faff, Gallagher and Looi [15]	To investigate if fund size affects performance. To identify the causes for the possible relation.	1998-2001 (40mths)	35 Australian equity funds	Regression analysis and simulation.	Fund size lowers performance, especially for funds with highly active trading approaches.	-
2008	Yan [16]	To examine the impact of liquidity and investment style on the relationship between fund size and fund performance.	1993-2002	1024 actively managed U.S. mutual funds.	Cross-sectional regression analysis and a portfolio approach. Performance measured with Alpha, CAPM, three- and four-factor models.	A negative relationship between fund size and fund performance. Liquidity is proposed as an important reason to cause this relation.	-
2004	Chen, Hong, Huang and Kubik [17]	To investigate if fund size affects fund performance.	1962-1999	3439 funds from the U.S.	Regression analysis. Performance measured with CAPM, three- and four-factor models.	A negative relationship between fund size and fund performance mainly caused by the lack of liquidity.	-
2001	Beckers and Vaughan [18]	To examine how fund size affects investment performance.	1996-1999	250 stocks from an Australian Index; Daily prices and trading volumes	Historical real-life simulation.	Bigger funds are less flexible in implementing their ideas and thus creating value-added is harder as the number of assets under management grow.	-
1997	Tufano and Sevick [19]	To research the relationship between fund board structure and fund fees. Also the relationship between fund size and fees is examined.	1991-1992 (12mths)	1587 U.S. open-end mutual funds.	Regression analysis.	Fund fees are inversely related to fund size, and thus larger funds have economies of scale.	+
1996	Golec [20]	To study if mutual fund manager's features affect fund fees, performance and risks. Also the effect of fund size is examined.	1988-1990	530 mutual funds; geographically not specified.	Regression analysis. Alpha and yield as performance measures.	Larger funds discover economies of scale. Large funds' fees are lower leading to larger yields.	+
1991	Perold and Salomon [21]	To detect the right amount of assets under management for financial maximization.	1982	Examples from [22] 1200 observations.	A mathematical analysis using a wealth-maximizing tradeoff. Alpha as performance measure.	The optimal fund size is when trading costs exceed the opportunity cost of not trading. A larger asset base than that leads to higher opportunity costs and lower returns.	-

[a] + indicates a positive relationship, - indicates a negative relationship, 0 indicates no relationship between the size of the fund and its performance; adapted from [23]

TABLE II
SUMMARY OF THE REVIEWED PAPERS DEALING WITH THE RELATIONSHIP OF THE LENGTH OF THE TENURE OF FUND MANAGER AND THE PERFORMANCE OF THE FUND.

Year	Paper, Authors	Declared objective(s)	Period	Data characteristics	Methodology	Results	Assumed effect[a]
2016	Kjetsaa and Kieff [24]	To explore the effect of manager tenure, expenses and turnover on blend fund performance.	2002-2012	559 blend funds; geographically not specified.	Regression analysis for three time horizons (3, 5 and 10 years). Returns as a performance measure.	There is a positive relation between manager tenure and mutual fund returns.	+
2006	Costa, Jakob and Porter [25]	To examine how market trends and fund managerial experience affect the ability to outperform the market.	1990-2001	1249 mutual equity funds from the U.S.	Regression analysis. Alpha from a four-factor model as a performance measure.	Longer-tenured managers do not outperform shorter tenured managers.	0
2004	Filbeck and Tompkins [26]	To investigate if there is a relation between manager tenure and risk-adjusted returns.	1990-2000	sample size or geographical area not specified.	Regression analysis. M-squared as a measure of risk-adjusted performance.	Longer-tenured managers outperformed the market more than shorter-tenured managers. Long-tenured managers were able to manage funds on lower expenses and thus more efficiently.	+
2002	Brooks and Tompkins [27]	To investigate the effect of mutual fund characteristics on mutual fund performance.	1989-1999	474 mutual funds; geographically not specified.	A two-tailed Z-test and regression analysis. M-squared as a measure of risk-adjusted performance.	A slight adverse relationship between manager tenure and risk-adjusted returns.	-
1999	Fortin, Michelson and Jordan-Wagner [28]	To research how manager tenure affects mutual fund performance across all investment classes.	1985-1995	800 bond and equity funds; geographically not specified.	Comparison of short-term and long-term fund managers' performance and regression analysis. Alpha as a performance measure.	Manager tenure does not affect mutual fund performance. There is an adverse relation between manager tenure and fund turnover.	0
1996	Golec [20]	To study if mutual fund manager's features affect fund fees, performance and risks. Also the effect of fund size is examined.	1988-1990	530 mutual funds; geographically not specified.	A three-stage least squares (3SLS) regression analysis. Yield and Jensen's Alpha as performance measures.	There is a positive connection between manager tenure and fund performance.	+
1996	Lemak and Satish [29]	To examine the differences in mutual fund performance and risk between longer-tenured mutual fund managers (>10 years) and shorter tenured managers (<10 years).	1984-1994	313 mutual funds; geographically not specified.	Comparison of short-term and long-term fund managers' performance. Regression analysis. Return as a performance measure.	Longer-tenured (10 years or more) fund managers performed better than shorter tenured managers.	+

[a] + indicates a positive relationship, - indicates a negative relationship, 0 indicates no relationship between the length of tenure and the performance of the fund; adapted from [23]

B. Relationship between the length of the tenure of mutual fund's manager and the performance of the fund

As can be seen in Table II, manager tenure and its effect on the performance of the mutual funds is a more actual topic with periods being analyzed stretching at least to 2012. Also in this context, the majority of the research is based on regression (statistical) models that in many cases involve the assumption of linearity of the relationship in one way or another. Also, the results of the research are a bit less consistent. We can find research that did not discover any sort of relationship between the length of manager tenure and fund performance, also some weak evidence of a negative-type of relationship can also be found; the prevailing result, however, seems to be one that confirm the existence of a positive relationship between the length of manager's tenure and fund performance. The positive relationship can be expressed by the manager's experience and ability to manage the fund more efficiently, while the negative relationship might be stemming from the inability of long-term managers to "think out of the box" and thus missing some opportunities.

Based on the presented summary of previous research, we consider the relationship *If fund manager's tenure is high, then the risk-adjusted returns of the fund are high* to be the one to validate on our data. Again, we will also investigate the validity of the opposite relationship *If fund manager's tenure is high, then the risk-adjusted returns of the fund are not high.* The definition of the fuzzy-number representation of high tenure will be provided further on in the data section.

C. The relationship between the sustainability rating of the fund and its performance

Out of all the variables, whose potential effect on fund performance is being studied in this paper, sustainability is definitely the one that has been receiving researchers' attention most recently (see Table III). From the conducted literature review it is obvious, that even though the topic is being currently researched, the findings are far from being unanimous. One main issue in the scientific investigation of the effect of sustainability on (or the relationship thereof with) other variables suffers from the multitude of possible approaches to sustainability and its definition. We can see the terms sustainable, responsible, green and many others being used interchangeably, and we can also frequently encounter the "environmental, social and governance" (ESG) label denoting those funds (companies) that either consider these factors in the composition of their investment portfolios or set explicit goals concerning these areas. In older literature mainly the predecessor of ESG - the corporate social responsibility (CRS) - can be found. Even though all these terms and concepts might share some goals or an ultimate vision, their definitions are not identical, the measures for the fulfillment of all the necessary criteria to use some of these labels are not widely available and there are also some potential methodological issues with the measurement of a "sustainability level" of a mutual fund or a company. Sustainability as a concept requires such behavior, goals and actions that allow for the continuous existence of all the elements of the system (all the stakeholders) or at least give a chance for "survival" to most. Even though this is a very simplified summary of the concept of sustainability, it helps us point out the key methodological issues connected with the concept: first of all sustainability is by definition a system issue - it is difficult to measure without the inputs concerning all the elements of the system, second it is a forward-looking concept meaning that its assessment needs to rely on predictions, and third there seem to be many ways to assess sustainability, most of which sooner or later degenerate to binary ones (sustainable/unsustainable, ESG/nonESG, etc.) or are at least interpreted as such.

There are, on the other hand, some indices for sustainability like the Morningstar Sustainability Ratings (MSR) [30] which allow for some graduality in the transition from non-sustainable to sustainable labels. It is also good to note that many ratings such as the one provided by Morningstar are intrinsically relative, i.e. they identify the "most sustainable" and the "least sustainable" units in the given set. Nothing guarantees that the most sustainable units are "sustainable enough" as well as nothing says that the least sustainable units are "not sustainable at all". It is also interesting to note that for a portfolio to obtain a Morningstar Sustainability Score, only 2/3 of its assets under management need to have the ESG risk rating. This means that the MSR might not reflect the full ESG risk and full information concerning the funds being assessed. It also considers the environmental, social and governance issues as proxies for sustainability, without an explicitly declared overall sustainability focus. Still, as evidenced also by the literature review conducted by us (see Table III), MSR is a frequently used proxy for fund sustainability.

Given the issues we have discussed above (which are just some of the issues connected with the measurement of something as complex and ill-defined as sustainability), it is not surprising that one can find research papers that do not find any relationship between funds' sustainability ratings and their performance, research that suggests the existence of a positive relationship between these two variables, but also research that points out the inability of sustainable (ESG) funds to outperform the market during non-crisis periods. Again, the prevalence of regression methods in the research is high, which only stresses the need for validation of these nonuniform findings by another approach. Given the results presented in Table III, we will further investigate the consistency of the following relationship with our data: *If the Morningstar Sustainability Rating of the fund is high, then the risk-adjusted returns are high.* Also, in this case, we will investigate the opposite relationship *If the Morningstar Sustainability Rating of the fund is high, then the risk-adjusted returns are not high.* Now that we know what relationships are expected based on the previous research, we can describe the dataset used in our analysis and also provide the fuzzy-number meanings of the linguistic terms used in the relationships to be investigated by the tools of fsQCA.

IV. Data and implied definitions of the fuzzy-number meanings of high/low values of the fund features

For our analysis, we have obtained a set of 429 mutual equity growth ESG funds from the European area from the Morningstar Mutual Fund Screener. Out of the over 31 000 mutual funds available in the database at the time of data retrieval (March 2021) we strived to get a compact sample by limiting our scope to

- "Europe Developed" or "Europe Developing" which limited the number of funds available for the analysis to 3583
- "Growth" funds ruling out funds that would be dividend-paying to simplify the performance assessment of the funds
- "Euro" as the currency to further facilitate the intercomparability of the funds and their performance
- at least three years old funds to ensure sufficient history of the analyzed funds; more specifically we required the funds to be in the database for the whole March 2018 - March 2021 period
- funds for which the MSR value is available
- equity funds; the reason for this is that other than equity funds were very infrequent in the resulting sample and their different characteristics might not be strong enough to have significant effect in the results, but might have biased the results for the equity funds.

TABLE III
SUMMARY OF THE REVIEWED PAPERS DEALING WITH THE RELATIONSHIP OF SUSTAINABILITY OF THE FUND (MEASURED IN VARIOUS WAYS) AND ITS PERFORMANCE.

Year	Paper, Authors	Declared objective(s)	Period	Data characteristics	Methodology	Results	Assumed effect[a]
2020	Steen, Moussawi and Gjolberg [31]	To analyze the relationship between the Morningstar Sustainability rating and fund performance	2014-2018	146 mutual funds domiciled in Norway.	Fama-French regression, geographical bias of the ratings considered. Sustainability measured with the MSR. Alpha as a performance measure.	Among categorized European funds (to avoid geographical bias) the performance improves in parallel with improving ESG ratings.	+
2019	Dolvin, Fulkerson and Krukover [32]	To investigate the effect of sustainable investing on investment performance.	2012-2016	1853 U.S. mutual funds.	Performance measured with Carhart alpha. Sustainability measured with the Morningstar Sustainability scores.	No difference in risk-adjusted returns between sustainable and conventional funds. However, sustainable funds limited to large-cap funds and thus can feature a higher risk and weaker diversification.	0
2016	Henke [33]	To examine the financial effect of screening ESG criteria on corporate bond fund portfolios.	2001-2014	103 socially responsible and 309 matched conventional bond mutual funds from the U.S. and Eurozone.	Regression analysis. Comparing socially responsible funds with their conventional pairs. Performance measured with risk-adjusted returns (a five-factor model). Sustainability is measured with ESG ratings based on information provided by the US Sustainable Investment Forum and the European Social Investment Forum.	Socially responsible bond mutual funds performed better than their conventional pairs annually.	+
2016	Nagy, Kassam and Lee [34]	To investigate if ESG factors of an investment affect investment performance.	2007-2015	global MSCI stock data.	Back-testing two global model portfolios that regard ESG criteria: "ESG tilt" and "ESG momentum." Alpha as a performance measure. MSCI ESG ratings as a sustainability measure.	Both tested portfolios that consider ESG criteria beat the global benchmark index MSCI World Index.	+
2014	Nofsinger and Varma [35]	To examine the performance of socially responsible funds during periods of market crisis and periods of non-crisis.	2000-2011	240 U.S. equity mutual funds and their 209 conventional pairs.	Regression analysis. CAPM, three-factor and four-factor models as performance measures.	Socially responsible mutual funds outperform their conventional pairs in periods of market crisis and underperform conventional funds during periods of non-crisis.	+/-
2005	Bello [36]	To examine the effects of socially responsible investing on portfolio diversification and fund performance.	1994-2001	42 socially responsible funds provided by Morningstar and 84 conventional funds from the U.S.	Regression analysis. Comparing socially responsible funds with their conventional pairs. Performance measured with Jensen's Alpha, Sharpe Ratio and excess standard deviation adjusted return.	There is no notable difference between the performance or diversification of socially responsible and conventional funds.	0
1993	Hamilton, Jo amd Statman [37]	To evaluate the financial effect of socially responsible investing in mutual fund performance.	1981-1990	32 socially responsible funds and 150 conventional funds.	Performance comparison between socially responsible and conventional funds. Jensen's Alpha as a performance measure. The selected funds were identified as socially responsible funds by their managers.	There is no practical difference between the performance of socially responsible and conventional funds.	0

[a] + indicates a positive relationship, - indicates a negative relationship, 0 indicates no relationship between the fund's sustainability rating and its performance; adapted from [23]

- funds having no missing values of the relevant variables (assets under management, manager tenure etc.) in the investigated period

After the selection of the dataset and ensuring that all the funds within do not have any missing values of the variables relevant for our research, the fuzzy numbers representing the meanings (denoted by the M operator) of "high", "middle" and "low" values of the variables were defined in the following way.

Fund performance measures values

For Jensen's alpha the prototype of the middle value representing "middle value" of alpha can be considered to be 0, which is the natural middle value of this variable. All values within the $(-3, 3)$ interval were considered at least partially fitting for the description "middle". These thresholds are set by the authors and can be modified if needed in future analyses. The idea of not setting the definition of "middle alpha" around the median of the data is that the alpha has a natural middle (neutral) point at zero. The minimum and maximum values of alpha were set relative to the available values of the funds. This resulted in:

- M("high alpha") $\sim (0, 3, 14.23, 14.23)$
- M("middle alpha") $\sim (-3, 0, 0, 3)$
- M("low alpha") $\sim (-12.44, -12.44, -3, 0)$

which implies

- M("not high alpha") $\sim (-12.44, -12.44, 0, 3)$
- M("not low alpha") $\sim (-3, 0, 14.23, 14.23)$.

For the Sharpe ratio, there is no natural minimum, middle or maximum value prototype. We have therefore identified the minimum, first, second and third quartile and the maximum value of the Sharpe ratios available in the given sample, which were -0.19, 0.25, 0.39, 0.63 and 1.58 respectively. We have used these values to define the meanings of "high", "middle" and "low" values of Sharpe ratio in the following way:

- M("high Sharpe ratio") $\sim (0.39, 0.63, 1.58, 1.58)$
- M("middle Sharpe ratio") $\sim (0.25, 0.39, 0.39, 0.63)$
- M("low Sharpe ratio") $\sim (-0.19, -0.19, 0.25, 0.39)$

which implies

- M("not high Sharpe ratio") $\sim (-0.19, -0.19, 0.39, 0.63)$
- M("not low Sharpe ratio") $\sim (0.25, 0.39, 1.58, 1.58)$.

It is clear that for variables without specific natural middle points, maxima or minima, the definitions of the meanings of the linguistic terms used in the investigated relationships need to be defined either relatively to the available values of the variables, or based on experience or expert knowledge.

Fund size values

Fund size was measured by assets under management (in millions of EUR). This variable has a natural minimum at 0, but no natural middle or maximum values. Therefore the first, second and third quartiles as well as the maximum value of this variable were determined: 78.39, 235.03, 664.91 and 7124.65 respectively. The meanings of "large", "middle" and "small" values of fund size were thus defined in the following way:

- M("large size") $\sim (235.03, 664.91, 7124.65, 7124.65)$
- M("middle size") $\sim (78.39, 235.03, 235.03, 664.91)$
- M("small size") $\sim (0, 0, 78.39, 235.03)$.

Manager tenure values

The length of manager tenure (measured in years) also has a natural minimum at 0, but no natural middle or maximum values. We have thus again decided to use the first, second and third quartiles as well as the maximum value of this variable, which were 3.58, 7.83, 12.08 and 23.58 respectively. The meanings of "long", "middle" and "short" values of manager tenure length were thus defined in the following way:

- M("long tenure") $\sim (7.83, 12.08, 23.58, 23.58)$
- M("middle tenure") $\sim (3.58, 7.83, 7.83, 12.08)$
- M("short tenure") $\sim (0, 0, 3.58, 7.83)$.

Sustainability rating values

The values of the MSR are always from the $\{1, 2, 3, 4, 5\}$ set, in other words, there are only five possible ratings to be assigned. As such the scale has a natural maximum, minimum and middle point which can be used for the definitions of the fuzzy-number meanings of the linguistic values used in the investigated relationships. Given the limited number of numerical values of this scale, we have decided to distinguish only between "low" and "high" sustainability defined in the following way:

- M("high sustainability") $\sim (2, 4, 5, 5)$
- M("low sustainability") $\sim (1, 1, 2, 4)$.

V. Results of the analysis using fsQCA methods

For the assumed relationships between each of the individual variables (fund size, manager tenure, fund sustainability rating) and fund performance (measured using Jensen's alpha and Sharpe ratio), we have calculated all four fuzzified consistency and coverage measures (3)-(10). To gain additional insights into the relationships between the variables, we have investigated not only the assumed relationships and their negations, but also relationships that lead to the outcome represented by the opposite linguistic term on the scale than the one that was postulated. In other words, we investigate (Long Tenure $\Rightarrow$ High risk-adjusted returns), (Long Tenure $\Rightarrow$ not High risk-adjusted returns), but also (Long Tenure $\Rightarrow$ Low risk-adjusted returns) and (Long Tenure $\Rightarrow$ not Low risk-

TABLE IV
SUMMARY OF THE RESULTS OBTAINED BY APPLYING THE F_1 - F_4 CONSISTENCY AND COVERAGE MEASURES ON THE INVESTIGATED RELATIONSHIPS.

A = Large Size, B = High Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.318	F1 consistency	0.709
F1 coverage	0.398	F1 coverage	0.355
F2 consistency	0.236	F2 consistency	0.626
F2 coverage	0.316	F2 coverage	0.290
F3 consistency	0.000	F3 consistency	0.390
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.305	F4 consistency	0.695
F4 coverage	0.381	F4 coverage	0.348
SUP1(A=>B) =	0.185	DISP1(A=>B) =	0.449
SUP0.9(A=>B) =	0.212	DISP0.9(A=>B) =	0.532
SUP0.8(A=>B) =	0.222	DISP0.8(A=>B) =	0.606
SUP0.7(A=>B) =	0.249	DISP0.7(A=>B) =	0.674
SUP0.6(A=>B) =	0.272	DISP0.6(A=>B) =	0.708
SUP0.5(A=>B) =	0.279	DISP0.5(A=>B) =	0.721
SUP0.4(A=>B) =	0.292	DISP0.4(A=>B) =	0.728
SUP0.3(A=>B) =	0.332	DISP0.3(A=>B) =	0.751
SUP0.2(A=>B) =	0.394	DISP0.2(A=>B) =	0.778
SUP0.1(A=>B) =	0.468	DISP0.1(A=>B) =	0.788
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Large Size, B = High Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.438	F1 consistency	0.590
F1 coverage	0.430	F1 coverage	0.332
F2 consistency	0.358	F2 consistency	0.510
F2 coverage	0.354	F2 coverage	0.266
F3 consistency	0.000	F3 consistency	0.152
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.424	F4 consistency	0.576
F4 coverage	0.417	F4 coverage	0.324
SUP1(A=>B) =	0.302	DISP1(A=>B) =	0.449
SUP0.9(A=>B) =	0.322	DISP0.9(A=>B) =	0.474
SUP0.8(A=>B) =	0.342	DISP0.8(A=>B) =	0.484
SUP0.7(A=>B) =	0.370	DISP0.7(A=>B) =	0.529
SUP0.6(A=>B) =	0.394	DISP0.6(A=>B) =	0.559
SUP0.5(A=>B) =	0.431	DISP0.5(A=>B) =	0.582
SUP0.4(A=>B) =	0.441	DISP0.4(A=>B) =	0.606
SUP0.3(A=>B) =	0.471	DISP0.3(A=>B) =	0.630
SUP0.2(A=>B) =	0.516	DISP0.2(A=>B) =	0.658
SUP0.1(A=>B) =	0.526	DISP0.1(A=>B) =	0.678
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Large Size, B = Low Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.293	F1 consistency	0.746
F1 coverage	0.307	F1 coverage	0.405
F2 consistency	0.196	F2 consistency	0.649
F2 coverage	0.229	F2 coverage	0.333
F3 consistency	0.000	F3 consistency	0.454
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.273	F4 consistency	0.727
F4 coverage	0.287	F4 coverage	0.394
SUP1(A=>B) =	0.132	DISP1(A=>B) =	0.551
SUP0.9(A=>B) =	0.149	DISP0.9(A=>B) =	0.591
SUP0.8(A=>B) =	0.170	DISP0.8(A=>B) =	0.638
SUP0.7(A=>B) =	0.207	DISP0.7(A=>B) =	0.679
SUP0.6(A=>B) =	0.244	DISP0.6(A=>B) =	0.703
SUP0.5(A=>B) =	0.281	DISP0.5(A=>B) =	0.719
SUP0.4(A=>B) =	0.297	DISP0.4(A=>B) =	0.762
SUP0.3(A=>B) =	0.321	DISP0.3(A=>B) =	0.793
SUP0.2(A=>B) =	0.362	DISP0.2(A=>B) =	0.830
SUP0.1(A=>B) =	0.409	DISP0.1(A=>B) =	0.851
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Large Size, B = Low Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.316	F1 consistency	0.717
F1 coverage	0.314	F1 coverage	0.401
F2 consistency	0.230	F2 consistency	0.631
F2 coverage	0.243	F2 coverage	0.329
F3 consistency	0.000	F3 consistency	0.401
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.300	F4 consistency	0.700
F4 coverage	0.297	F4 coverage	0.392
SUP1(A=>B) =	0.192	DISP1(A=>B) =	0.565
SUP0.9(A=>B) =	0.202	DISP0.9(A=>B) =	0.578
SUP0.8(A=>B) =	0.210	DISP0.8(A=>B) =	0.609
SUP0.7(A=>B) =	0.235	DISP0.7(A=>B) =	0.643
SUP0.6(A=>B) =	0.257	DISP0.6(A=>B) =	0.681
SUP0.5(A=>B) =	0.294	DISP0.5(A=>B) =	0.715
SUP0.4(A=>B) =	0.319	DISP0.4(A=>B) =	0.743
SUP0.3(A=>B) =	0.357	DISP0.3(A=>B) =	0.765
SUP0.2(A=>B) =	0.391	DISP0.2(A=>B) =	0.790
SUP0.1(A=>B) =	0.422	DISP0.1(A=>B) =	0.798
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Long Tenure, B = High Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.288	F1 consistency	0.746
F1 coverage	0.381	F1 coverage	0.395
F2 consistency	0.194	F2 consistency	0.652
F2 coverage	0.285	F2 coverage	0.331
F3 consistency	0.000	F3 consistency	0.458
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.271	F4 consistency	0.729
F4 coverage	0.359	F4 coverage	0.387
SUP1(A=>B) =	0.133	DISP1(A=>B) =	0.501
SUP0.9(A=>B) =	0.166	DISP0.9(A=>B) =	0.567
SUP0.8(A=>B) =	0.178	DISP0.8(A=>B) =	0.627
SUP0.7(A=>B) =	0.197	DISP0.7(A=>B) =	0.686
SUP0.6(A=>B) =	0.236	DISP0.6(A=>B) =	0.726
SUP0.5(A=>B) =	0.246	DISP0.5(A=>B) =	0.754
SUP0.4(A=>B) =	0.274	DISP0.4(A=>B) =	0.764
SUP0.3(A=>B) =	0.320	DISP0.3(A=>B) =	0.803
SUP0.2(A=>B) =	0.373	DISP0.2(A=>B) =	0.822
SUP0.1(A=>B) =	0.433	DISP0.1(A=>B) =	0.840
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Long Tenure, B = High Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.355	F1 consistency	0.670
F1 coverage	0.369	F1 coverage	0.399
F2 consistency	0.278	F2 consistency	0.593
F2 coverage	0.293	F2 coverage	0.332
F3 consistency	0.000	F3 consistency	0.315
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.343	F4 consistency	0.657
F4 coverage	0.356	F4 coverage	0.391
SUP1(A=>B) =	0.217	DISP1(A=>B) =	0.513
SUP0.9(A=>B) =	0.237	DISP0.9(A=>B) =	0.538
SUP0.8(A=>B) =	0.252	DISP0.8(A=>B) =	0.570
SUP0.7(A=>B) =	0.299	DISP0.7(A=>B) =	0.613
SUP0.6(A=>B) =	0.320	DISP0.6(A=>B) =	0.637
SUP0.5(A=>B) =	0.340	DISP0.5(A=>B) =	0.662
SUP0.4(A=>B) =	0.363	DISP0.4(A=>B) =	0.680
SUP0.3(A=>B) =	0.387	DISP0.3(A=>B) =	0.701
SUP0.2(A=>B) =	0.430	DISP0.2(A=>B) =	0.748
SUP0.1(A=>B) =	0.462	DISP0.1(A=>B) =	0.763
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Long Tenure, B = Low Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.346	F1 consistency	0.680
F1 coverage	0.385	F1 coverage	0.390
F2 consistency	0.252	F2 consistency	0.585
F2 coverage	0.311	F2 coverage	0.321
F3 consistency	0.000	F3 consistency	0.334
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.333	F4 consistency	0.667
F4 coverage	0.370	F4 coverage	0.383
SUP1(A=>B) =	0.189	DISP1(A=>B) =	0.499
SUP0.9(A=>B) =	0.207	DISP0.9(A=>B) =	0.534
SUP0.8(A=>B) =	0.233	DISP0.8(A=>B) =	0.555
SUP0.7(A=>B) =	0.265	DISP0.7(A=>B) =	0.609
SUP0.6(A=>B) =	0.297	DISP0.6(A=>B) =	0.658
SUP0.5(A=>B) =	0.321	DISP0.5(A=>B) =	0.679
SUP0.4(A=>B) =	0.342	DISP0.4(A=>B) =	0.710
SUP0.3(A=>B) =	0.398	DISP0.3(A=>B) =	0.735
SUP0.2(A=>B) =	0.445	DISP0.2(A=>B) =	0.767
SUP0.1(A=>B) =	0.466	DISP0.1(A=>B) =	0.793
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = Long Tenure, B = Low Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.391	F1 consistency	0.644
F1 coverage	0.410	F1 coverage	0.382
F2 consistency	0.304	F2 consistency	0.558
F2 coverage	0.337	F2 coverage	0.307
F3 consistency	0.000	F3 consistency	0.253
F3 coverage	0.000	F3 coverage	0.000
F4 consistency	0.373	F4 consistency	0.627
F4 coverage	0.391	F4 coverage	0.371
SUP1(A=>B) =	0.253	DISP1(A=>B) =	0.505
SUP0.9(A=>B) =	0.266	DISP0.9(A=>B) =	0.509
SUP0.8(A=>B) =	0.286	DISP0.8(A=>B) =	0.536
SUP0.7(A=>B) =	0.326	DISP0.7(A=>B) =	0.575
SUP0.6(A=>B) =	0.340	DISP0.6(A=>B) =	0.594
SUP0.5(A=>B) =	0.376	DISP0.5(A=>B) =	0.642
SUP0.4(A=>B) =	0.406	DISP0.4(A=>B) =	0.660
SUP0.3(A=>B) =	0.425	DISP0.3(A=>B) =	0.674
SUP0.2(A=>B) =	0.464	DISP0.2(A=>B) =	0.714
SUP0.1(A=>B) =	0.491	DISP0.1(A=>B) =	0.734
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = High Sustainability, B = High Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.357	F1 consistency	0.698
F1 coverage	0.690	F1 coverage	0.541
F2 consistency	0.263	F2 consistency	0.604
F2 coverage	0.464	F2 coverage	0.304
F3 consistency	0.000	F3 consistency	0.341
F3 coverage	0.272	F3 coverage	0.038
F4 consistency	0.329	F4 consistency	0.671
F4 coverage	0.636	F4 coverage	0.519
SUP1(A=>B) =	0.203	DISP1(A=>B) =	0.483
SUP0.9(A=>B) =	0.230	DISP0.9(A=>B) =	0.530
SUP0.8(A=>B) =	0.247	DISP0.8(A=>B) =	0.589
SUP0.7(A=>B) =	0.270	DISP0.7(A=>B) =	0.633
SUP0.6(A=>B) =	0.302	DISP0.6(A=>B) =	0.669
SUP0.5(A=>B) =	0.314	DISP0.5(A=>B) =	0.686
SUP0.4(A=>B) =	0.331	DISP0.4(A=>B) =	0.698
SUP0.3(A=>B) =	0.371	DISP0.3(A=>B) =	0.730
SUP0.2(A=>B) =	0.411	DISP0.2(A=>B) =	0.753
SUP0.1(A=>B) =	0.470	DISP0.1(A=>B) =	0.774
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = High Sustainability, B = High Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.473	F1 consistency	0.587
F1 coverage	0.718	F1 coverage	0.510
F2 consistency	0.378	F2 consistency	0.493
F2 coverage	0.498	F2 coverage	0.265
F3 consistency	0.000	F3 consistency	0.114
F3 coverage	0.344	F3 coverage	0.000
F4 consistency	0.443	F4 consistency	0.557
F4 coverage	0.672	F4 coverage	0.484
SUP1(A=>B) =	0.323	DISP1(A=>B) =	0.441
SUP0.9(A=>B) =	0.340	DISP0.9(A=>B) =	0.460
SUP0.8(A=>B) =	0.357	DISP0.8(A=>B) =	0.481
SUP0.7(A=>B) =	0.390	DISP0.7(A=>B) =	0.502
SUP0.6(A=>B) =	0.418	DISP0.6(A=>B) =	0.523
SUP0.5(A=>B) =	0.456	DISP0.5(A=>B) =	0.553
SUP0.4(A=>B) =	0.477	DISP0.4(A=>B) =	0.582
SUP0.3(A=>B) =	0.498	DISP0.3(A=>B) =	0.610
SUP0.2(A=>B) =	0.519	DISP0.2(A=>B) =	0.643
SUP0.1(A=>B) =	0.540	DISP0.1(A=>B) =	0.660
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = High Sustainability, B = Low Jensen's alpha			
	A=>B		A=>notB
F1 consistency	0.335	F1 consistency	0.727
F1 coverage	0.543	F1 coverage	0.609
F2 consistency	0.225	F2 consistency	0.616
F2 coverage	0.289	F2 coverage	0.381
F3 consistency	0.000	F3 consistency	0.392
F3 coverage	0.000	F3 coverage	0.166
F4 consistency	0.304	F4 consistency	0.696
F4 coverage	0.493	F4 coverage	0.583
SUP1(A=>B) =	0.173	DISP1(A=>B) =	0.521
SUP0.9(A=>B) =	0.179	DISP0.9(A=>B) =	0.559
SUP0.8(A=>B) =	0.194	DISP0.8(A=>B) =	0.589
SUP0.7(A=>B) =	0.234	DISP0.7(A=>B) =	0.639
SUP0.6(A=>B) =	0.268	DISP0.6(A=>B) =	0.688
SUP0.5(A=>B) =	0.297	DISP0.5(A=>B) =	0.703
SUP0.4(A=>B) =	0.312	DISP0.4(A=>B) =	0.732
SUP0.3(A=>B) =	0.365	DISP0.3(A=>B) =	0.766
SUP0.2(A=>B) =	0.411	DISP0.2(A=>B) =	0.806
SUP0.1(A=>B) =	0.441	DISP0.1(A=>B) =	0.821
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

A = High Sustainability, B = Low Sharpe ratio			
	A=>B		A=>notB
F1 consistency	0.331	F1 consistency	0.725
F1 coverage	0.507	F1 coverage	0.627
F2 consistency	0.240	F2 consistency	0.634
F2 coverage	0.283	F2 coverage	0.387
F3 consistency	0.000	F3 consistency	0.394
F3 coverage	0.000	F3 coverage	0.205
F4 consistency	0.303	F4 consistency	0.697
F4 coverage	0.464	F4 coverage	0.603
SUP1(A=>B) =	0.211	DISP1(A=>B) =	0.586
SUP0.9(A=>B) =	0.213	DISP0.9(A=>B) =	0.599
SUP0.8(A=>B) =	0.222	DISP0.8(A=>B) =	0.614
SUP0.7(A=>B) =	0.249	DISP0.7(A=>B) =	0.639
SUP0.6(A=>B) =	0.264	DISP0.6(A=>B) =	0.677
SUP0.5(A=>B) =	0.304	DISP0.5(A=>B) =	0.711
SUP0.4(A=>B) =	0.323	DISP0.4(A=>B) =	0.736
SUP0.3(A=>B) =	0.361	DISP0.3(A=>B) =	0.751
SUP0.2(A=>B) =	0.386	DISP0.2(A=>B) =	0.778
SUP0.1(A=>B) =	0.401	DISP0.1(A=>B) =	0.787
SUP0.0(A=>B) =	1.000	DISP0.0(A=>B) =	1.000
alpha-SUP =	1.000	alpha-DISP =	1.000

adjusted returns) and analogously for the other two features and the funds' performance.

Let us start with the relationship of the size of the fund and its performance. The original assumed relationship was that a large fund size should result in low risk-adjusted returns. The results for this relationship are available in the last two subtables in the top row of subtables in Table IV. We can clearly see that the consistency of the "large fund size resulting in low fund performance" (both measured by Jensen's alpha and Sharpe ratio) is much lower than the consistency of "large fund size resulting in not low performance". The values of F_1 consistencies of "not low performance" resulting from "large fund size" being higher than F_1 consistencies of "low performance" resulting from "large fund size", as well as the values of F_4 consistencies being higher than 0.5 (0.727 for alpha, 0.700 for Sharpe) suggest that there is much more evidence in favor of the *If the size of the fund is large, then its risk-adjusted returns are not low* relationship in the data than there is for the originally assumed one. The originally assumed relationship even does not have any pure non-ambivalent excess evidence in its favor meaning that there is no evidence in its favor as defined by the F_3 consistency measure (both values of F_3 consistency are zero for alpha and for Sharpe). By the same logic, looking at the first two subtables in the top row of Table IV, we can see that there is also no pure non-ambivalent excess evidence for high size resulting in high performance of the funds. Given the fact that all values of unconditional support and disproof are nonzero in the first row of subtables in Table IV and that the largest values of the unconditional support/disproof are for $\text{DISP}_1(A \Rightarrow B)$ in the Large size $\rightarrow$ Low Sharpe ratio (0.565) and in the Large size $\rightarrow$ Low Jensen's alpha (0.551) with comparatively lower values of the unconditional support $\text{SUP}_1(A \Rightarrow B)$, we can conclude that Large fund size being related to not low performance seems to be the most plausible of the investigated relationships. Note, that "not low performance" covers the "middle or high" performance in this case. The F_3 consistency of "Large fund size $\rightarrow$ not High Sharpe ratio" being rather low (0.152) prevents us from claiming that large fund size would be related with not high performance of the fund in general, though. We, however, do not see any clear support for the claim that the large funds are high-performing either. Still large fund size seems to be preventing low performance.

As far as the relationship between manager tenure and fund performance is concerned, we need to look at the middle row of subtables in Table IV. By the same logic applied here, we can see that the most viable relationship that can be found in the data is *If fund manager's tenure is high, then the risk-adjusted returns of the fund are not high*. Again, here "not high" covers "middle or low". It is, however, true to say that the relationship *If fund manager's tenure is high, then the risk-adjusted returns of the fund are not low* has a similar support by the data. Overall, we can see that high manager tenure does not seem to guarantee high performance and it also does not guarantee low performance of the fund.

Now we can focus on the bottom row of subtables in Table IV that investigates the relationships between the sustainability rating of the funds and their performance. Applying the same logic in this case, we can clearly see that the most supported relationship in the data is the one of "High MSR $\rightarrow$ not Low performance of the fund" both measured by Sharpe ratio and by Jensen's alpha. If we have a close look at the values, the F_1 consistencies and coverages are rather high for them, the F_3 consistencies are nonzero and reasonably high implying that there is pure non-ambivalent excess evidence for these relationships and there is also nonzero F_3 coverage for these relationships. These relationships are the only ones (except for "High MSR $\rightarrow$ not High Jensen's alpha") with nonzero F_3 consistency and coverage, but for the "not Low performance of the fund" the values of F_3 consistencies and coverages are such that one can see clear evidence in favor of the given relationship in the data. Given all this, we can conclude that the data supports the relationship *If the Morningstar sustainability rating of the fund is high, then the risk-adjusted returns of the fund are not low*. There is not enough evidence to conclusively prove the validity of the claims that high sustainability ratings would be related with high fund performance. The evidence in favor of claiming that high sustainability is related with not high fund performance is inconclusive.

VI. Conclusions

In this paper, we have analyzed the relationship between three selected characteristics of mutual funds, namely their size, length of their manager's tenure and their sustainability ratings, and the performance of these funds. The analysis was carried out on a sample of European growth mutual funds using the fsQCA tools, mainly the recently proposed fuzzified versions of consistencies and coverages.

Overall the strongest relationship found in the data can be expressed in general terms by the statement *If the sustainability rating of the fund is high, then its performance is not low.* This is well in line with the previous research that suggests that sustainable/responsible funds might overperform the non-sustainable ones in crises periods, but at the same time there seems to be evidence that they might underperform during calmer times (see Table III and its discussion). Our findings suggest, on the given sample and under the given definitions of the variables, that although the high sustainability rating does not guarantee the high performance of the fund, it seems to indicate that low performance of the fund is not to be expected.

There are several ways in which to continue this research. First of all this paper focused on the drivers of high performance of the European growth mutual funds and thus the potential reasons for low performance etc. were not analyzed. An analogous analysis can be performed with the intention of identifying potential sources of low performance for these funds even using the same dataset. We have also analyzed only isolated effects of single features on the performance. The fsQCA methodology allows for the investigation of combined effects (for example of the type "IF sustainability rating is High and manager tenure is Not low, THEN the performance of the fund is High"). These combined effects were left out of

the scope of this paper and can constitute a research direction that sheds more light on the drivers of the performance of mutual funds.

VII. Acknowledgement

This paper is an extended summary of the analysis and results obtained in the thesis by Fanni Welling [23].

References

[1] J. Tomar, S. Agarwal, and K. R. Chaturvedi, "Relationship between Employer Branding and Corporate Social Responsibility," in S. Agarwal, D. N. Burrell, and V. K. Solanki (eds.) *Proceedings of the 2020 International Conference on Research in Management & Technovation*, *ACSIS*, vol. 24, pp. 123-127, 2020, doi: http://dx.doi.org/10.15439/2020KM19

[2] H. D. Hai, and C. N. Huu, "The Effect of Corporate Social Responsibility Factor on the Sustainable Development of Industrial SMEs: A Case Study in Hanoi-Vietnam," in S. Agarwal, D. N. Burrell, and V. K. Solanki (eds.) *Proceedings of the 2020 International Conference on Research in Management & Technovation*, *ACSIS*, vol. 24, pp. 57-62, 2020, doi: http://dx.doi.org/10.15439/2020KM23

[3] P. C. Fiss, "Building better causal theories: A fuzzy set approach to typologies in organization research," *Acad. Manag. J.*, vol. 54, no. 2, pp. 393-420, 2011.

[4] P. C. Fiss, "A Set-Theoretical Approach to Organizational Configurations," *Acad. Manag. Rev.*, vol. 32, no. 4, pp. 1180-1198, 2007.

[5] J. Stoklasa, P. Luukka, and T. Talášek, "Set-theoretic methodology using fuzzy sets in rule extraction and validation - consistency and coverage revisited," *Inf. Sci.*, vol. 412-413, pp. 154-173, 2017, doi: 10.1016/j.ins.2017.05.042

[6] J. Stoklasa, T. Talášek, and P. Luukka, "On consistency and coverage measures in the fuzzified set-theoretic approach for social sciences: dealing with ambivalent evidence in the data," in *Proceedings of the 36th International Conference on Mathematical Methods in Economics*, 2018, pp. 521-526.

[7] M. M. Kumbure, A. Tarkiainen, P. Luukka, J. Stoklasa, and A. Jantunen, "Relation between managerial cognition and industrial performance : An assessment with strategic cognitive maps using fuzzy-set qualitative comparative analysis," *J. Bus. Res.*, vol. 114, no. June 2020, pp. 160-172, 2020, doi: 10.1016/j.jbusres.2020.04.001

[8] M. M. Kumbure, P. Luukka, A. Tarkiainen, J. Stoklasa and A. Jantunen, "An investigation of hidden shared linkages among perceived causal relationships in cognitive maps," in P. Luukka and J. Stoklasa (eds.), *Intelligent Systems and Applications in Business and Finance,* Springer, in press.

[9] S. Tomer, and G. Rana, "Green Human Resource Management: A Conceptual Study," in S. Agarwal, D. N. Burrell, and V. K. Solanki (eds.) *Proceedings of the 2020 International Conference on Research in Management & Technovation*, *ACSIS*, vol. 24, pp. 147-150, 2020, doi: http://dx.doi.org/10.15439/2020KM10

[10] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[11] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Aplications*. New Jersey: Prentice Hall, 1995.

[12] C. C. Ragin, *Redesigning Social Inquiry: Fuzzy sets and Beyond*. Chicago: University of Chicago Press, 2008.

[13] M. C. Jensen, "The Performance of Mutual Funds in the Period 1945-1964", *J. Financ.*, vol. 23, no. 2, pp. 389-416, 1968.

[14] W. F. Sharpe, "Mutual Fund Performance", *J. Bus. (Chicago, Ill.)*, vol. 39, no. 1, pp. 119-138, 1966.

[15] H. W. H. Chan, R. W. Faff, D. R. Gallagher and A. Looi, "Fund Size, Transaction Costs and Performance: Size Matters", *Aust. J. Manage.*, vol. 34, no. 1, pp. 73-96, 2009.

[16] X. Yan, "Liquidity, Investment Style, and the Relation between Fund Size and Fund Performance", *J. Financ. Quant. Anal.*, vol. 43, no. 3, pp. 741-767, 2008.

[17] J. Chen, H. Hong, M. Huang and J. D. Kubik, "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization", *Am. Econ. Rev.*, vol. 94, no. 5, pp. 1276-1302, 2004.

[18] S. E. Beckers and G. Vaughan, "Small Is Beautiful", *J. Portfolio Manage.*, vol. 27, no. 4, pp. 9-17, 2001.

[19] P. Tufano and M. Sevick, "Board structure and fee-setting in the U.S. mutual fund industry", *J. Financ. Econ.*, vol. 46, no. 3, pp. 321-355, 1997.

[20] J. Golec, "The effects of mutual fund managers' characteristics on their portfolio performance, risk and fees", *Financ. Serv. Rev. (Greenwich, Conn.)*, vol. 5, no. 2, pp. 133-147, 1996.

[21] A. F. Perold and R. S. Salomon, "The Right Amount of Assets under Management", *Financ. Analysts J.*, vol. 47, no. 3, pp. 31-39, 1991.

[22] T. F. Loeb. "Trading Cost: The Critical Link between Investment Information and Results." *Financ. Analysts J.*, vol. 39, no. 3, pp. 39-44, 1983.

[23] F. Welling, "The role of sustainability, manager tenure, and fund size in European mutual growth fund performance," BSc thesis, LUT University, 2021.

[24] R. Kjetsaa and M. Kieff, "Impact of Expenses, Turnover and Manager Tenure on Blend Fund Performance", *J. Bus. Account.*, vol. 9, no. 1, pp. 99-115, 2016.

[25] B. A. Costa, K. Jakob and G. E. Porter, "Mutual Fund Performance and Changing Market Trends 1990-2001: Does Manager Experience Matter?", *J. Investing*, vol. 15, no. 2, pp. 79-86, 2006.

[26] G. Filbeck and D. L. Tompkins, "Management Tenure and Risk-Adjusted Performance of Mutual Funds", *J. Investing*, vol. 13, no. 2, pp. 72-80, 2004.

[27] M. Brooks and D. L. Tompkins, "Mutual funds' risk adjusted performance", *J. Commercial Bank. Financ.*, Vol. 1, pp. 111-121, 2002.

[28] R. Fortin, S. Michelson and J. Jordan-Wagner, "Does mutual fund manager tenure matter?", *J. Financ. Plan.*, vol. 12, no. 7, pp. 72-79, 1999.

[29] D. Lemak and P. Satish, "Mutual Fund Performance and Managers' Terms of Service: Are There Performance Differences?" *J. Investing*, Winter 1996, pp. 59-63, 1996.

[30] Morningstar Research, "Morningstar Sustainability Rating Methodology," *Morningstar*, 2019. [Web document] [Accessed 20.3.2021]. Available: https://www.morningstar.com/content/dam/marketing/shared/research/methodology/744156_Morningstar_Sustainability_Rating_for_Funds_Methodology.pdf

[31] M. Steen, J. T. Moussawi and O. Gjolberg, "Is there a relationship between Morningstar's ESG ratings and mutual fund performance?", *J. Sustain. Financ. Invest.*, vol. 10, no. 4, pp. 349-370, 2020.

[32] S. Dolvin, J. Fulkerson and A. Krukover, "Do "Good Guys" Finish Last? The Relationship between Morningstar Sustainability Ratings and Mutual Fund Performance", *J. Investing*, vol. 28, no. 2, pp. 77-91, 2019.

[33] H. Henke, "The effect of social screening on bond mutual fund performance", *J. Bank. Financ.*, vol. 67, pp. 69-84, 2016.

[34] Z. Nagy, A. Kassam and L. Lee, "Can ESG Add Alpha? An Analysis of ESG Tilt and Momentum Strategies", *J. Investing*, vol. 25, no. 2, pp. 113-124, 2016.

[35] J. Nofsinger and A. Varma, "Socially responsible funds and market crises", *J. Bank. Financ.*, vol. 48, pp. 180-193, 2014.

[36] Z. Y. Bello, "Socially Responsible Investing and Portfolio Diversification", *J. Financ. Res.*, vol. 28, no. 1, pp. 41-57, 2005.

[37] S. Hamilton, H. Jo and M. Statman, "Doing Well While Doing Good? The Investment Performance of Socially Responsible Mutual Funds", *Financ. Analysts J.*, vol. 49, no. 6, pp. 62-66, 1993.

Author Index