Stochastic Multiattribute Acceptability Analysis: an application to the ranking of Italian regions

Greco, Salvatore and Ishizaka, Alessio and Matarazzo, Benedetto and Torrisi, Gianpiero

22 December 2015

Sigma-Mu efficiency analysis: A methodology for evaluating units through composite indicators

Salvatore Greco [a,b], Alessio Ishizaka[b], Menelaos Tasiou[c], and Gianpiero Torrisi [a,c]

[a] *Department of Economics and Business, University of Catania, Corso Italia 55, 95129 Catania, Italy*
[b] *University of Portsmouth, Centre of Operations Research and Logistics, PO1 3DE, Portsmouth, UK*
[c] *University of Portsmouth, Portsmouth Business School, PO1 3AH, Portsmouth, UK*

ABSTRACT

We propose a methodology to employ composite indicators for performance analysis of units of interest using and extending the family of Stochastic Multiattribute Acceptability Analysis. We start evaluating each unit by means of weighted sums of their elementary indicators in the whole set of admissible weights. For each unit, we compute the mean, μ, and the standard deviation, σ, of its evaluations. Clearly, the former has to be maximized, while the latter has to be minimized as it denotes instability in the evaluations with respect to the variability of weights. We consider a unit to be Pareto-Koopmans efficient with respect to μ and σ if there is no convex combination of μ and σ of the rest of the units with a value of μ that is not smaller, and a value of σ that is not greater, with at least one strict inequality. The set of all Pareto-Koopmans efficient units constitutes the first Pareto-Koopmans frontier. In the spirit of context-dependent Data Envelopment Analysis, we assign each unit to one of the sequence of Pareto-Koopmans frontiers. We measure the local efficiency of each unit with respect to each frontier, but also its global efficiency taking into account all frontiers in the $\sigma-\mu$ plane, thus enhancing the explicative power of the proposed approach. To illustrate its potential, we present a case study of 'world happiness' based on the data of the homonymous report that is annually produced by the United Nations' Sustainable Development Solutions Network.

Keywords: Data Envelopment Analysis · Composite Indicators · Sigma-Mu efficiency · Stochastic Multiattribute Acceptability Analysis · neo-Benthamite approach.

JEL Classification: C43, C44, I31.

✉ Menelaos Tasiou
menelaos.tasiou@port.ac.uk

Salvatore Greco
salgreco@unict.it

Alessio Ishizaka
alessio.ishizaka@port.ac.uk

Gianpiero Torrisi
gianpiero.torrisi@port.ac.uk

1 Introduction

In recent years, composite indicators are witnessed as increasingly popular tools for evaluating the performance of units such as countries and institutions (Becker et al., 2017). In fact, there are over 500 official composite indicators evidenced to date, mainly produced by institutions, scholars and universities, with the aim of assessing countries in a complex socio-economic phenomenon (Bandura, 2011; Yang, 2014). Understandably, their adoption by global institutions (e.g. the OECD, UN, World Bank etc.) over the past years has gradually drawn the attention of the media and policy-makers around the globe (Saltelli, 2007), and the number of applications in the literature has surged ever since (Greco et al., 2018b). This spiral of attention raises several flags on issues that are still debated in the literature, mainly regarding two stages in the construction of an indicator; namely, the weighting and aggregation. There is a wide variety of methods available in these steps, and there is no documented approach without a single drawback (Gan et al., 2017). Undeniably, the choice of the proper method lies in the developer's craftsmanship and the objective of the indicator (OECD, 2008). Nonetheless, these issues are still in great need of consideration; especially when something as crucial as a policy is to be drawn on the basis of a synthetic measure that could easily be 'manipulated' (Grupp and Schubert, 2010; Abberger et al., 2017).

A fundamental step in the construction of composite indicators regards the weighting of the elementary indicators. Very often, this point is not taken into account and an equally-weighted mean -typically the arithmetic mean (e.g. see, among others, the *Index of Economic Freedom* (Miller et al., 2018) and the *Inclusive Development Index* (Samans et al., 2018)), but sometimes also the geometric mean (see, e.g., the 2010 HDI; UNDP, 2010)- is considered, mainly due to simplicity, or a lack of framework to suggest otherwise (Freudenberg, 2003). This oversimplifying choice, however, is "*obviously convenient but also universally considered to be wrong*" (Chowdhury and Squire, 2006, p.762). By contrast, sometimes the dimensions are weighted by taking into account reasonable differences in the importance of the considered dimensions (Decancq and Lugo, 2013). Either way, at first sight, this procedure of weighting the indicators -with, or without equal weights- could appear as a neutral approach to the problem of aggregating the different dimensions, given a single, well-determined vector of weights. Of course, this implicitly assumes a representative agent (Hartley and Hartley, 2002), summing up in itself the preferences of all the individuals potentially interested in the composite indicator. However, one has to admit that, in a miscellaneous group of people, each one may assign a radically different importance to the considered dimensions and, consequently, in order to ensure that the composite indicator is meaningful, the diversity of existing viewpoints has to be considered (Decancq et al., 2013).

Undeniably, the hypothesis of the representative agent is rather stringent. Moreover, it has been long criticized in economics with the so-called *"fallacy of composition"*, proposed by Kirman (1992), who gave an example in which the representative agent disagrees with all individuals in the economy (a similar point can be found in Blackburn and Ukhov (2013), examining the relationship between individual and aggregate risk preferences in the financial markets). Besides the observation of a plurality of preferences corresponding to the individuals interested in the composite indicator; one has to take into account that each individual can be seen as a multiplicity of 'selves' that she is composed of (see, e.g., Elster, 1987). Several researchers have acknowledged the relevance of this point in economics (see, e.g., Ainslie, 2001; Schelling, 1980; McClure et al., 2004), so that even to represent an individual's preferences, we need to consider a set of weight vectors for the considered dimensions. Something similar happens in Multiple Criteria Decision Aiding (MCDA) (for an updated survey, see Greco et al., 2016). In particular, some recently-introduced MCDA models consider a plurality of value functions compatible with the preferences expressed by a decision maker (see, e.g., Greco et al., 2008, 2010; Corrente et al., 2013), or even a probability distribution in the set of value functions (see,

e.g., Corrente et al., 2016b). This can be interpreted as a plurality of selves for each individual, from the point of view that each considered value function is a specific 'self'. Similar arguments hold for multi-prior models proposed for decisions under uncertainty, where each individual takes a decision considering a plurality of probability distributions on the state of the worlds (see, for example, Gilboa and Schmeidler, 1989; Bewley, 2002; Gilboa et al., 2010; Cerreia-Vioglio et al., 2018).

On a more general note, there is a growing interest in multi-utility models (see e.g. Evren and Ok, 2011; Giarlotta and Greco, 2013) representing preferences of individuals with a set of utility functions $\mathscr{U}$, such that an alternative a is at least as good as alternative b if the value assigned to a is not smaller than the value assigned to b by all utility functions $u \in \mathscr{U}$. Multi-utility models are appreciated, because they permit to represent incomplete preferences, which are considered more realistic for individual preferences than the classical models assuming perfect comparability among all alternatives (see e.g. Aumann (1962), but also Von Neumann and Morgenstern (1944, pp.19-20)). This point is also related to the question of interpersonal comparability. In fact, apart from the extremely egalitarian approach (Rawls, 2009), between the two extreme positions of perfect comparability, i.e. between single individual preferences (see e.g. Marshall, 1961), and of absolute interpersonal incomparability (see e.g. Robbins, 1935); there could be an intermediate position, such as the one proposed by Sen (1970), which is based on the idea that the preferences of each individual are represented by a set of welfare functions rather than a single one. In the context of composite indicators, in which the utility function is represented by the weighted sum of single indicators, these arguments suggest to abandon the idea of a single, allegedly well-defined weighting of dimensions corresponding to a single utility function.

Indeed, by taking into account the whole set of admissible weight vectors, one can consider the whole spectrum of preferences of individuals, as well as multiple selves within each individual interested in the composite indicator. With respect to the domain of composite indicators, this approach was recently proposed by Greco et al. (2018a) using Stochastic Multiattribute Acceptability Analysis (SMAA) (Lahdelma et al., 1998; Lahdelma and Salminen, 2001). More specifically, by considering a probability distribution on the set of feasible weight vectors, SMAA reveals the probability that a unit attains a given ranking position, as well as the probability that a given unit is better than another. It is worth noting that the above consideration of multiple selves also suggests to consider a plurality of weight vectors for composite indicators not only at the level of a collectivity of individuals, but also at the level of single individuals. In this case the typical results of SMAA, which are the probability that an alternative a is the most preferred, or the probability that a is preferred to alternative b, can be interpreted in terms of random choices (Luce, 1959; McFadden, 1981). In fact, this is perfectly in line with the prevailing application of SMAA within MCDA, that is to support decision problems with a single decision-maker.

The use of SMAA in this context seems alluring. Indeed, the difficulty that is intrinsically associated to the choice of a single, well-defined vector of preferences is moderated through the use of SMAA. Yet, this comes at the expense of the ability to produce a single composite indicator value. Moreover, up to this point, SMAA has been put to use to provide ordinal information, whereas composite indicators are cardinal in nature (Booysen, 2002). This motivated us to consider another use of SMAA in conjunction with renowned methods in the field of Operations Research to construct composite indicators that encapsulate a more holistic evaluation in a single value that, instead of a ranking, provide information about the magnitude of the performance of each alternative. We call this method "$\sigma-\mu$ efficiency analysis".

Last but not least, let us now point out two remarks related again to the above recalled interpretation of a plurality of weight vectors in terms of a plurality of utility functions for an individual with multiple selves:

1. The proposed concept of $\sigma-\mu$ efficiency can be also applied to represent evaluations of single individuals whose preferences can be represented in terms of a plurality of weight vectors. In this case, the set of

considered weight vectors can be elicited by interacting with the decision maker, following the basic idea of robust ordinal regression (Greco et al., 2008; Kadziński and Tervonen, 2013). Note that, in this context, it is also possible to elicit a distribution of probabilities in the space of feasible weight vectors (Corrente et al., 2016b). Of course this probability can be used to define the mean and the variance of the approach we are proposing.

2. The proposed methodology can be also seen as a different SMAA approach, that is, instead of computing the probability that each considered alternative could obtain a given rank position and the probability of being preferred to another alternative; one could compute the mean μ and the standard deviation σ of the values assigned by the weighted sum to each alternative. Afterwards, μ and σ could be used to arrive at a single overall evaluation using the overall (global) efficiency measure that we propose in this study. In fact, this represents a new method in the SMAA family (Tervonen and Figueira, 2008) that we call $\sigma-\mu$—SMAA.

The aim of this paper is to introduce a methodology for constructing composite indicators that we call "σ-μ efficiency analysis", illustrating its potential in a case study of world happiness, based on the homonymous report by Helliwell et al. (2017). In what follows: Section 2 describes in more detail the issues of weighting in the construction of a composite indicator. Section 3 introduces the σ-μ efficiency analysis, followed by a brief didactic example given in Section 4 to illustrate its application on real-world data on a step-by-step basis. Section 5 contains the case study of world happiness and a robustness analysis of the obtained results. Section 6 contains a discussion about further considerations and generalization of the proposed approach. Section 7 provides conclusive remarks and future direction of research.

2 Composite indicators: Some methodological issues

2.1 Weighting dimensions in composite indicators

The use of composite indicators is constantly growing by the day. This can be witnessed by an ever-increasing number of composite measures produced every year by global institutions, academics and media around the world (Bandura, 2011; Yang, 2014), despite the severe criticism these synthetic measures received in their inauguration (see Sharpe, 2004, pp.9-11). This is mainly owed to their irresistible property of summarizing complex phenomena with a sole number that can be easily interpreted as a benchmark (Saisana et al., 2005). Of course, this can be seen as both an asset and a liability at the same time. More specifically, lack of transparency in their construction allows significant room for 'manipulation' (Grupp and Schubert, 2010; Abberger et al., 2017). The reason is that there exists a sequence of steps in the construction of a composite indicator and, admittedly, different choices in each step might radically alter the final outcome. As one would expect, not a single step in the construction process lacks criticism (Booysen, 2002); nonetheless, the paramount critique lies in two stages, namely the weighting and aggregation of the underlying indicators. The former refers to the process of declaring the importance of indicator dimensions, whereas the latter refers to the final synthesis of the overall measure. In this paper we are engrossed with the former, thus the discussion of this section will solely revolve around it.

The basic model of composite indicators is the following. There exists a set of units $I = \{1,\dots,n\}$ to be evaluated with respect to the set of dimensions $J = \{1,\dots,m\}$, the values of which are x_{ij}. For each unit $i \in I$, the vector $\mathbf{x}_i = [x_{i1},\dots,x_{im}]$ collects the values assigned to that unit in the dimensions from J. To each dimension $j \in J$, a weight, w_j, is attached such that $w_j \geq 0$ for all $j \in J$ and $\sum_{j=1}^{m} w_j = 1$. Given a weight vector $\mathbf{w} = [w_1,\dots,w_m]$, the composite indicator assigns the following value to each unit $i \in I$:

$$CI(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{n} x_{ij} w_j.$$

The authoritative *Handbook on Constructing Composite Indicators* (OECD, 2008) lists several approaches regarding the weighting procedure in the construction of a composite indicator (for a recent review of existing methodologies, criticism and proposed solutions, see Greco et al., 2018b), with equal weighting being the most common scheme on the grounds of equal importance (Paruolo et al., 2013, p.627). This, however, also appears to be the most criticized (Decancq and Lugo, 2013). More specifically, assignment of equal weights can be seen as a convenient solution of the last resort that is "*obviously convenient, but also universally considered to be wrong*" (Chowdhury and Squire, 2006, p.762). It is mainly used when there is no scientific basis to justify peculiar weighting (OECD, 2008), or when an alleged objectivity, or simplicity (Babbie, 1995; Freudenberg, 2003) is desired; the latter often justified using the principle that is known as 'Occam's Razor' (Hopkins, 1991, as cited in Cherchye et al. (2007)). Nonetheless, this rationale could be contradicted for the following reasons. First, equal weights could be reasonably considered subjective as they are considered objective (see, e.g., Ray, 2008; Mikulić et al., 2015). The reason being equal weights consist a specific weight vector that could represent a specific type of person who equally prefers all attributes of a composite indicator. Second, as far as the uncertainty around the lack of a framework to support differential weighting is concerned, there are more realistic solutions to equal weights that have been proposed in the literature to deal with this issue (see, e.g., Doumpos et al., 2016, 2017; Greco et al., 2018a). Third, in response to the argument corresponding to Occam's parsimony (i.e. "*since it is probably impossible to obtain agreement on weights, the simplest arrangement [equal weighting] is the best choice*", Hopkins (1991, p.1471)), we could argue that, perhaps, a better principle to abide by would be Einstein's parsimony that "*things should be made as simple as possible - but no simpler*[1]". In addition, in contradicting Babbie (1995)'s argument that equal weighting is the virtue of simplicity; Cherchye et al. (2007, p.141) add: "*our own opinion regarding Babbie's statement is, hence, the other way around: the burden of the proof should be on equal weighting whereas the norm should be differential [benefit of the doubt] weighting*".

Other past solutions revolve around two sets of approaches, often characterized as 'subjective', and 'objective' respectively (Booysen, 2002). The former set involves participatory techniques such as the Budget Allocation Process (BAP) (see OECD, 2008, p.96) or Analytic Hierarchy Process (AHP) (Saaty, 1977, 1980). These engage a single, or a number of stakeholders (e.g. a panel of experts) to decide upon the weights to be assigned, according to their beliefs/expertise (hence, the term 'subjective'). These approaches appear to be ideal where a well-defined framework for national policy exists (see Munda, 2005b). Still, they might yield radically different results (see Saisana et al., 2005, p.314, for a comparison between AHP and BAP), while in the presence of many criteria, they can give decision-makers 'cognitive stress' that is amplified in the AHP due to the number of pairwise comparisons required (Ishizaka and Nemery, 2013). The second set of approaches are awarded their epithet ('objective') from the fact that they do not rely on human judgment, but rather on the use of data-driven techniques (e.g. Multiple linear regression analysis, Principal Component Analysis (Pearson, 1901), Factor Analysis (Spearman, 1904), or Data Envelopment Analysis (Charnes et al., 1978)). These have been conceptually criticized for being disoriented from the objective at hand, or that they provide non-reasonable weight vectors (Decancq and Lugo, 2013), while at the same time they have a few methodology-related drawbacks that need to be addressed (Greco et al., 2018b).

Irrespective of classification though (i.e. 'subjective', or 'objective'), the above approaches produce a sin-

[1] A reputed paraphrase of Einstein's following phrase: "*It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience*" (Einstein, 1934, p.165) (also known as Einstein's razor) by Rogers Sessions (1950).

gle weight vector overall -or, in the case of DEA, a single weight vector for each unit- that is then used in the stage of aggregation to synthesize the composite indicator. While this procedure is common practice in the domain of composite indicators (OECD, 2008), either unwittingly or deliberately, the developer assumes that the obtained set of weights is representative of the whole population interested in the composite indicator. Understandably, one could argue that this is a rather stringent assumption, as in a miscellaneous group of people, each individual may assign a radically different importance to each dimension, and the representativeness assumption may be only valid for a very small portion of the population or it could even become infeasible overall. Decancq et al. (2013) argue that when a policy-maker chooses a weight-vector, there are several individuals who are inevitably 'worse-off'. This situation highly resembles the case of the representative agent in economics (see e.g. Hartley and Hartley, 2002), which has been long criticized in the literature by Kirman (1992). Kirman provides an example in which, quaintly to his title, the 'representative' agent disagrees with all the individuals in the economy. Acknowledging this confounding situation, Greco et al. (2018a) recently proposed the use of SMAA (Lahdelma et al., 1998; Lahdelma and Salminen, 2001) to take into account the whole set of possible weight vectors in the evaluation process. According to the authors, the standard procedure of choosing a single weight vector produces a single, allegedly 'representative' ranking for the evaluated units that "*amalgamates different preferences in the population*" (p.6). SMAA essentially permits the inclusion of several potential viewpoints in the decision-making process, e.g. in the form of weight vectors, enriching this way the single ranking that is obtained from a single preference. In terms of output, probabilistic rankings are assigned to each unit, expressing its probability to be ranked first, second etc.; or, its probability to be preferred to another unit. The use of SMAA in this exercise seems alluring, whether it is applied to take into account potential representations of citizens' preferences (Greco et al., 2018a), or simply to deal with uncertainty in the lack of information about decision-makers' preferences (see e.g. Doumpos et al., 2016, 2017). Since SMAA is the fundamental framework that we take into account in this paper, we present it in more detail in the following subsection.

2.2 Stochastic multiattribute acceptability analysis (SMAA)

SMAA offers a solid solution to real-world decision-making that is surrounded by any source of uncertainty. In the domain of composite indicators, such an example would involve a decision-maker that is unable to provide the parameters required for the evaluation process (see e.g. Doumpos et al., 2016, 2017). In this paper we are engrossed with the step of weighting, hence, we are solely considering this source of uncertainty. Essentially, SMAA takes it into account by considering a probability distribution f_w over the space of all weight vectors

$$W = \left\{ \mathbf{w} = [w_1, \ldots, w_m] : w_j \geq 0, j = 1, \ldots, m, \sum_{j=1}^{m} w_j = 1 \right\}.$$

Understandably, if a different importance has to be assigned to the dimensions from J, the space W is transformed accordingly. For instance, if dimension $j_{(1)}$ is the most important, $j_{(2)}$ is the second most important and so on until the least important, e.g. $j_{(m)}$, and we have to assign higher weights to the more important dimensions; then the space W is transformed as follows:

$$W = \left\{ \mathbf{w} = [w_1, \ldots, w_m] : w_{j_{(1)}} \geq w_{j_{(2)}} \geq \ldots w_{j_{(m)}} \geq 0, j = 1, \ldots, m, \sum_{j=1}^{m} w_j = 1 \right\}.$$

SMAA (Lahdelma et al., 1998; Lahdelma and Salminen, 2001) proposes to compute the following meaningful values:

- The rank acceptability index $b_i^r, i \in I$ and $r = 1, \ldots, n$, that gives the probability that randomly picking a weight vector $\mathbf{w} \in W$, unit i is r^{th} in the final rank provided by CI;
- The central weight vector $\mathbf{w}_i$ that, in case $b_i^1 \neq 0$, gives the barycenter of the set of weight vectors for which unit $i \in I$ is the the best according to CI;
- The pairwise winning index Tervonen et al. (2009); Leskinen et al. (2006) $p_{ii'}$ that gives the probability that, according to CI, unit i is better than unit i' randomly picking a weight $\mathbf{w}$ from W.

SMAA was only recently introduced in the field of composite indicators. More specifically, Doumpos et al. (2016) use it to deal with the uncertainty arising from the lack of information regarding the parameters to be used in the evaluation process of some financial institutions. Using 10,000 uniformly distributed random weights and marginal value functions, the authors evaluate the overall financial strength of 1,200 commercial banks through an additive value function setting, given five financial characteristics from the CAMEL framework. A similar application is found in Doumpos et al. (2017), comparing the overall financial strength of Islamic and conventional banks. Greco et al. (2018a) propose the use of SMAA in the context of composite indicators as a way to deal with the issue of representativeness inherent in the single weight vector. The authors evaluate the 20 regions of Italy based on 65 socio-economic criteria. By enlarging the space of weight vectors, they refrain from the classic setting of the univocal set of weights, including 1,000,000 uniformly distributed weight vectors. In an alternative interpretation, this could be potentially seen as an expression of several decision-makers' preferences, e.g. ranging from policymakers to citizens, regarding the importance of the indicator's dimensions. This involvement of a 'multiplicity of participants', or even 'selves' (see Elster, 1987) could indeed be enriching to consider in such an exercise. Quoting Munda (2005a, p.132): *"when science is used in policy, the appropriate management of quality has to be enriched to include this multiplicity of participants and perspectives"*. While the author's point refers to the context of a sustainability policy exercise (regarding the objectives and scales of such an analysis and the set of dimensions to be used in the evaluation process), the intended allegory is astonishingly fit to the context of the decision-makers' number and preferences respectively.

3 The σ-μ efficiency

We stand by the principle that a meaningful composite indicator should ideally reflect a multiplicity of viewpoints. Technically speaking, this can be achieved in the weighting stage, in which individuals that the indicator concerns can participate, by expressing their preferences on the importance of indicator dimensions. These individuals could constitute different clusters, e.g. experts, policy-makers, or even citizens at whom policies are addressed. Therefore, the main driver of this concept refrains from the classic scheme of a single, allegedly representative weight vector in the construction of an indicator, by taking into account all these individuals' viewpoints. In the past, this has been feasible with the use of SMAA (see, e.g., Greco et al., 2018a). Still, SMAA comes at the expense of a single composite indicator, given the fact that its outputs are probabilistic indicators for a unit to be ranked at a given place, or to dominate/be dominated from another unit. In Section 3.1, we re-consider the framework of SMAA to obtain the two main parameters of the $\sigma - \mu$ approach that serve as its starting point. In Section 3.2 we present the definitions of dominance as well as the measures of local and global efficiencies that we obtain with the proposed approach. Section 3.3 shows how this approach can be used in real-life problems, as well as how it compares to other measures of efficiency in the literature.

3.1 The starting point: μ and σ

We re-consider the framework of SMAA, and for each unit, $i \in I$, we synthesize the distribution of its composite indicators values, $CI(\mathbf{x}_i, \mathbf{w})$, by computing its mean value μ_i and standard deviation σ_i in the weight vector space W as follows:

$$\mu_i = \int_{\mathbf{w} \in W} f_w(\mathbf{w}) CI(\mathbf{x}_i, \mathbf{w})\, d\mathbf{w}, \tag{1}$$

$$\sigma_i = \sqrt{\int_{\mathbf{w} \in W} f_w(\mathbf{w}) \left[CI(\mathbf{x}_i, \mathbf{w}) - \mu_i \right]^2 \, d\mathbf{w}}. \tag{2}$$

As it will become clear towards the end of this section, but mainly in Section 4, where we go through the steps using a didactic example, the integrals defining the values of μ_i and σ_i can be approximated in a Monte Carlo simulation environment. This analogy between the original inferential problem and such techniques (e.g. bootstrap) is greatly described in Daraio and Simar (2007a, p.53), in terms of "*an analogy between the real world, where we want to make inference about [a parameter of interest] but most of the desired quantities are unknown, and the bootstrap world, where we mimic the real world but where everything is known and so can be computed or simulated by Monte-Carlo methods*".

These two -μ and σ- will be our parameters of interest and the main input to the remaining part of the proposed approach that we present in this section. Understandably, μ_i is intended to be maximized, because it represents the average evaluation of a unit taking into account the variability of the weight vectors $\mathbf{w}$. Instead, σ_i has to be minimized, as it exhibits the instability in the overall evaluations with respect to the variability of weights. In fact, as it will forthwith become apparent, the rationale for minimizing σ is manifold.

On abstract and general grounds, it is worth stressing that -once the variety of perspectives on the dimensions under analysis has been fully considered in the preceding weighting stage- the dispersion is a measure such that the lower it is the better. Thus, the dispersion of the CIs is an inverse measure of the robustness of the performance of a given unit as to the weighting choice. On a conceptual ground, it somehow reflects how balanced is the performance of a given unit among the considered dimensions. If its performance depends on one or very limited number of dimensions to a greater extent, that unit will achieve very different overall performances according to those dimensions being valued most or least in relative terms (i.e. according to different vectors of weights). The dependence on a given (eventually) favorable vector of weights is something that needs to be minimized in the construction of an overall efficiency measure, in order to pursuit robustness in the evaluation process.

The above argument about the opportunity of the methodological choice to minimize σ can be expanded on economic grounds. For example, assuming that the evaluation exercise involves the creation of a composite indicator intended to measure multidimensional well-being in an attempt to go beyond GDP (Stiglitz et al., 2010), our approach can be interpreted in the following neo-Benthamite perspective (see e.g. Collard, 2006). The value given by the composite indicator when the weight vector representing a given individual is adopted can be seen as the "happiness" of that individual. Consequently, the distribution of the values assumed by the composite indicators computed in the space of the considered weight vectors can be seen as an estimate of the distribution of the well-being among the considered population. In this perspective, the average, μ, and the standard deviation, σ, of the distribution can be seen as two parameters describing it. Moreover, if we suppose that the distribution is approximately normal (which is reasonable, considering the relatively great number of weight vectors we extract with a random sampling), then, σ and μ unambiguously characterize the distribution. In this context, μ should be clearly maximized because multiplying μ for the number of individuals in the considered population we get an estimate of the sum of individual "happiness".

Since Bentham's social welfare function (SWF) is simply additive with equal weights, substituting the

mean to the the actual values will not change the overall SWF level. Instead, σ can be seen as a measure of inequality in the distribution of well-being in the population, which is an important issue in the "GDP economics" discussion (see e.g. Piketty, 2014). Moreover, the argument about the perverse effects of excessive levels of inequality has been connected to the recent financial crisis using the 'suspension bridge' figurative narrative (see e.g. Reich, 2010). Thus, the discussion on inequality with respect to the distribution of well-being seems to us quite relevant in this neo-Benthamite "beyond GDP economics" perspective. In this respect, the standard deviation, σ, can be regarded just as a common measure of inequality used in the economic literature (Atkinson, 1970). Once transposed to the multidimensional well-being setting, the dispersion between different CIs maintains its conceptual nature of inequality. Consistent with this conceptual framework, we argue that σ has to be minimized. Put differently, expanding to the multidimensional setting at hand, Atkinson (2015, p.9)'s argument with respect to the single measure of inequality based on income is: "[We are] not seeking to eliminate all differences in economic outcomes. [We are] not aiming for total equality. Indeed, certain differences [...] may be quite justifiable. Rather, the goal is to reduce inequality [...]".

Therefore, it is reasonable that, *ceteris paribus*, the performance of units showing higher levels of dispersion will be considered worse than the performance of units registering lower levels of dispersion around the average well-being of the hypothetical community under investigation. Of course, this comparative static can be extended to include the dynamic case accordingly. Indeed, building upon Barro and Sala-i Martin (1992)'s seminal contribution in terms of (σ-)convergence of GDP, a given set of units could, in principle, be observed and evaluated at regular intervals (e.g. years) to check whether a more balanced multidimensional performance (still according to a variety of different weighting choices) is occurring over time.

On the same premises, a higher dispersion of the measure of performance -as a result of an unbalanced endowment along the considered dimensions- is undesirable when dealing with capital endowment. For example, Hansen (1965, p.13), with reference to the case of regional development, argued that "*persons benefited most by SOC [Social Overhead Capital] may migrate to other regions in the absence of supplementary policy measure*". More recently, Martin (2011, p.14), in analyzing the resilience of UK regions to economic shocks, pointed out how an unbalanced economic structure and, "*especially the relative dependence on production industry, is generally regarded as having a major influence on the sensitivity of regional economies to recessionary shock*". A similar argument has been made by Collins et al. (2017) with reference to the effects of smartness on resilience at city level. Indeed, the study of Collins et al. (2017) shows that the unbalance between different dimensions of 'smartness' does increase the cities' vulnerability to shocks. Hence, for example, in measuring the competitiveness of these regions via a CI considering the different economic sectors, the unbalance towards the production industry clearly has to be penalized. In terms of our proposed measure, the heavy dependence on the production industry would result in higher levels of dispersion generated by extremely high [low] CIs depending on the weights randomly assigning a relative higher [lower] importance to this sector. Nonetheless, this high dispersion will be taken into account by the methodological choice setting the minimization of σ as an objective of the evaluation exercise.

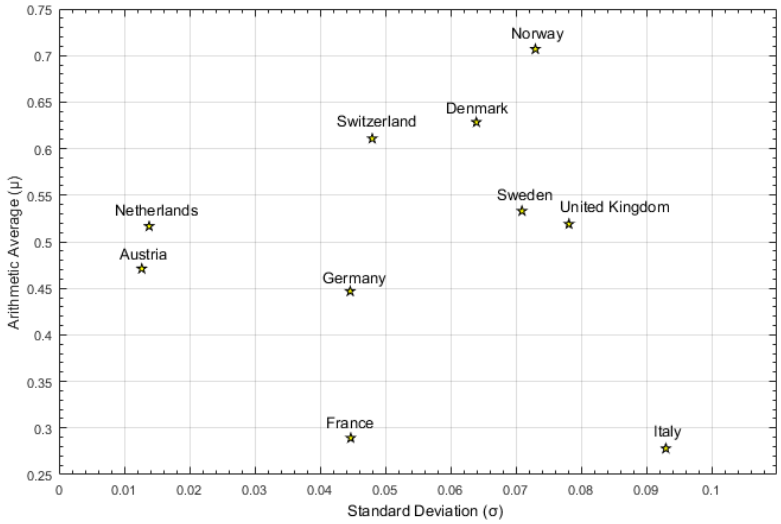
3.2 Defining dominance relationships: Local and global efficiencies

Consideration of the mean value and the standard deviation along with the related dominance and efficiency concepts clearly reminds the Markowitz mean-variance analysis (Markowitz, 1952), which formed the foundations of modern portfolio theory (Elton et al., 2009). Following his influential theory, taking into account the mean, μ_i, and the standard deviation, σ_i, one can draw a plane that units $i \in I$ are plotted on, pending evaluation. To be consistent with the proposed concept of $\sigma - \mu$ efficiency analysis, we will refer to this throughout the text as 'The $\sigma - \mu$ plane', which is illustrated in Figure 1 and shows the standard deviation σ (on the horizontal axis) and the mean μ (on the vertical axis) of ten European countries with respect to the

data of the 2017 World Happiness Report (WHR) (Helliwell et al., 2017) that will be detailed in Section 4. One can define a $\sigma-\mu$ Pareto dominance relation on the set of units I as follows: for all $i, i' \in I$, unit i is Pareto dominating unit i' if $\mu_i \geq \mu_{i'}$ and $\sigma_i \leq \sigma_{i'}$, with at least one of the two inequalities being strict. A unit $i \in I$ is $\sigma-\mu$ Pareto efficient if there is no other unit dominating it. The set of all Pareto efficient units constitutes the Pareto frontier.

Figure 1: The $\sigma-\mu$ plane

Units $i \in I$ are plotted on the plane with coordinates (σ_i,μ_i). The $\sigma-\mu$ analysis hereby presented concerns ten EU countries evaluated with respect to the data of the 2017 World Happiness Report (WHR) (Helliwell et al., 2017) as explained in Section 4.



However, we are not only interested in finding dominating solutions (i.e. alternatives lying on the Pareto-efficiency frontier), but in measuring the efficiency of each unit with respect to the frontier. In the domain of Operations Research this naturally leads to the consideration of Data Envelopment Analysis (DEA) (Charnes et al., 1978; Cooper et al., 2011), which brings us to acknowledge another definition of efficiency, taking into account this time the possibility to combine different units. This permits us to define a concept stricter than $\sigma-\mu$ Pareto efficiency that was defined above: That is the $\sigma-\mu$ Pareto-Koopmans efficiency (Charnes and Cooper, 1962). In particular, a unit $i \in I$ is $\sigma-\mu$ Pareto-Koopmans efficient if there is no convex combination of $\mu_{i'}$ and $\sigma_{i'}$ of the remaining units, $i' \neq i$, with a mean value μ that is not smaller, and a standard deviation σ that is not greater, with at least one of these inequalities being strict. Formally, a unit $i \in I$ is $\sigma-\mu$ Pareto-Koopmans efficient if for all vectors $[\lambda_{i'}, i' \neq i]$, with $\lambda_{i'} \geq 0$, for all $i' \neq i$ and $\sum_{i' \neq i} \lambda_{i'} = 1$, neither (3) nor (4) hold:

$$\sum_{i' \neq i} \lambda_{i'} \mu_{i'} > \mu_i \text{ and } \sum_{i' \neq i} \lambda_{i'} \sigma_{i'} \leq \sigma_i \tag{3}$$

$$\sum_{i' \neq i} \lambda_{i'} \mu_{i'} \geq \mu_i \text{ and } \sum_{i' \neq i} \lambda_{i'} \sigma_{i'} > \sigma_i. \tag{4}$$

The set of all $\sigma-\mu$ Pareto-Koopmans efficient units constitutes the $\sigma-\mu$ Pareto-Koopmans frontier. The membership of a unit $i \in I$ to the Pareto-Koopmans efficiency frontier can be verified with a direct or an indirect procedure described below.

The direct procedure verifies that there exists no unit -obtained as linear combination of mean $\mu_{i'}$ and standard deviation $\sigma_{i'}$- dominating unit i. This is obtained by considering the following LP problem:

$$
\begin{aligned}
&\varepsilon_i^* = \text{Max } \varepsilon \\
&s.t. \\
&\begin{cases}
\sum_{i' \neq i} \lambda_{i'} \mu_{i'} \geqslant \mu_i + \varepsilon \\
\sum_{i' \neq i} \lambda_{i'} \sigma_{i'} \leqslant \sigma_i - \varepsilon \\
\lambda_i' \geqslant 0, \ \forall i' \neq i \\
\sum_{i' \neq i} \lambda_{i'} = 1
\end{cases}
\end{aligned}
$$

where a unit, i, is $\sigma - \mu$ Pareto-Koopmans efficient if $\varepsilon_i^* \leqslant 0$.

The indirect procedure to test the $\sigma - \mu$ Pareto-Koopmans efficiency requires to consider the following LP problem:

$$
\begin{aligned}
&\delta_i^* = \text{Max } \delta \\
&s.t. \\
&\begin{cases}
\alpha \mu_i - \beta \sigma_i \geqslant \alpha \mu_{i'} - \beta \sigma_{i'} + \delta, \ \forall i' \neq i \\
\alpha, \beta \geqslant 0 \\
\alpha + \beta = 1
\end{cases}
\end{aligned} \tag{5}
$$

which can be interpreted as follows. An evaluation $\alpha \mu_{i'} - \beta \sigma_{i'}$, with $\alpha, \beta \geqslant 0$ and $\alpha + \beta = 1$, is assigned to all units $i' \in I$. The non-negative coefficient α for the mean $\mu_{i'}$ and the non-positive coefficient β for the standard deviation $\sigma_{i'}$ are coherent with the idea that $\mu_{i'}$ is intended to be maximised and $\sigma_{i'}$ is intended to be minimised. Therefore, ideally, the greater $\alpha \mu_{i'} - \beta \sigma_{i'}$, the better the unit i' performs with respect to $\mu_{i'}$ and $\sigma_{i'}$. The LP problem verifies whether a pair (α, β) exists, for which unit $i \in I$ receives an evaluation that is not worse than the remaining units, $i' \neq i$, that is if $\alpha \mu_i - \beta \sigma_i \geqslant \alpha \mu_{i'} - \beta \sigma_{i'} + \delta$, $\forall i'$, with a non-negative value of δ. This happens if $\delta_i^* \geqslant 0$, which, for the units belonging to the $\sigma - \mu$ Pareto-Koopmans efficiency frontier, represents the margin that can be subtracted from the overall evaluation $\alpha \mu_i - \beta \sigma_i$ of unit i maintaining the maximality of its evaluation with respect to all other units $i' \neq i$. For all units $i \in I$ that do not belong to the $\sigma - \mu$ Pareto-Koopmans efficiency frontier, the greater the absolute value of δ_i^*, the greater the margin that has to be added to $\alpha \mu_i - \beta \sigma_i$, in order to attain the evaluation $\alpha \mu_{i'} - \beta \sigma_{i'}$ of the units belonging to the $\sigma - \mu$ Pareto-Koopmans efficiency frontier. In this sense, the value of δ_i^* can be interpreted as a measure of efficiency of unit $i \in I$ with the following characteristics:

- if δ_i^* is non-negative, then unit i is efficient, with higher values of δ_i^* indicating greater efficiency for i,
- if δ_i^* is non-positive, then unit i is inefficient, with higher values of $|\delta_i^*|$ indicating greater inefficiency for i.

For this reason, in the following we shall refer to δ_i^* as the $\sigma - \mu$ Pareto-Koopmans efficiency score of unit i.

The following proposition enunciates the equivalence between the direct and the indirect test of the $\sigma - \mu$ Pareto-Koopmans efficiency.

Proposition 1. $\delta_i^* \geqslant 0$ if and only if $\varepsilon_i^* \leqslant 0$

Proof. Let us start by proving that if $\delta_i^* \geqslant 0$ then $\varepsilon_i^* \leqslant 0$.

If $\delta_i^* \geqslant 0$, then there exists $\alpha, \beta \geqslant 0$, with $\alpha + \beta = 1$, for which:

$$
\alpha \mu_i - \beta \sigma_i \geqslant \alpha \mu_{i'} - \beta \sigma_{i'} \text{ for all } i' \neq i.
$$

Therefore, for all $\lambda = [\lambda_{i'}, i' \neq i]$ with $\lambda_{i'} \geqslant 0$, for all $i' \neq i$, and $\sum_{i' \neq i} \lambda_{i'} = 1$, we have:

$$\lambda_{i'}(\alpha\mu_i - \beta\sigma_i) \geqslant \lambda_{i'}(\alpha\mu_{i'} - \beta\sigma_{i'}) \text{ for all } i' \neq i \tag{6}$$

By (6) we can get the following:

$$\sum_{i' \neq i} \lambda_{i'}(\alpha\mu_i - \beta\sigma_i) \geqslant \sum_{i' \neq i} \lambda_{i'}(\alpha\mu_{i'} - \beta\sigma_{i'}), \qquad \text{and, consequently,}$$

$$\alpha\mu_i - \beta\sigma_i \geqslant \alpha \sum_{i' \neq i} \lambda_{i'}\mu_{i'} - \beta \sum_{i' \neq i} \lambda_{i'}\sigma_{i'}.$$

This implies that the following condition is not verified

$$\begin{cases} \sum_{i' \neq i} \lambda_{i'}\mu_{i'} \geqslant \mu_i \\ \sum_{i' \neq i} \lambda_{i'}\sigma_{i'} \leqslant \sigma_i \end{cases}$$

with at least one strict inequality. This amounts to the Pareto-Koopmans efficiency of unit i, so that we have $\varepsilon^* \leq 0$. Thus, we proved that if $\delta_i^* \geqslant 0$, then $\varepsilon_i^* \leqslant 0$. Let us now prove that if $\varepsilon_i^* \leqslant 0$, then $\delta_i^* \geqslant 0$.
For a given unit, i, let us consider the pair (σ_i, μ_i) and the two following sets:

- the set $P^+(\sigma_i, \mu_i)$ of all the pairs $(\sigma, \mu) \in \mathbf{R}_+^2$ Pareto dominating (σ_i, μ_i), that is

$$P^+(\sigma_i, \mu_i) = \{(\sigma, \mu) \in \mathbf{R}_+^2 : \sigma \leq \sigma_i \text{ and } \mu \geq \mu_i \text{ with at least one strict inequality}\}$$

- the set $P^-(\sigma_i, \mu_i)$ given by the convex hull of the pairs $(\sigma_{i'}, \mu_{i'})$ with $i' \neq i$, that is

$$P^-(\sigma_i, \mu_i) = \left\{ \left(\sum_{i' \neq i} \lambda_{i'}\mu_{i'}, \sum_{i' \neq i} \lambda_{i'}\sigma_{i'} \right) : \lambda_{i'} \geq 0 \text{ for all } i' \neq i \text{ and } \sum_{i' \neq i} \lambda_{i'} = 1 \right\}.$$

Let us note that the condition $\varepsilon_i^* \leqslant 0$ implies that (σ_i, μ_i) is Pareto-Koopmans efficient. This means that there exists no pair $(\sigma, \mu) \in \mathbf{R}_+^2$ being a convex combination of the pairs $(\sigma_{i'}, \mu_{i'}) \in \mathbf{R}_+^2$, $i' \neq i$ that is dominating (σ_i, μ_i). As the set of pairs $(\sigma, \mu) \in \mathbf{R}_+^2$ dominating (σ_i, μ_i) is $P^+(\sigma_i, \mu_i)$ and the set of convex combinations of the pairs $(\sigma_{i'}, \mu_{i'})$, $i' \neq i$, is $P^-(\sigma_i, \mu_i)$, the Pareto-Koopmans efficiency of (σ_i, μ_i) amounts to the condition that $P^+(\sigma_i, \mu_i)$ and $P^-(\sigma_i, \mu_i)$ are disjoint. Let us point out that both $P^+(\sigma_i, \mu_i)$ and $P^-(\sigma_i, \mu_i)$ are convex sets in $\mathbf{R}^2$. Therefore, for the hyperplane separating theorem (see e.g. Boyd and Vandenberghe (2004), there must be a hyperplane separating $P^+(\sigma_i, \mu_i)$ from $P^-(\sigma_i, \mu_i)$ in the $\sigma - \mu$ space. In fact, this means that there exists a straight line $\alpha\mu - \beta\sigma = \gamma$, such that:

$$\alpha\mu - \beta\sigma > \gamma, \qquad \text{for all } (\sigma, \mu) \in P^+(\sigma_i, \mu_i), \text{ and}$$

$$\alpha\mu - \beta\sigma < \gamma, \qquad \text{for all } (\sigma, \mu) \in P^-(\sigma_i, \mu_i).$$

For contradiction, suppose now that $\delta_i^* < 0$. This means that for all $\alpha, \beta \geq 0$ we have

$$\alpha\mu_i - \beta\sigma_i < \alpha\mu_{i'} - \beta\sigma_{i'}$$

for at least one $i' \neq i$. Thus, for all $\gamma \in \mathbf{R}$

$$\alpha\mu_i - \beta\sigma_i > \gamma$$

implies

$$\alpha\mu_{i'} - \beta\sigma_{i'} > \gamma$$

for at least one $i' \neq i$. But $(\sigma_{i'}, \mu_{i'}) \in P^-(\sigma_i, \mu_i)$ and therefore, there cannot exist any hyperplane

$$\alpha\mu - \beta\sigma = \gamma$$

separating $P^+(\sigma_i, \mu_i)$ from $P^-(\sigma_i, \mu_i)$. Thus, in this case the pair (σ_i, μ_i) is not $\sigma - \mu$ Pareto-Koopmans efficient. So, if $\varepsilon_i^* \leqslant 0$ and, consequently (σ_i, μ_i) is efficient, then $\delta_i^* \geqslant 0$.

□

The $\sigma - \mu$ Pareto-Koopmans efficiency δ_i^* of unit $i \in I$ refers to the $\sigma - \mu$ Pareto-Koopmans efficiency frontier. However, for a unit that is quite remote from the $\sigma - \mu$ Pareto-Koopmans efficiency frontier, it might not be very meaningful to compare it with units of that frontier, as they could be seen as potentially implausible benchmarks. Instead, it could be useful to compare these remote units with their counterparts that are closer to them in the $\sigma - \mu$ plane, and as such, constitute more realistic benchmarks. This suggests taking into consideration the idea of a sequence of efficiency frontiers considered within the celebrated evolutionary multi-objective optimization algorithm NSGA-II (Deb et al., 2002).

A first sequence of $\sigma - \mu$ efficiency frontiers can be defined by taking into consideration the Pareto dominance. In this perspective, the set of all $\sigma - \mu$ Pareto-efficient units constitutes the first $\sigma - \mu$ Pareto efficiency frontier, denoted by PF_1. Removing PF_1 from I and computing again the $\sigma - \mu$ Pareto efficiency frontier for the remaining units, we get the second $\sigma - \mu$ Pareto-efficiency frontier denoted by PF_2. The third $\sigma - \mu$ Pareto efficiency frontier, PF_3, and the following ones can be computed analogously.

The sequence of Pareto efficiency frontiers $PF_1, PF_2, \ldots, PF_p$ based on the concept of Pareto dominance is used in NSGA-II (Deb et al., 2002). However, for the sake of our analysis, an analogous sequence of efficiency frontiers based on the concept of Pareto-Koopmans dominance seems more appropriate. The idea of a series of Pareto-Koopmans frontiers has been originally introduced by Seiford and Zhu (2003) as "*context-dependent*" data envelopment analysis. It was developed to show the 'attractiveness' or 'progress' of each evaluated DMU, according to each frontier in the sequence. The reason being is that the authors assume each efficiency frontier (or 'level') to be an alternative 'evaluation context' that, measuring the 'attractiveness' of each unit from, greatly facilitates identifying DMUs with outstanding performance, or simply to differentiate between efficient DMUs. In the spirit of their study, we suggest decomposing the set of evaluated DMUs into a sequence of Pareto-Koopmans frontiers that illustrate the $\sigma - \mu$ efficient DMUs on each level. We call the efficiency frontiers of this new sequence first $\sigma - \mu$ Pareto-Koopmans efficiency frontier, denoted by PKF_1, second $\sigma - \mu$ Pareto-Koopmans efficiency frontier, denoted by PKF_2, and so on and so forth. Let us denote by $\mathbf{PKF} = \{PKF_1, \ldots, PKF_p\}$ the set of all the $\sigma - \mu$ Pareto-Koopmans efficiency frontiers. For each unit $i \in I$, and for each $\sigma - \mu$ Pareto-Koopmans efficiency frontier $PKF_k \in \mathbf{PKF}$, we can define a 'local' $\sigma - \mu$ Pareto-Koopmans efficiency δ_{ik} with respect to PKF_k as follows:

$$
\begin{aligned}
&\delta_{ik} = \text{Max } \delta \\
&s.t. \\
&\begin{cases}
\alpha\mu_i - \beta\sigma_i \geqslant \alpha\mu_{i'} - \beta\sigma_{i'} + \delta, \ \forall i' \in I \setminus \bigcup_{h=1}^{k-1} PKF_h \\
\alpha, \beta \geqslant 0 \\
\alpha + \beta = 1
\end{cases}
\end{aligned} \tag{7}
$$

The above LP problem verifies whether there exists a pair (α, β), for which unit $i \in I$ receives an evaluation $\alpha\mu_i - \beta\sigma_i$ which is not worse than the analogous evaluation of the rest of the units $i' \in I \setminus \bigcup_{h=1}^{k-1} PKF_h$, that is, all the units i' belonging to the k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier, or to a better $\sigma-\mu$ Pareto-Koopmans efficiency frontier. This happens if $\delta_{ik} \geqslant 0$. Instead, if $\delta_{ik} < 0$, then unit i belongs to a $\sigma-\mu$ Pareto-Koopmans efficiency frontier worse than PKF_k, that is, $i \in PKF_h$ with $h = k+1, \ldots, p$. The interpretation of δ_{ik} with respect to the k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier is analogous to the interpretation of δ_i^* with respect to the overall $\sigma-\mu$ Pareto-Koopmans efficiency frontier. More precisely, for the units in the k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier or better, $\delta_{ik} \geqslant 0$ represents the margin that can be subtracted from the overall evaluation $\alpha\mu_i - \beta\sigma_i$ of unit i maintaining an evaluation that is superior to all units in the k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier or worse. Instead, for all units $i \in I$ belonging to the k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier or worse, the absolute value of $\delta_i^* < 0$ represents the margin that has to be added to $\alpha\mu_i - \beta\sigma_i$, in order to obtain the same evaluation of at least one unit belonging to k-th $\sigma-\mu$ Pareto-Koopmans efficiency frontier or better. Therefore, as δ_i^* constitutes an efficiency measure with respect to the overall $\sigma-\mu$ Pareto-Koopmans efficiency frontier (that, in fact, corresponds to the first $\sigma-\mu$ Pareto-Koopmans efficient frontier), δ_{ik} constitutes an efficiency measure with respect to the overall k^{th} $\sigma-\mu$ Pareto-Koopmans efficiency frontier. For this reason, in the following we shall refer to δ_{ik} as $\sigma-\mu$ Pareto-Koopmans efficiency of unit i with respect to the k^{th} frontier.

The following proposition gives a simple, yet useful result with respect to the $\sigma-\mu$ Pareto-Koopmans efficiency corresponding to the k^{th} frontier.

Proposition 2. The $\sigma-\mu$ Pareto-Koopmans efficiency respects the $\sigma-\mu$ Pareto dominance, that is, for all $i, i' \in I$ if $\mu_i \geqslant \mu_{i'}$ and $\sigma_i \leqslant \sigma_{i'}$, then $\delta_{ik} \geqslant \delta_{i'k}$ for any $k = 1, \ldots, p$.

Proof. As $\mu_i \geqslant \mu_{i'}$ and $\sigma_i \leqslant \sigma_{i'}$, $\alpha\mu_i - \beta\sigma_i \geqslant \alpha\mu_{i'} - \beta\sigma_{i'}$ for all $\alpha, \beta \geqslant 0$ with $\alpha + \beta = 1$. Consequently,

$$
\alpha\mu_{i'} - \beta\sigma_{i'} \geqslant \alpha\mu_{i''} - \beta\sigma_{i''} + \delta
$$

implies

$$
\alpha\mu_i - \beta\sigma_i \geqslant \alpha\mu_{i''} - \beta\sigma_{i''} + \delta
$$

for any $i'' \in I$ and any $\delta \in \mathbf{R}$. Therefore

$$
\alpha\mu_{i'} - \beta\sigma_{i'} \geqslant \alpha\mu_{i''} - \beta\sigma_{i''} + \delta_{i'k}, \ \forall i'' \in I \setminus \bigcup_{h=1}^{k-1} PKF_h
$$

implies

$$
\alpha\mu_i - \beta\sigma_i \geqslant \alpha\mu_{i''} - \beta\sigma_{i''} + \delta_{i'k}, \ \forall i'' \in I \setminus \bigcup_{h=1}^{k-1} PKF_h.
$$

Consequently, since δ_{ik} is the maximum δ satisfying

$$\alpha\mu_i - \beta\sigma_i \geqslant \alpha\mu_{i''} - \beta\sigma_{i''} + \delta,\ \forall i'' \in I \setminus \bigcup_{h=1}^{k-1} PKF_h,$$

we have to conclude that $\delta_{ik} \geqslant \delta_{i'k}$. □

Augmenting the above analysis and the classic concept of context-dependent DEA, we may proceed to a more holistic evaluation as follows. To all units $i \in I$, we can assign an overall, 'global' $\sigma-\mu$ Pareto-Koopmans efficiency score, denoted by sm_i, that reflects its efficiency with respect to all frontiers from **PKF**, as follows:

$$sm_i = \sum_{k=1}^{p} \delta_{ik}. \tag{8}$$

The following corollary of Proposition 2 ensures that overall $\sigma-\mu$ Pareto - Koopmans efficiency score sm_i respects the $\sigma-\mu$ Pareto dominance.

Proposition 3. For all $i, i' \in I$ if $\mu_i \geqslant \mu_{i'}$ and $\sigma_i \leqslant \sigma_{i'}$, then $sm_{ik} \geqslant sm_{i'k}$.

Proof. By Proposition 2: $\mu_i \geqslant \mu_{i'}$ and $\sigma_i \leqslant \sigma_{i'}$ implies $\delta_{ik} \geqslant \delta_{i'k}$ for all $k = 1, \ldots, p$. Consequently, we have

$$sm_i = \sum_{k=1}^{p} \delta_{ik} \geqslant \sum_{k=1}^{p} \delta_{i'k} = sm_{i'}.$$

□

3.3 Applying σ-μ analysis to real life problems and alternative measures of efficiency

In this section we provide a couple remarks related to the application of our approach in real life problems, as well as some definitions of efficiency. In particular, we refer to the technical parts of how our approach can be applied in real life problems and how our proposed measure of efficiency compares with other respective measures. For a non-technical, step-by-step analysis on real-world data, we refer the reader to Section 4, where we provide a didactic example using a sub-set of the data set analysed in its entirety as a case study in Section 5.

As usual for the other indicators of SMAA, the integrals defining the mean value μ_i and the standard deviation σ_i, $i \in I$, can be approximated by numerical methods or via the use of a Monte-Carlo simulation, which, as noted in Daraio and Simar (2005, 2007a) (as applied to the computation of the m, or a-order efficiency measures), is a usual and convenient way to avoid numerical integration. In fact, as the authors acknowledge (Daraio and Simar, 2005, p.103), "*the quality of the approximation can be tuned*" by increasing the number of simulations (in our particular case, this would refer to the number of random draws of the weight vectors). Therefore, using a random sampling of q vectors of weights - with q being a relatively large number; for instance, following the suggestions of Tervonen and Lahdelma (2007), q could equal 10,000- we may approximate the two parameters of interest. The q random extracted weight vectors $\mathbf{w_h} = [w_{1h}, \ldots, w_{mh}], h = 1, \ldots, q$ can be collected in the following $m \times q$ **RW** matrix:

$$\underset{m\times q}{\mathbf{RW}} = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1q} \\ w_{21} & w_{22} & \cdots & w_{2q} \\ \vdots & \vdots & \ldots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mq} \end{pmatrix}$$

Using the weight vector matrix **RW**, a composite indicator $CI(\mathbf{x}_i, \mathbf{w}_h)$ can be computed for each unit $i \in I$ and each weight vector $\mathbf{w}_h$, and the obtained results can be ordered in the following $n \times q$ matrix **CI** shown below:

$$\underset{n\times q}{\mathbf{CI}} = \begin{pmatrix} CI(\mathbf{x}_1,\mathbf{w}_1) & CI(\mathbf{x}_1,\mathbf{w}_2) & \cdots & CI(\mathbf{x}_1,\mathbf{w}_q) \\ CI(\mathbf{x}_2,\mathbf{w}_1) & CI(\mathbf{x}_2,\mathbf{w}_2) & \dots & CI(\mathbf{x}_2,\mathbf{w}_q) \\ \vdots & \vdots & \cdots & \vdots \\ CI(\mathbf{x}_n,\mathbf{w}_1) & CI(\mathbf{x}_n,\mathbf{w}_2) & \cdots & C(\mathbf{x}_n,\mathbf{w}_q) \end{pmatrix}$$

Using the values collected in **CI**, for each unit $i \in I$ one can compute the approximated values $\widetilde{\mu}_i$ and $\widetilde{\sigma}_i$ for the mean μ_i and the standard deviation σ_i as follows:

$$\widetilde{\mu}_i = \frac{1}{q}\sum_{h=1}^{q} CI(\mathbf{x}_i,\mathbf{w}_h), \quad \widetilde{\sigma}_i = \sqrt{\frac{1}{q}\sum_{h=1}^{q}(CI(\mathbf{x}_i,\mathbf{w}_h)-\widetilde{\mu_i})^2}.$$

It is worth noting that, when it comes to real-world applications, the existence of outliers is a constant struggle and an issue that appears more often than not (Hawkins, 1980). The presence of outliers in a working data set could seriously impact the obtained estimators of local and global efficiency measures respectively. In such case, the preceded analysis could greatly benefit from established robust frontier techniques (e.g. see, among others, the studies of Simar and Wilson, 1998; Daraio and Simar, 2005, 2007b). In this study we will consider the use of 'partial' frontier techniques, such as the m-order frontiers (Cazals et al., 2002; Daraio and Simar, 2005) to obtain robust estimators for the local and global $\sigma-\mu$ efficiencies. An extended discussion and application is presented in sub-section 5.1.

Last but not least, before concluding this section, let us comment on the concept of efficiency we are proposing, comparing it with other efficiency measures proposed in the literature. First, note that we are considering a non-parametric frontier approach for the "production set" Ψ of pairs (σ,μ). In fact, in our approach, Ψ is the set of all pairs (σ,μ) obtained as convex combination of pairs (σ_i,μ_i), $i=1,\dots,n$, that is

$$\Psi = \left\{ \left(\sum_{i=1}^{n}\lambda_i\mu_i, \sum_{i=1}^{n}\lambda_i\sigma_i \right) : \lambda_i \geq 0, i=1,\dots,n, \text{ and } \sum_{i=1}^{n}\lambda_i = 1 \right\},$$

which has the following efficient frontier:

$$\widehat{\Psi} = \left\{ (\sigma,\mu) : \text{ there is no } (\sigma',\mu') \in \Psi \text{ such that } (\sigma',\mu') \neq (\sigma,\mu), \sigma' \leq \sigma \text{ and } \mu' \geq \mu \right\}.$$

The Pareto-Koopmans efficiency δ_i^* we compute can be interpreted as a distance from the efficient frontier $\widehat{\Psi}$. Indeed, we can imagine to scalarize the vectors (σ,μ) introducing the scalarization function $F_{\alpha,\beta}(\sigma,\mu) = \alpha\mu-\beta\sigma, \alpha,\beta \geq 0, \alpha+\beta=1,$, measuring the distance $D\left((\sigma,\mu),\widehat{\Psi}\right)$ between (σ_i,μ_i), $i=1,\dots,n$, and the efficient frontier $\widehat{\Psi}$ as:

$$D\left((\sigma_i,\mu_i),\widehat{\Psi}\right) = min_{(\sigma',\mu')\in\widehat{\Psi}} F_{\alpha,\beta}(\sigma_i,\mu_i) - F_{\alpha,\beta}(\sigma',\mu'),$$

and, finally, taking into account all the feasible pairs (α,β) we get:

$$\delta_i^* = min_{\alpha,\beta\geq 0,\alpha+\beta=1} D\left((\sigma_i,\mu_i),\widehat{\Psi}\right).$$

In fact, practically all the measures of efficiency proposed in the literature can be expressed in terms of a distance from a frontier. In this sense, the Debreu-Farrell efficiency measure (Debreu, 1951; Farrell, 1957) gives the radial distance of the point with respect to the efficiency frontier, which in the context of the $\sigma-\mu$–efficiency analysis amounts to the following two efficiency measures:

- a μ-oriented efficiency measure that provides the value $\theta_\mu(\sigma_i,\mu_i)$, which shall be multiplied by the average μ_i to permit unit i to become Pareto-Koopmans $\sigma-\mu$—efficient, that is:

$$\theta_\mu(\sigma_i,\mu_i)=min\{\theta|(\sigma_i,\theta\mu_i)\in\widehat{\Psi}\}, \tag{9}$$

  so that, the smaller $\theta_\mu(\sigma_i,\mu_i)$, the more efficient is unit i that can be considered Pareto-Koopmans efficient if $\theta_\mu(\sigma_i,\mu_i)=1$;

- a σ-oriented efficiency measure that provides the value $\theta_\sigma(\sigma_i,\mu_i)$ to be multiplied by the standard deviation σ_i to permit unit i becoming Pareto-Koopmans $\sigma-\mu$—efficient, that is:

$$\theta_\sigma(\sigma_i,\mu_i)=max\{\theta|(\theta\sigma_i,\mu_i)\in\widehat{\Psi}\}, \tag{10}$$

  so that, the greater $\theta_\sigma(\sigma_i,\mu_i)$, the more efficient is unit i that can be considered Pareto-Koopmans efficient if $\theta_\sigma(\sigma_i,\mu_i)=1$.

Of course, in such case the LP problem formulation for the μ and σ-oriented efficiency measures (eq.9 &10 respectively) would be the following:

$$\theta_i^\mu=\text{Max }\theta \qquad s.t. \qquad \begin{cases}\theta\mu_i\leq\sum_{j=1}^{n}\lambda_j\mu_j\\ \sigma_i\geq\sum_{j=1}^{n}\lambda_j\sigma_j\\ \lambda_j\geqslant 0\\ \sum\lambda_j=1\end{cases} \tag{11a}$$

$$\theta_i^\sigma=\text{Min }\theta \qquad s.t. \qquad \begin{cases}\mu_i\leq\sum_{j=1}^{n}\lambda_j\mu_j\\ \theta\sigma_i\geq\sum_{j=1}^{n}\lambda_j\sigma_j\\ \lambda_j\geqslant 0\\ \sum\lambda_j=1\end{cases} \tag{11b}$$

while, in the spirit of Andersen and Petersen (1993), one could compute the 'super-efficiency' of each unit not only with respect to the first, but with respect to each Pareto-Koopmans frontier in the sequence (e.g. 'lifting' each time the units lying on a PKF from the constraints and re-computing the LP formulation). This would permit to have an efficiency measure in the $[0,1]$ space for local efficiencies, and in the $[0,\infty)$ space for global efficiencies. Yet, the drawback associated with these measures of efficiency is that, in our proposed model, we consider a twofold kind of a trade-off between μ and σ (see Section 6 for a discussion of this point) that is hereby lost.

4 The σ-μ efficiency analysis step by step: A didactic example

The present section illustrates the application of $\sigma-\mu$ efficiency analysis with a concise didactic example. We consider a sample of the dataset supplied by the 2017 World Happiness Report (WHR) (Helliwell et al., 2017) that will be analyzed in its entirety as a case study in Section 5. The WHR provides an evaluation of life satisfaction in more than 150 countries, based on citizens' responses to a Gallup World Poll survey. The report further supplies data on six key variables, analysing their relation with life satisfaction. For this didactic example, we take into consideration a sub-set of ten European countries (namely, *Austria, Denmark, France, Germany, Italy, Netherlands, Norway, Sweden, Switzerland and United Kingdom*) for the latest available year (data regarding the year 2016) to be evaluated through $\sigma-\mu$ efficiency analysis. For the sake of simplicity, we

only consider three of the six key variables, and more precisely, *GDP per capita, Social support* and *Perceptions of corruption*. We report these in Table 1.

Table 1: Raw and normalized values of the considered dimensions

	Raw Data				Normalized values		
Country	**Log of GDP per capita**	**Social support**	**Perceptions of corruption**	**Country**	**Log of GDP per capita**	**Social support**	**Corruption free**
Austria	10.69	0.93	0.52	Austria	0.48	0.49	0.44
Denmark	10.68	0.95	0.21	Denmark	0.47	0.70	0.71
France	10.54	0.88	0.62	France	0.33	0.18	0.35
Germany	10.70	0.91	0.45	Germany	0.49	0.34	0.51
Italy	10.43	0.93	0.90	Italy	0.23	0.50	0.11
Netherlands	10.76	0.93	0.43	Netherlands	0.54	0.49	0.52
Norway	11.07	0.96	0.41	Norway	0.84	0.74	0.54
Sweden	10.74	0.91	0.25	Sweden	0.53	0.38	0.68
Switzerland	10.92	0.93	0.30	Switzerland	0.70	0.50	0.63
United Kingdom	10.57	0.95	0.46	United Kingdom	0.37	0.70	0.50
Average	10.71	0.93	0.46				
Standard Deviation	0.17	0.02	0.19				

Data: 2017 World Happiness Report (WHR), obtained from: http://worldhappiness.report/ed/2017/. The data regard the year 2016. The detailed description and the sources of the considered dimensions can be found in Helliwell et al. (2017, p.17).

Normalization is an essential part of data aggregation to avoid adding-up "*apples and oranges*" (OECD, 2008, p.27). The reason is that indicators often come in a variety of ranges or scales that might render them incomparable in the stage of aggregation (Freudenberg, 2003). According to the author, the most common approach is standardization due to its desirable characteristics that we forthwith quote:
It converts all variables to a common scale and assumes a "normal" distribution; it has an average of zero, meaning that it avoids introducing aggregation distortions stemming from differences in variable means. In the other approaches, the scaling factor is the range of the distribution, rather than the standard deviation, which means that extreme values can have a large effect on the composite indicator (Freudenberg, 2003, p.11).
We start by standardizing the raw data reported in Table 1. As Booysen (2002, p.123) argues, "*standard scores can be further adjusted if calculations yield awkward values*". Adjustment of these values is in fact a reasonable exercise. De Muro et al. (2011) choose to adjust these values around the range $[70,130]$ with the value of a 100 being a good reference point (mean around which the standard deviations will revolve). In their spirit, Greco et al. (2018a, see online Appendix A.2) choose a different adjustment range for the standardized values. In particular, they set it to $[0,1]$, with 0.5 being the mean around which the standard deviations will revolve. Values falling outside this range (3 standard deviations away from the mean) will be replaced with the lower or upper bound accordingly, as they could generally be considered extreme given that within this range lie 99.73% of the values in the case of a normal distribution, and 89% of the values in the case of any distribution (Chebyshev's inequality). We will hereby adopt this normalization that we describe in the following:

Let us denote by y_{ij}, $i \in I, j \in J$ the raw value assumed for unit i with respect to dimension j. For each

dimension $j \in J$, the mean value M_j and the standard deviation s_j can be computed as follows:

$$M_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}, \qquad s_j = \sqrt{\frac{\sum_{i=1}^{n}(y_{ij} - M_j)^2}{n}}.$$

Using the mean M_j and the standard deviation s_j, for each $i \in I$ and $j \in J$ we obtain the z-score :

$$z_{ij} = \frac{y_{ij} - M_j}{s_j}.$$

Finally, we compute the normalized values x_{ij} as follows:

$$x_{ij} = \begin{cases} 0, & \text{if } y_{ij} \leqslant M_j - 3s_j \\ 0.5 + \dfrac{z_{ij}}{6}, & \text{if } M_j - 3s_j < y_{ij} < M_j + 3s_j \\ 1, & \text{if } y_{ij} \geqslant M_j + 3s_j \end{cases}$$

The normalization is applicable to positively-oriented dimensions, that is, dimensions for which the greater the raw value the better (e.g. *GDP per capita* and *Social Support*). Instead, for negatively-oriented dimensions, for which the greater the raw value the worse for a unit's performance (e.g. *Perception of corruption*), the normalization is formulated as follows:

$$x_{ij} = \begin{cases} 0, & \text{if } y_{ij} \geqslant M_j + 3s_j \\ 0.5 - \dfrac{z_{ij}}{6}, & \text{if } M_j - 3s_j < y_{ij} < M_j + 3s_j \\ 1, & \text{if } y_{ij} \leqslant M_j - 3s_j \end{cases}$$

Let us explain the general idea behind this normalization. Let us denote by y_{j*} and y_j^* the worst and best values respectively that are taken under consideration, such that, beyond these values we consider the evaluation y_{ij} with respect to dimension $j \in I$ an outlier. This means that, if the dimension j is positively-oriented, then $y_{j*} < y_j^*$, and all the values $y_{ij} \leq y_{i*}$ are assigned a value $x_{ij} = 0$, as well as all the values $y_{ij} \geq y_{i*}$ are assigned a value $x_{ij} = 1$. Instead, if the dimension j is negatively-oriented, then $y_{j*} > y_j^*$, and all the values $y_{ij} \leq y_{i*}$ are assigned a value of $x_{ij} = 1$, while all the values $y_{ij} \geq y_{i*}$ are assigned a value of $x_{ij} = 0$. We consider as outlier a value y_{ij} which extends $\gamma \times s_j$ beyond/above the mean M_j, and, since we hereby fixed $\gamma = 3$ (though, of course, other values of γ can be assigned according to the nature of the problem), this amounts to $y_{j*} = M_j - 3s_j$ and $y_j^* = M_j + 3s_j$ if j is positively-oriented, and $y_{j*} = M_j + 3s_j$ and $x_j^* = M_j - 3s_j$ if j is negatively-oriented. Then, in case the value of y_{ij} lies between the values of y_{j*} and y_j^*, it can be normalized as follows (where $\pm$ means $+$ in case j is positively-oriented and $-$ in case j is negatively oriented, and vice versa for $\mp$):

$$x_{ij} = \frac{y_{ij} - y_{j*}}{y_j^* - y_{j*}} = \frac{y_{ij} - (M_j \mp 3s_j)}{(M_j \pm 3s_j) - (M_j \mp 3s_j)} =$$

$$\frac{y_{ij} - M_j \pm 3s_j}{\pm 6s_j} = 0.5 \pm \frac{y_{ij} - M_j}{6s_j} = 0.5 \pm \frac{z_{ij}}{6s_j}.$$

If the value of y_{ij} lies outside the interval of y_{j*} and y_j^*, then the normalized value of y_{ij} (i.e. x_{ij}) is either 0 or 1 as explained above.

With respect to the creation of the weight vector matrix **RW**, in this didactic example we consider the following two scenarios, where $w_{GDP}, w_{Soc}, w_{Corr}$ denote weights for *GDP per capita*, *social support* and

perception of corruption respectively:

- Scenario 1: No definite ranking importance for the three considered dimensions, so that the set of feasible weight vectors is

$$\mathbf{W} = \{[w_{GDP}, w_{Soc}, w_{Corr}] : w_{GDP} \geqslant 0, w_{Soc} \geqslant 0, w_{Corr} \geqslant 0, w_{GDP} + w_{Soc} + w_{Corr} = 1\};$$

- Scenario 2: Social support is more important than perception of corruption that in turn is more important than GDP per capita, so that the set of feasible weight vectors is

$$\mathbf{W} = \{[w_{GDP}, w_{Soc}, w_{Corr}] : w_{Soc} \geqslant w_{Corr} \geqslant w_{GDP} \geqslant 0, w_{GDP} + w_{Soc} + w_{Corr} = 1\}.$$

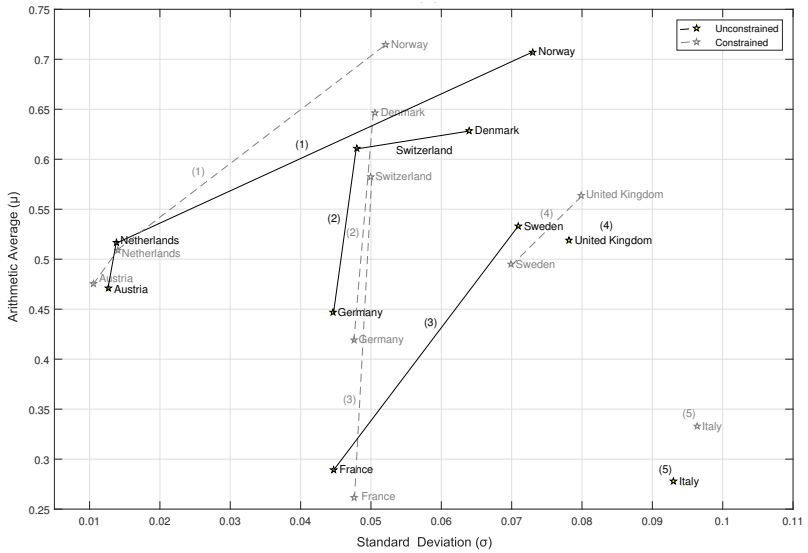
For both scenarios, a set of 10,000 weight vectors $\mathbf{w}_h$, $h = 1, \ldots, 10,000$, was randomly sampled from a uniform distribution on the feasible set of weight vectors $\mathbf{W}$ and collected in the matrix $\mathbf{RW} = [w_{jh}, j = 1,2,3, h = 1, \ldots, 10,000]$. The weight vectors from $\mathbf{RW}$ and the normalized values x_{ij}, $i = 1, \ldots, 10$, $j = 1,2,3$, are then used to compute the composite indicators:

$$CI(\mathbf{x}_i, \mathbf{w}_h) = w_{GDP} x_{i,GDP} + w_{Soc} x_{i,Soc} + w_{Corr} x_{i,Corr}, \; h = 1, \ldots, 10,000.$$

Using the values $CI(\mathbf{x}_i, \mathbf{w}_h)$, $i = 1, \ldots, 10, h = 1, \ldots, 10,000$, the approximation of the mean value $\widetilde{\mu}_i$ and the standard deviation $\widetilde{\sigma}_i$ of composite indicators were calculated for each considered country. For the sake of simplicity, we refer to them as μ_i and σ_i, respectively. These two measures are reported for both considered scenarios in Table 2 and plotted, along with the respective Pareto-Koopmans frontiers, on Figure 2.

Figure 2: The $\sigma - \mu$ plane in the two scenarios

Black colour represents $\sigma - \mu$ efficiency analysis output in the unconstrained case (scenario 1), grey colour represents respective output in the constrained case (scenario 2). Numbers in parentheses denote respective $\sigma - \mu$ Pareto-Koopmans efficiency frontier (PKF_i).



The $\sigma - \mu$ Pareto-Koopmans local efficiencies δ_{ik} of the considered countries with respect to the different $\sigma - \mu$ Pareto-Koopmans efficiency frontiers are given in Table 3. In both examined scenarios, the

Table 2: Evaluating the units with $\sigma-\mu$ under the two alternative scenarios

	Scenario 1			**Scenario 2**		
	Unconstrained weights			Constrained weights		
Country	$\boldsymbol{\mu_i}$	$\boldsymbol{\sigma_i}$	$\mathbf{sm_i}$	$\boldsymbol{\mu_i}$	$\boldsymbol{\sigma_i}$	$\mathbf{sm_i}$
Austria	0.471	0.013	0.338	0.475	0.011	0.281
Denmark	0.628	0.064	0.561	0.646	0.051	0.514
France	0.289	0.045	0.076	0.262	0.048	0.037
Germany	0.447	0.045	0.188	0.419	0.048	0.074
Italy	0.278	0.093	-0.188	0.333	0.096	-0.209
Netherlands	0.517	0.014	0.393	0.509	0.014	0.303
Norway	0.707	0.073	0.948	0.715	0.052	0.802
Sweden	0.533	0.071	0.219	0.495	0.070	0.081
Switzerland	0.611	0.048	0.512	0.582	0.050	0.287
United Kingdom	0.519	0.078	0.394	0.564	0.080	0.204

$\boldsymbol{\mu_i}$ and $\boldsymbol{\sigma_i}$ are the means and standard deviations of the composite indicator $CI(\mathbf{x}_i,\mathbf{w})$ in the 10,000 extractions accordingly. $\mathbf{sm_i}$ is the overall score computed as in eq.8.

$\sigma-\mu$ Pareto-Koopmans family of frontiers consists of five frontiers. For the first scenario, that without a definite ranking of importance for the considered dimensions, the five frontiers are the following: PKF_1 = {Norway, the Netherlands, Austria} , PKF_2 = {Denmark, Switzerland, Germany}, PKF_3 = {Sweden, France}, PFK_4 = {United Kindom}, PKF_5 = {Italy}. In the second scenario, the $\sigma-\mu$ Pareto-Koopmans frontiers remain the same with the exceptions of Switzerland, that was in the second $\sigma-\mu$ Pareto-Koopmans efficiency frontier in the first scenario but descended to the third frontier in the second scenario. Similarly, Sweden, which was in the third frontier in the first scenario has been now descended to the fourth frontier.

Table 3: Measuring $\sigma-\mu$ Pareto-Koopmans efficiency

Unconstrained weights						**Constrained weights**					
	$\sigma-\mu$ Pareto-Koopmans efficiency						$\sigma-\mu$ Pareto-Koopmans efficiency				
	PKF1	**PKF2**	**PKF3**	**PKF4**	**PKF5**		**PKF1**	**PKF2**	**PKF3**	**PKF4**	**PKF5**
Country	δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	δ_{i5}	**Country**	δ_{i1}	δ_{i2}	δ_{i3}	δ_{i4}	δ_{i5}
Austria	0.001	0.032	0.047	0.065	0.193	Austria	0.003	0.037	0.038	0.059	0.143
Denmark	-0.012	0.018	0.095	0.110	0.350	Denmark	-0.009	0.064	0.064	0.083	0.313
France	-0.032	0.000	0.026	0.033	0.048	France	-0.037	0.000	0.002	0.022	0.049
Germany	-0.032	0.002	0.015	0.034	0.169	Germany	-0.037	0.001	0.001	0.022	0.086
Italy	-0.080	-0.048	-0.045	-0.015	0.000	Italy	-0.086	-0.049	-0.048	-0.026	0.000
Netherlands	0.008	0.032	0.050	0.064	0.239	Netherlands	0.002	0.034	0.035	0.056	0.176
Norway	0.078	0.078	0.174	0.188	0.429	Norway	0.068	0.068	0.132	0.151	0.382
Sweden	-0.040	-0.024	0.014	0.014	0.255	Sweden	-0.049	-0.021	-0.020	0.010	0.162
Switzerland	-0.004	0.013	0.078	0.092	0.333	Switzerland	-0.019	0.000	0.028	0.028	0.249
United Kingdom	-0.049	-0.031	-0.008	0.241	0.241	United Kingdom	-0.047	-0.030	-0.019	0.069	0.231

PKF1-5 denote respective $\sigma-\mu$ Pareto-Koopmans frontiers illustrated in Figure 2. $\delta_{\mathbf{ik}}$ shows the (in)efficiency of Country i, with respect to the k^{th} frontier.

In terms of their overall, global efficiencies (sm_i), Norway presents the highest score, while the second highest score is attributed to Denmark in both scenarios. It is worthwhile to observe that Denmark is not in the first $\sigma-\mu$ Pareto-Koopmans efficiency frontier, which, instead, is the case for the Netherlands and Austria. Therefore, we can say that even if Denmark is in a worse Pareto-Koopmans efficiency frontier with

respect to the Netherlands and Austria, overall it compares better relative to the whole set of efficiency frontiers (as shown by the global efficiency scores, sm_i). The reason being can be better explained in the following. Let us compare Austria and Denmark in the unconstrained case. First, it is apparent that none of these countries is dominating the other in both parameters. In particular, Denmark has a greater average score ($\mu_{\text{Denmark}} = 0.628$, $\mu_{\text{Austria}} = 0.471$), while Austria has a lower deviation ($\sigma_{\text{Austria}} = 0.013$, $\sigma_{\text{Denmark}} = 0.064$). Second, by breaking down their global scores ($sm_{\text{Denmark}} = 0.561$, $sm_{\text{Austria}} = 0.338$), it appears that Austria has a greater local score as to the first two frontiers, which is reasonable given that it lies on a higher frontier ($\delta_{\text{Austria1}} = 0.001$, $\delta_{\text{Denmark1}} = -0.012$, $\delta_{\text{Austria2}} = 0.032$, $\delta_{\text{Denmark2}} = 0.018$); still, Denmark is 'catching-up' and, in fact, surpassing Austria by being more efficient with respect to the remaining three frontiers and, in particular, boasting almost twice the Austria's efficiency ($\delta_{\text{Austria3}} = 0.047$, $\delta_{\text{Denmark3}} = 0.095$, $\delta_{\text{Austria4}} = 0.065$, $\delta_{\text{Denmark4}} = 0.11$, $\delta_{\text{Austria5}} = 0.193$, $\delta_{\text{Denmark5}} = 0.35$). Understandably, the same applies also when it comes to the comparison of Denmark and the Netherlands, as well as Switzerland and the Netherlands or Austria. Of course, as proven in proposition 3, the same could not apply to Germany, which, despite the fact that it shares the frontier with Switzerland and Denmark, it is dominated by both Austria and the Netherlands in both parameters. Additionally, let us also observe that in both scenarios Italy is the only country for which the efficiency score, sm_i, is negative. On the other hand, Italy is also the only country in the worst efficiency frontier.

Observe, finally, that the $\sigma - \mu$ efficiency analysis described above can be interpreted as the application of a multiple criteria decision aid method to evaluate the attractiveness of the considered countries. In this perspective this procedure can be seen as a new method in the SMAA family. We call this method $\sigma-\mu-$SMAA (for another method taking into account mean and variance of the evaluations of alternatives, but in another context see Ishizaka and Kunsh (2018))."

5 Case study: World Happiness Index

In this section, we apply $\sigma - \mu$ efficiency analysis to the whole set of data supplied by the 2017 Report of 'World Happiness'. The age-old concept of happiness can be traced back to Aristotle's 'eudaimonia', a word commonly translated as 'welfare' (Shin and Johnson, 1978). Central concept of the Aristotelian ethics, welfare was seen as the ultimate human good (Robinson, 1989), which, more than two millennia after Aristotle's era, appears to be at the centre of academics and policy-makers' discussions. More specifically, world-renowned economists have recently criticized the use of traditional, economic output measures like the GDP as a proxy for welfare (see e.g. Costanza et al., 2009; Stiglitz et al., 2009). In April 2012, an initiative of a group of independent experts -in support of the United Nations' High Level Meeting on happiness and well-being- further paved this way. Through the Sustainable Development Solutions Network of the UN, they published the first 'World Happiness Report' (Helliwell et al., 2012). Since 2012, these reports have gained considerable attention, while, in the authors' words (Helliwell et al., 2017, p.3): "*happiness is now increasingly considered the proper measure of social progress and the goal of public policy*". In fact, on a recent OECD meeting at the ministerial level (OECD, 2016, p.12), the OECD committed to "*redefine the growth narrative to put people's well-being at the center of governments' efforts*".

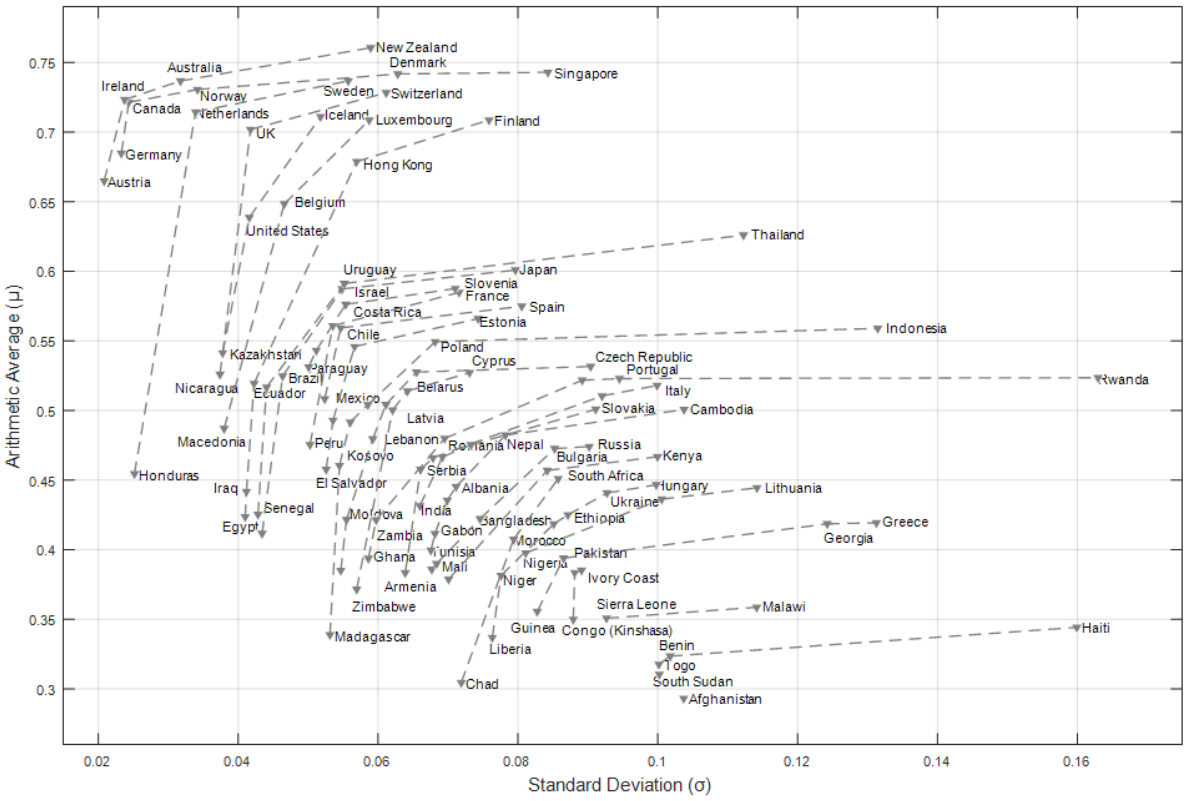
The 'World Happiness' report (WHR) presents and analyses the data of a survey question conducted by the Gallup World Poll. More specifically, 3,000 respondents in each of the -roughly- 150 countries considered, evaluate their lives on a 0-10 scale which is known as 'Cantril Ladder' (see Helliwell et al., 2017, p.123). The authors use a three-year rolling window of the average response in each country (Subjective Well-Being; SWB) to rank them accordingly. For instance, the 2016 ranking is based on the average response of the three-year period 2014-2016. According to the report, 6 key variables (namely *GDP per capita, healthy life expectancy at*

birth, social support, freedom to make life choices, generosity and *perceptions of corruption*) used as proxies for 6 socio-economic aspects respectively, may on average explain 75% of the respondents' subjective evaluations (Pooled OLS regression). Detailed information about the description and sources of the 6 key variables can be found in Helliwell et al. (2017, Technical Box 2, p.17). We applied $\sigma-\mu$ efficiency analysis adopting the same procedure extensively described in the previous section (which considered a sub-sample of 10 European countries) apart from the following step. We use a three-year rolling-window for the six variables, in order to be consistent with the procedure used by the World Happiness Report for the subjective evaluation. This means that the values we consider in each dimension in year 2016 are in fact non-weighted arithmetic averages of the period 2014-2016. We restrict the sample to only these countries that possess data for all 6 dimensions for the 2016 and at least one of the years 2014 and 2015. After this data cleaning procedure we are left with a final sample of 119 countries.

In applying the proposed approach, we find that the family of $\sigma-\mu$ Pareto-Koopmans frontiers consists of 31 frontiers, which are illustrated in Figure 3. We computed the local (δ_{ik}) and global (sm_i) $\sigma-\mu$ Pareto-Koopmans efficiencies for each country. However, due to a large number of countries and frontiers in our sample, we will hereby discuss and report only the efficiency of the top-10 ranked countries of the 2017 'World Happiness' report. The results for the rest of the countries (e.g. local/global efficiencies and rankings) are disclosed in the on-line supplementary appendix (available here: `https://goo.gl/URBRuC`). According to the 2017 report, the countries found in the top ten rankings are the following: *Norway, Denmark, Iceland, Switzerland, Finland, the Netherlands, Canada, New Zealand, Australia and Sweden*, which are ranked in this exact order. In our analysis, these 10 countries are found to be spread in the first seven frontiers, which will therefore be the focus of our analysis for the rest of this section.

Figure 3: Family of $\sigma-\mu$ Pareto-Koopmans frontiers

The 119 countries in our sample are spread over 31 $\sigma-\mu$ Pareto-Koopmans efficiency frontiers (PKF). Further details about the coordinates, efficiency with respect to each PKF, overall $\sigma-\mu$ efficiency and rankings of each country are given in the on-line supplementary appendix.



The countries spread over the first seven frontiers are reported in Table 4, ordered according to their attributed rankings by the WHR (denoted 'WHR rank' respectively). Also reported in the table are the mean score (μ_i) and the standard deviation (σ_i) of the countries' scores in the 10,000 extractions, the $\sigma-\mu$ Pareto-Koopmans local efficiency (δ_{ik}) of each country with respect to the efficient frontiers $PKF_k, k = 1,\ldots,7$, and the global efficiency score (sm_i) with its corresponding ranking (denoted '$\sigma-\mu$ rank').

Table 4: Case study results for the first seven frontiers

Country	WHR rank	μ_i	σ_i	$\mathbf{sm_i}$	$\sigma-\mu$ rank	$\sigma-\mu$ Pareto-Koopmans efficiency						
						PKF1 δ_{i1}	**PKF2** δ_{i2}	**PKF3** δ_{i3}	**PKF4** δ_{i4}	**PKF5** δ_{i5}	**PKF6** δ_{i6}	**PKF7** δ_{i7}
Norway	1	0.731	0.034	6.040	6	-0.004	0.003	0.008	0.017	0.020	0.024	0.034
Denmark	2	0.742	0.063	6.312	3	-0.012	0.003	0.005	0.013	0.031	0.033	0.033
Iceland	3	0.711	0.052	5.445	11	-0.022	-0.019	-0.011	-0.002	0.006	0.006	0.016
Switzerland	4	0.728	0.061	5.922	7	-0.018	-0.009	-0.007	0.017	0.017	0.020	0.020
Finland	5	0.709	0.076	5.335	13	-0.036	-0.027	-0.024	-0.017	-0.002	-0.000	0.030
Netherlands	6	0.714	0.034	5.619	10	-0.010	-0.008	0.009	0.010	0.016	0.022	0.028
Canada	7	0.721	0.024	5.843	9	-0.001	0.006	0.009	0.018	0.025	0.031	0.036
New Zealand	8	0.761	0.059	6.904	1	0.018	0.018	0.024	0.032	0.050	0.052	0.052
Australia	9	0.737	0.032	6.218	4	0.002	0.005	0.012	0.021	0.026	0.028	0.038
Sweden	10	0.737	0.056	6.173	5	-0.011	-0.002	0.009	0.009	0.026	0.028	0.028
Austria	13	0.665	0.021	4.496	17	0.002	0.002	0.011	0.020	0.021	0.025	0.032
United States	14	0.639	0.042	3.726	19	-0.021	-0.018	-0.010	-0.001	0.004	0.004	0.011
Ireland	15	0.723	0.024	5.891	8	0.001	0.001	0.010	0.019	0.026	0.032	0.038
Germany	16	0.685	0.023	4.955	15	-0.001	0.001	0.009	0.018	0.022	0.025	0.031
Belgium	17	0.648	0.047	3.925	18	-0.025	-0.023	-0.014	-0.006	-0.003	0.006	0.007
Luxembourg	18	0.709	0.059	5.358	12	-0.027	-0.023	-0.015	-0.007	-0.002	0.010	0.010
United Kingdom	19	0.702	0.042	5.252	14	-0.018	-0.017	-0.008	0.008	0.008	0.013	0.018
Singapore	26	0.743	0.084	6.341	2	-0.018	0.001	0.006	0.015	0.032	0.034	0.034
Nicaragua	41	0.526	0.037	1.668	33	-0.017	-0.014	-0.010	0.000	0.000	0.003	0.005
Ecuador	44	0.519	0.042	1.496	38	-0.021	-0.019	-0.014	-0.005	-0.004	-0.002	0.002
Kazakhstan	60	0.541	0.038	1.871	30	-0.017	-0.014	-0.009	0.000	0.001	0.003	0.006
Hong Kong	71	0.679	0.057	4.592	16	-0.034	-0.033	-0.023	-0.015	-0.008	-0.004	0.012
Honduras	91	0.455	0.025	1.359	40	-0.004	-0.002	0.009	0.012	0.013	0.013	0.016
Macedonia (F.Y.R.)	92	0.487	0.038	1.272	41	-0.017	-0.015	-0.011	-0.001	0.000	0.004	0.004
Egypt	111	0.424	0.041	0.786	55	-0.020	-0.018	-0.016	-0.004	-0.003	-0.003	0.000
Iraq	117	0.442	0.041	0.876	54	-0.020	-0.018	-0.016	-0.004	-0.003	-0.003	0.000

WHR is the rank attributed to Country i by the 'World Happiness' report using the Gallup World Poll surveys (i.e. 'Cantril Ladder'). $\boldsymbol{\mu_i}$ and $\boldsymbol{\sigma_i}$ are the means and standard deviations of the composite indicator $CI(\mathbf{x}_i, \mathbf{w})$ in the 10,000 extractions accordingly. $\mathbf{sm_i}$ is the overall score computed as in eq.8. $\boldsymbol{\sigma}$**-**$\boldsymbol{\mu}$ **rank** is the rank obtained based on the overall score **sm**. **PKF1-7** denote respective frontiers and $\delta_{\mathbf{ik}}$ exhibits the (in)efficiency of Country i, with respect to the k^{th} $\sigma-\mu$ Pareto-Koopmans frontier.

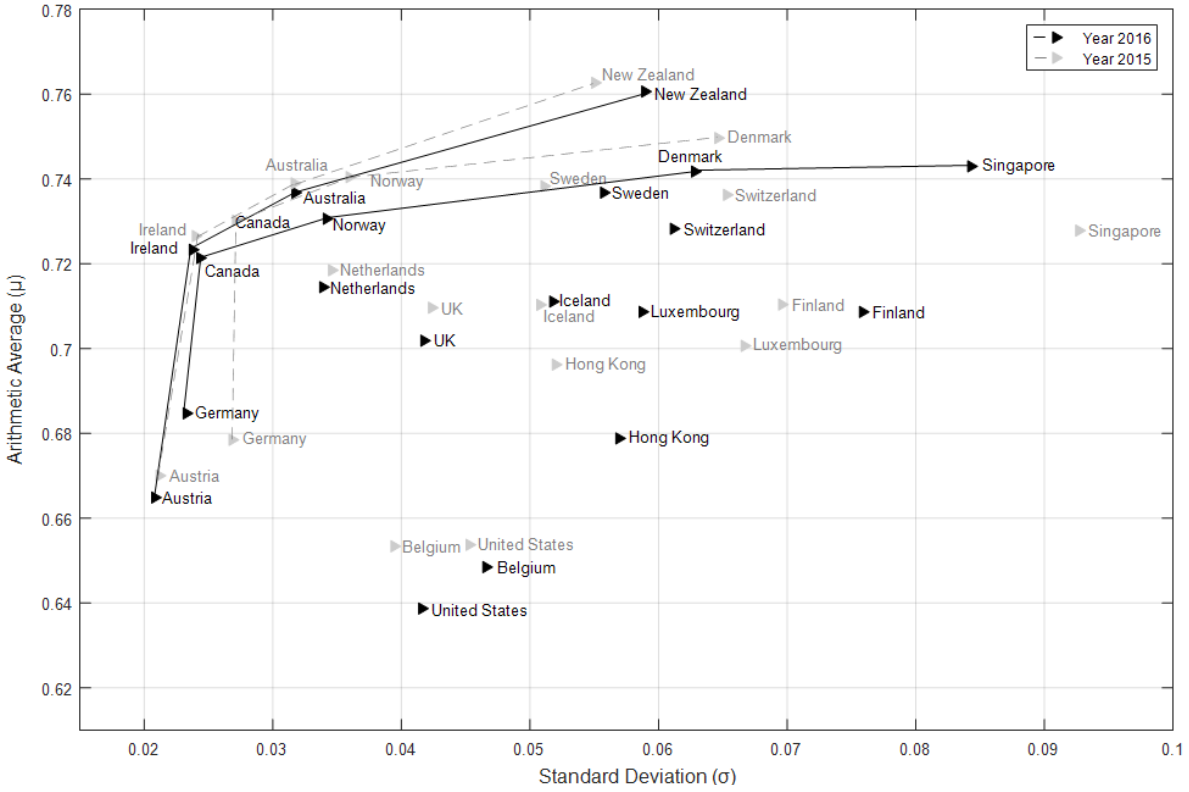
First of all, we should note that it is by definition reasonable to observe a shuffle, or even entirely different patterns between the SWB ('WHR rank') and the $\sigma-\mu$ efficiency rankings ('$\sigma-\mu$ rank'). The first expresses peoples' own subjective beliefs, while the latter refers to the aggregation of 6 variables that are considered key determinans of the average SWB. Moreover, there is a whole ongoing discussion between the difference of SWB and objective conditions attributed to psychological reasons and cultural differences (see Kroll and Delhey, 2013). In other words, the two rankings are not directly comparable, nor should they necessarily be; though one could make a few interesting inferences. To start with, it is notable, that the countries which are self-claimed to be ranked in the top-10 positions (i.e. having the top-10 highest subjective evaluation) are positioned in our top-10 list as well, with the exception of Iceland and Finland, which we position in the 11^{th} and 13^{th} places accordingly.

A second interesting point relates to the measurement of efficiency with respect to the frontiers, and how the dynamics of these might change under some circumstances. Consider for instance Finland, a country that is ranked 13^{th} according to our overall $\sigma-\mu$ Pareto-Koopmans efficiency, and which participates in the $\sigma-\mu$ Pareto-Koopmans family by lying on the 7^{th} frontier. The reason Finland is not participating in the previous frontier (i.e. PKF6) can be better clarified when it is compared to Luxembourg. The latter clearly dominates the former in terms of standard deviation ($\sigma_{Luxembourg}=0.059$ versus $\sigma_{Finland}=0.076$), but only marginally dominates in terms of average performance ($\mu_{Luxembourg}=0.70865$ versus $\mu_{Finland}=0.70864$ - in Table 4 both are rounded to three decimals). Therefore, if Finland slightly increases its average performance to surpass that of Luxembourg, it will then, *ceteris paribus*, move to frontier 6. This is also clear by looking at the efficiency of Finland with respect to the 6^{th} frontier (Table 4: $\delta_{Finland,6}=-0.00001$), which is almost zero. Following this line of reasoning, one could be interested to compare Finland with Iceland ($\mu_{Finland}=0.70864$ versus $\mu_{Iceland}=0.7111$), e.g. by looking at the inefficiency of the former with respect to the frontier that the latter is lying on (Table 4: $\delta_{Finland,5}=-0.002$).

Another interesting point arises from tracking the frontiers' formation from a dynamic viewpoint. More specifically, one could be interested in tracing changes in the performance of units in the $\sigma-\mu$ plane within a time period and thus, how were the frontiers re-structured accordingly. This could be accomplished in several ways. For instance, one could trace all, or a subset of the $\sigma-\mu$ PKF, or even trace the frontiers and performance of only certain countries. An example is given in Figure 4, which illustrates how the first two frontiers were changed from 2015 (illustrated in gray) to the following year (illustrated in black). To some extent, this augments the analysis of Färe et al. (1994, see Fig.3, p.77) by visualizing the dynamic formations of all subsequent frontiers. It quickly becomes obvious that Singapore did not participate in the first two frontiers in 2015, but it joined the second one in 2016. Moreover, one can distinguish how the performance of the countries lying in the first two $\sigma-\mu$ PKF changed during this time period. For instance, as it is apparent in Fig.4, almost all countries exhibit a drop as to their mean values in 2016. This is less noticeable in some countries and more apparent in others. Exception to this rule are Germany, Luxembourg and Singapore, with the latter meeting with such an improvement that positioned the country in the second frontier. Of course this can be attributed to both a remarkable improvement in the elementary indicators, and the fact that the performance of the surrounding countries was deteriorated (e.g. see Denmark in Fig.4). This highlights the fact that even if a unit's performance remains steady through a time period examined, the distance with respect to other frontiers might alter either due to an improvement, or a downturn of the surrounding units. Understandably, this reminds of the decomposition of total factor productivity (see Färe et al., 1997). In this sense, it is possible to directly measure the change in the overall relative efficiency (EC) by considering a ratio in the spirit of the efficiency change component of Malmquist Productivity Index (see Färe et al., 1994, p.71). Although it extends beyond the scope of this study, it is worth noting that such an analysis from a dynamic viewpoint could greatly benefit the explanation of results, by decomposing the total productivity into relative efficiency and technical change. In fact, an interesting study in the domain of composite indicators is presented by Kortelainen (2008), constructing an Environmental Performance Index in which they exhibit how changes in the environmental performance of 20 EU member states over the period 1990-2003 may be decomposed into shifts in relative efficiency and environmental technology respectively. Additionally, in this particular example we have used two consecutive years, which, from a policy-maker's perspective might not be enough; thus, the time period examined in the plane could be re-considered to that of specific 'goalposts' (i.e. the start and end dates of a scheduled policy period, see Mazziotta and Pareto, 2016, p.989).

Figure 4: Dynamic illustration of the frontiers

An interesting feature of $\sigma-\mu$ analysis is the comparison of units or frontiers from a dynamic viewpoint. A developer might be keen on tracking the formation of a frontier of interest, or the performance of a unit through time (e.g. either consecutive years, or a policy period of interest). This figure delineates the formation of the first two $\sigma-\mu$ Pareto-Koopmans efficiency frontiers (PKF) in two consecutive years. Black colour represents the year 2016 while grey colour represents the year 2015.



Consequently, there are several points that could be noted from the outputs of our proposed approach. From an overall score/ranking that takes into account all potential viewpoints (i.e. space of weight vectors) and all potential benchmarks (as denoted by the family of $\sigma-\mu$ Pareto-Koopmans frontiers), to the analysis of the dynamic performance of a unit. These could be all advantageous to both the developer of an indicator and the individuals interested in it. We should hereby note again that subjective evaluations (i.e. those of the WHR in this case) and our own output (i.e. sm_i global efficiency scores and $\sigma-\mu$ rankings accordingly) cannot be directly compared due to the intrinsic differences in their representation.

5.1 Robustness of results to outliers

A crucial question at this point relates to the sensitivity of the obtained results from the preceded $\sigma-\mu$ analysis. Such question is mainly driven from the fact that extreme points (outliers) could be distorting the results. The problem of outliers is one of the oldest in Statistics that is constantly reemerging (Hawkins, 1980). The intention to explore the mechanisms driving the outliers extends beyond the scope of this paper (for a comprehensive analysis, we refer the reader to the book of Hawkins, 1980), though it is of interest to explore the steps in which outliers could distort our analysis, along with ways to make our inferences more robust to them. In particular, we can identify two stages in which outliers could pose a threat, and which we forthwith explain in more detail along with ways to mitigate their impact.

The first stage is in the process of normalizing the sub-indicators, the chosen method of which could distort the transformed indicators in the presence of 'extreme' units (see some common normalization tech-

niques and their drawbacks in OECD, 2008, sect. 1.5). Distorted transformed indicators could in turn affect the computed composite indicators' values, on which our two measures of interest (σ and μ) rely upon for the subsequent part of the analysis. We believe that, up to some extent, the normalization procedure that we follow (Greco et al., 2018a) takes this issue into account by replacing the values of extreme units (see Section 4 for a detailed description of the procedure). Moreover, the fact that in our method, a variety of weight vectors are involved (hereby, 10,000) -contrary to the classic scheme involving a unique weight vector- means that it could alleviate this issue even more. The reason is that, in the case of a single weight vector, it could happen that this particular vector favors the dimension(s) which are affected the most from the existence of an 'extreme' unit in the set of DMUs. On the contrary, 10,000 weight vectors could even out this issue, of course, always up to some extent.

The second stage in which outliers could pose a threat comes after the computation of the parameters of interest (σ_i and μ_i) has taken place. Outliers in this stage could affect the local efficiency scores (δ_{ik}), which in turn would distort the global efficiencies (sm_i). The reason is that in DEA the addition or removal of efficient DMUs would alter the efficiencies of the remaining DMUs (Seiford and Zhu, 2003). This means that, if an extreme unit exists in the $\sigma-\mu$ plane, it could compromise the results up to some extent, as the overall (global) efficiency scores do not solely rely on the first Pareto-Koopmans efficient frontier, but also on all the remaining frontiers in the sequence. In such a case, our analysis could benefit from well-established approaches in the literature of 'robust' (or 'partial') frontiers, such as the order-m (Cazals et al., 2002; Daraio and Simar, 2005, 2007b) or order-α (Aragon et al., 2003; Daouia and Simar, 2007) frontiers that we explore in this section. In brief[2], although slightly different in their principles, the advantage of both above-mentioned techniques is that they are more robust to outliers than the classic efficient estimators, as they do not simultaneously envelop all the data points but rather a sub-sample of them (the choice of which consists the fundamental difference among the two approaches). In this paper we consider the order-m robust frontiers, originally introduced by Cazals et al. (2002) and later generalized and extended by Daraio and Simar (2005, 2007b), although the intuition could be similar in applying the order-a robust frontiers (Aragon et al., 2003; Daouia and Simar, 2007).

The procedure to obtain robust DEA estimators of order-m -which, we hereby use to obtain robust local and global $\sigma-\mu$ Pareto-Koopmans efficiency scores- is extensively covered in the study of Daraio and Simar (2007b, pp.18-19). The authors provide a simple Monte-Carlo simulation implemented in four steps, which we adopt to be fitted to our proposed approach. We implement it in two ways, described in the following. First, if one is solely interested in taking into account a single frontier, we adopt it without any modification. That is, for each unit $i \in I$, we randomly draw a sample of size m (in this case we choose a 'strict' value of $m = 10$) with replacement so that it satisfies the following conditions: $\mu_l \geq \mu_i$ and $\sigma_l \leq \sigma_i$; $l = 1,\ldots,m$. We then proceed by solving the LP formulation given in equation 5 to obtain the efficiency score, δ_{i1} for the evaluated unit i with respect to $PKF1$. We repeat this procedure B times for every unit $i \in I$, with B being a relatively large number, averaging the results afterwards. Following the suggestions of Daraio and Simar (2007a, p.72), we use a value of $B = 200$. Understandably, this analysis could be extended to include the case where additional information is provided by other variables $Z \in \mathbb{R}^r$ that are exogenous to the process but could explain part of it. In such a case, the conditional order-m efficient estimators could be used (see e.g. Daraio and Simar, 2005, 2007b).

We avoid using the exact procedure of robust m-order frontiers for the case of multiple-frontier evaluation, as this is only 'forward-looking' for competitors in the sense of trying to find competitors from only the dominating choices (i.e. $\mu_l \geq \mu_i$, and $\sigma_l \leq \sigma_i$, $l = 1,\ldots,m$). We believe that the concept of global scores (sm_i)

[2]For a comprehensive review of the intuition behind the robust frontier techniques and a set of empirical applications, we refer the reader to the book of Daraio and Simar (2007a).

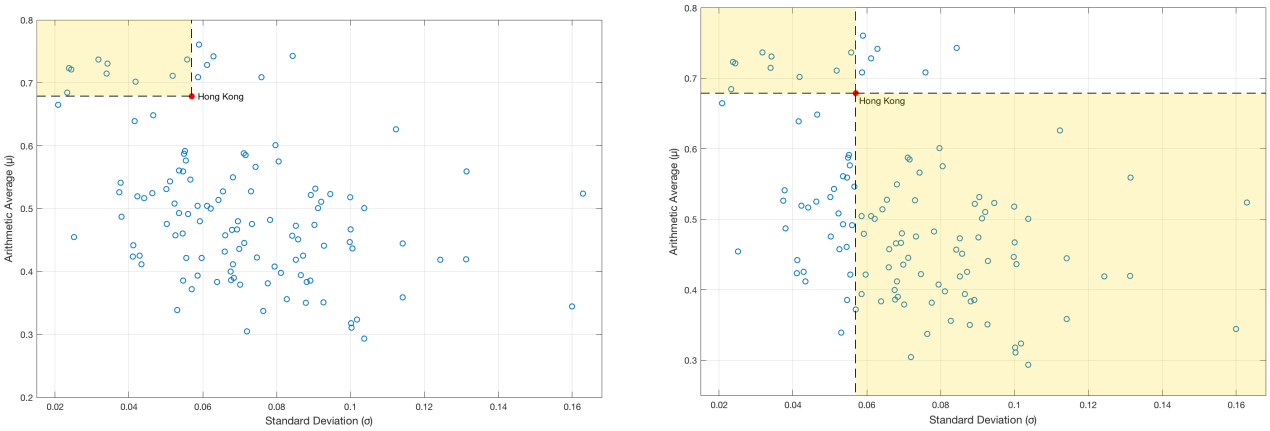
should not only take into account the frontiers that lie ahead of a unit, but also to be 'backward-looking', giving a sort of 'net position' evaluation for a unit with respect to the 'competitors' in front and back of that unit in the plane. Thus, a second way in which we apply the m-order robust frontiers is by modifying this procedure to equally look for the exact opposite scenario; that is, for each unit $i \in I$, we randomly draw (with replacement) m units exactly as before (i.e. $\mu_l \geq \mu_i$, and $\sigma_l \leq \sigma_i$ $l = 1,\ldots,m$), but also m units dominated by the evaluated unit (i.e. $\mu_l \leq \mu_i$, and $\sigma_l \geq \sigma_i$ $l = 1,\ldots,m$). Then we solve the LP formulation as given in eq.7 and compute the global scores (sm_i) as in eq.8. A visual interpretation of the two above-mentioned procedures is given in Fig.5 for the case that we evaluate a random unit of interest (e.g. Hong-Kong).

Figure 5: Didactic illustration of computing the m-order efficiency estimators in the proposed method.

This figure illustrates the computation of m-order robust efficiency estimators (Cazals et al., 2002; Daraio and Simar, 2005, 2007b) for a randomly chosen country (hereby, Hong-Kong).

The un-adjusted case is presented in the left sub-plot, where in evaluating Hong-Kong, a randomly sampled set of countries of order m (hereby $m = 10$) is used from the highlighted area to find the efficiency with respect to the single frontier (or δ_{i1}), solving the LP formulation presented in eq.5. This procedure is repeated B times (hereby $B = 200$), and the expected estimator is used as an m-order robust estimator for this country, taking into account only the first Pareto-Koopmans frontier.

The adjusted case (right sub-plot) involves the same procedure, sampling this time a set of order m from the highlighted area above the evaluated country (dominating solutions) and a set of units of order m from the highlighted area beneath it (dominated solutions), solving the LP formulation presented in eq.7 and computing the global scores (sm_i) as in eq.8. This procedure is repeated B times and the expected estimator is used as an m-order robust global estimator for this country, taking into account all potential Pareto-Koopmans frontiers in the sampling space.



To compare our results with both above-mentioned applications of the m-order partial frontiers, we normalize the original and robust scores to the $[0,1]$ space. The diagonal in Fig.6 shows perfect equality, while deviations from it show under or over-evaluation of units with respect to each set of estimators (robust or non-robust to outliers). Understandably, in the case of a single frontier (Fig.6, left sub-plot) the deviations are very small and negligible. Taking into account the multiple-frontier case (Fig.6, right sub-plot) though, we can clearly see the existence of three outliers (Thailand, Indonesia and Rwanda) that were affecting the original set of estimators. With respect to the 'scoreboards' of the evaluated countries, 17 of them (approx. 14% of the sample) do not present any change whatsoever, while another 17 of them only change by a single ranking. 32 countries (approx. 27%) present a change of between 2 and 3 rankings (median change is 3), while another 27 (approx. 22.7%) change between 4 and 7 rankings (that completes the 3rd quartile). The fourth quartile contains changes between 8 and 19 rankings with only the outliers exceeding this range, changing 35 rankings. As it is also visually apparent from Fig.7, the biggest changes are presented at and around the frontiers in which the outliers are participating. In this respect, the robust m-order frontiers aid significantly in adjust-

ing the estimators to account for these outliers, and given these noticeable differences (especially around the frontiers containing the outliers), we strongly encourage their use alongside our proposed approach.

Figure 6: Robustness checks.

This figure delineates the robustness of the obtained results using the unconditional m-order robust estimators Daraio and Simar (2005, 2007b) to the single frontier case (δ_{i1}) [left], or adjusted to the multiple-frontier case [right]. In both figures, vertical axis represents non-robust measures of efficiency (δ_{i1} left and sm_i right) and the horizontal axis represents the robust m-order ($m = 10$) efficiency estimators. To render them completely comparable (adjusting their scales), we normalize them (using the 'min-max' method). The diagonal thus represents perfect equality among the two, with units lying above (below) the diagonal being favored more (less) in the case of non-robust estimators.

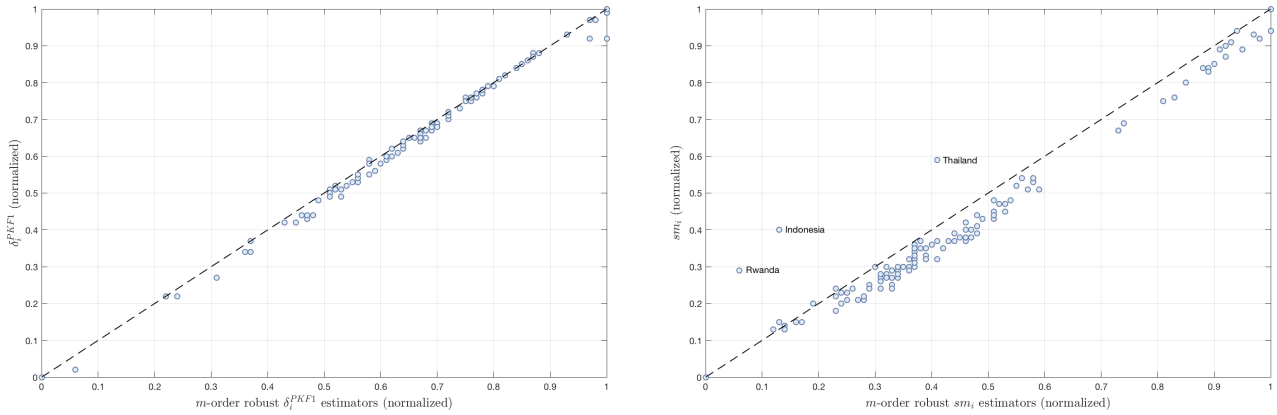
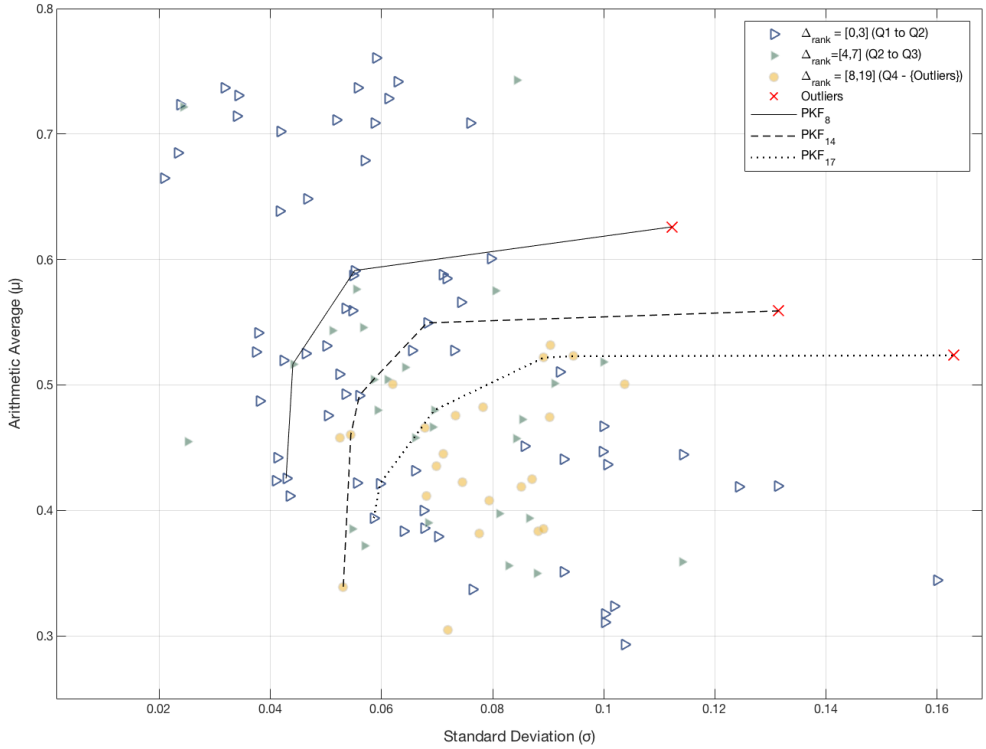


Figure 7: Outliers & rank reversals.

This figure delineates the absolute changes in the rankings of the evaluated countries with respect to the robust and non-robust global estimators produced up to this point. Δ_{rank} denotes absolute change of a country's ranking with respect to the two compared set of estimators, Q denotes quartile with respect to the whole range of rank reversals. The PKFs of the outliers are plotted to delineate how the units at and around these frontiers in which outliers participate can distort the global efficient (non-robust) estimators.



6 Further considerations and generalizations

A basic and natural question arising from our approach is the following: What is the trade-off between μ and σ? To answer this, let us first note the following main general interpretations of a 'trade-off':

- Trade-off as *rate of technical substitution*; that is, taking into account the "production frontier", how much can we increase μ and decrease σ to remain in the same "isoquant"?
- Trade-off as *rate of substitution*; that is, taking into account the "preferences" of the stakeholder, the policy-maker or the 'expert' considered in the composite indicator, how much can we increase μ and decrease σ to maintain the same level of "utility"?

Our approach permits to take into consideration both interpretations of a trade-off. In fact, on the one hand, the Pareto-Koopmans frontier can be interpreted as the isoquant between σ and μ, so that, in this perspective, the weights α and β attached to μ and σ respectively in the solution of eq.5 can be interpreted as the rate of technical substitution between them. On the other hand, our approach based on the Pareto-Koopmans frontier in the $\sigma-\mu$ space can be considered as a specific application of the Benefit of Doubt (BoD) method (Cherchye et al., 2007) in that space. BoD is a well-known methodology in the domain of composite indicators assigning to each unit the most favorable set of weights that maximize its performance. Therefore, 'weights' α and β obtained from the solution of eq.5 can be interpreted analogously to the weights of BoD. That is, they define a rate of substitution in the case that the most awarding evaluation is adopted for the considered unit.

Another interpretation of the trade-off between μ and σ in terms of a rate of substitution relates to their use in evaluating units to give an approximate value to the p-th percentile of the distribution of values assumed by the composite indicator $CI(\mathbf{x}_i,\mathbf{w})$ in the space of weight vectors $\mathbf{w}\in W$. Indeed, one can assume that this distribution is approximately normal and therefore we can compute the p-th percentile as $\mu-\phi^{-1}(p)\sigma$ where $\phi^{-1}(p)$ is the percentile of the standard normal distribution, so that, for example, $\phi^{-1}(0.1)=1.645, \phi^{-1}(0.05)=1.960$ and $\phi^{-1}(0.01)=2.576$. Suppose now that a stakeholder is interested in evaluating units on the basis of a specific percentile, e.g. 0.05. Since each unit $i\in I$ will be attached a value $\mu_i-1.960\sigma_i$, implicitly weights α and β such that $\frac{\beta}{\alpha}=1.960$ are adopted and, consequently, a trade-off in terms of substitution rate such that each decrease of an amount, say Δ, in terms of μ has to be compensated by a decrease of 1.960Δ in terms of σ is adopted.

In this study we have considered the development of a composite indicator in terms of a weighted sum that, in fact, is a weighted arithmetic mean of the underlying sub-indicators. Nonetheless, one may easily generalize the weighted sum by considering the weighted quasi-arithmetic mean that is

$$CI(\mathbf{x}_i,\mathbf{w})=f^{-1}\left(\sum_{j=1}^{n}w_j f(x_{ij})\right),$$

with $f:[0,1]\rightarrow[0,1]$ being a strictly increasing function. A typical example of the weighted quasi arithmetic mean is the weighted geometric mean that is obtained as: $f(x)=log\,x$. Notice that, our current proposal formulating a composite indicator of the form $CI(\mathbf{x}_i,\mathbf{w})$ can thus be straightforwardly extended to the general formulation in terms of weighted quasi-arithmetic mean. It is also worth noting that, independently of the formulation of $CI(\mathbf{x}_i,\mathbf{w})$, also the utility function

$$U(\sigma,\mu)=\alpha\mu-\beta\sigma$$

that we considered to define our $\sigma-\mu$ efficiency, can be written as a weighted quasi arithmetic mean, that is

$$U(\sigma,\mu)=f^{-1}\left(\alpha f(\mu)-\beta f(\sigma)\right).$$

In case of $f(x)=log\,x$, we get

$$U(\sigma,\mu)=\mu^{\alpha}\cdot\sigma^{-\beta}.$$

In any case, whatever the function f is, the whole procedure we proposed to define the $\sigma-\mu$ efficiency can be easily extended accordingly, substituting μ and σ with $f(\mu)$ and $f(\sigma)$.

7 Conclusion

There is a long discussion in the literature of composite indicators regarding the issue of weighting in their construction. Years of disputes and past solutions revolve around the use of a weight vector that allegedly perfectly represents a specific unit or all evaluated units overall. Still, quite different results can be obtained even by slightly changing this vector, the choice of which resembles a quest for the "holy grail". Extending this argument from a conceptual point of view, this set of weights (commonly univocal) could be never representative for the population interested in this synthetic measure. Therefore, it seems reasonable to take into account for each unit the distribution of values assumed by the composite indicator on the whole set of feasible weight vectors. Our proposed methodology called '$\sigma-\mu$ efficiency analysis' synthesizes such distributions for each unit with its mean value, μ, intended to be maximized, and its standard deviation, σ, intended to be minimized, as it denotes instability in the evaluations with respect to the variability of weights. We further defined the concepts of $\sigma-\mu$ Pareto-Koopmans dominance and efficiency, which permitted us to define for each unit under analysis, several types of meaningful efficiency measures. This way we outlined the $\sigma-\mu$ efficiency analysis, which finds its basis in some well-known Operational Research methodologies listed below:

- Stochastic Multiattribute Acceptability Analysis (SMAA), for the idea of considering the whole set of feasible weight vectors. With respect to this point, let us remark that our proposed approach can be seen as another method in the SMAA family: the $\sigma-\mu-$SMAA;
- Data Envelopment Analysis (DEA), for the idea of measuring efficiency;
- Markowitz modern portfolio theory, for the idea of representing distributions in terms of mean and standard deviation.
- NSGA-II, for the idea of a sequence of Pareto frontiers.
- Context-dependent DEA, for the idea of a sequence of Pareto-Koopmans frontiers.

Additionally, the $\sigma-\mu$ analysis can be seen as being at the crossroads of the following three prominent research domains in economics:

- Well-being economics in a neo-Benthamite perspective, because consideration of the whole set of feasible weight vectors can be seen as a means of taking into account the utility of all individuals in the population.
- Research on inequality in economics, because in a "post-GDP" perspective, the standard deviation of the distribution of composite indicators values in the space of weight vectors can be seen as the counterpart of an income inequality measure in a standard, "GDP economics" perspective.

- Efficiency analysis taking into account, among others, the contributions of Koopmans, Debreu and Farrell, because it permits fruitful investigation and scrutiny of mean and standard deviation of the composite indicator values' distribution.

With respect to its merits, the proposed methodology permits the inclusion of all potential viewpoints in the construction of a composite indicator, while it takes into account the distances of units from all the $\sigma-\mu$ Pareto-Koopmans frontiers lying on the plane, collapsed into a global efficiency score. In addition, the use of robust order-m or order-α efficient frontiers could greatly benefit the proposed approach by providing more accurate estimators that are robust to outliers. While there is no particular scope in this study to treat compensatory issues in the construction of a composite indicator; we should note that our methodology permits the use of non-compensatory aggregation techniques such as PROMETHEE methods (see Brans et al., 1986) or ELECTREE methods (for a survey see Figueira et al., 2016 and for a review of recent developments see Figueira et al., 2013) to be applied instead of the additive utility model illustrated in the paper. In this case, to apply the SMAA to PROMETHEE and ELECTRE methods, see the approaches proposed in Corrente et al. (2014) and Corrente et al. (2016a) respectively. Moreover, interaction and hierarchy of dimensions can be considered through the use of Choquet integral and Multiple Criteria Hierarchy Process (see e.g. Angilella et al., 2018).

We attempted to show the potential of $\sigma-\mu$ efficiency analysis by applying it to the data supplied by the 'World Happiness' report, obtaining some interesting results and insights. Of course, our methodology cannot be considered a 'panacea' for the many problems affecting the adoption of composite indicators, in general, and the 'World Happiness' in particular (see e.g. the critical discussion on composite indicators applied to wellbeing measures in Kroll and Delhey, 2013). However, we hope that this case study can convince on the many interesting insights that $\sigma-\mu$ efficiency analysis permits in this domain. Finally, as far as its future direction of research is concerned, we believe that our methodology can be fruitfully applied to all the domains in which composite indicators are considered, ranging from the ranking of universities to the measurement of competitiveness of geographical regions and beyond.

Acknowledgments

We are grateful to three anonymous referees for constructive comments and suggestions that greatly helped in improving this paper. We thank Salvatore Corrente for a rich discussion and comments on an earlier version of the paper, as well as Michalis Doumpos for remarks particular to our approach. All errors remain our own. Salvatore Greco would like to acknowledge the funds "*Data analytics for entrepreneurial ecosystems, sustainable development and wellbeing indicators*" (PTR-DEI 2016/2018) and "Chance" (DR 28/04/2017 - Rep. n. 1393) of the University of Catania.

References

Abberger, K., Graff, M., Siliverstovs, B., and Sturm, J.-E. (2017). Using rule-based updating procedures to improve the performance of composite indicators. *Economic Modelling*, 68:127–144.

Ainslie, G. (2001). *Breakdown of will*. Cambridge, United Kingdom: Cambridge University Press.

Andersen, P. and Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management science*, 39(10):1261–1264.

Angilella, S., Catalfo, P., Corrente, S., Giarlotta, A., Greco, S., and Rizzo, M. (2018). Robust sustainable development assessment with composite indices aggregating interacting dimensions: The hierarchical-SMAA-Choquet integral approach. *Knowledge-Based Systems*, 158:136 – 153.

Aragon, Y., Daouia, A., and Thomas-Agnan, C. (2003). *A Conditional Quantilebased Efficiency Measure*. Technical report, Discussion paper, GREMAQ et LSP, Université de Toulouse.

Atkinson, A. (2015). *Inequality*. Harvard University Press.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of economic theory*, 2(3):244–263.

Aumann, R. J. (1962). Utility theory without the completeness axiom. *Econometrica: Journal of the Econometric Society*, pages 445–462.

Babbie, E. R. (1995). *The practice of social research*. Wadsworth Publishing Company.

Bandura, R. (2011). *Composite Indicators and Rankings: Inventory 2011*. Technical report, New York: Office of Development Studies, United Nations Development Programme (UNDP).

Barro, R. J. and Sala-i Martin, X. (1992). Convergence. *Journal of political Economy*, 100(2):223–251.

Becker, W., Saisana, M., Paruolo, P., and Vandecasteele, I. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological Indicators*, 80:12–22.

Bewley, T. F. (2002). Knightian decision theory. Part I. *Decisions in Economics and Finance*, 25(2):79–110.

Blackburn, D. W. and Ukhov, A. D. (2013). Individual vs. aggregate preferences: The case of a small fish in a big pond. *Management Science*, 59(2):470–484.

Booysen, F. (2002). An overview and evaluation of composite indices of development. *Social Indicators Research*, 59(2):115–151.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge, United Kingdom: Cambridge university press.

Brans, J.-P., Vincke, P., and Mareschal, B. (1986). How to select and how to rank projects: The PROMETHEE method. *European Journal of Operational Research*, 24(2):228–238.

Cazals, C., Florens, J.-P., and Simar, L. (2002). Nonparametric frontier estimation: a robust approach. *Journal of econometrics*, 106(1):1–25.

Cerreia-Vioglio, S., Giarlotta, A., Greco, S., Maccheroni, F., and Marinacci, M. (2018). Rational preference and rationalizable choice. *Economic Theory*. In print.

Charnes, A. and Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics (NRL)*, 9(3-4):181–186.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Cherchye, L., Moesen, W., Rogge, N., and Puyenbroeck, V. T. (2007). An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research*, 82(1):111–145.

Chowdhury, S. and Squire, L. (2006). Setting weights for aggregate indices: An application to the commitment to development index and human development index. *The Journal of Development Studies*, 42(5):761–771.

Collard, D. (2006). Research on well-being: Some advice from Jeremy Bentham. *Philosophy of the Social Sciences*, 36(3):330–354.

Collins, A., Leonard, A., Cox, A., Greco, S., and Torrisi, G. (2017). PERCEIVE Deliverable 4.3 Report on Smart Cities and Resilience, available: http://amsacta.unibo.it/5721/. Technical report.

Cooper, W. W., Seiford, L. M., and Zhu, J. (2011). *Handbook on data envelopment analysis*. International Series in Operations Research & Management Science. 2nd edition, Springer US.

Corrente, S., Figueira, J. R., and Greco, S. (2014). The SMAA-PROMETHEE method. *European Journal of Operational Research*, 239(2):514–522.

Corrente, S., Figueira, J. R., Greco, S., and Słowiński, R. (2016a). A robust ranking method extending ELECTRE III to hierarchy of interacting criteria, imprecise weights and stochastic analysis. *Omega*, 73:1–17.

Corrente, S., Greco, S., Kadziński, M., and Słowiński, R. (2013). Robust ordinal regression in preference learning and ranking. *Machine Learning*, 93(2-3):381–422.

Corrente, S., Greco, S., Kadziński, M., and Słowiński, R. (2016b). Inducing probability distributions on the set of value functions by subjective stochastic ordinal regression. *Knowledge-Based Systems*, 112:26–36.

Costanza, R., Hart, M., Posner, S., and Talberth, J. (2009). *Beyond GDP: The Need for New Measures of Progress*. Pardee Center for the Study of the Longer-Range Future, Boston.

Daouia, A. and Simar, L. (2007). Nonparametric efficiency analysis: a multivariate conditional quantile approach. *Journal of Econometrics*, 140(2):375–400.

Daraio, C. and Simar, L. (2005). Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of productivity analysis*, 24(1):93–121.

Daraio, C. and Simar, L. (2007a). *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications*. Springer Science & Business Media.

Daraio, C. and Simar, L. (2007b). Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach. *Journal of Productivity Analysis*, 28(1-2):13–32.

De Muro, P., Mazziotta, M., and Pareto, A. (2011). Composite Indices of Development and Poverty: An Application to MDGs. *Social Indicators Research*, 104(1):1–18.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2):182–197.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, 19:273–292.

Decancq, K. and Lugo, M. A. (2013). Weights in Multidimensional Indices of Wellbeing: An Overview. *Econometric Reviews*, 32(1):7–34.

Decancq, K., Van Ootegem, L., and Verhofstadt, E. (2013). What if we voted on the weights of a multidimensional well-being index? An illustration with Flemish data. *Fiscal Studies*, 34(3):315–332.

Doumpos, M., Gaganis, C., and Pasiouras, F. (2016). Bank Diversification and Overall Financial Strength: International Evidence. *Financial Markets, Institutions & Instruments*, 25(3):169–213.

Doumpos, M., Hasan, I., and Pasiouras, F. (2017). Bank overall financial strength: Islamic versus conventional banks. *Economic Modelling*, 64:513–523.

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of science*, 1(2):163–169.

Elster, J. (1987). *The Multiple self*. Cambridge, United Kingdom: Cambridge University Press.

Elton, E. J., Gruber, M. J., Brown, S. J., and Goetzmann, W. N. (2009). *Modern portfolio theory and investment analysis*. John Wiley & Sons.

Evren, Ö. and Ok, E. A. (2011). On the multi-utility representation of preference relations. *Journal of Mathematical Economics*, 47(4-5):554–563.

Färe, R., Grifell-Tatjé, E., Grosskopf, S., and Knox Lovell, C. (1997). Biased technical change and the malmquist productivity index. *The Scandinavian Journal of Economics*, 99(1):119–127.

Färe, R., Grosskopf, S., Norris, M., and Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American economic review*, 84(1):66–83.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3):253–290.

Figueira, J. R., Greco, S., Roy, B., and Slowinski, R. (2013). An overview of ELECTRE methods and their recent extensions. *Journal of Multi-Criteria Decision Analysis*, 20(1-2):61–85.

Figueira, J. R., Mousseau, V., and Roy, B. (2016). ELECTRE methods. In Greco, S., Ehrgott, M., and Figueira, J., editors, *Multiple criteria decision analysis: State of the art surveys*, pages 155–185. Springer.

Freudenberg, M. (2003). *Composite Indicators of Country Performance: A critical assessment*. OECD Science, Technology and Industry Working Papers. OECD Publishing.

Gan, X., Fernandez, I. C., Guo, J., Wilson, M., Zhao, Y., Zhou, B., and Wu, J. (2017). When to use what: Methods for weighting and aggregating sustainability indicators. *Ecological Indicators*, 81:491–502.

Giarlotta, A. and Greco, S. (2013). Necessary and possible preference structures. *Journal of Mathematical Economics*, 49(2):163–172.

Gilboa, I., Maccheroni, F., Marinacci, M., and Schmeidler, D. (2010). Objective and subjective rationality in a multiple prior model. *Econometrica*, 78(2):755–770.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153.

Greco, S., Ehrgott, M., and Figueira, J. (2016). *Multiple Criteria Decision Analysis: State of the Art Surveys*. International Series in Operations Research & Management Science. 2nd edition, New York: Springer.

Greco, S., Ishizaka, A., Matarazzo, B., and Torrisi, G. (2018a). Stochastic multi-attribute acceptability analysis (SMAA): an application to the ranking of Italian regions. *Regional Studies*, 52(4):585–600.

Greco, S., Ishizaka, A., Tasiou, M., and Torrisi, G. (2018b). On the methodological framework of composite indices: A review of the issues of weighting, aggregation and robustness, *Social Indicators Research*, doi: 10.1007/s11205-017-1832-9, advance online publication.

Greco, S., Mousseau, V., and Słowiński, R. (2008). Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):416–436.

Greco, S., Słowiński, R., Figueira, J., and Mousseau, V. (2010). Robust ordinal regression. *Trends in multiple criteria decision analysis*, pages 241–283.

Grupp, H. and Schubert, T. (2010). Review and new evidence on composite innovation indicators for evaluating national performance. *Research Policy*, 39(1):67–78.

Hansen, N. M. (1965). Unbalanced growth and regional development. *Economic inquiry*, 4(1):3–14.

Hartley, J. E. and Hartley, J. E. (2002). *The representative agent in macroeconomics*. London, United Kingdom: Routledge.

Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.

Helliwell, J., Layard, R., and Sachs, J. (2012). *World Happiness Report 2012*. New York: Sustainable Development Solutions Network.

Helliwell, J., Layard, R., and Sachs, J. (2017). *World Happiness Report 2017*. New York: Sustainable Development Solutions Network.

Hopkins, M. (1991). Human development revisited: A new undp report. *World Development*, 19(10):1469–1473.

Ishizaka, A. and Nemery, P. (2013). *Multi-Criteria Decision Analysis: Methods and Software*. Chichester, United Kingdom: John Wiley & Sons.

Kadziński, M. and Tervonen, T. (2013). Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. *European Journal of Operational Research*, 228(1):169–180.

Kirman, A. P. (1992). Whom or what does the representative individual represent? *The Journal of Economic Perspectives*, 6(2):117–136.

Kortelainen, M. (2008). Dynamic environmental performance analysis: a malmquist index approach. *Ecological Economics*, 64(4):701–715.

Kroll, C. and Delhey, J. (2013). A happy nation? Opportunities and challenges of using subjective indicators in policymaking. *Social Indicators Research*, 114(1):13–28.

Kunsch, P. L. and Ishizaka, A. (2018). Multiple-criteria performance ranking based on profile distributions: An application to university research evaluations. *Mathematics and Computers in Simulation*. Advance online publication, DOI: 10.1016/j.matcom.2018.05.021.

Lahdelma, R., Hokkanen, J., and Salminen, P. (1998). SMAA - Stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143.

Lahdelma, R. and Salminen, P. (2001). SMAA-2 : Stochastic Multicriteria Acceptability Analysis for Group Decision Making. *Operations Research*, 49(3):444–454.

Leskinen, P., Viitanen, J., Kangas, A., and Kangas, J. (2006). Alternatives to incorporate uncertainty and risk attitude in multicriteria evaluation of forest plans. *Forest Science*, 52(3):304–312.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

Marshall, A. (1961). *Principles of Economics. 9th (variorum) ed.* Macmillan.

Martin, R. (2011). Regional economic resilience, hysteresis and recessionary shocks. *Journal of economic geography*, 12(1):1–32.

Mazziotta, M. and Pareto, A. (2016). On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. *Social Indicators Research*, 127(3):983–1003.

McClure, S. M., Laibson, D. I., Loewenstein, G., and Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, 306(5695):503–507.

McFadden, D. (1981). Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272.

Mikulić, J., Kožić, I., and Krešić, D. (2015). Weighting indicators of tourism sustainability: A critical note. *Ecological Indicators*, 48:312–314.

Miller, T., Kim, A. B., and Roberts, J. M. (2018). *2018 Index of Economic Freedom*. Technical report, The Heritage Foundation.

Munda, G. (2005a). "Measuring sustainability": A multi-criterion framework. *Environment, Development and Sustainability*, 7(1):117–134.

Munda, G. (2005b). Multiple Criteria Decision Analysis and Sustainable Development. In Greco, S., Ehrgott, M., and Figueira, J., editors, *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 953–986.

OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing, Paris.

OECD (2016). *Strategic Orientations of the Secretary-General for 2016 and beyond*. Technical report, Meeting of the OECD Council at Ministerial Level. Retrieved from https://www.oecd.org/mcm/documents/strategic-orientations-of-the-secretary-general-2016.pdf.

Paruolo, P., Saisana, M., and Saltelli, A. (2013). Ratings and rankings: voodoo or science? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3):609–634.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Cambridge, MA: Harvard University Press.

Rawls, J. (2009). *A theory of justice*. Harvard university press.

Ray, A. K. (2008). Measurement of social development: an international comparison. *Social Indicators Research*, 86(1):1–46.

Reich, R. B. (2010). *Aftershock: The next economy and America's future*. Vintage.

Robbins, L. (1935). *An essay on the nature and significance of economic science*. McMillan.

Robinson, D. N. (1989). *Aristotle's psychology*. New York: Columbia University Press.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281.

Saaty, T. L. (1980). *The Analytic Hierarchy Process*. McGraw-Hill, New York.

Saisana, M., Saltelli, A., and Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 168(2):307–323.

Saltelli, A. (2007). Composite indicators between analysis and advocacy. *Social Indicators Research*, 81(1):65–77.

Samans, R., Blanke, J., Corrigan, G., and Drzeniek, M. (2018). The Inclusive Growth and Development Report 2018. In *Geneva: World Economic Forum*.

Schelling, T. C. (1980). The intimate contest for self-command. *The Public Interest*, 64:94–118.

Seiford, L. M. and Zhu, J. (2003). Context-dependent data envelopment analysis?measuring attractiveness and progress. *Omega*, 31(5):397–408.

Sen, A. (1970). Interpersonal aggregation and partial comparability. *Econometrica: Journal of the Econometric Society*, pages 393–409.

Sessions, R. (1950). How a 'difficult' composer gets that way; harpsichordist. *The New York Times*, Online article, Retrieved from: https://www.nytimes.com/1950/01/08/archives/how-a-difficult-composer-gets-that-way-harpsichordist.html, Accessed on 23 July 2018.

Sharpe, A. (2004). *Literature Review of Frameworks for Macro-indicators*. Centre for the Study of Living Standards, Ottawa.

Shin, D. C. and Johnson, D. M. (1978). Avowed happiness as an overall assessment of the quality of life. *Social indicators research*, 5(1-4):475–492.

Simar, L. and Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management science*, 44(1):49–61.

Spearman, C. (1904). "General Intelligence", Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

Stiglitz, J., Sen, A. K., and Fitoussi, J.-P. (2009). *The measurement of economic performance and social progress revisited: Reflections and Overview*. Commission on the Measurement of Economic Performance and Social Progress, Paris.

Stiglitz, J. E., Sen, A., and Fitoussi, J.-P. (2010). *Mismeasuring our lives: Why GDP doesn't add up*. The New Press.

Tervonen, T., Figueira, J., Lahdelma, R., and Salminen, P. (2009). SMAA-III: A simulation-based approach for sensitivity analysis of ELECTRE III. In *Real-Time and Deliberative Decision Making*, pages 241–253. Springer.

Tervonen, T. and Figueira, J. R. (2008). A survey on stochastic multicriteria acceptability analysis methods. *Journal of Multi-Criteria Decision Analysis*, 15(1-2):1–14.

Tervonen, T. and Lahdelma, R. (2007). Implementing stochastic multicriteria acceptability analysis. *European Journal of Operational Research*, 178(2):500–513.

UNDP (2010). *Human Development Report (HDR) 2010: The Real Wealth of Nations: Pathways to Human Development*. Technical report, United Nations Development Programme (UNDP). Retrieved from http://hdr.undp.org/en/content/human-development-report-2010.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Yang, L. (2014). *An Inventory of Composite Measures of Human Progress*. Technical report, UNDP Human Development Report Office.