Performance metrics in multi-objective optimization

Nery Riquelme
National University of Asuncion
Asuncion, Paraguay
Email: neryriquelme90@gmail.com

Christian Von Lücken
National University of Asuncion
Asuncion, Paraguay
Email: clucken@pol.una.py

Benjamín Barán
National University of Asuncion
Universidad Nacional del Este
Email: bbaran@pol.una.py

***Abstract*—In the last decades, a large number of metrics has been proposed to compare the performance of different evolutionary approaches in multi-objective optimization. This situation leads to difficulties when comparisons among the output of different algorithms are needed and *appropriate* metrics must be selected to perform those comparisons. Hence, no complete agreement on what metrics should be used exists. This paper presents a review and analysis of 54 multi-objective-optimization metrics in the specialized literature, discussing the usage, tendency and advantages/disadvantages of the most cited ones in order to give researchers enough information when choosing metrics is necessary. The review process performed in this work indicates that the hypervolume is the most used metric, followed by the generational distance, the epsilon indicator and the inverted generational distance.**

I. Introduction

Evolutionary Algorithms (EAs) proved to be capable of finding a good approximation to the Pareto optimal front in Multi-objective Optimization Problems (MOPs) where there exist two or more conflicting objective functions. In consequence, many EAs have been proposed during the last decades, giving rise to the need of establishing comparison methods in order to measure the quality of the solution sets obtained by different algorithms. In general, the performance of an EA is evaluated using experimental tests and, as a consequence, several performance metrics have been defined for this purpose. Metrics consider mainly three aspects of a solution set [1]:

- the convergence, i.e. the closeness to the theoretical Pareto optimal front;
- the diversity: distribution as well as spread; and
- the number of solutions.

Hence, it becomes intuitive to classify metrics considering the three aspects they account for. Alternatively, metrics usually are also categorized taking into consideration the number of solution sets they can simultaneously evaluate. In this regard, a metric may be unary or binary (h-ary in general) as will be discussed in Section II.

The real importance of studying metrics resides on their extended acceptance in the specialized community to perform experimental studies that are necessary to reflect in some way the output quality of different algorithms as well as to compare various approaches. Previous analysis and reviews on metrics can be found in [1], [2] and [3].

This paper presents a detailed study of performance metrics, analyzing their usage and behavior throughout last years, examining and taking as reference the works published in EMO (*Evolutionary Multi-Criterion Optimization*) conferences [4], [5], [6], [7] and [8]. EMO is one of the most relevant event specialized in evolutionary multi-objective optimization. The main motivation behind this study is the lack of research works that cover in details the usage and tendency of performance metrics for the multi-objective optimization field, their advantages and disadvantages. In fact, it is important for researchers in the multi-objective optimization field to know which metrics are the most convenient to quantify a given behavior. The idea of this work was born when discussing which metric should be used to prove the advantages of a new algorithm developed by some of the authors [9].

The rest of the paper is organized as follows: definitions and multi-objective optimization concepts are introduced in Section II. Then, the general methodology followed in this work and the key research questions are presented in Section III. Next, Section IV covers the results obtained and the corresponding analysis considering usage and tendency of performance metrics. Finally, Section V presents the conclusions and future works.

II. Definitions

Before analyzing in detail the information concerning the characteristics, usage and tendency of performance metrics, we formalize the basic concepts and terms used throughout the rest of the paper.

A. Decision variables

The decision variables are the numerical quantities for which values are to be chosen in an optimization problem [10]. These independent variables can be denoted as x_j, $j = \{1, 2...., n\}$.

Then, a vector $\mathbf{x}$ containing n decision variables can be represented by:

$$\mathbf{x}= [x_1, x_2...., x_n]^T$$

B. Objective functions

Objective functions are computable functions applied over the decision variables in order to have some criteria to evaluate the quality of a certain solution. They can be denoted as $f_1(\mathbf{x}), f_2(\mathbf{x})..., f_K(\mathbf{x})$, where K is the number of objective functions in the optimization problem being solved [10]. Then, a vector function containing K objective functions can be represented by:

978-1-4673-9143-6/15/$31.00

$$\mathbf{F}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})...., f_K(\mathbf{x})]^T$$

C. Spaces

An Euclidean n-space is defined as the set of all n-tuples of real numbers $\mathbb{R}^n$. When dealing with MOPs, two Euclidean spaces are covered [10]:

- the n-dimensional decision space, denoted as Ω, in which decision variables coexist and where each coordinate axis corresponds to a component of vector $\mathbf{x}$; and
- the K-dimensional objective space, denoted as Λ, in which objective functions coexist and where each coordinate axis corresponds to a component of vector $\mathbf{F}(\mathbf{x})$.

Notice that each point in Ω has its corresponding point in Λ. The former point represents a solution and the latter represents the quality of this solution. It is worth mentioning that more than one point in Ω may be mapped to the same point in Λ. For practical purposes, we consider that Ω contains only the feasible solutions.

D. Multi-objective Optimization Problem (MOP)

A general MOP is defined as:

$$\begin{cases} \textit{optimize } \mathbf{F}(\mathbf{x}) & \\ \textit{subject to} & g_s(\mathbf{x}) \leq 0 \qquad s = 1, 2, 3...S; \\ & h_j(\mathbf{x}) = 0 \qquad j = 1, 2, 3...J; \end{cases}$$

For a MOP, an EA optimizes (minimizes/maximizes) the K objective functions contained in the vector $\mathbf{F}(\mathbf{x})$.

$g_s(\mathbf{x}) \leq 0$ and $h_j(\mathbf{x}) = 0$ represent constraints that must be fulfilled while optimizing $\mathbf{F}(\mathbf{x})$ and Ω contains all feasible $\mathbf{x}$ satisfying all restrictions that can be used to optimize an objective function $\mathbf{F}(\mathbf{x})$.

E. Pareto Dominance

Given two vectors $\mathbf{x}, \mathbf{x}' \in \Omega$, the vector $\mathbf{x}$ is said to dominate $\mathbf{x}'$, $\mathbf{x} \succ \mathbf{x}'$, iff $\mathbf{x}$ is not worse than $\mathbf{x}'$ in any objective function and it is strictly better in at least one objective function [11].

If neither $\mathbf{x}$ dominates $\mathbf{x}'$, nor $\mathbf{x}'$ dominates $\mathbf{x}$, $\mathbf{x}$ and $\mathbf{x}'$ are said to be no-comparable, denoted as $\mathbf{x} \sim \mathbf{x}'$.

F. Pareto optimal set and Pareto optimal front

For a given MOP, the Pareto optimal set (P^*) is the set containing all the solutions that are non-dominated with respect to Ω. It can be denoted [12]:

$$P^* := \{\mathbf{x} \in \Omega | \neg\exists \mathbf{x}' \in \Omega \textit{ such that } \mathbf{x}' \succ \mathbf{x}\} \qquad (1)$$

Then, for a given MOP and its corresponding Pareto optimal set P^*, the Pareto optimal front (PF^*) is the result of mapping P^* to Λ. PF^* is defined as [12]:

$$PF^* := \{\mathrm{F}(\mathbf{x}) \in \Lambda \mid \mathbf{x} \in P^*\} \qquad (2)$$

There are cases, specially in real-world problems, when the Pareto optimal front cannot be calculated for diverse reasons. Then, a reference set R can be considered. A reference set is an approximation to the Pareto optimal front that is used when PF^* is unknown, containing all non-dominated solutions already known.

It is worth emphasizing that for practical real-world applications, it is not always necessary to calculate the Pareto optimal set being enough the calculation of a good approximation to it with its corresponding Pareto front approximation.

G. Approximation set

An approximation set is defined by Zitzler *et al.* as follows [2]: let $A \subseteq \Lambda$ be a set of objective vectors. A is called an approximation set if any element of A does not dominate or is not equal to any other objective vector in A. The set of all approximation sets is denoted as Z. As mentioned above, the result of solving a real-world problem usually is an approximation set A and not the Pareto optimal front PF^*.

H. Performance metric

Given h approximation sets $A_i \in Z, i = 1, 2...h$, an h-ary performance metric or quality indicator, I, is defined by Zitzler *et al.* in [2] as a function $I : Z^h \mapsto \mathbb{R}$, which assigns to each set $(A_1, A_2, ..., A_h)$ a real value $I(A_1, A_2, ..., A_h)$ usually used to compare the *quality* of different algorithms when solving MOPs.

I. Metrics classification

As said before, there are two main ways of categorizing metrics. The first classification criterion considers the aspects that metrics measure when evaluating the approximation sets in Z. Considering an approximation set A, as it will be shown in Table IV, the metrics can be grouped as [1]:

- **Cardinality metrics:** the cardinality of A refers to the number of solutions that exists in A. Intuitively, a larger number of solutions is preferred.
- **Accuracy metrics:** this aspect refers directly to the convergence of A. In other words, it indicates *how distant* is A from the theoretical Pareto optimal front PF^*. Notice that when the Pareto optimal front is unknown, a reference set R is considered instead.
- **Diversity metrics:** distribution and spread are two very closely related facets, yet they are not completely the same [1]. The distribution refers to the relative distance among solutions in A [11] while the spread refers to the range of values covered by the solutions in A [3]. The spread is also known as "the extent" of an approximation set.

The second classification criterion can be easily deduced from the formal definition of performance metric given in Section II-H. It takes into consideration the number h of approximation sets that will be evaluated by the metric. Two types of metrics have been used in the specialized literature:

- **Unary metrics:** The metric is said to be unary if it receives as parameter only one approximation set A to be evaluated. Formally, an unary metric is a function denoted as $I(A) : Z \mapsto \mathbb{R}$.
- **Binary metrics:** The metric is said to be binary if it receives as parameter two approximation sets, A and B,

to be compared. Formally, a binary metric is a function denoted as $I(A, B) : Z^2 \mapsto \mathbb{R}$.

Notice that unary metrics give a real value after considering one or more of the three aspects mentioned above while binary metrics consider mainly the relationship between two approximation sets in terms of dominance to give an idea of which one is better. It is worth mentioning that unary metrics that measure accuracy need one of the following parameters:

1) a reference point. This parameter is specifically used to compute the hypervolume metric [13] and it is employed to calculate the space covered by solutions in the objective space Λ;
2) metrics like GD, IGD, Υ, etc. require the points in the Pareto optimal front PF^* to calculate convergence. When PF^* is unknown, a reference set R is used to estimate the closeness to PF^*.

It is important to remark that the vast majority of existing metrics are unary. This fact will be easily noticed in Table IV.

A more detailed discussion on unary and binary metrics, their advantages and limitations, can be found in [2].

III. Methodology

This section describes the general methodology followed in this work to analyze the state-of-the-art usage of performance metrics in evolutionary multi-objective optimization benchmarking.

First, it was necessary to define an universe of articles to be analyzed. EMO is a bi-annual international conference series, dedicated to advances in the theory and practice of evolutionary multi-criterion optimization [8]. EMO conference was selected to serve as source of information for this study since it is a top event in the field and hence it can reflect the real state of the art. This survey covers the 3rd (2005) [4], 4th (2007) [5], 5th (2009) [6], 6th (2011) [7] and 7th (2013) [8] editions of EMO. The 8th (2015) edition [54] was not included since its proceedings were published after this survey was completed.

After defining the universe of articles and since the scope of EMO is quiet extensive, an inclusion criterion was established to have the samples of relevant articles for this study. Let U be the universe of articles containing all the research works published in the five mentioned editions of EMO and x an article, the set of relevant articles W was fulfilled according to the following inclusion criterion:

$$W := \{x \in U | x \textit{ uses at least one performance metric}\}$$

With the inclusion criterion properly defined, the research questions that this work aims to answer can be presented:

- *What is the set of performance metrics used by the EMO community? how has that set of performance metrics evolved?*
- *Considering each edition of EMO individually, what is the percentage of articles that apply at least one performance metric? how has that percentage varied throughout EMO editions?*

TABLE I
All performance metrics cited in EMO conference from 2005 to 2013

Id	Performance metrics	Symbol
1	Attainment functions approach metric [14]	-
2	Convergence metric [15]	CM
3	Cluster metric [16]	Cl_μ
4	Consolidation Ratio [17]	CR
5	Contribution metric [18]	-
6	Convergence index [19]	-
7	Convergence measure [20]	Υ
8	Coverage [21]	-
9	Coverage of the front [22]	-
10	D metric [23]	D
11	D quantifier [24]	-
12	D_1R [25]	-
13	Distance-based indicator [26]	$D_{(p)}$
14	Diversity metric [15]	DM
15	Dominance-based quality [27]	DQ_p
16	Entropy metric [28]	-
17	Epsilon family [2]	ϵ
18	Generational distance [12]	GD
19	Hypervolume [13]	HV
20	Hypercube-based diversity metric [20]	-
21	Inverted generational distance [29]	IGD
22	M_1^* metric [3]	M_1^*
23	M_3^* metric [3]	M^*3
24	M_3^* metric Improved [30]	-
25	Maximum crowding distance [31]	MCD
26	Mean Absolute Error [32]	-
27	Minimal distance graph [33]	MDG
28	Mutual domination rate [34]	MDR
29	Non-dominated evaluation metric [11]	-
30	ONVG [35]	-
31	Overall Pareto Spread [16]	-
32	Population per run [36]	-
33	R-metric [37]	R
34	Ratio of non-dominated individuals [38]	RNI
35	Ratio of non-dominated solutions [39]	-
36	Relations between non-dominated fronts [40]	-
37	Spacing [35]	Sp
38	Sparsity Index [19]	-
39	Spread measure [41]	-
40	Spread metric [42]	-
41	Spread: Delta indicator [11]	Δ
42	Spread: Generalized Spread [43]	Δ^*
43	Success Counting [44]	SCC
44	Sum of maximum objective values [45]	Smax
45	The adjusted Rand Index [46]	-
46	The average distance of points to the PF [47]	-
47	The delineation metric [48]	-
48	The diversity metric [48]	-
49	The front Spread [49]	FS
50	The front to Set Distance [49]	(DPF¿S)
51	Two set coverage [50]	C
52	μ_d metric [51]	μ_d
53	Volume measure [52]	V_p(A)
54	ϵ-performance metric [53]	-

- *What metrics were used the most at each EMO edition?*
- *What is the tendency of the most-used metrics?*
- *What metrics tend to decrease/increase in usage?*
- *What is the average number of metrics used per article in EMO?*

To calculate the tendency of the performance metrics we used the least squares regression approach.

From a total of 262 articles published in the five considered

TABLE II
TOTAL OF WORKS PUBLISHED IN EMO EDITIONS AND THE NUMBER OF WORKS THAT USES PERFORMANCE METRICS

Articles	EMO 2005	EMO 2007	EMO 2009	EMO 2011	EMO 2013	Total
Published articles (U)	59	65	39	42	57	262
Articles using metrics (W)	33	29	23	21	33	139
Percentage (%)	**55.93**	**44.62**	**58.97**	**50.00**	**57.89**	**53.05**

editions of EMO only 139 satisfied the inclusion criterion (see details in Table II). 54 metrics were found in EMO as a result of applying the above described methodology. Table I summarizes all those 54 metrics and their standard symbols.

In the next section, results and discussion about ranking and tendency of performance metrics are presented.

IV. RESULTS

This section presents the answers to the key research questions stated in section III.

A. Usage of metrics

First, it is important to show how the usage of metrics behave during the last years. Table II reflects the usage of metrics from 2005 to 2013 by presenting the number of research articles that used at least one metric per EMO edition.

Remember that the total number of articles published in EMO is the universe U of papers and the total number of articles that used performance metrics corresponds to the set W of articles that satisfied the condition applied by the inclusion criterion. Then, let U_{year} and W_{year} be the total number of articles published in a $year$ and the number of articles that used metrics in that $year$. Then, from Table II: W_{2005} represents the 55.93% of U_{2005}, W_{2007} represents the 44.61% of U_{2007}, W_{2009} corresponds to the of 58.97% U_{2009}, W_{2011} is exactly the 50% of U_{2011}, W_{2013} represents 57.89% of U_{2013} and finally W corresponds to the 53.05% of U. It is interesting how the citation of metrics reached the lowest point in 2007 but two years later it reached its highest point in 2009. Notice that the usage of metrics has not varied meaningfully from 2005 to 2013 being a relevant comparison tool for most EMO publications, proving its acceptance in the research community.

Table III shows how many performance metrics were cited per article in EMO. It is clear that the vast majority of articles used between 1 and 2 performance metrics. Also, it is notable how the number of articles using more than 2 metrics have decreased since 2007. The average of performance metrics used per article considering all EMO editions is 2.11. Note also that number of papers clearly decreases with the number of metrics.

Finally it should be mentioned that Wagner *et al.* [55] cited 9 performance metrics, the maximum number in the studied set of articles W.

TABLE III
NUMBER OF METRICS USED PER ARTICLE FROM 2005 TO 2013

EMO editions	Quantity of metrics employed per article					
	1	2	3	4	>4	Total
2005	12	9	7	2	3	33
2007	7	9	8	4	1	29
2009	12	4	5	1	1	23
2011	10	4	4	1	2	21
2013	19	10	2	1	1	33
Total	**60**	**36**	**26**	**9**	**8**	**139**
Percentage (%)	**43.17**	**25.90**	**18.71**	**6.47**	**5.76**	**100**

B. Ranking

Now, the ranking of metrics throughout the years can be presented. For space reasons, the following ranking only covers the TOP metrics considering the number of citation achieved by the metrics in W.

Table IV presents the top-ten number of citations in EMO and the corresponding metrics. The most used performance metric was definitively the hypervolume metric with 91 citations. The hypervolume (HV) [13], also known as S metric, hyper-area or Lebesgue measure, is an unary metric that measures the size of the objective space covered by an approximation set. A reference point must be used to calculate the mentioned covered space. HV considers all three aspects: accuracy, diversity and cardinality, being the only unary metric with this capability. It has been widely accepted since it offers the following unique and desirable properties [56]: i) whenever one approximation set completely dominates another approximation set, the hypervolume of the former will be greater than the hypervolume of the latter. As a consequence, HV is said to be *Pareto compliant*; ii) as a result from the just mentioned property, hypervolume guarantees that any approximation set A that achieves the maximum possible quality value for a particular MOP, contains all Pareto optimal solutions. Besides, a binary version of this metric was proposed in [23] to give HV the capability of assessing the dominance relationship between two approximation sets.

The generational distance metric (GD) occupies the second position of the ranking with 26 citations. GD takes as reference an approximation set A and calculates *how far* it is from the Pareto optimal front PF^* (or reference set R). This unary measure considers the average Euclidean distance between the members of A and the nearest member of PF^* [12]. It can be noticed that GD considers only one aspect of A: the accuracy.

The third most-used metric is epsilon (ϵ) [2]. Epsilon is a binary indicator that gives a factor by which an approximation set is worse than another considering all objectives. Formally, let A and B be two approximation sets, then $\epsilon(A, B)$ equals the minimum factor ϵ such that for any solution in B there is at least one solution in A that is not worse by a factor of ϵ

TABLE IV
TOP TEN OF THE MOST USED METRICS IN EMO FROM 2005 TO 2013

Ranking	Citations	Metrics	Classification	
			Aspects	Sets
1°	91	Hypervolume (HV)	- *Accuracy* - *Diversity*	Unary
2°	26	Generational distance (GD)	- *Accuracy*	Unary
3°	23	Epsilon family (ϵ)	all	Binary
4°	17	Inverted generational distance (IGD)	- *Accuracy* - *Diversity*	Unary
	17	Spread: Delta indicator (Δ)	- *Diversity*	Unary
	17	Two set coverage (C)	all	Binary
5°	9	ONVG	- *Cardinality*	Unary
	9	R-metric	all	Binary
6°	8	Convergence measure (Υ)	- *Accuracy*	Unary
7°	6	Convergence metric (CM)	-*Accuracy*	Unary
	6	D_1R	- *Accuracy* - *Diversity*	Unary
	6	Spacing (Sp)	- *Diversity*	Unary
8°	5	M_3^* metric	- *Diversity*	Unary
9°	4	M_1^* metric	- *Accuracy*	Unary
10°	3	Diversity metric (DM)	- *Diversity*	Unary
	3	Entropy metric	- *Diversity*	Unary
	3	Spread measure	- *Diversity*	Unary

considering all objectives [2].

In the fourth position there are three metrics with 17 citations each: inverted generational distance (IGD), spread (Δ) and two set coverage (C). Each of these metrics considers different criteria of an approximation set. IGD [10] is an inverted variation of GD but it presents significant differences with GD: i) it calculates the minimum Euclidean distance (instead of the average distance) between an approximation set A and the Pareto optimal front PF^*, ii) IGD uses as reference the solutions in PF^* (and not the solutions in A) to calculate the distance between the two sets and iii) if sufficient members of PF^* are known, IGD could measure both the diversity and the convergence of A [57].
The Δ metric is an unary indicator that measures the distribution and extent of spread achieved among the solutions in A. As shown in Table IV, Δ considers only one aspect: diversity of solutions.
Two set coverage [23], usually named C-metric or simply "coverage", is a binary indicator that can be described as follows: let A and B be two approximation sets. $C(A, B)$ gives the fraction of solutions in B that are dominated by at least one solution in A [2]. Hence, $C(A, B) = 1$ means that all solutions in B are dominated by at least one solution in A while $C(A, B) = 0$ implies that no solution in B is dominated by a solution in A.

The fifth position in Table IV is occupied by two metrics, ONVG and R-metric, having 9 citations each. ONVG [35] is an unary cardinality metric that gives the total number of solutions found in an approximation set [12]. R-metric [37] is a family of three binary indicators: R_1 , R_2 and R_3 that are based on a set of utility functions. Basically, the idea is that for any two approximation sets (A and B), these metrics use a set of utility functions and determine the expected number of occasions the solutions of one set are better than the solutions of the other. That is, these R-metrics declare as *the winner* the set that will be the choice of most decision-makers [15]. R indicators give the relative quality of two approximation sets. A further reading on R-metric can be done in [37] and [58].

The sixth position is for the convergence measure (Υ) with 8 citations. This unary metric measures the extent of convergence to a known set of Pareto optimal solutions. Υ uses the average Euclidean distance among solutions in an approximation set and Pareto optimal solutions to calculate the convergence. It is important to mention that Υ can also provide some information about the spread in the solutions [20]. This accuracy metric need as parameter the Pareto optimal front PF^* or a reference set R.

The seventh position in Table IV belongs to three unary metrics with 6 citation each: convergence metric (CM), D_1R and spacing (Sp). CM calculates per each run of an EA the normalized minimum Euclidean distance from an approximation set to the Pareto optimal front [15]. The average of these distances is the value given by CM. D_1R [25] is very similar to IGD, but it measures the mean distance, over the points of the Pareto optimal front (or reference set), of the nearest point in an approximation set [59] and, as IGD, D_1R can be used to measure convergence and diversity. On the other hand, Spacing (Sp) [35] was designed to measure how evenly the members of an approximation set are distributed. A value of 0 for Sp means that all members of the approximation set are equidistantly spaced.

Having 5 citations, the eighth position is for the M_3^* diversity metric. This unary function was defined in [3] along with two other complementary metrics in the objective space: M_1^* and M_2^*. This metric considers the extent of an approximation set. M_3^* is usually named as "extent metric" or "maximun spread".

In the ninth position resides the convergence metric M_1^* with 4 citations. As just mentioned, M_1^* complements with

two other metrics, M_2^* and M_3^*, to asses the quality of an approximation set A [3]. M_1^* gives the average distance of A to the Pareto optimal front (or reference set). This metric is also named "Average distance".

The last position in this ranking is shared by three unary metrics: diversity metric (DM), the entropy metric and the spread measure with 3 citations each. The entropy metric [28] and DM [15] are two very related diversity metrics that apply an entropy concept to calculate the diversity of solutions. DM is based on the entropy metric. Both of them basically attempt to project the solutions of an approximation set on a suitable hyperplane assigning them entropy functions that later will be added together to compose a normalized entropy function. Further explanation on these metrics can be found in [28] and [15]. The spread measure is an alternative to calculate the diversity in an approximation set. Basically, this metric gives a notion of the spread by using the sum of the width of each objective [41].

Finally, the top-five metrics of each EMO edition are shown with more details in Table V. Note that for every metric the number of citations is specified below. Metrics with equal number of citations occupy the same positions in the ranking. Once again, HV beats all other metrics in every EMO edition. It is remarkable how ϵ appeared in 2007 and since then it has been one of the most cited metrics. A similar situation can be observed with respect to IGD. On the other hand, metrics like C, Υ and Δ have decreased in usage dramatically. GD metric achieved its highest point in 2009 and then fell, but it is still a widely-used metric.

An interesting phenomenon can also be appreciated in Table V: the variety of cited metrics was reduced notably throughout the years. In 2005 nine metrics conformed the top five of the most cited metrics, in 2007 and 2009 the number decreased to 8, in 2011 the number was reduced to 7 and finally in 2013 only 6 metrics conformed the top five. This indicates that the process of establishing an "ideal" set of metrics to asses the performance of EAs is somehow advancing, showing a process of maturity in the filed.

C. Tendency

This part of the paper estimates how the usage of top metrics may behave in the following years. For space reasons, we included the analysis of tendency of the first five positions of the ranking displayed in Table IV (the 8 most cited metrics). As stated in Section III, the least squares method was employed to calculate the regression line showed in the following figures. Note that figures show normalized values to make the numbers of citations per EMO edition comparable. To normalize, we divided (for each metric) the number of citation achieved in a $year$ by the number of articles in the set W_{year}. The trend line equation is also displayed with each figure.

1) HV: The usage and tendency of Hypervolume is displayed in figure 1. The number of citations in every EMO edition shows that HV is by far the most accepted metric in the community. Notice that in 2013 almost 82% of the articles used this metric for experimental purposes, implying an enormous growth of almost 43% in citations with respect to 2005. Hence, the regression line for the hypervolume is ascendant. This means in practical terms the HV will still be used in the next years at least as much as it has been used until 2013. This indicator gained special attention for the properties mentioned in Subsection IV-B but it has some biases that must be considered: first of all, HV requires preference information for choosing the reference point and even though there is some freedom to choose this point, that could imply in some cases information that may not be appropriate in specific situations [56]. Second, HV may be misleading if the approximation set A is non-convex. This bias and its possible solution are covered in detail in [12]. Third, many-objective problems are gaining attention in the multi-objective optimization field i.e. multi-objective problems with large number K of objective functions. The worst-case computational complexity of this metric is exponential with respect to the number of objective functions K [60]. In consequence, it becomes unsuitable to apply HV for many-objective problems. Chan [60], Beume *et al.* [61] and Fonseca *et al.* [62] presented alternatives to deal with this situation.

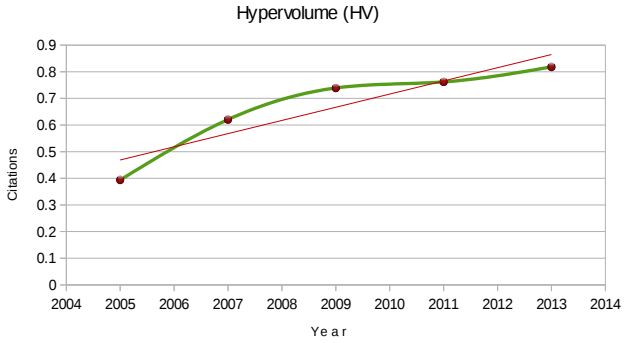

Fig. 1. Tendency of hypervolume metric (the trend line is defined by the equation: $citations = 0.049\ year - 98.748$)

2) GD: Figure 2 shows the tendency of the second most-used metric: the generational distance. It is interesting to note that while HV was increasing in use, the opposite happened with GD. However, the proportion of citations in 2013 improved for GD with respect to the years 2011 and (specially) 2009. Generational distance presents the advantage of being *lightweight* (if compared with the hypervolume, for example) in terms of computational costs and, combined with others metrics, can give a very acceptable notion of quality for an approximation set A. Yet, two main drawbacks can be identified [11]: i) if there exist an approximation set A with a large fluctuation in the distance values that are calculated per generation, the metric may not reveal the true distance; ii) it is necessary to know the Pareto optimal front or at least a large number of Pareto optimal solutions.

Although the regression line clearly indicates that this metric is decreasing in use, GD is still employed in many studies and hence considerable amount of time shall pass before researchers desist from using it.

TABLE V
TOP-FIVE METRICS CONSIDERING EACH EMO EDITION INDIVIDUALLY

Ranking	EMO 2005	EMO 2007	EMO 2009	EMO 2011	EMO 2013
1°	HV 13 citations	HV 18 citations	HV 17 citations	HV 16 citations	HV 27 citations
2°	GD and C 7 citations	ϵ 9 citations	ϵ 5 citations	ϵ 5 citations	IGD 7 citations
3°	Δ 6 citations	GD and Δ 6 citations	GD 4 citations	IGD 4 citations	GD 6 citations
4°	Υ 4 citations	C 5 citations	IGD and Δ 3 citations	GD and C 3 citations	ϵ 4 citations
5°	CM, D_1R, M_1^* and Sp 3 citations	Υ, IGD and M_3^* 3 citations	C, ONVG and R 2 citations	Δ and ONVG 2 citations	ONVG and R 2 citations
Total number of metrics	**9**	**8**	**8**	**7**	**6**
Total number of citations	**33**	**41**	**31**	**30**	**46**

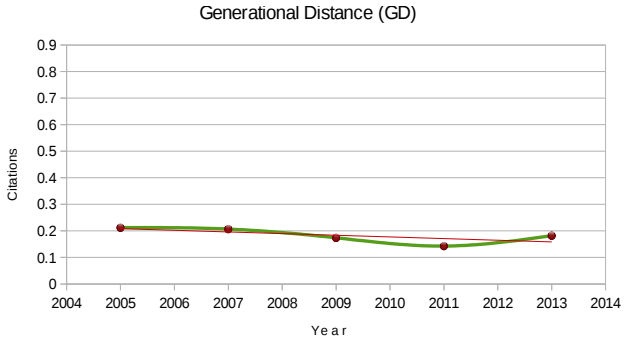

Fig. 2. Usage and tendency of generational distance (the trend line is defined by the equation: $citations = -0.006\ year + 12.704$)

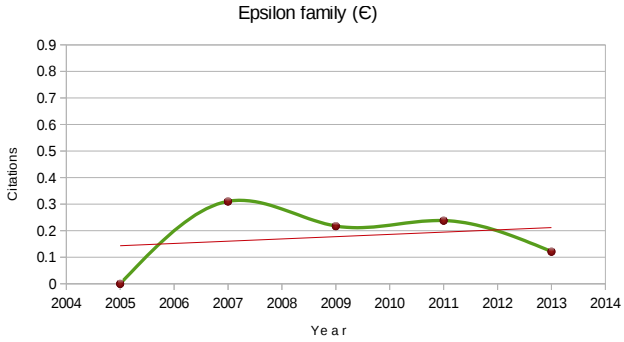

Fig. 3. Usage and tendency of epsilon family (the trend line is defined by the equation: $citations = 0.008\ year - 16.916$)

3) ϵ: The epsilon family consists in a variety of unary and binary indicators, and it appeared remarkably in EMO 2007. Since then, this metric has gained a place among the well-accepted approaches to asses the quality of solution sets. The usage and tendency of ϵ can be seen in Figure 3. Although it has reached its highest point it 2007 (31% of citations) and then decreased in usage, what the regression line suggests is interesting and expectable considering that binary ϵ has a valuable feature: ϵ can give complete information about the relationship of two approximation set. In fact, this indicator can detect whether an approximation set is better than another. ϵ indicator is inexpensive to compute [2] and thus it represents a very viable alternative specially when dealing with many-objective problems.

4) IGD: Figure 4 reflects how the inverted generational distance metric has been used and how its usage tends to increase in the future. It seems very clear that IGD has gained more and more attention throughout the years, showing a total increase of approximately 18% in citations from 2005 to 2013. What makes IGD attractive are basically its low computational cost and its capability of considering not only the convergence of an approximation set, but also the diversity if sufficient optimal solutions are known. The latter feature may lead to a drawback for this metric when the Pareto optimal front is unavailable and the reference set contains only few optimal solutions. This drawback is probable to occur

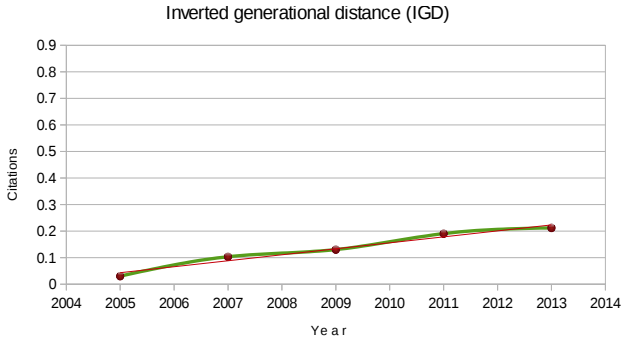

Fig. 4. Usage and tendency of inverted generational distance (the trend line is defined by the equation: $citations = 0.022\ year - 45.136$)

when dealing with many-objective problems. This difficulty and some strategies to overcome it are discussed in [63]. Still, IGD represents an excellent alternative in many-objective problems to other computationally expensive metrics like the hypervolume. Consequently, the considerable slope of the IGD regression line and its simplicity compared to HV may suggest that IGD usage will grow as many-objective problems become more popular considering the complexity of calculating HV.

5) Δ: This metric is the preferred diversity indicator in the literature and it was widely used in previous years but Figure 5 shows that Δ experimented an important decrease

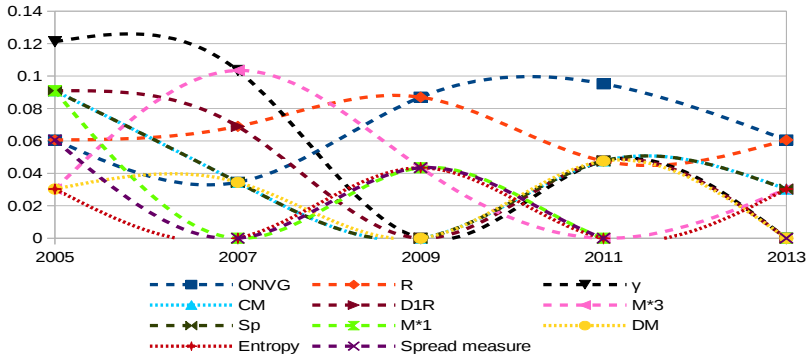

Fig. 7. Usage of the rest of the top-ten metrics

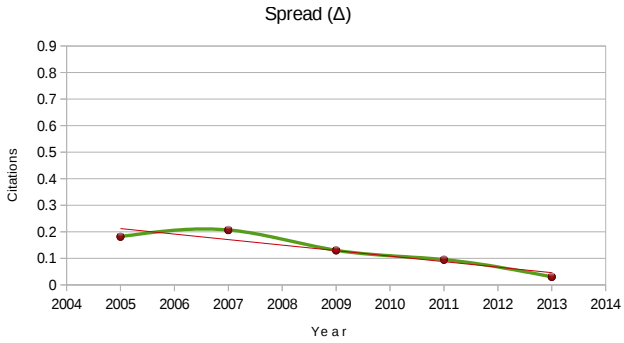

Fig. 5. Usage and tendency of Δ.(the trend line is defined by the equation: $citations = -0.021\, year + 41.784$)

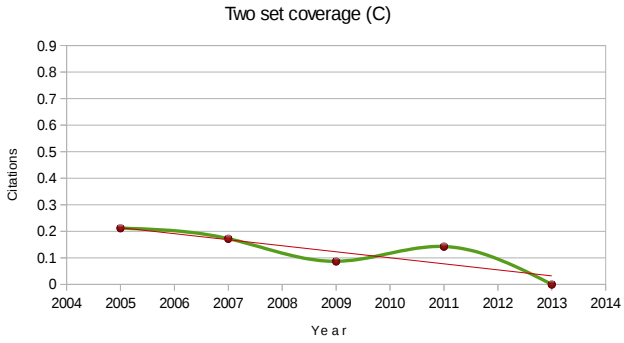

Fig. 6. Usage and tendency of two sets coverage (the trend line is defined by the equation: $citations = 0.023\, year + 45.707$)

in usage in the last years. This diversity indicator diminished in citations due to the existence of other metrics that provide more complete information on the quality of solution sets by considering more than one approximation set (binary metrics) or by taking into account more than one aspect. Then, the major drawback of Δ is that it measures only the diversity of the approximation set A. Yet, Δ combined with some other convergence metrics is still an interesting approach to asses quality at a low computational cost.

6) C: Similar to the spread's situation, Figure 6 displays how C metric was decreasing in citations in the late years. The percentage of citations decreased from approximately 21% to 9% in the lapse 2005-2009; in 2013 no article cited this metric. C was used in a vast quantity of studies because the idea of applying a performance metric that could reflect information on relative quality between two approximation sets A and B was very interesting. In fact, this quality indicator was the first widely-accepted binary metric. Apparently, the rise of metrics like hypervolume and (specially) epsilon has impacted directly in the usage of this metric given that ϵ indicator can reflect more aspects of the relative quality of A and B at a low computational cost. Although coverage is capable of detecting dominance between approximation sets, it does not provide additional information about this dominance. Furthermore, C values are often difficult to interpret when solution sets A and B are no-comparable [2]. This scenario commonly takes place in many-objective problems. Regarding many-objective problems and the algorithms used to solve them, Von Lücken *et al.* published a complete survey.

Finally, the usage of the remaining metrics in Table IV can be seen in Figure 7. Note that in Figure 7 a different Y-axis scale is used to easily appreciate the behavior of the less-used metrics. All these metrics clearly tend to continue decreasing in citations or at best maintain their low usage. There may be several explanations on why these performance metrics were used in many studies in the past years but in the late years researchers somehow desisted from using them. An important drawback for Sp, DM, entropy metric, M_3^* and spread measure is that they only consider the diversity of solutions in A. In fact, Sp only considers the distribution (and not the spread) of solutions at a high computational cost [11]. With M_3^* occurs the opposite, M_3^* considers only the extent of A, leaving out information about the distribution. CM, Υ, M_1^* and D_1R are all convergence indicator (like GD and IGD) that were

successfully applied in many articles and they implemented diverse strategies to estimate in some way the closeness to the Pareto optimal front. However, having a set conformed by many indicators measuring the same aspect requires at some point the selection of representative *members*. These representative convergence metrics were with no doubt the GD and IGD indicators. Both became almost standard when measuring convergence is needed, leaving the other metrics in background. R-metric and ONVG somehow have maintained their citation level during the years. They never achieved an important number of citations but the community has not desisted from using them. This may indicate for ONVG that counting the number of solutions obtained for a MOP is still a useful information for experimental purposes. R-metric has been accepted as a valid alternative binary metric by a small part of the community and that it still can evolve in order to gain more acceptance in the future, even though its usage is not relevant nowadays.

V. Conclusion and future works

This paper presented an extensive review of the performance metrics based on the research articles published in a specialized event: Evolutionary Multi-criterion Optimization conference. We revealed all the performance metrics used by the EMO community during 8 years, from 2005 to 2013 (see Table I). We identified the most used metrics in terms of the citations they have received and built a top-ten ranking (see Table IV). The top-ten metrics were defined and, for the first fifth positions of the mentioned ranking, an analysis of usage and tendency was performed.

Considering the results obtained in this work, we can summarize the following conclusions:

- the hypervolume (HV) is the unary metric preferred by the research community. In fact, hypervolume was the most used metric beyond the classification method chosen.
- With regard to convergence metrics, generational distance (GD) and inverted generational distance (IGD) are the most used metrics. While GD is decreasing in usage, IGD is clearly getting more acceptance.
- With regard to diversity metrics, the spread (Δ) was the most accepted diversity metric. However, in the last years, the community has not shown interest in measuring only the diversity of solutions anymore, decreasing the relevance of this metric.
- With regard to cardinality metrics, ONVG was the metric used and it is still used in some works, even though it is not very relevant for most researchers.
- With respect to binary metrics, two set coverage (C) was widely used before but now the community has paid more attention to the ϵ indicator. On the other hand, R-metric is a binary metric that was cited in every EMO edition but it is still in process of being accepted.
- HV and IGD are the metrics that show the most notable in-ascendant tendency. Moreover, the regression line for IGD indicates that this metric is experimenting a considerable growth in usage. Even though the hypervolume also shows a remarkable growth, in the following years it may tend to stop its growth in citations given the rise of many-objective problems where HV is not suitable due to its exponential computational complexity.
- Finally, it is important to state that the variety of performance metrics has been interestingly reduced. For every measured aspect in the solution sets, there exist metrics that have already became standard to perform the measurement, denoting a maturity in the multi-objective optimization field.

At the moment of the current redaction, the authors are working in a more complete survey not presented in this paper for space reasons.

References

[1] T. Okabe, Y. Jin, and B. Sendhoff, "A critical survey of performance indices for multi-objective optimisation," in *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, vol. 2. IEEE, 2003, pp. 878–885.

[2] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. Da Fonseca, "Performance assessment of multiobjective optimizers: an analysis and review," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 2, pp. 117–132, 2003.

[3] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evolutionary computation*, vol. 8, no. 2, pp. 173–195, 2000.

[4] C. C. Coello, A. H. Aguirre, and E. Zitzler, *Evolutionary Multi-Criterion Optimization: Third International Conference, EMO 2005, Guanajuato, Mexico, March 9-11, 2005, Proceedings*. Springer, 2005, vol. 3410.

[5] H. Nakayama, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata, *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007. Proceedings*, 1st ed., ser. Lecture Notes in Computer Science 4403. Springer-Verlag Berlin Heidelberg, 2007.

[6] M. Laguna, M. Ehrgott, C. M. Fonseca, X. Gandibleux, J.-K. Hao, and M. Sevaux, *Evolutionary Multi-Criterion Optimization: 5th International Conference, EMO 2009, Nantes, France, April 7-10, 2009. Proceedings*, 1st ed., ser. Lecture Notes in Computer Science 5467 : Theoretical Computer Science and General Issues. Springer-Verlag Berlin Heidelberg, 2009.

[7] S. Bandaru, K. Deb, R. H. C. Takahashi, E. F. , Wanner, and S. Greco, *Evolutionary Multi-Criterion Optimization: 6th International Conference, EMO 2011, Ouro Preto, Brazil, April 5-8, 2011. Proceedings*, 1st ed., ser. Lecture Notes in Computer Science 6576. Springer-Verlag Berlin Heidelberg, 2011.

[8] R. C. Purshouse, P. J. Fleming, C. M. Fonseca, S. Greco, and J. Shaw, *Evolutionary Multi-criterion Optimization: 7th International Conference, EMO 2013, Sheffield, UK, March 19-22, 2013. Proceedings*. Springer, 2013.

[9] C. von Lücken, C. Brizuela, and B. Barán, "Clustering based parallel many-objective evolutionary algorithms using the shape of the objective vectors," in *Evolutionary Multi-Criterion Optimization*. Springer, 2015, pp. 50–64.

[10] C. A. C. Coello, D. A. Van Veldhuizen, and G. B. Lamont, *Evolutionary algorithms for solving multi-objective problems*. Springer, 2002, vol. 242.

[11] K. Deb, *Multi-objective optimization using evolutionary algorithms*. John Wiley & Sons, 2001, vol. 16.

[12] D. A. Van Veldhuizen, "Multiobjective evolutionary algorithms: classifications, analyses, and new innovations," DTIC Document, Tech. Rep., 1999.

[13] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach," *evolutionary computation, IEEE transactions on*, vol. 3, no. 4, pp. 257–271, 1999.

[14] C. M. Fonseca, V. G. Da Fonseca, and L. Paquete, "Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 250–264.

[15] K. Deb and S. Jain, "Running performance metrics for evolutionary multi-objective optimizations," in *Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning (SEAL'02),(Singapore)*. Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning (SEAL'02),(Singapore), 2002, pp. 13–20.

[16] J. Wu and S. Azarm, "Metrics for quality assessment of a multiobjective design optimization solution set," *Journal of Mechanical Design*, vol. 123, no. 1, pp. 18–25, 2001.

[17] T. Goel and N. Stander, "A non-dominance-based online stopping criterion for multi-objective evolutionary algorithms," *International Journal for Numerical Methods in Engineering*, vol. 84, no. 6, pp. 661–684, 2010.

[18] H. Meunier, E.-G. Talbi, and P. Reininger, "A multiobjective genetic algorithm for radio network optimization," in *Evolutionary Computation, 2000. Proceedings of the 2000 Congress on*, vol. 1. IEEE, 2000, pp. 317–324.

[19] M. Nicolini, "Evaluating performance of multi-objective genetic algorithms for water distribution system optimization," in *Proceedings of the 6th International Conference*, vol. 21. World Scientific, 2004, p. 24.

[20] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182–197, 2002.

[21] T. Hiroyasu, M. Miki, and S. Watanabe, "Divided range genetic algorithms in multiobjective optimization problems," *Proc. of IWES*, vol. 99, pp. 57–65, 1999.

[22] D. Greiner, G. Winter, J. M. Emperador, and B. Galván, "Gray coding in evolutionary multicriteria optimization: application in frame structural optimum design," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 576–591.

[23] E. Zitzler, *Evolutionary algorithms for multiobjective optimization: Methods and applications*. Citeseer, 1999, vol. 63.

[24] O. M. Shir, M. Preuss, B. Naujoks, and M. Emmerich, "Enhancing decision space diversity in evolutionary multiobjective algorithms," in *Evolutionary Multi-Criterion Optimization*. Springer, 2009, pp. 95–109.

[25] P. Czyżżak and A. Jaszkiewicz, "Pareto simulated annealinga metaheuristic technique for multiple-objective combinatorial optimization," *Journal of Multi-Criteria Decision Analysis*, vol. 7, no. 1, pp. 34–47, 1998.

[26] Q. Zhang, A. Zhou, Y. Jin *et al.*, "Modelling the regularity in estimation of distribution algorithm for continuous multi-objective evolutionary optimization with variable linkages," *IEEE Transactions on Evolutionary Computation*, vol. 5, pp. 1–40, 2006.

[27] L. T. Bui, S. Wesolkowski, A. Bender, H. A. Abbass, and M. Barlow, "A dominance-based stability measure for multi-objective evolutionary algorithms," in *Evolutionary Computation, 2009. CEC'09. IEEE Congress on*. IEEE, 2009, pp. 749–756.

[28] A. Farhang-Mehr and S. Azarm, "Diversity assessment of pareto optimal solution sets: an entropy approach," in *Computational Intelligence, Proceedings of the World on Congress on*, vol. 1. IEEE, 2002, pp. 723–728.

[29] C. A. C. Coello and N. C. Cortés, "Solving multiobjective optimization problems using an artificial immune system," *Genetic Programming and Evolvable Machines*, vol. 6, no. 2, pp. 163–190, 2005.

[30] C. K. Goh and K. C. Tan, "An investigation on noisy environments in evolutionary multiobjective optimization," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 3, pp. 354–381, 2007.

[31] O. Roudenko and M. Schoenauer, "A steady performance stopping criterion for pareto-based evolutionary algorithms," 2004.

[32] H. Kaji and H. Kita, "Individual evaluation scheduling for experiment-based evolutionary multi-objective optimization," in *Evolutionary Multi-Criterion Optimization*. Springer, 2007, pp. 645–659.

[33] K. P. Yoon and C.-L. Hwang, *Multiple attribute decision making: an introduction*. Sage Publications, 1995, vol. 104.

[34] L. Martí, J. García, A. Berlanga, and J. M. Molina, "A cumulative evidential stopping criterion for multiobjective optimization evolutionary algorithms," in *Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation*. ACM, 2007, pp. 2835–2842.

[35] J. R. Schott, "Fault tolerant design using single and multicriteria genetic algorithm optimization." DTIC Document, Tech. Rep., 1995.

[36] A. Berry and P. Vamplew, "The combative accretion model–multiobjective optimisation without explicit pareto ranking," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 77–91.

[37] M. P. Hansen and A. Jaszkiewicz, *Evaluating the quality of approximations to the non-dominated set*. IMM, Department of Mathematical Modelling, Technical Universityof Denmark, 1998.

[38] J. D. Knowles and D. W. Corne, "Approximating the nondominated front using the pareto archived evolution strategy," *Evolutionary computation*, vol. 8, no. 2, pp. 149–172, 2000.

[39] H. Ishibuchi and T. Murata, "A multi-objective genetic local search algorithm and its application to flowshop scheduling," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 28, no. 3, pp. 392–403, 1998.

[40] C. A. Brizuela and E. Gutiérrez, "Multi-objective go with the winners algorithm: a preliminary study," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 206–220.

[41] H. Ishibuchi and Y. Shibata, "Mating scheme for controlling the diversity-convergence balance for multiobjective optimization," in *Genetic and Evolutionary Computation–GECCO 2004*. Springer, 2004, pp. 1259–1271.

[42] M. Li and J. Zheng, "Spread assessment for evolutionary multi-objective optimization," in *Evolutionary Multi-Criterion Optimization*. Springer, 2009, pp. 216–230.

[43] A. Zhou, Y. Jin, Q. Zhang, B. Sendhoff, and E. Tsang, "Combining model-based and genetics-based offspring generation for multi-objective optimization using a convergence criterion," in *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*. IEEE, 2006, pp. 892–899.

[44] M. R. Sierra and C. A. C. Coello, "Improving pso-based multi-objective optimization using crowding, mutation and-dominance," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 505–519.

[45] H. Ishibuchi, N. Tsukamoto, and Y. Nojima, "Evolutionary many-objective optimization: A short review." in *IEEE congress on evolutionary computation*. Citeseer, 2008, pp. 2419–2426.

[46] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[47] M. T. Emmerich and A. H. Deutz, "Test problems based on lamé superspheres," in *Evolutionary Multi-Criterion Optimization*. Springer, 2007, pp. 922–936.

[48] H. Eskandari, C. D. Geiger, and G. B. Lamont, "Fastpga: A dynamic population sizing approach for solving expensive multiobjective optimization problems," in *Evolutionary Multi-Criterion Optimization*. Springer, 2007, pp. 141–155.

[49] P. A. Bosman and D. Thierens, "The naive $\{\backslash$ mathbb M$\}$ id $\{\backslash$ mathbb E$\}$ a: A baseline multi–objective ea," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 428–442.

[50] E. Zitzler and L. Thiele, "Multiobjective optimization using evolutionary algorithmsa comparative case study," in *Parallel problem solving from naturePPSN V*. Springer, 1998, pp. 292–301.

[51] A. Riise, "Comparing genetic algorithms and tabu search for multi-objective optimization," in *Abstract conference proceedings*, 2002, p. 29.

[52] J. E. Fieldsend, R. M. Everson, and S. Singh, "Using unconstrained elite archives for multiobjective optimization," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 3, pp. 305–323, 2003.

[53] J. B. Kollat and P. M. Reed, "The value of online adaptive search: a performance comparison of nsgaii, ε-nsgaii and εmoea," in *Evolutionary Multi-Criterion Optimization*. Springer, 2005, pp. 386–398.

[54] A. Gaspar-Cunha, C. H. Antunes, and C. C. Coello, *Evolutionary Multi-Criterion Optimization: 8th International Conference, EMO 2015, Guimarães, Portugal, March 29–April 1, 2015. Proceedings*. Springer, 2015, vol. 9019.

[55] T. Wagner, H. Trautmann, and L. Martí, "A taxonomy of online stopping criteria for multi-objective evolutionary algorithms," in *Evolutionary Multi-Criterion Optimization*. Springer, 2011, pp. 16–30.

[56] E. Zitzler, D. Brockhoff, and L. Thiele, "The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration," in *Evolutionary multi-criterion optimization*. Springer, 2007, pp. 862–876.

[57] Q. Zhang, A. Zhou, and Y. Jin, "Rm-meda: A regularity model-based multiobjective estimation of distribution algorithm," *Evolutionary Computation, IEEE Transactions on*, vol. 12, no. 1, pp. 41–63, 2008.

[58] D. Brockhoff, T. Wagner, and H. Trautmann, "On the properties of the r2 indicator," in *Proceedings of the 14th annual conference on Genetic and evolutionary computation*. ACM, 2012, pp. 465–472.

[59] J. Knowles and D. Corne, "On metrics for comparing nondominated sets," in *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, vol. 1. IEEE, 2002, pp. 711–716.

[60] T. M. Chan, "Klee's measure problem made easy," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 410–419.

[61] N. Beume, C. M. Fonseca, M. López-Ibáñez, L. Paquete, and J. Vahrenhold, "On the complexity of computing the hypervolume indicator," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 5, pp. 1075–1082, 2009.

[62] C. M. Fonseca, L. Paquete, and M. López-Ibánez, "An improved dimension-sweep algorithm for the hypervolume indicator," in *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*. IEEE, 2006, pp. 1157–1163.

[63] H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima, "Difficulties in specifying reference points to calculate the inverted generational distance for many-objective optimization problems," in *Computational Intelligence in Multi-Criteria Decision-Making (MCDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 170–177.