A Correlation Analysis of Set Quality Indicator Values in Multiobjective Optimization

Arnaud Liefooghe
Univ. Lille, CNRS, UMR 9189 – CRIStAL, France
Dolphin, Inria Lille – Nord Europe, France
arnaud.liefooghe@univ-lille1.fr

Bilel Derbel
Univ. Lille, CNRS, UMR 9189 – CRIStAL, France
Dolphin, Inria Lille – Nord Europe, France
bilel.derbel@univ-lille1.fr

ABSTRACT

A large spectrum of quality indicators has been proposed so far to assess the performance of discrete Pareto set approximations in multiobjective optimization. Such indicators assign, to any solution set, a real-value reflecting a given aspect of approximation quality. This is an important issue in multiobjective optimization, not only to compare the performance and assets of different approximate algorithms, but also to improve their internal selection mechanisms. In this paper, we adopt a statistical analysis to experimentally investigate by how much a selection of state-of-the-art quality indicators agree with each other for a wide range of Pareto set approximations from well-known two- and three-objective continuous benchmark functions. More particularly, we measure the correlation between the ranking of low-, medium-, and high-quality limited-size approximation sets with respect to inverted generational distance, additive epsilon, multiplicative epsilon, R2, R3, as well as hypervolume indicator values. Since no pair of indicators obtains the same ranking of approximation sets, we confirm that they emphasize different facets of approximation quality. More importantly, our statistical analysis allows the degree of compliance between these indicators to be quantified.

1. INTRODUCTION

Set quality indicators have been initially proposed in the late 1990s, and are still refined nowadays, in order to compare the output of approximate multiobjective optimization algorithms. By defining a total order between Pareto set approximations, they are particularly relevant when the partial order induced by the Pareto dominance relation is not sufficiently qualified to discriminate between different approximation sets. However, given their different background, structural properties and focus in terms of quality, it is with no surprise that the order obtained with respect to different set quality indicators are sometimes contradictory. For instance, it is often the case that the approximation set obtained by an Algorithm A is pictured to be better than the one obtained by an Algorithm B for some indicator, while the opposite is true for another indicator; see e.g. [14]. In addition, set quality indicators can also be seen as a support for multicriteria decision making, in the sense that they allow to provide the decision maker with a representative subset of a potentially very large set of trade-offs for presenting a compact and reliable "picture" of the Pareto front for the problem at hand. In this regards, any indicator actually makes some assumptions about the decision maker preferences [26]. More recently, those quality indicators have been plugged onto the design principles of evolutionary and other approximate multiobjective optimization algorithms; see e.g. [2, 4, 6]. This class of indicator-based approaches seeks an approximation set of a given or bounded cardinality that maximizes or minimizes the indicator value, thus explicitly formalizing the goal of the search process [3, 20, 28].

The properties of state-of-the-art quality indicators have been studied in terms of computational complexity, parameter dependency, scaling invariance, and monotonicity with respect to set dominance relations [16, 26, 29]. In particular, the proportion of non-compliant decisions made by quality indicators with respect to dominance relations has also been experimentally investigated in [16], and the absolute difference in indicator values was investigated in [12]. However, the relation between any two quality indicators is far from being well understood. Actually, we usually do not know precisely what are the differences in terms of quality or in terms of interpretation each indicator is able to provide. Intuitively, this also depends on many factors such as the shape of the Pareto front, the distribution of non-dominated vectors in the objective space, or some user-defined parameters. For instance, the hypervolume is known to be largely affected by the choice of the reference point [1, 15], particularly in the lexicographically optimal regions of the Pareto front. As well, the hypervolume is believed to favor convex regions over concave regions [27], and to give more focus on knee points [1, 4]. Similarly, the distribution of solutions from an approximation set optimizing the epsilon indicator clearly depends on the shape of the Pareto front [5].

For all these reasons, it might be interesting to measure the agreements and disagreements those quality indicators have in assessing one approximation set better than another, depending on the problem characteristics, and given a large-picture of approximation set quality. In this paper, we propose to adopt a statistical analysis in order to experimentally investigate by how much (unary) quality indicators agree with each other on the induced ranking of approximation sets. Indeed, we do not aim at highlighting the difference between quality indicators on some particular examples, but rather to *quantify* this difference, i.e. by *how much* do they vary and *not* why they do so. Our analysis extends results from [12, 23] by systematically analyzing the nonparametric rank correlations between a selection of quality indicators, and by contrasting their association across a large spectrum of approximation quality and problem classes. More particularly, we are interested in the inverted generational distance [7], the additive and multiplicative versions of the

GECCO '16, July 20 - 24, 2016, Denver, CO, USA

ISBN 978-1-4503-4206-3/16/07...$15.00

DOI: http://dx.doi.org/10.1145/2908812.2908906

epsilon indicator [29], the R2 and R3 indicators from the R-metric family [10], and the hypervolume [27]. We believe that this selection provides a representative sample of popular indicators from today's literature, most of them being monotonic with respect to the conventional Pareto dominance relation [26]. However, in the past, many researchers were still using indicators capable of contradicting the Pareto dominance relation; see e.g. [16]. For this reason, notice that a statistical analysis as conducted in the paper is likely to result in much lower correlations for such non-monotonic indicators. Based on this selection, we then compute the indicator values of samples of possible low-, medium- and high-quality approximation sets over a representative subset of multiobjective optimization problems, particularly in terms of the Pareto front shape. For this, we rely on the well-known multiobjective continuous functions from the CEC 2009 special session and competition on the performance assessment of multiobjective optimization algorithms [25]. Based on this sample of approximation sets, we measure the obtained value for each indicator and each approximation set from our sample, and we experimentally investigate the correlation between indicator values. This allows us to quantify the degree of compliance between any pair of quality indicators, and to highlight their differences depending on the problem characteristics and on the properties of approximation sets. This analysis gives a first step towards a better understanding of the relations between set quality indicators, and might provide important implications in terms of performance assessment, algorithm design and decision making in multiobjective optimization.

The remainder of the paper is organized as follows. In Section 2, we recall some definitions related to multiobjective optimization and we describe the quality indicators under consideration in our study. In Section 3, we present the setup of the experiments. In Section 4, we provide a throughout correlation analysis on the CEC 2009 benchmark functions. Finally, we conclude the paper and discusses further research in the last section.

2. BACKGROUND

In this section, we introduce the necessary definitions and provide a selection of conventional quality indicators from the multiobjective optimization literature.

2.1 Multiobjective Optimization

Let us assume that we are given an arbitrary multiobjective optimization problem (X, f), where X is the *solution space*, and $f = (f_1, \ldots, f_i, \ldots, f_d)$ is an objective function vector such that f_i is to be minimized for all $i \in \{1, \ldots, d\}$. Let $Z = f(X)$ be the *objective space*, $Z \subseteq \mathbb{R}^d$. Each solution $x \in X$ is associated with an objective vector $z \in Z$ such that $z = f(x)$. An objective vector $z \in Z$ is *dominated* by an objective vector $z' \in Z$ ($z \prec z'$) iff $\forall i \in \{1, \ldots, d\} : z'_i \leqslant z_i$ and $\exists i \in \{1, \ldots, d\}$ such that $z'_i < z_i$. Two objective vectors $z, z' \in Z$ are *mutually non-dominated* iff $z \not\prec z'$ and $z' \not\prec z$. An objective vector $z^\star \in Z$ is *Pareto optimal* or *non-dominated* iff there does not exists a $z \in Z$ such that $z^\star \prec z$. Analog definitions can be formalized for solutions $x \in X$ by using the associated objective vectors $z \in Z$, such as $z = f(x)$. The *Pareto front* $Z^\star \subseteq Z$ is the set of non-dominated objective vectors; the *Pareto set* $X^\star \subseteq X$ is a set of solutions that maps to the Pareto front, i.e. $f(X^\star) = Z^\star$. One of the most challenging issue in multiobjective optimization is to identify the Pareto set/front, or a good approximation of it for complex problems. More particularly, EMO and other approximate algorithms aim to identify an approximation set of limited cardinality, ideally a subset of the exact Pareto set/front, that is to be presented to the decision maker for further consideration. For the sake of clarity, we will focus on Pareto front approximations in the following sections. This can be easily extended by considering the mapping of a Pareto set approximation to the objective space.

2.2 Quality Indicators

A (unary) *quality indicator* is a function $2^Z \rightarrow \mathbb{R}$ that assigns each approximation set to a (scalar) value reflecting its quality [26]. In the following, we select and introduce a subset of conventional quality indicators from the multiobjective literature. The reader is referred to [14, 16, 26, 29] for a broader review. Let $A \subseteq Z$ be a set of mutually non-dominated objective vector (i.e. a Pareto front approximation, or approximation set), and $R \subseteq Z$ be a reference set (ideally the exact Pareto front when it is discrete, i.e. $R = Z^\star$). In the following, we assume that there does not exist any vector in A that dominates a vector in R; i.e. $\forall r \in R$, $\nexists a \in A$ such that $r \prec a$. In other words, the reference set R weakly dominates any approximation set A [29].

IGD: The inverted generational distance [7] is an inverted version of the generational distance [22], see also [21] for a detailed explanation. It gives the average distance between any point from the reference set R and its closest point from the approximation set A.

$$\text{IGD}(A) := \frac{1}{|R|}\sqrt{\sum_{r \in R} \min_{a \in A} ||a - r||_2^2}$$

The euclidean distance (L2-norm) in the objective space is usually used for distance calculation. Obviously, the smaller the IGD value, the closer the approximation set from the reference set. An indicator value of 0 actually implies $R \subseteq A$.

EPS: The epsilon indicator family [29] gives the minimum factor by which the approximation set has to be translated in the objective space in order to (weakly) dominate the reference set. The *additive* epsilon indicator ($\text{EPS}_{(+)}$) is based on an additive factor.

$$\text{EPS}_{(+)}(A) := \max_{r \in R} \min_{a \in A} \max_{i \in \{1,\ldots,d\}} (a_i - r_i)$$

The *multiplicative* version ($\text{EPS}_{(\times)}$) is based on a multiplicative factor, and assumes that all objective function values are strictly positives.

$$\text{EPS}_{(\times)}(A) := \max_{r \in R} \min_{a \in A} \max_{i \in \{1,\ldots,d\}} (a_i / r_i)$$

Both epsilon indicator versions are to be minimized; and $\text{EPS}_{(+)}(A) = 0$ or $\text{EPS}_{(\times)}(A) = 1$ implies that $R \subseteq A$.

R: The family of R-metrics [10] are based on a set of utility functions. A utility function $u : Z \rightarrow \mathbb{R}$ maps an objective vector to a scalar value based on specified parameters. A typical example is the weighted Chebyshev scalarizing function defined below.

$$u_\lambda(z) = \max_{i \in \{1,\ldots,d\}} \lambda_i \cdot \left| z_i^\star - z_i \right|$$

where $z \in Z$ is a candidate objective vector, $z^\star \in \mathbb{R}^d$ is the ideal point (i.e. $z_i^\star = \min_{z \in Z} z_i$, $i \in \{1, \ldots, d\}$) and $\lambda \in \mathbb{R}^d$ is a weighting coefficient vector. By defining a set of uniformly-defined weighting coefficient vectors Λ such that for all $\lambda \in \Lambda$, $\lambda = (\lambda_1, \ldots, \lambda_i, \ldots, \lambda_d)$, $\lambda_i \geq 0$ and $\sum_{i=1}^d \lambda_i = 1$, the R2 and R3 indicators can be defined as follows.

$$\text{R2}(A) := \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \Big(\min_{r \in R} u_\lambda(r) - \min_{a \in A} u_\lambda(a) \Big)$$

$$\text{R3}(A) := \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \frac{\min_{r \in R} u_\lambda(r) - \min_{a \in A} u_\lambda(a)}{\min_{r \in R} u_\lambda(r)}$$

Once again, both R2 and R3 indicators are to be minimized; and $\text{R2}(A) = 0$ or $\text{R3}(A) = 0$ implies $R \subseteq A$.

RHV: The hypervolume [27, 29] gives the multidimensional volume of the portion of the objective space that is weakly dominated by an approximation set.

$$\text{HV}(A) := \int_{z^{min}}^{z^{\max}} \alpha_A(z)dz$$

such that:

$$\alpha_A(z) := \begin{cases} 1 & \text{if } \exists a \in A \text{ such that } z \prec a \\ 0 & \text{otherwise} \end{cases}$$

In practice, only the upper-bound vector $z^{\max} \in \mathbb{R}^d$ is required to compute the hypervolume; this parameter is called *reference point*. In the following, we will be interested in the *relative* hypervolume indicator (RHV), that is the relative deviation of the approximation set's hypervolume to the reference set's hypervolume.

$$\text{RHV}(A) := \frac{\text{HV}(R) - \text{HV}(A)}{\text{HV}(R)}$$

This allows us to consider minimizing indicator values as well, such that $\text{RHV}(A) = 0$ means that $R \subseteq A$.

2.3 Properties

In this section, we summarize a number of properties from [16, 26, 29] that describe the quality indicators presented above.

Monotonicity: An indicator is monotonic with respect to the weak Pareto dominance relation (Pareto-compliant in [16, 29]) if for any approximation set that dominates another approximation set, its indicator value is better; i.e. a monotonic indicator does not disagree with the (partial) order induced by the dominance relation [26]. All the indicators presented in the previous section are monotonic, with the exception of IGD, despite its regular use as an absolute performance metric. A strict version of monotonicity can also be defined by considering the standard Pareto dominance relation and a strict inequality between indicator values. The hypervolume is the only known indicator that satisfies the strict monotonicity property [29]. Notice that an empirical analysis of the degree of monotonicity for some non-monotonic indicators are reported in [11, 16].

Scaling invariance: An indicator is scaling invariant if the order of approximation sets induced by the indicator values remains the same when applying a monotonic transformation of the objective function values. However, as the indicators under consideration all explicitly exploit the objective function values, none of them actually satisfies this scaling invariance property, except of course if the hypervolume reference point is transformed accordingly.

Parameters and problem knowledge: In our definitions of the quality indicators, a reference set R is always required, although hypervolume could be used without any reference set. In addition, the definition of R2 and R3 is based on the ideal point and on a user-given number of weighting coefficient vectors, while the definition of RHV is based on a reference point that must be specified by the practitioner. Actually, the ordering of the approximation sets induced by the hypervolume is known to by affected be the setting of this reference point [26].

Computational complexity: Since an in-depth experimental analysis may require the comparison of a large number of approximation sets, and given that an indicator can potentially be integrated into the search process of an approximate algorithm, the computational resources required to compute an indicator value constitute and important feature of the indicator characteristics. Obviously, the computational complexity for IGD, EPS and the R-metrics is polynomial in the objective space dimension, the approximation set cardinality and the reference set cardinality (as well as the number of weighting coefficient vectors for R2 and R3), whereas it is exponential in the number of objectives for the hypervolume [26].

3. EXPERIMENTAL SETTING

In this section, we shall describe the benchmark functions, the approximation set samples, the parameter setting, and the correlation measure of our experimental analysis. All the experiments have been conducted in `R` [19], using the `ggplot2` [24], `emoa` [17], and `mco` [18] packages.

3.1 CEC 2009 Benchmark Functions

In order to analyze the indicator values of approximation sets and their correlation, we consider nine multiobjective continuous functions from the CEC 2009 special session and competition on the performance assessment of constrained and bound-constrained multiobjective optimization algorithms [25]. This set of benchmark functions has been specifically designed to resemble complicated real-life optimization problems. They present different properties in terms of dimension, separability, multi-modality, and shape of the Pareto front. More particularly, we consider all the unconstrained (bound-constrained) functions UF01–10, with the exception of UF05 which contains a very limited number of points belonging to the Pareto front. The first six problems consist of two-objective functions, whereas the last three problems consist of three-objective functions, all to be minimized. The Pareto front from UF01, UF02 and UF03 is convex, the one from UF04, UF08 and UF10 is concave, and the one from UF06, UF07 and UF09 is a line or plane. In addition, there are gaps in the Pareto front of UF06 and UF09. Notice that, for all problems, all objective functions roughly have the same range, and the objective function values of solutions from the Pareto set all lie in $[0, 1]$. The formulation of these test functions can be found in [25]; we consider them under their original setting.

During the CEC 2009 competition, the competing algorithms were run multiple times for a maximum number of function evaluations. For each problem instance, the average IGD indicator value of the final approximation sets was the *only* figure of merit for comparing the algorithms. Notice, however, that the IGD used in [25] is a slight variation of the original IGD [7] that we use in the current paper. This results in having different IGD-values but the ranking of approximation sets is exactly the same in both definitions; i.e. the Kendall rank correlation coefficient $\tau = 1$, see Section 3.4. In addition, the organizers provided a source code to generate a set of uniformly distributed points along the Pareto front in the objective space, which is available at the following URL: http://dces.essex.ac.uk/staff/qzhang/moeacompetition09.htm. We used it for computing a reference set R for each function in our analysis. The cardinality of this reference set is provided in Table 1 for each function.

3.2 Sampling Strategy

We consider the following strategies in order to sample a subset of all possible approximation sets for each function.

low-Q: We generate a number of $\mu = 100$ solutions at random in the solution space, i.e. following a uniform distribution within the boundary provided for each problem variable [25], from which we extract the subset of non-dominated vectors.

med-Q: We run a black-box (randomized) EMO algorithm with

Table 1: Description of the nine benchmark functions used in the experimental analysis and of the average cardinality of the sample of approximation sets.

problem	# objectives (d)	feasible f_i-range	Pareto front f_i-range	Pareto front structure		reference set size	avg. approximation set size low-Q	med-Q	high-Q
UF01	2	$[0, 07]$	$[0, 1]$	convex	no gap	1 000	8.43	100.00	100.00
UF02	2	$[0, 05]$	$[0, 1]$	convex	no gap	1 000	11.55	100.00	100.00
UF03	2	$[0, 10]$	$[0, 1]$	convex	no gap	1 000	6.92	99.95	100.00
UF04	2	$[0, 02]$	$[0, 1]$	concave	no gap	1 000	46.97	99.93	100.00
UF06	2	$[0, 25]$	$[0, 1]$	line	gaps	1 000	6.51	83.72	100.00
UF07	2	$[0, 07]$	$[0, 1]$	line	no gap	1 000	7.31	100.00	100.00
UF08	3	$[0, 25]$	$[0, 1]$	concave	no gap	2 025	18.6	100.00	100.00
UF09	3	$[0, 25]$	$[0, 1]$	plane	gaps	2 025	17.87	100.00	100.00
UF10	3	$[0, 99]$	$[0, 1]$	concave	no gap	2 025	17.54	99.91	100.00

a population size $\mu = 100$, and consider the subset of mutually non-dominated approximate solutions identified by the algorithm as an approximation set. In our experiments, NSGA-II [8] is performed for 1 000 generations, using the SBX crossover operator with a rate 0.7 and a polynomial mutation with a rate 0.2.

high-Q: We sample uniformly at random a subset of $\mu = 100$ solutions from the reference set. This means that the obtained approximation set does not contain any dominated solutions, but actually contains ten times less elements than within the reference set for two-objective problems, and approximately twenty times less elements for three-objective problems, respectively.

Each sampling strategy is repeated 1 000 times for each multiobjective problem under consideration. The average cardinality of the obtained approximation sets is reported in Table 1. In Section 4, we analyze the correlation between the indicator values obtained by these approximation set samples.

3.3 Parameter Setting

As reported in Table 1, each approximation set contains at most $\mu = 100$ solutions. For each function, we consider a fixed reference set of 1 000 solutions for $d = 2$ and 2 025 for $d = 3$. Notice that, for all problems, the objective function values of all solutions lie in $[0, f^{\max}]$. In order to avoid any issue in the computation of the indicators, in particular for $\text{EPS}_{(\times)}$, we simply shift the objective function values in the hyper-box $[1, f^{\max} + 1]^d$ without modifying the shape of the Pareto front. The ideal point $z^\star \in \mathbb{R}^d$ is then defined such that $z^\star_i = 1$ for all $i \in \{1, \ldots, d\}$.

For computing the R-metrics, we generate a set of $|\Lambda| = 100$ uniformly-defined weighting coefficient vectors, and we use the ideal point $z^\star$ as a reference point. At last, we analyze the impact of the reference point $z^{\max}$ on the hypervolume indicator with two different settings: (i) $z^{\max}_i = f^{\max}$, and (ii) $z^{\max}_i = 1.1 \times f^{\text{worst}}$ for all $i \in \{1, \ldots, d\}$, such that $f^{\max}$ is the maximum objective function value for the problem under consideration, and f^{worst} is the worst objective function value found for a given problem and a given sampling strategy. In other words, the first hypervolume setting with $f^{\max}$, denoted by RHV (max), corresponds to the anti-ideal point and only depends on the problem under consideration. On the contrary, the second hypervolume setting with f^{worst}, denoted by RHV (worst), is more tight and depends not only on the problem, but also on the sampling strategy. For high-quality approximation sets, the hypervolume reference point based on the setting with f^{worst} actually corresponds to the nadir point, shifted by a factor of 1.1 in order to ensure that it is strictly dominated by any objective vector from the approximation sets.

3.4 Measuring Correlation

In order to measure the association between the indicator values obtained by a given sample of approximation sets, we consider the Kendall rank correlation coefficient τ [13], which is a rank-based nonlinear correlation coefficient measure. Indeed, we do not provide a more conventional Pearson correlation coefficient, which gives the *linear* relationship between the indicator values. Instead, we focus on the *ranking* of approximation sets obtained within each indicator, i.e. by how much do the indicators rank the approximation sets similarly. In other words, we are not interested in the correlation between the values obtained by each indicator, but rather on the underlying ranking they obtain within the sample of approximation sets. This is also the reason why we decided not to perform a (multiple) linear regression analysis.

Let us consider two arbitrary indicators I_1 and I_2 to be minimized, and a pair (A_1, A_2) of approximation sets from our sample. The pair is said to be *concordant* if $I_1(A_1) > I_1(A_2) \wedge I_2(A_1) > I_2(A_2)$, or if $I_1(A_1) < I_1(A_2) \wedge I_2(A_1) < I_2(A_2)$. On the contrary, the pair is said to be *discordant* if $I_1(A_1) > I_1(A_2) \wedge I_2(A_1) < I_2(A_2)$, or if $I_1(A_1) < I_1(A_2) \wedge I_2(A_1) > I_2(A_2)$. If $I_1(A_1) = I_1(A_2)$ or $I_2(A_1) = I_2(A_2)$, the pair is neither concordant nor discordant. The Kendall coefficient τ quantifies the difference between the proportion of concordant and discordant pairs among all possible pairwise approximation sets. It is defined as follows:

$$\tau = \frac{(\%\text{ concordant pairs}) - (\%\text{ discordant pairs})}{\%\text{ pairs}}$$

The coefficient τ ranges in $[-1, 1]$, from perfect disagreement ($\tau = -1$), to perfect agreement ($\tau = 1$). When τ is approximately zero, the indicator values are said to be independent.

4. CORRELATION ANALYSIS

Descriptive statistics (average and standard deviation) on the obtained indicator values are provided in Table 2. Although those indicator values are not directly comparable since their distribution obviously depends on the indicator and the test case (benchmark function and sampling strategy) under consideration, we can still notice that the average value and the standard deviation typically decrease with the sampling quality. This means that the better the sample of approximation sets, the smaller the expected indicator value, and the smaller the variation around the mean. The only notable exception is for the hypervolume with a tight reference point RHV (worst), because the setting of the reference point is different for each sampling strategy in this case. The table also reports the number of tied indicator values over all pairs of approximation

Table 2: Average indicator value (avg) and standard deviation (sd) for each benchmark function, approximation quality level (low-Q, med-Q, high-Q) and set quality indicator. For each setting, the proportional number of occurrences where a pair of approximation sets obtained the same indicator value is also reported (t). All values are rounded to 10^{-2}, t is given as a percent.

		IGD ($\times 10$)		**EPS** ($+$)		**EPS** ($\times$)		**R2**		**R3**		**RHV (max)**		**RHV (worst)**	
		avg $_{(sd)}$	t	avg $_{(sd)}$	t	avg $_{(sd)}$	t	avg $_{(sd)}$	t	avg $_{(sd)}$	t	avg $_{(sd)}$	t	avg $_{(sd)}$	t
UF01	low-Q	0.41 $_{(0.04)}$	0	1.19 $_{(0.09)}$	0	1.97 $_{(0.08)}$	0	0.57 $_{(0.05)}$	0	0.65 $_{(0.06)}$	0	0.23 $_{(0.02)}$	0	0.36 $_{(0.03)}$	0
	med-Q	0.03 $_{(0.01)}$	0	0.17 $_{(0.04)}$	0	1.14 $_{(0.04)}$	0	0.03 $_{(0.01)}$	0	0.03 $_{(0.01)}$	0	0.02 $_{(0.01)}$	0	0.05 $_{(0.02)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	2	1.02 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0
UF02	low-Q	0.20 $_{(0.02)}$	0	0.61 $_{(0.04)}$	0	1.52 $_{(0.05)}$	0	0.28 $_{(0.02)}$	0	0.32 $_{(0.03)}$	0	0.18 $_{(0.02)}$	0	0.28 $_{(0.02)}$	0
	med-Q	0.02 $_{(0.01)}$	0	0.10 $_{(0.03)}$	0	1.09 $_{(0.03)}$	0	0.01 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0	0.01 $_{(0.01)}$	0	0.03 $_{(0.01)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	2	1.02 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0
UF03	low-Q	0.35 $_{(0.02)}$	0	1.13 $_{(0.08)}$	0	2.09 $_{(0.09)}$	0	0.62 $_{(0.04)}$	0	0.70 $_{(0.04)}$	0	0.18 $_{(0.01)}$	0	0.35 $_{(0.02)}$	0
	med-Q	0.07 $_{(0.02)}$	0	0.33 $_{(0.08)}$	0	1.33 $_{(0.08)}$	0	0.09 $_{(0.03)}$	0	0.10 $_{(0.03)}$	0	0.05 $_{(0.01)}$	0	0.24 $_{(0.05)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	1	1.02 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0
UF04	low-Q	0.07 $_{(0.00)}$	0	0.22 $_{(0.01)}$	0	1.20 $_{(0.02)}$	0	0.08 $_{(0.00)}$	0	0.10 $_{(0.00)}$	0	0.21 $_{(0.01)}$	0	0.34 $_{(0.01)}$	0
	med-Q	0.02 $_{(0.00)}$	0	0.08 $_{(0.01)}$	0	1.05 $_{(0.01)}$	0	0.01 $_{(0.00)}$	0	0.02 $_{(0.00)}$	0	0.03 $_{(0.01)}$	0	0.08 $_{(0.01)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	1	1.02 $_{(0.01)}$	1	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.01)}$	0	0.02 $_{(0.01)}$	0
UF06	low-Q	1.74 $_{(0.20)}$	0	4.60 $_{(0.44)}$	0	4.69 $_{(0.31)}$	0	2.40 $_{(0.22)}$	0	2.90 $_{(0.26)}$	0	0.25 $_{(0.02)}$	0	0.39 $_{(0.03)}$	0
	med-Q	0.10 $_{(0.04)}$	0	0.42 $_{(0.16)}$	0	1.41 $_{(0.16)}$	0	0.11 $_{(0.06)}$	0	0.13 $_{(0.06)}$	0	0.02 $_{(0.01)}$	0	0.19 $_{(0.07)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.02 $_{(0.01)}$	2	1.01 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0
UF07	low-Q	0.42 $_{(0.04)}$	0	1.36 $_{(0.12)}$	0	2.30 $_{(0.15)}$	0	0.67 $_{(0.06)}$	0	0.79 $_{(0.07)}$	0	0.27 $_{(0.02)}$	0	0.41 $_{(0.03)}$	0
	med-Q	0.06 $_{(0.06)}$	0	0.31 $_{(0.25)}$	0	1.28 $_{(0.27)}$	0	0.07 $_{(0.08)}$	0	0.08 $_{(0.09)}$	0	0.04 $_{(0.04)}$	0	0.11 $_{(0.12)}$	0
	high-Q	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	4	1.02 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0
UF08	low-Q	0.67 $_{(0.07)}$	0	2.47 $_{(0.20)}$	0	3.34 $_{(0.19)}$	0	0.85 $_{(0.06)}$	0	0.92 $_{(0.07)}$	0	0.15 $_{(0.02)}$	0	0.28 $_{(0.02)}$	0
	med-Q	0.06 $_{(0.00)}$	0	0.69 $_{(0.09)}$	0	1.69 $_{(0.10)}$	0	0.06 $_{(0.01)}$	0	0.06 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.04 $_{(0.01)}$	0
	high-Q	0.02 $_{(0.00)}$	0	0.18 $_{(0.05)}$	5	1.17 $_{(0.06)}$	6	0.01 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.09 $_{(0.01)}$	0
UF09	low-Q	0.68 $_{(0.07)}$	0	2.41 $_{(0.22)}$	0	3.30 $_{(0.20)}$	0	0.84 $_{(0.06)}$	0	0.90 $_{(0.06)}$	0	0.14 $_{(0.01)}$	0	0.24 $_{(0.02)}$	0
	med-Q	0.05 $_{(0.02)}$	0	0.41 $_{(0.11)}$	0	1.41 $_{(0.11)}$	0	0.05 $_{(0.01)}$	0	0.05 $_{(0.01)}$	0	0.00 $_{(0.00)}$	0	0.03 $_{(0.01)}$	0
	high-Q	0.01 $_{(0.00)}$	0	0.10 $_{(0.02)}$	9	1.08 $_{(0.02)}$	5	0.01 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.04 $_{(0.01)}$	0
UF10	low-Q	3.38 $_{(0.32)}$	0	10.55 $_{(0.93)}$	0	10.81 $_{(0.73)}$	0	4.06 $_{(0.26)}$	0	4.45 $_{(0.28)}$	0	0.17 $_{(0.01)}$	0	0.28 $_{(0.02)}$	0
	med-Q	0.08 $_{(0.02)}$	0	0.85 $_{(0.11)}$	0	1.85 $_{(0.11)}$	0	0.12 $_{(0.06)}$	0	0.13 $_{(0.06)}$	0	0.00 $_{(0.00)}$	0	0.04 $_{(0.04)}$	0
	high-Q	0.02 $_{(0.00)}$	0	0.17 $_{(0.05)}$	5	1.16 $_{(0.06)}$	6	0.01 $_{(0.00)}$	0	0.01 $_{(0.00)}$	0	0.00 $_{(0.00)}$	0	0.09 $_{(0.01)}$	0

sets for each test case, i.e. the proportional number of occurrences where a pair of approximation sets is neither concordant nor discordant in the computation of the Kendall rank correlation coefficient (τ). For all indicators but the epsilon indicator variants, we actually never observe any tied indicator values. For the epsilon indicators, the number of pairs that are neither concordant nor discordant are always below 10%.

Let us now analyze the correlation between the indicator values obtained by the sample of approximation sets for the different problem functions. It is obviously not possible to report the scatter plot for each test case and each pair of indicators due to the large amount of data. Instead, Figure 1 reports the Kendall rank correlation coefficient between all pairs of set quality indicators for each benchmark function and each sampling strategy. We decided to split the correlation values into different groups, from a very high negative correlation ($\tau < -0.75$) to a very high positive correlation ($\tau > 0.75$), as well as an additional group containing test cases which were reported to be non-significant by the Kendall coefficient test with a p-value of 0.05. The figure provides the correlation between any pair of indicators (on the $x-$ and $y-$axes) for each problem function (from top to bottom) and each sampling strategy (from left to right). The higher the correlation degree, the higher the agreement between the two corresponding indicators, the darker the corresponding area on the heat-map.

Overall, the indicators under consideration are never in conflict one against another, as there is never a significant amount of negative correlation. Indeed, for all the test cases, the τ value is actually always larger than 0.07 when it is significant. However, there does not exist any two indicators that fully agree with each other on any of the problem functions ($\tau < 1.00$). The few test cases where the τ value is larger than 0.98 actually correspond to indicators from the same family. This confirms that the performance of multiobjective optimizers cannot be assessed properly by a single set quality indicator, and that each performance metric actually measures a different facet of approximation quality. We analyze those correlations in detail, and more importantly, we quantify them for each indicator below.

IGD: Let us start with the inverted generational distance. For low-quality approximation sets, the correlation degree between IGD and any other indicator is quite low ($\tau < 0.5$, expect for $\text{EPS}_{(+)}$ where $\tau < 0.75$). For medium-quality approximation sets, this correlation seems to get higher, but τ remains below 0.75 for all the instances but two-objective problems with a linear Pareto front (UF06 and UF07). It is also worth noticing the strong effect of concavity on the correlation of IGD with other indicators for medium-quality approximation sets. Indeed, for concave problems, those correlation values drop substantially. For high-quality approximation sets, IGD is actually slightly correlated with RHV (worst) for two-dimensional convex Pareto front ($\tau > 0.5$), but not for other problems ($\tau < 0.5$). This means that one could be a reasonable estimator of the other on those cases. This trends is roughly the same for all other indicators but $\text{EPS}_{(+)}$, which is moderately correlated with IGD for all two-objective problems, but not as much for three-objective problems.

Overall, the IGD indicator is fairly correlated with $\text{EPS}_{(+)}$ and

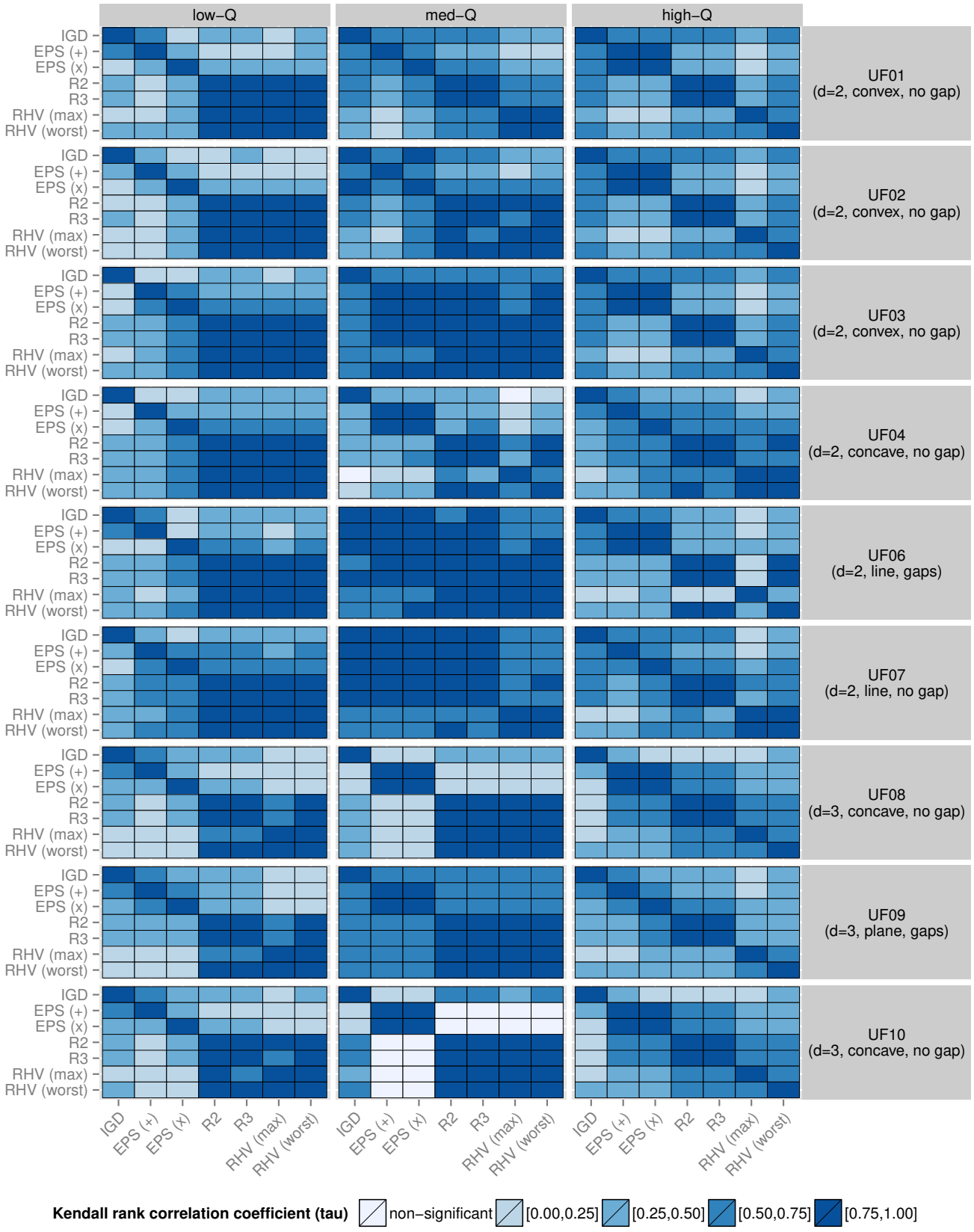

Figure 1: Heat-map Kendall rank correlation τ for each pair of set quality indicators (displayed on both axes), each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–UF10).

RHV, especially when the later is based on the (slightly shifted) worst-found objective function values for the sampling strategy under consideration. On the contrary, the correlation is very low for $\text{EPS}_{(\times)}$ and RHV (max). The correlation with the remaining indicators is lower for low-quality approximation sets than for medium- and high-quality approximation sets. Let us remind that IGD is the only indicator considered in our analysis which is *not* (weakly) monotonic with respect to the Pareto dominance relation. This means that IGD agrees more with monotonic indicators for good approximation sets than for bad ones. As a consequence, IGD might actually be an acceptable measure for algorithm performance assessment. Notice that some experiments on the number of non-compliant measures made by IGD with respect to Pareto dominance have been recently reported in [11].

As a side remark, the results of the CEC 2009 competition, which were based on IGD only, might actually be different if another indicator was used to assess the performance of the competing algorithms. It would be worth revisiting those results with a set of complementary quality indicators. Indeed, the competition winner, and more importantly the understandings we have from the competing algorithms, might change while using another, or several others, indicator(s) to assess the quality of the identified approximation sets.

EPS: Unsurprisingly, $\text{EPS}_{(+)}$ and $\text{EPS}_{(\times)}$ are highly correlated with each other for medium- and high-quality approximation sets (τ is always larger than 0.5, and most of time than 0.75). However, they are only slightly correlated with each other for low-quality approximation sets, as with any other indicator. With respect to the remaining indicators, there is a low correlation between the EPS indicators and R2, R3 or RHV ($\tau < 0.75$), except for medium-quality approximation sets with a linear or planar Pareto front (UF06, UF07, UF09). EPS is also moderately correlated with IGD, as already mentioned above. Let us remind that $\text{EPS}_{(+)}$ and $\text{EPS}_{(\times)}$ are the only indicators for which we observed tied indicator values. This might explain the tendency to have smaller Kendall rank correlation coefficient τ compared with other indicators.

R: The R-metrics globally show higher correlation degrees. As expected, R2 and R3 are highly correlated with each other for all functions and all types of approximation set samples. The τ value is actually larger than 0.91 on all the test cases. As mentioned before, the R-metrics are only moderately correlated with IGD and EPS. In fact, the correlation degree seems to be particularly small for low-quality approximation sets and for medium-quality approximation sets with a concave Pareto front (UF04, UF08, UF10). At last, the correlation between the R-metrics and RHV is particularly high for low- and medium-quality approximation sets for all problem functions (τ is always higher than 0.5, except for UF04 and medium-quality approximation sets where it is below). However, for high-quality approximation sets, this correlation degree drops substantially, even if the correlation with RHV (worst) remains significant for some of the problem instances, with two objectives and a Pareto front which is not convex. But overall, and compliant with the results reported in [23], the correlation between R2 or R3 and RHV is is in average the highest we obtained for a pair of indicators belonging to two different families.

RHV: Finally, we focus on the correlation coefficients for RHV. Both settings of RHV, with a tight and a wide reference point, denoted as RHV (worst) and RHV (max) respectively, are highly correlated with each other for low- and medium-quality approximation sets on all functions ($\tau > 0.75$, except for UF04). This correlation largely decreases for high-quality approximation sets, particularly for UF06 whose Pareto front is discontinuous. As also pointed out in [1, 15], this means that the hypervolume indicator might rank high-quality approximation sets quite differently depending on the position of the reference point, in our case either as the (shifted) nadir point or at the maximum objective function vector. In fact, additional experiments, not reported here due to space restriction, reveal that the correlation between RHV indicator values with different settings of the reference point is always very high for low- and a medium-quality approximation sets (except, once again, for UF04, i.e. the only instance with a two-dimensional concave Pareto front), while the setting appears to be more sensitive for high-quality approximation sets, particularly when there are gaps on the Pareto front (UF06 and UF09).

In addition, as reported above, RHV is slightly to moderately correlated with IGD and EPS, whereas it is significantly correlated with R2 and R3 for low- and medium-quality approximation sets, but not as much for high-quality approximation sets. This would actually suggest that the R2 or R3 indicator could potentially be used to approximate the hypervolume indicator at the early stages of an indicator-based search process, as does the R2-based multiobjective search algorithm from [6], while computationally expensive hypervolume calculations would be dedicated to the latest refinements, when the approximation set gets closer to the Pareto front. Of course, this would need to be carefully calibrated depending on data like the number of objectives, the number of points in the reference set and in the approximation set, but also on practitioner parameters like the number of weighting coefficient vectors, which all polynomially increase the complexity of R2 and R3, whereas the hypervolume complexity increases exponentially with the objective space dimension. It would then be worth measuring the impact of the number of weighting coefficient vectors on the ability of the R-metrics to estimate the ranking induced by the hypervolume.

5. CONCLUSIONS

In this paper, we experimentally investigated the degree of correlation between the order induced by different set quality indicators. Our analysis highlights important insights for the performance assessment, the interpretation of preferences, and the design of algorithms in multiobjective optimization. First, our findings clearly confirms the well-known fact that there does not exist a single set quality indicator which is able to capture all the aspects of approximation quality, even if none of them are clearly in conflict with another. Second, the correlation of the epsilon indicator with the other indicators from our analysis is overall very low. This means that this indicator actually focus on complementary aspects with respect to other indicators, but also that it does not allow to capture all the facets of approximation quality. The same reasoning applies for the inverted generational distance. For this reason, we plan to revisit the data from the CEC 2009 competition, where the inverted generational distance was the single performance measure under consideration, in order to enhance our knowledge and understandings of the competing algorithms by means of supplementary indicators. Moreover, the hypervolume shows a high correlation with the R-metrics for completely random solution sets to better approximations identified by some evolutionary algorithm. As a consequence, it would be worth investigating more thoroughly the estimation of the computationally prohibitive hypervolume with the R2 or R3 indicator, as it might for instance enable to speed up the selection process of an indicator-based approach using the hypervolume, such as SMS-EMOA [4] or HypE [2]. At last, we plan to extend our analysis with (i) additional indicators such as the averaged Hausdorff distance [21], (ii) additional sampling strategies potentially mapping to populations maintained by EMO algorithms while the search process evolves by taking inspiration from [9], and (iii) additional problem classes, in particular with respect to the

number of objectives. This will hopefully allow us to increase our knowledge on the relations between set quality indicators in multi-objective optimization, as well as the underlying mechanisms that explain their differences.

Acknowledgements. The authors would like to acknowledge Fabio Daolio and Joshua Knowles for fruitful discussions related to the results presented in the paper. This work was partially supported by the Japanese-French JSPS-Inria project "Threefold Scalability in Any-objective Black-Box Optimization" (2015-2017).

6. REFERENCES

[1] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications. *Theoretical Computer Science*, 425:75–103, 2012.

[2] J. Bader and E. Zitzler. HypE: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation*, 19(1):45–76, 2011.

[3] M. Basseur, A. Goëffon, A. Liefooghe, and S. Verel. On set-based local search for multiobjective combinatorial optimization. In *Conference on Genetic and Evolutionary Computation (GECCO 2013)*, pages 471–478. ACM, 2013.

[4] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, 2007.

[5] K. Bringmann, T. Friedrich, and P. Klitzke. Efficient computation of two-dimensional solution sets maximizing the epsilon-indicator. In *Congress on Evolutionary Computation (CEC 2015)*, pages 970–977, Sendai, Japan, 2015. IEEE Press.

[6] D. Brockhoff, T. Wagner, and H. Trautmann. R2 indicator-based multiobjective search. *Evolutionary Computation*, 23(3):369–395, 2015.

[7] C. A. Coello Coello and N. C. Cortés. Solving multiobjective optimization problems using an artificial immune system. *Genetic Programming and Evolvable Machines*, 6(2):163–190, 2005.

[8] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[9] J. Derrac, S. García, S. Hui, P. N. Suganthan, and F. Herrera. Analyzing convergence performance of evolutionary algorithms: A statistical approach. *Information Sciences*, 289:41–58, 2014.

[10] M. P. Hansen and A. Jaszkiewicz. Evaluating the quality of approximations of the non-dominated set. Technical Report IMM-REP-1998-7, Institute of Mathematical Modeling, Technical University of Denmark, 1998.

[11] H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima. Modified distance calculation in generational distance and inverted generational distance. In *Evolutionary Multi-Criterion Optimization (EMO 2015)*, volume 9019 of *Lecture Notes in Computer Science*, pages 110–125, Guimarães, Portugal, 2015. Springer.

[12] S. Jiang, Y.-S. Ong, J. Zhang, and L. Feng. Consistencies and contradictions of performance metrics in multiobjective optimization. *IEEE Transactions on Cybernetics*, 44(12):2391–2404, 2014.

[13] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[14] J. Knowles and D. Corne. On metrics for comparing non-dominated sets. In *Congress on Evolutionary Computation (CEC 2002)*, pages 711–716, Piscataway, NJ, USA, 2002. IEEE Press.

[15] J. Knowles and D. Corne. Properties of an adaptive archiving algorithm for storing nondominated vectors. *IEEE Transactions on Evolutionary Computation*, 7(2):100–116, 2003.

[16] J. Knowles, L. Thiele, and E. Zitzler. A tutorial on the performance assessment of stochastic multiobjective optimizers. TIK Report 214, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland, 2006. (revised version).

[17] O. Mersmann. *emoa: Evolutionary Multiobjective Optimization Algorithms*, 2012. R package version 0.5-0.

[18] O. Mersmann. *mco: Multiple Criteria Optimization Algorithms and Related Functions*, 2014. R package version 1.0-15.1.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[20] G. Rudolph, O. Schütze, C. Grimme, C. Domínguez-Medina, and H. Trautmann. Optimal averaged Hausdorff archives for bi-objective problems: theoretical and numerical results. *Computational Optimization and Applications*, (to appear).

[21] O. Schutze, X. Esquivel, A. Lara, and C. A. Coello Coello. Using the averaged Hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 16(4):504–522, 2012.

[22] D. A. V. Veldhuizen and G. B. Lamont. Evolutionary computation and convergence to a Pareto front. In *Genetic Programming (GP 1998)*, pages 221–228, 1998.

[23] S. Wessing and B. Naujoks. Sequential parameter optimization for multi-objective problems. In *Congress on Evolutionary Computation (CEC 2010)*, pages 1–8, Barcelona, Spain, 2010.

[24] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer, New York, USA, 2009.

[25] Q. Zhang, A. Zhou, S. Zhao, P. N. Suganthan, W. Liu, and S. Tiwari. Multiobjective optimization test instances for the CEC 2009 special session and competition. Working Report CES-887, School of Computer Science and Electrical Engineering, University of Essex, 2008.

[26] E. Zitzler, J. Knowles, and L. Thiele. Quality assessment of Pareto set approximations. In *Multiobjective Optimization – Interactive and Evolutionary Approaches*, volume 5252 of *Lecture Notes in Computer Science*, chapter 14, pages 373–404. Springer, 2008.

[27] E. Zitzler and L. Thiele. Multiobjective optimization using evolutionary algorithms — a comparative case study. In *Parallel Problem Solving from Nature (PPSN V)*, volume 1498 of *Lecture Notes in Computer Science*, pages 292–301. Springer, Amsterdam, The Netherlands, 1998.

[28] E. Zitzler, L. Thiele, and J. Bader. On set-based multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 14(1):58–79, 2010.

[29] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. Grunert da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003.