A Correlation Analysis of Set Quality Indicator Values in Multiobjective Optimization

Arnaud Liefooghe

A Correlation Analysis of Set Quality Indicator Values in Multiobjective Optimization

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

ABSTRACT
A large spectrum of quality indicators have been proposed so far to assess the performance of discrete Pareto set approximations in multiobjective optimization. Such indicators assign a real-value to any approximation set that reflects a given aspect of its 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. However, identifying fine-grained theoretical properties between different classes of indicators is generally out of reach due to the high complexity of the approximation set structures. In this paper, we adopt a statistical analysis to experimentally investigate by how much a subset 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 benchmark continuous test 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, epsilon, R-metric and hypervolume indicator values. Since none of them obtains the exact same ranking of approximation sets, we show that they actually emphasize different facets of approximation quality. Moreover, our statistical analysis allows us to quantify the degree of compliance between these quality indicators.

KEY WORDS: multiobjective optimization, set quality indicators, performance assessment, correlation analysis, inverted generational distance, epsilon indicator, R-metrics, hypervolume.

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 dominance relations are 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. Knowles and Corne (2002). In addition, they 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, and also given that any indicator actually makes some assumptions about the deci-

sion maker preferences (Zitzler et al., 2008). More recently, those quality indicators have been plugged onto the design principles of evolutionary and other approximate multiobjective optimization algorithms; see e.g. Zitzler and Künzli (2004); Beume et al. (2007); Bader and Zitzler (2011). This class of indicator-based approaches seeks an approximation set of a given or bounded cardinality that maximizes or minimizes the indicator value, then explicitly formalizing the goal of the search process (Zitzler et al., 2010; Basseur et al., 2013).

The properties of state-of-the-art quality indicators have been studied by Zitzler et al. (2003, 2008) and Knowles et al. (2006) in terms of computational complexity, parameter dependency, scaling invariance, and monotonicity with respect to dominance relations between approximation sets. The proportion of mistakes made by quality indicators in terms of dominance relations has also been experimentally investigated by Knowles et al. (2006). 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 (Knowles and Corne, 2003; Auger et al., 2012), particularly in the lexicographically optimal regions of the Pareto front. As well, the hypervolume is believed to favor convex regions over concave regions (Zitzler and Thiele, 1998), and to give more focus on knee points (Beume et al., 2007). Similarly, the distribution of solutions from an approximation set optimizing the epsilon indicator clearly depends on the shape of the Pareto front (Bringmann et al., 2015).

For all these reasons, it might be interesting to quantify the agreements and disagreements those quality indicators have by 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. More particularly, we are interested in the inverted generational distance (Coello Coello and Cortés, 2005), the additive and multiplication versions of the epsilon indicator (Zitzler et al., 2003), the R2 and R3 indicators from the R-metric family (Hansen and Jaszkiewicz, 1998), and the hypervolume (Zitzler and Thiele, 1998). We compute the indicator value for a sample of possible low-, medium- and high-quality approximation sets over a representative subset of multiobjective optimization problems, particularly in terms of the shape of the Pareto front. 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 (Zhang et al., 2008). 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

This section introduces the necessary definitions and provides a subset 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 $\nexists z \in Z$ 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 that $z = f(x)$. 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 (Deb, 2001; Coello Coello et al., 2007; Branke et al., 2008). 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 in 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 (Zitzler et al., 2008). In the following, we introduce a subset of conventional quality indicators from the multiobjective literature. The reader is referred to Knowles and Corne (2002); Knowles et al. (2006) or Zitzler et al. (2003, 2008) 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 (Zitzler et al., 2003).

IGD: The inverted generational distance (Coello Coello and Cortés, 2005) is an inverted version of the generational distance (Veldhuizen and Lamont, 1998). It gives the average distance between any point from the reference set R and its closest point from the approximation set A.

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

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 $A = R$.

EPS: The epsilon indicator family (Zitzler et al., 2003) 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.

$$\text{EPS}_{(+)}(A) := \max_{r \in R} \min_{a \in A} ||a_i - r_i||_\infty$$

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

$$\text{EPS}_{(\times)}(A) := \max_{r \in R} \min_{a \in A} ||a_i / r_i||_\infty$$

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

R: The family of R-metrics (Hansen and Jaszkiewicz, 1998) 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 $\lambda = (\lambda_1,\ldots,\lambda_i,\ldots,\lambda_d) \in \Lambda$, $\lambda_i \geq 0$ and $\sum_{i=1}^d \lambda_i = 1$, the R2 and R3 indicators can be defined as follows.

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

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

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

RHV: The hypervolume (Zitzler and Thiele, 1999) 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 $A = R$.

2.3 Properties

In this section, we summarize a number of properties from Knowles et al. (2006) and Zitzler et al. (2008) that describe the quality indicators presented above.

Monotonicity: An indicator is monotonic with respect to the weak Pareto dominance relation (Pareto-compliant in Knowles et al. (2006)) 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 (Zitzler et al., 2008). 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 (Zitzler et al., 2007). Notice that an empirical analysis of the degree of monotonicity for some non-monotonic indicators are reported by Knowles et al. (2006).

Scaling invariance: An indicator is scaling invariant if the order of approximation sets induced by the indicator values remain 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.

Parameters and problem knowledge: In our definitions of quality indicators, a reference set R is always required. In addition, the definition of R2 and R3 is based on the ideal point and on a user-given number 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 be affected be the setting of this reference point (Zitzler et al., 2008).

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 is also an 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 and the reference set cardinalities (and the number of weighting coefficient vectors for R2 and R3), whereas it is exponential in the number of objectives for the hypervolume; see e.g. While (2005) or Chan (2013).

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` (R Core Team, 2013), using the `ggplot2` (Wickham, 2009), `emoa` (Mersmann, 2012), and `mco` (Mersmann, 2014) packages.

3.1 CEC 2009 Benchmark Functions

In order to analyse 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 (Zhang et al., 2008). 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, multimodality, 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 in 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 on 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 Zhang et al. (2008); 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* merit of figure for comparing the algorithms. In addition, the organizers provided a source code to generate a set of uniformly distributed points along the Pareto front in the objective space, 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. Notice that there exists some restriction on the values the size of the reference set can take, which explains the difference in terms of cardinality for two- and three-objective problems.

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 (Zhang et al., 2008), from which we extract the subset of non-dominated vectors.

med-Q: We run a black-box (randomized) EMO algorithm with 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, we perform NSGA-II (Deb et al., 2002) for $1\,000$ generations, using the SBX crossover operator with a rate 0.7 and a polynomial mutation with a rate 0.2.

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

function	# objectives (d)	f_i-values	Pareto front structure		reference set size	avg. approximation set size low-Q	med-Q	high-Q
UF01	2	$[0,\ 7]$	convex	no gap	1 000	8.43	100.00	100.00
UF02	2	$[0,\ 5]$	convex	no gap	1 000	11.55	100.00	100.00
UF03	2	$[0, 10]$	convex	no gap	1 000	6.92	99.95	100.00
UF04	2	$[0,\ 2]$	concave	no gap	1 000	46.97	99.93	100.00
UF06	2	$[0, 25]$	line	gaps	1 000	6.51	83.72	100.00
UF07	2	$[0,\ 7]$	line	no gap	1 000	7.31	100.00	100.00
UF08	3	$[0, 25]$	concave	no gap	961	18.6	100.00	100.00
UF09	3	$[0, 25]$	plane	gaps	961	17.87	100.00	100.00
UF10	3	$[0, 99]$	concave	no gap	961	17.54	99.91	100.00

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 around ten times less elements than within the reference set.

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 analyse the correlation between the indicator values obtained by these samples of approximation sets.

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 961 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 $|\Lambda| = 100$ 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 $z^{\max}$ for 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.

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 (Kendall, 1938), 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.

More particularly, 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 independent.

4 Correlation Analysis

In this section, we analyze the correlation between the indicator values obtained by the sample of approximation sets for the different problem functions. Figures 1 — 4 report the Kendall rank correlation coefficient between all pairs of set quality indicators for each benchmark function and each sampling strategy. A given figure provides the correlation between a particular indicator (written on top) and each other indicator (corresponding to colored curves), for each problem function (on the x-axis) and each sampling strategy (low-Q, med-Q, high-Q, from left to right). The higher the correlation degree, the higher the agreement between the two corresponding indicators.

Overall, the indicators under consideration are never in conflict one against another, as there is always some positive amount of correlation ($\tau > 0$), even if it is sometimes insignificant. However, we clearly see that there does not exists any two indicators that fully agree with each other on any of the problem function. This highlights that the performance of multiobjective optimizers cannot be analyzed properly within a single set quality indicator, and that each performance metric actually measures a different facet of approximation quality. We analyze those correlations in details for each indicator below.

IGD: Let us start with the inverted generational distance (IGD) in Figure 1. For low-quality approximation sets, the correlation degree between IGD and any other indicator is quite low ($\tau < 0.7$). For medium-quality approximation sets, this correlation gets higher, but τ is always below 0.75, except for two-objective problems with a linear Pareto front (UF06 and UF07), and for all indicators but RHV. For high-quality approximation sets, IGD is actually slightly correlated with RHV with a tight reference point (denoted as RHV(worst)) for two-dimensional convex Pareto front ($\tau \approx 0.7$), but not for other problems ($\tau \approx 0.3$). 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 to 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 RHV, 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 the RHV setting based on the absolute maximum objective function values for the problem at hand. 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* monotonic with respect to the (weak) 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 mistakes made by IGD with respect to Pareto dominance have been recently reported by Ishibuchi et al. (2015).

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

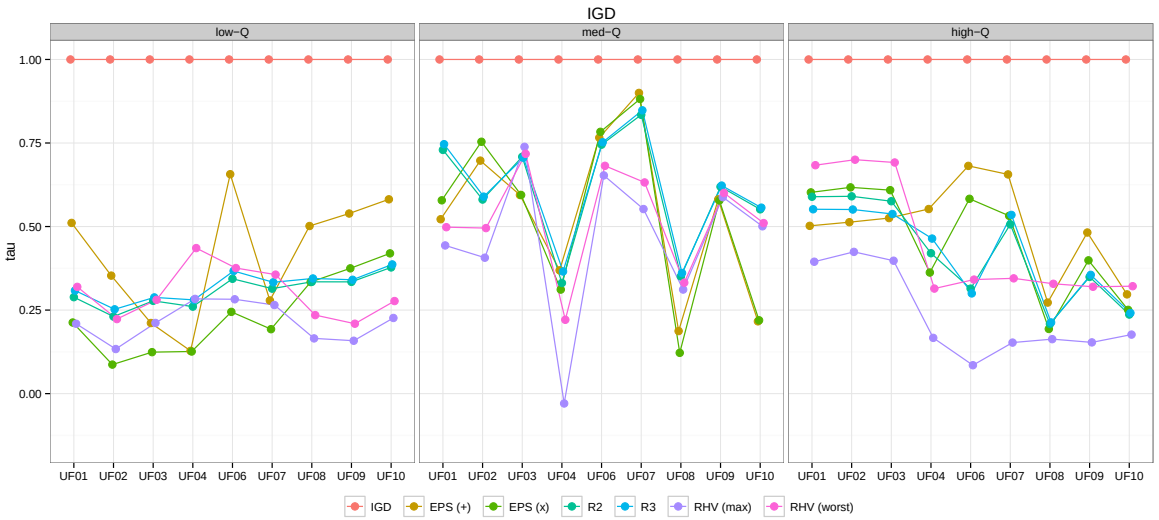

Figure 1: Kendall rank correlation coefficient τ between IGD and any other quality indicator for each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10).

change while using another, or several others, indicator(s) to assess the quality of the identified approximation sets.

EPS: The results for $\text{EPS}_{(+)}$ and $\text{EPS}_{(\times)}$ are reported in Figure 2. Unsurprisingly, those two indicators are highly correlated with each other for medium- and high-quality approximation sets. 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.

R: The R-metrics globally show higher correlation degrees, as reported in Figure 3. As expected, R2 and R3 are highly correlated with each other for all functions and all types of approximation set samples ($\tau > 0.9$). As mentioned before, the R-metrics are only moderately correlated with IGD and EPS. In fact, the correlation seems to be particularly low 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 0.65, except for UF04 and medium-quality approximation sets where it is around 0.5). However, for high-quality approximation sets, this correlation degree drops substantially, even if the correlation with the RHV setting with a tight reference point remains significant for some of the problem instances, with two objectives and a Pareto front which is not convex. But overall, 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, Figure 4 reports the correlation coefficients for RHV. Both settings of RHV, with a tight and a wide reference point, are highly correlated with each other for low- and medium-quality approx-

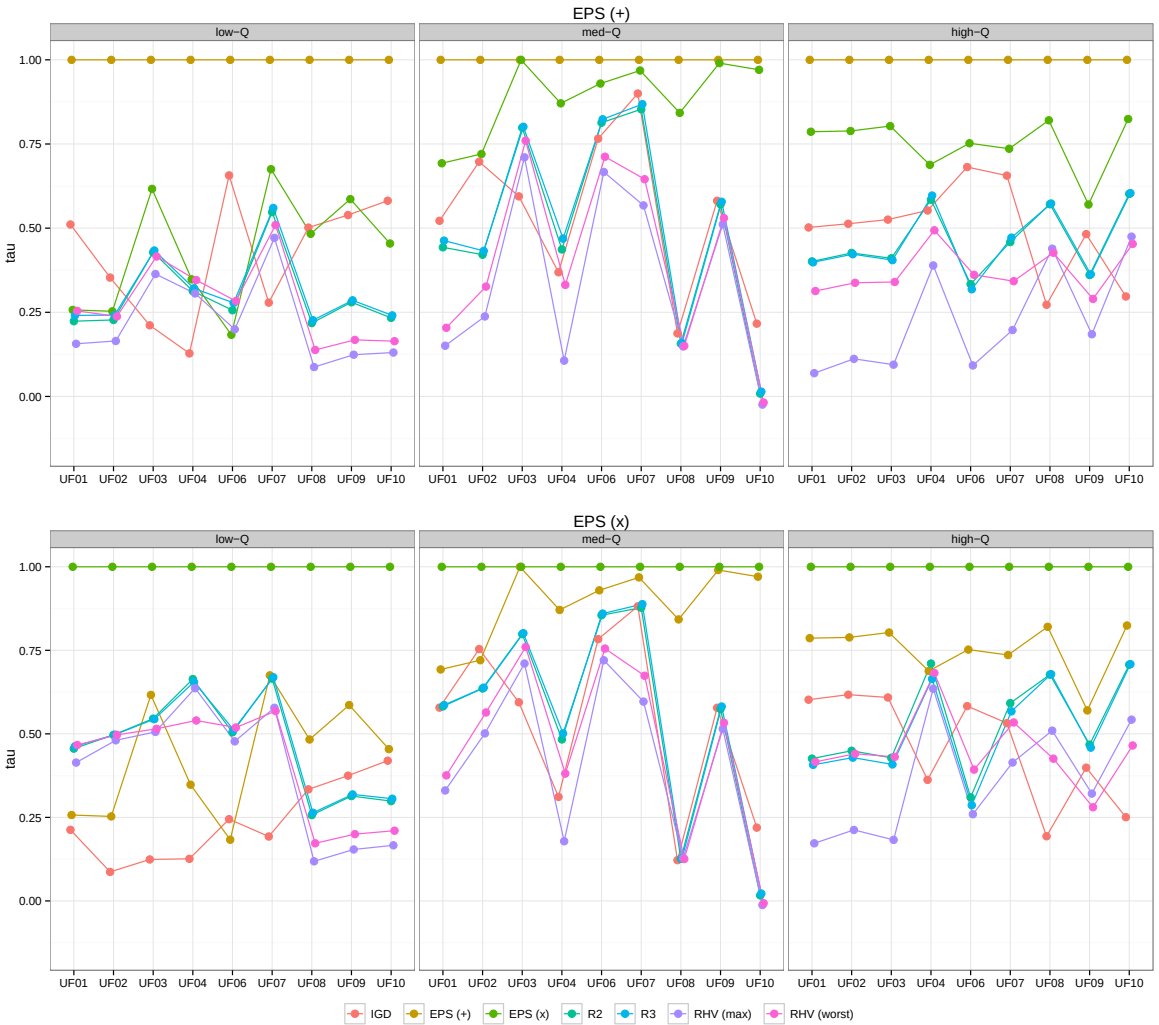

Figure 2: Kendall rank correlation coefficient τ between EPS (+) (top), EPS (x) (bottom) and any other quality indicator for each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10).

imation sets on all functions ($\tau > 0.8$, except for UF04). This correlation largely decreases for high-quality approximation sets, particularly for UF06 and UF09, whose Pareto front is discontinuous. As also pointed out by Knowles and Corne (2003) or Auger et al. (2012), 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 provided in Figure 5 reveal that the correlation between RHV indicator values with different settings of the reference point is always very high ($\tau > 0.7$) for low- and a medium-quality approxima-

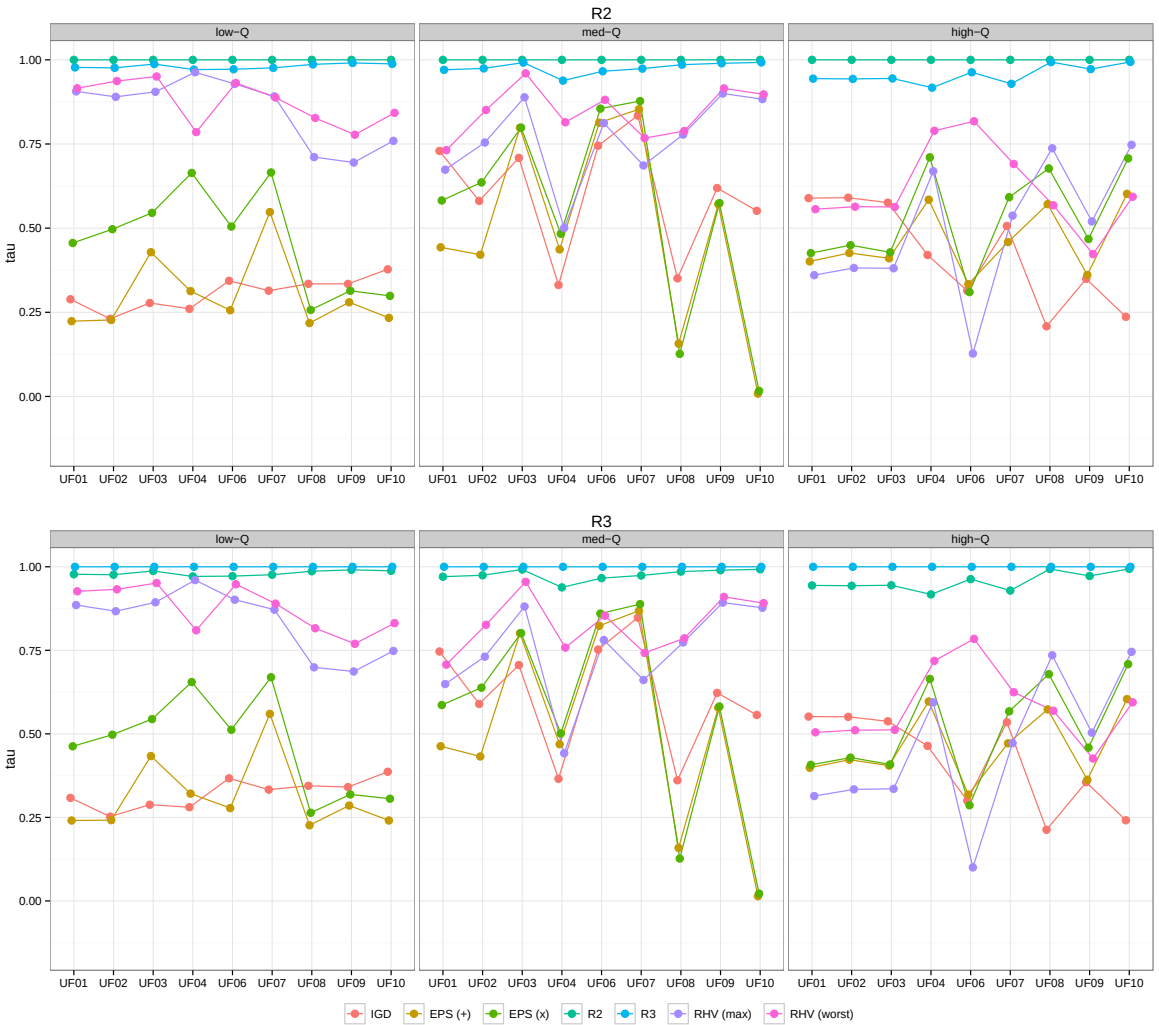

Figure 3: Kendall rank correlation coefficient τ between R2 (top), R3 (bottom) and any other quality indicator for each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10).

tion 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, which is relatively cheap to compute, could potentially be used to approximate the hypervolume indicator at the early stages of an indicator-based search process, while computa-

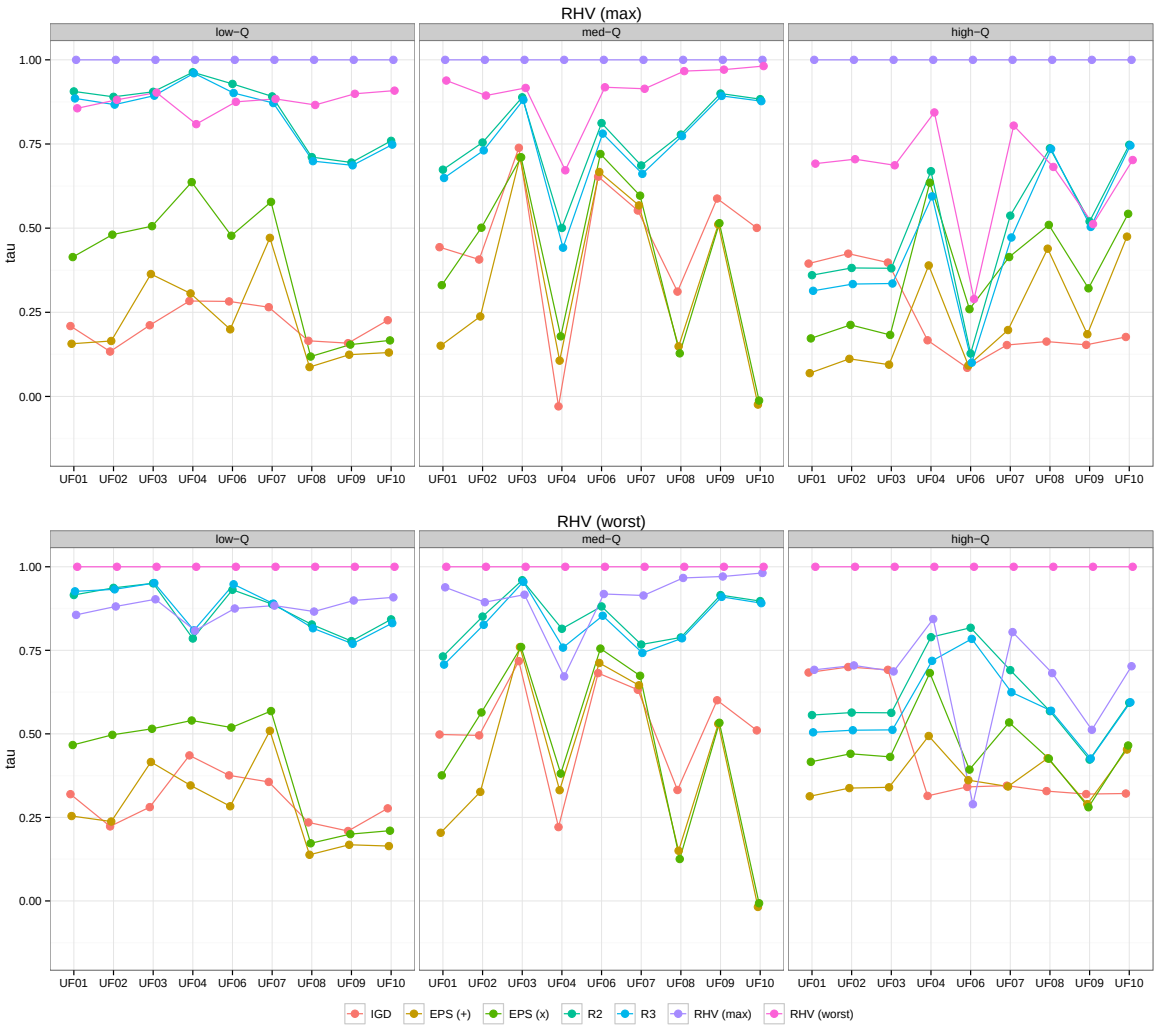

Figure 4: Kendall rank correlation coefficient τ between RHV using the problem's maximum objective function vector as a reference point (top), RHV using the sampling and problem's worst-seen objective function vector shifted by a factor 1.1 as a reference point (bottom), and any other quality indicator for each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10).

tionally expensive hypervolume calculations would be dedicated to the latest refinements, when the approximation set gets closer to the Pareto front.

Impact of scaling: Figure 6 reports, for each indicator, the correlation between the original indicator value and the indicator value computed over a monotonic transformation of the objective function values. In our experiments, the normalized objective function

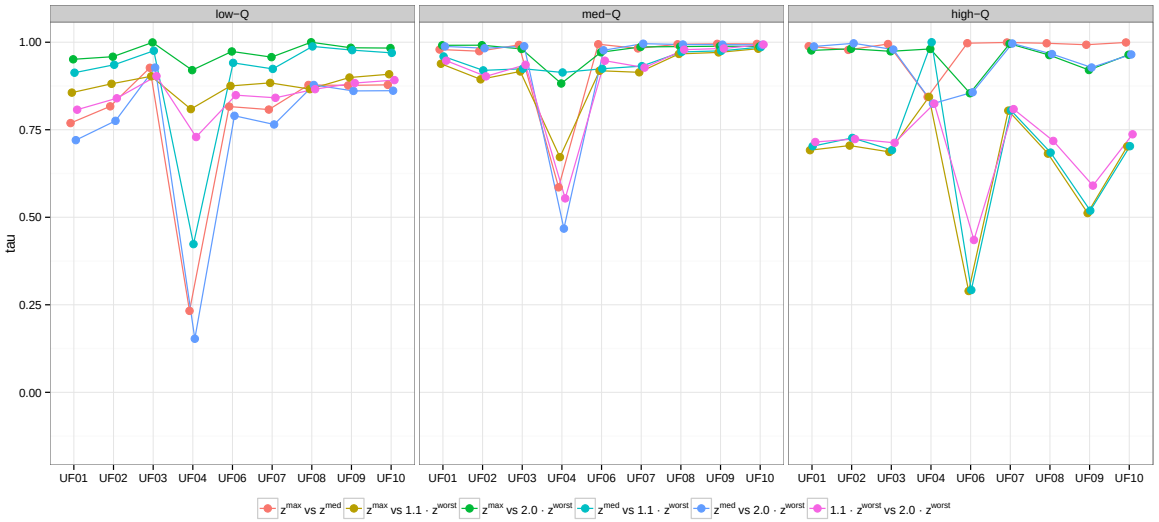

Figure 5: Kendall rank correlation coefficient τ between the RHV indicator value with different settings of the reference point for each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10), such that $z_i^{\max} = f^{\max}$, $z_i^{\text{med}} = (f^{\max} - f^{\min})$, and $z_i^{\text{worst}} = f^{\text{worst}}$ for all $i \in \{1, \ldots, d\}$, $f^{\min}$ (resp. $f^{\max}$) being the minimum (resp. maximum) objective function value for the problem under consideration, and f^{worst} being the worst objective function value found for a given problem and a given sampling strategy.

values follow a transformation of the form $f_i'(x) = 1 + (f_i(x) - f^{\min})/(f^{\max} - f^{\min})$, $i \in \{1, \ldots, d\}$, where $f^{\min}$ (resp. $f^{\max}$) is the lower (resp. upper) bound of the objective function values, such that each normalized objective vector lies in $[1, 2]^d$. As defined in Section 2.3, by obtaining the same order over the approximation sets for normalized and unnormalized objective function values, a given indicator would be scaling invariant. In such a case, the correlation coefficient would be $\tau = 1$. Despite none of the indicators under consideration satisfies this property in the general case, the degree of invariance is actually quite acceptable for most of them, as reveled by our experiments. Indeed, the correlation coefficient is always larger than 0.9, except for $\text{EPS}_{(\times)}$ whatever the approximation set quality, and for the RHV setting with a tight reference point for high-quality approximation sets. For $\text{EPS}_{(\times)}$, this might be explained by the use of a multiplicative factor over the objective function values, whereas for RHV, this might actually be an artefact of the absolute position of the reference point within this particular setting, the lexicographically optimal regions of the Pareto front having more impact in one case than in the other.

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 that there does not exist a single set quality indicator which is able to capture all the aspects of approximation quality, even if all of them are at least slightly correlated. Second, the correlation of the epsilon indicator with the other indicators from our analysis is overall very low. This means that

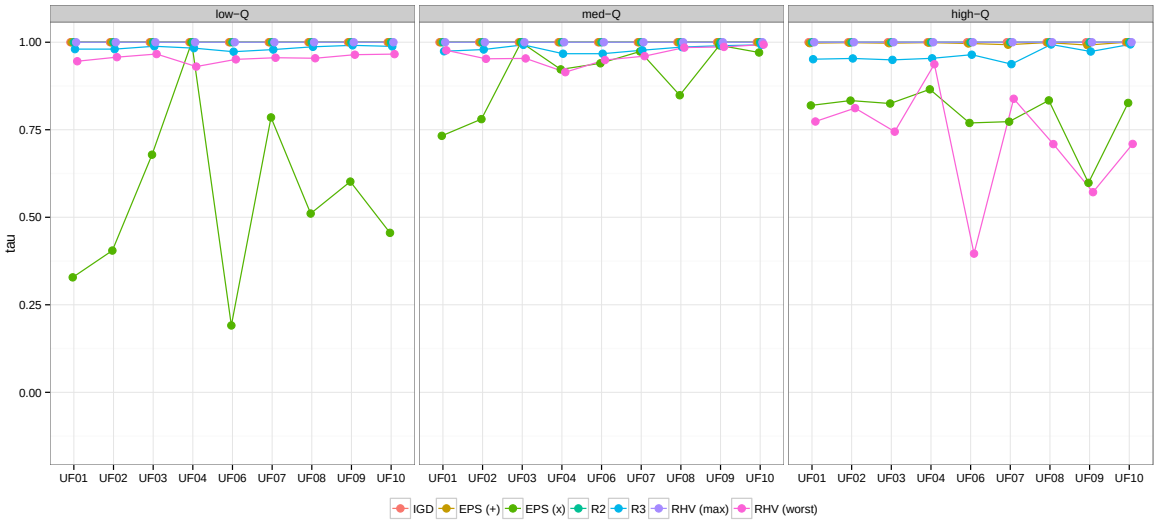

Figure 6: Kendall rank correlation coefficient τ between the indicator value over the original objective function values and the indicator value over a monotonic transformation of the objective function values for each quality indicator, each sampling strategy (low-Q, med-Q, high-Q) and each problem function (UF01–10).

this indicator actually focus on complementary aspects with respect to other indicators. The same 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 algorithms by means of supplementary indicators. Next, as already pointed out by Knowles and Corne (2002); Zitzler et al. (2008); Auger et al. (2012), our statistical analysis reveals that the hypervolume is sensitive to the setting of the reference point, especially for high-quality approximation sets, i.e. smaller subsets of the (exact) Pareto front. 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 affordable 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 (Beume et al., 2007) or HypE (Bader and Zitzler, 2011). At last, a similar analysis with additional indicators and for other classes of problem instances, in particular with respect to the number of objectives, would allow us to increase our knowledge on the relations between set quality indicators in multiobjective optimization.

Acknowledgements. The author would like to acknowledge Fabio Daolio and Joshua Knowles for fruitful discussions related to the results presented in this paper. This work was partially supported by the Japanese-French JSPS project “Global Research on the Framework of Evolutionary Solution Search to Accelerate Innovation” (2013-2016), and by the Japanese-French JSPS-Inria project “Threefold Scalability in Any-objective Black-Box Optimization” (2015-2017).

References

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

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

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

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

Branke, J., Deb, K., Miettinen, K., and Slowinski, R., editors (2008). *Multiobjective Optimization – Interactive and Evolutionary Approaches*, volume 5252 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Germany.

Bringmann, K., Friedrich, T., and Klitzke, P. (2015). Efficient computation of two-dimensional solution sets maximizing the epsilon-indicator. In *IEEE Congress on Evolutionary Computation (CEC 2015)*. IEEE Press. (to appear).

Chan, T. (2013). Klee's measure problem made easy. In *LAnnual Symposium on Foundations of Computer Science (FOCS 2013)*, pages 410–419. IEEE Press.

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

Coello Coello, C. A., Lamont, G. B., and Van Veldhuizen, D. A. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Genetic and Evolutionary Computation Series. Springer, New York, USA, second edition.

Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, Chichester, UK.

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

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

Ishibuchi, H., Masuda, H., Tanigaki, Y., and Nojima, Y. (2015). 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. Springer.

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

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

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

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

Mersmann, O. (2012). *emoa: Evolutionary Multiobjective Optimization Algorithms*. R package version 0.5-0.

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

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

While, L. (2005). A new analysis of the LebMeasure algorithm for calculating hypervolume. In *Evolutionary Multi-Criterion Optimization (EMO 2005)*, volume 3410 of *Lecture Notes in Computer Science*, pages 326–340. Springer.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer, New York, USA.

Zhang, Q., Zhou, A., Zhao, S., Suganthan, P. N., Liu, W., and Tiwari, S. (2008). 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.

Zitzler, E., Brockhoff, D., and Thiele, L. (2007). The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization (EMO 2007)*, volume 4403 of *Lecture Notes in Computer Science*, pages 862–876. Springer.

Zitzler, E., Knowles, J., and Thiele, L. (2008). Quality assessment of Pareto set approximations. In Branke et al. (2008), chapter 14, pages 373–404.

Zitzler, E. and Künzli, S. (2004). Indicator-based selection in multiobjective search. In *Conference on Parallel Problem Solving from Nature (PPSN VIII)*, volume 3242 of *Lecture Notes in Computer Science*, pages 832–842, Birmingham, UK. Springer-Verlag.

Zitzler, E. and Thiele, L. (1998). 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.

Zitzler, E. and Thiele, L. (1999). Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271.

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

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