Chapter

On model evaluation, indices of importance, and interaction values in rough set analysis[⋆] [⋆⋆]

Günther Gediga[1] and Ivo Düntsch[2]

[1] Institut für Evaluation und Marktanalysen, Brinkstr. 19, 49143 Jeggen, Germany, gediga@eval-institut.de
[2] Department of Computer Science, Brock University, St. Catherines, Ontario, L2S 3AI, Canada, duentsch@cosc.brocku.ca

Abstract. As most data models, “Computing with words” uses a mix of methods to achieve its aims, including several measurement indices. In this paper we discuss some proposals for such indices in the context of rough set analysis and present some new ones.

In the first part we investigate several classical approaches based on approximation quality and the drop of approximation quality when leaving out elements. We show that using the approximation quality index is sensible in terms of admissibility, and present additional indices for the usefulness and significance of an approximation. The analysis of a “drop” is reinterpreted in terms of model comparison, and a general framework for all these concepts is presented.

In the second part of the paper we present an example how using similar nomenclature in the theory of Choquet-type aggregations of fuzzy measurements and rough set approximation quality, without regard to the fine structure of the underlying model assumptions, can suggest connections where there are none.

On a more positive note, we show that so called qualitative power and interaction indices, which are structurally similar to quantitative Choquet-type aggregations can be used in the context of rough set analysis. Furthermore, we propose an entropy-based measure which allows the use of qualitative power and interaction indices as an approximation.

[⋆] Co-operation for this paper was supported by EU COST Action 274 “Theory and Applications of Relational Structures as Knowledge Instruments” (TARSKI), www.tarski.org

[⋆⋆] Equal authorship is implied

> “Mesmerized by a single-purpose, mechanised ‘objective’ ritual in which we convert numbers into other numbers and get a yes-no answer, we have come to neglect close scrutiny of where the numbers come from”. [3]

1 Introduction

Notwithstanding a widely spread fascination with numbers, it is recognised that human behaviour is often guided by imprecise concepts. In recognition of this fact, the direction “From computing with numbers to computing with words”, put forward by Zadeh [36], calls for a manipulation of perceptions instead of measurements:

> “The rationale for computing with words rests on two major imperatives: 1. computing with words is a necessity when the available information is too imprecise to justify the use of numbers and 2. when there is tolerance for imprecision which can be exploited to achieve tractability, robustness, low solution cost and better rapport with reality” [36, p. 111].

This is not to say that computing with words does not use numbers; for example, fuzzy events and constraint propagation assume a continuous scale of truth values for propositions. Leaving aside the question where these initial parameters come from, one finds that subsequent processing may generate other numbers such as averages or maxima, and the question arises in which sense these measurements are “meaningful”, and what, if anything, these numbers actually measure. Wrong interpretations of numbers may lead to scaling artifacts – constructs which apply operations to a model which are not justified. A classical instance is a study by Miller [20][1], who, in an investigation of traffic deaths, assigned numbers to four groups of people (0 - white male, 1 - black male, 2 - white female, 3 - black female), and proceeded to take averages and variances. It turned out that over the investigated population, the average person to cause a fatal accident was a black male ($\bar{x} = 1$). While this example is certainly extreme, the danger to fall into the trap so adequately described above by Cohen is always present when relations among numbers are used for representing relations among objects.

The main topic of this paper is the classical problem how to build meaningful indices from data or from other indices. Both processes need a scaling theory, which enables the user to understand what the indices mean in terms of the data and their context. It easy to declare a concept “useful” if it fits a few small examples, but usually quite hard to substantiate such claim theoretically or experimentally. Even though some systems work “in spite of the erroneous assumptions that underly them” [1], without such a theory a sound basis for interpretation of numbers as measures is not given, and, more often than not, such an interpretation will at some stage lead to wrong results. Data analysis methods which are not primarily quantitative (and,

[1] Reported by Vogel [34, p.64]

consequently, do not have a “built in” scaling theory) need to pay particular attention to this fact.

In the present paper, we will be concerned with a basic statistic of rough set data analysis, the approximation quality, and its interpretation as an ordinal or interval scale. A main theme will be the distinction between *admissibility* and *usefulness* of a measure. In short, we call a measure *admissible* if it satisfies a given standard for the description of a situation, and *useful*, if it can be used to give advice for decision making. We will see that these situations need different tools to cope with the specific circumstances.

As an example, we will exhibit an instance where the application of one method (Choquet-like measures) to another (rough set approximation quality) leads to measurements without measure – numbers which cannot be meaningfully interpreted. We will focus on the influence of one (or more) feature(s) on a decision process as investigated in multi-criteria decision making (MCDM), and the influence of one (or more) feature(s) on the approximation quality of rough set data analysis (RSDA) with respect to a decision attribute. We also suggest that capacities with a maximum operator instead of $\sum$, similar to the Sugeno integral, can be useful in the context of RSDA, and that these measures can be applied to (rough) entropy as well in a meaningful way. It is not an aim of the paper to criticise “ad hoc” methods for data analysis *per se* which are certainly useful in specific cases. We also do not claim that the methods we suggest are universally better than others. Indeed, results of Wolpert and Macready [35] show that no classification algorithm can always outperform another one. What we do want to show is that data modelling has to consider the scaling assumptions of the applied measures and indices, and that things may be more complicated than they seem at first glance. It is necessary to guarantee the meaningfulness of measurements in the first place; the question of whether one is computationally better than another – or even more successful – can be posed after the first step. As expressed in the context of weighted voting in the seminal paper by Banzhaf [1],

> “. . . its intent is only to explain the effects which necessarily follow once
> the mathematical model and the rules of its operation are established . . . ”

The paper is organised as follows: We will first develop the necessary machinery from rough sets, followed by a discussion of several frequently used indices and relations in RSDA, and we show how the model of a proportional error reduction helps to unify the different approaches. Afterwards, we recall various capacities and aggregation measures and show how they differ. Section 4.3 will investigate the connection between the two contexts, based on a discussion of some examples. In Section 4.4 we will present some thoughts on a reconciliation. Finally, we will discuss some examples, in which we show how the different techniques behave, and will close the paper with an outlook.

2 Rough set theory and approximation quality

Rough set data analysis (RSDA) was introduced by Z. Pawlak [22], and has since gained importance as an instrument for non-invasive data analysis. We invite the reader to consult [24] for a short introduction to traditional RSDA, and [9] for a more detailed presentation.

Knowledge representation in rough set data analysis is done via *information systems*. These are structures of the form

$$I = \langle U, \Omega, \{V_x : x \in \Omega\} \rangle, \tag{1}$$

where

- U is a finite set of objects.
- Ω is a finite set of mappings $x : U \to V_x$; each $x \in \Omega$ is called an *attribute*.
- V_x is the set of *attribute values* of attribute x.

Each set P of attributes defines an equivalence relation θ_P (and an associated partition) on U by

$$x \theta_P y \iff a(x) = a(y) \text{ for all } a \in P. \tag{2}$$

Note that $\theta_\emptyset = U \times U$.

Given an information system I, a basic construction of RSDA is the approximation of a given partition $\mathcal{R}$ of U by the partition generated by a set P of attributes. For example, one may want to reduce the number of attributes necessary to identify the class of an object $x \in U$ by its feature vector – i.e. looking for a key in a relational database –, or generate rules from a decision system (which is an information system enhanced by a decision attribute).

By its very nature, RSDA operates on the lowest level of data modelling, namely, on what is commonly called “raw data”, captured in an information table as defined above. Thus, in the language of measurement theory [32], RSDA resides on a *nominal scale* where meaningful operations are those that preserve equality. This parsimony[2] of model assumptions can be seen both as a strength and a weakness: Having only few model assumptions allows the researcher to investigate a greater variety of situations than, for example, data samples with a pre-assumed distribution such as normal or iid, which are required by many statistical methods. On the other hand, the results obtained by RSDA may be too weak to allow meaningful inference about the situation at hand. A case in point is the basically logical (or algebraic) nature of RSDA: Its outcome is a set of “if - then” rules, which are logically true,

[2] It is sometimes claimed, that RSDA has no (pre-)requisites; a moment’s reflection shows that this cannot be true: To be able to distinguish objects by their feature vector, one has to assume that the data contained in the vector are accurate.

and can be well used in a descriptive situation or for *deductive* knowledge discovery, i.e. analysing (the rules in) a given datatable. However, if one looks at *inductive* reasoning such as prediction or classification of previously unseen objects, such rules are not necessarily useful. If, say, each rule is based on only one instance (i.e. each object determines a unique rule), then the rule set will not be helpful for classifying new elements. Therefore, in these situations, additional tools are required to supplement the results obtained by basic RSDA. As a first step, well within the RSDA philosophy, one uses information given by the data themselves, usually in form of counting parameters. One of the first (and most frequent) to have been used is the *approximation quality* γ, which, roughly speaking, measures the goodness of fit of expressing knowledge about the world, which is given by one set of features, by another set of features in the following way: Suppose that θ is an equivalence relation on a set U with associated partition $\mathcal{P}$, and that $X \subseteq U$. Then, we first set

$$\pi_\theta(X) = \frac{|\bigcup\{Y \in \mathcal{P} : Y \subseteq X\}|}{|X|}.$$

This index measures the relative number of elements in X which can be classified as certainly being in X, given the granularity provided by $\mathcal{P}$. If P is a set of attributes and $\mathcal{R}$ a fixed partition of U, then the *approximation quality of P* (with respect to $\mathcal{R}$) is defined as

$$\gamma_{\mathcal{R}}(P) = \sum_{Y \in \mathcal{R}} \frac{|Y|}{|U|} \cdot \pi_{\theta_P}(Y). \tag{3}$$

$\gamma_{\mathcal{R}}$ measures the relative cardinality of correctly classified elements of U as being in a class of $\mathcal{R}$ with respect to the indiscernability relation θ_P. In the situations which we are going to consider, the partition $\mathcal{R}$ arises from a decision attribute d; in the sequel, we assume that $\mathcal{R}$ is fixed, and we will just write γ instead of $\gamma_{\mathcal{R}}$. To avoid trivialities, we assume that $\mathcal{R}$ has at least two classes. By a *model* we understand a set of attributes P along with the set of deterministic rules which are generated by P with respect to the partition $\mathcal{R}$ generated by a decision attribute d. For reasons of brevity, we sometimes just call P a model.

3 Relations based on approximation quality

Since $\gamma(P)$ is a real number between 0 and 1, one can, in principle, apply transformations and form relations with the γ values as real numbers. However, in order that such operations are meaningful and to result in a valid interpretation, one has to have a theory, sometimes called a *scaling model*, which justifies the use of the transformation. In this Section, we will discuss some "standard" approaches of handling γ, while more complex transformations will be discussed later.

3.1 Comparing approximation qualities

One approach is the comparison of γ with a fixed number $0 \lneq c \leq 1$. Choosing the constant c is up to the user or investigator, and is driven by practical necessities; having scanned the literature on applied RSDA, we have found that c is never chosen less than 0.5 and often close to 1.0. The usual interpretation says that any set of attributes P with $\gamma(P) \lneq c$ shows a "bad" approximation of $\mathcal{R}$, and any P with $\gamma(P) \geq c$ is "admissible" for the approximation of $\mathcal{R}$; this approach is used to define "iso-gamma" reducts and core [24, p. 51].

This is a straightforward and seemingly unequivocal interpretation of γ, and thus, this technique is frequently used. However, the question arises whether it is based on a meaningful interpretation of the approximation quality. The answer is: It depends on the context, and we have to distinguish between *admissibility* and *usefulness*: Suppose that $c \leq \gamma(P)$.

1. If P is claimed to be admissible for the approximation of $\mathcal{R}$, we do not run into problems, because the approximation quality induced by P is not less than the required relative number c of elements which are correctly classified with respect to $\mathcal{R}$; therefore, the approximation of $\mathcal{R}$ is admissible with respect to c. This interpretation is based purely on the algebraic structure of the equivalence relations, and assumes that the data is correct "as given". In other words, the approximation quality counts the relative number of elements which can be captured by the deterministic rules associated with P. In this sense, γ counts what is "logically true".
2. If P is claimed to be useful (i.e. it can or should be used) for the approximation of $\mathcal{R}$ – such as in decision support and medical diagnosis –, the situation is more complicated: One has to take into account that attaining the standard c may have come about by random influences ("noise", "error"), and that, therefore, the application of P for the approximation of $\mathcal{R}$ is not necessarily useful [7]. A simple example of divergence of admissibility and usefulness is an information system consisting of a running number and a decision attribute. Here, we have $\gamma = 1$, and the running number is helpful to identify any case without error. However, knowing the running number without knowing the value of the decision attribute does not help – for this purpose $\gamma = 1$ is not useful.

A further example, shown in Table 1 on the facing page, shall illustrate how random processes may influence the results in case of very low approximation qualities. If $\mathcal{R}$ is the partition associated with the decision attribute d, then

$$\gamma(p) = \gamma(q) = 0. \tag{1}$$

Whereas p is essential to predict d with only a class switch of 4 and 8 achieving perfect approximation quality, q is only required to "separate" 4 and 8 from their respective p-classes. In other words, a mis-classification of 4 and 8 may well have taken place, owing to random influences in representing the data.

Table 1. A simple decision system [7]

U	*p*	*q*	*d*	U	*p*	*q*	*d*
1	0	0	0	5	1	0	1
2	0	2	0	6	1	2	1
3	0	2	0	7	1	2	1
4	1	1	0	8	0	1	1

The interpretation as "usefulness" needs an additional tool to cope with randomness. There are – at least – two different ways to evaluate the usefulness of P: Cross validation methods or statistical testing of "random admissibility". We prefer the latter approach, because cross validation methods require additional (non–rough!) model assumptions for the classification of unseen cases in the learning sample, while testing usefulness – as "random admissibility" – does not.

The γ statistic is also used to define an ordinal relation among subsets of the attribute set by setting

$$P \preceq_\gamma Q \Longleftrightarrow \gamma(P) \leq \gamma(Q).$$

The interpretation "Q is at least as admissible as P" causes no difficulty, because choosing a criterion c with $\gamma(P) \leq c \leq \gamma(Q)$ results in such an interpretation. On the other hand, interpreting $\preceq_\gamma$ by assigning the term "better than" (in the sense of "more useful") to the defined relation is not meaningful: Since this interpretation only makes sense in a numerical system in which a $\leq$ relation is established[3], the notation "better than" cannot be filled by an empirical interpretation.

3.2 Comparing differences of approximation qualities

From its early days, identification of "important features" has been an integral part of RSDA:

> "The idea of attribute reduction can be generalised by an introduction of the concept of *significance*[4] *of attributes*, which enables an evaluation of attributes ... by associating with an attribute a real number from the $[0,1]$ closed interval; this number expresses the importance of the attribute in the information table." [16]

[3] Note, that the numbers on the shirts of soccer players are $\leq$ related, but this does not induce a "better than" relation among soccer players.

[4] The terminology "significance" is somewhat unfortunate, because the name has a fixed (and quite diverging) meaning in Statistics. We think the name "importance of an attribute" would be a better choice.

The first approach to analyse the "significance" of an attribute with respect to a given attribute set was the inspection of the "drop" of the approximation quality when leaving out an attribute a from a set P defined by

$$d_1(P,a) = \gamma(P) - \gamma(P \setminus \{a\}), \tag{2}$$

see e.g. [23], p. 59. The analysis of the drop starts after P has been chosen as an admissible attribute set, for example, a reduct. Therefore, the difference cannot, in general, be interpreted in terms of admissibility, because $\gamma(P \setminus \{a\})$ may have a smaller value than the chosen standard c, which sets the standard for admissibility.

One can argue that d_1 is meaningful, because γ forms an interval scale. In terms of admissibility, this is an additional condition which has as one consequence that a difference from $\gamma(P) = 0.1$ to $\gamma(P \setminus \{a\}) = 0.0$ is assumed to be identical (in terms of set approximation) to the difference among $\gamma(P) = 1.0$ and $\gamma(P \setminus \{a\}) = 0.9$. One has to be aware, however, that – contrary to the approximation quality comparison, which relies only on a monotone relationship among the numerical measures – the interpretation of differences needs much stronger assumptions than a comparison of numerical values. The problem that "differences ignore the base rate" occurs in other disciplines as well; in descriptive statistics, for instance, the odd-ratio quotient of two probabilities is often used to describe the "difference" among probabilities in a meaningful way. Building the odd-ratio of probabilities requires an extra scaling theory, which serves as a foundation for the interpretation of the odd-ratio. Obviously, something comparable would be needed in the context of RSDA as well. As far as we are aware, there has been no attempt to provide such a theory.

Simple differences should not be interpreted in terms of usefulness, because the step from 0.0 to 0.1 can easily explained by random processes, whereas it is much harder – given comparable distributions of attribute values – to result in a step from 0.9 to 1.0.

The problem of dealing with differences has been addressed by a newer approach for measuring the "significance" of an attribute [16, 31]:

$$d_2(P,a) = \frac{\gamma(P) - \gamma(P \setminus \{a\})}{\gamma(P)} = 1 - \frac{\gamma(P \setminus \{a\})}{\gamma(P)}$$

It is not at all clear how this function should be interpreted, and it has some peculiar properties: Given identical linear differences with different base rates, one can observe that the d_2 differences will be smaller, with an increasing base rate. We have, for example,

$\gamma(P)$	$\gamma(P \setminus \{a\})$	$d_1(P,a)$	$d_2(P,a)$
1.0	0.9	0.1	0.1
0.6	0.5	0.1	0.167
0.2	0.1	0.1	0.5

This property of $d_2(P,a)$ is not satisfactory, because – as the preceeding discussion shows – a constant gain should be result in larger values, if the base rate is at the upper end of the scale.

3.3 Averaging of approximation qualities

Another way of treating interval scale information is forming averages. If this operation is applied to set approximation qualities, we again run into problems. Suppose that $\mathcal{A}$ is a collection of attribute sets. The average

$$E[\mathcal{A}] = \frac{1}{|\mathcal{A}|} \sum_{P \in \mathcal{A}} \gamma(P)$$

computes the expectation of the approximation quality when each of the sets $P \in \mathcal{A}$ has the same probability to be used for the approximation of a decision attribute. Unlike the case in which sampling properties can be described by the principle of indifference, forming of attribute sets is not a random choice and is under control of the researcher. Therefore, building an expectation value does not make much sense, because the population for the sample cannot be properly defined. Furthermore, any researcher would agree that $\max\{\gamma(P) : P \in \mathcal{A}\}$ is a characteristic value for the set $\mathcal{A}$ – and that the expectation may offer strange results as Table 2 demonstrates. There, the collection $\mathcal{A}$ consists of six non-admissible attribute sets if $c \gneq 0.2$ for the approximation of $\mathcal{R}$, whereas one set in $\mathcal{B}$ is admissible for any c. The maximum of the approximation qualities will point to $\mathcal{B}$ as the "better set". In contrast, using average values, one sees that the "mean admissibility" of $\mathcal{A}$ is higher then the "mean admissibility" of $\mathcal{B}$, although no element in $\mathcal{A}$ is admissible at all.

Table 2. Strange results using expectations of approximation qualities

$\mathcal{A}$	γ	$\mathcal{B}$	γ
A_1	0.2	B_1	0.0
A_2	0.2	B_2	0.0
A_3	0.2	B_3	0.0
A_4	0.2	B_4	0.0
A_5	0.2	B_5	0.0
A_6	0.2	B_6	1.0
Maximum	0.2		1.0
Mean	0.2		0.167

3.4 Indices of proportional error reduction as a general concept

In the preceeding section, both indices d_1 and d_2 are based on transformations of the γ index by assuming that the transformation somehow fits to the semantics (the “significance” or “importance”). We have shown that d_1 and d_2 are not necessarily meaningful, either for admissibility or usefulness, due to a lack of a sound theory which guides the index building process.

In [10] we have introduced the PRE (Proportional Reduction of Errors) approach of Hildebrand et al. [15] into RSDA, which – in the general case – describes the error reduction when a model is applied, based on the errors of a given benchmark model. In the context of RSDA, we say that an “error” is an object which cannot be explained a deterministic rule. In line with this interpretation, the approximation quality becomes

$$\gamma(P) = 1 - \frac{1-\gamma(P)}{1-\gamma(\emptyset)},$$

which means that $\gamma(P)$ measures the proportional error reduction of a model using the attribute set P in comparison to the “worst case” benchmark model in which every object is counted as an error.

Adapting the PRE approach to the “importance” problem, we find that

$$d_3(P,a) = \begin{cases} 0, & \text{if } \gamma(P \setminus \{a\}) = 1, \\ 1 - \frac{1-\gamma(P)}{1-\gamma(P\setminus\{a\})}, & \text{otherwise,} \end{cases} \tag{3}$$

is a suitable index for the comparison of a model using the attribute set P against a model using the attribute set $P \setminus \{a\}$. The index $d_3(P,a)$ measures the error reduction when using the set of attributes P compared with the benchmark model which uses the set of attributes $P \setminus \{a\}$. This value can be compared to a threshold value c_g, and therefore $d_3(P,a)$ can be interpreted as a measure for the admissibility of the gain.

Comparing the measures d_1, d_2 and d_3, we observe that the behaviour of d_3 is as it should be:

$\gamma(P)$	$\gamma(P \setminus \{a\})$	$d_1(P,a)$	$d_2(P,a)$	$d_3(P,a)$
1.0	0.9	0.1	0.1	1.0
0.6	0.5	0.1	0.167	0.2
0.2	0.1	0.1	0.5	0.111

The evaluation of differences given a small base rate is lower than the same differences when given high base rate.

Because admissibility gain does not take into account random influences, an index for a *usefulness gain* has to be defined as well. To this end, the results of [7] can be used to derive a descriptive measure for a usefulness gain

Descriptive indices such as admissibility or usefulness estimate the actual size of an effect, given a fixed set of model assumptions; therefore, such indices are often called *effect size measures*. In statistical applications, one often considers one kind of effect size measure. However, this can only be done under rather restrictive assumptions. For instance, if it is assumed that two variables represent a bivariate normal distribution, the admissibility and usefulness of the correlation coefficient are identical. If this assumption is dropped, this identity does not hold in general.

Because effect sizes in terms of PRE-measures are used in many contexts, there exists a rule of thumb how to assess effect sizes for expectation – based benchmark models [2]:

Effect size (ES)	Interpretation
$ES \lnsim 0.1$	no effect
$0.10 \leq ES \lnsim 0.3$	small effect
$0.30 \leq ES \lnsim 0.5$	medium effect
$ES \geq 0.5$	large effect

Effect size measures must rely on empirical data in order to estimate the range of effects in real life data – this is just the way, Cohen [2] arrived at the interpretation of effect sizes. But until there exists a database for empirical studies which have been done on the basis of RSDA, the given rule of thumb can be used as a first approximation.

The PRE-measures discussed in this section show a peculiar behaviour in case of $\gamma(P) = 1$. In this situation any PRE-measure will only result in 0 (if the benchmark model results in 1 as well) or 1. In terms of admissibility, this binary nature of the index cannot be resolved, but if a statistical benchmark model is used, it is easy to replace the descriptive PRE-measure by a measure from inference statistics by computing the position of the observed error in the distribution of expected errors given the benchmark model. This position is called (statistical) *significance* and the value should be small (conventionally smaller than 5%) for a good model. It should be noted, that usability and significance are two different concepts – which may dissociate –, although both are using identical random processes. Whereas significance is changed when increasing the number of observations, the usability remains unchanged.

Table 3 on the next page collects all approaches – PRE-measures and significance – discussed so far, enhanced by *set gains*, which are a simple generalisation of the preceeding indices by choosing $S = P$ or $S = \{a\}$ respectively.

4 Capacities, power indices and values of interaction

The influence and power of an attribute, as well as the interaction of several attributes, have been extensively studied in Game Theory and Multicriteria Decision Analysis (MCDA). For an overview of earlier work, we invite the reader to consult the collection of essays edited by Roth [26], and for more recent advances the article by Grabisch [11]. Since the approximation quality has the same mathematical properties as a capacity or fuzzy measure (explained below), it has been claimed that

Table 3. PRE indices as descriptive measures and significance values in RSDA

<table>
<tr><td colspan="4">$\mathcal{P}^\sigma$:= Set of attribute sets, which are constructed by random assignment of elements to the attributes P.
$\mathcal{S}^\sigma$:= Set of attribute sets, which are constructed by random assignment of elements to the attributes S.
$\mathcal{A}^\sigma$:= Set of attribute sets, which are constructed by random assignment of elements to the attribute a.</td></tr>
<tr><th>Error of the model</th><th>Error of the benchmark model</th><th>Interpretation</th><th>Source</th></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\gamma(\emptyset)=1$</td><td>admissibility</td><td>[10]</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\mathcal{E}\left[\gamma(R)\vert R\in\mathcal{P}^\sigma\right]$</td><td>usefulness</td><td>[10]</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\gamma(P\setminus\{a\})=1-\gamma((P\setminus\{a\})\cup\emptyset)$</td><td>admissible gain</td><td>this text</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\mathcal{E}\left[\gamma((P\setminus\{a\})\cup R)\vert R\in\mathcal{A}^\sigma\right]$</td><td>usable gain</td><td>[7]</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\gamma(P\setminus S)=1-\gamma((P\setminus S)\cup\emptyset)$</td><td>admissible set gain</td><td>this text</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$1-\mathcal{E}\left[\gamma((P\setminus S)\cup R)\vert R\in\mathcal{S}^\sigma\right]$</td><td>usable set gain</td><td>this text</td></tr>
<tr><th>Error of the model</th><th>Position of the error given the benchmark model</th><th>Interpretation</th><th>Source</th></tr>
<tr><td>$1-\gamma(P)$</td><td>$p\left[1-\gamma(R)\leq 1-\gamma(P)\vert R\in\mathcal{P}^\sigma\right]$</td><td>significance</td><td>[10]</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$p\left[1-\gamma\big((P\setminus\{a\})\cup R\big)\leq 1-\gamma(P)\vert R\in\mathcal{A}^\sigma\right]$</td><td>significant gain</td><td>[10]</td></tr>
<tr><td>$1-\gamma(P)$</td><td>$p\left[1-\gamma\big((P\setminus S)\cup R\big)\leq 1-\gamma(P)\vert R\in\mathcal{S}^\sigma\right]$</td><td>significant set gain</td><td>this text</td></tr>
</table>

> “Due to this equivalence, it is possible to use different indices defined on fuzzy measures to assess the relative value of information supplied by each attribute and to analyze interaction between attributes.” [13].

After introducing the necessary machinery, we shall show in this Section that using γ as an interval scaled capacity leads to scaling artifacts which do not take into account the basic model assumptions of the indices, and thus, they provide a “measurement without measure”. Alternatives, which are more promising evaluation tools, are presented as well.

4.1 Quantitative indices of power and interaction

In decision theory, the aggregation of criteria is usually done by weighted arithmetic means, or, as they are sometimes called, discrete integrals, and we will define below the most often used indices. Throughout, we suppose that $U=\{1,...n\}$ is a finite set, which, in the present context, can be interpreted as a set of criteria or a set of players.

A function $\mu:2^U\rightarrow[0,1]$ is called a *capacity* or *fuzzy measure* if for all $X\subseteq U$

$$\mu(\emptyset)=0,\quad \mu(X)\leq 1. \tag{1}$$

$$A\subseteq B\subseteq X \text{ implies } \mu(A)\leq\mu(B). \tag{2}$$

We will usually identify singletons with the element they contain, e.g. we will write $\mu(p)$ instead of $\mu(\{p\})$.

The set function μ takes into account that the contribution of one criterion to a set S of criteria may vary, depending on the choice of S. In other words, μ is chosen in such a way that it respects the interaction among criteria according to the belief of the investigator.

The common quantitative aggregation functions, resulting in power indices, rely on a simple difference construction: If $K \subseteq U$ and $m \notin K$, we let

$$\Delta^{\mu}(K,m) = \mu(K \cup \{m\}) - \mu(K) \tag{3}$$

denote the (unweighted) marginal contribution of m to $\mu(K \cup \{m\})$. Two well known power indices are based on Δ^{μ}: The *Shapley value* [28] is defined by

$$\varphi_S^{\mu}(m) = \sum_{K \subseteq U \setminus \{m\}} \frac{(n-|K|-1)!|K|!}{n!} \Delta^{\mu}(K,m). \tag{4}$$

It is usually interpreted as a measure of the weighted marginal average contribution of m to sets of the form $\mu(K \cup \{m\})$ under the assumption that "all orders in which an individual enters any coalition are equiprobable" [29], and it is often called the *importance of m with respect to the weighting μ.*

Another value frequently considered is the *Banzhaf value* [1], given by

$$\varphi_B^{\mu}(m) = \frac{1}{2^{n-1}} \sum_{K \subseteq U \setminus \{m\}} \Delta^{\mu}(K,m), \tag{5}$$

Both indices φ^{μ} make several model assumptions including the following:

A1. Taking averages of differences is meaningful.

A2. If μ is a capacity and σ a permutation of U, then, for all $m \in U$,

$$\varphi^{\mu}(m) = \varphi^{\sigma\mu}(\sigma(m)).$$

Here, $\sigma\mu(K) = \{\sigma(k) : k \in \mu(K)\}$.

We invite the reader to consult [4, 5, 17] for axiomatisations of the Shapley and Banzhaf values.

Apart from the “first order” power indices more complicated indices can be built as well. If $K \subseteq U$ and $i, j \notin K$, let

$$\Delta^{\mu}(K,\{i,j\}) = \mu(K \cup \{i,j\}) - \mu(K \cup \{i\}) - \mu(K \cup \{j\}) + \mu(K). \tag{6}$$

Weighted averages of these “second order” differences – called *interaction values* – result in the Shapley interaction index

$$\varphi_S^{\mu}(\{i,j\}) = \sum_{K \subseteq U \setminus \{i,j\}} \frac{(n-|K|-2)!|K|!}{(n-1)!} \Delta^{\mu}(K,\{i,j\}) \tag{7}$$

and the Banzhaf interaction index

$$\varphi_B^{\mu}(\{i,j\}) = \frac{1}{2^{n-2}} \sum_{K \subseteq U \setminus \{i,j\}} \Delta^{\mu}(K,\{i,j\}), \tag{8}$$

respectively [21, 27]. Following [18], we say that $i, j \in U$ show a *negative interaction* if

$$\varphi^{\mu}(\{i,j\}) \lneq 0, \tag{9}$$

they *do not interact* or are *uncorrelated* , if

$$\varphi^{\mu}(\{i,j)\} = 0, \tag{10}$$

and they *show a positive interaction*, if

$$\varphi^{\mu}(\{i,j)\} \gneq 0. \tag{11}$$

4.2 Qualitative indices of power and interaction

As an alternative to the assumption of an interval scale for $\mu(K)$, Grabisch [12] and Dubois et al. [6] offer an index which assumes only ordinal scaling and uses neither differences nor the mean to calculate the influence of an element. Let $\bigoplus$ denote the maximum of an ordered set (if it exists), and set

$$a \ominus b = \begin{cases} a, & \text{if } a \gneq b, \\ 0, & \text{otherwise.} \end{cases}$$

The *qualitative power value of i* is defined by

$$\varphi_Q(i) = \bigoplus_{K \subseteq U \setminus \{i\}} \mu(K \cup \{i\}) \ominus \mu(K). \tag{12}$$

This expression can be handled formally as the indices introduced in Section 4. The value $\varphi_Q(i)$ is the largest value of μ for a set containing i, which will drop, when i is left out. Based on investigations of Grabisch [12], Dubois et al. [6] note that this value “seems to be the only reasonable definition for a qualitative Shapley value”. It is based only on the ordinal scaling assumption, because φ_Q is monotone invariant with respect to μ, since for any monotone mapping $T : \mathbb{R} \to \mathbb{R}$ we have

$$\varphi_Q^{T(\mu)}(m) = T\left(\varphi_Q^{\mu}(m)\right).$$

Therefore, φ_Q is an instance of the ordinal meaningful aggregation functions proposed by Marichal and Mathonet [19].

Although φ_Q shares some structural properties with φ_B and φ_S, the interpretation of φ_Q is quite different: Whereas φ_Q computes a maximum, the quantitative counterparts compute an average (based on different weights). Another difference is due the monotone invariance of φ_Q: Because any monotone transformation of μ is admissible, there is no need to restrict μ to the interval $[0,1]$.

It is interesting to note, that the application of φ_Q is not restricted to capacities, but is meaningful with any set function: Even if we drop the assumption of monotony of μ with respect to $\subseteq$, the value of $\varphi_Q(i)$ is of interest, because its value is the largest one for which i shows a non-redundant contribution to μ.

Let us note the following “relative commutativity" of the operation $\ominus$:

Lemma 1. *$(a \ominus b) \ominus c = (a \ominus c) \ominus b$ for all a, b, c.*

Proof. The conclusion follows from the following table:

$a \ominus b$	$a \ominus c$	$(a \ominus b) \ominus c$	$(a \ominus c) \ominus b$
a	a	a	a
a	0	0	0
0	a	0	0
0	0	0	0

It is now straightforward to define a *qualitative value of interaction* as well, using the $\bigoplus$ and $\ominus$ operators[5]:

$$\varphi_Q(\{i,j\}) = \bigoplus_{K \subseteq U \setminus \{i,j\}} (\mu(K \cup \{i,j\}) \ominus \mu(K \cup \{i\})) \ominus \mu(K \cup \{j\}). \tag{13}$$

By Lemma 1, $\varphi_Q(\{i,j\})$ is well defined, and it can be shown that this expression is formally the same as the quantitative interaction indices. The index $\varphi_Q(\{i,j\})$ addresses the largest value of μ for which i, j truly interact in the sense that both elements contribute a part to μ, that cannot be expressed by the other element. If $\varphi_Q(i,j) = 0$, no such interaction among i, j is observable. Because φ_Q uses monotone invariant operators, it is monotone invariant as well, and the name “qualitative value of interaction” is justified.

4.3 Applications of capacities in Rough Set Data Analysis

Choquet-type aggregation measures have been considered in the context of RSDA:

> “All these indices can be useful to study the informational dependence among the considered attributes and to choose the best reducts. . . . the Shapley values . . . can be interpreted as measures of importance of the corresponding attributes in the rough approximation.” [13, p. 102f]

[5] This construction may be known

We will investigate below, whether the aims addressed in the quote can be fulfilled by quantitative power indices in question.

We start with a minor observation: If quantitative power indices are used, one has to differentiate between the interpretation of the Shapley and the Banzhaf index, since it is possible that $\varphi_S(p) \lneqq \varphi_S(q)$ and $\varphi_B(p) \gneqq \varphi_B(q)$ [33, for an example see Table 4]. Therefore, it is not clear which of these indices, if any, tells us something about the "informational dependence of the considered attributes" [14].

Table 4. A non-monotone relationship of Banzhaf- and Shapley-values

K	μ	$w_S(K)$	$\Delta(K,p)$	$\Delta(K,q)$	$w_S(K)\cdot\Delta(K,p)$	$w_S(K)\cdot\Delta(K,q)$
$\emptyset$	0	0.25	0.5	0	0.125	0
$\{r\}$	0	0.0833	0.5	0.8	0.0417	0.0667
$\{s\}$	0	0.0833	0.5	0.8	0.0417	0.0667
$\{p\}$	0.5	0.0833		0		0
$\{q\}$	0	0.0833	0.5		0.0417	
$\{r,s\}$	0	0.0833	1	1	0.0833	0.0833
$\{r,p\}$	0.5	0.0833		0.5		0.0417
$\{r,q\}$	0.8	0.0833	0.2		0.0167	
$\{s,p\}$	0.5	0.0833		0.5		0.0417
$\{s,q\}$	0.8	0.0833	0.2		0.0167	
$\{r,s,p\}$	1	0.0833		0		0
$\{r,s,q\}$	1	0.0833	0		0	
		Sum:	3.4	3.6	0.3668	0.3

A more severe problem is the fact that comparing Banzhaf – or Shapley – values leads directly to a comparison of averages based on a collection of sets , which may lead to "strange" results, as the example in Section 3.3 demonstrates. This can also be seen quite easily as follows: If we compare set functions of the type

$$\varphi^{\mu}(i) = \sum_{K \subseteq U \setminus \{i\}} w(|K|) \cdot \big(\mu(K \cup \{i\}) - \mu(K)\big),$$

a straightforward calculation shows that

$$\varphi^{\mu}(i) - \varphi^{\mu}(j) = \sum_{K \subseteq U \setminus \{i,j\}} \big(w(|K|) + w(|K|+1)\big) \cdot \big(\mu(K \cup \{i\}) - \mu(K \cup \{j\})\big).$$

Therefore, the difference of Banzhaf values can be rewritten as

$$\varphi^{\mu}_B(i) - \varphi^{\mu}_B(j) = \Big(\frac{1}{2^{n-2}} \sum_{K \subseteq U \setminus \{i,j\}} \mu(K \cup \{i\})\Big) - \Big(\frac{1}{2^{n-2}} \sum_{K \subseteq U \setminus \{i,j\}} \mu(K \cup \{j\})\Big).$$

This leads to a very simple interpretation of Banzhaf value differences: $i \prec_B j$ iff $\varphi^{\mu}_B(i) - \varphi^{\mu}_B(j) \lneqq 0$, which means that the average of μ based on the sets $K \cup \{i\}$ is less than the average of μ based on the sets $K \cup \{j\}$.

For differences of Shapley values we find

$$\varphi^{\mu}_{S}(i) - \varphi^{\mu}_{S}(j) = \sum_{K \subseteq U \setminus \{i,j\}} \frac{(n-|K|-2)!|K|!}{(n-1)!} \Big(\mu(K \cup \{i\}) - \mu(K \cup \{j\})\Big),$$

which addresses the comparison of Shapley-weighted averages of μ-values based on different sets.

Comparing Banzhaf or Shapley values leads to a comparison of mean values based on disjoint sets – which is exactly the situation we have discussed in Section 3.3. One might argue that this cannot be observed in RSDA, but a more refined example can be constructed . Using the data in Table 5 we observe $\gamma(\{p\}) = \gamma(\{a,p\}) =$ $\gamma(\{b,p\}) = \gamma(\{c,p\}) = \gamma(\{a,b,p\}) = \gamma(\{a,c,p\}) = \gamma(\{b,c,p\}) = 0.125, \gamma(\{a,b,c,p\}) = 0.25$, $\gamma(\{q\}) = \gamma(\{a,q\}) = \gamma(\{b,q\}) = \gamma(\{c,q\}) = \gamma(\{a,b,q\}) = \gamma(\{a,c,q\}) = \gamma(\{b,c,q\}) = 0.0$, and $\gamma(\{a,b,c,q\}) = 1$. Now it is easy to calculate that $0.140 = \varphi^{\mu}_{B}(p) \gneqq \varphi^{\mu}_{B}(q) = 0.125$, and once again, a set of attribute sets with very low approximation qualities dominates another set with one perfect attribute combination.

Table 5. Information system which approximation qualities should not be averaged

condition attributes					decision attr.	condition attributes					decision attr.
a	b	c	p	q	d	a	b	c	p	q	d
1	1	0	0	1	0	1	1	1	1	1	4
2	1	0	0	2	0	2	1	1	1	2	4
1	1	0	1	2	1	1	1	1	1	2	5
2	2	0	1	1	1	2	2	1	1	1	5
1	2	0	1	1	2	1	2	1	1	1	6
2	2	0	1	2	2	2	2	1	1	2	6
1	2	0	1	2	3	1	2	1	1	2	7
2	1	0	1	1	3	2	1	1	1	1	7

Whatever power index will be used, it should be noted that this index cannot be interpreted in terms of usefulness. As an example, consider the case where $\Omega = \{p,q\}$, $\gamma(p) = c$ with $c \lneqq 1$, and $\gamma(q) = 1$. The Banzhaf values and qualitative Shapley values are

$$\begin{aligned} \varphi_B(p) &= \frac{1}{2}((1-1)+(c-0) = \frac{c}{2}, \\ \varphi_B(q) &= \frac{1}{2}((1-c)+(1-0) = 1-\frac{c}{2}, \\ \varphi_Q(p) &= \max\{1 \ominus 1, c \ominus 0\} = c, \\ \varphi_Q(q) &= \max\{1 \ominus c, 1 \ominus 0\} = 1. \end{aligned}$$

Now, $\varphi_x(p) \lneqq \varphi_x(q)$ holds for every $x \in \{B,S,Q\}$ and any value of c strictly less than 1, regardless of the “quality” of the attribute q – it may be a running number or an attribute with low entropy. It is easy to construct situations in which attribute p is more useful than attribute q, but neither of the power indices would detect this.

Recall from (10) that in MCDM two objects p,q are “uncorrelated” or “independent”, if $\mu(\{p,q\}) = \mu(p) + \mu(q)$. The application of the quantitative interaction values in RSDA are problematic as well, but the term “uncorrelated” should be used with caution: Assume an attribute q with $\gamma(q) = 0$ and use another another attribute q', which generates the same partition on U. Then q and q' show no interaction with d and can be called “uncorrelated”, but their dependency is maximal.

4.4 A qualitative power value based on rough entropy

In [8] we have noted that γ is a conditional measure, and therefore, comparisons of γ values are only valid in so called *nested models*, which means that $\gamma(P)$ and $\gamma(Q)$ are only comparable in a meaningful manner if either $P \subseteq Q$ or $Q \subseteq P$ holds. To allow model selection from all possible attribute sets within RSDA, we have presented a measure called *entropy of deterministic rough approximation* which is based on the maximum entropy principle as a worst case. Suppose we have a fixed decision attribute d generating the equivalence relation θ_d on U. If Q is a set of attributes generating θ_Q, we define a new equivalence θ_Q^{det} by

$$x\theta_Q^{\mathsf{det}}y \iff \begin{cases} x\theta_Q y, & \text{if } \theta_Q(x) = \theta_Q(y) \text{ and } \theta_Q(x) \subseteq \theta_d(z) \text{ for some } z \in U, \\ x = y, & \text{otherwise.} \end{cases}$$

Its associated probability distribution, based on the principle of indifference, is given by $\{\hat{\psi}_K : K \in \mathcal{P}(\theta_Q^{\mathsf{det}})\}$ with

$$\hat{\psi}(K) = \frac{|K|}{n} \tag{14}$$

The *entropy of deterministic rough approximation* (with respect to Q and d) is now defined by

$$H^{\mathsf{det}}(Q) = \sum_K \hat{\psi}(K) \cdot \log_2(\frac{1}{\psi(K)}).$$

If

$$H(d) = \sum_{L \in \mathcal{P}_d} \frac{|L|}{n} \cdot \log_2 \frac{n}{|L|},$$

$H^{\mathsf{det}}(Q)$ can be standardised by

$$\mathrm{NRE}(Q) := 1 - \frac{H^{\mathsf{det}}(Q) - H(d)}{\log_2(n) - H(d)}, \tag{15}$$

assuming $H(d) \lneq \log_2(n)$. We obtain a measure of approximation success within RSDA, which can be used to compare different models in terms of the combination of coding complexity and uncertainty outside the approximation in the sense that a perfect approximation results in $\mathrm{NRE}(Q) = 1$, and the worst case is at $\mathrm{NRE}(Q) = 0$. Unlike γ, NRE is an unconditional measure, because both, the complexity of the rules generated by the independent attributes and the uncertainty after approximation, are merged into one measure.

We are now able to define the *qualitative power index of an attribute m using* NRE by

$$\varphi(m) = \max\{\mathrm{NRE}(K \cup \{m\}) \ominus \mathrm{NRE}(K) : K \subseteq U \setminus \{m\}\}. \tag{16}$$

This value is meaningful, because it addresses the value of the maximum NRE (or minimum rough entropy) for which attribute m contributes a non-zero amount of additional information. It is easy to see that

$$\max_m \varphi(m) = \max_K \mathrm{NRE}(K)$$

so that the maximum of φ also is the highest NRE over all subsets of U.

5 An example

One of the first published applications of RSDA was a study which describes patients after highly selective vagotomy for duodenal ulcer [25]. An enhanced data set of 122 patients was used in [30], and this data set will be used in the sequel.

The information system consisted of 11 condition attributes and a decision attribute “Visick grading” . Comparing approximation qualities, it was decided that the attribute set P, consisting of

3: Duration of disease
4: Complication
5: Basic HCI concentration
6: Basic Vol. of gastric juice
9: Stimulated HCI concentration
10: Stimulated Vol. of gastric juice

is a good basis to approximate attribute “Visick grading” with an approximation quality $\gamma(P) = 0.795$. Inspecting the decline of the approximation quality (2) under the assumption of an admissibility threshold of $c = 0.55$, it was found that the attribute sets

$$A = \{4,5,6,9,10\},\ B = \{3,4,6,9,10\},\ C = \{3,4,5,6,9\}$$

Table 6. Analysis of the duodenal ulcer data, I

Attribute set	γ	Interpretation	d_3	Interpretation
3,4,5,6,9,10 (P)	0.795	admissible		
·,4,5,6,9,10 (A)	0.590	admissible	0.500	non admissible gain
3,·,5,6,9,10	0.516	not admissible	0.576	admissible gain
3,4,·,6,9,10 (B)	0.680	admissible	0.359	non admissible gain
3,4,5,·,9,10	0.549	not admissible	0.545	admissible gain
3,4,5,6,·,10 (D^*)	0.631	admissible	0.444	non admissible gain
3,4,5,6,9,· (C)	0.648	admissible	0.418	non admissible gain

are candidates for future research. These are presented in Table 6; it turns out the attribute set $D^* = \{3,4,5,6,10\}$ with $\gamma = 0.631$ should have been included as well for further analysis.

The gain analysis with d_3 can be interpreted in the same way: Leaving out attribute 4 or 6 results in models, which are not admissible, if we set $c_g = 0.5$. The analysis of gain complements the set admissibility: If a subset of P is labelled as admissible, the gain is labelled to be non admissible and vice versa. This result is not a triviality, because the admissibility labels for sets and gains are driven by the different constants c and c_g.

As we have discussed above, the analysis of usefulness and significance requires a simulation frame: The results, based on 1000 simulated randomisations for each analysis, are gathered in Table 7. Column 1 shows the attributes under consideration, column 2 the observed approximation quality γ of this set, column 3 the expectation of γ given the benchmark model, column 4 the corresponding PRE-measure (usefulness), and column 5 the estimated position of γ in the distribution of the random matching assumption (significance).

The first part of Table 7 presents the results of usefulness and significance for sets. The admissible sets (P,A,B,C,D^*) are significant as well. P offers a medium effect size, whereas the usefulness of the admissible subsets of P is smaller.

The analysis of gain within P (second part of Table 7) achieves an astonishing result: All attributes are conditional casual within P. This means that there are always only a few of the 122 observations which can be approximated additionally by introducing the attribute under study into the set. Thus, one can argue that the number of observations in the duodenal ulcer information system is too small, and the good results of P are pushed due to overfitting. Because rough entropy is helpful to check the complexity of the rule system based on the attributes – and therefore helpful to prevent against overfitting –, inspection of the normed Rough Entropy (NRE) values in Table 8 provides further insight: Among the given alternatives, set C has the highest NRE (or the lowest complexity) , and from this point of view can be regarded as a favourable model.

Table 7. Analysis of the duodenal ulcer data, II

	Analysis of the attribute set				
Attribute set	γ	$\mathcal{E}[\gamma]$	Usefulness	Significance	Interpretation
3,4,5,6,9,10	0.795	0.703	0.311	0.013	significant, medium usefulness
·,4,5,6,9,10 (A)	0.590	0.554	0.081	0.153	not significant , not useful
3,·,5,6,9,10	0.516	0.484	0.063	0.199	not significant, not useful
3,4,·,6,9,10 (B)	0.680	0.579	0.241	0.018	significant, small usefulness
3,4,5,·,9,10	0.549	0.487	0.121	0.084	not significant, low usefulness
3,4,5,6,·,10 (D*)	0.631	0.515	0.240	0.008	significant, small usefulness
3,4,5,6,9,· (C)	0.648	0.524	0.259	0.011	significant, small usefulness

	Analysis of the gain within $\{3,4,5,6,9,10\}$				
Attribute	γ	$\mathcal{E}[\gamma]$	Usefulness	Significance	Interpretation
3	0.795	0.769	0.112	0.182	not significant, low usefulness
4	0.795	0.751	0.178	0.099	not significant, low usefulness
5	0.795	0.792	0.017	0.394	not significant
6	0.795	0.760	0.145	0.107	not significant, low usefulness
9	0.795	0.763	0.137	0.127	not significant, low usefulness
10	0.795	0.786	0.044	0.310	not significant

Table 8. Analysis of the duodenal ulcer data, III

Attribute set	γ	*NRE*
3,4,5,6,9,10	0.795	0.063
·,4,5,6,9,10 (A)	0.590	0.046
3,·,5,6,9,10	0.516	0.070
3,4,·,6,9,10 (B)	0.680	0.079
3,4,5,·,9,10	0.549	0.064
3,4,5,6,·,10 (D*)	0.631	0.076
3,4,5,6,9,· (C)	0.648	0.092

Finally, we present the results of the qualitative power index analysis for this example in Table 9. In terms of NRE, the unique optimal set is $\{3,4,6,10\}$ and there-

Table 9. Analysis of the duodenal ulcer data, IV

Attribute	φ_Q	Due to attribute set
3	0.1006	3,4,6,10
4	0.1006	3,4,6,10
5	0.0982	3,4,5,9
6	0.1006	3,4,6,10
9	0.0947	3,4,9
10	0.1006	3,4,6,10

fore, by construction of φ_Q in Section 4.2, we result in $\varphi_Q(3) = \varphi_Q(4) = \varphi_Q(6) = \varphi_Q(10) = NRE(\{3,4,6,10\})$. For the elements 5 and 9 there exist two further unique conditional optimal sets, which are used to determine their qualitative power indices.

To sum up, we conclude that set C seems to be the optimal choice: It shows an admissible approximation quality, its usefulness is near the optimal value, it is significant, and its complexity is close to the optimal value as well.

6 Conclusion

The paper demonstrates that there is more than one way to evaluate a non-numeric model, but it shows as well that an "anything goes" approach does not work: Forming simple differences or averages poses problems for interpretation and the Choquet-type aggregation schemes will achieve strange results under certain circumstances. This does not mean that the such approaches will not work most of the time, but there is no guarantee that they do. It is notable, that Banzhaf [1] begins his analysis of voting schemes with exactly the same ideas. The replacement of quantitative indices by their qualitative counterparts is a cure, but the results of these qualitative indices are not overwhelming: They are simply pointers to maximal values of some basic evaluation function (γ, NRE) and most of the results are achieved without the need of an extra theory, simply by reading out the results of the optimisation of the basic evaluation function.

Because there are several indices, one has to find a guideline when and how these indices should be applied. The examples demonstrate that in RSDA a reasonable start of evaluation is the inspection of the approximation quality, because it is very easy to set a first restriction for a good model. A further restriction can be set by comparing the usefulness with the given standards of effect sizes – models with very low effect sizes ($\lesssim 0.1$) have to be excluded, and the final model should be not too far away from the maximum effect size. The approximation quality of the final model must be significant, and its complexity (NRE) should be nearly optimal. The examples demonstrate as well, that these four qualities need not be present in one model. Furthermore, there need not even be a successful combination at all: We have shown that the indices may dissociate, because they are looking only at partially overlapping features of a model. It may happen that approximation quality is high, but either usefulness is very low or significance is lacking; in such cases, the data do not vote for using RSDA – and this is a fair result as well.

Bibliography

[1] Banzhaf, J. F. (1965). Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317–343.

[2] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum, Hillsdale, NY.

[3] Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45:1304–1312.

[4] Dubey, P. (1975). On the uniqieness of the Shapley value. *International Journal of Game Theory*, 4:131–139.

[5] Dubey, P. and Shapley, L. (1979). Mathematical properties of the Banzhaf power index. *Mathematics of Operations Research*, 4(2):99–131.

[6] Dubois, D., Grabisch, M., Modave, F., and Prade, H. (1997). Relating decision under uncertainty and multicriteria decision making models. Technical report, IRIT-CNRS Université P. Sabatier, Toulouse Cedex.

[7] Düntsch, I. and Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, 46:589–604.

[8] Düntsch, I. and Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence*, 106(1):77–107.

[9] Düntsch, I. and Gediga, G. (2000). *Rough set data analysis: A road to non-invasive knowledge discovery*, volume 2 of *Methoδos Primers*. Methoδos Publishers (UK), Bangor.

[10] Gediga, G. and Düntsch, I. (2001). Rough approximation quality revisited. *Artificial Intelligence*, 132:219–234.

[11] Grabisch, M. (1996). The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89:445–456.

[12] Grabisch, M. (1997). k-additive and k-decomposable measures. Proceedings of the Linz Seminar.

[13] Greco, S., Matarazzo, B., and Słowinski (1998). Fuzzy measure technique for rough set analysis. In Zimmermann, H.-J., editor, *Proceedings of the 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98)*, pages 99–103, Aachen. Mainz Wissenschaftsverlag.

[14] Greco, S., Matarazzo, B., and Słowinski, R. (2001). Rough sets theory for multicriteria decision analysis. *European Journal of Operational Research*, 129:1–47.

[15] Hildebrand, D., Laing, J., and Rosenthal, H. (1974). Prediction logic and quasi-independence in empirical evaluation of formal theory. *Journal of the Mathematical Sociology*, 3:197–209.

[16] Komorowski, J., Pawlak, Z., Polkowski, L., and Skowron, A. (1999). Rough sets: A tutorial. In Pal, S. and Skowron, A., editors, *Rough Fuzzy Hybridization*, pages 3–98. Springer–Verlag.

[17] Laruelle, A. and Valenciano, F. (2001). Shapley-Shubik and Banzhaf indices revisited. *Mathematics of Operations Research*. To appear.

[18] Marichal, J.-L. (2000). An axiomatic approach of the discrete choquet integral as a tool to aggregate interacting criteria. *IEEE Transactions on Fuzzy Systems*, 8(6).

[19] Marichal, J.-L. and Mathonet, P. (2000). On comparison meaningfulness of aggregation function. *Journal of Mathematical Psychology*, 45:213–223.

[20] Miller, F. (1966). Computer study into the causes of 1965-1966 traffic deaths in Jacksonville, Florida. Unpublished.

[21] Murofushi, T. and Soneda, S. (1993). Techniques for reading fuzzy measures iii: Interaction index (in japanese). In *Proc. 9th Fuzzy System Symposium*, pages 693–696. Sapporo, Japan.

[22] Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, 11:341–356.

[23] Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*, volume 9 of *System Theory, Knowledge Engineering and Problem Solving*. Kluwer, Dordrecht.

[24] Pawlak, Z. (1997). Rough set approach to knowledge-based decision support. *European Journal of Operational Research*, 99(1):48–57.

[25] Pawlak, Z., Słowiński, K., and Słowiński, R. (1986). Rough classification of patients after highly selective vagotomy for duodenal ulcer. *International Journal of Man-Machine Studies*, 24:413–433.

[26] Roth, A. E., editor (1976). *The Shapley value – Essays in honor of Llloyd S. Shapley*. CUP, Cambridge.

[27] Roubens, M. (1996). Interaction between criteria through the use of fuzzy measures. Technical Report 96.007, Institute de Mathématique, Université de Liège, Liège.

[28] Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

[29] Shubik, M. (1997). Game theory, complexity, and simplicity.

[30] Słowiński, K. (1992). Rough classification of HSV patients. In Słowiński, R., editor, *Intelligent decision support: Handbook of applications and advances of rough set theory*, volume 11 of *System Theory, Knowledge Engineering and Problem Solving*, pages 77–94. Kluwer, Dordrecht.

[31] Stepaniuk, J. (2000). Knowledge discovery by application of rough set models. In Polkowski, L., Tsumoto, S., and Lin, T. Y., editors, *Rough Set Methods and Applications*. Physika, Heidelberg.

[32] Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In Stevens, S. S., editor, *Handbook of Experimental Psychology*. Wiley, new York.

[33] Straffin, P. D. (1976). The Shapley – Shubik and Banzhaf power indices as probabilities. In [26], pages 71–81.

[34] Vogel, F. (1975). *Probleme und Verfahren der numerischen Klassifikation*. Vandenhoeck & Ruprecht, Göttingen.

[35] Wolpert, D. H. and Macready, W. G. (1995). No free lunch theorems for search. Technical Report SFI-TR-95-02-010, Santa Fe Institute, Santa Fe, NM.

[36] Zadeh, L. A. (1999). From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. *IEEE Transactions in Circuits and Systems*, 45(1):105–119.