Darmstadt Discussion Papers in Economics

rough set methodology in meta-analysis a comparative and exploratory analysis

Thomas Rupp

Nr. 157

Arbeitspapiere
des Instituts für Volkswirtschaftslehre
Technische Universität Darmstadt

Applied Research in Economics

rough set methodology in meta-analysis

a comparative and exploratory analysis

Thomas Rupp*

Keywords: rough data set, RSDA, meta analysis, data mining,

pattern recognition, deterrence, criminometrics

K14, K42, C49 (EconLit Classification), 62-07 (AMS 1991 Classification)

November 25, 2005

We study the applicability of the pattern recognition methodology *rough set data analysis* (RSDA) in the field of meta analysis.

We give a summary of the mathematical and statistical background and then proceed to an application of the theory to a meta analysis of empirical studies dealing with the deterrent effect introduced by Becker and Ehrlich. Results are compared with a previously devised meta regression analysis.

We find that the RSDA can be used to discover information overlooked by other methods, to preprocess the data for further studying and to strengthen results previously found by other methods.

*Department of Applied Economics and Microeconometrics, Institute of Economics
Darmstadt University of Technology
Tel.: +49 (0)6151 16-5512, Fax: +49 (0)6151 16-5652, email: rupp@vwl.tu-darmstadt.de

Contents

1 Introduction

Up to now several hundreds of empirical studies[1] have been conducted to explore the relationship between crime and deterrence either explicitly or implicitly. To this date no unambiguous answer can be given to the question whether and by how much raising the probability or severance of punishment reduces crime.
Only a few literature surveys (for example Nagin (1978) and Cameron (1988)) tried to summarize the current knowledge of their time and still fewer studies (Antony and Entorf (2003); Müller (2003)) performed a meta analysis. Basically a meta analysis is a statistical analysis of the literature about a specific topic (refer to Stanley (2001), who also describes the basic procedure of a meta (regression) analysis.). Its goal is to accumulate all existing knowledge about a specific subject and then to statistically extract sound conclusions from it, especially concerning the reliability of certain effects.

In the beginning meta analysis was used (particulary in medical research) to reevaluate various treatment effects. In the meantime new components and goals entered the vicinity of meta analysis and its esteem in the field of economics rises steadily (economic examples, beside many others, are Stanley (2001); Baaijens and Nijkampf (2000); Weichselbumer and Winter-Ebmer (2005) and Rose (2004)). Aside from the examination of presumed effects it has become more important to empirically understand which factors are responsible for the existing variety of outcomes.
To our knowledge no thorough meta analysis has been published in the field of criminometrics. Antony and Entorf (2003) reported only descriptive statistics since their data base was too small for more profound methods while Müller (2003) performed a simple meta regression analysis but was not published. Both found evidence of a deterrent effect and discovered some key relationships – like the decreasing significance as the number of socioecomic explanatory variables rises or whether the simultaneity problem was accounted for – between study characteristics and the outcome of a study.

Although meta analysis has also drawbacks like the handling of "publication bias" (insignificant studies tend to be not published and therefore might be biased), omitted variable bias and the incorporation of different levels of quality of the underlying studies. But even with these problems, meta analysis is still a prominent idea to strengthen or refute existing theories, find unknown or only suspected connections between attributes and to summarize the existing knowledge on a robust basis at a low cost.
It should be helpful to apply different methods in a meta analysis to find strong and robust[2] results.

[1] A list of these studies, covering the fields of economic, sociology, criminology and others, is available upon request.
[2] in the sense that the results are not only independent of changes in parameters but also do

In this paper we present a detailed introduction of such a methodology to the field of economics with an application as a meta analysis: the rough set data analysis (RSDA). RSDA is, in principal, a non parametric pattern recognition methodology which operates without any assumptions about the data. Its theory was developed by (Pawlak, 1982) and is generating an increasing amount of literature (like the further development of the theory as in (Slezak and Ziarko, 2003), as well as its applications – refer to Düntsch and Gediga (2000) for such studies). Although it does not produce numerical coefficients, it should still be very useful to study the interaction of data on a nominal level. Additionally it can be utilized as a pre filter for other methods and to strengthen results found by other statistical procedures.
After presenting the required mathematical foundation and the statistical methods we use the RSDA on a data set which comprises 30 criminometric studies about the deterrent effect[3].

Each study provides many of explanatory variables, ranging from explicitly used variables in a study, to implicitly given ones due to the study design itself[4] (such as the author himself, his origin, the kind of paper, samplesizes, methods used, etc.). Most of these variables as well as the dependent variable will be nominal however, and hence standard regression procedures (linear, logit, etc.) are either not suited very well for this problem or require assumptions which won't most probably be met, not to mention the problem of missing values.
We show that the existing methods and our extensions are useful tools to find and describe various significant influences of variables on the outcome of a study and to judge their importance.

We compare our results with a linear regression analysis which was done with the aforementioned data set. Besides replicating many of the results, the most prominent results are that there exists a lot of contradictory information which indicates that there are no dominating key characteristics and that the distinction between property and violent crime is important but does not influence the outcome of a study on its own.
In contrast to results from the linear regression we find that the importance of the distinction between youth- and adult crime is artificial and that publication year of a study is of non-linear importance.

The rest of this paper is arranged in the following way: section 2 gives the mathematical basis of the RSDA and some key characteristics. In section 3 and 4 we present the statistical methods and a more practical extension of the rough

not depend on the kind of analysis itself

[3] In fact this study is meant as a proof of concept, paving the way for an extensive criminometric meta analysis of more than 700 empirical studies. The data set used here was provided by Müller (2003)

[4] The meta analysis we will perform on the existing empirical studies will consist of more than 300 such variables.

set methodology.
We will conclude in section 5 with a thorough application of the presented methods to a criminometric data set which has been subject to a meta regression analysis, and then end with the conclusion in section 6.

2 Rough Set data analysis

The principles of rough set data analysis were introduced and extended by Pawlak (1982, 1983) and is used for classification (pattern recognition, artificial intelligence) in a variety of fields of science (computer science, medicine, automation, psychology, etc.); refer to Düntsch and Gediga (2000) for a large list of references.

Some economists have already used some methods of the RSDA (e.g. Baaijens and Nijkampf (2000) in a meta analysis) but only to a limited extent – far from exhausting its potential.

2.1 Principles

One important property of RSDA is it's complete lack of assumptions about the data. Only the structure given by the observed data is used. Therefore it can be used with almost any kind of data, especially when little is known about the distribution of the data or when the required assumptions of other methods are not sufficiently met.

Let us assume we have a set of objects, each object possessing values in the same set of attributes (a missing value is also a value). Let one attribute be marked as a decision attribute (the equivalent to the endogenous variable). The question is: given the values of all other attributes of an object, which value should the decision attribute of this object have and is this classification correct?
Instead of one attribute an arbitrary number of attributes can be simultaneously chosen as decision attributes.

The basic idea is to describe the data with rough sets. Each such a set consists of objects which share the same values in respect to a specific set of attributes – they are regarded to be in one equivalence class and are therefore indistinguishable. Each set contains a specific property certainly, certainly not or perhaps. The key lies in the recognition of the most important patterns in the data set, if they exist at all, which can characterize the data sufficiently well.

The primary aims are the elimination of redundancy (distilling the relevant information), the finding of small sets of relevant attributes (extracting the most important characteristics) and the generation of classifying rules (learn how these characteristics influence the decision attribute(s)).
Basically RSDA is, as pointed out by Düntsch and Gediga (1998), a reversal of the standard statistical procedures. While standard methods decribe uncertainty and

often add variables to distinguish the influencing variables from the insignifcant ones, RSDA describes redundancy and reduces all available information to reducts and cores (which will be defined further below).

2.2 Mathematical basis

In the following we use Düntsch and Gediga (2000); Slezak and Ziarko (2003); Pawlak (1983) and Yin *et al.* (2001) to present the basics of the rough set theory.

Definition 2.1 (information system) *An information system*

$$I =< U, A, V_a, f_a >_{a \in A}$$

consists of

1. *a finite set of objects* $U = \{x_1, x_2, \ldots, x_n\}$,
2. *a finite set of attributes* $A = \{a_1, a_2, \ldots, a_m\}$,
3. *for every attribute* $a \in A$ *exists*
   - *a set* $V_a = \{v_{a_1}, v_{a_2}, \ldots, v_{a_s}\}$ *of values,*
   - *an information function* $f_a : U \to V_a, \quad x_i \mapsto v_{a_j}$.

Definition 2.2 (decision system) *A decision system is an information system whose set of attributes can be disjointly split in a set of state- (*$\{A \backslash D\}$*) and decision attributes (*$\{D\}$*).*

Example

U	a_1	a_2	a_3	d
x_1	red	0	x	3
x_2	green	1	y	1
x_3	green	1	y	1
x_4	red	1	y	0
x_5	red	1	y	0
x_6	blue	0	x	1
x_7	green	1	y	1
x_8	blue	0	x	1
x_9	blue	0	y	2
x_{10}	red	0	x	3

This example is derived from Pawlak (1991) and is referred to throughout this paper. With $U = \{x_1, x_2, \ldots, x_{10}\}$, $A = \{a_1, a_2, a_3, d\}$, $V_{a_1} = \{\text{green, red, blue}\}$,$V_{a_2} = \{0, 1\}$, $V_{a_3} = \{x, y\}$ and $V_d = \{0, 1, 2, 3\}$, a decision system looks like the table on the left.

Of course the values of a_1, a_3 could be recoded into numerical values since they are all interpreted as nominal values.

■

The already mentioned equivalence classes are given by the

Definition 2.3 (Equivalence relation θ_Q) *For every subset $Q \subseteq A$ we associate an equivalence relation of* ***indistinguishability in respect to*** *Q θ_Q on U:*

$$\forall x, y \in U : \quad x\theta_Q y \Leftrightarrow \big(\forall a \in Q : f_a(x) = f_a(y)\big).$$

If $x\theta_Q y$ then x and y are both in the same equivalence class: $x \equiv y \mod Q$. These residue classes resemble the granularity[5] of the data.

Definition 2.4 (elementary classes) *Each subset $Q \subseteq A$ induces a partition on U. Its elements E_Q consist of those objects of U which possess the same values in regard to Q:*

$$U = \{E_{Q_1}, E_{Q_2}, \dots, E_{Q_m}\}, \forall i \neq j : E_{Q_i} \cap E_{Q_j} = \emptyset \text{ and } x\theta_Q y \Leftrightarrow \exists_1 i : x, y \in E_{Q_i}$$

Hence all elements of each elementary class can't be distinguished and are in the same equivalence class U*modulo* Q.
We omit the index Q if $Q = A$ or Q is understood and get

$$\forall x \in U \exists_1 i : x \in E_i = \{y \in U : x\theta y\}.$$

We abbreviate $\{y \in U : x\theta y\}$ by θx.

The principle of the rough set theory can be derived from these definitions: all our knowledge about the given data is represented via the elementary sets. Only these sets are distinguishable, not the objects contained in each set (in fact, according to Düntsch and Gediga (2001), the objects are randomly distributed in these sets).

Example In our example the elementary sets $E_i \in E_A$ are $E_1 = \{x_1, x_{10}\}, E_2 = \{x_2, x_3, x_7\}, E_3 = \{x_4, x_5\}, E_4 = \{x_6, x_8\}$ and $E_5 = \{x_9\}$. For example x_1 and x_{10} are equivalent in respect to A (e.g. $x_1 \equiv x_{10} \mod A$).

■

Definition 2.5 (Upper and lower approximation) *For a subset $X \subseteq U$ we define*

1. *the lower approximation $\underline{X} = \bigcup\{\theta x : \theta x \subseteq X\}$,*
2. *the upper approximation $\overline{X} = \bigcup\{\theta x : x \in X\}$.*

X is definable (i.e. the sum of elementary classes) iff $\underline{X} = \overline{X}$.

[5] more classes lead to a finer structure, i.e. more granularity

The lower approximation is the sum of all elementary classes which are completely contained in X, while the upper approximation is the sum of all elementary classes containing at least one element of X.
A *rough set* is the set $(\underline{X}, \overline{X})$; the knowledge about the certain $(\underline{X} \cup -\overline{X})$ and uncertain $(\overline{X} - \underline{X})$ area of classification (in regard to X) of an object[6] – hence the term *rough set data analysis*.
Example Let X_1 be $\{x_1, x_2, x_3, x_7, x_{10}\}$, $X_2 = \{x_1, x_2, x_9\}$ and $D = d$.
Then $\underline{X_1} = \overline{X_1} = E_1 \cup E_2$ and X_1 is therefore definable.
X_2 is not definable because $\underline{X_2} = E_5 \neq \overline{X_2} = E_1 \cup E_2 \cup E_5$.

■

If $X \subseteq U$ is, for example, determined[7] by the attribute d, we can conclude for every $x \in U$:

- if $x \in \underline{X_i}$, x has **certainly** the value d_i;
- if $x \in \overline{X_i} \backslash \underline{X_i}$, x has **perhaps** the value d_i;
- if $x \in U \backslash \overline{X_i}$, x has **certainly not** the value d_i.

A reduct of Q is a locally minimal choice of attributes from Q which generates the same equivalence classes as Q.

Definition 2.6 (reduct) *Let $P \subseteq Q \subseteq A$.*
P is a Q-reduct iff $\theta_P = \theta_Q$ and $\forall R \subsetneq P : \theta_R \neq \theta_Q$.
P is a X-reduct iff X is defined by P but by no true subset of P.
The set of all reducts of Q is Red(Q).

As long as understood we omit the prefix of the reduct.
Given a reduct all of its elements are indispensable for classification.

Definition 2.7 (core(Q)) *The core of Q is the intersection of all Q-reducts.*

core(Q) is unique and normally no reduct (unless there's only one reduct) and all its attributes are indispensable for a classification in regard to Q.
Example Let Q be $\{A \backslash d\}$. The induced partition is identical to the elementary classes E_A. The only Q-reduct is $\{a_1, a_3\}$ because the equivalence relation induced by $\{a_1, a_2\}$ cannot differentiate E_4 from E_5 and the equivalence relation induced by a_2, a_3 merges the partitions E_2 and E_3.
The core of Q consists therefore of a_1 and a_3 which are essential to generate the same partitions as the whole set Q – a_2 is unnecessary.

■

[6] whether $x \in U$ belongs to X or not
[7] e.g. $X = \{X_1, \ldots, X_m\}$ and $X_i = \{x \in U | f_d(x) = d_i\}$

2.3 Discretesizing and filtering data

Since the RSDA is a pattern recognition it relies heavily on repeated patterns in the data. Continues data can lead to too many classes (in the worst case every value and therefore every pattern is unique) it is reasonable to discretize the data before doing the analysis.
We have different possibilities at our disposal:

- Gather all values which do not change the equivalence classes. This may lead to big changes in the values of the attributes but does not influence the analysis since the patterns in the data remain the same. The advantage lies in keeping the structure of the data unchanged, the disadvantage is that they do not simplify the algorithm and could be hard to interpret although they may lead to new conclusions at the end of the analysis.

- Gather all values which seem reasonable. This relies on the knowledge of the analyst and introduces some external influence into the data but is a good method to condense the number of the values of an attribute. A simple example is to code a metrical attribute into "small","medium" and "large" by the choice of the analyst.

- Create a number of new classes which either contain the same amount of values or, if possible, each class represents some common measure.

  This is less arbitrary than the previous method but the resulting categories may be a bit harder to interpret. For example, instead of arbitrarily choosing the three classes, these could be chosen in such a way that each class contains the same amount of objects.

Additionally the number of attributes may be reduced by removing superfluous attributes (either all values are the same or, properly encoded, are equal to another attribute). These and further possibilities are discussed by Düntsch and Gediga (2001).

2.4 Reduction of attributes and deduction of decision rules

In regard to size every minimal reduct is a best choice to reduce the set of attributes. Therefore the set of minimal reducts is a natural choice to reduce the amount of attributes to be used in the further analysis.
Since the problem of locating all reducts is NP-hard (Skowron and Stepaniuk, 1991) and thus has an exponential worst-case runtime, this may pose a serious problem in any analysis with numerous attributes.

With d as a decision attribute we may deduct classification rules of the form
If $f_{a_1}(x) = v_1$ and $\dots$ and $f_{a_t}(x) = v_t$,
then $f_d(x) = b_1$ or $\dots$ or $f_d(x) = b_j$, for a $j > 0$.

In the case of $j = 1$ we call a rule *deterministic*. A set of rules is written as[8] $Q \rightarrow D$. Therefore a minimal X-reduct can be interpreted as a set of classification rules with minimal size. The goal is to find all minimal reducts with the "best" ("best" can be interpreted in various ways and will be specified later) ability to classify.

The *support* of a rule is the number of objects which meet the rule; a rule is *deterministic casual* if it has a support of one (it seems not suited to predict other objects since it appears only once in the data).

Example

In our example exist four A-reducts: $\{a_1, d\}$, $\{a_1, a_3\}$, $\{a_2, d\}$ and $\{a_3, d\}$. The core is empty. In each case two attributes are sufficient for the classification (of the five elementary sets).

If we are interested in the classification of d we must consider the unique $\{A \backslash d\}$-reduct $\{a_1, a_3\}$.

a_1	$\wedge$	a_3	$\Rightarrow$	d
r	$\wedge$	x	$\Rightarrow$	3
g			$\Rightarrow$	1
r	$\wedge$	y	$\Rightarrow$	0
b	$\wedge$	x	$\Rightarrow$	1
b	$\wedge$	y	$\Rightarrow$	2

Therefore only the two attributes a_1, a_3 are required for the classification of d. The deducted rulesystem is given on the right side.

■

2.5 Choice of relevant attributes

There can be many different (minimal) reducts and therefore we need methods to measure the relevance of such sets of attributes.

First of all we keep all attributes in the core since they are all indespensable for classification. But often this is not enough since the core may be very small or even empty.

2.5.1 dynamic reducts

One possibility to determine important sets is the usage of dynamic reducts. The idea (refer to Düntsch and Gediga (2000)) is to delete subsets of U randomly and then recalculate all reducts. These reducts which reappear considerable more often than others are considered to be more stable and therefore more important. Let $\mathbf{F}$ be the set of all subsets of A. The stability coefficients SC can be interpreted relatively to each other; they are defined by

$$SC_Q := \frac{|\{J \in \mathbf{F} : Q \in \text{Red(J)}\}|}{|\mathbf{F}|}.$$

[8] and consists of the rules $(X, Y) \in (\theta_Q, \theta_D)$ with $X \cap Y \neq \emptyset$ and $Q \cap D = \emptyset$. All elements of class X are mapped to a class Y. A rule is deterministic if $X \subseteq Y$.

Those reducts with a high SC are more robust and therefore more trustworthy.

2.5.2 wnf-benchmark

Furthermore we propose to use an additional benchmark for the individual attributes by calculating the **w**eighted and **n**ormalized **f**requency **wnf** for every attribute in a reduct:

$$\mathrm{wnf}_a := \frac{\mathrm{t}_a - \min\limits_{a\in A}(\mathrm{t}_a)}{\max\limits_{a\in A}(\mathrm{t}_a) - \min\limits_{a\in A}(\mathrm{t}_a)} \text{ with } \mathrm{t}_a := \sum_{Q\in\mathrm{Red(A)}} 1_{[a\in Q]} SC_Q.$$

Other methods are plain considerations of the frequencies of attributes in reducts (section 3.1.3), sets of attributes which share a special relation (see section 3.4) or sets of attributes with good entropy-characteristics (refer to section 3.2).

Although rough set methodology is based on symbolic characteristics several numerical functions exist.

2.5.3 Approximation functions

Let $\mathcal{P}$ be a partition of U. The function

$$\gamma_Q(\mathcal{P}) = \frac{\sum_{X\in\mathcal{P}} |\underline{X}_{\theta_Q}|}{|U|}$$

tells us, how much certain knowledge exists of $\mathcal{P}$ through Q.
In the case of $\gamma_Q(\mathcal{P}) = 1$ all $X \in \mathcal{P}$ are definable by Q and all rules are deterministic.

For $\mathcal{P} = \{X, -X\}$ is $\gamma_Q(\{X, -X\}) = \gamma_Q(X) = \frac{|\underline{X}|+|\underline{-X}|}{|U|}$; the global knowledge about X through Q.

Let Q be a reduct of P and $\mathcal{P}$ be the partition induced by P, writing $\gamma_Q(P)$ instead of $\gamma_Q(\mathcal{P})$. One measure of the importance of an attribute is the drop from $\gamma_Q(P)$ to $\gamma_{Q\setminus\{q\}}(P)$ – the importance of q grows with this difference.

The local knowledge – how much is known about X through Q – of a set X is measured by:

$$\alpha_Q(X) = \frac{|\underline{X}|}{|\overline{X}|}.$$

Example The reduct $\{a_1, a_3\} \in A\backslash d$ in our example has two attributes. The partition $\mathcal{P}$ induced by $A\backslash d$ is equal to E_A (see 2.2). Since $\gamma_{\{a_1,a_3\}}(\mathcal{P}) = 1$ and

$$\gamma_{a_1}(\mathcal{P}) = 0.3 \text{ resp. } \gamma_{a_3}(\mathcal{P}) = 0$$

the attribute a_1 is most important while a_3 is also important but to a lesser extent. Without a_1 (e.g. with a_3 alone) no object can be classified (since $f_{a_3}(u) = x \Rightarrow f_d(u) = 1 \vee 3$ and $f_{a_3}(u) \neq x \Rightarrow f_d(u) = 0 \vee 1 \vee 2$). Whereas a_1 can still classify the objects x_2, x_3, x_7 on its own.

Since $\underline{X_1} = \overline{X_1}$ we know X_1 perfectly: $\alpha_A(X_1) = 1$ – each object can be assigned correctly, whether it belongs to X_1 or not.
However $\alpha_A(X_2) = \frac{1}{6}$ – the area of lower approximation covers just 16.67% of the upper approximation.

■

We have to stress that such functions cannot be used on their own to compare different sets of attributes (refer to Düntsch and Gediga (2001) or section 3.2 for a function which can) since they ignore the complexity[9] of Q.

2.5.4 Jackknife

A robust classification shouldn't suffer much when one object is left out before learning the rules.
We leave one object out and relearn the rules and then classify all objects again (if more than one rule can be applied we guess the rule – if no rule fits an object it cannot be assigned to any category). We record two values: j_a is the fraction of correctly classified objects and j_w is the fraction of correctly classified objects when all guesses go astray.

Example

	a_1	a_2	d
x_1	0	1	1
x_2	1	1	0
x_3	0	0	1

If we leave x_1 out, a_1 or a_2 is superfluous and the rules are either $a_1 = 1 \Rightarrow d = 0, a_1 = 0 \Rightarrow d = 1$ or $a_2 = 1 \Rightarrow d = 0, a_2 = 0 \Rightarrow d = 1$. Hence x_1 will be correctly classified if, with probability 0.5, a_1 and therefore rule $a_1 = 0 \Rightarrow d = 1$ is chosen.
If x_2 is left out, the rule is $a_1 = 0 \Rightarrow d = 1$ and x_2 is not classified because $a_1 = 1$ cannot be assigned.
If we leave x_3 out the rules are $a_1 = 0 \Rightarrow d = 1, a_1 = 1 \Rightarrow d = 0$ and x_3 will be correctly assigned to 1.
In the worst case we will always guess wrong and therefore

$$j_a = \frac{1}{3}\left(\frac{1}{2} \cdot 1 + 0 + 1\right) = \frac{1}{2} \text{ and } j_w = \frac{1}{3}\left(\frac{1}{2} \cdot 0 + 0 + 1\right) = \frac{1}{3}.$$

If we had analyzed the reduct a_1 we would have gotten $j_a = j_w = \frac{2}{3}$.

■

[9] If we have one rule for every object we have perfect classification but still gained nothing.

3 Statistical methods

Besides quantitative measures various statistical tools can be derived.

3.1 statistical significance of decision rules

Düntsch and Gediga (1997) proposed a permutation test to decide whether a given[10] set of decision rules $Q \rightarrow P$ should be considered significant.

Let H_0 be the hypothesis that the values of Q and P (normally P consists of one attribute d) are distributed randomly on Q and P (i.e. the rules are completely random, containing no information). Define

$$f_a^{\sigma,Q}(x) = \begin{cases} f_a(\sigma(x)) & \text{if } a \in Q, \\ f_a(x) & \text{else.} \end{cases}$$

With σ being a random permutation of U. Hence, the function $f_a^{\sigma,Q}(x)$ delivers a random value of V_a if a is in Q – otherwise it delivers the proper value $f_a(x)$. In other words: all the values of the attributes in Q are chosen randomly according to their frequencies..

We can now calculate (with $f_a^{\sigma,Q}$ instead of f_a) the approximations $\gamma_{\sigma,Q}(P)$ for all permutations and look whether the fraction of those approximations, which are better than the original, is above a significance level α or not. If it is below α we can reject H_0.

3.1.1 Assessment of attributes

In section 2.5.3 we have seen how we can measure the importance of an individual attribute by the drop of its associated γ-value. This drop is not necessarily enough since it could be coincidental.

Let Q be a reduct since all superfluous attributes have no γ drop at all.

Düntsch and Gediga (1997) call an attribute $q \in Q$ *conditional casual*, if the attribute is hardly necessary to predict P (i.e. its γ-drop is casual).

Let $\gamma_{Q,\sigma,q}(P)$ be the usual approximation function with

$$f_a(x) = f_a^{Q,\sigma,q}(x) = \begin{cases} f_a(\sigma(x)) & \text{if } a = q, \\ f_a(x) & \text{else.} \end{cases}$$

In this case only the values of attribute q are permutated.

If the fraction of $\gamma_{Q,\sigma,q}(P) \geq \gamma_Q(P)$ is smaller than α we can reject the hypothesis that q is *conditional casual*.

Example We assume that in our example only a_1, a_3, d exist. The reduct of $\{A \backslash d\}$ is $Q = \{a_1, a_3\}$. Since all objects can be classified $\gamma_Q(P) = 1$.

[10]not derived from any optimization of classification; for such cases see 3.1.1 below

We now permutate all values of the attributes a_1 and a_3 and see that the fraction of the improved (or equal) γ-values is 0.005. If we permutate only a_1 it's 0.05 and in the case of a_3 it is 0.005.
Therefore we can reject all three hypotheses (both are casual, or one of them is conditional casual) on a 5% level (two of them even on a 0.5% level).

■

In the case of already optimized sets Düntsch and Gediga (1997) present a slight modification of 3.1.1: mark randomly half of the data (i.e. introducing a new binary attribute) and test whether this marking is conditional casual. If it is "the hypothesis that the rules in both sets of objects are identical should be kept".

3.1.2 Validating decision rules

A common method, which is also implemented in GROBIAN (1999), is the following bootstrapping method:
Chose randomly $n' < n$ objects and use them to calculate all rules. With these rules determine the prediction quality of all objects[11]. Repeat this sufficiently often to determine an average prediction quality.

To assess the calculated average prediction quality we compare it to its lower bound (mere guessing) using the frequencies[12] p_i of the values of d:

$$1 - \sum_i p_i(1 - p_i) = \sum_i p_i^2.$$

3.1.3 Binomial-Test

Even if we know all (minimal) reducts we have to decide on the one hand which reducts will be further investigated and on the other hand which attributes deserve more attention. Since in many cases there are too many (minimal) reducts to evaluate we have to provide a method to assign priorities. For the former problem we have the dynamic reducts, for the latter we propose the wnf-benchmark and the following test.
A very simple method to assign priorities of attributes in reducts of equal length[13] is to look if they appear more often in those reducts than others.

We consider the hypothesis[14] that all attributes are of equal importance and therefore should be equally represented.

[11] the fraction of all objects which can be classified correctly; either by a deterministic rule or by guessing

[12] i.e. $|\{x \in U : f_d(x) = v_i\}|/|U|$

[13] Since we will be interested in the smallest reducts we have to discern different lengths of reducts; otherwise we could overestimate attributes stemming from reducts of lesser interest.

[14] Of course this hypothesis can be rejected since we know exactly how often any attribute appears – we're solely interested in the deviation from this hypothesis.

Under this hypothesis the frequency H_i of an attribute A_i is binomially distributed with the parameters $(\frac{1}{m}, n)$. If the hypothesis can be rejected on a certain α-level ($b := P(X \leq H_i) > 1 - \frac{\alpha}{2}$ or $< \frac{\alpha}{2}$) we can be confident that A_i is of some importance.

Since we already know that the attributes of the core are indispensable those could dampen the results and the procedure should be repeated after removing these attributes from the set of reducts.

3.2 Entropy

The entropy of a dataset is a measure of the amount of information contained therein – the minimal number of bits required to describe the data (devised by Shannon (1948)).

Düntsch and Gediga (1998) carried this measure forward to the rough set methodology.
Let $\mathcal{P}$ be a partition of U with classes $M_i, i = 0, 1, \ldots, k$ of the size m_i. Since in each class the objects are randomly distributed the probability that a randomly drawn object is in class M_i is $\frac{|M_i|}{|U|}$. Therefore the entropy function is

$$H(\mathcal{P}) := -\sum_{i=0}^{k} \frac{m_i}{|U|} \cdot \log_2 \left(\frac{m_i}{|U|} \right).$$

If $\mathcal{P}$ is induced by θ we will write $H(\theta)$.

Düntsch and Gediga (1998) carried this measure forward to the rough set methodology.
Let $\mathcal{P}$ be a partition of U with classes $M_i, i = 0, 1, \ldots, k$ of the size m_i. Since in each class the objects are randomly distributed the probability that a randomly drawn object is in class M_i is $\frac{|M_i|}{|U|}$. Therefore the entropy function is

$$H(\mathcal{P}) := -\sum_{i=0}^{k} \frac{m_i}{|U|} \cdot \log_2 \left(\frac{m_i}{|U|} \right).$$

If $\mathcal{P}$ is induced by θ we will write $H(\theta)$.

H measures the granularity of the data. In the case that no information is contained in the data the entropy is maximal: $-\log_2 \left(\frac{1}{|U|} \right)$; when $\mathcal{P}$ the identity ($M_0 = U$) the entropy is zero.

3.2.1 NRE – A measure of comparison

With this entropy function Düntsch and Gediga (1998) define a measure which allows us to compare the classification quality of different sets of attributes by

mixing approximation qualities (minimal statistical uncertainty ($H(d|Q)$) and coding complexity ($H(Q)$).
They propose three different methods of which we use and recapitulate only the one closest to the RSDA principle: the only assumption needed is the representativeness of the classes θ_Q (i.e. the probability of a class is given by its size).
In principle we assume that these objects, which cannot be classified, bear no information at all and are therefore random, building a class on their own. Thus we have classes which either classify objects with certainty or by chance.

The required equivalence relation is: two objects are equivalent iff they are equal or are in the same residue class of Q and classify d with certainty.
Let $\{X_1, X_2, \ldots, X_c\} := \{X_i \in \theta_Q | \exists j : X_i \subseteq Y_j, Y_j \in \theta_d\}$. The equivalence relation is then $\theta^{\mathrm{det}} : x \equiv y \Leftrightarrow \exists i \leq c : x, y \in X_i$ or $x = y$.
The distribution of the generated classes is

$$\hat{\psi}_i := \begin{cases} \hat{\pi}_i = \frac{|X_i|}{|U|} & \text{if } i \leq c, \\ \frac{1}{n} & \text{else.} \end{cases}$$

The *entropy of deterministic rough prediction* is defined as

$$\begin{aligned} H^{\mathrm{det}}(Q \to d) &:= H\left(\theta^{\mathrm{det}}\right) = -\sum_i \hat{\psi}_i \log_2(\hat{\psi}_i) \\ &= -\sum_{i \leq c} \hat{\pi}_i \log_2(\hat{\pi}_i) + |U \setminus \bigcup_{i \leq c} X_i| \frac{\log_2(n)}{n} \\ &= \underbrace{-\sum_{i \leq c} \hat{\pi}_i \log_2(\hat{\pi}_i)}_{\text{certain classification}} + \underbrace{(1 - \gamma(Q)) \log_2(n)}_{\text{random guessing}}. \end{aligned}$$

The **NRE** (normalized rough entropy) $S(\cdot)$ is then defined by

$$S^{\mathrm{det}}(Q \to d) = S(Q) = 1 - \frac{H^{\mathrm{det}}(Q \to d) - H(d)}{log_2(n) - H(d)} \in [0, 1]$$

which can be applied as a measure of comparison.
A low value indicates either randomness (low c and low $\gamma(Q)$) or low granularity[15] and therefore high coding complexity (high c with small $\hat{\pi}_i$ and low $1 - \gamma(Q)$). The larger the value the better is Q.

Example

[15]the induced equivalence relation is fine – the finest beeing the idenity

	A	B	C	D	d
	1	0	0	0	0
	2	0	0	0	0
	3	1	0	1	0
	4	2	0	1	1
	5	2	0	1	1
γ	1	1	0	0.4	–

$S(\{B,\cdot\}) = 0.592$ except $S(\{A,B,\cdot\}) = 0$. $S(D) = 0.296$. All other S are zero.
The null-vector is easy to code ($H(C) = 0$) but contains no information. A explains everything but θ_A is the identity (highest coding complexity). In respect to D both effects are mixed.
The attribute B is the best candidate – even D is still better than A although its γ is much lower.

■

3.3 Fisher's exact test

In the following we will present another method to gather attributes for further analysis. The presentation of the underlying principle of Fisher's exact test is derived from Tsumoto (2002) and Eibe and Witten (1998) and is applied in a similar way in the context of decision trees.
We must stress that at this stage we introduce no additional information about the objects at hand – we just make use of methods to reject hypotheses.

Let $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_m$ be values of the attributes A and B and $x_{ij} := |\{x \in U : f_A(x) = A_j \text{ and } f_B(x) = B_i\}|$. Its contingency table is shown in figure 1.

Figure 1: general contingency table

	A_1	A_2	$\cdots$	A_n	sum
B_1	x_{11}	x_{12}	$\cdots$	x_{1n}	b_1
B_2	x_{21}	x_{22}	$\cdots$	x_{2n}	b_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_m	x_{m1}	x_{m2}	$\cdots$	x_{mn}	b_m
sum	a_1	a_2	$\cdots$	a_n	N

According to Tsumoto (2002) the χ^2-approximation is not reliable in the case of small samplesizes or small cell-values and therefore we will rely on the exact probabilities.

With Fisher's exact test (also called Freeman-Halton-test in dimensions larger than two) we can test the hypothesis that the x_{ij} are distributed independently – i.e. whether the values of A (columns) do not depend on attribute B (rows). If the hypothesis is true they are distributed according to their marginal frequencies and the probability of the observed contingency table follows the multivariate hypergeometrical distribution (consult Martin (1995) for the proof):

Figure 2: contingency tables of the example

$a_1\backslash a_2$	0	1	
1	2	2	4
2	0	3	3
3	3	0	3
	5	5	10

$a_2\backslash a_3$	0	1	
0	4	1	5
1	0	5	5
	4	6	10

$a_1\backslash a_3$	0	1	
1	2	2	4
2	0	3	3
3	2	1	3
	4	6	10

$$p_x := P(x_{11}, x_{21}, \ldots, x_{mn} | a_1, a_2, \ldots, b_m) = \frac{\prod_{i=1}^{m} a_i! \prod_{j=1}^{n} b_j!}{N! \prod_{i=1}^{m} \prod_{j=1}^{n} x_{ij}!}.$$

Let p_0 be the probability of the contingency table under consideration. We will not use p_x as a statistic because Eibe and Witten (1998) show that it has a strong bias in favor of attributes with high values. The Freeman-Halton-test evaluates the statistic $p = \sum_T I_{[p_x \leq p_0]} p_x$ with $\sum_T$ being the sum of all contingency tables with equal marginal frequencies (i.e. the same column- and row-sums).
The hypothesis can be rejected on an α-level if p is below α since then there are barely any other contingency tables with equal marginal frequencies which are equally (or less) likely. In the case of bigger p's we cannot derive any conclusions about the dependance between the attributes as long as we do not know whether their values are distributed according to their marginal frequencies or not.

In 3.4 we will discuss how we can make use of this to gather attributes for further investigation.

Example The contingency tables of our example (with red = 1, green = 2 and blue = 3, as well as $x = 0, y = 1$) of our example are shown in figure 2.
The individual probabilities p_0 are 0.0238 resp. 0.0238 resp. 0.0875. The Freeman-Halton-statistics are $p = 0.095, p = 0.0476$ and $p = 0.4$.
Therefore we conclude that the attributes a_1 and a_3 might be independent while there is a statistical significant dependance between a_2 and a_3 (and to some extent between a_1 and a_2).

■

Since the calculation of all contingency tables with the same marginal frequencies has exponential runtime we use a Monte-Carlo-algorithm presented by Eibe and Witten (1998) which enables us to approximate the Freeman-Halton-statistics in most cases to an arbitrary precision in a very short time.

3.3.1 Contingency tables (c.t.)

We think the same procedure is applicable for the individual values of two attributes.
Consider again the contingency table in figure 1. Under the hypothesis H_0 that the individual values of the attributes are distributed according to their marginal frequencies we can describe these conditional probabilities[16] as

$$P(x_{ij}|f_A = i) = \text{dbinom}\left(x_{ij}, a_i, \frac{b_j}{N}\right) \text{ and } P(x_{ij}|f_B = j) = \text{dbinom}\left(x_{ij}, b_j, \frac{a_i}{N}\right).$$

Whereas $\text{dbinom}(x, n, p)$ is the density of the binomial distribution with x successes in n Bernoulli trails with probability p.
We reject H_0 when the p-value $\text{pbinom}(x, n, p)$ is below α in the case of $x \leq np$ or $(1 - \text{pbinom}(x - 1, n, p))$ is below α otherwise, while pbinom is the binomial cumulative distribution function.
Example

$A\backslash d$	$f_d = 0$	$f_d = 1$	$f_d = 2$	
$f_A = 0$	5	11	2	18
$f_A = 1$	1	2	6	9
	6	13	8	27

Given the hypothesis of independence between the attributes A and d the observed frequencies are distributed according to their marginal frequencies.

p-values given A

$A\backslash d$	0	1	2
0	0.351	0.139	0.020
1	0.351	0.139	0.020

p-values given d

$A\backslash d$	0	1	2
0	0.370	0.194	0.064
1	0.372	0.109	0.238

Given the attribute A the only significant observations ($\alpha = 0.05$) are 2 and 6 (its p-values are $\text{pbinom}(2, 8, 18/27) = 0.020 = 1 - \text{pbinom}(5, 8, 9/27)$).
In the reverse case however, given d, no significant observation remains.
Therefore we have always to keep in mind which direction of influence is subject to the analysis.

■

3.4 RFH-sets

In the following section we will call the matrice with the pairwise (testing attribute i against attribute j) Freeman-Halton-statistics p_{ij} the FH-matrice.

[16] with some abuse of language we write $f_A = i$; which means that we consider the objects x with $f_A(x) = v_{a_i}$

3.4.1 RFH$_1^A(\alpha)$-sets

A possible strategy to search promising attributes is to look in those sets (without the decision attribute) which show no significant dependency – sets with insignificant Freeman-Halton-statistics.
It is reasonable to believe that these attributes provide a wide range of information[17] while the other attributes contain more redundant information.

Definition 3.1 *Let I be an index function (i.e. an automorphism mapping $\{1, 2, \ldots, |A|\}$) on A.*
A set $R = \{r_{I(1)}, r_{I(2)}, \ldots, r_{I(c)}\} \subseteq A$ is a $RFH_1^A(\alpha)$-set iff $c \geq 2$ and every $i, j \in [1, c], i \neq j$ satisfies $p_{I(i)I(j)} \geq \alpha$.

In other words: each R consists of attributes which have, to the chosen α-level, insignificant pairwise Freeman-Halton-statistics.

After finding all RFH$_1^A(\alpha)$-sets we merge some of these since these sets are usually quite small. We merge those sets which contain the most frequent attribute and those with the second most frequent attribute.

Since this search process has obviously an unusable runtime we implemented a simple probabilistic algorithm which randomly explores the search space.

3.4.2 RFH$_0^A(\alpha)$-sets

Opposite to the definition just given the $p_{I(i)I(j)}$ must now be $< \alpha$.
Assume $d = a_m$. We are interested in those RFH$_0^A(\alpha)$-sets which contain d. The attributes in these sets *could* have an important linkage with d – the same reasoning applies to all individual attributes a_i with $p_{im} < \alpha$.
We emphasize *could* because attributes which often assume the same values will be recognized as significant although they contain little information and should be treated carefully.

4 Model extensions

Until now we implicitly dealt only with deterministic relations between attributes. As soon as noise, measurement errors, etc. exist, these effects will blur the relations and hamper our analysis in the rough set framework. To cope with this problem Slezak and Ziarko (2003) presents various extensions from which we will use one.
We must emphasize that at this moment we relax the basic principle of "let the data speak for itself" (Düntsch and Gediga, 2001) since parameters, terms not provided by the data, are introduced.

[17] Attributes with no information (i.e. random values) are included in those sets and must therefore, afterwards, be filtered out (finding reducts in those sets, the binomial test, etc.).

4.1 Assumptions

1. Each subset $X \subseteq U$ possesses a (uncertain) prior probability $0 < P(X) < 1$.
2. A conditional probability $P(X|E)$ with regard to an elementary set E can be assigned to each such a X.
3. For all X: $P(\neg X) = 1 - P(X)$.
4. $$P(X|E) = \frac{P(E|X)P(X)}{P(E)} \text{ and } P(E) = P(E|X)P(X) + P(E|\neg X)P(\neg X).$$

A possibility to estimate $P(X)$ and $P(E|X)$ is to use their fractions:

$$P(X) = \frac{|\{X\}|}{|\{U\}|} \text{ and } P(E|X) = \frac{|\{E \cap X\}|}{|\{X\}|}.$$

With these assumptions then follows

$$P(E) = \frac{|\{E\}|}{|\{U\}|} \text{ and } P(X|E) = \frac{|\{X \cap E\}|}{|\{E\}|}.$$

4.2 Variable Precision Rough Set Model (VPRS)

This parametric method allows us to classify an object up to a given precision.

Definition 4.1 (VPRS) *Chose l and u satisfying $0 \leq l < P(X) < u \leq 1$ and $Q \subseteq A$.*

1. $POS_u^Q(X) = \bigcup\{E \in E_Q : P(X|E) \geq u\}$
2. $NEG_l^Q(X) = \bigcup\{E \in E_Q : P(X|E) \leq l\}$
3. $BND_{l,u}^Q(X) = \bigcup\{E \in E_Q : l < P(X|E) < u\}$

We will omit Q if it is understood or $Q = A$.
The parameters u (resp. l) control the margin of the classification precision. They define the range where X is still regarded to be significantly influenced by E compared to the uninformative prior[18].
The VPRS-model is equivalent to the original model if $u = 1$ and $l = 0$ – because then $\mathrm{POS}_1(X) = \underline{X}$, $\mathrm{BND}_{0,1}(X) = \overline{X} \backslash \underline{X}$ and $\mathrm{NEG}_0(X) = U \backslash \overline{X}$.

[18] the equivalent to a random guess

5 Application

This paper is a preparation for a large-scale meta-analysis of empirical studies about the generally assumed deterrent effect, searching for factors which play a role in the outcome of a study.
The economic framework of the deterrence theory was first introduced by Becker (1968) and Ehrlich (1973) who interpreted criminal behavior as the result of an rational choice process weighting the expected benefit against the expected penalty.
The core result was that the probability of being convicted and the severance of the penalty have a negative influence on the willingness to commit a crime.

The methodology presented in this paper enables us to explore, reveal and verify the underlying factors in those studies. However we cannot judge whether the deterrent effect is present, and if, how strong it is. This will be subject to a numerical analysis.

We applied the methodology to a data set provided by Müller (2003); a meta regression analysis covering exactly the same subject on a smaller scale. The data set consists of 84 cases from 30 studies while we will be working with more than 700 studies.
In the following application we adopted only 74 data because 10 had many missing values of important attributes.
The attributes of this data set is listed in table 1.

5.1 Preprocessing the data

Hereunder are gathered all steps which are done before the main analysis can take place.

5.1.1 discretisizing the data

Since the variables Datats, Datapan and Datacs characterize the kind of datastructure we merge this variable back into one: Datatpc (1 for time series, 2 for panaldata and 3 for cross-sectional data).
Similarly we merge
{Dataviol, Dataprob, Dataall} → {Datavpa},
{Datanat, Datastate, Datacounty, Datamicro} → {Datasrc} and
{Datayear, Datamonth, Dataweek, Dataday} → {Datatime}.

Then we discretisize the date by creating groups which contain an equal amount of objects[19].

[19] In the case of manually created groups the results achieved remained the same or were only slightly clearer.

The only exceptions are NOEXVAR which is the maximum of the transformed variables NODETVAR and NOSOZVAR to maintain the relation (since it was the sum of these two).

The decision attribute d was created in the following way (a positive t-value, a two sided significance test of the respective deterrence hypothesis, stands for approval):

$$d = \begin{cases} 3 & \text{at least one t-value is} < -2, \\ 0 & \text{all t-values are} < 2, \\ 1 & \text{only } 1/3 \text{ to } 2/3 \text{ of all t-values are} < 2, \\ 2 & \text{no t-value is} < 2. \end{cases}$$

This should be more accurate than the simple average used by Müller (2003).

5.1.2 Datafilter

Since all reasonable candidates have already been merged no more attributes are superfluous or can be merged.

5.2 Analysis – choice of attributes

Let Q be $\{A \backslash d\}$.
Since we have reduced the number of attributes to twenty we can easily find all reducts in a short time[20].
Since $\max\limits_{Q' \subseteq Q} \gamma_{Q'}(d) = 0.527$ is smaller than one we have no *real* reducts and must be content with those which maximise γ.
Table 2 shows these reducts with minimal length or with a stability coefficient greater than 0.6, calculated by RSES (2004) with standard settings.
The *core*[21] contains DATAYOUTH (12) and DATAVPA (18).
This led to the characteristics of the attributes which are shown in table 3 (f is the frequency, b' and wnf' are the statistics calculated without the *core*).
Therefore we can temporarily omit the attributes $01, 04, 05, 08, 11, 16, 19, 20$ and mark the attributes $02, 03, 06, 07, 10, 12, 14, 17, 18$ as important.
This means that it is not necessary to consider the total number of explanatory variables and simultaneously their distinction between social and deterrence variables. Also it is not important whether a study considers unemployment or income. It also does not matter whether a study was published in the US or

[20] The runtime depends heavily upon the sizes of the reducts and the number of attributes. If the reducts can become small and we have more than 50 attributes the runtime may exceed weeks on a standard computer.

[21] all attributes which are indispensable for $\gamma(Q') = 0.527$; i.e. the intersection of all γ maximizing Q'

concentrates on deterrence, nor is its samplesize[22], the level of aggregation or the frequency of data collection of importance. The attributes and their numerical abbreviations are depicted in table 1.
Additionally it seems to be that attribute 15 is nearly as important as 06 or 07 – at least more important than 09.

5.2.1 other candidates

The $\mathrm{RFH}_1^{A\setminus d}(0.05)$-sets are

$\{7,11,15,18\}$	$\{6,14,18\}$	$\{1,11,18\}$	$\{6,7,18\}$	$\{4,6,8\}$	$\{7,20\}$
$\{7,12,15,18\}$	$\{6,16,18\}$	$\{1,5,18\}$	$\{6,8,18\}$	$\{10,16\}$	$\{8,10\}$
$\{5,6,17,18\}$	$\{7,10,12\}$	$\{3,5,10\}$	$\{2,6,18\}$	$\{12,13\}$	$\{9,14\}$
$\{6,7,11,18\}$	$\{7,10,19\}$	$\{3,5,18\}$	$\{4,5,6\}$	$\{4,20\}$	

The attributes 18 and 6 appear most often (12× and 9×) and thus we get, combining all sets which contain at least one of these,

$$\{1,2,3,5,6,7,8,11,12,14,15,16,17,18\} \text{ and } \{2,4,5,6,7,8,11,14,16,17,18\}.$$

The attributes a_i with $p_{im} < 0.05$ are $1,2,3,5,8,9,11,13,14,16,17,19,20$.

The best attribute sets GROBIAN (1999) determined with the entropy-function (γ-Cut 0.4, maximum length 10) were

set	rel. entropy	γ
$\{05,07,08,09,10,11,14,17,18\}$	0.113	0.432
$\{05,06,08,09,10,11,14,17,18,20\}$	0.107	0.486
$\{05,06,07,09,10,14,17,18,20\}$	0.107	0.486
$\{05,06,07,09,10,11,14,17,18\}$	0.107	0.486
$\{05,06,07,08,09,10,11,14,17,18\}$	0.107	0.486
$\{05,06,07,09,10,11,14,17,18,20\}$	0.107	0.486
$\{06,07,08,09,10,11,14,17,18,20\}$	0.107	0.473
$\{05,09,10,11,12,14,17,18,20\}$	0.107	0.473
$\{05,08,09,11,14,17,18,20\}$	0.107	0.405
$\{03,05,06,07,08,09,10,11,18,19\}$	0.101	0.459

5.2.2 Final preselection

According to the binomial test and the wnf-values we plan to eliminate the attributes $01,04,05,08,11,16,19,20$; our opinion about $09,13,15$ remains undetermined by these tests.
These exclusions are supported by their small γ-drops – the only exception is 16 which is frequently observed in the minimal reducts and has larger drops of γ as shown in table 4.

[22] This is an unpleasant result, since it is an indicator that no effect exists (refer to Stanley (2005)) but is replicated by regression analysis with more data.

We decide to keep the attributes 05, 09, 15, 16 and reject 01 (its drops are never above 10% and 01 is already taken care of by 02 and 03). We also keep 13 as a precaution, since it only appears two times (although both times with very low drops).

After removing the just mentioned attributes we can have a final look on the remaining sets in table 5.
To these we add those subsets with the highest NRE-value (as defined on page 3.2.1).
The attributes 02, 03, 06, 07, 10, 14, 17 appear to be important, 12 and 18 seem to be very important.
Our final choice of sets are the remaining minimal reducts, the RFH-sets and the sets found by GROBIAN (1999) by maximising the NRE-value (NRE-sets); all without the attributes previously declared unimportant.

5.3 Analysis – statistics

Most of the following statistics were calculated with *GROBIAN*.
In table 6 we depict the statistics of our chosen sets – first the minimal reducts, then the RFH-sets and finally the NRE-sets.
All attributes and attribute sets were found to be conditional casual (see 3.1, the values are not shown here) which indicates that there is neither a true dominating attribute or set (there must be much redundancy in those sets) nor is there an indication that the rules depend on a specific set of objects. However, although not significant on any reasonable level, attribute 18 had in almost all cases the lowest statistic (with a large gap to the attribute with the second lowest value).

If we draw random sets of six attribute the average γ is 0.22 (sd 0.11).

5.3.1 Validating

Without any information (i.e. predicting with the null-vector) the expected fraction of correctly classified objects is

$$\left(\frac{22}{74}\right)^2 + \left(\frac{18}{74}\right)^2 + \left(\frac{29}{74}\right)^2 + \left(\frac{5}{74}\right)^2 = 0.3057.$$

The calculated prediction qualities vary around 51.02% (sd 1.83%) (those with a NRE above 0.1 around 52.4% (sd 1.3%)) and therefore surpass the random guessing by 66.9% (respectively 71.4%).
When we randomly[23] draw 6 attributes for prediction the mean quality is 49.3% (sd 2.5%) (their standard deviation varies around 12%, thus being somewhat smaller). Since most of our sets are very similar the comparison with randomly drawn attributes is a bit artificial.

[23] all comparisions were conducted with the same 10 randomly chosen sets of attributes

5.3.2 Jackknife

6 randomly drawn attributes lead to a mean classification of 49.63% (sd 2.6%) compared to our 51.24% (sd 1.95%) (respectively 52.67% (sd 1.23%)).
Much more interesting is the mean worst-case classification of 15% (sd 7.9%) compared to our observed 32.05% (sd 4.14%) (respectively 34.6% (sd 1.92%)) which indicates that our sets reduce the need to guess by a substantial amount.

5.3.3 Rules

Rules with a support greater than 4 are mostly driven by several studies. Hence we collect in table 7 all distinct rules from the sets with a NRE above 0.1 and a support greater or equal than four, which deterministically classify an object.
Since only two studies with $d = 3$ are correctly classified and they are from the same author, the rule $\mathrm{NODETVAR} = 3, \mathrm{NOSOZVAR} = 4, \mathrm{INCOME} = 1 \Rightarrow d = 3$ should be considered with care. It is worth mentioning, however, that 3 of 5 negative significant studies used many variables (more than 13) whereas the other two used five or more.

Table 7 must be interpreted with caution since many attributes reappear in many sets and therefore may be repeatedly represented in the table in combination with different attributes. To soften this effect we take a look at the correlations between the important attributes and d.

5.3.4 Contingency tables

Significant observations (therefore unlikely to be distributed according to the marginal frequencies) are the following:

Nodetvar,02 Studies with only one variable have often (positive) significant ($p = 0.021$) and seldom partly insignificant[24] results ($p = 0$). Studies with a medium number show no significance and with many have often partly insignificant results ($p = 0.002$). Negative results are evenly spread across the quantity and given the outcome almost the same results are achieved.

Nosozvar,03 The studies with the lowest number of variables had significantly often mixed ($p = 0.011$) and seldom ($p = 0.024$) insignificant results. Studies using more than ten variables (4) had very often ($p = 0.006$) insignificant results; all studies in between show nothing special. Negative results are evenly spread across the quantity and given the outcome the same results follow.

Income,05 Studies not considering the income have often insignificant ($p = 0.035$) and seldom mixed ($p = 0.035$) results, otherwise they show a contrary

[24] a partly insignificant study means $d = 1$

behavior: often mixed ($p = 0.035$) and seldom insignificant ($p = 0.035$). No significant correlation towards negative results can be found and given the outcome the same results follow only for the omission of income.

Cultyear,09 Old studies have more mixed ($p = 0$) and less insignificant results ($p = 0.053$), while newer studies (> 1980) show the contrary ($p = 0$ and 0.053) and the newest studies show no correlation at all. Negative results are evenly spread across time and given the outcome nothing changes.

Dataage,13 The studies with the oldest data have often mixed ($p = 0$) and rarely insignificant results ($p = 0.018$). Studies with the newest data (> 1989) have very often insignificant ($p = 0$) and seldom mixed results ($p = 0.001$). Data from $69 - 89$ (i.e. 2,3) show no significant behavior; negative results are evenly spread across the age of the data and given the outcome almost the same results follow.

Dataus,14 Studies with US-data have very often insignificant ($p = 0.018$) results and those without show these results only seldom ($p = 0.018$). Negative results are evenly spread and given the outcome, only the studies with insignificant results remain suspicious ($p = 0.022$).

Datano,16 Data with the smallest samples have seldom insignificant ($p = 0.009$) and often mixed results ($p = 0$), therefore all other studies show the opposite effect (same p-values). Distinctively analyzed almost all the p-values show little significance, negative results are evenly spread across the sample sizes and given the outcome almost the same results follow.

Datatpc,17 Studies with time series have often mixed ($p = 0$) and seldom ($p = 0.027$) positive results. Studies with panels have seldom ($p = 0.001$) mixed and often ($p = 0.001$) positive results while studies with cross-sectional data show slightly more insignificant results ($p = 0.061$). Negative results are evenly spread across the sampletype and given the outcome the same results follow.

Datavpa,18 No significant Freeman-Halton-statistic and no significant connection to the specific subject (all $p > 0.14$). The only exception: negative results are conspicuously high within the combined studies ($p = 0.02$) and given the outcome the same results follow (all $p > 0.19$).

Whereas the type of crime under consideration is indispensable for classification it is not significantly correlated with the outcome of a study.

5.4 Findings

The results end up being not as straightforward as we had hoped but nonetheless we manage to draw some interesting conclusions.

1. Most important:

    - since the highest γ-values are well below one, there are many contradictive configurations in the observed data; this means that the data contains many similar patterns with different outcomes and that there is no decisive set of attributes which could explain most of the observed t-values
    - the most decisive coherence is the connection between the subject of the study (whether it concentrates on violence or property crime or both) and its outcome: DATAVPA (18) is in the core, has large (often above 50%) drops of γ throughout the sets and appears in many deterministic rules[25]. Since its contingency table provides no significant insights (except that negative results tend to appear in studies with no focus on violence or property crime) we draw the conclusion that the target of the study at hand does not significantly favor an outcome on its own but only in combination with other variables;
    - whether the study concentrates on youths or not (DATAYOUTH,12) is important for classification (included in the core) but this is constricted to a few cases only and thus bears only little general information (the γ-drop is almost always low, the contingency table is insignificant and it appears only in five deterministic rules which is relatively low since it is contained in every set).

    Other important conclusions and attributes are:

    - attributes with strong dependance with the outcome are not sufficient to predict the outcome; therefore considering only the highly correlated attributes is not enough
    - the number of variables (NODETVAR, NOSOZVAR,2,3) used in a study (third and first highest wnf', significant contingency tables, mostly big γ-drops (sometimes of more than 50%), appearance in many rules;
    - the structure (DATATPC,17) of the data (high wnf', significant contingency table (c.t.), medium to high γ-drops, appears in many rules, high representation in sets with high NRE);
    - whether the study came from an economic (CULTECO,10) background (second highest wnf' and appears in many sets with high NRE but insignificant c.t., low γ-drops and little representation in the rules);
    - whether simultaneity (SIMUL,6) was accounted for (medium wnf' and appears in many sets with high NRE but insignificant c.t., low γ-drops and only small representation in the rules);

[25] we concentrate on deterministic rules and omit all rules with a support below four

- whether fixed effects models (FIXEFF,7) were used (medium wnf' and appears in many sets with high NRE but insignificant c.t., low γ-drops and only small representation in the rules);
- whether American (CULTUS,8) data were used (high wnf', significant c.t. and appears in many sets with a high NRE but low drops of γ and no rule with a support ≥ 4).

Attributes which should also be considered are (medium to low wnf'-values):

- age (DATAAGE,13) of the data (medium wnf', significant c.t., appears in some rules but low γ-drops);
- age (CULTYEAR,9) of the study (bigger γ-drops in set with high NRE, significant c.t., contained in most set with high NRE, appears in many rules);
- whether urban (DATAURBAN,15) data were used (big γ-drops appears in many rules, but insignificant c.t.);
- (DATANO,16) – the sample size (significant c.t., big γ-drops, appears in many rules);
- whether the INCOME (5) was accounted for (significant c.t., is contained in most sets with high NRE, medium γ-drops but does not appear in many rules).

2. The following attributes appear to be irrelevant for the outcome (least important first):

   - the time interval of the data (DATATIME,20);
   - whether the deterrent effect was studied (DETFOCUS,11);
   - whether the author is American (CULTUS ,8);
   - the source (DATASRC, 19) of the data;
   - whether unemployment (UNEMP,4) was accounted for.

3. The combined insights from the rulesets with support ≥ 4 and the contingency tables seem to be:

   a) the more variables a study uses (especially social variables) the more probable are insignificant results;

   b) newer studies (resp. studies with newer data) tend to get clearer results;

   c) studies using urban data tend to get less mixed results (all t-values tend to be either significant or insignificant)

d) studies with time series get mixed results (especially those concentrating on property crime) while those with cross-section data tend to insignificant results. Studies with panel data get more significant results (especially those concentrating on property crime);

e) the outcome of studies which user property crimes is easier to predict

4. The best attribute sets seem to be:

   a) $\{05, 09, 10, 14, 17, 18\}$ (NRE, number of rules, %),

   b) $\{07, 09, 10, 14, 17, 18\}$ (NRE, number of rules, %),

   c) $\{02, 03, 09, 15, 18\}$ (γ, NRE, $\%, \alpha'$s, number of rules),

   d) $\{03, 10, 12, 15, 17, 18\}$ (γ, NRE, $\%, \alpha'$s, number of rules).

5.4.1 Comparison with the crime data

Are these results compatible with the conclusions from a standard meta regression analysis?
The comparison is tabulated in table 9.
Conclusions which were not discovered by the linear regression are:

1. the importance of the data structure (time series, panel, cross-section);

2. the explanatory power of the kind of crime studied works only in combination with other attributes (in particular DATATPC);

3. oldest and newest studies (and data) have plainer results while all others have not;

4. studies with urban data show a similar behavior;

5. the influence of studies which concentrate on youths is not systematic.

We see that both types of analyses agree on many findings but also differ on some. While RSDA cannot give us precise information about the strength about an effect, it can tell us much about relations between attributes which remained hidden or were probably misinterpreted in a linear regression analysis.

6 Conclusion

In this paper we presented the mathematical concept and statistical tools of the Rough Set Data Analysis (RSDA) which, to our best knowledge, has not yet been used in the fields of economics or meta-analysis to such an extent. It relies on no prior assumptions about the data and refers largely to the principle of pattern recognition. It does not produce coefficients of the strength of relationships between attributes but, however, provides qualitative characteristics about these. Therefore it can be used to discover overlooked dependencies, filter the set of attributes for further inspection with other methods, make already known results more robust or refute existing results.

The theory of deterrence has already been studied for decades but still neither a common agreement could be reached nor could the attributes, which are decisive for this diversity, be isolated – the underlying mechanism, if it exists, is still empirically not understood.
We applied the methodology of RSDA to a data set of empirical studies about the deterrent effect of the probability and severity of punishment which had already been subject to a small meta regression analysis.

We found that both methods agreed upon many relations between recorded variables and the outcome of a study but also found some contradictions. The most interesting conflicting results were the influence of the distinction between property and violent crime and whether the empirical study concentrated on youths, which are both positive and significant in the meta-regression analysis (i.e. studying property or juvenile crime tended to significant positive t-values, thus supporting the theory that property crime fits the rational criminal theory more than violent crime and that youths appear more often in the crime statistics than adults). However, we discovered evidence that the distinction between property and violent crime is important but only in conjunction with other variables and not ceteris paribus. Also the the importance of the juvenile crime is artificial.
We also discovered the non-linear influence of the publication year which was not studied in the meta-regression analysis. The attributes which were declared by both methods to be not important enough for a deeper inspection were in large the same.

To sum it up we think that RSDA can be a useful tool in the field of data analysis, especially when the assumptions of the standard econometric methods are not sufficiently met or the data structure is not easily accessible to them. In particular the ability to deal with nearly any kind of data (continuous, ordinal or nominal) and its complete lack of prior assumptions (the distribution of the data, its covariance matrix and other often undesirable properties are of no importance) recommend it to be added to the pool of statistical methods in econometrics.

Table 1: attributes from the studied data set

abbr.	attribute	description
	Tarrest	t-value of the arrest rate
	Tconvict	t-value of the conviction rate
	Tsevere	t-value of the penalty severence
01	Noexvar	number of explanatory variables
02	Nodetvar	number of explanatory deterrence-variables
03	Nosozvar	number of explanatory social-variables
04	Unemp	whether unemployment was considered
05	Income	whether the income was considered
06	Simul	whether simultaneity was considered
07	Fixeff	whether a fixed effect model was used
08	Cultus	whether it is an U.S. study
09	Cultyear	year the study was published
10	Culteco	whether the author is an economist
11	Detfocus	study focuses on deterrence
12	Datayouth	whether only youth were studied
13	Dataage	average year of the data
14	Dataus	whether the data a from the U.S.
15	Dataurban	whether urban data were studied
16	Datano	samplesize
17	Datats	study used time series
17	Datapan	study used panel data
17	Datacs	study used cross-sectional data
18	Dataviol	study focused on violent crime
18	Dataprop	study focused on property crime
18	Dataall	study used both types
19	Datanat	data from a nation
19	Datastate	data from a state
19	Datacounty	data from a county
19	Datamicro	data from individuals
20	Datayear	data was surveyed yearly
20	Datamonth	data was surveyed monthly
20	Dataweek	data was surveyed weekly
20	Dataday	data was surveyed daily

Table 2: reducts of the dataset

Q'	$\gamma_{Q'}(d)$	SC	Q'	$\gamma_{Q'}(d)$	SC
$\{02, 03, 06, 12, 15, 18\}$	0.527	0.686	$\{02, 03, 05, 08, 12, 18\}$	0.527	0.549
$\{02, 03, 10, 12, 15, 18\}$	0.527	0.667	$\{02, 03, 05, 09, 12, 18\}$	0.527	0.549
$\{01, 10, 12, 14, 15, 17, 18\}$	0.527	0.647	$\{02, 03, 05, 12, 14, 18\}$	0.527	0.549
$\{03, 08, 10, 12, 15, 17, 18\}$	0.527	0.647	$\{02, 03, 12, 15, 18, 19\}$	0.527	0.549
$\{01, 02, 12, 14, 15, 17, 18\}$	0.527	0.627	$\{02, 03, 12, 16, 17, 18\}$	0.527	0.549
$\{03, 10, 11, 12, 15, 17, 18\}$	0.527	0.627	$\{02, 03, 08, 12, 16, 18\}$	0.527	0.529
$\{02, 03, 12, 15, 17, 18\}$	0.527	0.608	$\{02, 03, 12, 14, 16, 18\}$	0.527	0.529
$\{02, 03, 05, 12, 17, 18\}$	0.527	0.588	$\{02, 03, 12, 13, 15, 18\}$	0.527	0.49
$\{02, 03, 08, 12, 15, 18\}$	0.527	0.588	$\{02, 03, 09, 12, 15, 18\}$	0.527	0.471
$\{02, 03, 12, 14, 15, 18\}$	0.527	0.588	$\{02, 03, 09, 12, 16, 18\}$	0.527	0.412
$\{03, 07, 10, 12, 16, 18\}$	0.527	0.588	$\{03, 09, 10, 12, 16, 18\}$	0.527	0.412
$\{03, 10, 12, 15, 16, 18\}$	0.527	0.588			

Q' is the chosen subset of attributes, $\gamma_{Q'}(d)$ is the fraction of complete knowledge of d through Q' (see subsection 2.5.3) and SC is the stability coefficient (see 2.5.1) of Q'.

Table 3: characteristics of the attributes in the dataset

attr.	f	b	b'	wnf	wnf'	attr.	f	b	b'	wnf	wnf'
01	42	0.00	0.06	0.10	0.15	11	21	0.00	0.00	0.00	0.00
02	99	1.00	1.00	0.42	0.63	12	193	1.00	–	1.00	–
03	126	1.00	1.00	0.67	1.00	13	58	0.04	0.77	0.20	0.30
04	43	0.00	0.08	0.11	0.17	14	90	0.98	1.00	0.36	0.54
05	43	0.00	0.08	0.10	0.14	15	51	0.00	0.41	0.22	0.33
06	69	0.36	0.99	0.26	0.39	16	22	0.00	0.00	0.01	0.01
07	68	0.32	0.98	0.24	0.36	17	91	0.99	1.00	0.35	0.53
08	22	0.00	0.00	0.02	0.04	18	193	1.00	–	1.00	–
09	56	0.02	0.68	0.15	0.22	19	39	0.00	0.02	0.09	0.13
10	100	1.00	1.00	0.49	0.72	20	24	0.00	0.00	0.00	0.00

attr. is the numerical abbrevation of an attribute (see table 1), f is the number of appearances of an attribute in all reducts of minimal length, b is the p-value under the hypothesis that all attributes are equally distributed across these reducts (see 3.1.3) and wnf is the weighted normalized frequency (refer to 2.5.2). b' and wnf' are calculated without the core.

Table 4: additional evidence about the irrelevant attributes

set	γ	$\gamma(\cdot\backslash 01)$	$\gamma(\cdot\backslash 04)$	$\gamma(\cdot\backslash 05)$	$\gamma(\cdot\backslash 08)$	$\gamma(\cdot\backslash 09)$	$\gamma(\cdot\backslash 11)$	$\gamma(\cdot\backslash 13)$	$\gamma(\cdot\backslash 15)$	$\gamma(\cdot\backslash 16)$	$\gamma(\cdot\backslash 19)$	$\gamma(\cdot\backslash 20)$
{02, 03, 06, 12, 15, 18}	0.527								−26%			
{02, 03, 10, 12, 15, 18}	0.527								−26%			
{01, 10, 12, 14, 15, 17, 18}	0.527	−15%							−38%			
{03, 08, 10, 12, 15, 17, 18}	0.527				−5%				−33%			
{01, 02, 12, 14, 15, 17, 18}	0.527	−13%							−15%			
{03, 10, 11, 12, 15, 17, 18}	0.527						−5%		−33%			
{02, 03, 12, 15, 17, 18}	0.527								−21%			
{02, 03, 05, 12, 17, 18}	0.527			−21%								
{02, 03, 08, 12, 15, 18}	0.527				−3%				−10%			
{02, 03, 12, 14, 15, 18}	0.527								−10%			
{03, 07, 10, 12, 16, 18}	0.527									−41%		
{03, 10, 12, 15, 16, 18}	0.527								−18%	−41%		
{02, 03, 05, 08, 12, 18}	0.527			−10%	−8%							
{02, 03, 05, 09, 12, 18}	0.527			−10%								
{02, 03, 05, 12, 14, 18}	0.527			−10%								
{02, 03, 12, 15, 18, 19}	0.527										−3%	
{02, 03, 12, 16, 17, 18}	0.527									−21%		
{02, 03, 08, 12, 16, 18}	0.527				−8%					−10%		
{02, 03, 12, 14, 16, 18}	0.527									−10%		
{02, 03, 12, 13, 15, 18}	0.527							−3%	−15%			
{02, 03, 09, 12, 15, 18}	0.527					−7%			−10%			
{02, 03, 09, 12, 16, 18}	0.527					−7%				−10%		
{03, 09, 10, 12, 16, 18}	0.527					−18%				−26%		
{1, 2, 3, 5, 6, 7, 8, 11, 12, 14, 15, 16, 17, 18}	0.527	−0%		−5%	−0%	−0%	−0%		−0%	−0%		
{2, 4, 5, 6, 7, 8, 11, 14, 16, 17, 18}	0.487		−0%	−0%	−0%		−0%			−5%		
{1, 2, 3, 5, 8, 9, 11, 13, 14, 16, 17, 19, 20}	0.297	−0%		−9%	−0%	−0%	−0%	−0%		−0%	−0%	−0%
{05, 07, 08, 09, 10, 11, 14, 17, 18}	0.432			−3%	−0%	−25%	−0%					
{05, 06, 08, 09, 10, 11, 14, 17, 18, 20}	0.486			−14%	−0%	−22%	−0%					−0%
{05, 06, 07, 09, 10, 14, 17, 18, 20}	0.486			−3%		−14%						−0%
{05, 06, 07, 09, 10, 11, 14, 17, 18}	0.486			−3%		−11%	−0%					
{05, 06, 07, 08, 09, 10, 11, 14, 17, 18}	0.486			−3%	−0%	−11%	−0%					
{05, 06, 07, 09, 10, 11, 14, 17, 18, 20}	0.486			−3%		−11%	−0%					−0%
{06, 07, 08, 09, 10, 11, 14, 17, 18, 20}	0.473				−0%	−9%	−0%					−0%
{05, 09, 10, 11, 12, 14, 17, 18, 20}	0.473			−11%		−26%	−0%					−0%
{05, 08, 09, 11, 14, 17, 18, 20}	0.405			−10%	−0%	−33%	−0%					−0%
{03, 05, 06, 07, 08, 09, 10, 11, 18, 19}	0.459			−0%	−0%	−0%	−0%				−5%	

set is the chosen set of attributes, γ is the complete knowledge of d through the chosen set, $\gamma(\cdot\backslash x)$ is the γ-drop caused by leaving out the attribute x, as described in section 2.5.3.

Table 5: importance of the remaining attributes

set	γ	$\gamma(\cdot\backslash 02)$	$\gamma(\cdot\backslash 03)$	$\gamma(\cdot\backslash 05)$	$\gamma(\cdot\backslash 06)$	$\gamma(\cdot\backslash 07)$	$\gamma(\cdot\backslash 09)$	$\gamma(\cdot\backslash 10)$	$\gamma(\cdot\backslash 12)$	$\gamma(\cdot\backslash 13)$	$\gamma(\cdot\backslash 14)$	$\gamma(\cdot\backslash 15)$	$\gamma(\cdot\backslash 16)$	$\gamma(\cdot\backslash 17)$	$\gamma(\cdot\backslash 18)$
{02, 03, 05, 09, 12, 18}	0.527	−23%	−38%	−10%			−8%		−8%						−43%
{02, 03, 05, 12, 14, 18}	0.527	−53%	−53%	−10%					−8%		−8%				−46%
{02, 03, 05, 12, 17, 18}	0.527	−35%	−26%	−21%					−8%					−8%	−41%
{02, 03, 06, 12, 15, 18}	0.527	−49%	−67%		−3%				−18%			−26%			−64%
{02, 03, 09, 12, 15, 18}	0.527	−13%	−44%				−3%		−8%			−10%			−44%
{02, 03, 09, 12, 16, 18}	0.527	−5%	−18%				−8%		−8%				−10%		−44%
{02, 03, 10, 12, 15, 18}	0.527	−41%	−82%					−3%	−18%			−26%			−49%
{02, 03, 12, 13, 15, 18}	0.527	−10%	−23%						−8%	−3%		−15%			−49%
{02, 03, 12, 14, 15, 18}	0.527	−54%	−69%						−18%		−3%	−10%			−59%
{02, 03, 12, 14, 16, 18}	0.527	−18%	−21%						−8%		−8%		−10&		−54%
{02, 03, 12, 15, 17, 18}	0.527	−23%	−28%						−18%			−21%		−3%	−53%
{02, 03, 12, 16, 17, 18}	0.527	−18%	−10%						−8%				−21%	−8%	−38%
{02, 12, 14, 15, 17, 18}	0.459	−41%							−24%		−18%	−18%		−65%	−59%
{03, 07, 10, 12, 16, 18}	0.527		−33%			−18%		−15%	−8%				−41%		−51%
{03, 09, 10, 12, 16, 18}	0.527		−28%				−18%	−5%	−8%				−26%		−54%
{03, 10, 12, 15, 16, 18}	0.527		−41%					−15%	−8%			−18%	−41%		−56%
{03, 10, 12, 15, 17, 18}	0.500		−36%					−23%	−23%			−41%		−41%	−49%
{10, 12, 14, 15, 17, 18}	0.446							−39%	−36%		−24%	−52%		−73%	−70%
{02, 03, 05, 06, 07, 12, 14, 15, 16, 17, 18}	0.527	−5%	−0%	−0%	−0%	−0%			−8%		−0%	−0%	−0%	−0%	−38%
{02, 05, 06, 07, 14, 16, 17, 18}	0.486	−8%		−0%	−6%	−14%					−6%		−6%	−14%	−54%
{02, 03, 05, 09, 13, 14, 16, 17}	0.297	−5%	−0%	−9%			−0%			−0%	−0%		−0%	−0%	
{05, 07, 09, 10, 14, 17, 18}	0.432			−3%		−0%	−38%	−6%			−6%			−22%	−38%
{05, 06, 09, 10, 14, 17, 18}	0.486			−25%	−11%		−25%	−6%			−6%			−19%	−39%
{05, 06, 07, 09, 10, 14, 17, 18}	0.486			−3%	−11%	−0%	−14%	−6%			−6%			−19%	−39%
{06, 07, 09, 10, 14, 17, 18}	0.473				−11%	−23%	−3%	−6%			−6%			−20%	−49%
{05, 09, 10, 12, 14, 17, 18}	0.473			−11%			−43%	−9%	−9%		−6%			−26%	−34%
{03, 05, 06, 07, 09, 10, 18}	0.459		−26%	−0%	−12%	−0%	−9%	−6%							−53%

set is the chosen set of attributes, γ is the complete knowledge of d through the chosen set, $\gamma(\cdot\backslash x)$ is the γ-drop caused by leaving out the attribute x, as described in section 4.

Table 6: benchmarks of the chosen sets

set	γ	NRE	s-Attr.	$J\%$	$A\%$	#rules	$(1, \geq 4, \%)$	$\#/22, \alpha_0$	$\#/18, \alpha_1$	$\#/29, \alpha_2$	$\#/5, \alpha_3$
{02, 03, 05, 09, 12, 18}	0.527	0.094	–	52, 34	51, 15	19	(6, 4, 43%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 05, 09, 18}	0.486	0.101	03, 18	54, 36	53, 15	17	(7, 3, 38%)	12, 35	8, 31	14, 33	2, 15
{02, 03, 05, 12, 14, 18}	0.527	0.094	02, 03	48, 30	49, 14	20	(7, 3, 49%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 05, 14, 18}	0.486	0.101	03, 02, 18	51, 32	50, 14	18	(7, 3, 41%)	12, 35	8, 31	14, 33	2, 15
{02, 03, 05, 12, 17, 18}	0.527	0.094	–	51, 34	51, 15	19	(7, 3, 46%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 05, 17, 18}	0.486	0.101	–	54, 35	54, 15	17	(6, 3, 42%)	12, 35	8, 31	14, 33	2, 15
{02, 03, 06, 12, 15, 18}	0.527	0.094	03, 18	49, 30	49, 14	21	(10, 3, 41%)	14, 42	8, 31	15, 36	2, 17
{02, 03, 09, 12, 15, 18}	0.527	0.094	–	52, 35	52, 15	20	(10, 4, 42%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 09, 15, 18}	0.527	0.101	18	53, 36	53, 15	17	(7, 4, 50%)	12, 36	8, 31	14, 33	2, 15
{02, 03, 09, 12, 16, 18}	0.527	0.088	–	49, 32	49, 15	21	(12, 4, 41%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 09, 16, 18}	0.486	0.094	18	49, 32	50, 15	18	(10, 3, 36%)	12, 36	8, 31	14, 33	2, 15
{02, 03, 10, 12, 15, 18}	0.527	0.094	03	50, 30	50, 13	21	(10, 3, 41%)	14, 42	8, 31	15, 36	2, 17
{02, 03, 12, 13, 15, 18}	0.527	0.094	–	53, 36	52, 15	18	(8, 4, 45%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 13, 15, 18}	0.486	0.101	18	53, 38	53, 15	15	(5, 4, 43%)	12, 36	8, 31	14, 33	2, 15
{02, 03, 12, 14, 15, 18}	0.527	0.094	03, 18, 02	49, 28	48, 14	21	(10, 3, 41%)	14, 42	8, 31	15, 36	2, 17
{02, 03, 12, 14, 16, 18}	0.527	0.088	18	49, 31	49, 14	22	(12, 3, 41%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 14, 16, 18}	0.486	0.094	18	49, 31	49, 14	19	(10, 3, 36%)	12, 36	8, 31	14, 33	2, 15
{02, 03, 12, 15, 17, 18}	0.527	0.094	18	52, 32	51, 14	19	(9, 4, 42%)	14, 42	8, 31	15, 36	2, 17
{02, 03, 12, 16, 17, 18}	0.527	0.088	–	49, 32	49, 15	20	(11, 4, 41%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 16, 17, 18}	0.486	0.094	–	50, 32	50, 14	17	(9, 4, 38%)	12, 36	8, 31	14, 33	2, 15
{02, 12, 14, 15, 17, 18}	0.459	0.082	17, 18	49, 26	49, 13	20	(11, 2, 32%)	14, 30	8, 31	10, 22	2, 10
{03, 07, 10, 12, 16, 18}	0.527	0.088	18	52, 31	50, 14	20	(10, 3, 43%)	14, 42	8, 31	15, 36	2, 17
↪ {03, 07, 10, 16, 18}	0.486	0.094	16, 18	52, 31	51, 14	17	(7, 3, 39%)	12, 36	8, 31	14, 33	2, 15
{03, 09, 10, 12, 16, 18}	0.527	0.088	18	48, 32	48, 15	21	(11, 4, 42%)	14, 42	8, 31	15, 36	2, 17
↪ {03, 09, 10, 16, 18}	0.486	0.094	18	50, 32	51, 15	19	(10, 3, 36%)	12, 36	8, 31	14, 33	2, 15
{03, 10, 12, 15, 16, 18}	0.527	0.088	18	50, 31	50, 14	21	(12, 4, 39%)	14, 42	8, 31	15, 36	2, 17
↪ {03, 10, 15, 16, 18}	0.486	0.094	18, 16	50, 31	50, 14	18	(10, 4, 35%)	12, 36	8, 31	14, 33	2, 15
{03, 10, 12, 15, 17, 18}	0.500	0.103	–	52, 32	51, 14	17	(7, 4, 42%)	13, 38	8, 31	14, 32	2, 15
{10, 12, 14, 15, 17, 18}	0.446	0.082	17, 18, 15	48, 23	48, 13	19	(10, 2, 30%)	14, 30	7, 23	10, 22	2, 10
{02, 03, 05, 06, 07, 12, 14, 15, 16, 17, 18}	0.527	0.094	–	52, 34	52, 14	18	(7, 4, 46%)	14, 42	8, 31	15, 36	2, 17
↪ {02, 03, 05, 14, 18}	0.486	0.101	03, 02, 18	51, 32	51, 14	18	(7, 3, 41%)	12, 35	8, 31	14, 33	2, 15
{02, 05, 06, 07, 14, 16, 17, 18}	0.486	0.094	18	53, 38	53, 15	17	(6, 3, 43%)	12, 35	8, 31	14, 33	2, 15
↪ {05, 06, 07, 14, 16, 17, 18}	0.446	0.101	18	52, 36	53, 14	15	(6, 3, 38%)	12, 34	7, 26	13, 30	1, 8
{02, 03, 05, 09, 13, 14, 16, 17}	0.297	0.075	–	52, 17	51, 12	10	(5, 1, 23%)	8, 16	4, 15	8, 16	2, 15
↪ {02, 05, 13, 14}	0.257	0.098	13	55, 22	55, 12	8	(3, 1, 22%)	8, 15	3, 10	6, 11	2, 10
{05, 07, 09, 10, 14, 17, 18}	0.432	0.113	–	53, 34	53, 14	13	(4, 4, 39%)	11, 27	8, 31	11, 24	2, 15
{05, 06, 09, 10, 14, 17, 18}	0.486	0.107	–	53, 36	53, 14	16	(6, 4, 42%)	12, 35	8, 31	14, 33	2, 15
↪ {05, 09, 10, 14, 17, 18}	0.432	0.113	09	53, 34	53, 14	13	(4, 4, 39%)	11, 27	8, 31	11, 24	2, 15
{06, 07, 09, 10, 14, 17, 18}	0.473	0.107	–	54, 36	53, 14	15	(5, 4, 42%)	12, 34	7, 27	14, 33	2, 15
↪ {07, 09, 10, 14, 17, 18}	0.419	0.113	–	54, 34	54, 14	13	(4, 4, 38%)	12, 34	7, 27	14, 33	2, 15
{05, 09, 10, 12, 14, 17, 18}	0.473	0.107	–	50, 32	50, 15	15	(5, 5, 45%)	12, 31	8, 31	13, 28	2, 17
{03, 05, 06, 07, 09, 10, 18}	0.459	0.101	18	53, 36	52, 15	15	(5, 3, 41%)	12, 35	7, 26	13, 30	2, 15

set: the chosen set of reducts ,γ: the complete knowledge of d through the chosen set (section 2.5.3), NRE: the normalized rough entropy (section 3.2.1), s-Attr.: those attributes which cause a γ-drop of more than 50%, J% the two values from the Jackknife test (expected and worst-case prediction, see 2.5.4), A%: the average prediction from 3.1.2, # rules: the number of deterministic rules, (1,$\geq$ 4, %): the number of rules with a support of one resp. larger than 4 and the percentage of correctly classified objects by all rules which are non deterministic casual, $\#\backslash X$,α_Y: the number of correctly classified objects of $d = Y$ (maximum of X), the local knowledge about the values of d in % (refer to 2.5.3).

Table 7: correct classifying rules with support ≥ 4

Nodetvar	Nosozvar	Income	Simul	Fixeff	Cultyear	Culteco	Datayouth	Dataage	Dataurban	Datano	Datatpc	Datavpa	d	support
2								4					0	8
		0						4					0	8
2							1			4			0	5
2	4									4			0	5
2	4	0											0	5
2					3								0	5
		0			3								0	5
				1						4	3		0	5
				1						4		2	0	5
				1	3						3		0	5
					3					4			0	5
	4						1		0			2	0	4
	4											1	0	4
			0				0			3		1	0	4
			0								3	1	0	4
					3						3	1	0	4
							1		0		3	2	0	4
									0	4		2	0	4
3									0		1	2	1	4
3	3											2	1	4
	3									1		2	1	4
	3										1	2	1	4
	3				1							2	1	4
				1						1	1	2	1	4
					1						1	2	1	4
						1			1		1	2	1	4
1					4							2	2	4
	3							3				2	2	4
	3								0		2	2	2	4
					2	1	1						2	4
					4				0			2	2	4
									0	3		2	2	4
										3	2	2	2	4

d: the values of the decision attribute,
support: the number of objects which support the rule.

Table 8: dependencies between categories; significance **given** A or *given* d

	0	1	2	3
1	8	***0***	***16***	2
2	12	7	9	1
3	*2*	***11***	4	2
$02 \times d$		**p=0**		

	0	1	2	3
1	***1***	***9***	7	0
2	4	3	7	2
3	4	5	9	0
4	***13***	***1***	6	3
$03 \times d$		**p=0.003**		

	0	1	2	3
0	***10***	***1***	8	0
1	**12**	**17**	21	5
$05 \times d$		**p=0.021**		

	0	1	2	3
0	19	11	20	5
1	3	7	9	0
$06 \times d$		$p = 0.16$		

	0	1	2	3
0	8	7	8	3
1	14	11	21	2
$07 \times d$		$p = 0.558$		

	0	1	2	3
1	*2*	***13***	4	0
2	7	***2***	11	3
3	***11***	***0***	7	2
4	2	3	7	0
$09 \times d$		**p=0**		

	0	1	2	3
0	5	1	3	1
1	17	17	26	4
$10 \times d$		$p = 0.384$		

	0	1	2	3
0	16	18	25	4
1	6	*0*	4	1
$12 \times d$		$p = 0.095$		

	0	1	2	3
1	***1***	***12***	5	0
2	4	3	*7*	3
3	4	3	13	2
4	***13***	***0***	4	0
$13 \times d$		**p=0**		

	0	1	2	3
0	***3***	7	15	2
1	**19**	11	14	3
$14 \times d$		**p=0.045**		

	0	1	2	3
1	***1***	***13***	6	0
2	6	***1***	12	2
3	9	3	9	2
4	*6*	1	2	1
$16 \times d$		**p=0**		

	0	1	2	3
1	5	***12***	***3***	0
2	4	***0***	***18***	2
3	13	6	8	3
$17 \times d$		**p=0.001**		

	0	1	2	3
1	10	5	10	1
2	10	12	15	1
3	2	1	4	***3***
$18 \times d$		$p = 0.167$		

Table 9: comparison of the results

Observation / effect	RSDA	meta regression analysis
number of moderators	negative	negative
simultaneity	low and unsigned	positive
fixed-effects	low and unsigned	insignificant
deterrence effect	n.a.	yes
property crime	distinction important	positive
	unclear on its own	important
income significant	yes	yes
unemployment insignificant	yes	yes
data source important	no	yes
data structure	important	unimportant
background of author	unimportant	unimportant
publication year	yes	no
	not linear	only linearity studied
focus on deterrence	unimportant	unimportant
urban data	low and unsigned	no
youth	importance artificial	positive
US-data	low	insignificant
sample size	low and negative	low and negative

References

Antony, Jürgen; Entorf, Horst (2003), "Zur Gültigkeit Der Abschreckung im Sinne der ökonomischen Theorie der Kriminalität: Grundzüge einer Meta-Studie", *Kriminalität, Ökonomie und der europäische Sozialstaat*, Heidelberg, Physica-Verlag, pp. 167-185.

Baaijens, Sef; Nijkampf Peter (2000), "Meta-Analytic Methods for Comparative and Exploratory Policy Research: An Application to the Assessment of Regional Tourist Multipliers", *Journal of Policy Modelling*, vol. 22, no. 7, pp. 821-858

Becker, Gary (1968), "Crime and Punishment: An Economic Approach", *The Journal of Political Economy*, vol. 76, no. 2, pp. 169-217

Cameron, Samuel (1988), "The Economics Of Crime Deterrence: A Survey Of Theory And Evidence", *Kyklos*, vol. 41, no. 2, pp. 301-323

Düntsch, Ivo; Gediga, Günther (1997). "Statistical evaluation of rough set dependency analysis", *International Journal of Human-Computer Studies*, vol. 46, no. 5, pp. 589-604

Düntsch, Ivo; Gediga, Günther (1998). "Uncertainty measures of rough set prediction" *Artificial Intelligence*, vol. 106, no. 1, pp. 109-137

Düntsch, Ivo; Gediga, Günther (1999, November 29 – last update) "Grobmengen Informations-Analysator", Available: http://www.infj.ulst.ac.uk/~cccz23/grobian/grobian.html (Accessed: 2005, February 8).

Düntsch, Ivo; Gediga, Günther (2000). "Rough set data analysis" *Encyclopedia of Computer Science and Technology*, vol. 43, no. 28, pp. 281-301

Düntsch, Ivo; Gediga, Günther (2001). "ROUGHIAN: Rough information analysis" *International Journal of Intelligent Systems*, vol. 16, no. 1, pp. 121-147

Ehrlich, Isaac (1973), "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation", *The Journal of Political Economy*, vol. 81, no. 3, pp. 521-565

Eibe, Frank; Witten, Ian H. (1998). "Using a Permutation test for Attribute Selection in Decision Trees" In: *Machine Learning: Proceedings of the Fifteenth International Conference, Madison, Wisconsin*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 152-160

Martin, J. Kent (1995), "An Exact Probability Metric for Decision Tree Splitting and Stopping", *University of California, Irvine, Technical Report* 95-16

Müller, Collin (2003), "Die Abschreckungshypothese im ökonomischen Modell der Kriminalität – Eine Meta-Analyse", Unpublished diploma thesis

Nagin, Daniel (1978), "General Deterrence: A Review of the Empirical Evidence", *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates*, Washington, DC: National Academy of Sciences, pp. 95-139

Pawlak, Zdzislaw (1982), "Rough sets", *International Journal of Information and Computer Science*, vol. 11, no. 5, pp. 341-356

Pawlak, Zdzislaw (1991). "Rough Sets – Theoretical aspects of reasoning about data", Kluwer Academic Publisher

Pawlak, Zdzislaw (1999), "Rough classification", *International Journal of Human-Computer Studies*, vol. 51, no. 2, pp. 369-383

Rose, Andrew K. (2004), "A Meta-Analysis of the effect of common currencies on international trade", Working Paper 10373, http://www.nber.org/papers/w10373

Shannon, C. E. (1948), "A Mathematical Theory of Communication", *The Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656

Skowron, A.; Stepaniuk, J. (1991). "Towards an approximation theory for discrete problems: Part I", *Fundamenta Informaticae*, vol. 15, no. 2, pp. 331-362

Stanley, Tom (2001), "Wheat From Chaff – Meta-Analysis As Quantitative Literature Review", *Journal of Economic Perspectives*, vol. 15, no. 3, pp. 131-150

Stanley, Tom (2005), "Beyond Publication Bias", *Journal of Economic Surveys*, vol. 19, no. 3, pp. 309-345

Logic Group, Institute of Mathematics, Warsaw University, Poland (1994-2004), "Rough Set Exploration System, ver. 2.1.1, 1994-2004", Availale: http://logic.mimuw.edu.pl/∼ rses/start.html (Accessed: 2005, February 8).

Ślęzak, Dominik; Ziarko, Wojciech (2003), "Attribute Reduction in the Bayesian Version of Variable Precision Rough Set Model", *Electronic Notes in Theoretical Computer Science*, vol. 82, no. 4

Tsumoto, Shusaku (2002), "Statistical Test for Rough Set Approximation Based on Fisher's Exact test", In: *Lecture Notes in Artificial Intelligence 2475, Springer Verlag* pp. 381-388

Weichselbaumer, Doris; Winter-Ebmer, Rudolf (2003), "A Meta-Analysis of the international Gender Wage gap", *Journal of Economic Surveys*, vol. 19, no. 3, pp. 479-511

Yin, Xuri; Zhou, Zhihua; Li, Ning; Chen, Shifu (2001), "An Approach For Data Filtering Based On Rough Set Theory", In: *Lecture Notes in Computer Science 2118, Springer Verlag*, pp. 367-374

ISSN: 1438-2733