Maximum consistency of incomplete data via non-invasive imputation*

Günther Gediga†
Institut für Evaluation und Marktanalysen
Brinkstr. 19
49143 Jeggen, Germany
gediga@eval-institut.de

Ivo Düntsch†
Department of Computer Science
Brock University
St. Catherines, Ontario, Canada, L2S 3AI
duentsch@cosc.brocku.ca

Abstract

We present an algorithm to impute missing values from given data alone, and analyse its performance. The proposed procedure is based on non–numeric rule based data analysis, and aims to maximise consistency of imputation from known values. In contrast to the prevailing statistical imputation algorithms, it does not make representational assumptions or presupposes other model constraints. Therefore, it is suitable for a wide variety of data–sets, and can be used as a pre-processing step before resorting to harder numerical methods.

Keywords: Imputation, maximal consistency, non–invasive data analysis

1 Introduction

Incomplete data is a major problem in data analysis. There are several ways to "impute" missing values, all of which are based on statistical procedures. The basic idea of these procedures is the estimation of the missing values by minimising a loss function such as the least square measure or the negative likelihood. For an overview of the current practice of working with missing data we invite the reader to consult Acock [1]; for a more complete treatment we recommend the excellent book by Schafer [18].

One major problem of every data mining method is the huge state complexity, even when there are only a few attributes with a few possible values (Table 1).

One observes that, in most cases, real life problems have relatively few data points relative to the number of all possible observations, and therefore, distributional assumptions are often hard to justify; additionally, such problems often show many attribute dependencies. Thus, traditional statistical models are not always optimal tools for data mining.

The famous EM-algorithm [4] was one of the first effective approaches to handle missing data problems on the basis of the likelihood measure. One drawback of the EM algorithm is that it is slow and

*Co-operation for this paper was supported by EU COST Action 274 "Theory and Applications of Relational Structures as Knowledge Instruments" (TARSKI); http://www.tarski.org/

†Equal authorship is implied.

Table 1: State complexity

# of attr. values	# of attributes		
	10	20	30
	$\log_{10}$(states)		
2	3.01	6.02	9.03
3	4.77	9.54	14.31
4	6.02	12.04	18.06
5	6.99	13.98	20.97

costly. Furthermore, it only works if we assume a restricted model class for the data, namely, when the distribution of the data is assumed to be a sample from a fixed distribution family such as the multivariate normal distribution. This is problematic, because the model assumption itself influences the estimation of the missing values.

We have discussed elsewhere [8, 9] that in many instances of data analysis the use of hard statistical methods is neither appropriate nor, indeed, necessary. It is often sufficient to use non–invasive tools which do not make any distributional or other strong assumptions: Dependency rules in the spirit of rough set data analysis (RSDA) [15], statistical significance of symbolic rules by randomisation methods [5], data discretisation by using only classification information ([6], [19]), and model selection by entropy minimisation [7].

In this paper we describe an algorithm to impute missing values from given data alone, and analyse its performance. Our approach is based on non–numeric rule based data analysis. In contrast to statistical procedures, such analysis offers no straightforward way to define loss functions or a likelihood function; these are based on statistical assumptions, which are not given in rule based data analysis. Therefore, other optimisation criteria must be used. A simple criterion is the demand that the rules of the system should have a maximum in terms of consistency, which means if we fill a missing entry with a value, we should result in a rule which is consistent with the other rules of the system. Our algorithm imputes missing values in an attribute vector $\vec{x}$ by presenting a list of possible values drawn from the set of all vectors $\vec{y}$ which do not contradict $\vec{x}$, i.e. they have the same entries wherever both are defined.

The structure of the paper is as follows: We begin with a list of definitions and notation (Section 2), which is followed by a brief overview of current statistical imputation methods (Section 3). The next Section presents our algorithm, and Section 5 analyses its performance by a simulation study. We close with a discussion of our results.

2 Definitions and notation

An *information system* is a structure

$$I = \langle U, \Omega, \{V_a : a \in \Omega\} \rangle,$$

where

- U is a finite set of objects.

- Ω is a finite set of partial functions $a : \operatorname{dom} a \subseteq U \to V_a$; each $a \in \Omega$ is called an *attribute*.
- V_a is the set of *attribute values* of attribute a.

If each a is a function, i.e. $\operatorname{dom} a = U$, then we call I a *complete information system*, otherwise, it is called *incomplete.* If $x \in U \setminus \operatorname{dom} a$, then we say that x *has a missing value at* a. Such a system can be pictured as a data matrix such as Table 2. We write "?" in cell $\langle x, a\rangle$, if a is not defined at x. In our interpretation, the ? is a placeholder for any value in V_a.

For each $x \in U$ and each $\emptyset \neq Q \subseteq \Omega$ we let

$$\operatorname{rel}_Q(x) = \{a \in Q : x \in \operatorname{dom} a\}$$

be the set of *Q-relevant attributes for x.* Let $\vec{x}_Q$ be the feature vector of x with respect to the attributes in Q, i.e.

$$\vec{x}_Q = \langle a(x) : a \in Q\rangle.$$

Here, we assume that the attributes in Ω, and thus in Q, are suitably indexed. Each $\vec{x}_Q$ is called a *Q-granule*; if $Q = \Omega$ or Q is understood, we just speak of *granules.* Observe that

$$\vec{x}_Q = \vec{x}_{\operatorname{rel}_Q(x)}.$$

For each $\emptyset \neq Q \subseteq \Omega$ we define a relation Q_I on U by

$$xQ_Iy \iff a(x) = a(y) \text{ for all } a \in \operatorname{rel}_Q(x) \cap \operatorname{rel}_Q y.$$

If xQ_Iy, we say that x and y are *consistent.* This terminology is justified by the fact that xQ_Iy just in case that whenever a is defined on both x and y, it does not distinguish between them. For example, x and y with

$$\vec{x}_Q = \langle 1, ?, 3\rangle,\ \vec{y}_Q = \langle 1, 4, ?\rangle$$

are consistent, while in case

$$\vec{x}_Q = \langle 1, ?, 3\rangle,\ \vec{y}_Q = \langle 1, ?, 2\rangle$$

they are not. Consistency is a generalisation of the indiscernability concept which is used in RSDA: Two objects x, y are Q-indiscernible, if $\operatorname{rel}_Q(x) = Q = \operatorname{rel}_Q(y)$, and their induced granules are equal. The granules of two consistent objects can be made equal on the union of their relevant attributes by filling in missing values in one granule by values which are defined in the other granule.

It is clear from the definition that Q_I is reflexive and symmetric, but not necessarily transitive. Such relations are sometimes called *similarity relations.*

For $x \in U$ we set $Q_I(x) = \{y \in U : xQ_Iy\}$. The sets $Q_I(x)$ are called *similarity classes.* Clearly,

$$Q_I(x) = \{y \in U : (\forall a \in Q)[a(x) = a(y) \text{ or } a(x) \text{ is not defined or } a(y) \text{ is not defined}]\},$$

We call $x \in U$ *a – casual* (with respect to Q and I), if $a \in Q$, and

$$(\forall y)[y \in Q_I(x) \Rightarrow y \notin \operatorname{dom}(a)]. \tag{2.1}$$

In this case, there is no information for x with respect to attribute a from any granule compatible with $\vec{x}_Q$.

The next result is crucial for our imputation algorithm:

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	?	4.00	5.00
x_4	?	2.00	?	3.00
x_5	2.00	2.00	?	?
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	?

Table 2: Missing data table

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	4.30	3.00	2.00	1.00
x_3	2.00	2.00	4.00	5.00
x_4	2.39	2.00	2.28	3.00
x_5	2.00	2.00	1.62	3.11
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	4.17

Table 3: Single imputation via iterated linear regression

Lemma 2.1. *If* $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$ *and* $a(x) = a(y)$ *for all* $a \in \mathrm{rel}_Q(x)$*, then* $Q_I(y) \subseteq Q_I(x)$.

Proof. Let $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$, $a(x) = a(y)$ for all $a \in \mathrm{rel}_Q(x)$, and assume $Q_I(y) \not\subseteq Q_I(x)$. Then, there is some $z \in U$ such that

1. Whenever $a \in Q$ is defined for y and z, then $a(y) = a(z)$.
2. There is some $a_0 \in Q$ such that $a_0(x)$ and $a_0(z)$ exist, and $a(x) \neq a(z)$.

By the assumption $\mathrm{rel}_Q(x) \subseteq \mathrm{rel}_Q(y)$, a_0 is defined for y as well. Since $a_0(y) = a_0(z)$ by 1., we have $a_0(x) \neq a_0(y)$, contradicting $a(x) = a(y)$ for all $a \in \mathrm{rel}_Q(x)$. □

3 Imputation methods

Suppose that $I = \langle U, \Omega, \{V_a : a \in \Omega\}\rangle$ is an information system which is not complete. In order to cope with the missing data entries several strategies can be used. The simplest one is ignorance, which means that any observation with missing entries is skipped from further analysis. There is well-documented evidence to show that ignorance is usually a bad strategy [13].

A second stream of methods uses a *single imputation* strategy. Here, a missing value is replaced by the best suited replacement, where "best suited" is defined in terms of a statistical model. Such models usually make some distributional assumption such as a multinomial or a multivariate normal

distribution, e.g. the treatment of missing data in the AMOS system [2], or the EQS system [3], or they use mixed multinomial/normal models [18].

Even though the approach has come under criticism [10], we use the *iterated linear regression algorithm for single imputation* as an illustrative example to show how missing data can be replaced by a simple statistical technique.

With the data of Table 2, this results in the underlined imputed values in Table 3. This method estimates the missing values by linear regression of the other variables in a form such as

$$\begin{aligned}
\hat{a}_1 &= b_{0,1} + b_{2,1}a_2 + b_{3,1}a_3 + b_{4,1}a_4, \\
\hat{a}_2 &= b_{0,2} + b_{1,2}a_1 + b_{3,2}a_3 + b_{4,2}a_4, \\
\hat{a}_3 &= b_{0,3} + b_{1,3}a_1 + b_{2,3}a_2 + b_{4,3}a_4, \\
\hat{a}_4 &= b_{0,4} + b_{1,4}a_1 + b_{2,4}a_2 + b_{3,4}a_3,
\end{aligned}$$

where $b_{i,j}$ are constants which are optimal to fit the regression line for the prediction of a_j by the attributes $a_i,\ i \neq j$. Because replacing the missing entries $a_i(x_k)$ by their expectation $\hat{a}_i(x_k)$ – assuming the validity of the linear regression model – the entries will change, and therefore the basis of the model estimates changes as well. In most applications many iteration steps are necessary to result in a stable optimal configuration.

Furthermore, the critical assumption of this imputation model, namely, that there is a linear relationship among the variables, cannot be assumed in general. Even if the relationship is linear, the procedure faces another problem: The existence of only one outlier will bias the estimation of the *b*-values dramatically.

Of course, many non-linear relationships can be modelled simply by a non-linear transformation of the variables, but a sound model of the data is necessary to achieve a good description of the missing value. A badly chosen model can make the things even worse than the ignorance strategy, and hence, statistical imputation should be used with care. Detailed discussions of the interplay between the model used for imputation and the model used for analysis can be found in [14], and [17].

A third stream of methods employs the *multiple imputation* strategy [16]. Here, the data–set is replaced by a number of "mutual" data sets in order to simulate the uncertainty about the missing values. Every data set can then be used within the data analysis, and a model based aggregation scheme enables the results to be combined.

4 Non-invasive imputation

The procedure proposed in this paper is located in the broad class of rule based reasoning algorithms. This class of procedures generates rules for prediction and can be very simple – for example, the multiplication of values [11] – or more complex – for example, cluster analysis or decision trees [12]. We call our approach a *non-invasive* imputation method, because it takes all its information from the given data and makes no additional dependency or distributional assumptions. A seeming disadvantage of this approach is that the procedure cannot inter-/extrapolate into unknown regions like the regression method, because there is no mechanism for inter-/extrapolation. The procedure uses only rules and dependencies within the body of the data, and therefore missing values can only replaced by known facts. In our view this is no disadvantage at all: If the procedure cannot replace the

missing value it will signal a "do not know"- sign, which seems more reliable (and honest) than any extrapolation based on strong model assumptions when these assumptions may not be fully justified.

It is our aim to transform an incomplete system I into a complete system. If the granule $\vec{x}_Q$ has a missing value at, say, $a \in Q$, we will try to impute it from the a-values of the objects in the similarity class of x. This will not always be possible, and, if it is, there may not be a unique value. Thus, the result of the imputation process will in some (or many) cases be a list of values from which a value may be picked, possibly by other methods, without violating consistency.

Let us define a mapping $m : U \times \Omega \to \bigcup_{a\in\Omega} 2^{V_a}$ which will give us the possible imputable values by collecting for each $x \in U$ and each $a \in Q$ those entries which appear as entries $a(y)$ in the granules induced by a $y \in U$ which is consistent with x.

$$m(x,a) = \begin{cases} a(x), & \text{if } a(x) \text{ is defined,} \\ \{a(y) : y \in Q_I(x)\}, & \text{if } a \text{ is not defined at } x, \\ & \text{but } a \text{ is defined for some } y \in Q_I(x), \\ ?, & \text{otherwise.} \end{cases}$$

We see that m does not change unique values; furthermore, if a is not defined at any $y \in Q_I(x)$, i.e. if x is a – casual, then we will not be able to fill the entry $\langle x,a\rangle$; in this case, there is no "collateral knowledge" for $\langle x,a\rangle$. This is the case, when a rule is based on only one granule; we have discussed this briefly in Düntsch and Gediga [7].

Based on Lemma 2.1, we can now give a non-invasive imputation algorithm.

Algorithm 1. Define a sequence of information systems as follows:

1. $I_0 = I$.
2. Suppose that $I_k = \langle U, \Omega_k, \{V_{a^k} : a^k \in \Omega_k\}\rangle$ is defined for some $k \geq 0$.
   (a) Find the similarity classes $Q_{I_k}(x)$.
   (b) For each $a^k \in \Omega_k,\ x \in U$, let
   $$a^{k+1}(x) = \begin{cases} m(x,a^k), & \text{if } |m(x,a^k)| = 1, \\ ?, & \text{otherwise.} \end{cases}$$
   (c) Set $\Omega_{k+1} \stackrel{\text{def}}{=} \{a^{k+1} : a^k \in \Omega_k\}$ and $V_{a^{k+1}} \stackrel{\text{def}}{=} V_{a^k}$.

With this procedure, we successively extend the attribute mappings; in other words, we increase (or leave constant) $\mathrm{rel}_\Omega(x)$. In this process, we reduce (or leave constant) the similarity classes $Q_{I_k}(x)$, and Lemma 2.1 now tells us that there is an n such that $Q_{I_n}(x) = Q_{I_r}(x)$ for all $x \in U$ and $n \leq r$. At this step we report the matrix $\langle m(x,a)\rangle_{a\in\Omega_n}$. If $m(x,a)$ has more than one element, this set will give us the possibilities for a value $a(x)$, based on previous experience.

The steps taken by this method for the example data are given in Table 4. We have listed the intermediate sets $m(x,a)$ to indicate how the compatibilities change after we have extended attribute functions in one step. After the first step we note that we cannot make the table complete, since x_2 is a_1-casual.

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	$\{a_2(x_5)\}$	4.00	5.00
x_4	$\{a_1(x_5), a_1(x_8)\}$	2.00	$\{a_3(x_8)\}$	3.00
x_5	2.00	2.00	$\{a_3(x_3)\}$	$\{a_4(x_3), a_4(x_4)\}$
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	$\{a_4(x_4)\}$

Table 4: Non-invasive imputation I

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	$\{2.00\}$	4.00	5.00
x_4	$\{a_1(x_8)\}$	2.00	$\{5.00\}$	3.00
x_5	2.00	2.00	$\{4.00\}$	$\{a_4(x_3)\}$
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	$\{3.00\}$

Table 5: Non-invasive imputation II

There is some ambiguity for the observations x_4 and x_5 which cannot be resolved in the first step. The reason is that there are – at this step – consistent replacements of the missing values in x_3, x_5, and x_8 respectively, which allow us to build a suitable granule for the prediction of the missing values of x_4 and x_5. Since the similarity classes are reduced in step 2, there are fewer possibilities for replacement, and, indeed, all ambiguities can be resolved.

One might argue that there is a bias towards one element sets $m(x,a)$, since we always fill those in first, and then compute the similarity classes. If we have committed ourselves to minimise the number of remaining ? and fill in whatever we can, we have no other choice: We must impute singletons first, since they are all we have in such an instance. This procedure leads to the least number of remaining ? cells.

	a_1	a_2	a_3	a_4
x_1	5.00	4.00	3.00	2.00
x_2	?	3.00	2.00	1.00
x_3	2.00	$\{2.00\}$	4.00	5.00
x_4	$\{3.00\}$	2.00	$\{5.00\}$	3.00
x_5	2.00	2.00	$\{4.00\}$	$\{5.00\}$
x_6	5.00	4.00	3.00	2.00
x_7	3.00	2.00	1.00	1.00
x_8	3.00	2.00	5.00	$\{3.00\}$

Table 6: Non-invasive imputation III: Final state

The simple multiplication of Grzymała-Busse [11] replaces one observation by a list of new observations each of which fills in the missing value(s) by one of the possibilities; this procedure assumes that every value is "legitimate". The first step of our algorithm, though seemingly similar, differs in one important aspect: We replace a missing value with a set of possible values, and thus, implicitly, assume a disjunctive connection among the possible observations, while adding new observations to the information system induces a conjunctive relationship.

5 A simulation study

In this Section we explore the behaviour of our imputation procedure by a simulation study. While a concrete example may be an indicator of a procedure's usefulness and limitation – but nothing more –, a simulation study systematically varies the variables which may influence the quality of results, and thus it sheds a broader light on the issues under consideration.

The variable parameters considered are

Number of different granules	$n = 500, 400, 300, 200, 100$
Number of attributes	$k = 10, 8, 6, 4$
Number of attribute values	$m = 6, 5, 4, 3, 2$
Percentage of missing values	$p = 0.2, 0.15, 0.10, 0.05, 0.02.$

Not all combinations were possible, since we have to observe

$$n \leq m^k.$$

For each combination we have randomly generated 100 information systems, each with $N = 500$ objects. We have used the following additional variables:

$NOG = \frac{N}{n},$	Average observations per granule
s	Number of missing values
e	Number of replacement errors
$E = \frac{e}{s},$	Error rate
$SC = k \cdot \ln(m).$	State complexity

A replacement error occurs, when the original value is not in the list of suggested values. Observe that an error is only possible, if each compatible instance of the granule contains a missing value; in other words, if there is a compatible granule without missing value, an error cannot occur.

Furthermore, we let $T = \{t_i : 1 \leq i \leq s\}$ be the set of all suggested lists of replacements, including the one-element lists.

We have the considered the mean (μ) and variance (σ) of the following functions[1]:

[1]The tables are available from `http://www.roughian.co.uk/papers/missdata/misshtml.html`

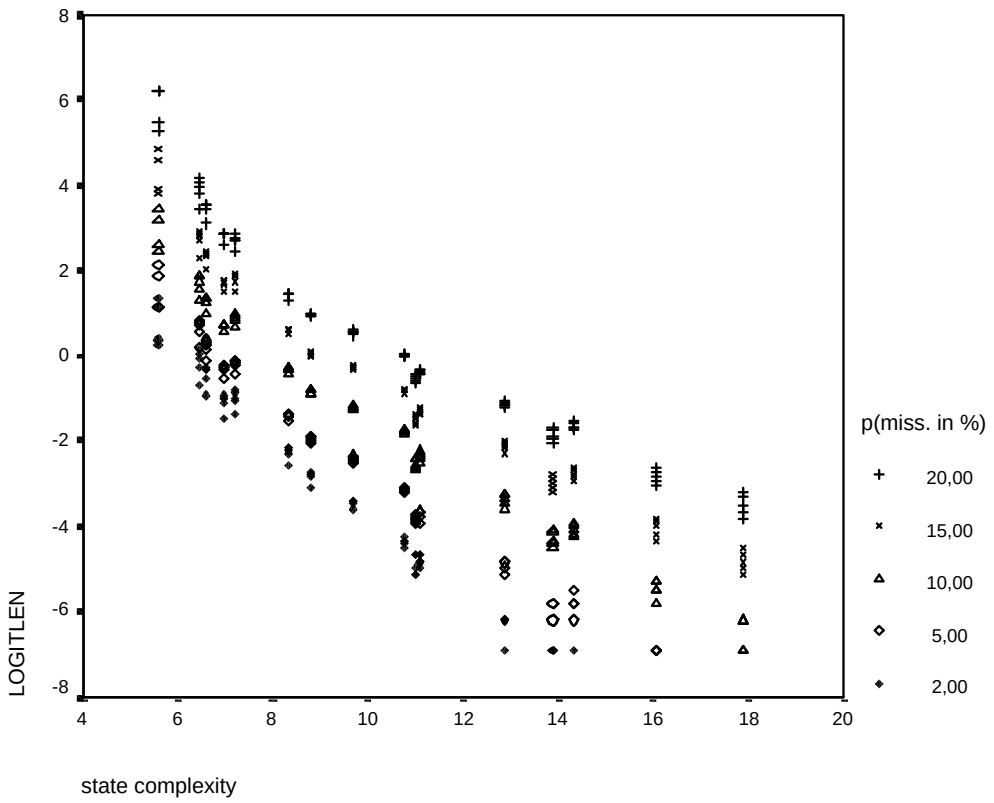

Figure 1: Plot of logit(len) with state complexity and p

1. The percentage of missing values which were replaced by one value.
2. len $= \frac{\sum_{i=1}^{s} |t_i|}{s \cdot m}$, the average size of a replacement list when taken over all lists, relative to the number m of possible values. This function serves as a measure of uncertainty.
3. The number of replacement errors.

We have replaced every remaining ? by a list of all values which have occurred in the appropriate column, which expresses the observation that no other data entry can be used to reduce uncertainty for this case, and therefore every entry will be consistent.

The simulation results give rise to the question which variables determine the uncertainty len and the prediction error rate E. Because E and len are percentages in a range from near 0 to near 1, we use the logit transformation ($\text{logit}(u) = \ln(\frac{u}{1-u})$) of E as an error indicator, and logit(len) as an indicator of uncertainty. This enables us to use a linear model based on the transformed variables.

Figure 1 shows that only two variables are needed to describe the results of logit(len): The regression line

$$\text{logit}(\text{len}) = 3.372 - 0.723SC + 0.244p, \qquad R^2 = 0.959$$

fits the data up to a small amount of error, because the R^2 measure indicates that the regression line expresses 95.9% of the squared differences of logit(len) and only the amount of 4.1% of the

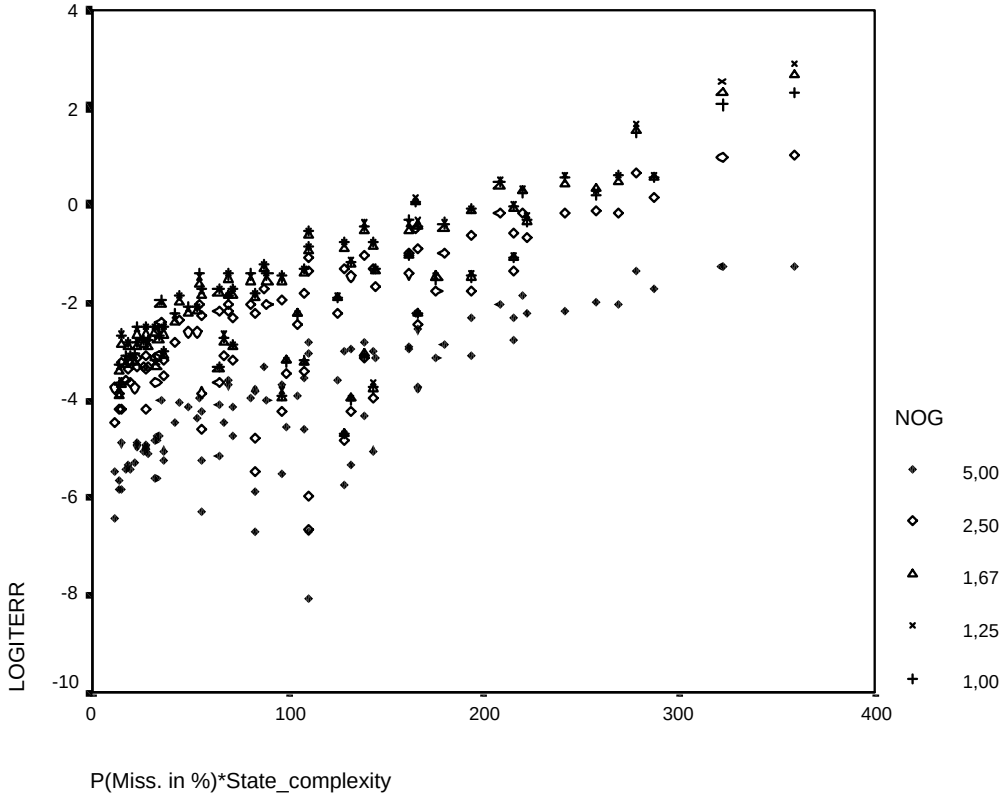

Figure 2: Plot of logit(E) with number of observations per granule (NOG) and $p \cdot SC$

squared differences of logit(len) remains unexplained. Inspection of Figure 1 shows that the remaining unexplained variance may be caused by an additional small non-linear effect. The prediction success (in terms of R^2) of the entropy of the underlying distributions is by far lower than the prediction success of *SC*.

The error data are more complex. In order to find out which variables are necessary to predict logit(E) we have performed a multiple regression analysis resulting in

$$\text{logit}(E) = -1.992 + 0.02674 p \cdot SC - 0.0594 NOG - 0.185 p, \qquad R^2 = 0.892.$$

The R^2-measure demonstrates that the model fits the data up to small non-linear departures and a small amount of additional noise. Some significant enhancements of the R^2-measure are possible, but the impact of the additional variables in the enhanced regression equations is by far lower than the one reported above.

Figure 2 shows the dependency of E on the number of observations per granule (*NOG*) and the variable $p \cdot SC$. The figure shows that despite the variation in the simulation, a rule of thumb "5 observations per granule" leads to acceptable behaviour of the procedure – note that this means "5 observations per granule including the full list at remaining missing values". Once again: It is the state complexity and not the entropy which is a suitable predictor for the results.

In order to describe most of the variance of logit(E), the relative number p of missing values has to be

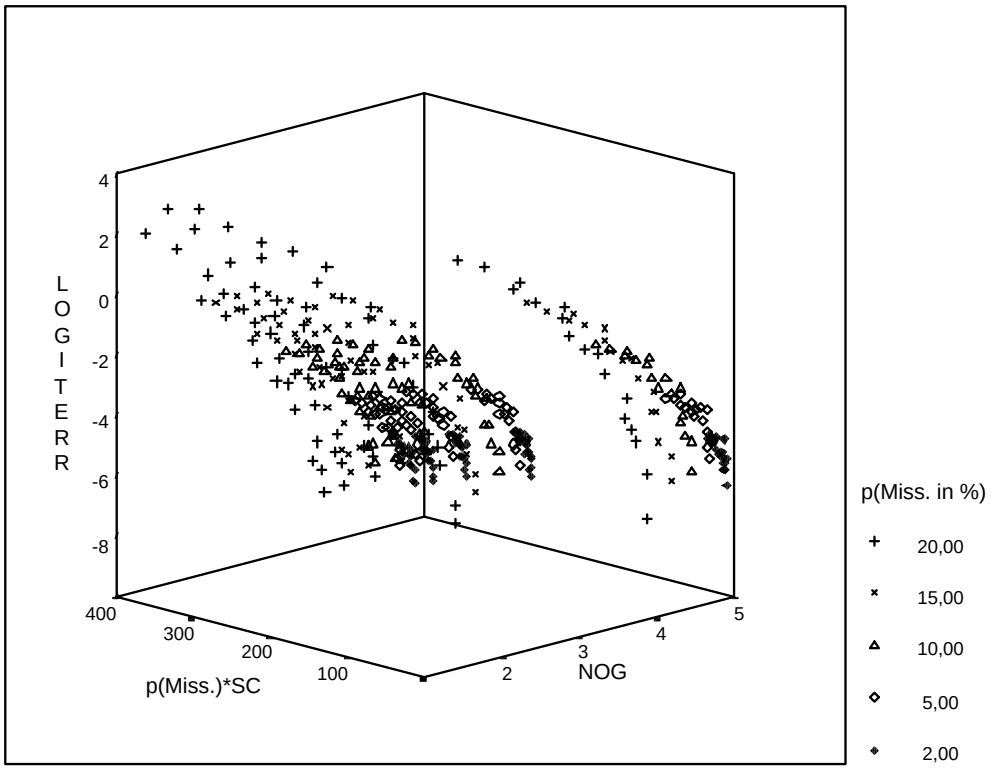

Figure 3: Plot of logit(E) with number of observations per granule (NOG), $p \cdot SC$, and p

used as an additional predicting variable. The 4-dimensional plot of these relationships is presented in Figure 3.

As the figures have indicated, the error measure E depends negatively on the number of observations per granule, and positively on the state complexity and relative number of missing values.

6 Conclusion

The technique of replacing missing values by maximising the consistency of the data–set offers a non-invasive alternative to the missing value estimations by statistical methods, which is very quick in comparison to the time required by statistical procedures: For every instance of the simulated data we observe convergence after at most 5 cycles.

The advantage (and perhaps the drawback) of the procedure is that it will result in all possibilities of replacements by known values which allow interpretations consistent with the observed data, and only in those. It may serve as a first instrument to look at admissible replacements of missing data in terms of consistency. The number of possible replacements is negatively correlated with state complexity. This is a price one has to pay if one wants to know the possible (and not the probable best) replacements: If the state complexity grows, the chance of mutual replacements which are in

the underlying distribution of the data rises; if the number of missing values grows as well, the risk of resulting in many replacements grows multiplicatively.

The intention of the proposed procedure is to present to the user a picture of what happens if missing values are imputed that are consistent with the given data – a different goal to finding a (statistical) procedure to estimate a model among variables. This interplay of non-invasive computing and more demanding statistical modelling is intended: Non-invasive computing shows which results are possible from (and consistent with) the obtained data – statistical modelling offers the most probable solution of the problem.

Acknowledgement

We would like to thank the referees for their constructive remarks, which helped to improve the quality of the paper.

References

[1] Acock, A. (1997). Working with missing data. *Family Science Review*, 10:76–102.

[2] Arbuckle, J. (1996). *Amos Users Guide: Version 3.6*. SmallWaters Corp., Chicago.

[3] Bentler, P. (1996). *EQS: Structural Equations Program Manual*. BMDP Statistical Software, Los Angeles.

[4] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society (B)*, 39:1–38.

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

[6] Düntsch, I. and Gediga, G. (1998a). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, 18(1–2):93–106.

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

[8] 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.

[9] Düntsch, I. and Gediga, G. (2001). Roughian – Rough Information Analysis. *International Journal of Intelligent Systems*, 16(1):121–147.

[10] Graham, J. W., Hofer, S. M., and Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In Collins, L. M. and Seitz, L., editors, *Advances in Data Analysis for Prevention Intervention Research*, Washington. NIDA Research Monograph. Series 142.

[11] Grzymała-Busse, J. (1991). On the unknown attribute values in learning from examples. In *Proc of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems*, volume 542 of *Lecture Notes in Artificial Intelligence*, pages 368–377. Charlotte.

[12] Lakshminarayan, K., Harp, S. A., Samad, T., and Goldman, R. P. (1996). Imputation of missing data using machine learning techniques. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 140–145, Menlo Park. American Association for Artificial Intelligence.

[13] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[14] Meng, X. L. (1995). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 10:538–573.

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

[16] Rubin, D. B. (1987). *Multiple Imputations for Nonresponse in Surveys*. Wiley, New York.

[17] Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91:473–489.

[18] Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

[19] Wang, H., Düntsch, I., and Gediga, G. (2000). Classificatory filtering in decision systems. *International Journal of Approximate Reasoning*, 23:111–136.