KYBERNETIKA — VOLUME 16 (1980), NUMBER 2

An Alternative Approach to Missing Information in the GUHA Method

TOMÁŠ HAVRÁNEK

In the present paper some new approaches to the treatment of missing data items in the GUHA procedures are presented. The first one is the dual approach to the usual conservative approach used in present GUHA procedures. Both these approaches are independent of the pollution proces destroying the original data matrix with no missing data items. The second new approach is principially based on erasing objects with missing data items in just processed parts of data.

In the GUHA method, since [2], [1] a "pesimistic" way for the treatment of missing information has been applied, i.e. generalized quantifiers and/or underlying two-valued decision functions are extended to data containing missing information in the following conservative way:

Data are finite structures of the form $\boldsymbol{M} = \langle M, f_1, \ldots, f_n \rangle$, where $M \neq \emptyset$, $M \subseteq \mathcal{N}$ and each $f_i : M \to V$; n is called the type of data (elements of M are understood as names of observed objects and $f_i(o)$ for an object $o \in M$ as observed value of some investigated quantity). Data containing missing information are such that $f_i : M \to V \cup \{\times\}$, where the value $\times$ means "unspecified", "unobserved" etc. Now for given M, V denote $\mathscr{M}^{\times}_{V,M}$ the class of all structures (data) containig (possibly) missing information. By $\mathscr{M}_{V,M}$ we denote structures with no missing values. Denote $\mathscr{D} = \mathscr{P}_{\mathrm{fin}}(\mathcal{N}) - \{\emptyset\}$. Then put $\mathscr{M}_V = \bigcup_{M \in \mathscr{D}} \mathscr{M}_{V,M}$. Similarly for $\mathscr{M}^{\times}_V$. If $V = \{0, 1\}$, we write only $\mathscr{M}_M$ etc. Let now $\boldsymbol{M} \in \mathscr{M}^{\times}_{V,M} - \mathscr{M}_{V,M}$; data $\boldsymbol{N} \in \mathscr{M}_{V,M}$ are called a *completion* of $\boldsymbol{M}$ if all values $\times$ in $\boldsymbol{M}$ are in $\boldsymbol{N}$ replaced by some values from V and the rest is left unchanged.

The underlying philosophy is, that data with no missing information have been damaged by a "pollution" process and we have to make some statistical decisions without knowledge of the original data – "true" completion. No assumptions concerning the pollution process are done.

A decision function $f : M_V \to \{0, 1\}$ (i.e. defined on data with complete information can be *conservatively* extended to incomplete data (i.e. onto $\mathcal{M}_V^{\times}$) as follows:

Put

$$f(M) = \begin{cases} 1 & \text{if for all completions } N \text{ of } M, f(N) = 1\,, \\ \times & \text{otherwise}\,, \\ 0 & \text{if for all completions } N \text{ of } M, f(N) = 0\,. \end{cases}$$

This is in accordance with the meaning of $\times$ as "unknown". Usually an ordering of values $0 < \times < 1$ is supposed. Then conservative extension can be obtained by using a "treshold" value 1. In [1] such an extension is called *secured* extension (some decision is accepted iff $f(M) \geqq 1$). Cf. also [8].

Logical connectives $\vee$, & and $\neg$ can be extended using the "securedness" principle: Put the value 1 (or 0) if you are sure that the value must be 1 (or 0). Then we obtain for $\vee$, &, $\neg$ the following tables:

$\neg$	
0	1
$\times$	$\times$
1	0

,

&	0	$\times$	1
0	0	0	0
$\times$	0	$\times$	$\times$
1	0	$\times$	1

,

$\vee$	0	$\times$	1
0	0	$\times$	1
$\times$	$\times$	$\times$	1
1	1	1	1

.

The above conservative way can be criticised at least from the following two reasons (think decision 1 as acceptance of an alternative hypothesis):

(i) It is in contradiction with the philosophy of the GUHA methods (c. f. for example the introductory paper [4] – searching of all "interesting" in data and interestingness is equivalent with acceptance); it overestimates local reliability of results on the account of the global power of the method.

(ii) In practice, one can see that even relatively small "pollution" of data can lead to heavy consequences in omitting many results, which can be obtained e.g. if we exclude all objects with some missing data item from the processed data.

Hence some alternative ways for the treatment of missing information are to be suggested. The critical point of such suggestions is the possibility to implement new alternative decision functions into present computer GUHA procedures without destroying their architecture. The further suggested extensions will be discussed from this point of view, hence the present paper is rather technical. A preliminary version of this paper is obtained in the report [7].

Some important questions concerning so called helpful quantifiers connected with the presentend approach is postponed into the forthcomming paper [5].

1.0. The "optimistic" extension of a decision function is defined as follows: $f(M) = 1$ if there is a completion N of M such that $f(N) = 1$, otherwise $f(M) = 0$.

1.1. Or, in three valued logic we use threshold value $\times$: set of designed values (interesting values) is defined as $\{1, \times\}$. It means that wee seek for "potentially" true sentences. For the sake of simplicity, we follow only the way 1.0. without any lost of generality.

1.2. In the rest of paper we use freely technical means of observational functor calculi as described in [3] (or partially in [1]). We are concerned in generalized quantifiers defined by some decision functions. If q is such a quantifier, its defining function is denoted Asf_q and it is called the associated function of q. We shall use open formulas with one variable x (and hence we do not write the variable explicitely).

For example if q joins two open formulas (names of composite qualities) φ, ψ, the value of $q(\varphi, \psi)$ on a structure of an appropriate type is defined by $\|q(\varphi, \psi)\|_M = \mathrm{Asf}_q(M_{\varphi,\psi})$, where $M_{\varphi,\psi} = \langle M, \|\varphi\|_M, \|\psi\|_M\rangle$ (note, that $\|\varphi\|_M$ is a function on M, the domain of M). Moreover, we shall assume that $V = \{0, 1\}$.

1.3. For associational quantifiers (cf. 3.2.2 of [3]) we have to check its "associotionality" for optimistic extension. Consider quantifiers of the type 2 (i.e. joining two open formulas). Their associated functions are defined on structures of the form $M = \langle M, f_1, f_2\rangle$. For each pair $u = \langle u, v\rangle \in \{0, \times, 1\}^2$, define $\mathrm{fr}(M, u) = \mathrm{card}\{o \in M; \langle f_1(o), f_2(o)\rangle = u\}$. Then we denote $a(M) = \mathrm{fr}(M, \langle 1, 1\rangle)$, $b(M) = \mathrm{fr}(M, \langle 1, 0\rangle)$, $c(M) = \mathrm{fr}(M, \langle 0, 1\rangle$, $d(M) = \mathrm{fr}(M, \langle 0, 0\rangle)$.

A *quantifier is associational* if $\mathrm{Asf}_q(M) = 1$ and $N \geqq^a M$ implies $\mathrm{Asf}_q(N) = 1$, where $\leqq^a$ is the a – better relation, i.e. $M \leqq^a N$ if $a(M) \leqq a(N)$, $b(M) \geqq b(N)$, $c(M) \geqq c(N)$, $d(M) \leqq d(N)$. Similarly for $\geqq^i$, the i – better ordering: $M \leqq^i N$ if $a(M) \leqq a(N)$, $b(M) \geqq b(N)$, we can define *implicational quantifiers.*

For incomplete data, define $i(M) = \mathrm{fr}(M, \langle 1, \times\rangle)$, $o(M) = \mathrm{fr}(M, \langle \times, 1\rangle)$, $n(M) = \mathrm{fr}(M, \langle \times, \times\rangle)$, $p(M) = \mathrm{fr}(M, \langle \times, 0\rangle)$, $q(M) = \mathrm{fr}(M, \langle \times, 0\rangle)$.

1.4. Let $M \in M^\times$. Let M^b is any completion of M such that $a(M^b) = a(M) + i(M) + o(M) + n(M)$ and $b(M^b) = b(M)$. Then M^b is called the *best implicational completion of M.*

Theorem. For each implicational quantifier $\to$ and its optimistic extension $\to^*$ we have $\mathrm{Asf}_{\to^*}(M) = 1$ iff $\mathrm{Asf}_{\to}(M^b) = 1$ for the best implicational completion of M.

Proof. Clearly, for each other completion N of M, $a(N) \leqq a(M^b)$ and $b(N) \geqq b(M^b)$.

1.5. Our first aim is now to inspect observational inference rules (immediate

consequence rules) used in the implicational GUHA procedure as concerns their soundness for the optimistic extension of implicational quantifiers.

1.6. Theorem. The despecifying – dereducing inference rule (3.2.20 [3]) is sound for optimistic extension of any implicational quantifier.

Proof. Let $\rightarrow^*$ be an optimistic extension of an implicational quantifier. The inference rule in question is defined as

$$\mathrm{SpRd} = \left\{ \frac{\varphi \,\&\, \psi \rightarrow^* \delta}{\varphi \rightarrow^* \delta \vee \neg\psi \vee \chi} \; ; \begin{array}{l} \varphi, \psi \text{ elementary conjunctions,} \\ \delta, \chi \text{ elementary disjunctions and} \\ \varphi, \psi, \delta, \chi \text{ mutually disjoint .} \end{array} \right\}$$

Consider two steps (using transitivity of the rule SpRd):

(i) Let $a, b, \ldots$ concern the structure $\boldsymbol{M}_{\varphi,\delta}$, $a', b', \ldots$ the structure $\boldsymbol{M}_{\varphi,\delta\vee\chi}$ (both derived from a structure $\boldsymbol{M}$). Inspect the following Table 1:

Table 1.

frequencies	φ	δ	χ	frequencies
a	1	1	everything	a'
	1	0	1	
	1	×	1	
i	1	×	0	i'
	1	×	×	
	1	0	×	
	×	1	0	
	×	1	×	
o	×	1	1	o'
	×	0	1	
	×	×	1	
	×	×	0	
	×	×	×	
n	×	×	×	n'
	×	0	×	

Then $a(\boldsymbol{M}^b_{\varphi,\delta\vee\chi}) = a' + i' + o' + n' \geqq a + i + o + n = a(\boldsymbol{M}^b_{\varphi,\delta})$ and $b(\boldsymbol{M}^b_{\varphi,\delta\vee\chi}) \leqq$ $\leqq b(\boldsymbol{M}^b_{\varphi,\delta})$.

(ii) Similarly for $\varphi \,\&\, \psi \rightarrow^* \delta$ and $\varphi \rightarrow^* \delta \vee \neg\psi$: let $a, b, \ldots$ concern $\boldsymbol{M}_{\varphi\&\psi,\delta}$ and $a', b', \ldots$ concern $\boldsymbol{M}_{\varphi,\delta\vee\neg\psi}$. Inspect the Table 2:

Table 2.

frequencies	cards φ	ψ	δ	$\neg\psi$		frequencies
a	1	1	1	0		a'
	1	ev.	1	ev.		
	1	0	×	1		
	1	0	0	1		
i	1	1	×	0		i'
	1	×	×	×		
	1	×	0	×		
o	×	1	1	0		o'
	1	×	1	×		
	×	×	1	×	$\to a'$	
	×	1	1	0		
	×	0	1	1		
	×	0	0	1		
	×	×	0	×		
n	×	×	×	×		n'
	×	1	×	0		
	1	×	×	×	$\to i'$	
	×	ev.	×	0		
	×	ev.	0	×		
	×	ev.	×	×		

Then $a(M^b_{\varphi,\delta\vee\neg\psi}) = a' + i' + o' + n' \geqq a + i + o + n = a(M_{\varphi\&\psi,\delta})$ and $b(M^b_{\varphi,\delta\vee\neg\psi}) = b(M^b_{\varphi\&\psi,\delta})$.

1.7. For a $\{0, \times, 1\}$ – structure M of the type 2, define the *best symmetric completion* as any completion M^b such that: $a(M^b) = a(M) + i(M) + o(M) + n_1$, $d(M^b) = d(M) + j(M) + p(M) + n_2$, $c(M^b) = c(M)$, $b(M^b) = b(M)$ with $n_1, n_2 \in N$ such that $a(M^b)\,d(M^b) = \max_{g+h=n(M),\, g,h\in N} (a(M) + i(M) + o(M) + g)\,(d(M) + j(M) + p(M) + h)$.

1.8. Theorem. Let $\sim$ be one of the quantifiers $\sim$, $\sim_\alpha$, $\sim^2_\alpha$ (i.e. in words, the simple associational, Fisher and chi-square quantifier). Then for each M and each best symmetric completion M^b of M, we have $\mathrm{Asf}_{\sim^*}(M) = 1$ iff $\mathrm{Asf}_{\sim}(M^b) = 1$.

1.9. We first prove the following lemma: For each associational quantifier $\sim$ we have the following: for each M, $\mathrm{Asf}_{\sim^*}(M) = \mathrm{Asf}_{\sim^*}(M^*)$, where $M^* = M(\langle 1, \times\rangle : \langle 1, 1\rangle)(\langle \times, 0\rangle : \langle 0, 0\rangle)(\langle \times, 1\rangle : \langle 1, 1\rangle)(\langle 0, \times\rangle : \langle 0, 0\rangle)$, i.e. M^* is obtained from M replacing each card $\langle 1, \times\rangle$ by $\langle 1, 1\rangle$ etc.

Proof. (i) for each completion of M^* we can construct an a-better completion of M; clearly each completion of M^* is a completion of M, (ii) Let N is a completion of M, then changing completed as cards suggested above we obtain an completion of M^* a-better than N.

1.10. Proof of Theorem 1.1. For quantifiers $\sim$ and $\sim_\alpha^2$ the proof is now trivial, due their symmetry w. r. t. $|a-d|$ and $|b-c|$. We have to treat the case of $\sim_\alpha$; namely prove the monotonicity of the statistic

$$\Delta(a, r, k, m) = \sum_{i=a}^{\min(r,k)} \sigma(i, r, k, m),$$

where $\sigma(i, r, k, m) = r!\, s!\, k!\, l!/m!\, i!\, b!\, c!\, d!$ with $b = k - i$, $c = r - i$, $s = m - r$, $l = m - k$, $d = m + i - r - k$. Consider now tables completed due to the Lemma 1.9.:

$a + n'$	b	$k + n'$
c	$d + n - n'$	$l + n - n'$
$r + n'$	$s + n - n'$	m

where $n' \in [0, n]$. We prove first an auxiliary inequality for the contingency table

a	b	k
c	d	l
r	s	m

with $a \geqq d$ and $ad > bc$. Consider inequality

$$sl/d \leqq rk/a . \tag{1}$$

We can use the following equivalent forms of (1):

$$(a + d)(d + c)/d \leqq (a + c)(a + b)/a ; \tag{2}$$

$$d + bc/d \leqq a + bc/a ; \tag{3}$$

$$d(ad) + a(bc) \leqq a(ad) + d(bc) \tag{4}$$

and, due $d - a \leqq 0$, from $(d - a)\, ad \leqq bc\, (d - a)$ we have $ad \geqq bc$, which is true. Due to the symmetry of the statistic we can suppose that in any case $a + n' \geqq$ $\geqq d + n - n'$ and prove the monotonicity of $H(n') = \Delta(a + n', k + n', r + n', m)$ for n' increasing. I.e. we have to prove $H(n') \leqq H(n' + 1)$ for each n' such that $a + n' \geqq d + n - n'$ and $n' + 1 \leqq n$. Both sums have the same number of members; we can compare $\sigma(i + n', r + n', k + n', m)$ and $\sigma(i + n' + 1,\ r + n' + 1,$

$k + n' + 1, m)$ for $i = a, \ldots, \min(k, r)$. Denote $i' = i + n'$, $r' = r + n'$ etc. The desired inequality is then: $\sigma(i', r', k', m) \leqq \sigma(i' + 1, r' + 1, k' + 1, m)$ for $i' = a' = a + n', \ldots, \min(k', r')$. We can reduce

$r!\,(m - r')!\,(m - k')!\,k'!/m!\,i'!\,(k' - i')!\,(r' - i')!\,(m + i' - r' - k')! \leqq (r' + 1)!\,(m - r' - 1)!\,(m - k' - 1)!\,(k' + 1)!/m!\,(i' + 1)!\,(k' + 1 - i' - 1)!\,(r' + 1 - i' - 1)!\,(m + i' + 1 - r' - 1 - k' - 1)!$ to

$$(5) \qquad (m - r')\,(m - k')/(m + i' - r' - k') \leqq (r' + 1)\,(k' + 1)/(i' + 1)$$

for $i' = a', \ldots, \min(r', k')$. First, note that

$r'k'/i' \leqq (r' + 1)\,(k' + 1)/(i' + 1)$; the last inequality is equivalent to $r'k'i' + r'i' + k'i' + i' \geqq r'k'i' + r'k'$ which is equivalent to $i'(r' + k') + i \geqq r'k'$ and then $a'((a' + b') + (a' + c')) + a' \geqq (a' + b')\,(a' + c')$. It is equivalent to simple inequality $a'^2 + a' \geqq bc'$; using $a' \geqq d'$ we obtain $a'd' \geqq b'c'$ which is true, and from which the previous inequality follows. Now we can consider instead of (5) the inequality

$$(6) \qquad (m - r')\,(m - k')/(m + i' - r' - k') \leqq r'k'/i' .$$

For $m - r' - k' \geqq 0$ the ratio $(m + i' - r' - k')/i'$ is decreasing function of i', hence we can test (6) for $i' = k'$ or $i' = r'$. Then we obtain $(m - k') \leqq r'$ (or $(m - r') \leqq k'$) which is true due the fact $a' \geqq d'$. For $m - r' - k' < 0$ we must test the inequality for $i' = a'$ and use the inequality (1).

2. THE UNBIASED APPROACH

2.1. We can extend an quantifier q into a new quantifier q^* by the following way: $\mathrm{Asf}_{q^*}(\boldsymbol{M}) = \mathrm{Asf}_q(\boldsymbol{M}^e)$, where $\boldsymbol{M}^e$ is obtained from $\boldsymbol{M}$ by errasing all rows containing at least one value $\times$.

2.2. Particularly, in GUHA procedures with quantifiers of the type 2, we evaluate a sentence $q(\varphi, \psi)$ on $\boldsymbol{M}^e_{\varphi,\psi}$obtained form $\boldsymbol{M}_{\varphi.\psi}$ by erasing all cards $\langle \times, \times \rangle$, $\langle 1, \times \rangle$, $\langle 0, \times \rangle$, $\langle \times, 1 \rangle$, $\langle \times, 0 \rangle$. It means that objects with missing information are not omitted from $\boldsymbol{M}$ before computing, but only in evaluating particular sentence $q^*(\varphi, \psi)$.

2.3. Theorem. Let $\to$ be an implicational quantifier. Then the inference rule SpRd is sound for its unbiased extension $\to^*$.

Proof. Consider the same steps as in 1.6. First: Clearly, by Table 1, $a' \geqq a$. Similarly $b' \leqq b$:

$$\begin{array}{cc} & \begin{array}{ccc} \varphi & \delta & \chi \end{array} \\ b & \left\{\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & \times \\ 1 & 0 & 1 \end{array}\right\} \quad b' \end{array}$$

Second: by Table 2, we have $a' \geqq a$ again. Similarly $b' = b$:

$$\begin{array}{cc} & \begin{array}{cccc} \varphi & \psi & \delta & \neg\psi \end{array} \\ b & \{1 \quad 1 \quad 0 \quad 0\} \quad b' \end{array}$$

2.4. Note that this way of extension of implicational and associational quantifiers is enabled by the new definition of associativity and implicativity used in [3] in the opposition to the old one used in [1], [2], that demanded for $\boldsymbol{M} \leqq^{a} \boldsymbol{N}$ to have card $(\boldsymbol{M})$ = card $(\boldsymbol{N})$.

3. DISCUSSION

3.1. We shall summarize here some (more or less evident) properties of the three mentioned ways of extending quantifiers for the three valued $\{0, \times, 1\}$ – calculus.

Consider quantifiers defined on $\{0, 1\}$ structures of the type 2. Denote, for such an quantifier q, by q_s, q_e and q_o its secured, unbiased and optimistic extension.

3.2. Fact. Let q be an associational quantifier. Then $\mathrm{Asf}_{q_s} \leqq \mathrm{Asf}_{q_e} \leqq \mathrm{Asf}_{q_o}$.

Proof. Let $\boldsymbol{M} \in \mathscr{M}^{\times}$ and $a, b, \ldots$ be corresponding frequencies. Denote by $a_c, b_c, \ldots$ the frequencies obtained by completion. For the secured case consider frequencies of completions with $a_c = a$, $d_c = d$, $b_c \geqq b$, $c_c \geqq c$. If q_s is secured and $\mathrm{Asf}_{q_s}(\boldsymbol{M}) = 1$ then for all the above completions $\mathrm{Asf}_{q}(\boldsymbol{M}^c) = 1$. By associationality $\mathrm{Asf}_{q_c}(\boldsymbol{M}) = 1$. The case of q_o is analogous.

3.3. Note. If q is associational and saturable (or executive), then for each $\boldsymbol{M} \in' \mathscr{M}^{\times} - \mathscr{M}$, at least one of the inequalities in 3.2 is strict (for models with domains greater then a treshold value m_0).

3.4. Suppose now that data (models, structures) are obtained by some random experiments, i.e. we have a probability space $\langle \Sigma, \mathscr{R}, P \rangle$ and $\boldsymbol{Q} : \mathscr{P}_{\mathrm{fin}}(\mathbb{N}) \times \Sigma \to \mathscr{M}$ ($\boldsymbol{Q}$ parcialized on M, $M \in \mathscr{P}_{\mathrm{fin}}(\mathbb{N})$, is measurable). Moreover, assume that for each $M \in \mathscr{P}_{\mathrm{fin}}(\mathbb{N})$ and each $\boldsymbol{M} \in \mathscr{M}_M$ we have $P_M(\{\omega \in \Sigma;\ \boldsymbol{Q}(\omega, M) = \boldsymbol{M}\}) > 0$ (e.g. data are obtained by independent multinomial sampling with no zero probabilities).

3.5. Denote, for each M, by $\mathscr{C} \subseteq \mathscr{M}_M \times \mathscr{M}_M^{\times}$ the completion relation, i.e. $\langle \mathscr{M}, \boldsymbol{N} \rangle \in \mathscr{C}$ iff $\boldsymbol{M}$ is a completion of $\boldsymbol{N}$.

Consider now a (deterministic) pollution mechanism **pol** as a mapping from $\mathscr{M}_M$ into $\mathscr{M}_M^\times$, **pol** $\subseteq \mathscr{C}$.

For each such a mapping **pol**, we can define for any $\mathscr{N} \subseteq \mathscr{M}_M^\times$, $P(\mathscr{N})$ as $P(\mathscr{N}) = P_M(\mathbf{pol}^{-1}(\mathscr{N}))$. Write now $q(M)$ instead of $\mathrm{Asf}_q(M)$. Consider now sets

$$\mathscr{N}_s = \{N \in \mathscr{M}_M^\times;\ q_s(N) = 1\} \quad \text{and} \quad \mathscr{N}_o = \{N \in \mathscr{M}_M^\times;\ q_o(N) = 1\}$$

and denote $p_s = P(\mathscr{N}_s)$ and $p_o = P(\mathscr{N}_o)$. Pedantically, we should write $p_s(\mathbf{pol}, M)$ etc. Put $p = P_M\{M \in \mathscr{M}_M;\ q(M) = 1\}$.

3.6. Theorem. Under assumptions of 3.3–3.5, we have for $\mathrm{card}(M) \leqq m_0$, $p_s = p = p_o = 0$ and, for $\mathrm{card}(M) > m_0$, $p_s \leqq p \leqq p_o$.

Proof. The first part is clear; for $\mathrm{card}(M) \leqq m_0$ there is no completion with $q(M) = 1$. For the second part, we have to prove that for each **pol** the following is true:

$$\mathbf{pol}^{-1}(\mathscr{N}_s) \subseteq \{M \in \mathscr{M}_M;\ q(M) = 1\} \subseteq \mathbf{pol}^{-1}(\mathscr{N}_o)\,.$$

Clearly: if $N \in \mathscr{N}_s$ then for each M such that $\langle M, N\rangle \in \mathscr{C}$, $q(M) = 1$, hence for each $M \in \mathbf{pol}^{-1}(\mathscr{N}_s)$, $q(M) = 1$. The second inclusion can be proved similarly.

3.7. There are pollutions mappings, for which $p_s < p < p_o$; moreover, there is an pollution mapping giving $p_o = 1$ and $p_s = 0$. For example, a necessary condition for a mapping to satisfy $p_s < p < p_o$ is: for each M, $\mathrm{card}(M) > m_0$, there are $M_1, M_2 \in \mathscr{M}_M$ such that $\mathbf{pol}(M_1)$ has a completion M_1' with $q(M_1') = 0$ and $q(M_1) = 1$ and $\mathbf{pol}(M_2)$ has a completion M_2' such that $q(M_2') = 1$ and $q(M_2) = 0$.

For the constant mapping $\mathbf{pol}_\times : \mathscr{M}_M \to X$, where X is a model containing only $\times$-values, clearly $p_s = 0$ and $p_o = 1$.

3.8. We can consider a probabilistic model of the pollution process: Consider for each M a pair of (generally dependent) random variables $\langle Q, \mathbf{pol}^*\rangle : \Sigma \to \mathscr{C}$. Then for each M we obtain a conditional distribution on

$$\mathbf{pol}^*(M) = \{N;\ \exists \omega \in \Sigma,\ \langle Q, \mathbf{pol}^*\rangle(\omega) = \langle M, N\rangle\}\,.$$

It can be extended onto $\mathscr{M}_M^\times$ putting $P(N/M) = 0$ for each $\langle M, N\rangle \notin \mathscr{C}$. Assume that for each $\langle M, N\rangle \in \mathscr{C}$, $P(\langle M, N\rangle) > 0$.

Now we can clearly speak about (for a given domain M) $p_0 = P(\mathscr{N}_0)$ and $p_s = P(\mathscr{N}_s)$.

3.9. Theorem. Under assumptions of 3.3, 3.4 and 3.8, we have $p_o < p < p_s$ (for $\mathrm{card}(M) > m_0$, else $p_o = p = p_s = 0$).

Proof. Consider the case of $\mathrm{card}(M) > m_0$. Then $p > 0$. For each $\mathscr{N} \subseteq \mathscr{M}_M^\times$ denote $\mathscr{N}^M = \{N \in \mathscr{N};\ \langle M, N\rangle \in \mathscr{C}\}$. Consider $P(\mathscr{N}_o) = \sum_{M \in \mathscr{M}_M} P(\mathscr{N}_o/M)\ P(M) =$

$= \sum_{M \in \mathcal{M}_M} P(\mathcal{N}_o^M/M)\, P(M)$. Note that $p = P_M(\{M \in \mathcal{M}_M;\ q(M) = 1\}) = \sum_{M \in \mathcal{M}_M} q(M)\, P(M)$ and that if $q(M) = 1$ then $P(\mathcal{N}_o^M/M) = 1$; clearly $p < p_o$. Similarly for $p_s < p$ (if $q(M) = 1$ then $P(\mathcal{N}_s^M/M) = 0$). The strict inequality follows from the fact that for each M, $P(\langle M, X\rangle) > 0$.

3.10. In the present general framework the properties of q_e can be hardly discussed. But under some particular conditions, q_e has a great advantage, namely $p_e = p$.

Consider the special case in which for each M, $\langle \boldsymbol{Q}, \mathbf{pol}^*\rangle$ is a sequence of independent and identically and independently of M distributed random variables (d-homogeneity and d-independence condition of [3]). Moreover, let $\boldsymbol{Q}$ and $\mathbf{pol}^*$ be mutually independent.

3.11. Under assumptions of 3.3, 3.4 and 3.10 we have the following: if p is independent of M, then $p_e = P(\mathcal{N}_e) = p$ for each M.

Proof. Note that in the present context, all probabilities depend only on cardinality m of M, not on M. We can write $P_m(\mathcal{N}_e) = \sum_{k=0}^{m} P_k(\{M \in \mathcal{M}_m;\ q(M) = 1)\, P_m(\{N;\ N$ has exactly k rows without $\times\}) = p \sum_{k=0}^{m} P_m(\{N;\ N$ has exactly k rows without $\times) = p$.

3.12. For the usual quantifiers, the condition that p is independent of m is satisfied not quite exactly (usually $p_m \to p$). This fact is due mainly to the non continuity of underlying distributions. The condition could be "easily" satisfied by randomization (under for example independency hypothesis and considering exact distributions).

3.13. The considerations of the present chapter give clearly some quidelines showing to us when use different ways of extension of quantifiers in the GUHA procedures.

(Received August 14, 1979.)

REFERENCES

[1] P. Hájek: Automatic listing of important observational statements I—III, Kybernetika *9* (1973), 187—205, 251—271 and *10* (1974), 95—124.

[2] P. Hájek, K. Bendová, Z. Renc: The GUHA method and the three valued logic. Kybernetika *7* (1971), 13—21.

[3] P. Hájek, T. Havránek: Mechanizing hypothesis formation — mathematical foundations for a general theory. Springer-Verlag, Berlin—Heidelberg—New York, 1978.

[4] P. Hájek, T. Havránek: The GUHA method — its aims and techniques. Int. J. Man-Machine Studies *10* (1978), 3—22.

[5] P. Hájek, T. Havránek: A theory of helpful quantifiers and its application to approaches to missing information in the GUHA method. (In preparation for Kybernetika.)

[6] P. Hájek, I. Havel, T. Havránek, M. Chytil, Z. Renc, J. Rauch: Metoda GUHA. Dům techniky, České Budějovice, sec. ed. 1977.

[7] T. Havránek: Some alternative ways of treatment of missing information in GUHA pro-

cedures I—III, GUHA res. rep. 124, 125, 126, Czech. Acad. Sci. Cent. of Biomathematics, Prague 1977.
[8] W. Lipski: Information systems with incomplete information. Proceedings of the third international symposion on automata theory, languages and programs, S. Michaelson, R. Milner (eds.). University Press, Edinburgh 1976, 120—130.

RNDr. Tomáš Havránek, CSc., Matematické středisko biologických ústavů ČSAV (Center of Biomathematics — Czechoslavak Academy of Sciences), Vídeňská 1083, 142 20 Prague 4. Czechoslovakia.