Automatic Generation of Hypotheses Using the GUHA Method

K. Ramírez-Amaro, V. Ortega-González and J. Figueroa-Nazuno

Centro de Investigación en Computación
Instituto Politécnico Nacional
Unidad Profesional "Adolfo López Mateos"
Colonia Lindavista, C. P. 07738
México D. F.
kramirezb05@sagitario.cic.ipn.mx, eortegag631@ipn.mx, jfn@cic.ipn.mx

Abstract. The GUHA method for automatic generation of hypotheses is presented in this paper. This technique is based on mathematical logics and one of its advantages is not to assume any statistical distribution between the data. With the rules generated from GUHA it is possible to automatically extract models from data. This article is focused on the study of the interacting variables which are considered as indicators of social inequality, such as potable water and electricity availability among others. Using this data the results indicate us that some variables have a higher incidence than others. Furthermore, we demonstrate that GUHA is an interesting approach for obtaining automatic models from variables. This is experimentally evaluated on variables related to social inequality.

1. Introduction

The General Unary Hypotheses Automaton (GUHA) was first introduced by P. Háyek, I. Havel and M. Chytil [1]. GUHA is a method for automatic generation of hypotheses based on empirical data. GUHA is one of the oldest methods of data mining [1]. The principle of this method is to let the computer generate and evaluate all hypotheses and select those that are interesting from the point of view of the given data and the studied problem. It is important to mention that GUHA is not a method to verify previously formulated hypotheses.

GUHA systematically finds "all interesting hypotheses" from the point of view of a specific problem based on given data. This contains a dilemma: "all" means "as many as possible", and "interesting" implies to create "not too many" rules. To cope with this dilemma, one may try systematically different GUHA procedures. Once a specific procedure has been selected, it is necessary to adjust the values of its different parameters. All the specifications and results referred in this paper were obtained using the specific procedure known as GUHA-ASSOC.

This article is focused on the study of the interacting variables which are considered as indicators of social inequality. In order to evaluate experimentally the GUHA method, we analyzed 27 different social variables measured for each of the 32

Data Mining and Information Systems.
Research in Computer Science 22, 2006, pp. 31-41

states of Mexican Republic. These variables are considered as indicators of social inequality. The data base used in this analysis was taken from the INEGI (Instituto Nacional de Estadística, Geografía e Informática) [10, 11 and 12].

2. GUHA Method

GUHA is a method for automatic formulation of interesting hypotheses supported by given data, and this is done by means of computer procedures. These hypotheses express statements concerning all variables from our sample. In general, the data can not guarantee the truthfulness of such hypotheses, but offer support and make them plausible.

The data to be processed can be represented as a rectangular matrix:

$$D := (d_{i,j})_{m \times n} \tag{1}$$

where $d_{i,j}$ is the value of the j-th attribute for the i-th object. Thus, the rows of matrix D correspond to the objects belonging to our sample and each column stand as a variable of interest, e.g., objects may be the states and attributes may be social variables.

It is important to keep in mind that GUHA produces multifactorial hypotheses. Therefore, these hypotheses express relations among single variables, pairs, triples, quadruples and further; and not only one-on-one relations.

3. General Procedure of GUHA

The hypotheses in GUHA-ASSOC exhibit the following structure "$A \sim S$" (properties of A are associated with S), e.g., smoking and cancer; where "~" stands as some notation of association for generalized quantifiers. A is called *antecedent* and S the *succedent* of the statement "$A \sim S$". A special case of association $A \sim S$ is the implication of the form $A \rightarrow S$ ("A makes S likely"). Therefore, implicational quantifiers in some sense estimate the conditional probability $P(S \mid A)$ [4].

Each generated hypothesis is evaluated as a statement on the data matrix. If the processed data matrix has no missing data items[1], each pair A and S produces the corresponding four-fold table:

Table 1. Four-fold table

Variable	S	¬S	Total
A	a	b	r:= a+b
¬A	c	d	s:= c+d
Total	k:= a+c	l:= b+d	n

[1] There exist some special considerations for handling missing values in GUHA; for more information see [5].

where a, b, c and d are the observed frequencies, defined as follows:

a:= *Freq(A & S)*; the number of objects in the data satisfying both A and S;

b:= *Freq(A & ¬S)*; satisfying A but not satisfying S;

c:= *Freq(¬A & S)*; not satisfying A but satisfying S;

d:= *Freq(¬A & ¬S)*; not satisfying both A and S;

and n is the number of objects in our data base, such that $n = a+b+c+d = k+l = r+s$.

On given data, each pair of boolean attributes (A, S) determines its own four-fold frequency table; the association of A with S is defined by choosing an associational quantifier "~".

GUHA uses generalized binary quantifiers which are sometimes referred to as "operators". The semantics of quantifiers are given by their *associated functions*: for each quantifier "~", there is an associated function $Tr_{\sim}$ with the values 0 or 1. The associated functions operate on the frequencies of the different objects satisfying or not the given statement. For example, the associated function of the quantifier $\forall$ yields *1* if all the objects satisfy the statement; otherwise its value is 0.

There are several types of associational quantifiers, among them: *implicational quantifiers* (e.g., FIMPL) which formalize the association "many A are S"; *comparative quantifiers* (e.g., SIMPLE) which express the association "A makes S more likely, than ¬S does"; *symmetric associational quantifier* such as CHI-SQUARE $\sim_{\alpha}^{\chi^2}$, corresponding to the χ^2 asymptotic test of independence in four-fold tables with the significance level α [3]. Some quantifiers just express observations on the data; and some others serve as tests of statistical hypotheses with unknown probabilities.

Since we are interested in symmetric associations, the quantifier used in this article is CHI-SQUARE test (χ^2). The associational quantifiers are symmetric if satisfy the following:

If *(a,b,c,d)* is a four-fold table, $Tr_{\sim}(a,b,c,d) = 1$, then $Tr_{\sim}(a,c,b,d) = 1$

where "~" is the quantifier with the associated function $Tr_{\sim}$[4].

This quantifier has two input parameters: s and α, where s is the number of valid hypotheses and α is the level of significance.

Considering the input values of the associational quantifier CHI-SQUARE: $s \geq 2$, $\alpha \in (0, 0.5]$, an hypothesis is valid iff satisfies the conditions (2), (3) and (4) consecutively:

$$a \geq s \tag{2}$$

$$ad > bc \tag{3}$$

$$\chi^2 = \frac{n(ad-bc)^2}{k \cdot l \cdot r \cdot s} \geq \chi_1^2(1-2\alpha) \tag{4}$$

where $\chi_1^2(1-2\alpha)$ is the $1-2\alpha$ quantile of the χ^2 distribution with one degree of freedom [6]. Otherwise, the hypothesis is not considered valid.

Matrices with some missing data items can be processed. The user can choose one of three possible techniques for treatment of missing information: secured (the default choice), deleting or optimistic. These three possibilities in fact give triple meaning to

a sentence $A \sim S$ in a data matrix D with incomplete information. For more details see [7]. In this article data with missing items are not considered.

Now we briefly describe the GUHA-ASSOC procedure working with associational quantifiers such as CHI-SQUARE. The application of the procedure takes place in three main steps:

- *Preprocessing* – In this first step there is needed to define the following:
  - the data matrix,
  - it is necessary to establish whether a variable is considered as a antecedent or as a succedent and to define which rules of inference are to be selected
  - parameters determining syntactic form of antecedents and succedents to be generated,
  - minimal and maximal length of antecedents/succedents (number of literals),
  - the quantifier and its parameters,
  - preparing the internal representation of the data matrix in a suitable form for a quick generation and evaluation of hypotheses.

  For a whole comprehension on this step see Fig 1a and 1b.
- *Processing* – The main program produces all associations $A \sim S$ satisfying the syntactic restrictions. The evaluation of these associations is supervised avoiding exhaustive search; hence, a group of "interesting" rules is produced. Semantics of hypotheses is determined by the selection of a quantifier and its parameters see Fig. 1c and 1d.
- *Postprocessing* – The output is formed by all generated hypotheses that have been found true and are not immediate consequences of previously found hypotheses. See Fig. 1e.

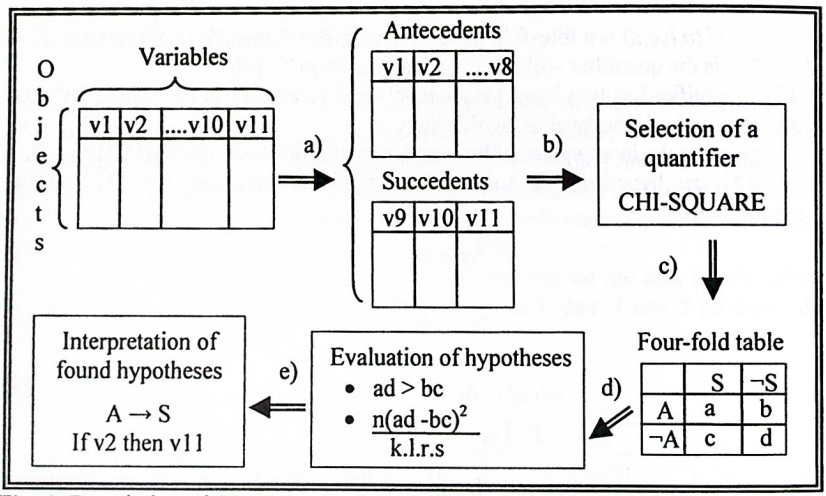


Fig. 1. Description of GUHA method. In the figure are shown the main stages of the GUHA method where the input is the matrix D and the output are the found hypotheses.

4. Experimental Analysis with Social Variables

There are many underlying factors for social inequality. Among them are: labour market, mortality rate, education, race, gender, culture, wealth accumulation, and development patterns. Social inequality refers to disparities in the distribution of economic assets and incomes.

As we saw in section 2, the data to be processed is represented as a rectangular matrix D, where the objects (rows) correspond to the 32 states of the Mexican Republic and the attributes (columns) stand as the social variables which are 27 [10, 11 and 12]. From these 27 variables (see Table 2), we define the first 24 as the antecedents and the last three as succedents. This division between antecedents and succedents is based on the literature [9]. After that, for each variable the objects are grouped into ten intervals according to their values. Subsequently, we select the quantifier CHI-SQUARE and its input parameters: $s = 2$ and $\alpha = 0.05$.

Table 2. Social Variables. The white boxes refer to antecedents and the gray ones to succedents.

% Urban Population	Number of libraries	Total population (in thousands)	Annual rate of growth	% Illiteracy population
% Population without drainage	% Population without electricity	% Population with ground floors	% Population that speaks Indigenous languages	% Population without potable water
% Internal product	% Immigration rate	% Woman population earning up to 2 minimal wages	% Man illiteracy population	% Man population earning up to 2 minimal wages
% Emigration Rate	Index of corruption	Density of Population	% Woman illiteracy population	% Population up to 3rd grade of primary education
% Population up to 6th grade of primary education	% Population up to 1st grade of secondary education	% Population up to 3rd grade of secondary education	% Population beyond high school education	
Infant mortality rate	Degree of marginalization	Extreme poverty		

The next step is to generate the hypotheses and their corresponding four-fold table, for example:

Antecedent (*A*): *%Urban_popul* (53.9, 65.4]
Succedent (*S*): *Inf_mortality_rate* [5.9, 10]

Where () – defines an open interval and [] - defines a closed interval.

The following four-fold table is obtained from the data matrix and the antecedent and succedent previously defined:

	S	¬S	
A	3	3	6
¬A	2	24	26
	5	27	32

Antecedent frequency

Seceedent frequency

As we can see there are three hypotheses that satisfying both *A* and *S*, i.e. three states of the Mexican republic satisfy if *%Urban_popul* (53.9, 65.4] then *Inf_mortality_rate* [5.9, 10].

The whole table indicate us that the hypotheses generated from the previous antecedent and succedent has a confidence of the 50% due to the fact that there are three objects that satisfy both *A* and *S*, but on the other hand there are three other objects that satisfy *A* but not *S* from the total of antecedents. The hypothesis that give us a remarkable information is the first one (*A*~*S*) with 3/6 of meaning, which is equivalent to 50% of confidence.

5. Experimental Results

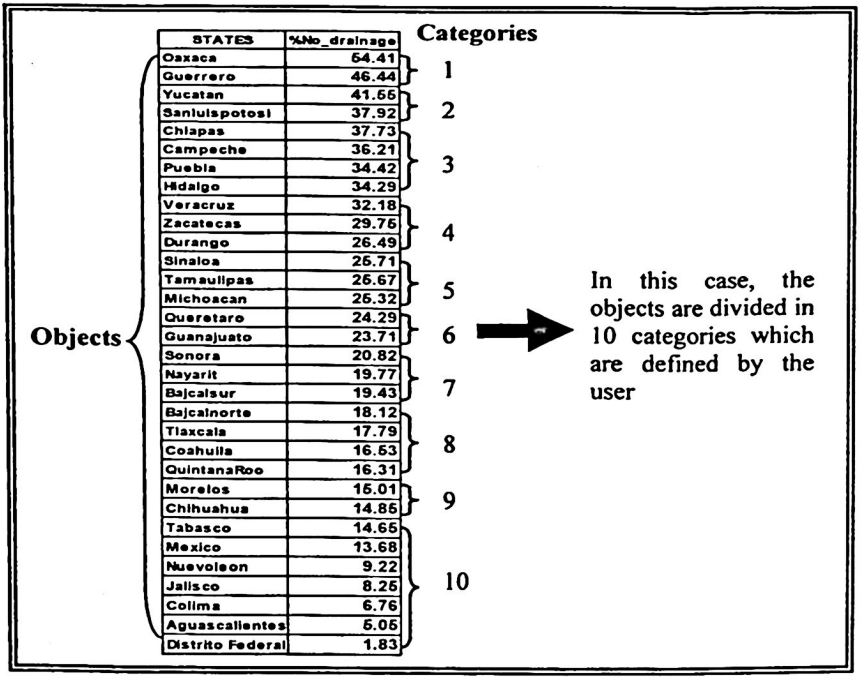
The objects re divided in ten intervals for each antecedent and succedent variables; each group was labeled for a better interpretation of the results. The categorization could be defined by the user or the program can do it by itself in two different ways: equidistant or equiprobable. In this analysis we decided to manually categorize the objects trying to conserve together those with similar properties. Fig. 2 can be seen as an example of the categorization for the %No_drainage variable.

After that, the ten categories are labeled as shows the Table 3.

Table 3. Labels for the Categories

1. Extremely low	2. Very low	3. Low	4.Middle low	5. Medium1
6. Medium2	7. Middle high	8. High	9. Very high	10. Extremely high

The results will be explained in the following subsections, they are divided in three parts due to the fact that we have previously defined three succedents. These hypotheses were selected arbitrarily among those labeled with the highest and lowest levels (corresponding to categories 1, 2 or 3 and 8, 9 or 10).



STATES	%No_drainage	Categories
Oaxaca	54.41	1
Guerrero	46.44	
Yucatan	41.66	2
Sanluispotosi	37.92	
Chiapas	37.73	3
Campeche	36.21	
Puebla	34.42	
Hidalgo	34.29	
Veracruz	32.18	4
Zacatecas	29.76	
Durango	26.49	
Sinaloa	25.71	5
Tamaulipas	25.67	
Michoacan	25.32	
Queretaro	24.29	6
Guanajuato	23.71	
Sonora	20.82	7
Nayarit	19.77	
Bajcalsur	19.43	
Bajcalnorte	18.12	8
Tlaxcala	17.79	
Coahuila	16.53	
QuintanaRoo	16.31	9
Morelos	15.01	
Chihuahua	14.85	
Tabasco	14.66	10
Mexico	13.68	
Nuevoleon	9.22	
Jalisco	8.25	
Colima	6.76	
Aguascalientes	5.05	
Distrito Federal	1.83	

Fig. 2. Example of the categorization of the variable %No_drainage.

As we mentioned in section 3, although an hypothesis satisfies the conditions (2), (3) and (4), GUHA does not guarantee its truthfulness. Some causes for this sort of situations could be:

- the quantity of objects from the matrix is insufficient,
- the categorizations are not appropriate,
- the confidence rate is less than 50% , etc.

In the following results it is possible to observe some hypotheses affected by one or a combination of the former causes.

5.1 Results of the succedent Infant Mortality Rate

The GUHA method generates 317 valid hypotheses from the succedent infant mortality rate. The most significant hypotheses are the following:

A) A = *Men_2wage* (low) $\rightarrow$ S = *Inf_death* (very low)

	S	¬S	
A	2	1	3
¬A	3	26	29
	5	27	32

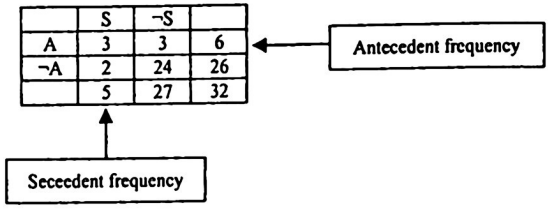
Confidence = 2/3= 66.6%

The next step is to generate the hypotheses and their corresponding four-fold table, for example:

Antecedent (*A*): *%Urban_popul* (53.9, 65.4]
Succedent (*S*): *Inf_mortality_rate* [5.9, 10]

Where () – defines an open interval and [] - defines a closed interval.

The following four-fold table is obtained from the data matrix and the antecedent and succedent previously defined:



As we can see there are three hypotheses that satisfying both *A* and *S*, i.e. three states of the Mexican republic satisfy if *%Urban_popul* (53.9, 65.4] then *Inf_mortality_rate* [5.9, 10].

The whole table indicate us that the hypotheses generated from the previous antecedent and succedent has a confidence of the 50% due to the fact that there are three objects that satisfy both *A* and *S*, but on the other hand there are three other objects that satisfy *A* but not *S* from the total of antecedents. The hypothesis that give us a remarkable information is the first one (*A*¬*S*) with 3/6 of meaning, which is equivalent to 50% of confidence.

5. Experimental Results

The objects re divided in ten intervals for each antecedent and succedent variables; each group was labeled for a better interpretation of the results. The categorization could be defined by the user or the program can do it by itself in two different ways: equidistant or equiprobable. In this analysis we decided to manually categorize the objects trying to conserve together those with similar properties. Fig. 2 can be seen as an example of the categorization for the %No_drainage variable.

After that, the ten categories are labeled as shows the Table 3.

Table 3. Labels for the Categories

1. Extremely low	2. Very low	3. Low	4.Middle low	5. Medium1
6. Medium2	7. Middle high	8. High	9. Very high	10. Extremely high

The results will be explained in the following subsections, they are divided in three parts due to the fact that we have previously defined three succedents. These hypotheses were selected arbitrarily among those labeled with the highest and lowest levels (corresponding to categories 1, 2 or 3 and 8, 9 or 10).

If the percentage of urban population and the immigration rate are extremely low *then* the degree of marginalization is very high.

F) *A* = *%No_potable* & *%No_drainage* & *%No_electricity* & *%Ground_floors* & *%Illiteracy_P* (Extremely high) → *S* = *Marginal* (very high)

	S	¬S	
A	3	1	4
¬A	3	25	28
	6	26	32

Confidence = 3/4= 75%

	S	¬S	
A	2	0	2
¬A	4	26	30
	6	26	32

Confidence = 2/2= 100%

	S	¬S	
A	3	1	4
¬A	3	25	28
	6	26	32

Confidence = 3/4= 75%

	S	¬S	
A	3	0	3
¬A	3	26	29
	6	26	32

Confidence = 3/3= 100%

	S	¬S	
A	3	0	3
¬A	3	26	29
	6	26	32

Confidence = 3/3= 100%

If the percentage of population without potable water and the percentage of population without drainage and the percentage of population without electricity and the percentage of population with ground floors and the percentage of illiteracy population are extremely high *then* the marginalization grade is very high.

G) *A* = *%No_drainage* & *%No_electricity* & *%Ground_floors* & %IlliteracyP (extremely low) → *S* = *Marginal* (very low)

	S	¬S	
A	2	5	7
¬A	1	24	25
	3	29	32

Confidence= 2/7= 28.5%

	S	¬S	
A	2	0	2
¬A	1	29	30
	3	29	32

Confidence= 2/2= 100%

	S	¬S	
A	3	1	4
¬A	0	28	28
	3	29	32

Confidence= 3/4= 75%

	S	¬S	
A	3	0	3
¬A	0	29	29
	3	29	32

Confidence= 3/3= 100%

If the percentage of population without drainage and the percentage of population without electricity and the percentage of population with ground floors and the percentage of illiteracy population are extremely low *then* the degree of marginalization is very low.

H) *A* = *%Urban_p* & *%GIB* (extremely high) → *Marginal* (very low)

	S	¬S	
A	3	0	3
¬A	0	29	29
	3	29	32

Confidence= 3/3= 100%

	S	¬S	
A	2	1	3
¬A	1	28	29
	3	29	32

Confidence= 2/3= 66.6%

If the percentage of urban population and the gross internal product are extremely high *then* the degree of marginalization is very low.

5.3 Results of the succedent Extreme Poverty

The GUHA method generates 317 valid hypotheses from the succedent Extreme Poverty. The most significant hypotheses of this succedent are the following:

I) *A = %No_potable & %Ground_floors & %Men_2wage & illiteracy_M & illiteracy_W & P3_Sec & %No_drainage* (extremely low) → *S = E_poverty* (extremely low)

	S	¬S	
A	3	1	4
¬A	1	27	28
	4	28	32

Confidence = 3/4= 75%

	S	¬S	
A	2	2	4
¬A	2	26	28
	4	28	32

Confidence = 2/4= 50%

	S	¬S	
A	2	1	3
¬A	2	27	29
	4	28	32

Confidence = 2/3= 66.6%

	S	¬S	
A	2	1	3
¬A	2	27	29
	4	28	32

Confidence = 2/3= 66.6%

	S	¬S	
A	2	1	3
¬A	2	27	29
	4	28	32

Confidence = 2/3= 66.6%

	S	¬S	
A	2	2	4
¬A	2	26	28
	4	28	32

Confidence = 2/4= 50%

	S	¬S	
A	3	4	7
¬A	1	24	25
	4	28	32

Confidence = 3/7= 42.8%

If the percentage of population without potable water and the percentage of population with ground floors and the man population earning up to two minimal wages and the percentage of man illiteracy population and the percentage of woman illiteracy population and the population up to 3rd grade of secondary education and the percentage of population without drainage are extremely low *then* the extreme poverty is extremely low.

J) *A = %No_potable & %No_electricity & Illiteracy_M* (extremely low) → S = *E_poverty* (extremely high)

	S	¬S	
A	2	2	4
¬A	1	27	28
	2	29	32

Confidence = 2/2= 50%

	S	¬S	
A	2	2	4
¬A	1	27	28
	3	29	32

Confidence = 2/4= 50%

	S	¬S	
A	2	1	3
¬A	1	28	29
	3	29	32

Confidence = 2/3= 66.6%

If the percentage of population without potable water and the percentage of population without electricity and the percentage of man illiteracy population are extremely low *then* the extreme poverty is extremely high.

6. Conclusions

By means of GUHA method we are able to obtain several hypotheses which relate different variables (in this specific case social inequality variables).

For each hypothesis we compute a percentage of confidence. Using this measure is possible to define if the hypotheses can be consider as truth or could be rejected, e.g., the hypotheses D states "*If* percentage of population without drainage is extremely low *then* infantile mortality rate is very high". As we can see, this particular hypothesis make no sense, and if we evaluate the value of its *% of confidence* we can see that shows a very low value compared with the other hypotheses. Thus, we define this hypothesis as non reliable.

On the other hand, the hypothesis E "*If* the percentage of urban population and immigration rate are extremely low *then* the degree of marginalization is very high" shows the highest reliability and can be considered to model some of these social variables.

Using the hypotheses with higher reliability we can design models that allow us to characterize these sorts of phenomena. Therefore, the GUHA method implicitly gives rise to these important models through the determination of specific rules with any assumption of statistical distribution between the data.

References

1. P. Hájek, I. Havel, and M. Chytil, *"The GUHA method of automatic hypotheses determination"*, Computing, no. 1, pp. 293–308, 1966.
2. R. Agrawal, H. Manilla, R. Sukent, A. Toivonen, and A. Verkamo, *"Fast Discovery of Association Rules"* in Advance in Knowledge Discovery and Data Mining, AAA Press, 1996, pp.307-328.
3. P. Hájek, M. Holeña, J. Rauch, *"The GUHA method and foundations of (relational) data mining"*, Springer 2003, pp. 17-37.
4. P. Hájek, A. Sochorová, and J. Zvárová, "GUHA for personal computers", *Comp. Stat. Data Anal.*, no. 19, pp. 149–153, 1995.
5. P. Hájek, "Briefly on the GUHA method of data mining", Journal of Telecommunications and Information Technology, no. 3, pp. 112-114, 2003.
6. P. Hájek, T. Havránek, *"Mechanizing hypothesis formation- Mathematical foundations for a general theory"*, Springer Verlag, Heidelberg, 1978.
7. P. Hájek, *"The new version of the GUHA procedure ASSOC (generating hypotheses on associations)"*, Mathematical Foundations, COMPSTAT 1984, Proceedings in Computational statistics, Physica-Verlag, Wien, pp. 360-365.
8. T. Havránek, *"The statistical modification and interpretation of the GUHA method"*, Kybernetika 7, 1971, pp.13-21.
9. Documentación técnica de los indicadores Sociodemográficos, Archivo de metadatos, Consejo Nacional de Población Press., Diciembre de 2005. (www.conapo.gob.mx/)
10. INEGI. XII Censo General de Población y Vivienda 2000. Resultados preliminares. México, 2000.
11. INEGI. Estados Unidos Mexicanos. XII Censo General de Población y Vivienda, 2000. Tabulados Básicos. Tomo I. Aguascalientes, Ags., México, 2001.
12. INEGI. Encuesta Nacional de Ocupación y Empleo 2005, Indicadores estratégicos. Aguascalientes, Ags., 2005.