REDUCING IMPRECISION IN A HUMAN RESOURCE DATABASE THROUGH ROUGH SET THEORY [1]

AYRTON BENEDITO GAIA DO COUTO - LUIZ FLAVIO AUTRAN MONTEIRO GOMES
Centro Rio de Janeiro - BRASIL
acouto@bndes.gov.br - autran@ibmecrj.br

Fecha Recepción: Octubre 2010 - Fecha Aceptación: Agosto 2012

ABSTRACT

This study deals with decision-making using replicated and inconsistent data, relating to the universe of Human Resources, within a domestic/local financial institution. Replication occurs because of technical and/or economic questions, and seeks to meet the corporate and departmental requirements of such an institution. As research methodology, direct observation of such inconsistencies was used as well as a simulation based on actual data which would reflect replication with inconsistencies. Application of a multi-criteria method became necessary in view of the need to render the decision-making process rational, and was transformed into an element that stimulated this study. The method used was Rough Set Theory (RST), inasmuch as there existed no other information on the occurrence of such inconsistencies. An algorithm was developed to indicate the major data sources and was subsequently implemented into a software to facilitate research of such sources.

KEY WORDS: Muticriteria decision aiding – Decision-making – Inconsistency - Rough Set Theory

RESUMEN

Este estudio aborda la toma de decisión con datos reproducidos e inconsistentes dentro del ámbito Recursos Humanos, en una importante institución financiera y social brasileña. La reproducción proviene de cuestiones técnicas o económicas, buscando la adecuación a las exigencias corporativas y departamentales de esa institución. Como metodología, optamos por la observación directa de las inconsistencias y el simulacro, basándonos en datos reales reflejando la reproducción con inconsistencias.

[1] The first version of this article was presented in the XLI Brazilian Symposium of Operations Research held in Porto Seguro, State of Bahia, Brazil, from September 1st to September 4th, 2009

Fue necesario el uso de un método analítico de multicriterio, para convertir en realidad y hacer más racional ese proceso de toma de decisión. Se usó la Teoría de los Conjuntos Aproximativos, porque no quedaba disponible ninguna información sobre la ocurrencia de inconsistencias. Para eso, desarrollamos un algoritmo que indicase las principales fuentes de datos reproducidos e inconsistentes. Ese algoritmo fue subsecuentemente implementado con un *software* usado para facilitar la investigación sobre aquellas fuentes de datos.

PALABRAS CLAVE: Apoyo multicriterio a la decision – Toma de decisión – Inconsistencia – Teoría de los Conjuntos Aproximativos

1. INTRODUCTION

Based on the observation of consultations from replicated data from the Human Resources (HR) department database of a domestic financial institution, it was noted that there were inconsistencies in the results obtained. Consequently, the question, “How should one choose a data source in the face of replicated and inconsistent data?” came to be the motivating element of this study. As the cause of the inconsistency does not come into the scope of this study, it became necessary to seek a tool (method) which made the decision making process rational. As the inconsistencies were of a sporadic nature, they were then simulated on an electronic spreadsheet, with the intention of illustrating situations which generate replication commonly found in the work environment. The use of RST against other theories for treatment of vagueness (ex. Fuzzy Set Theory), due to the fact that there is no need for preliminary information about the data under analysis. That was the situation in this study: there was no preliminary information.

2. DEFINITION OF THE PROBLEM

When carrying out research (consultations) in databases (data storage systems in computers), a commonly found situation is the replication of data, that is, when multiple copies of the same set of data are made available for consultation, for example, with the aim of decentralizing access (Son, 1988). However, if the updating of these copies is not carried out under some form of control (controlled redundancy), there will be occasions in which the copies are not concordant, in other words, when at least one copy has not been completely updated. In this case, the database is said to be “inconsistent” (Codd, 1970; Date, 1984; Son, 1988).

In a particular domestic financial institution, specifically in relation to HR data, it was observed that inconsistencies sporadically occurred in the results obtained from the same consultation (ex. quantitative research of personnel) as a result of replicated data. This replication is due to technical issues, the main one being the need to meet corporate and departmental demands arising from different technological platforms (environments). The choice of Rough Set Theory arises from the lack of any need for any preliminary information on the data in question (ex. probability distribution). Other theories could be used ─ ex. Fuzzy Set Theory proposed by Lotfi Asker Zadeh, in 1965, as an extension of conventional Boolean logic to introduce the concept of non-absolute truth (Gomes, Gomes and Almeida, 2006). In addition, this method was implemented in a computer program, in Borland ® Delphi/Pascal language, rendering the decision making rational by the indication of the possible "reducts"of data and the principal source ("nucleus"), if it exists. It is limited only to detecting the inconsistency, if it exists, and to applying a rational method in the choice of one or more sources of data.

3. ROUGH SET THEORY

Rough Set Theory, proposed by the Polish mathematician, Zdzislaw Pawlak, in 1982, is designed to deal with imprecise data, by means of "approximations" (lower and higher) of a set of data (Pawlak, 1991). It has the relation of indiscernibility as its starting point, in other words, that which identifies objects with the same property. Objects of interest which have the same properties are "indiscernible" and, consequently, are treated as identical or similar "granules". The granularity in the representation of the information, according to Pawlak and Slowinski (1994), can be the source of inconsistency in decisions, due to the ambiguity in explaining and prescribing based on inconsistent information. In Pawlak (1991) an example is found which illustrates some concepts. Given the set U = {x1, x2, x3, x4, x5, x6, x7, x8} of toys, classified according to color (red, blue, yellow), shape (square, round, triangular) and size (small, large), we have:

	color		shape		size
x_1, x_3, x_7	red,	x_1, x_5	round,	x_2, x_7, x_8	large,
x_2, x_4	blue,	x_2, x_6	square,	x_1, x_3, x_4, x_5, x_6	small.
x_5, x_6, x_8	yellow.	x_3, x_4, x_7, x_8	triangular.		

Three equivalence relations were defined, R_1, R_2 and R_3, for color, shape and size, respectively, with the following classes of equivalence:

$$U/R_1 = \{\{x_1, x_3, x_7\},\{x_2, x_4\},\{x_5, x_6, x_8\}\} \quad (1)$$
$$U/R_2 = \{\{x_1, x_5\},\{x_2, x_6\},\{x_3, x_4, x_7, x_8\}\} \quad (2)$$
$$U/R_3 = \{\{x_2, x_7, x_8\},\{x_1, x_3, x_4, x_5, x_6\}\} \quad (3)$$

which are elementary concepts (categories) in the knowledge base K= (U, {R_1, R_2, R_3}).

Thus, according to Pawlak (1991), knowledge is supported by the ability to classify objects. In this case, an object can be something real or abstract. In addition, it is not usual to deal with a single classification but instead with a family of basic classifications (ex. color, temperature etc.) over U. At this point, "equivalence relations" and "classifications" have the same meaning, indiscriminately. According to Grzymala-Busse (1988) and Ziarko (1993a), the "equivalence relation" is also known as the "indiscernibility relation"; and "classes of equivalence" are known as "elementary sets". Thus, if R is an equivalence relation over U, then U/R signifies the family of all classes of equivalence of R (Pawlak, 1991). Also according to Pawlak (1991), if $P \subseteq R$ and $P \neq \varnothing$, then $\cap$ P (intersection of all the equivalence relations pertaining to P) is also an equivalence relation, and is indicated by IND(P), and is known as an 'indiscernibility relation' over P. Pawlak (2000) uses a disposition of data in the form of a table (database) – with lines and columns to exemplify other concepts. This example (Table 1) is composed of six shops and four attributes (quantitative and qualitative aspects):

Shop	E	Q	L	P
1	High	Good	No	Profit
2	Average	Good	No	Loss
3	Average	Good	No	Profit
4	Without	Average	No	Loss
5	Average	Average	Yes	Loss
6	High	Average	Yes	Profit

Table 1 – Table-example
Source: Adapted from Pawlak (2000)

In Table 1 we have: E – autonomy of the salespeople; Q – quality of merchandise; L – location with intense movement; P – result (profit or loss). Each shop is characterized by the attributes E, Q, L and P. In this way, all the shops are "discernible" by the use of the information made available by these attributes. However, shops 2 and 3 are "indiscernible" in terms of the attributes E, Q and L, bearing in mind that they have the same values for these attributes.

Each subset of attributes determines a "partition" ("classification") of all the objects in "classes", which have the same description in terms of those attributes.

The following problem is now considered: what are the characteristics of the shops which make a profit (or make a loss) in terms of the attributes E, Q and L? In other words, the interest is in describing the set (concept) {1, 3, 6} (or {2,4,5}).

It is easy to identify that this question cannot be answered in a single way, as shops 2 and 3 have the same characteristics as regards attributes E, Q and L, but shop 2 made a loss, while shop 3 made a profit. Based on the previous table (Table 1), it can be stated that: shops 1 and 6 made a profit, shops 4 and 5 made a loss and shops 2 and 3 cannot be classified (in terms of profit or loss). Table 2 shows the result of this analysis.

Shop	E	Q	L	P	Result
1	High	Good	No	Profit	PROFIT
2	Average	Good	No	Loss	?
3	Average	Good	No	Profit	?
4	Without	Average	No	Loss	LOSS
5	Average	Average	Yes	Loss	LOSS
6	High	Average	Yes	Profit	PROFIT

Table 2 – Table-example
Source: Adapted from Pawlak (2000)

Using the attributes E, Q and L, it is deduced that: shops 1 and 6 certainly make a profit, that is, certainly belong to the set {1,3,6}; while shops 1, 2, 3 and 6 possibly make a profit, that is, possibly belong to the set {1,3,6}. The sets {1,6} and {1,2,3,6} respectively represent the "lower" and "higher" approximations of the set {1,3,6}. The set {2,3} represents the difference between the higher and lower approximations and characterizes the "borderline region" of the set {1,3,6}. In addition, an information set or knowledge representation system or database is a finite table in which the rows are identified by the objects and the columns by the attributes. In this way, a knowledge system can be seen as a collection of objects described by the values of the attributes (Pawlak, 1991; Pawlak and Slowinski, 1994; Pawlak, 2000).

According to Pawlak and Slowinski (1994), the information system is understood to be a tupla S = (U, Q, V, f), where U is a finite set of objects, Q is a finite set of attributes, $V = \cup_{q \in Q} Vq$, where Vq is the domain of the attribute q and, f: U χ Q → V is a total function so that, $f(x, q) \in Vq$ for each $q \in Q$, $x \in U$, known as the "information function". Given an information system, S = (U, Q, V, f), and $P \subseteq Q$, e $x,y \in U$, it is said that x and y are "indiscernible" by the set of attributes P in S, if f(x,q) = f(y,q) for all $q \in P$.

Therefore, all $P \subseteq Q$ generates a binary relation in U, known as the "indiscernibility relation", denoted by IND(P). Given that, $P \subseteq Q$ and $Y \subseteq U$, the lower rough set ($\underline{P}$ Y) and the higher rough set ($\overline{P}$ Y) are defined as

$$\underline{P}\,Y = \cup \{X \in U/\,P: X \subseteq Y\} \text{ and } \overline{P}\,Y = \cup \{X \in U/\,P: X \cap Y \neq \emptyset\} \quad (4)$$

Thus, Y is a 'rough' set in relation to P, if and only if, $\underline{P}\,Y \neq \overline{P}\,Y$ (Pawlak, 1991). The borderline region of the set Y is defined as

$$Bn_P(Y) = \overline{P}\,Y - \underline{P}\,Y \quad (5)$$

4. REDUCT AND CORE OF A KNOWLEDGE SYSTEM

There are two important concepts: the reduct and the core of a knowledge system. The reduct is the essential part, that is, the set of attributes which supplies the same quality of classification as the original set of attributes (Pawlak, 1991; Pawlak and Slowinski, 1994; Pawlak, 2000). According to Ziarko (1993a), the reduct of the attributes is one of the most useful ideas in Rough Set Theory. The core can be interpreted as the most important part of this knowledge, in other words, the collection of the most important attributes of a knowledge system (Pawlak, 1991; Pawlak and Slowinski, 1994; Pawlak, 2000). Consider that R is a family of relations and $R \in R$. It is said that R is "dispensable" in R if IND(R) = IND(R – {R}); otherwise, R is "indispensable" in R. The family R is "independent" if each $R \in R$ is indispensable in R; if not, R is "dependent" (Pawlak, 1991). Pawlak (1991) defines the following propositions:

a) If **R** is independent and $\mathbf{P} \subseteq \mathbf{R}$, then **P** is also independent.

b) CORE(**P**) = $\cap$ RED(**P**), where RED(**P**) is the family of all the "reducts" of **P**.

In Pawlak (1991) an example can be found which illustrates how to obtain the reducts and the core of a knowledge system: given the family R = {P, Q, R} of three equivalence relations P, Q and R, with the following equivalence classes:

$$U/P = \{\{x_1, x_4, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_6, x_7\}\} \quad (6)$$

$$U/Q = \{\{x_1, x_3, x_5\}, \{x_6\}, \{x_2, x_4, x_7, x_8\}\} \quad (7)$$

$$U/R = \{\{x_1, x_5\}, \{x_6\}, \{x_2, x_7, x_8\}, \{x_3, x_4\}\} \quad (8)$$

Thus, the relation IND(**R**) has the following equivalence classes:

$$U/IND(\mathbf{R}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} \quad (9)$$

The relation P is indispensable in **R**, given that:

$$U/IND(\mathbf{R} - \{P\}) = \{\{x_1, x_5\}, \{x_2, x_7, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}\} \neq U/IND(\mathbf{R}) \quad (10)$$

For the relation Q, we have:

$$U/IND(\mathbf{R} - \{Q\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} = U/IND(\mathbf{R}) \quad (11)$$

Thus, the relation Q is dispensable in **R**.

Similarly, for the relation R, we obtain:

$$U/IND(\mathbf{R} - \{R\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} = U/IND(\mathbf{R}) \quad (12)$$

The relation R is also dispensable in R. This means that the classification defined by the three equivalence relations P, Q and R is the same as the classification defined by the relation P and Q or P and R. With the intention of finding the reducts of the family R = {P, Q, R}, it is checked whether each relation pair "P,Q" and "P,R" are independent or not. Given that $U/IND(\{P,Q\}) \neq U/IND(Q)$ and $U/IND(\{P,Q\}) \neq U/IND(P)$, the relations P and Q are independent and, consequently, {P,Q} is a reduct of R. A similar procedure is used to find the reduct formed by the relation {P,R}. Thus there are two reducts in the family R, {P,Q} and {P,R}, and the intersection of these reducts ($\{P,Q\} \cap \{P,R\}$) is the core {P} (Pawlak, 1991). Using the previous example as a reference ("family R"), an algorithm was developed to obtain the (in)dispensable equivalence relations, the reducts and the core:

1- For each equivalence relation, a rank is placed (starting at "1"), for each class of equivalence found. From the expressions (13), (14) and (15)

$$U/P = \{\{x_1, x_4, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_6, x_7\}\} \quad (13)$$

$$U/Q = \{\{x_1, x_3, x_5\}, \{x_6\}, \{x_2, x_4, x_7, x_8\}\} \quad (14)$$

$$U/R = \{\{x_1, x_5\}, \{x_6\}, \{x_2, x_7, x_8\}, \{x_3, x_4\}\} \quad (15)$$

Table 3 is now obtained:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
U/P	1	2	3	1	1	4	4	2
U/Q	1	3	1	3	1	2	3	3
U/R	1	3	4	4	1	2	3	3

Table 3 – Equivalence relations
Source: Adapted from Pawlak (1991)

2- The main relation **R** is obtained in the following way: based on the previous table (Table 3) and, beginning in "x_1" (class of order "1,1,1", respectively U/P, U/Q and U/R), another class is sought which possesses the same order ("x_5"). In this class, the class $\{x_1, x_5\}$ was found; this is the class of order "1,1,1".

The process is repeated for the other classes. For the relation **R**, the order shown in Table 4 was obtained:

	x1	x2	x3	x4	x5	x6	x7	x8
U/R	1	2	3	4	1	5	6	2

Table 4 – Obtaining the main relation R
Source: Adapted from Pawlak (1991)

3- The main relation is then obtained according to the ranking presented in Table 4:

$$U/IND(\mathbf{R}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} \quad (16)$$

4- The process is repeated to obtain the other relations as shown in Tables 5, 6 and 7:

a) Obtaining {**R** – P}. From:

	x1	x2	x3	x4	x5	x6	x7	x8
U/Q	1	3	1	3	1	2	3	3
U/R	1	3	4	4	1	2	3	3
U/R – P	1	2	3	4	1	5	2	2

Table 5 – Obtaining the relation {R – P}
Source: Adapted from Pawlak (1991)

$$\text{Then: } U/IND(\mathbf{R} - \{P\}) = \{\{x_1, x_5\}, \{x_2, x_7, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}\} \quad (17)$$

b) Obtaining {**R** – Q}. From:

	x1	x2	x3	x4	x5	x6	x7	x8
U/P	1	2	3	1	1	4	4	2
U/R	1	3	4	4	1	2	3	3
U/R – Q	1	2	3	4	1	5	6	2

Table 6 – Obtaining the relation {R – Q}
Source: Adapted from Pawlak (1991)

$$\text{Then: } U/IND(R - \{Q\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} \quad (18)$$

c) Obtaining {**R** – R}:

	x1	x2	x3	x4	x5	x6	x7	x8
U/P	1	2	3	1	1	4	4	2
U/Q	1	3	1	3	1	2	3	3
U/R – R	1	2	3	4	1	5	6	2

Table 7 – Obtaining the relation {R – R}
Source: Adapted from Pawlak (1991)

$$\text{Thus obtaining: } U/IND(\mathbf{R} - \{R\}) = \{\{x_1, x_5\}, \{x_2, x_8\}, \{x_3\}, \{x_4\}, \{x_6\}, \{x_7\}\} \quad (19)$$

5- The rankings obtained are compared, as shown in Table 8:

	x1	x2	x3	x4	x5	x6	x7	x8
U/R	1	2	3	4	1	5	6	2
U/R - P	1	2	3	4	1	5	2	2
U/R - Q	1	2	3	4	1	5	6	2
U/R - R	1	2	3	4	1	5	6	2

Table 8 – Comparison of the relations
Source: Adapted from Pawlak (1991)

As the ranking {**R** – P} is different to the ranking **R**, P it is "indispensable". As the rankings of {**R** – Q} and {**R** – R} are equal to the rankings of **R**, Q and R, they are "dispensable".

6- The possible reducts between "P, Q" and "P, R" are checked, according to Tables 9 and 10:

a) for {P,Q}:

	x1	x2	x3	x4	x5	x6	x7	x8
U/P	1	2	3	1	1	4	4	2
U/Q	1	3	1	3	1	2	3	3
U/P,Q	1	2	3	4	1	5	6	2

Table 9 – Checking of reduct for {P,Q}
Source: Adapted from Pawlak (1991)

As the ranking {P, Q} is different to the rankings of P and of Q, {P, Q} is a reduct.

c) for {P,R}:

	x1	x2	x3	x4	x 5	x6	x7	x8
U/P	1	2	3	1	1	4	4	2
U/R	1	3	4	4	1	2	3	3
U/P,R	1	2	3	4	1	5	6	2

Table 10 – Checking of reduct for {P,R}
Source: Adapted from Pawlak (1991)

As the ranking of {P, R} is different to the orders of P and R, {P, R} is a reduct.

7- By the intersection of the reducts, it is seen that there is a core:

$$\{P,Q\} \cap \{P,R\} = \{P\} \tag{20}$$

5. EXAMPLES OF THE APPLICATION OF ROUGH SET THEORY

Since the beginning of its creation in 1982, various applications of Rough Set theory have been found in the literature, such as: the analysis and simplification of digital circuits (Pawlak, 1991); artificial intelligence (Pawlak, 1991; Pawlak et al., 1995); knowledge discovery in clinical databases (Tsumoto, 2000); the treatment of imprecision in information systems (Gomes and Gomes, 2001); the processing of large databases (Lin, 2008); robotic systems (Bit and Beaubouef, 2008).

6. RESTRICTIONS OF ROUGH SET THEORY

Ziarko (1993a; 1993b) shows some restrictions to Rough Set Theory when applied to a set of information ("classifications"): it is susceptible to minor errors of classification, caused by problems of dependence of attributes; thus the conclusions drawn from this set are only applicable to this set, which, in practice, limits the generalization of the conclusions for a larger set of information. As an alternative to these restrictions, Ziarko (1993a; 1993b) proposes the use of a model, VP (variable precision), in order to recognize the presence of dependence of data in situations in which it would be considered independent. The VP model is also useful in dealing with problems involving data sets with large proportions of boundary cases. Nowicki (2008) proposes an alternative model which combines Neuro-Fuzzy architecture with Rough Set Theory architecture. In addition, Greco, Matarazzo and Slowinski (2005) observe that the principle of indiscernibility is not enough to cover all the semantics of a set of information. Within Multicriteria Decision Aiding, the principle of indiscernibility must be substituted by the "principle of dominance": if x dominates y, that is, if x is at least as good as y in relation to all the criteria considered, then x must belong to a class not worse than the class y; if not, there is an inconsistency between x and y (Roy and Bouyssou, 1993). Thus, in order to make it possible to deal with multicriteria decision aiding problems, Rough Set Theory must be extended with the substitution of the indiscernibility relation by a "relation of dominance" (Greco, Matarazzo and Slowinski, 2005).

7. PRACTICAL APPLICATION IN THE ORGANIZATION

It was observed that, in a domestic financial institution (BNDES, Brazilian Economic and Social Development Bank), sporadic inconsistencies occurred in the replication of HR data originating from a "main database" (in this context, stored in a central computer), to other databases in different technological platforms (electronic mail, departmental use and data aggregation) (Figure 1).

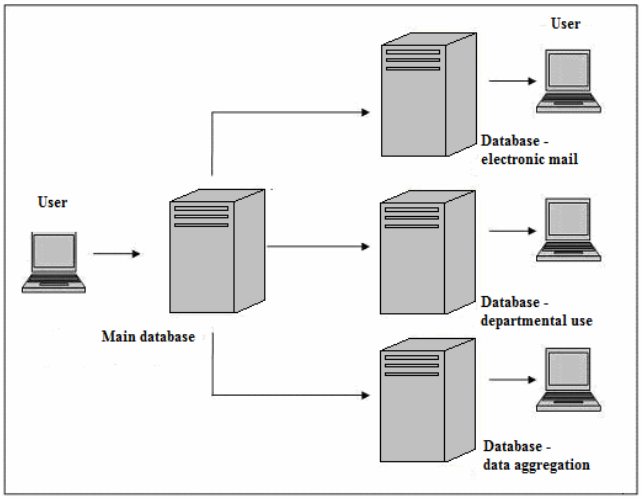

Figure 1 – Architecture of replication of data

The focal point of attention was “quantitative consultation of employees” performing the executive function “head of department”.

8. APPLICATION OF ROUGH SET THEORY

In this context, each register of a specific file of the database describes an employee (entity) and has, as well as other attributes, the registration which identifies him/her and the executive function performed (Figure 2).

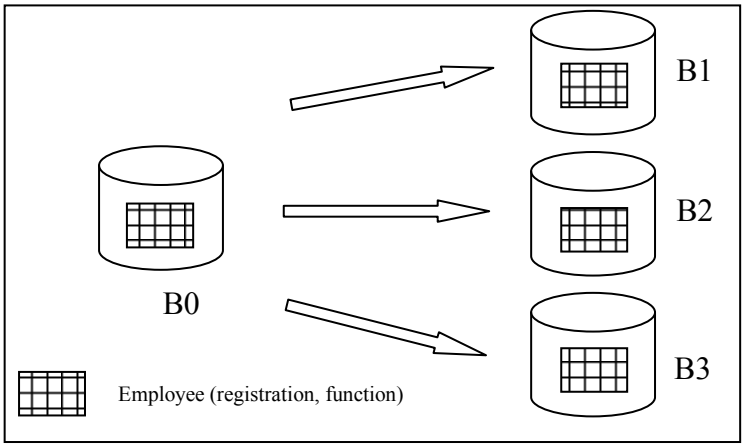

Figure 2 – Replication of the entity “Employee”

Updating is performed in the database B0 and the data is replicated in the other databases B1, B2 and B3, according to its purpose (electronic mail, departmental use and data aggregation, respectively). When the replicated data is consulted, there is a single vision of the database researched (B1, B2 or B3), which does not permit, immediately, the researcher to know if the replication occurred in a perfect manner. In order to simulate the replication with inconsistency of the attribute “executive function”, an electronic spreadsheet was used (Microsoft Excel), with the mathematical function RANDBETWEEN. Real data was considered: in a total of approximately 2,000 (two thousand) employees, 69% (or 1,381 employees) have studied in higher education, performing (or not) an executive function: services coordinator (CD), manager (GR), head of department (CH) and superintendent (SD). In the spreadsheet, “NN” indicates that an employee does not perform an executive function. It should be noted that there are 116 employees who perform the executive function of “head of department”. In the spreadsheet, each line simulates a specific employee – the original registration number has been substituted by a sequential number (column “Empl.”), with a column for each “executive function” attribute, originating from databases B0, B1, B2 and B3. The first ten registrations (or employees) were selected, shown in Table 11.

Empl.	B0	B1	B2	B3
0001	GR	GR	GR	X
0002	CH	CH	CH	CH
0003	CH	CH	X	CH
0004	CH	CH	X	CH
0005	CH	CH	CH	CH
0006	CH	CH	CH	CH
0007	CH	X	CH	CH
0008	SD	SD	SD	SD
0009	NN	NN	NN	NN
0010	GR	GR	GR	GR

Table 11 – Simulation of replication of data

This simulation was obtained by the use of the function RANDBETWEEN, at two moments: to select the registration (RANDBETWEEN (1;10)) and, to select the base (or database) B1, B2 or B3, to be replicated with inconsistency (RANDBETWEEN (1;3)), by the indication of an “X”. This procedure was carried out four times (inconsistencies in the registrations “0001”, “0003”, “0004” and “0007”) (Table 11). For the universe considered – employees who perform the executive function of “head of department” (CH), – “1” (one) was attributed to those which were replicated perfectly and “0” (zero) for those which were not or which did not belong to this universe, as shown in Table 12.

Empl.	B0	B1	B2	B3
0001	GR	0	0	0
0002	CH	1	1	1
0003	CH	1	0	1
0004	CH	1	0	1
0005	CH	1	1	1
0006	CH	1	1	1
0007	CH	0	1	1
0008	SD	0	0	0
0009	NN	0	0	0
0010	GR	0	0	0

Table 12 – Simulation of the replication of data

By Table 12, it is established that there is an inconsistency for employees "0003", "0004" and "0007", in relation to the replications in B2, B2 and B1, respectively. An analysis of "B1", for example, shows a relation of "indiscernibility" regarding employees "0002", "0003", "0004", "0005" and "0006", bearing in mind that all have the value "1". For the set of employees (E), those belonging to the lower rough set ($\underline{P}$) were identified, or, in other words, those which, certainly, were correctly replicated ("1" in B1, B2 and B3):

$$\underline{P}\,E = \{E2, E5, E6\} \quad (21)$$

For those which may have been correctly replicated in B1, B2 and B3, the higher rough set was obtained:

$$\overline{P}\,E = \{E1, E2, E3, E4, E5, E6, E7, E8, E9, E10\} \quad (22)$$

The borderline region is therefore:

$$\overline{P}\,E - \underline{P}\,E = \{E1, E3, E4, E7, E8, E9, E10\} \quad (23)$$

As the lower ($\underline{P}\,E$) and higher ($\overline{P}\,E$) rough sets are distinct, it is deduced that the set of employees in question can be dealt with by Rough Set Theory. The following equivalence relations were established for the set of employees considered:

$$R_{B1} = \{\{E2, E3, E4, E5, E6\},\{E1, E7, E8, E9, E10\}\} \quad (24)$$

$$R_{B2} = \{\{E2, E5, E6, E7\}, \{E1, E3, E4, E8, E9,E10\}\} \quad (25)$$

$$R_{B3} = \{\{E2, E3, E4, E5, E6, E7\}, \{E1, E8, E9, E10\}\} \quad (26)$$

These relations were obtained in the following way: firstly, a subset was formed for those employees which had the value "1" (one) and another subset for those with the value "0" (zero). Using the previous relations and the algorithm developed as a base, a main relation (R) was established, identifying first of all the employees with the value "1" and, next, the subsequent employees:

$$\mathbf{R} = \{\{E2, E5, E6\}, \{E1, E8, E9, E10\}, \{E3, E4\}, \{E7\}\} \quad (27)$$

To discover if each relation RB1, RB2 or RB3 is indispensable in relation to R (Pawlak, 1991), established a new common relation (RR), consecutively suppressing the relations R_{B1}, R_{B2} and R_{B3}:

$$\mathbf{RR}_{B1} = \{\mathbf{R} - \{R_{B1}\}\} = \{\{E2, E5, E6, E7\}, \{E1, E8, E9, E10\}, \{E3, E4\}\} \quad (28)$$

$$\mathbf{RR}_{B2} = \{\mathbf{R} - \{R_{B2}\}\} = \{\{E2, E3, E4, E5, E6\}, \{E1, E8, E9, E10\}, \{E7\}\} \quad (29)$$

$$\mathbf{RR}_{B3} = \{\mathbf{R} - \{R_{B3}\}\} = \{\{E2, E5, E6\}, \{E1, E8, E9, E10\}, \{E3, E4\}, \{E7\}\} \quad (30)$$

As the relations RR_{B1} and RR_{B2} are different to R, R_{B1} and R_{B2} are indispensable. As RR_{B3} is equal to the relation R, R_{B3} is dispensable. In order to find the "reducts", a new relation must be identified, (RT), for each pair of relations {B1, B3} and {B2, B3}:

$$\mathbf{RT}_{B1B3} = \{\mathbf{R} - \{R_{B2}\}\} = \{\{E2, E3, E4, E5, E6\}, \{E1, E8, E9, E10\}, \{E7\}\} \quad (31)$$

$$\mathbf{RT}_{B2B3} = \{\mathbf{R} - \{R_{B1}\}\} = \{\{E2, E5, E6, E7\}, \{E1, E8, E9, E10\}, \{E3, E4\}\} \quad (32)$$

As $\mathbf{RT}_{B1B3} \neq R_{B1}$ and $\mathbf{RT}_{B1B3} \neq R_{B3}$, the relation {B1, B3} is a "reduct". As $\mathbf{RT}_{B2B3} \neq R_{B2}$ and $\mathbf{RT}_{B2B3} \neq R_{B3}$, the relation {B2, B3} is also a "reduct".

Given the reducts {B1, B3} and {B2, B3}, we have:

$$\{B1, B3\} \cap \{B2, B3\} = \{B3\} \quad (33)$$

In other words, it has been identified that B3 is the "core" of this information system, according to the proposition CORE(P) = ∩ RED(P), where RED(P) is the family of all the "reducts" of P (Pawlak, 1991). By Table 12, it can be observed that there was no inconsistency in the replication for the database B3. The database B3 represents the best alternative for a sensitivity analysis in the face of the inconsistencies detected. In order to facilitate the application of Rough Set Theory, the algorithm was implemented in the computer language, Borland ® Delphi/Pascal, version 2007, by the first author of this study. In addition, two other simulations were carried out:

1 - with 1,381 registrations and 3 occurrences of inconsistency: employees "1313" (in B1), "0055" (in B3) and "0501" (in B1) (Figure 3). In this case, B2 was replicated correctly.

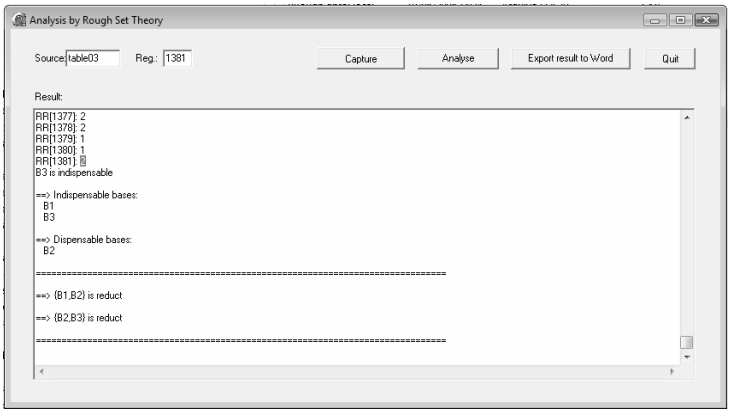

Figure 3 – Simulation with three occurrences of inconsistency – phase: results

By a sensitivity analysis, it was established that database B2 was replicated perfectly, comparing the indications of updating ("1") with what is registered in the main database B0: there are 116 employees with the executive function head of department (CH). In this way, the indication of the core (B2) by Rough Set Theory – intersection of the sets {B1, B2} and {B2, B3}, is in accordance with what was established. In addition, this result corroborated the sum of the cells with the value "1", in column "B2" (116 employees) (Figure 4).

Empl	B0	B1	B2	B3
1368	CH	1	1	1
1369	CH	1	1	1
1370	GR	0	0	0
1371	GR	0	0	0
1372	GR	0	0	0
1373	CH	1	1	1
1374	CH	1	1	1
1375	CH	1	1	1
1376	CH	1	1	1
1377	GR	0	0	0
1378	SD	0	0	0
1379	CH	1	1	1
1380	CH	1	1	1
1381	GR	0	0	0
	Σ	114	116	115

Figure 4 – Simulation with 3 occurrences of inconsistency – total of "heads of department"

For reasons of simplification, Figure 4 only reproduces the end of the spreadsheet.

2- With 1,381 registrations and 4 occurrences of inconsistency: employees “1313” (in B1), “0055” (in B3), “0501” (in B1) and “0202” (in B2) (Figure 5).

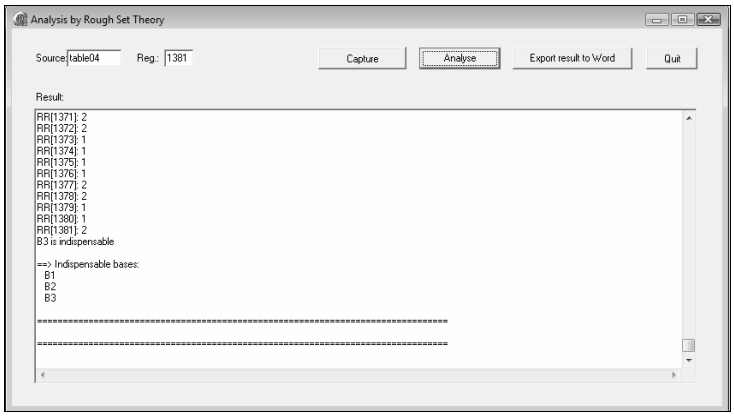

Figure 5 – Simulation with 4 occurrences of inconsistency – phase: results

As there was inconsistency in the replication in B1, B2 and B3, the result suggested indicates that all the bases are “indispensable” (Figure 5). Probably, the best result of the research (quantitative consultation of heads of department) should be a “mixture” of the results obtained from each database (Figure 6), or, simply be discarded and a new replication awaited.

Empl	B0	B1	B2	B3
1368	CH	1	1	1
1369	CH	1	1	1
1370	GR	0	0	0
1371	GR	0	0	0
1372	GR	0	0	0
1373	CH	1	1	1
1374	CH	1	1	1
1375	CH	1	1	1
1376	CH	1	1	1
1377	GR	0	0	0
1378	SD	0	0	0
1379	CH	1	1	1
1380	CH	1	1	1
1381	GR	0	0	0
	Σ	114	115	115

Figure 6 – Simulation with 4 occurrences of inconsistency - total of “heads of department”

For reasons of simplification, Figure 6 only reproduces the end of the spreadsheet.

9. CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

For the environment of the company in question (a domestic financial institution) and, in the face of the situation found – replicated and inconsistent databases, in which the inconsistency is of an unknown nature for the study in question and of sporadic occurrence, it falls to the “decision maker” (the executive responsible for the management and use of the information) to adopt, at first, a standard of checking the results obtained: manually, each result of the consultation (ex. quantitative consultation of personnel) was checked against the consultations taken from another database. It should be noted that, in the simulation, there is a complete vision of the replicated data and their origin (B0).

However, in the environment of the company under study, consultation of replicated data is restricted, in other words, it is carried out by means of a specific application with access to a single source of data (B1, B2 or B3). There is no way to compare the result of a consultation with the origin of the data (B0), other than in a manual way. The Rough Set Theory, as a method of decision aiding, was shown to be adequate for the treatment of “indiscernibility”, through the indications of “reducts” and a “core” of data, when possible. Its applicability arises from the very nature of the data researched (from the company HR department): in this case, there is no additional information on the occurrence of the inconsistencies, but only the data itself.

One role for Rough Set Theory which is conjectured would be as a support tool in monitoring data replications. The restrictions pointed out in Rough Set Theory and the proposal of an alternative model (“relation of dominance”), by Greco, Matarazzo and Slowinski (2005), are also shown to be an alternative to classic Rough Set Theory, which can be applied to the study in question. Specifically in this study, the extension to classic RST proposed by Ziarko (1993a, 1993b) - "model VP" (variable precision), was not used given the small proportion of occurrences of inconsistency in the simulations: both the first (3 occurrences for 1,381 registrations, or 0.22%) as in the second simulation (4 occurrences for 1,381 registrations, or 0.29%).

That is, the proportion of boundary regions (difference between the higher and lower approximations) was small. Therefore, given the situation (data replication with inconsistency), the RST was shown to be an appropriate tool for the detection of these inconsistencies, as well as suggesting the essential sources and, possibly, the source of the principal data, as options to the set of replicated data.

REFERENCES

- BIT, M.; BEAUBOUEF, T. (2008): "ROUGH SET UNCERTAINTY FOR ROBOTIC SYSTEMS". Journal of Computing Sciences in Colleges, Association for Computing Machinery (ACM) – n. 23 (i. 6) – pgs. 126-132.

- CODD, E. F. (1970): "A RELATIONAL MODEL OF DATA FOR LARGE SHARED DATA BANKS". Communications of the ACM, n. 13 (6), pgs. 377-387.

- GOMES, L. F. A. M.; GOMES, C. F. S. (2001): "UMA TÉCNICA DE DATA MINING: PRINCÍPIOS BÁSICOS DOS CONJUNTOS APROXIMATIVOS E SUAS APLICAÇÕES". Revista ANGRAD, 2 n. (1), pgs. 13-22.

- GOMES, L. F. A. M.; GOMES, C. F. S.; ALMEIDA, A. T. (2006): "TOMADA DE DECISÃO GERENCIAL: ENFOQUE MULTICRITÉRIO". Atlas, São Paulo, 289 pgs.

- GRECO, S.; MATARAZZO, B.; SLOWINSKI, R. (2005): "DECISION RULE APPROACH". In: Figueira, J.; Greco, S. and Ehrgott, M. (EDS.) Multiple criteria decision analysis state of the art surveys. Springer, New York. Science + Business media, cap. 13, pgs. 507-561.

- GRZYMALA-BUSSE, J. W. (1988): "KNOWLEDGE ACQUISITION UNDER UNCERTAINTY – A ROUGH SET APPROACH". Journal of Intelligent and Robotic Systems, n. 1, pgs. 3-16.

- LIN, T. Y. (2008): "ROUGH SET THEORY IN VERY LARGE DATABASES". Available on: http://www.cs.sjsu.edu/~tylin/publications/paperList/82_rs_dm8.pdf>.

- NOWICKI, R. (2008):"ON COMBINING NEURO-FUZZY ARCHITECTURES WITH THE ROUGH SET THEORY TO SOLVE CLASSIFICATION PROBLEMS WITH INCOMPLETE DATA". IEEE Transactions on Knowledge and Data Engineering. Available on: <http://ieeexplore.ieee.org/Xplore/login.jsp?

- URL=/IEL5/69/4358933/04487067.PDF?TP=&ARNUMBER=4487067&ISNUMBER=4358933>.

- PAWLAK, Z. (1991): "ROUGH SETS. THEORETICAL ASPECTS OF REASONING ABOUT DATA". Kluwer Academic Publishers, Dordrecht, 229 pgs.

- PAWLAK, Z. (2000): "ROUGH SETS AND DECISION ANALYSIS". Information Systems & Operational Research, n. 38 (3), pgs. 132-144.

- PAWLAK, Z.; GRZYMALA-BUSSE, J.; SLOWINSKI, R.; ZIARKO, W. (1995): “ROUGH SETS”. Communications of the ACM, n. 38 (11), pgs. 89-95.

- PAWLAK, Z.; SLOWINSKI, R. (1994): “ROUGH SET APPROACH TO MULTI-ATTRIBUTE DECISION ANALYSIS”. European Journal of Operational Research, Invited Review, n. 72, pgs. 443-459.

- ROY, B.; BOUYSSOU, D. (1993): "AIDE MULTICRITÈRE À LA DÉCISION: MÉTHODES ET CAS". Economica, Paris.

- SON, S. H. (1988): “REPLICATED DATA MANAGEMENT IN DISTRIBUTED DATABASE SYSTEMS”. Sigmod Record, n. 17 (4), pgs. 62-69.

- TSUMOTO, S. (2000): “AUTOMATED KNOWLEDGE DISCOVERY IN CLINICAL DATABASES BASED ON ROUGH SET MODEL”. Information Systems & Operational Research, n. 38 (3), pgs. 196-207.

- ZIARKO, W. (1993a): “ANALYSIS OF UNCERTAIN INFORMATION IN THE FRAMEWORK OF VARIABLE PRECISION ROUGH SETS”. Foundations of Computing and Decision Sciences, n. 18 (3-4), pgs. 381-396.

- ZIARKO, W. (1993b): “VARIABLE PRECISION ROUGH SET MODEL”. Journal of Computer and System Sciences, n. 46 (1), pgs. 39-59.

ACKNOWLEDGEMENTS

Research leading to this article was partially supported by the National Council for Scientific and Technological Research of Brazil (CNPq) through Project No. 310603/2009-9.