Using the KSDM Methodology for Knowledge Discovery from An Employment Domain where Repeated Very Short Serial Measures with a Blocking Factor are Present

JORGE RODAS
Knowledge & Software Engineering Unit. ITTC-IIT-UACJ. Cd. Juárez, México 32310
Email: jr.platon@gmail.com

GABRIELA ALVARADO
Business School. University of Barcelona. Barcelona, SPAIN 08034
Email: geac.sinai@gmail.com

Abstract

To present the Knowledge Discovery in Serial Measurement (KDSM) methodology for analyzing repeated very short serial measures with a blocking factor. An application to employment the domain is described using KDSM. Novel knowledge about labor domain's behavior was obtained once KDSM was applied to this specific domain. KDSM has shown that important information has been missed, especially when the kind of data explained in this paper is analyzed with common statistical methods and artificial intelligence techniques, independently employed.

Key Words: *KDSM Methodology, Employment Domain, Repeated very short Serial Measures, Knowledge Discovery.*

Introduction

Nowadays, whatever the task to be performed, there are many technological improvements in computational tools that make that performance easier. This is true for all fields of the human domain, for example: Economics, Education, Law, Medicine or any engineering areas. Some instances of the above are a computational way of monitoring a procedure, the follow up of a patient under therapy, the obtention of a chemical by synthesis, etc. This involves a great deal of information coming from the procedure, as well as from the *actors* or *events* in it. Very often it is possible to find that many data are obtained while monitoring the procedure, as a result of serial measures during the time the procedure lasts.

Research

Repeating the procedure as many times as necessary performs the research and the gathering of data. It appears that the characteristics of *repeated serial* measures do not constitute a serious problem when they are analyzed with classical time series techniques; nevertheless, the following question arises: What will happen when the number of measures is extremely small?

Frequently, in a situation like this, much additional information is available from actors of the procedure and the procedure itself. However, this information is non-serial, but it is closely related to what takes place in the procedure. In addition, the actors often constitute a blocking factor acting on serial measures, which gives rises to the following questions: How is it possible to take advantage from this additional information? How could this additional information be handled in relation to repeated serial measures if such information is not obtained by measuring the characteristics of actors taking part in the process or procedure?

To answer these questions, the KDSM methodology has been established. Such methodology allows the discovery of knowledge in domains where repeated serial measures are too short and present a blocking facto—composed of individuals—,and where additional information about the actors of the procedure, such as their characteristic attributes, is available.

The KDSM methodology allows the three following tasks to be performed:

1. Identification of individuals' relevant attributes in the first measures obtained which are used to establish the initial conditions;
2. analysis of the effect of each isolated event, thus eliminating the blocking factor; and
3. identification of the events' relevant characteristics, as well as its structure and its subsequent interpretation.

The origins and fundamentals of the KDSM methodology are duly documented in (Rodas and Rojo, 2005), and it is possible to say that such methodology has been very successful when applied in a psychiatric domain. In this paper, the application of the KDSM methodology to a domain not related to psychiatry, namely the employment domain, is presented.

Paper Structure

This paper is structured as follows: the next section introduces the problem formulation. In KDSM Methodology section, the steps of the methodology are described. In Study domain section the domain in which our case study is stated is summarized. Section Applying KDSM to employment domain shows a brief account of the results of the KDSM methodology application to this domain study. Finally, Conclusions about the application of the KDSM methodology are presented and the last section is devoted to Future work.

Problem Formulation

In Figure 1, the representation of a series of *Individuals* ($i_1 \ldots i_n$) can be seen, in which *m* occurrences of a given *event E* take place at different time points ($E_{i1} \ldots E_{im}$). Linked to each event occurrence, there exists an *attribute of interest Y* that affects the individual's behavior. The study of its evolution over a given period of time [t_1,t_r] following each occurrence of *E* is the objective of this research.

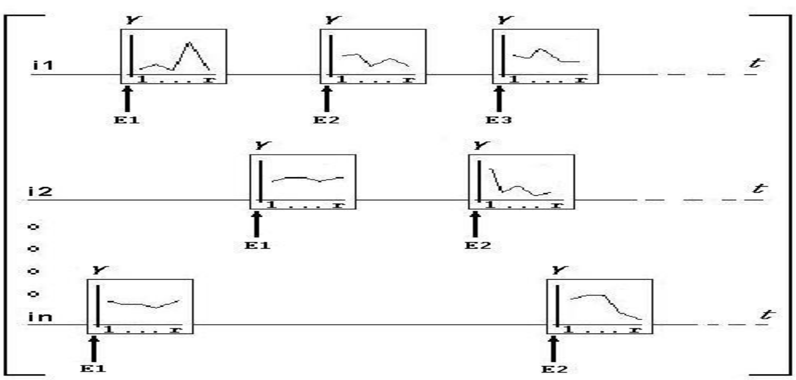

Figure 1. Measure of *Y* after each occurrence of *E* for each *i*.
Adapted from: Rodas, Gibert and Rojo 2002

Thus, a certain fixed number of measurements (r) of Y is taken at fixed time intervals, beginning from the occurrence of each E, for each individual and each occurrence of E.

This scenario generates information that can be structured as follows:

1. For each i a set of quantitative and/or qualitative characteristics $X_1 \ldots X_K$ is available. This gives rise to matrix X, which corresponds to $[x_{ik}]$, where $i=\{1 \ldots n\}$ and $k=\{1 \ldots K\}$.
2. For each occurrence of E, a sequence of serial measures of Y at all fixed time points is obtained. Let E_{ij}, $i=\{1 \ldots n\}$ and $j=\{1 \ldots m\}$ be the *j-th* occurrence of event E on individual i. Hence, for a given individual i there exists a number m of occurrences of E. Considering that time resets at 0 for each occurrence of E, it is possible to set $t_1 \ldots t_r$ as the time points when interest attribute Y will be measured after the occurrence of E. The measures of Y generate a second matrix Y corresponding to $[Y^t_{ij}]$; $i=\{1 \ldots n\}$ and $j=\{0 \ldots m\}$.
3. For each E a set of quantitative and/or qualitative characteristics $Z_1 \ldots Z_L$ is available. This gives rise to matrix Z, which corresponds to $[z_{ijl}]$, where $i=\{1 \ldots n\}$, $j=\{0 \ldots m\}$ and $l=\{1 \ldots L\}$.

The measures of the attribute of interest are given by Y^r_{ij}, where $i=\{1 \ldots n\}$ is the individual, $j=\{1 \ldots m\}$ indicates the *j-th* occurrence of E on individual i and $r=\{1 \ldots R\}$ indexes the time instant from the occurrence of E_{ij} when interest attribute Y was measured. It must be pointed out that *the measurement times are the same* with respect to the occurrence of all the events for all the individuals.
Once an i,j has been determined, the measures of interest attribute Y in the time period of $t_{1 \ldots} t_r$ may be represented by *very short curves* (*where r is small*) apparently independent of each other.

In fact, each individual is independent of the others. Consequently, the *number of events* and the *time instant* at which they occur may differ from individual to individual without any underlying pattern.

Nevertheless, all the events on the same individual are affected by his/her own characteristics, which causes all serial measures relative to a particular individual $\{Y^1_{ij} \ldots Y^r_{ij}\}$, $j=\{1 \ldots m\}$ to receive the individual's influence.

As a result, on matrix Y, individual i may be regarded as a *blocking factor* (Antony, 2003) defining *blocks* of curves which are not independent of each other (see Table 1).

Table 1. Serial measures blocks

	t_1	t_2	$\ldots$	t_R	
E_{11}	Y^1_{11}	Y^2_{11}	$\ldots$	Y^r_{11}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	block 1
E_{1m}	Y^1_{1m}	Y^2_{1m}	$\ldots$	Y^r_{1m}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
E_{n1}	Y^1_{n1}	Y^2_{n1}	$\ldots$	Y^r_{n1}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	block n
E_{nm}	Y^1_{nm}	Y^2_{nm}	$\ldots$	Y^r_{nm}	

Adapted from: Rodas, Gibert and Rojo 2002

A block is therefore constituted by all serial measures $\{Y^1_{ij} \ldots Y^r_{ij}\}$, $j=\{1 \ldots m\}$ following any occurrence of E on the same individual i. These series are composed of a *small number* of measurements over a specific time period where few observations are present. However, the number of measurements is the same after each event, and so is the distribution over time with respect to the event occurrence. In particular, *a set of very short serial measures over time with a blocking factor* will be analyzed.

The objective of KDSM is to find the pattern followed by the serial measures $\{Y^1_{ij}\ldots Y^r_{ij}\}$ and the characteristics of the individual $X_1 \ldots X_K$ and the event $Z_1 \ldots Z_L$ related to the temporal evolution of the attribute of interest *Y*. Nonetheless, considering the descriptions of matrices *X*, *Y*, and *Z* in previous paragraphs, it is obvious that they may not be related because of their incompatibility. Thus, a way to manipulate these matrices must be found to be able to analyze them.

If there were a common pattern for occurrences of *E* on all individuals, a single series per individual could be considered and analyzed by *means* of the *intervention policy* (Matarise, 2011) of statistical time series. However, this is not the case. For this reason, it is inadequate to resort to a classical temporal analysis. This situation would imply too rigid a hypothesis for many of the real situations to be covered.

For like problems, a method of analysis based on Matthews's ideas (Matthews, 1993) is often employed. The method consists in reducing the number of series of each individual to a single series that summarizes the whole set either through the *mean* of each instant (*thick line* in Figure 2), the *mean area* per series or *mean tendency* per series. This would allow the measures of interest attribute *Y* to be reduced to a single row per individual, and matrices *X*, *Y*, and *Z* would enable a classical analysis. Nevertheless, relevant information would often be lost, since variability depends on both each event and each *individual effect* (see Figure 2). The conclusions drawn from the study of such transformation may therefore be far from reality.

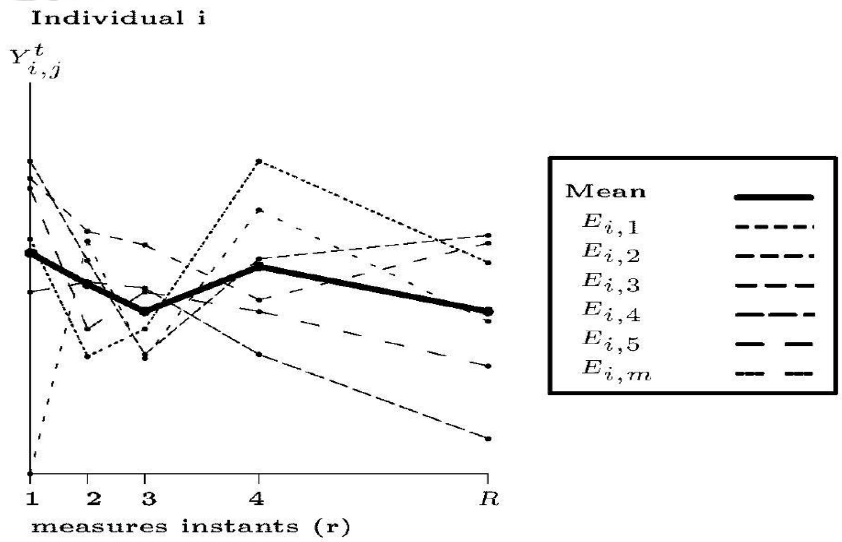

Figure 2. Example curves of serial measures related to an individual *i*
Adapted from: Rodas, Gibert and Rojo 2002

Figure 2 shows a hypothetical situation for an individual *i*, where the curves represent the recorded serial measures after each event *E* at the different measure instants *r*. Despite describing a general tendency of individual *i*, the mean (thick line) summarizes information excessively, and therefore details that might be significant are lost.

In summary, our formal problem is as follows:

Given a set of individuals $I=\{i_1\ldots i_n\}$, a set of attributes (quantitative and /or qualitative) $X_1 \ldots X_K$ which define *I*, a matrix $X=[x_{ik}]_{nK}$, $i=\{1\ldots n\}$, $k=\{1\ldots K\}$, x_{ik} being the value of X_K for *i*; given the attribute of interest configuring the serial measures Y^t_{ij}, $t=\{0\ldots r\}$, for $i=\{1\ldots n\}$, $j=\{1\ldots m\}$, the matrix containing all the serial measures over time $[Y^t_{ij}]_{Nr}$, $N=\Sigma^i_1 m$, $t=\{0...r\}$; a set of events $E=\{E_{i1}…E_{im}\}$, a set of attributes (quantitative and/or qualitative) $Z_1...Z_L$ which characterize E, matrix $Z=[z_{ijl}]_{nmL}$, $i=\{1...n\}$, $j=\{1...m\}$, $l=\{1...L\}$, z_{ijl} being the value of Z_L for E; and considering:

- The individuals $I=\{i_1…i_n\}$ act as a blocking factor on matrix Y,
- The measure points $t=\{0...r\}$ represent a *fixed* number over time for all the serial measures,
- The number of observations per serial measure r is small, and
- For each i there exists a *variable* number of serial measures m

A model for the behavior of serial measures will be established. The model will specify:

- The pattern(s) for the behavior of the serial measures $\{Y^t_{ij}\}$ and
- Their relationship with matrices X and Z.

KDSM Methodology

Based on our experience in real application, the steps of the KDSM methodology are presented in this section. Its justification is provided in (Rodas and Rojo, 2005).

KDSM performs three main tasks:

1. Identification of the individuals' different initial profiles by studying the baseline serial measures, Y^t_{i0}, and their relationship with matrix X,
2. from the knowledge induced in the previous task, the serial measures Y^t_{ij}, $t=\{0...r\}$ are studied and the possible patterns for these measures are obtained, and
3. the results of the previous task are crossed with matrix Z to find relationships between them and to obtain new knowledge.

The steps of the *methodology for knowledge discovery in serial measures* (KDSM) are the following:

1. *Extraction of a baseline matrix Y_0 from a database containing repeated very short serial measures over time (taking Y_0=[Y^t_{i0}], $i=\{1...n\}$, $t=\{0...r\}$).*
2. *Hierarchical clustering of I using the baseline matrix Y_0.*
3. *Interpretation of the resulting classes of Y_0* from the individuals' characteristics (matrix X).
4. *Rules induction from comparison between classes, on the basis of attributes X_k.*

The relevant characteristics which determine the *a priori* structure are searched for. In this step, multiple boxplots are used as a first alternative for comparison. The resulting rules, type *crisp* (Predicates' logic rules) are the basis for the initial and partial knowledge (KB_0) of the domain. Other methods are currently being analyzed, like the Automatic Generation of Conceptual Interpretation of Clustering (Pérez-Bonilla and Gibert, 2007) based on fuzzy rules and others developed from the Kruskal-Wallis method (Baguley, 2012).

5. *Construction of the matrix $D=[Y^t_{ij}- Y^t_{ij-1}]$ to measure the effect of a given event on the attribute of interest.*
6. *Classification Based on Rules of matrix D with knowledge base KB_0.*
7. *Interpretation of resulting classes*.

The general pattern of the curves of each class is characterized, and internal variability in each class and between them is studied and graphically represented by: a) the curves of each class plus the mean curve, and b) the mean curves of all the classes.

8. *Class analysis*.

The characteristics of matrix Z are projected over the classes generated from matrix D, and those which are relevant and determine a particular behavior of the individuals are searched for.

Study Domain

The State Employment Service emerges when the *Secretary of Labor and Social Prevention* (SLSP) (a dependence of the Executive Federal) and the *state governments* (Corresponding to the governor of each state) of Mexican Republic start and supervise the development of programs with a social mission, whose objective is to promote the design and application of policies devoted to the generation of employment.

The actions of the Department of State Employment Service are directed to connecting people without a job, and the needs of manpower of the productive sector, promoting the productive insertion of workers and the progressive interaction between them. To ensure the achievement of these objectives, among other relevant tasks, a program called "Training Scholarship Program for Unemployed Workers" (TSPUW) was implemented.

The KDSM methodology is applied to obtain knowledge from the evolution of TSPUW, as well as to detect the effectiveness of the analysis of the information gathered related to the assignment of the outcome of the training program, characteristics of the municipalities in the different states of Mexican Republic involved, and the needs of the productive sector. The information processed by KDSM will enable the Secretary of Labor and Social Prevention, (SLSP) to know the effect of each imparted training course in the municipality where TSPUW acts. This Secretary will then play a vital rol in the development of the training program, according to the global trends of labor markets, in a more effective and opportunistic way.

The analysis includes: all municipalities' characteristics (Matrix X), serial measures of employed-people quantity (Matrix Y) over a fixed period (3 months), taken in six times (one every 15 days), and all courses' characteristics (Matrix Z). This analysis is not trivial, since there are sets of measures of every course type (around 180 different specialties of courses) where each one of the 2427 municipalities acts as a blocking factor over serial measures and courses' characteristics. That is, there is one package of serial measures and another one of courses' characteristics—of that specific municipality—for each municipality.

This kind of data leads any data analyst to use some special technique or methodology that allows him/her to distinguish the really important information in order to achieve the main objective of this research, the TSPUW effectiveness. Since the expert who participates in this research work is familiar with the data of Chihuahua State in Mexico, the first step was to work with the information about this state. In fact, the study of Chihuahua began with only one course analysis—of the textile field—as Chihuahuas' State Employment Service is very interested in this analysis due to its popularity in the productive sector.

Applying KDSM to employment domain

In order to perform the data analysis of Chihuahua State, the KDSM methodology was applied. This section also shows a problem that arose during the performance of the first task of the KDSM, namely no relevant attribute was obtained that could describe the structure of individuals with the technics used up to date by the methodology. That is why a simple solution proposal together with the results obtained with it are introduced.

Identification of municipalities' relevant attributes

After steps 1—4 of the KDSM methodology section, a knowledge base constituted by rules that describe the structure of the municipalities was obtained.

When individuals from real and complex domains are described by quantitative attributes, finding attributes that characterize the different classes in a certain partition is often a very difficult task. Thus, in order to perform an adequate analysis of these domains, it is convenient to work with *attributes that* *partially characterize* (Pérez-Bonilla and Gibert, 2007) the classes; that is, the use of attributes common to all the classes. This situation may be represented by a rules system that handle a certain degree of fuzzy membership for each class in a certain partition (fuzzy rules). Then, obtaining attributes that really characterize classes becomes a serious problem. The characterization and quality-determination process of resulting classes from a reference classification is therefore no trivial task, since there is no objective criterium to determine class quality. That is why an expert's subjective point of view is frequently used in order to rate the classification quality based on the utility or the meaning of classes.

The following steps are our first analysis to an effective characterization process for obtaining a quality or useful classification obtention.

1. Descriptive statistics which provide preliminary information about all the data to be analyzed.

2. Experts' *a priori* knowledge must be included in order to obtain the semantic restrictions (rules) of the partition's resulting classes to facilitate the task of finding a useful meaning.

3. A system of rules from the characterize classes is necessary in order to obtain their relevant attributes using the Automatic Generation of Conceptual Interpretation of Clustering.

4. The determination of class quality in the expert's own terms about the "meaning"' and "usefulness" of a class based on the rules obtained in the previous step. If after a classification process the discovered structure is not useful, then repeat the whole process, including steps 2 and 3 of the KDSM methodology, in order to consider: a new classification where new semantic restrictions (rules) may or may not be included, the experts' additional knowledge, or fuzzy rules composed of combination of attributes to achieve a new structure with a better meaning for the objective study than the last structure. This process is repeated until a useful classification is obtained or the expert and data analyst decide to stop this process.

5. Interpretation of the obtained results and the discovered structure from data. This is a *new knowledge* that may be used in order to make a decision and stop this process or to continue with the second task of the KDSM methodology.

The rules obtained after the first 3 steps offered no useful knowledge. For this reason, the expert provided fuzzy rules from a combination of attributes—of interest to his/her study objective—in order to give usefulness to the obtained classes. To make these rules, he/she defined the ranges of values for the attributes in three linguistic labels, small, medium, and large, according to his/her own experience. The attributes of interest that he/she took in to account were: a) Territorial Stretching (TS) since it reflects an open opportunity for the business sector to establish new enterprizes or to maintain only the minimum number of already existent ones. This attribute is an indicator of how many and what kind of courses it is appropriate to offer for the Secretary of Labor and Social Prevention (SLSP). b) Non-working Population (NP), since it indicates the number of unemployed people.

Once the data analyst had obtained the fuzzy rules, the expert evaluated their level of structure representation and determined which ones are useful for his/her research objective. Further details may be found in (Rodas, Alvarado and Vázquez, 2007). Finally, the fuzzy rules—those that were considered by the expert—are used as a "bias" for the next KDSM methodology task.

Analysis of each course effect

The KDSM second task consists of steps 5─7 corresponding to the analysis of the effect of each isolated course, using *differences* to eliminate the blocking factor constituted by the municipalities over the courses.

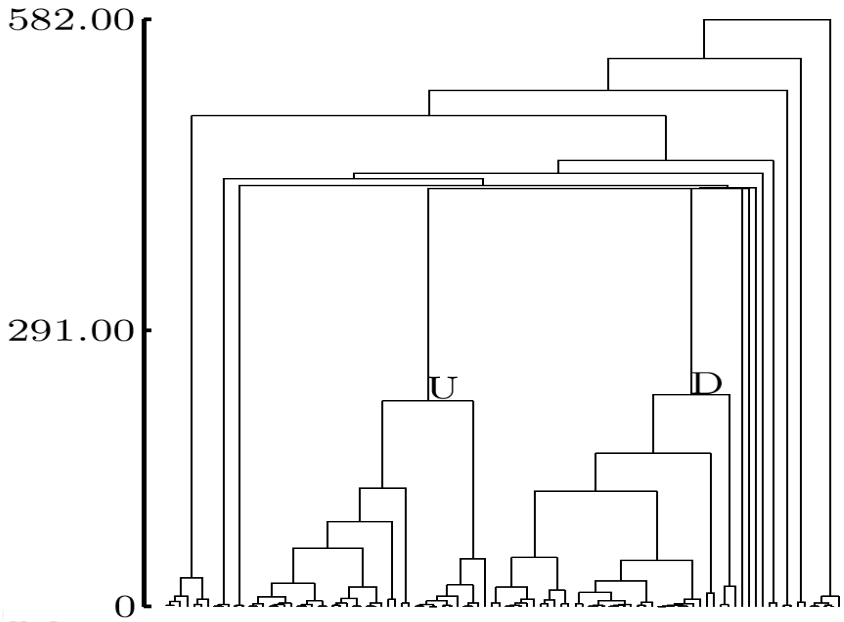

Figure 3. Structure of courses
Adapted from: Rodas, Alvarado and Vázquez, 2007

From Figure 3, the expert determined that the most appropriate cut was in 13 classes, since 2 of these classes are very important, interesting and useful for the expert's research objectives. The expert found a very important and novel results that are shown later. The 11 remaining classes are merged into only one class named residual class. The expert decided to omit this residual class from the analysis as it does not have information of interest to him/her. In addition, these classes represent only 20% of the data, which, in this particularly case study, is irrelevant.

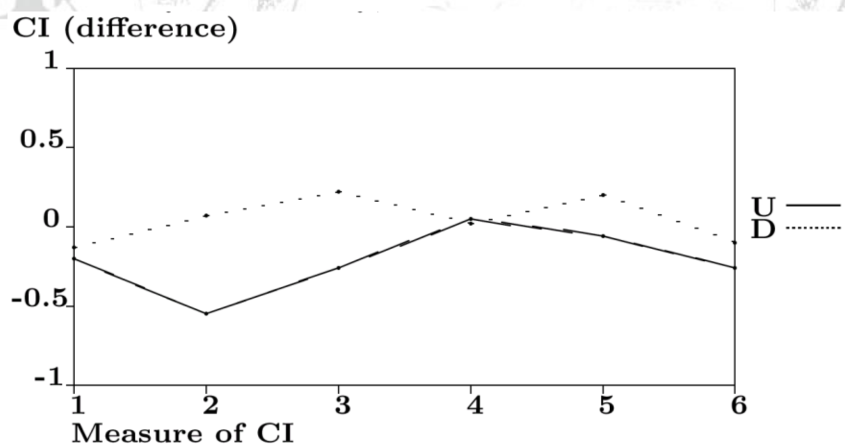

Figure 4. Two classes mean curves
Adapted from: Rodas, Alvarado and Vázquez, 2005

In Figure 4 the class U (line) has negative values, with only one exception in the 4th measure. This means that the contracting thickness occurs in this specific measure. Class D (dotted line) has positive values. Even though these values are discrete, they indicate that, in general, contracting is stable during the monitoring time of TSPUW.

Identification of courses' relevant features

The third and last KDSM methodology task consists of steps 8 and 9 corresponding the search for relevant attributes of the events.

The trend of class U (line), Figure 4, shows a belated recovery in the fourth measure of the Contracting Index (CI) difference, considering that this difference indicates the contracting index evolution without the municipality effect. This trend is very closely linked to the course modality (Figure 5), which shows that the majority of these courses are *self-employment*. That is why all participants in this kind of course─when graduate─require more time to find a job, or to establish their own business by themselves.

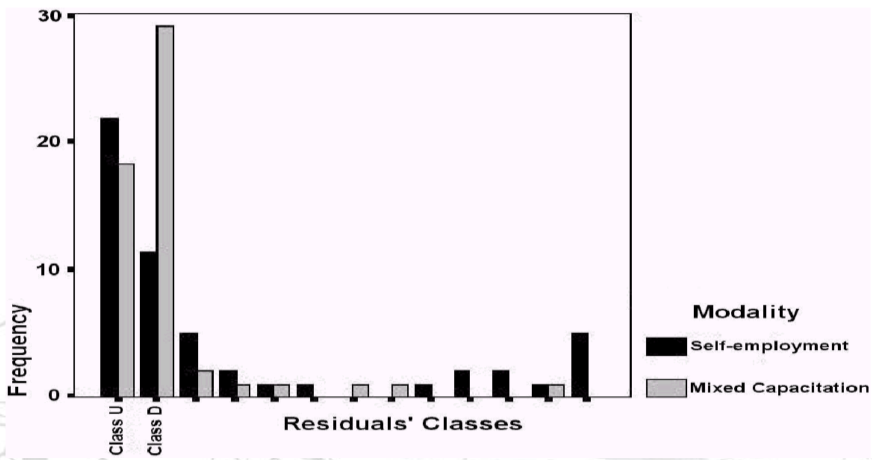

Figure 5. Courses modality per class
Adapted from: Rodas, Alvarado and Vázquez, 2007

On the other hand, in class D (dotted line) the majority of the differences are positive since the CI is directly related to the course modality, which can be seen in Figure 5, which shows that the majority of these courses are *mixed capacitation*. That is why all participants in this kind of course─when graduate─have a job in one of the enterprizes which participates in this mixed capacitation process.

Finally, there is a clear difference trend between both classes denoted by course modality, where each course has a particular effect which leads to the analysis of the attributes with a special influence in the behavior discovered. The Secretary of Labor and Social Prevention may then draw up a future planning of TSPUW on the basis of this information.

Conclusions

Three conclusions about the KDSM methodology may be drawn from this research work:

1. The KDSM methodology integrates common statistic tools, artificial intelligence and fuzzy logic in order to give a possible solution to the problem when there are no relevant attributes and it is necessary to characterize baseline matrix Y_0 in relation to data matrix X.
2. The use of fuzzy rules, constituted by a combination of attributes, allows a "useful" partition to be obtained for the expert's study objective.
3. While this kind of domains is being analyzed with a hybrid methodology (statistic and artificial intelligence such as the KDSM methodology) a great quantity of important data is recovered. Otherwise, this information would be highly summarized and many important characteristics would not be distinguished.

As for the application of KDSM in the employment domain, it must be born in mind that one of the function of the SLSP is to follow up the results obtained with TSPUW program by measuring their efficiency with a cost/benefit relationship, that is, the economic investment with respect to the graduate people that has joined to the labor sector. For this reason, the KDSM methodology has been applied in order to obtain important results from which the following conclusions are obtained:

The government's economic investment is low when the course modality is *mixed capacitation* due to a direct participation of enterprizes.

Thus, the SLSP could optimize the TSPUW program to make the most of the opportunity niche that the government-business coordination in capacitation represents for the government, business sector, and unemployed workers.

If business distribution, course modality and municipalities' features are known, the SLSP could draw up an optimum capacitation planning, determining which of the different course modalities, their contents, etc will meet the economic needs of municipalities and business sector, maximizing the cost/benefit relationship between all the involved sectors.

Finally, the SLSP will be able to face a situation in which the market no longer contracts enough graduate people and the balance of the cost/benefit relationship is broken. To this end, the SLSP makes use of the knowledge provided by the attributes relative to the population, total number of matriculations to a specific course, of graduated and of contracted people.

In summary, the conclusions obtained from the KDSM methodology application to the *employment* domain are the following:

Support to the Decision-Making Process.

KDSM carries out a feedback process with knowledge obtained from TSPUW so that the SLSP acts accordingly.

Assist in optimization and Planning Tasks.

KDSM bases the decisions on the type of course and other characteristics, which allows the SLSP to improve TSPUW performance in a constant and permanent way.

Supply New Knowledge.

KDSM provides the SLSP with the knowledge obtained from the courses monitoring as a support when the SLSP needs to face an imbalance in the cost/benefit relationship.

Future Work

This work represents the first stage of the KDSM methodology application to the employment domain, in which some tasks have been identified to:

- improve the knowledge representation and the process of utilization and interaction of knowledge with the Kruskal-Wallis test and the Automatic Generation of Conceptual Interpretation of Clustering method, mentioned in KDSM Methodology section,
- formalize a mechanism to improve the process of achievement of class quality in terms of “usefulness” and the multi-attribute analysis of the characterization process of the baseline matrix,

- define the mechanism for obtaining rules when none of the attributes completely characterizes the classes and they are not relevant statistically speaking, and
- establish a characterization system for class conceptual descriptions from a reference partition, as well as an automatic interpretation-generation model for classes from a reference partition.

Acknowledgements

We would like to thank the Economic and Social Information Center of the Government of the State of Chihuahua, for providing the data for the present study.

References

Antony, J. (2003). Design of Experiments for Engineers and Scientists. Great Britain: Butterworth-Heinemann. Elsevier.

Baguley, T.S. (2012). *Serious Stats: A guide to Advanced Statistics for the Behavioural Sciences*. Palgrave Macmillan.

Matarise, F. (2011). *Intervention Analysis in Time Series*. Springer Berlin Heidelberg.

Matthews, J.N.S. (1993). A refinement to the analysis of serial data using summary measures. *In Wiley, Statistics in Medicine, 12*, 27─37.

Pérez-Bonilla, A. & Gibert, K. (2007). Towards Automatic Generation of Conceptual Interpretation of Clustering. *CIARP 2007, In Rueda, Mery and Kittler (Eds.), Springer's Lecture Notes of Computer S:cience Series, 4756*, 653─663. DOI: 10.1007/978-3-540-76725-1_68

Rodas, J., Alvarado, G. & Vázquez, F. (2005). Using the KDSM methodology for knowledge discovery from a labor domain. In Proceedings of SNPD/SAWN 2005 (pp. 64-69). USA: IEEE. DOI: 10.1109/SNPD-SAWN.2005.79

Rodas, J., Alvarado, G. & Vázquez, F. (2007). KDSM: Effectiveness Detection on Government Programs. *Latin American and Caribbean Journal of Engineering Education, 1*(1), 12─19.

Rodas, J. & Rojo, J.E. (2005). Knowledge discovery in repeated very short serial measurements with a blocking factor. Application to a psychiatric domain. *International Journal of Hybrid Intelligent Systems, 2*(1), 57─87.

Rodas, J., Gibert, K. & Rojo, J.E. (2002). KDSM Methodology for Knowledge Discovery from Ill-Structured Domains Presenting Very Short and Repeated Serial Measures with Blocking Factor. *Springer's Lecture Notes of Computer Science Series, 2504*, 228─238. DOI: 10.1007/3-540-36079-4_20.