Data Mining Applied to the Instrumentation Data Analysis of a Large Dam

Rosangela Villwock, Maria Teresinha Arns Steiner,
Andrea Sell Dyminski and Anselmo Chaves Neto
Federal University of Paraná
Brazil

1. Introduction

Dams are conceived with the purpose of bringing great benefits to society. It is expected that their construction, operation and eventual decommissioning should occur safely. If a dam breaks down the destruction scale may be very high; it may put not only the environment in the surrounding areas at risk but also human lives. Therefore, adequate design, construction and operation of dams are a worldwide concern. International guidelines aiming at dams' safety and many productive discussions about this theme have been proposed by the ICOLD – International Commission on Large Dams (ICOLD, 2007).

An adequate auscultation system must be present in dams in order to monitor their structures and foundations during life cycle period. Generally an auscultation system is composed by a set of instruments installed in important points of a dam and of the subsoil where its foundation is based on. These instruments generate a large amount of data, which should be used to understand dam behavior and help engineers in decision making process involving dam safety. Usually the instrumentation readings compose a huge set where important information is mixed with non relevant data. So it would be very useful to have an automatic tool capable to point the significant information or hierarchically organize instrumentation data.

The objective of this work is to present a data mining based methodology to group and organize data from a dam instrumentation system aiming to assist dam safety engineers. The purpose with this work was to select, cluster and rank 72 rods of 30 extensometers located at the F stretch of Itaipu's dam, by means of Multivariate Statistical Analysis techniques. The Principal Components Analysis was used as a method to select the extensometers' rods, Clustering Analysis identified the extensometer rods that were similar and Factor Analysis was used to rank the extensometer rods.

This text is organized as follows: the second section is about Instrumentation system and its relevance to dam safety, the third section describes the KDD Process, the fourth section is a brief description of what Clustering Analysis is, the fifth section describes Itaipu's Dam, the sixth section introduces the Methodology, the seventh section shows the results and the eighth one has the conclusions.

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

2. Dam safety

The concept of "Dam Safety" should involve structural, hydraulic, geotechnical, environmental and operational aspects. All these features must be considered during a dam's lifespan. A proper instrumentation system capable of monitoring the dam's geotechnical and structural behavior is essential to assess its behavior and integrity. Good overviews about the relevance of instrumentation to evaluate dam safety can be found in Dibiagio (2000) and Duarte *et al.* (2006).

Some objectives of dam instrumentation and its relationship with structural safety are described in two Engineering Manuals published by U.S.Army Corp of Engineers (1987 and 1995). There, the main objectives of a geotechnical instrumentation plan were grouped into four categories: analytical assessment; prediction of future performance; legal evaluation; and development and verification of future designs. Instrumentation may achieve these objectives by providing quantitative data to assess useful information like groundwater pressure, deformation, total stress, and water levels. Combining visual and periodical inspections with careful data analysis a critical condition can be revealed (FEMA, 2004).

Dam stability must be firstly analyzed during design phase. The geometry of the structures and the properties of the involved materials must be considered as well as the loading conditions. Some basic loading conditions like dam weight (W), hydrostatic pressure acting against dam wall (which resultant is $F_{reservoir}$) and uplift pressure due to seepage in the foundation rock mass (resultant F_{uplift}) are shown in figure 1. The effects of the loads in dam failure process can be many and two of them are shown in figure 1: sliding and overturning. Loading conditions and materials properties can change along dam life cycle and the instrumentation can catch some of these changes.

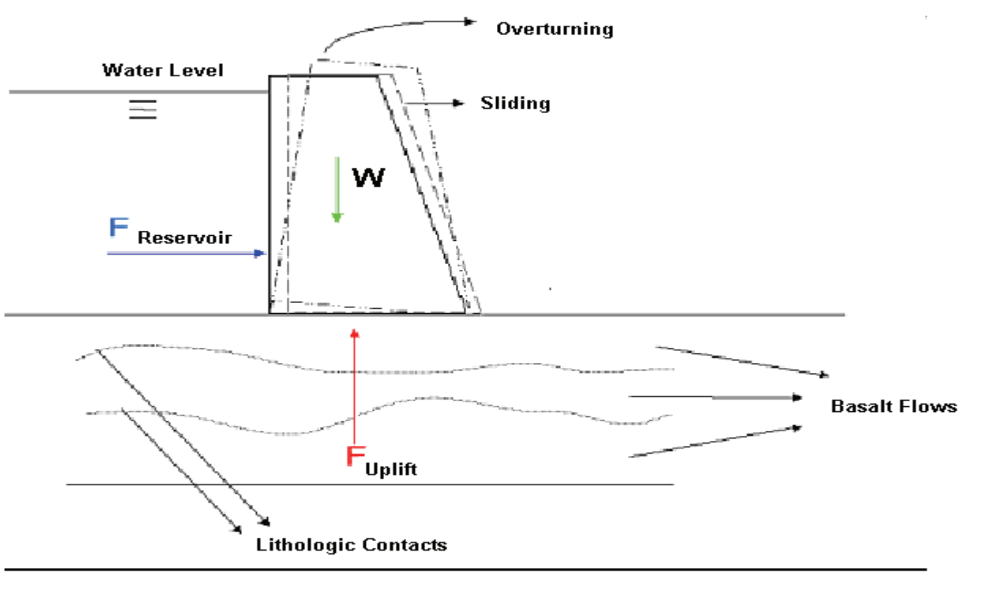


Fig. 1. Basic load condition and some failure processes in concrete gravity dams.

The instrumentation monitoring generates a large data set composed of periodical readings taken during several years. It is important to select the correct array of information to better understand the dam's behavior and solve occasional problems as soon as they occur. Decision making based on instrumentation data usually occurs throughout the lifespan of a dam. An interesting discussion about risk assessment and decision making in dam safety is presented in Bowles *et al*. (2003). Harrald *et al.* (2004) made a good review about some decision making systems and methodologies to help priorization of tasks and mitigation of failure risk.

Many times, a large amount of data contains useful information, called knowledge. Generally this information is not easily available or identified. Human analysts can spend weeks to discover this knowledge. Because of this fact some huge data sets never receive a detailed analysis (Tan *et al.*, 2005). The more the data volume increases, the more useful are the Data Mining techniques. According to Witten e Frank (2000), ingeniously analyzed data are an invaluable resource for decision-making.

3. The KDD process

The acronym KDD means "Knowledge Discovery in Databases" and Fayyad *et al*. (1996) define it as a non-trivial discovery process of valid, new, useful and accessible patterns. The main advantage of the discovery process is that no hypotheses are needed and knowledge is extracted from the data without previous knowledge.

KDD is related to the broad process of discovering information in a database in which there is an emphasis on a high-level application of the particular Data Mining (DM) method. While the DM step is characterized by the extraction of patterns hidden in the data, the whole KDD process is broader and includes all the processing (data selection, pre-processing and transformation) that is needed for this to occur, making it possible to evaluate and interpret the results that were obtained after the DM techniques were used.

The KDD process is a set of continuous activities that include five steps: Data Selection, Pre-processing, Formatting, Data Mining and Interpretation, as shown in Figure 2.

The process starts by understanding the application's domain and the targets that must be reached. Then, a selection can be drawn from these data so that one may work with the data that are of interest. The pre-processing step is the one in which missing or inconsistent data are analyzed and treated. During the formatting step data are prepared so Data Mining can be used as, for instance, to map categorical data among numerical data or using methods to reduce dimensions in the data. According to Silver (1996), pre-processing and formatting may take up to 80% of the time needed for the whole process.

Advancing along the process, there is the Data Mining step, the main one in the KDD process, in which several methods can be used to extract information, which are then presented to the last step, the interpretation, where knowledge is acquired.

If results are not satisfactory, the whole process may be fed back, changing some of the information, which may be reprocessed in the previous steps.

The main purpose with the KDD process is to obtain knowledge hidden in data that may be useful for decision-making, by using methods, algorithms and techniques from different scientific areas. According to Tan *et al*. (2005), these include Statistics, Artificial Intelligence, Machine Learning and Pattern Recognition.

According to Fayyad *et al*. (1996), Data Mining tasks are predictive and descriptive. The predictive ones use some variables to forecast unknown or future values of other variables, while the descriptive ones find patterns to describe the data. The main tasks of Data Mining are related to pattern Classification, Association and Clustering, In this work, the Data Mining task is to cluster patterns.

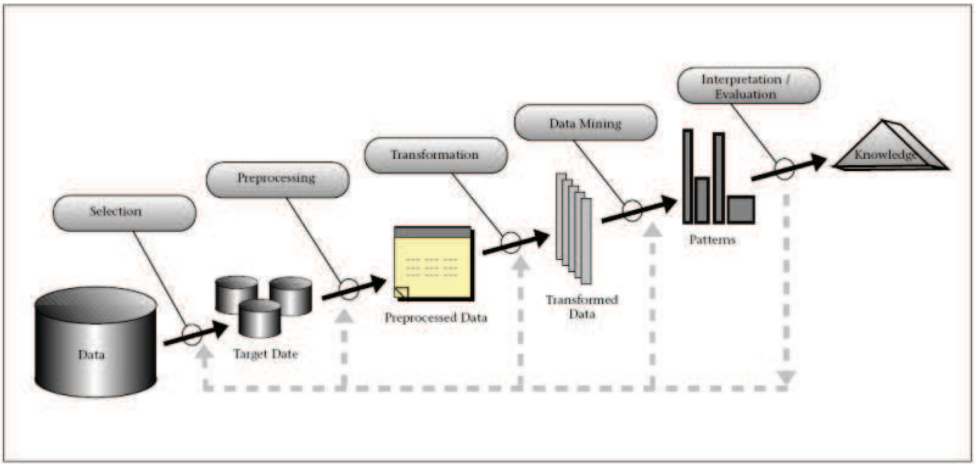


Fig. 2. Steps in the *KDD* process, Fayyad *et al.* (1996)

4. Clustering analysis

According to Tan *et al.* (2005), Clustering or Segmentation searches for clusters of patterns so that patterns that belong to a same cluster are more similar to one another and dissimilar to patterns in other clusters. According to Hair Jr *et al.* (2005), clustering analysis is an analytical technique to develop objects significant subclusters. Its purpose is to classify the objects into a small number of mutually excluding clusters. Freitas (2002) thinks that in Clustering Analysis it is important to favor a small number of clusters.

Clustering algorithms can be divided into categories, in several ways, according to some characteristics. The two main classes of clustering algorithms are the hierarchic and the partitioning methods. This work used the Ward Method, a hierarchical one.

Hierarchical methods include techniques that search clusters hierarchically and, for this, they admit that several clustering levels can be obtained. According to Diniz e Louzada-Neto (2000), hierarchical methods can be subdivided into dividing and agglomerative ones. The agglomerative hierarchical method first considers each pattern as a cluster and iteratively clusters the pair of clusters with greater similarity into a new cluster, until it forms one single cluster that contains all patterns. On the contrary, the dividing hierarchical method starts with one single cluster and runs a process with successive subdivisions.

The most usual way to represent a hierarchical cluster is through a dendrogram. A dendrogram represents the cluster of patterns and the similarity levels in which clusters are formed. According to Jain *et al.* (1999), a dendrogram can be "split" into different levels to show the different clusters. In the dendrogram showed in Figure 3, admitting a cut at the level shown in the figure, two clusters can be seen: the first one made up by patterns P1, P2 and P5, and the second one made up by patterns P3 and P4.

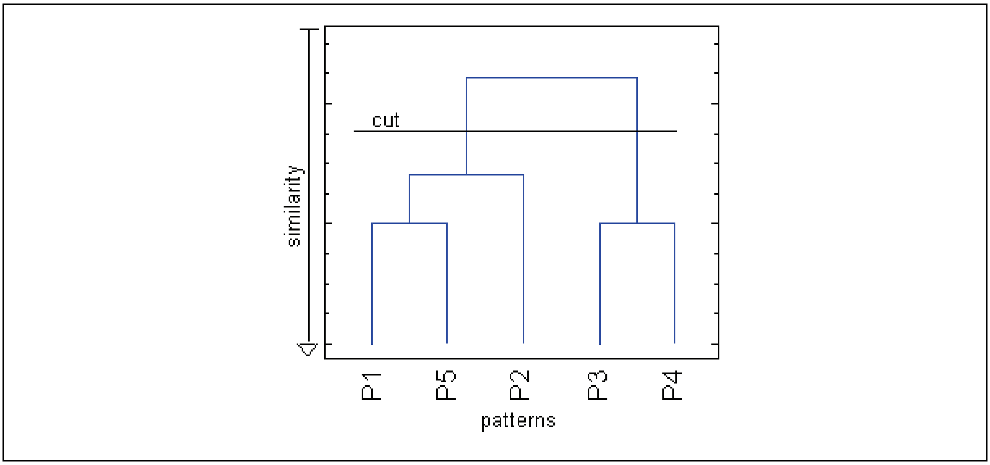


Fig. 3. Dendrogram example

5. The Itaipu Dam

According to ITAIPU (2008), ITAIPU Binacional, the largest hydroelectric power plant in the world, started to be built in 1973 at a part of the Paraná River known as ITAIPU, meaning "the singing rock" in tupy, and located at the heart of South America, at the border between Paraguay and Brazil. On 14th of November of 1978 were cast 7,207 m^3 of concrete, the equivalent to erecting 24 ten-story buildings in the same day, a Civil Engineering record in South America. In October 1982, the dam works were concluded and on the 5th of November, the same year, once the reservoir was filled, the presidents of Brazil, João Figueiredo, and of Paraguay, Alfredo Stroessner, put into operation the mechanism that automatically raises the spillway's 14 gatese, releasing the dammed waters of the Paraná River and officially inaugurating the largest hydroelectric power plant in the world.

Presently, ITAIPU has 20 generating units, with 700 MW (*megawatts*) each, adding up to a total installed capacity of 14,000 MW. In 2000, ITAIPU Binacional broke its own record of power production: approximately 93.4 billion kilowatts-hour (KWh) were generated that year. ITAIPU Binacional is responsible for supplying 95% of the power consumed in Paraguay and 24% of all the Brazilian demand.

The ITAIPU dam is 7,919 m long with a maximum height of 196 m, the equivalent to a 65-story building. It took 12.3 million m^3 of concrete and the iron and steel used to build it would be enough to build 380 Tower Eiffel, dimensions that have made the power plant a reference for studies in concrete and dam safety. Figure 4 shows ITAIPU dam's general structure and Table 1 shows the main features of the different stretches at the dam.

At the ITAIPU dam two segments are earth dams, one is a rockfill dam and another is made of concrete. Along it, and to follow-up the performance of concrete structures and foundations, there are 2,218 instruments (1,362 in concrete and 856 in the foundations and landfills). Of these, 210 are automated and 5,239 are drains (949 in the concrete and 4,290 in the foundations), and their readings occur at different frequencies, which may be, for instance, daily, weekly, biweekly or monthly, according to the type of instrument.

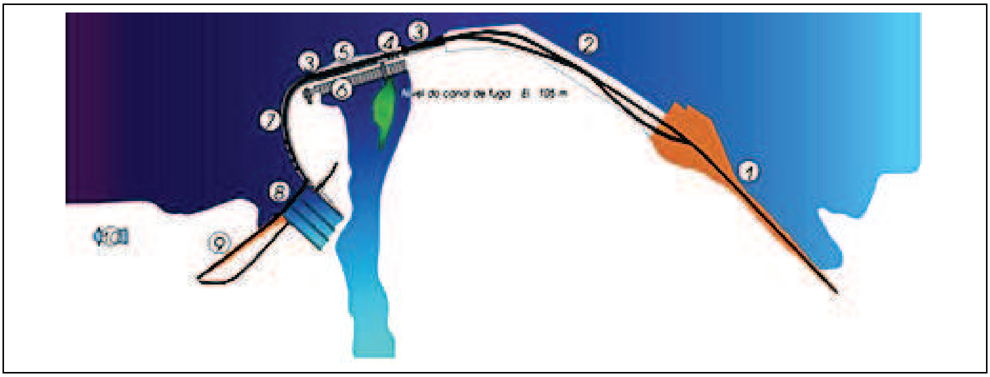


Fig. 4. General Structure of the ITAIPU complex (ITAIPU, 2008)

Stretch		Structure	Length (m)	Maximum Height (m)
1 (L)	Auxiliary Dam	Earth	2,294	30
2 (K)	Auxiliary Dam	Rockfill	1,984	70
3 (E e I) 7 (D)	Side Dams	Buttress	1,438	81
4 (H)	Deviation Structure	Solid Concrete	170	162
5 (F)	Main Dam	Hollow Gravity	612	196
9 (Q)	Auxiliary Dam	Earth	872	25
Others Stretches		Features		
8 (A)	Spillway	350 m wide		
6 (U)	Power House	20 Generating Units		

Table 1. Features of the stretches at the ITAIPU Dam

Although all nine stretches of the ITAIPU dam are instrumented and monitored, the Main Dam (stretch F) is a highlight and deserves a deeper study. Stretch F is where the power generating turbines are located and it is also where the highest water column and greater number of instruments are. Stretch F is constituted by several blocks and each one of them has instruments that supply data about their physical behavior, regarding their concrete structure and foundations.

In order to simplify, but without losing generality, once this same study may be carried out for the other instruments, the extensometer was the instrument that was chosen to apply the methodology in this work, because it is considered one of the most important instruments to monitor a dam and it is one of the instruments that ITAIPU's engineer team has automated. Measurements of settlements at a dam can be made with rod multiple extensometers (Figure 5) installed into probing holes. It is one of the most important observations to supervise the structure's behavior during the dam building, reservoir filling and operation periods. Installing extensometers upstream and downstream at the blocks where there are access galleries that are transversal to the axis allows, according to Silveira (2003), the measurement of angular displacements the dam may show close to the foundations.

By using the extensometers, it is possible to measure vertical displacements of the basaltic rock mass where dam foundation is based on. A typical geological profile of rock mass foundation can be observed in Figure 6. Settlement monitoring is very important and special attention is given to rock mass discontinuities, such as joints, faults and rock contacts. Each extensometer is installed at a specific location and can be composed by multiple rods of different lengths. Thus, it is possible to separately monitor the vertical displacement of each geological discontinuity.

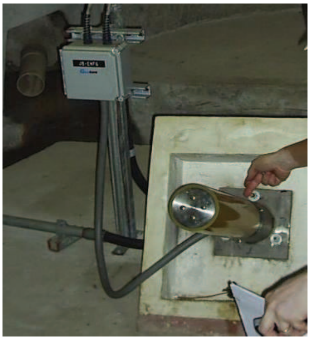


Fig. 5. Automated extensometer installed in Itaipu's rock foundation.

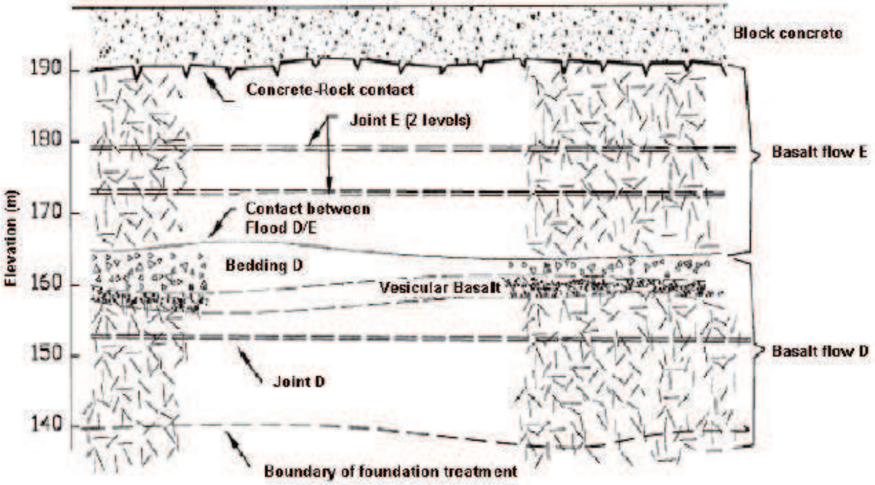


Fig. 6. Basaltic geological profile of Itaipu Dam rock foundation. (Adapted from OSAKO, 2002)

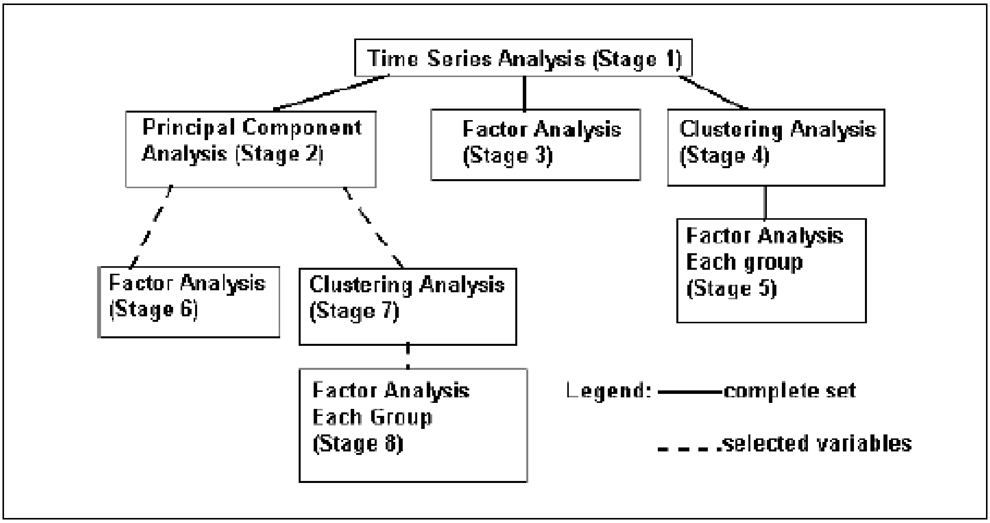
6. Methodology

The methodology that is presented in this work was applied to monitoring instruments at the ITAIPU dam, specifically at the dam's F stretch. The chosen instrument was the extensometer. Thirty extensometers are placed in the F stretch, each one bearing one, two or three rods, totaling 72 displacement measurements. These measurements will be identified as follows: equip4_1, means rod 1 of extensometer 4, for instance.

The data that will be used to develop this work will be monthly data, collected from January 1995 through December 2004, totaling 120 readings. This period was established after a suggestion of ITAIPU's engineering team, because the automatic data acquisition system was implanted after it. During this system's installation phase some instruments had no manual readings. Besides, the automated instruments suffered changes that may have influenced the later readings.

Most instruments render a monthly reading, but some of them render more then a reading per month and in these cases the monthly average was obtained. On the other hand some instruments had missing readings and in these cases interpolations were made by means of time series (stage 1, Flowchart 1, below), thus assuring that all instruments had exactly 120 readings.

Once the time series interpolations were made, were applied simultaneously: Principal Component Analysis (stage 2 – to select the extensometers' rods), Factor Analysis (stage 3 – to rank the extensometers' rods) and Clustering Analysis (stage 4 – to cluster the rods of similar extensometers). Factor Analysis was also applied to each cluster formed by the Clustering Analysis (stage 5). Steps 3, 4 and 5 were once again applied, considering only the extensometer rods selected in step 2, and these were called steps 6, 7 and 8, respectively. The several steps involved in producing this work are shown in Flowchart 1, below.


Time Series Analysis (Stage 1)
Principal Component Analysis (Stage 2)
Factor Analysis (Stage 3)
Clustering Analysis (Stage 4)
Factor Analysis Each group (Stage 5)
Factor Analysis (Stage 6)
Clustering Analysis (Stage 7)
Factor Analysis Each Group (Stage 8)
Legend: complete set
selected variables

Flowchart 1. Methodology application steps.

The Principal Component Analysis, as per Hair *et al*. (2005), was used to analyze the relationship between the variables of a set, by transforming the original set into a new one composed by non-correlated variables, called Principal Components, which have special properties in terms of variance. The Principal Components consist in linear combinations of the original variables and they were obtained in a descendant priority order. Most of the data variability can be explained by a small number of principal components.

The main objectives with the Principal Component Analysis were: to reduce the number of variables and indicate which variables, or variable sets, explain most part of the total variability, in order to reveal the type of relationship that exists between them. In the

Principal Components Analysis, it was possible to observe, for instance, that some components represent a non-significant part of the total variability (less than 1%) and that some variables are important (weights greater than 0.5 or smaller than –0.5) for these components. The most important variables that correspond to the least important components should not be selected.

Factor Analysis, as per Hair *et al.* (2005), was the technique that was used to explain the correlations between a large set of variables in terms of a set of few non-observable random variables that are called factors. Within the same cluster, variables can be highly correlated between one another and correlations can be only a few from one cluster to another. Each cluster represents a factor that is responsible for the observed correlations. Communality is the variable's variance part that is distributed throughout the factors.

There are several criteria to establish the number of factors. Kaiser's criterion is the mostly used one and it says that the number of factors should be equal to the number of eigenvalues that are greater than 1.

According to Johnson & Wichern (1998), by means of a rotation one may obtain a structure for the weights, such that each variable has a high weight in one single factor, and low or moderate weights in all the other factors. Kaiser suggested an analytical measure known as the Varimax criterion.

In Factor Analysis there are three types of variance: common, specific and of error. Common is the variance in a variable that is shared with all the other variables in the analysis. Specific variance is associated only with a specific variable. Error variance is the one that is due to the data clustering process' non-reliability, to a measuring error or to a random component in the measured phenomenon. Communalities are estimates of the common variance among the variables.

As communality is the part of the variable's variance that is ascribed to the factors and this represents a percentage of the variable's variation that is not random, the criterion to rank the extensometers' rods consists in sorting the extensometers' rods according to their communalities.

In the present work the method used for the Clustering Analysis was the Ward Method. The Ward Method is an agglomerative hierarchical method. According to Johnson e Wichern (1998), the Ward Method joins two clusters based on the "loss of information" criterion. This criterion can be represented by the Square Quadratic Error (SQE). For each cluster i are calculated the cluster's mean (or centroid) and square quadratic error (SQE_i). SQE_i is obtained through the sum of the square error of each variable in the cluster in relation to the cluster's mean. For k clusters there were SQE_1, SQE_2, ..., SQE_k, making possible the following definition:

$$SQE = SQE_1 + SQE_2 + \ldots + SQE_k \quad (1)$$

For each pair of clusters m and n, the mean of the new-formed cluster (cluster mn) was calculated. Then, the square quadratic error of cluster mn (SQE_{mn}) was calculated. The square quadratic error (SQE) could be recalculated by using:

$$SQE = SQE_1 + SQE_2 + \ldots + SQE_k - SQE_m - SQE_n + SQE_{mn} \quad (2)$$

Clusters m and n that resulted in the smallest SQE increase (i.e., the lowest "loss of information") could be merged. According to Hair Jr *et al.* (2005), this method tends to create clusters with the same size because of the minimization of the internal variation.

7. Results

As shown in the flowchart, for the 1st stage composed by Time Series, the model was automatically chosen according to the Akaike criterion (AIC) and also by observing the root mean squared error (RMSE). We observed the residual integrated periodogram and in some cases, after analyzing the p-values in the parameters "t" testing, the model was substituted by other considered more adequate.

Once the interpolations by Time Series were finished at the 2nd and 3rd steps, respectively, the Principal Component Analysis and the Factor Analysis were performed in order to select the most important extensometer rods among the 72.

The Principal Component Analysis (Stage 2) showed that 63 components represent a non-significant part of the total variability (less than 1%). Considering the extensometer rods that are important for these components ($- 0.5 \leq$ weight ≤ 0.5) nine extensometer rods are obtained: equip4_1, equip4_2, equip6_1, equip6_2, equip13_2, equip18_2, equip21_1, equip22_1 and equip29_1.

By means of this analysis the remaining 63 extensometer rods were considered important and should be used in steps 6, 7 and 8. This is not a good criterion to select the most important extensometer rods, once the number of selected rods is very big. However, this reduction of the number of rods may be interesting when other techniques are applied after that (such as in steps 6, 7 and 8).

In the Factor Analysis, stage 3, the variables with low communalities should have been discarded, but no variable had a communality that was smaller than 0.71. Communalities equal to 0.71 indicate that 71% of the variable's variance is distributed among the factors and only 29% is random. This meant that the corresponding instrument or rod was working well. Table 2 shows the 25 extensometer rods with the highest communalities.

Communality	Rod	Communality	Rod	Communality	Rod
0.988861	*equip29_1*	0.968213	equip23_2	0.957609	equip14_3
0.981763	*equip21_2*	0.968083	*equip4_1*	0.953036	equip25_1
0.976523	equip23_1	0.967029	*equip21_1*	0.950395	equip33_2
0.975655	*equip22_1*	0.966632	*equip4_2*	0.949394	equip24_2
0.972231	equip3_1	0.965999	*equip29_2*	0.949108	equip24_1
0.971971	*equip1_1*	0.965522	*equip34_3*	0.948646	equip5_1
0.971798	*equip22_3*	0.964925	*equip6_1*	0.943644	equip28_1
0.970804	equip11_1	0.963139	*equip6_2*		
0.970397	*equip1_2*	0.960121	*equip22_2*		

Table 2. 25 extensometer rods with the highest communalities

The dendrogram in Figure 7 shows how the Clustering Analysis formed the clusters (4th step). From the first cut, two clusters result. The first cluster, herein called Cluster A, is a cluster composed by extremely important instruments for dam monitoring. They are instruments installed at the block's axis, upstream the dam and inclined 60° upstream.

From the second cut we have how the two additional clusters were formed. The first one, called Cluster B, has most of the extensometer rods installed at Spillings B, C and D, and at Contacts B/C and C/D. The second cluster, called Cluster C, has most of the extensometer rods installed at Joints A and B, and at Contact A/B. Table 3 shows the extensometer rods in each cluster, the inclination, distance from the dam's axis and the attribute where the rod is installed.

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
A	EMF001/1	60° upstream	125.5 Meters upstream	Joint B
A	EMF001/2	60° upstream	105.4 Meters upstream	Contact B/C
A	EMF004/1	60° upstream	65.3 Meters upstream	Contact C/D
A	EMF004/2	60° upstream	60.4 Meters upstream	Fractured Rock
A	EMF006/1	60° upstream	150.8 Meters upstream	Joint A
A	EMF006/2	60° upstream	110.5 Meters upstream	Spilling B
A	EMF021/1	60° upstream	159.8 Meters upstream	Joint A
A	EMF021/2	60° upstream	135.1 Meters upstream	Spilling B
A	EMF026/1	60° upstream	139.2 Meters upstream	Joint B
A	EMF026/2	60° upstream	115.6 Meters upstream	Contact B/C
A	EMF031	60° upstream	64.7 Meters upstream	Contact C/D
B	EMF002/1	0	32.0 Meters upstream	Contact C/D
B	EMF002/2	0	32.0 Meters upstream	Fractured Rock
B	EMF003/1	0	32.0 Meters upstream	Spilling C
B	EMF003/2	0	32.0 Meters upstream	Spilling D
B	EMF005/1	0	13.0 Meters downstream	Contact C/D
B	EMF005/2	0	13.0 Meters downstream	Spilling D
B	EMF007/3	0	13.0 meters downstream	Spilling B
B	EMF008/2	0	84.0 meters upstream	Fractured Rock
B	EMF008/3	0	84.0 meters upstream	Spilling B
B	EMF012/1	60° downstream	47.2 meters downstream	Fractured Rock
B	EMF012/2	60° downstream	42.5 meters downstream	Dense Basalt
B	EMF013/2	0	44.0 meters downstream	Fractured Rock
B	EMF013/3	0	44.0 meters downstream	Spilling B
B	EMF014/2	0	54.0 meters downstream	Fractured Rock
B	EMF014/3	0	54.0 meters downstream	Spilling B
B	EMF015/1	0	80.0 meters upstream	Fractured Rock
B	EMF015/2	0	80.0 meters upstream	Spilling B
B	EMF018/3	0	33.0 meters downstream	Spilling B
B	EMF019/3	0	55.0 meters downstream	Spilling B
B	EMF020/2	0	82.0 meters upstream	Fractured Rock
B	EMF020/3	0	82.0 meters upstream	Spilling B
B	EMF023/3	0	36.0 meters downstream	Fractured Rock
B	EMF024/3	0	62.0 meters downstream	Spilling B
B	EMF025/2	0	75.0 meters upstream	Spilling B
B	EMF025/3	0	75.0 meters upstream	Spilling B
B	EMF027/1	30° upstream	16.6 meters downstream	Joint B
B	EMF027/2	30° upstream	22.6 meters downstream	Contact B/C
B	EMF029/2	30° downstream	55.7 meters downstream	Contact B/C
B	EMF032/1	30° upstream	36.5 meters upstream	Joint B

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
B	EMF032/2	30° upstream	14.6 meters upstream	Spilling C
B	EMF032/3	30° upstream	7.5 meters upstream	Contact C/D
B	EMF033/1	0	0.0	Joint B
B	EMF033/2	0	0.0	Spilling C
B	EMF033/3	0	0.0	Contact C/D
B	EMF034/3	30° downstream	7.5 meters downstream	Contact C/D
B	EMF035/1	90° upstream	0.0	Concrete
B	EMF035/2	90° upstream	0.0	Concrete
C	EMF007/1	0	13.0 meters downstream	Joint A
C	EMF007/2	0	13.0 meters downstream	Contact A/B
C	EMF008/1	0	84.0 meters upstream	Contact A/B
C	EMF011	0	81.0 meters upstream	Joint A
C	EMF013/1	0	44.0 meters downstream	Contact A/B
C	EMF014/1	0	54.0 meters downstream	Contact A/B
C	EMF018/1	0	33.0 meters downstream	Joint A
C	EMF018/2	0	33.0 meters downstream	Fractured Rock
C	EMF019/1	0	55.0 meters downstream	Joint A
C	EMF019/2	0	55.0 meters downstream	Fractured Rock
C	EMF020/1	0	82.0 meters upstream	Fractured Rock
C	EMF022/1	0	68.0 meters upstream	Joint A
C	EMF022/2	0	68.0 meters upstream	Fractured Rock
C	EMF022/3	0	68.0 meters upstream	Spilling B
C	EMF023/1	0	36.0 meters downstream	Joint A
C	EMF023/2	0	36.0 meters downstream	Fractured Rock
C	EMF024/1	0	62.0 meters downstream	Joint A
C	EMF024/2	0	62.0 meters downstream	Fractured Rock
C	EMF025/1	0	75.0 meters upstream	Joint A
C	EMF028/1	0	40.0 meters downstream	Joint B
C	EMF028/2	0	40.0 meters downstream	Contact B/C
C	EMF029/1	30° downstream	63.5 meters downstream	Joint B
C	EMF034/1	30° downstream	36.6 meters downstream	Joint B
C	EMF034/2	30° downstream	21.0 meters downstream	Spilling C

Table 3. Extensometer rods in each cluster

Observing the three clusters A, B and C obtained at the 4th step, at the 5th step Factor Analysis was applied within each cluster in order to rank the extensometer rods. Tables 4, 5 and 6 show the extensometer rods, and their communalities, for clusters A, B and C, respectively.

Considering the third cut, Cluster B was divided into two clusters called B1 and B2. The highlight is Cluster B2, which is formed mostly by extensometer rods installed at Spilling B. Cluster C was divided into two clusters called C1 and C2. Dyminski *et al*. (2008) introduces a methodology to identify the most important extensometer rods in these five clusters, by using Factor Analysis applied within each cluster. Visual Data Mining were used to examine relationships between extensometers in Silva Neto *et al*. (2008).

Communality	Rod	Communality	Rod
0.961839	*equip21_1*	0.881852	*equip26_1*
0.956979	*equip21_2*	0.854339	*equip6_2*
0.953791	*equip4_1*	0.809062	*equip1_2*
0.94278	*equip4_2*	0.798566	*equip26_2*
0.911982	*equip6_1*	0.677401	*equip31_1*
0.885099	*equip1_1*		

Table 4. Extensometer rods and their communalities – Cluster A

Communality	Rod	Communality	Rod	Communality	Rod
0.958451	equip3_1	0.899792	equip35_1	0.833353	equip19_3
0.957903	*equip29_2*	0.896858	equip2_2	0.832945	equip15_1
0.956985	*equip34_3*	0.895818	equip3_2	0.826735	equip18_3
0.949573	equip14_3	0.895145	equip8_2	0.818713	equip15_2
0.942628	equip33_3	0.894009	equip14_2	0.80722	equip33_1
0.938697	equip33_2	0.890046	equip23_3	0.77312	equip12_2
0.930857	*equip32_3*	0.888025	equip25_3	0.693537	*equip27_1*
0.92853	equip5_1	0.859488	equip5_2	0.692067	equip20_3
0.928364	*equip27_2*	0.854581	equip8_3	0.688618	equip25_2
0.925976	equip13_2	0.85328	*equip32_2*	0.635662	*equip32_1*
0.91628	equip20_2	0.847534	equip35_2	0.624787	equip7_3
0.905482	equip12_1	0.84403	equip2_1		
0.903051	equip13_3	0.843003	equip24_3		

Table 5. Extensometer rods and their communalities – Cluster B

Communality	Rod	Communality	Rod	Communality	Rod
0.975487	*equip29_1*	0.924851	*equip22_3*	0.783119	equip7_1
0.966626	equip23_1	0.917463	equip19_2	0.766859	*equip34_2*
0.952144	equip23_2	0.909212	equip11_1	0.732343	equip7_2
0.946772	equip24_1	0.899795	equip14_1	0.730472	equip18_1
0.945604	*equip22_1*	0.866861	equip28_2	0.680493	equip18_2
0.943369	equip24_2	0.853862	equip13_1	0.660694	equip28_1
0.942178	*equip34_1*	0.845538	equip25_1	0.647952	equip19_1
0.928596	*equip22_2*	0.812306	equip20_1	0.60656	equip8_1

Table 6. Extensometer rods and their communalities – Cluster C

Considering only the 63 rods selected by the Principal Component Analysis (2nd step), Factor Analysis and Clustering Analysis were preformed during the 6th and 7th steps, respectively.

Through the Factor Analysis that was applied to the 63 rods (6th step), it was observed that no extensometer rod showed a communality that was smaller than 0.7. A communality equal to 0.7 means that 70% of the rod's variance is ascribed to the factors and that only 30% of the variance is random, this is, the corresponding rods worked well. Table 7 shows the 25 extensometer rods with the highest communalities.

During the Clustering Analysis that was applied to the 63 rods (7th step) were considered the 3 clusters formed by the second cut, shown in Figure 8, through equivalence with the 4th step. Table 8 shows the extensometer rods in each cluster, the inclination, distance from the dam's axis and the attribute where the rod is installed.

Communality	Rods	Communality	Rods	Communality	Rods
0.974392	equip23_1	0.959987	*equip29_2*	0.939298	*equip34_1*
0.970051	*equip21_2*	0.956602	equip14_3	0.939156	equip33_3
0.96932	equip3_1	0.954664	equip23_2	0.938739	equip14_1
0.968294	*equip1_1*	0.952711	equip33_2	0.928829	*equip27_2*
0.968093	equip11_1	0.950157	equip24_1	0.928208	equip13_1
0.966944	*equip34_3*	0.947958	equip5_1	0.92547	*equip26_2*
0.96429	*equip1_2*	0.943773	equip19_2	0.92472	equip20_2
0.960271	*equip22_2*	0.941997	equip28_1		
0.96008	*equip22_3*	0.94105	equip24_2		

Table 7. 25 extensometer rods with yhe highest communalities – 63 rods

Cluster	Rod	Inclination	Distance from the dam's axis	Shape
A	EMF001/1	60° upstream	125.5 meters upstream	Joint B
A	EMF001/2	60° upstream	105.4 meters upstream	Contact B/C
A	EMF021/2	60° upstream	135.1 meters upstream	Spilling B
A	EMF026/1	60° upstream	139.2 meters upstream	Joint B
A	EMF026/2	60° upstream	115.6 meters upstream	Contact B/C
A	EMF031	60° upstream	64.7 meters upstream	Contact C/D
B	EMF002/1	0	32.0 meters upstream	Contact C/D
B	EMF002/2	0	32.0 meters upstream	Fractured Rock
B	EMF003/1	0	32.0 meters upstream	Spilling C
B	EMF003/2	0	32.0 meters upstream	Spilling D
B	EMF005/1	0	13.0 meters downstream	Contact C/D
B	EMF005/2	0	13.0 meters downstream	Spilling D
B	EMF007/3	0	13.0 meters downstream	Spilling B
B	EMF008/2	0	84.0 meters upstream	Fractured Rock
B	EMF008/3	0	84.0 meters upstream	Spilling B
B	EMF012/1	60° downstream	47.2 meters downstream	Fractured Rock
B	EMF012/2	60° downstream	42.5 meters downstream	Dense Basalt
B	EMF013/3	0	44.0 meters downstream	Spilling B
B	EMF014/2	0	54.0 meters downstream	Fractured Rock
B	EMF014/3	0	54.0 meters downstream	Spilling B
B	EMF015/1	0	80.0 meters upstream	Fractured Rock
B	EMF015/2	0	80.0 meters upstream	Spilling B
B	EMF018/3	0	33.0 meters downstream	Spilling B
B	EMF019/3	0	55.0 meters downstream	Spilling B
B	EMF020/2	0	82.0 meters upstream	Fractured Rock
B	EMF020/3	0	82.0 meters upstream	Spilling B
B	EMF023/3	0	36.0 meters downstream	Fractured Rock
B	EMF024/3	0	62.0 meters downstream	Spilling B

B	EMF025/2	0	75.0 meters upstream	Spilling B
B	EMF025/3	0	75.0 meters upstream	Spilling B
B	EMF027/1	30° upstream	16.6 meters downstream	Joint B
B	EMF027/2	30° upstream	22.6 meters downstream	Contact B/C
B	EMF029/2	30° downstream	55.7 meters downstream	Contact B/C
B	EMF032/1	30° upstream	36.5 meters upstream	Joint B
B	EMF032/2	30° upstream	14.6 meters upstream	Spilling C
B	EMF032/3	30° upstream	7.5 meters upstream	Contact C/D
B	EMF033/1	0	0.0	Joint B
B	EMF033/2	0	0.0	Spilling C
B	EMF033/3	0	0.0	Contact C/D
B	EMF034/3	30° downstream	7.5 meters downstream	Contact C/D
B	EMF035/1	90° upstream	0.0	Concrete
B	EMF035/2	90° upstream	0.0	Concrete
C	EMF007/1	0	13.0 meters downstream	Joint A
C	EMF007/2	0	13.0 meters downstream	Contact A/B
C	EMF008/1	0	84.0 meters upstream	Contact A/B
C	EMF011	0	81.0 meters upstream	Joint A
C	EMF013/1	0	44.0 meters downstream	Contact A/B
C	EMF014/1	0	54.0 meters downstream	Contact A/B
C	EMF018/1	0	33.0 meters downstream	Joint A
C	EMF019/1	0	55.0 meters downstream	Joint A
C	EMF019/2	0	55.0 meters downstream	Fractured Rock
C	EMF020/1	0	82.0 meters upstream	Fractured Rock
C	EMF022/2	0	68.0 meters upstream	Fractured Rock
C	EMF022/3	0	68.0 meters upstream	Spilling B
C	EMF023/1	0	36.0 meters downstream	Joint A
C	EMF023/2	0	36.0 meters downstream	Fractured Rock
C	EMF024/1	0	62.0 meters downstream	Joint A
C	EMF024/2	0	62.0 meters downstream	Fractured Rock
C	EMF025/1	0	75.0 meters upstream	Joint A
C	EMF028/1	0	40.0 meters downstream	Joint B
C	EMF028/2	0	40.0 meters downstream	Contact B/C
C	EMF034/1	30° downstream	36.6 meters downstream	Joint B
C	EMF034/2	30° downstream	21.0 meters downstream	Spilling C

Table 8. Extensometer rods in each cluster – 63 rods

Observing the three clusters obtained at the 7th step, at the 8th step Factor Analysis was applied within each cluster in order to rank the extensometer rods. Tables 9, 10 and 11 show the extensometer rods, and their communalities, for clusters 1, 2 and 3, respectively.

8. Conclusions

This work presents a methodology that can be included in the KDD area. The purpose in selecting, clustering and ranking the extensometers' rods is to be able to maximize the efficacy and efficiency of readings analyses, by identifying similar extensometer rods, as well as the main instruments.

In this work the methodology is applied to only one instrument: the extensometers located at the dam's Stretch F. There are a total of 30 extensometers with one, two or three rods, located at different points of Stretch F, totaling 72 displacement measurements. It is worth pointing out that from the 72 measurements, the company has automated 24.

During the Principal Component Analysis (2nd step), from the nine extensometer rods important for insignificant components, this is, they would not be considered in further analyses, seven are common with the ones Itaipu has automated. On the other hand, of the remaining 63 extensometer rods that were considered important, 17 are common with the ones Itaipu has automated. This is not a good criterion to select the important extensometer rods, once the number of selected rods is very big. However, this reduction of the number of rods may be interesting when other techniques are applied after that.

During the Factor Analysis (3rd step) it was observed that the extensometer rods worked well. Table 2 presents the 25 extensometer rods with the highest communalities. The 14 rods Itaipu's Engineering team automated are highlighted in italic.

During the Clustering Analysis (4th step) it was evidenced that it is possible to discover technical justifications for the cluster formation.

Observing the three clusters A, B and C obtained at the 4th step, at the 5th step Factor Analysis was applied within each cluster in order to rank the extensometer rods. The extensometer rods marked in italic in Tables 4, 5 and 6 are those Itaipu has automated. One can notice that the automated extensometer rods are, in most times, among the first of each cluster's ranking.

During the Factor Analysis applied to the 63 rods (6th step) it was observed that the extensometer rods worked well. Table 7 shows the 25 extensometer rods with the highest communalities. Highlighted in italic are the 10 rods Itaipu's Engineering team has automated, a figure that is smaller than the one found during the 3rd step.

Observing the three clusters obtained at the 7th step, at the 8th step Factor Analysis was applied within each cluster in order to rank the extensometer rods. The extensometer rods marked in italic in Tables 9, 10 and 11 are those Itaipu has automated. One can notice that the automated extensometer rods are, in most times, among the first of each cluster's ranking. The Principal Component Analysis (2nd step), however, excluded some of extensometer rods from this analysis. Therefore, the number of rods Itaipu's Engineering team has automated is smaller in this step than in the 5th step.

The results of steps 6 and 8 show the reduction of the number of rods that resulted from the Principal Component Analysis was not favorable when other techniques were applied after it. This happened because of the automated extensometer rods that were excluded.

9. Acknowledgements

The authors would like to thank FINEP for financially supporting the research project "AIEVC -"Uncertainty Analysis and Estimation of Control Values for the Geotechnical-structural Monitoring System at the Itaipu Dam", to CAPES for the author's first scholarship and to Itaipu's Civil Engineering team for instrumentation data and technical contributions.

Communality	Rods	Communality	Rods
0.949996	*equip1_1*	0.826814	*equip21_2*
0.901233	*equip1_2*	0.486577	*equip31_1*
0.845959	*equip26_2*	0.308738	*equip26_1*

Table 9. Extensometer rods and their communalities – Cluster 1

Communality	Rods	Communality	Rods	Communality	Rods
0.958367	equip3_1	0.89854	equip2_2	0.844444	equip2_1
0.957442	*equip29_2*	0.897657	equip14_2	0.833835	equip15_1
0.956486	*equip34_3*	0.895226	equip3_2	0.830048	equip19_3
0.948978	equip14_3	0.894602	equip8_2	0.827687	equip18_3
0.943927	equip33_3	0.89268	equip23_3	0.824894	equip15_2
0.937446	equip33_2	0.887947	equip25_3	0.806434	equip33_1
0.934316	*equip32_3*	0.877563	equip13_3	0.77671	equip12_2
0.929148	*equip27_2*	0.859427	equip5_2	0.702823	*equip27_1*
0.928345	equip5_1	0.855579	*equip32_2*	0.69919	equip20_3
0.916552	equip20_2	0.852541	equip8_3	0.687451	equip25_2
0.905877	equip12_1	0.849741	equip24_3	0.636324	*equip32_1*
0.899833	equip35_1	0.84772	equip35_2	0.625665	equip7_3

Table 10. Extensometer rods and their communalities – Cluster 2

Communality	Rods	Communality	Rods	Communality	Rods
0.965763	equip23_1	0.915313	*equip22_3*	0.791565	*equip34_2*
0.949473	equip23_2	0.911498	equip11_1	0.791028	equip7_1
0.945071	equip24_1	0.899258	equip14_1	0.755515	equip19_1
0.939808	*equip34_1*	0.869832	equip28_2	0.733123	equip7_2
0.939418	equip24_2	0.849429	equip13_1	0.70484	equip18_1
0.93064	equip19_2	0.83898	equip20_1	0.676714	equip28_1
0.921589	*equip22_2*	0.833704	equip25_1	0.594188	equip8_1

Table 11. Extensometer rods and their communalities – Cluster 3

10. References

Bowles, D.S., Anderson, L.R., Glover, T.F. e Chuhan, S.S. (2003) *Dam Safety Decision-Making: Combining Engineering Assessments With Risk Information*, Proc. of 2003 US Society on Dams Annual Lecture

CBGB (1983)- *Comitê Brasileiro de Grandes Barragens*. Diretrizes para a inspeção e avaliação de segurança de barragens em operação. Rio de Janeiro-Brasil

Dibiagio, E. (2000) *Question 78 - Monitoring of Dams and Their Foundations – General Report*. Proc. Of Twentieth Congress on Large Dams, ICOLD, 1459-1545, Beijing

Diniz, C. A. R. e Louzarda Neto, F. (2000) Data mining: uma introdução. ABE. São Paulo-Brasil

Duarte, J. M. G., Calcina, A. M. e Galván, V. R. (2006), *Instrumentação Geotécnica de Obras Hidrelétricas Brasileiras: Alguns Casos Práticos Atuais*. In: COBRAMSEG' 2006, Curitiba-Brasil

Dyminski, A. S., Steiner, M. T. A. e Villwock, R. (2008) *Hierarchical Ordering of Extensometers Readings from Itaipu Dam*. In: First International Symposium on Life-Cycle Civil Engineering -IALCCE' 8, Varenna – Italia

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. e Uthrusamy, R. (1996) *Advances in knowledge Discovery & Data Mining*. AAAI_MIT

FEMA – Federal Emergency Management Agency. (2004) *Federal Guidelines For Dam Safety*, U. S. Department Of Homeland Security, USA

Freitas, A. A. (2002) *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer. New York

Hair Jr, J.F., Anderson, R.E., Tatham, R.L. e Black, W.C. (2005) *Análise Multivariada de Dados (tradução)*. Bookman. São Paulo-Brasil

Harrald, J. R.; Renda-Tanali, I.; Shaw, G.L.; Rubin, C.B.; Yeletaysi, S. (2004) *Review of Risk Based Prioritization/Decision Making Methodologies for Dams*. Technical Report, US Army Corps of Engineers

ICOLD - International Commission on Large Dams. (2008) http://www.icold-cigb.org

ITAIPU. ITAIPU Binacional. (2008) http://www.itaipu.gov.br

Jain, A. K., Murty, M. N. e Flynn, P. J. (1999) *Data clustering: a review*. ACM Computing Surveys

Johnson, R.A. e Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis*. 4nd. Edition. Ed. Prentice Hall

Osako, C. I.. (2002) *A Manutenção dos Drenos nas Fundações de Barragens - O Caso da Usina Hidrelétrica de Itaipu*. Dissertação de Mestrado do Programa de Pós-Graduação em Construção Civil - PPGCC, UFPR, Curitiba-Brasil

Silva Neto, M. A., Villwock, R., Steiner, M. T. A., Dyminski, A. S. e Sccheer, S. (2008) *Mineração Visual de Dados Aplicada à Extração do Conhecimento nos Dados de Instrumentação da Barragem de Itaipu*. In: XL Simpósio Brasileiro de Pesquisa Operacional - SBPO, João Pessoa - Brasil.

Silveira, J. F. A. (2003) *Instrumentação e Comportamento de Fundações de Barragens de Concreto*. Oficina de Textos. São Paulo-Brasil

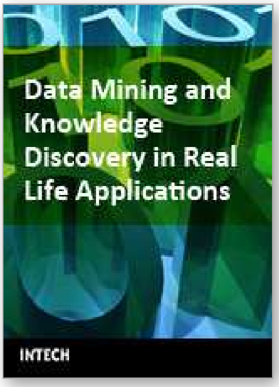
Silver, D.L. (1996) *Knowledge Discovery and Data Mining*. Technical Report MBA6522 CogNova Technologies London Health Science Center

Tan, P. N., Steinbach, M. e Kumar, V. (2005) *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co. Inc. Boston, MA, USA

U.S. Army Corps of Engineers. (1987) *Instrumentation for Concrete Structures*. Engineering and Design. Engineer Manual N^o No. 1110-2-4300, Washington, DC

U.S. Army Corps of Engineers. (1995) *Instrumentation of embankment dams and levees*. Engineering and Design. Engineer Manual N^o 1110-2-1908, Washington, DC

Witten, I. H. e Frank, E. (2000) *Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers. San Francisco, California



Data Mining and Knowledge Discovery in Real Life Applications
Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821