International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

The tasks of pre and post-processing in Data Mining applied to a real world problem

José Luis Díaz[1], Manuel Herrera[2], Joaquín Izquierdo[2], Rafael Pérez-García[2]
[1]Departamento de Ingeniería Hidráulica y Medio Ambiente
[2]Instituto de Matemática Multidisciplinar
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia, Spain
(jldiaz,mahefe,jizquier,rperez@gmmf.upv.es)

Abstract: Pre and post-processing are crucial tasks in Knowledge Discovery in Databases (KDD). In this contribution we present an application to a data set from a real water supply network (WSN) in the town of Calarcá (Colombia), located in the so-called "Eje Cafetero" coffee region. We use traditional and well-known techniques of pre and post-processing with the aim of showing its importance in Data Mining (DM), and of enhancing the need of results interpretability when dealing with real data set. Pre and post-processing tools, as well as other DM tasks implemented in Clementine 9.0 (SPSS), have been used. Clementine 9.0 has a number of pre and post-processing tools to work with records (rows) and fields (columns) in a database. Basically, we used *selection* and *deriving* operations for records, and *type and filter* operations for fields. The database consists of a record of requests, complains and claims (PQRs in Spanish), for the year 2006, remitted to the Calarcá Water Supply Company *Multipropósito, S.A. ESP*. Additionally, the database is also integrated by the network hydraulic model, some climatic variables, and thematic maps of vulnerabilities and risk areas for natural phenomena. The PQRs information consists of 846 records. First, the consistency of the PQRs was evaluated to determine outliers, and lost or missing information. Next, each point was located on the map of the town and its UTM coordinates were obtained. Then, each PQR was associated to its nearest pipe and node of the primary network. The graphical classification of variables shows trends that permit us to obtain *a priori* conclusions in KDD. These data were used to feed the model and to obtain relationships between different variables and the damage type on the network well within the post-processing task.

Keywords: Pre-processing; Post-Processing; Knowledge Discovery; Data Mining; Water Supply Networks.

1. INTRODUCTION

KDD corresponds to advanced data management processes that in the last few years have become very attractive for both water researchers in engineering and water supply managers. Unfortunately, the use of KDD in the water industry is not widespread yet. The steps in KDD processes are: domain understanding, data collection, data pre-processing, data mining, and post-processing of the derived knowledge. The data that are to be processed by a knowledge acquisition algorithm are usually noisy, incomplete, and often inconsistent. Many steps must be performed prior to data analysis itself. For this reason, pre-processing procedures are continuously developing. Also, a result from a machine learning algorithm, such as a decision tree, a set of decision rules, or an artificial neural network, may not be appropriate from the viewpoint of customers or for commercial applications. As a result, a conceptual description (model, knowledge base) produced by such an inductive process has to be usually post-processed. Post-processing procedures usually include varied pruning routines, rule filtering, or even knowledge integration, Bruha and Famili [2000]. Zhang *et al.* [2003] argue for the importance of data preparation because of three aspects: real world data is impure; high-performance mining systems require quality data; and quality data yields high-quality patterns. Real data can disguise useful patterns due to incomplete data, noisy data and inconsistent data. Data preparation generates new datasets that are smaller than the original one by selecting relevant data, recovering incomplete data, purifying data, reducing data, and resolving data conflicts.

The paper is organized as follows: sections 2 and 3 give a brief introduction to pre and post-processing, respectively; section 4 provides a description of the used data; finally, section 5 explains the performed tasks of pre and post-processing, and presents the results of the found relationships among different variables for the database used.

2. PREPROCESSING

This task usually takes substantial project time (between 70% and 80%), especially when many aggregations are required, Feelders *et al*. [2000]. The operations performed in a pre-processing process can be reduced to two main families of techniques, Gibert *et al.* [2008]: Detection Techniques (DT) to detect imperfections in data sets and Transforming Techniques (TT) oriented to obtain more manageable data sets. DT includes outlier's detection, missing data detection, influent observations detection, normality assessment, linearity assessment, and independence assessment. On the other hand, TT includes outlier treatment, missing data imputation, dimensionality reduction techniques or data projection techniques, deriving new attributes techniques, filtering and resampling. Additionally, the statistical technique of data cleaning, and the visualization techniques also play an important role in the pre-processing of data. *Clementine 9.0* (Clementine®,2004) has some data pre-processing tools to accomplish these tasks. Its tools include management of records (rows) and handling of fields or attributes (columns), as well as graphical representation tools in order to be able to preview the available information. These tools include selecting, sampling, merging, sorting, aggregating, filtering, deriving and typing.

3. POST-PROCESSING

KDD post-processing components can be categorized into the following groups, Buha and Famili [2000]: knowledge filtering; interpretation and explanation; evaluation; and knowledge integration. In the case of machine learning algorithms such as trees or decision rules trained with noisy data, the results are generated covering few training data. This is because the induction algorithms try to subdivide the training data set. To overcome this problem the decision trees or rules should be shrunk, by either post-pruning (decision trees) or truncation (decision rules). After obtaining new knowledge, this can be either implemented in an expert system or used by an end user. In this last case, the knowledge results should be documented for the end user interpretation. Another possibility is to display the knowledge and transform it into a form understandable to the end user. We can also check the new knowledge for potential conflicts with previously induced knowledge. In this step, we can also summarize the rules and combine them with a domain-specific knowledge provided for the given task. Once a learning system has induced concept hypotheses (models) from the training set, their evaluation (or testing) should take place. There are several criteria used for this purpose: classification accuracy, understanding, computational complexity, and so on.

4. THE DATABASE

The Calarcá water company provided the information used in this paper. The WSN serves the urban area of the municipality's 14229 equivalent users. The database consists of a record of requests, complains and claims (PQRs in Spanish) reported by users of Multipropósito, S.A. ESP (Calarcá Water Supply Company), 24 hours a day for the year 2006. Calarcá town is located in the Andean area of Colombia; it has an estimated population of 73500 inhabitants enjoying a varied climate, with temperatures ranging in accordance with the altitude, between 22°C and 4°C, depending on the influence of solar radiation and rainfall. Its average temperature is 20°C. The rainfall varies between 1700 and 2400 millimetres annually. The relative humidity, high and stable, is approximately 85%.

These PQRs are reports both in principal and domiciliary network sections. The PQRs indicate such information as type or description of damage, location, and relevant technical concepts and solutions. Every record was located in a map of the WSN according to the address (street) referenced in the PQR record; addresses were only in terms of street names and numbers, so a hard work had to be done for obtaining the UTM coordinates of every reported problem. Additionally, the database integrates the network hydraulic model for this year, and risks level and vulnerabilities by geologic hazards information represented in various maps. Moreover, climatic data for stations nearby were obtained. The database contains 846 records (rows) and 218 fields (attributes).

As said, the water company receives reports of PQRs 24 hours a day. A total of 846 PQRs were reported for the year 2006. In each report the following is registered: the date of damage notice; the person name submitting the PQR; the location of incident; the PQR description given by the reporter; the company technical assessment; the solution adopted; and the date and name of the officer providing the solution. This information was typed into a database for further processing.

The main pre-processing task involved decision about relevant and non-relevant fields in the database. This decision was made based on hydraulic criteria (domain knowledge). Since damage causes were the main objective, geographical locations suggesting causes and occurrence of water loses were devised of paramount importance. As a consequence, the used information included: pipe identification – to assess if it was subjected to a high or a low number of faults; upstream and downstream node identification – to evaluate concurrence of faulty pipes at the same spot pointing towards local pressure or demand problems; type of reported breakage, either domiciliary or on the main network; pipe diameter; pipe length; pipe material; roughness material; and magnitude of the leak. In addition, data obtained from the mathematical model of the network were included in the database, specifically, information related with patterns and demand characteristics. Finally, the UTM coordinates of the faulty points were also included. As a matter of fact, typical pre-processing tasks for identifying outliers, missing values, etc., were performed.

5. PRE AND POST-PROCESSING IN A REAL DATA SET

Our main objective was to discover damage management models in the WSN. To this purpose, relationships among a variety of variables and reported damage types in the network were investigated. Both Classification and Regression Trees (C&RT) (Breiman *et al*, 1984) and Artificial Neural Networks (ANN) algorithms for DM tasks were used. As said, the information contained in each of the 846 PQRs records is: claim date; claiming user name; damage location; damage description (user-supplied); specialist technical concept; solution; solution date; and the workers names (crew) that performed the repair. Also, hourly data flows and pressures in the network, diameters, materials, lengths, and roughness were obtained from the hydraulic model. The used climatic data correspond to rain (mm), humidity (%), and sunshine (hr), for some stations near the town. Finally, the thematic maps were used to obtain risk levels zoning (high, medium, low), and vulnerabilities in each area (earthquakes and/or landslides).

To begin with, the PQRs reports were typed up to a database. A first pre-processing task was to dismiss spurious and erroneous information; this task was carried out manually. Only 4% of the total data are missing and, specifically, they correspond to diameter and material type in the PQRs reports. Due to the difficulty to obtain these data, we decided not to take them into account. One of the C&RT algorithms advantages is their robustness to missing information. On the other hand, missing data are scattered throughout the study area and therefore can be assumed random. To spatially locate the PQRs, information was typed according to the reported address (street) on a digital map in Autocad format. Here, inconsistencies in terms of points clearly not identified or whose location PQRs report did not appear in the digital base were corrected. Based on the location of these points we obtained UTM coordinates for each PQR with its own identifier. This information was added to the database.

Other pre-processing tasks involved handling of the attribute *Date* (report day of the damage) and *Repair Date* (repair day), including the definition of a new field with the name *Repair Time* indicating the repair delay; these tasks were performed with node called *derive* of Clementine 9.0. *Damage Type* variable was harder to manage because it depends on each worker description without standard formats of damage types. Therefore, to facilitate the treatment one set of damages based on the reports were also numerized with the derive node. Also, a Boolean variable to indicate whether the reported damage was fixed or not was introduced (there are various reasons for not repairing a damage, such as not having authorization from the connection owner, among others). The field *Worker* was considered interesting as it might look for relationships between damages reported and the workers engaged in repairing tasks. The field was also numerized based on work crews according to PQRs (node derive). The variable *Diameter* was converted to mm to make comparisons between the reported diameter in a PQR and the diameter within the hydraulic model. To better understand and comprehend the available information, use was made, as a pre-processing tool, of visualization techniques through graphical representations, as shown below.

Figure 1 visualizes the number of PQRs per month for 2006. The highest number of incidents takes place in May, while in November and December fewer warnings appear.

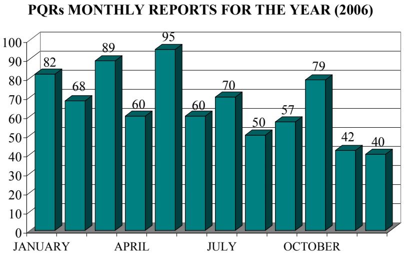

Figure 1. PQRs reports for 2006

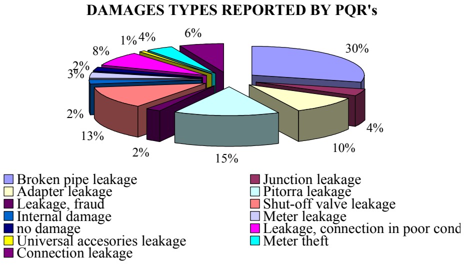

Figure 2. Damage type reported

Regarding damage types, 30% correspond to breaks in pipes followed by *pitorra* (a type of joint) leakage with 15%, and shut-off valve leakage with a 13% of reports, as visualized in Figure 2. Leakage due to fraud represents a small percentage (2%). The percentage of internal damage is also low, corresponding to 2% of PQRs. The supply system has deficiencies in terms of pipe age that had outlived their useful life, and problems of insufficient water for certain diameters, what could affect the amount of leakage present in the reports. Also, a greater number of incidents occur in plastic, metal or asbestos cement pipes, the latter two being older than the first.

Figure 3 visualizes the distribution per month of reported damage types. It can be observed that in January there is a leakage increase due to broken pipes, despite this being a month that corresponds to a holiday period. Also, with independence of the adopted numerization, water losses in the connections appear in a generalized way; this is a factor that directly influences both technical and economic management system, and that, compared with detected fraud by illegal extraction of fluid, affects in greater proportion the leaks caused to the system. Also, the proportion of reported internal damages is smaller, in general, than the damage reported in the connections. In the months of June and July a greater variety of damage types can be observed. Also, leakage due to broken pipes is reduced. Regarding claim distribution time (Figure 4), most of the claims appear in the morning between six and noon.

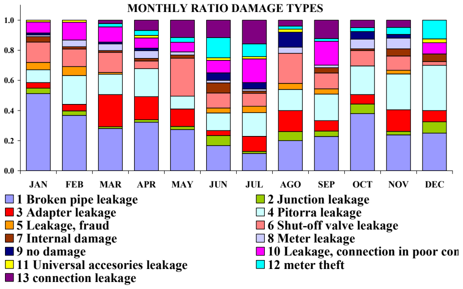

Figure 3. Per month reported damage relation

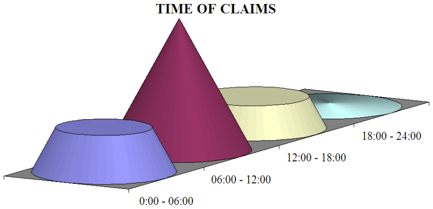

Figure 4. Daily claim distribution

Most of the PQRs, 799, correspond to the domiciliary network, against 32 for the main network (Figure 5). The greater percentage of network failures by diameter correspond to those of half-inch (12.7 mm) with a 60% of reports, followed by pipes of 16 mm with 20.6% of reports (Figure 6). Concerning materials (not always this field of the form was filled), 41.7% of the reports correspond to pipes of PVC, followed by 37.4% of pipes of polythene, 20.4% of galvanized iron pipes, and 0.4% of cast iron pipes, as can be observed in Figure 7.

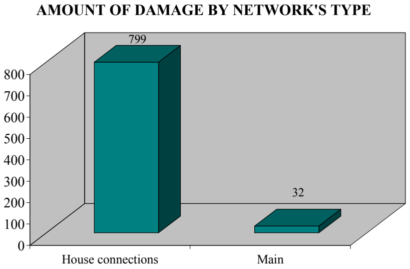

Figure 5. Network damage

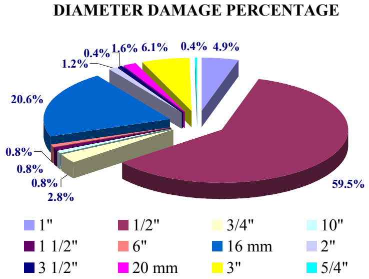

Figure 6. Damage distribution by diameter

The fact that the domiciliary network is more affected than the main pipes indicates more intensive management and maintenance in the primary network; although more complete information would be required to corroborate this affirmation. However, the distribution of diameters and materials for the reported damages of the network suggests that better maintenance in the domiciliary networks, replacement of old pipes, or improvement of installation techniques, could help reduce the number of reports.

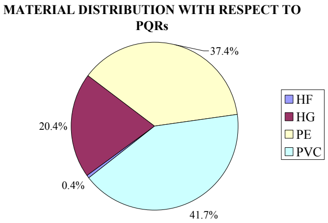

Figure 7. Material distribution

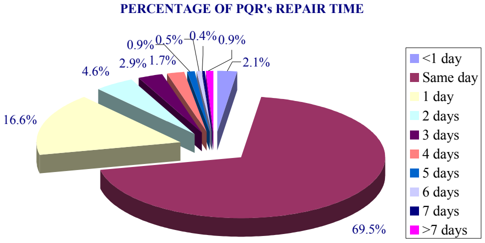

Figure 8. Repair times

In relation to report management, 69.5% of the PQRs are addressed and solved the same day, and 16.6% with a delay of one day (Figure 8); we have to mention in this regard that repair data with smaller values than one day are considered erroneous. A total of 768 reports were solved (Figure 9). Figure 10 shows the work crews conformation (company field workers), which is one of the variables used for the analysis.

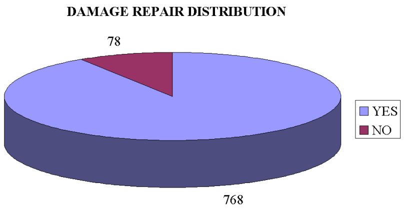

Figure 9. Damage repair

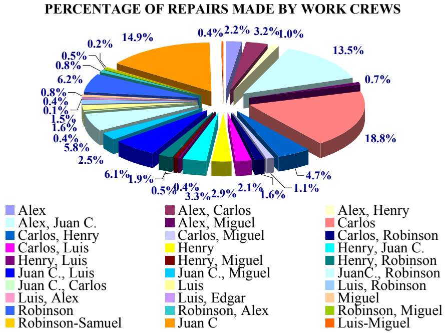

Figure 10. Work crews

Regarding the risk level from natural phenomena, 0.71% of PQRs correspond to zones with high level, 38.77% to mid level, and 52.72% of the data are located in zones of low risk; 7.8% of the data are located in no-risk zones. Considering threats, 0.71% of the registers of PQRs are located in earthquake and sliding threat zones and 38.77% in zones of earthquake risk. The remainder, 60.52%, are located in zones of no natural threat. The seismic risk for the majority of the WSN represents extra problems of management for the company.

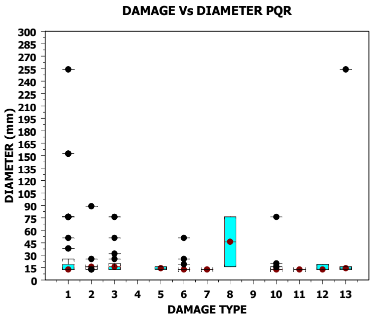

Figure 11. Damage vs diameter in PQRs

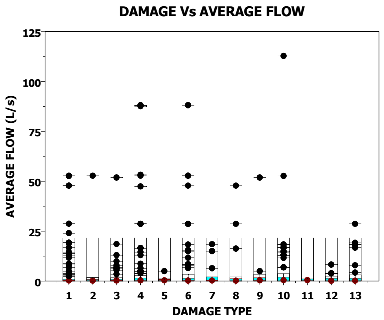

Figure 12. Damage vs average flow

Figures 11 and 12 visualize the relationship between different damage types (numerized), and diameter in PQRs as well as average flow in network sections on which a PQR report is located, respectively. As can be seen, there is no clear relationship between damage and these variables except for damage type 8 (leakage meter) where a uniform distribution by diameters is observed.

As can be appreciated, a basic characterization of the selected attributes can be obtained with the pre-processing performed by using visualization techniques. Also, first indications towards where the tasks of data mining should be directed are suggested. As mentioned previously, the aim of this work was to find relationships between the different variables and the damage type in order to find a damage management model for the WSN. To this end, various types of ANN and C&RT were trained based on the algorithms implemented in Clementine. Given that the used information corresponds to inherent variables for the operation of water supply network, previous variable selection was not deemed necessary. However, to get a more explanatory power, Kohonen network algorithms were used for a representative selection of attributes when significant differences between the variables were not found. Additionally, C&RT algorithms manage to find the best significant attributes for the database. For ANN algorithms Clementine 9.0 introduces a sensibility analysis for the input fields, once trained the network, which provides information about the input fields that are more decisive for the output field's prediction. Likewise, this analysis didn't find significant differences between the input data for de database used. The best results were obtained with the C&RT algorithms. Data were divided into separate training and test sets. Various models of C&RT were tested and various kinds of ANNs were trained (Table 1). In general, the data partition was 70% for training and 30% for testing; but in other cases, it was 60% for training, 30% for testing, and 10% for validation.

Model	Algorithm	Training Time			Correct Classifications (%)			Neurons		
		Hr	Min	Sec	Train.	Test.	Valid.	Input	Hidden	Output
Tree	C&RT	0	29	4	52.6	30.12				
ANN	RBFN	14	8	44	29.82	26.10		129	40	4
Tree	C&RT	0	24	46	55.44	29.32				
ANN	RBFN	24	32	13	31.32	26.10		122	40	4
Tree	C&RT	0	25	55	51.89	25.29	24.42			
ANN	RBFN	20	10	39	26.64	24.51	18.60	122	40	4
ANN	PRUNE	0	11	39	19.77	18.88		179	20-15-10	4
ANN	RBFN	0	9	42	29.15	32.13		210	20	4
ANN	Exhaustive	0	17	16	23.79	30.12		210	30-20	4
ANN	RBFN	0	19	23	28.48	29.72		210	20	4
ANN	RBFN	7	47	22	36.01	29.72		210	200	4
ANN	Multiple	0	12	18	27.81	31.73		210	2	4
ANN	Dynamic	0	8	32	27.81	31.73		210	3-4	4
ANN	Fast	0	12	4	27.81	31.73		210	20-15-10	4
ANN	RBFN	66	11	55	26.97	24.90		210	20	4
Tree	C&RT	0	35	38	57.12	28.11				
ANN	RBFN	165	36	32	28.14	26.91		89	40	4
Tree	C&RT	0	12	55	52.60	30.12				
ANN	RBFN	6	2	13	35.68	28.51		89	200	4
ANN	RBFN	22	31	12	36.68	30.92		89	200	4

Table 1. Training models

Table 1 shows different configurations of trees and ANN models. The tree models generally get better solutions than the ANN models in terms of data correctly classified, and also in the algorithm execution time. One of the greater benefits and strengths of C&RT algorithms is their final presentation and interpretability, either by rules or by the tree-shaped structures that help decision-making, an issue of great importance in the post-processing of the discovered knowledge. Pruning was used as a post-processing task in the case of trees to avoid model overfitting. Also the settings implemented in Clementine 9.0 for trees algorithms were considered: for handling missing data (maximum number of substitutes); for impurity management within the tree divisions, i.e. the creation of subgroups with similar output values (a node is considered pure if 100% of the node correspond to a specific category of field objective). Therefore, if the best division of a branch does not reduce the impurity to a specified value, this division is not performed. In the case of ANN models, overtraining was avoided by dividing the data randomly into training and testing independent sets. The network is trained by using the training set and the accuracy is estimated based on the test set. Additionally, one can choose whether to continue training the existing model or to start a new one. This is due to the fact that each time the algorithm is run a completely new network is created by default. It is also possible to manage the network architecture by choosing the number of hidden layers and neurons in each layer.

The set of rules in Table 2, which corresponds to a tree structure, shows one of the models for the damage type classification occurred in the network according to the selected variables. These variables correspond to the material reported in the PQRs (cast iron, galvanized iron, asbestos cement, high density polyethylene, flexible plastic, and PVC), binary variable damage repair (yes/no), numerized work crews, PQR point location using the distance to the nearest node, the average head loss (m/km) (obtained from the hydraulic model), the precipitation (mm) presented in nearby rainfall stations, water demand in nodes (L/S), PQR Cartesian coordinates, pressure in nodes (m) and section lengths (m).

The setting for model training was the following: the data were divided into partitions; a number of levels below the root of 100 were chosen; a maximum of five substitutes were chosen and 0.00000001 for minimum impurity change (the diversity Gini index was chosen as impurity measure for categorical objectives and for continuous objects was used the minimum square deviation as a impurity measure); tree pruning was used; record number in the parent branch was 2% and minimum number of records in the subsidiary branch was 1%, as stopping criterion. Also, as said, the data partition was 70% for training and 30% for testing. This tree classified 52.60% (314 records) of the training data and 30.12% (75 records) of the test data correctly. At the end of each line the branch summary is presented. This corresponds to statistical mode (most frequent value); number of records to which the rule applies (occurrences); and proportion of records for which the rule is true (confidence).

```
PQR MATERIAL in [ "" " " ] [ Mode: 1 ] (436)
   DAMAGE REPAIR in [ "No" ] [ Mode: 1 ] (49)
      Work Crews in [ "10" "16" "18" "20" "25" "26" "29" "7" ] [ Mode: 7 ] (32)
         PQR Distance from the starting node <= 95.855 [ Mode: 7 ] (24)
            Loss (m/Km) 0:00 <= 0.280 [ Mode: 1 ] (18)
               Final Node Demand  (L/s) 00:00 <= 0.065 [ Mode: 1 ] => 1 (9; 0,667)
               Final Node Demand  (L/s) 00:00 > 0.065 [ Mode: 7 ] => 5 (9; 0,222)
            Loss (m/Km) 0:00 > 0.280 [ Mode: 7 ] => 7 (6; 0,833)
         PQR Distance from the starting node > 95.855 [ Mode: 12 ] => 12 (8; 0,625)
      Work Crews in [ "11" "12" "14" "6" "9" ] [ Mode: 9 ] (17)
         Final Node Coordinate Y <= 992315.155 [ Mode: 9 ] => 9 (7; 1,0)
         Final Node Coordinate Y > 992315.155 [ Mode: 1 ] => 1 (10; 0,3)
   DAMAGE REPAIR in [ "Yes" ] [ Mode: 4 ] (387)
      Coordinate Y <= 990736.494 [ Mode: 1 ] => 1 (54; 0,463)
      Coordinate Y > 990736.494 [ Mode: 4 ] (333)
         Starting Node Coordinate X <= 1158706 [ Mode: 6 ] (62)
            Work Crews in [ "12" "16" "17" "18" "2" "3" "6" "7" ] [ Mode: 6 ] (39)
               Starting Node Pressure (WC) 6:00 <= 51.455 [ Mode: 6 ] => 6 (32; 0,688)
               Starting Node Pressure (WC) 6:00 > 51.455 [ Mode: 3 ] => 1 (7; 0,286)
            Work Crews in [ "10" "15" "19" "25" "29" "4" "9" ] [ Mode: 4 ] => 13 (23;
            0,217)
         Starting Node Coordinate X > 1158706 [ Mode: 4 ] (271)
            Section Length (m) <= 27.135 [ Mode: 10 ] => 10 (8; 0,75)
            Section Length (m) > 27.135 [ Mode: 4 ] (263)
               Final Node Coordinate Y <= 993953 [ Mode: 4 ] (256)
                  Starting Node Pressure (WC) 19:00 <= 41.025 [ Mode: 13 ] (26)
                     Work Crews in [ "1" "20" "21" "25" "3" "4" "5" "6" ] [ Mode: 6 ] => 4
                     (17; 0,294)
                     Work Crews in [ "11" "15" "24" "29" "7" ] [ Mode: 13 ] => 13 (9;
                     0,778)
                  Starting Node Pressure (WC) 19:00 > 41.025 [ Mode: 4 ] (230)
                     Network Material in [ "AC" “IG” ] [ Mode: 6 ] (106)
                        Work Crews in [ "11" "2" "3" ] [ Mode: 1 ] => 1 (13; 0,692)
                        Work Crews in [ "1" "12" "15" "16" "17" "18" "20" "21" "23" "24"
                        "25" "27" "29" "4" "5" "6" "7" ] [ Mode: 6 ] => 6 (93; 0,312)
                     Network Material in [ "" "AP" “HDPE” "PVC" ] [ Mode: 1 ] (124)
                        Section Length (m) <= 163.325 [ Mode: 1 ] (115)
                           Work Crews in [ "11" "12" "14" "15" "16" "17" "2" "21" "22"
                           "24" "25" "29" "3" "30" "4" "6" "7" "8" ] [ Mode: 1 ] (103)
                              Starting Node Demand (L/s) 00:00 <= 0.055 [ Mode: 6 ] (34)
                                 Work Crews in [ "12" "14" "2" "3" "4" "7" ] [ Mode: 6 ] =>
                                 6 (14; 0,571)
                                 Work Crews in [ "11" "16" "17" "29" "30" "6" "8" ] [ Mode:
                                 4 ] (20)
                                    R(mm) Bella <= 0.150 [ Mode: 4 ] => 4 (11; 0,545)
                                    R(mm) Bella > 0.150 [ Mode: 1 ] => 1 (9; 0,333)
                              Starting Node Demand (L/s) 00:00 > 0.055 [ Mode: 1 ] (69)
```

```
                      R(mm) Bella <= 9.850 [ Mode: 1 ] => 1 (51; 0,569)
                      R(mm) Bella > 9.850 [ Mode: 4 ] => 4 (18; 0,278)
                  Work Crews in [ "1" "10" "18" "23" "26" "9" ] [ Mode: 4 ] => 4
                  (12; 0,667)
                Section Length (m) > 163.325 [ Mode: 4 ] => 4 (9; 0,889)
          Final Node Coordinate Y > 993953 [ Mode: 4 ] => 4 (7; 0,857)
PQR MATERIAL in [ “CI” “IG” “HDPE” “FP” "PVC" ] [ Mode: 1 ] (161)
  PQR DIAMETER(mm) <= 22.700 [ Mode: 1 ] (132)
    PQR DIAMETER(mm) <= 14.350 [ Mode: 1 ] (79)
      Work Crews in [ "1" "20" "21" "26" "30" "5" ] [ Mode: 10 ] => 10 (14; 0,786)
      Work Crews in [ "10" "11" "16" "17" "18" "2" "25" "28" "29" "4" "6" "7" "8" "9" ]
      [ Mode: 1 ] (65)
        R(mm) Jardin <= 12.500 [ Mode: 1 ] => 1 (49; 0,653)
        R(mm) Jardin > 12.500 [ Mode: 3 ] => 3 (16; 0,438)
    PQR DIAMETER(mm) > 14.350 [ Mode: 3 ] (53)
      Work Crews in [ "12" "17" "18" "2" "21" "25" "29" "6" ] [ Mode: 1 ] (28)
        Section Length (m) <= 84.705 [ Mode: 1 ] => 1 (14; 0,714)
        Section Length (m) > 84.705 [ Mode: 2 ] => 2 (14; 0,5)
      Work Crews in [ "10" "11" "16" "27" "4" "7" "8" "9" ] [ Mode: 3 ] (25)
        Loss (m/Km) 0:00 <= 0.015 [ Mode: 2 ] => 2 (13; 0,462)
        Loss (m/Km) 0:00 > 0.015 [ Mode: 3 ] => 3 (12; 0,833)
  PQR DIAMETER(mm) > 22.700 [ Mode: 1 ] => 1 (29; 0,724)
```

Table 2. Obtained rules

The interpretation of the obtained rules is straightforward. The attribute or field that generates more information gain for the first tree division is the material reported in PQRS, which divides the tree into two branches. It is important here to note one of the strengths of these algorithms: when information is missing classifications are obtained despite the absence of this information. Decisions related to the classification of damage type correspond to the symbol => in the rules. Then, each of these two branches subdivides to reach the total depth of the tree, 13 in this case; only 16 of the initial attributes appear in the final model.

6. CONCLUSIONS

We have presented a real world practical application of some of the steps of pre-processing and post-processing information within a KDD process. The information used corresponds to failure events in the WSN of a municipality. Additionally, data mining tasks aimed at finding relationships between the different selected variables and the type of damage produced in the WSN.

The first problem was precisely to work with real data. The variety of uncertainties that impairs control over information in the sense of leading the process, as happens when working with experimental models, generates a lot of issues that require more ingenuity to address the problem.

Basically, the main tasks of pre-processing in this work consisted in attribute selection to take into account, handling missing values, numerizing attributes, and graphical representation of the selected variables to get a first approximation to the information. This visualization allows better understanding of the data that will be used in modelling.

For post-processing, some tasks were used such as pruning, data partition, and impurity management. Also, the model validation was executed by statistical values. This allows integration of discovered knowledge to improve the management of a supply network in terms of likely damage that the network could undergo.

Regarding the C&RT obtained model, its strength to missing values must be highlighted. This feature is especially interesting for WSN management. Moreover the solutions obtained with the C&RT algorithms are better than those of ANN algorithms.

7. ACKNOWLEDGEMENTS

This work has been carried out with the support of the project IDAWAS, DPI2009-11591, of the Dirección General de Investigación of the Ministerio de Ciencia e Innovación (Spain). A special gratitude must be acknowledged to Calarcá Water Supply Company Multipropósito, S.A. ESP for the collaboration with the data used in this work.

8. REFERENCES

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and regression trees, Monterey, CA: Wadsworth and Brooks-Cole, 1984.

Bruha, I., and A. Famili, Postprocessig in Machine Learning and Data Mining, *SIGKDD Explor. Newsl. 2*(2) (Dec. 2000), 110-114. DOI= http://doi.acm.org/10.1145/380995.381059. 110-114, 2000.

Clementine® 9.0 Algoritms Guide, Integral Solutions Limited. EE.UU, 2004.

Feelders, A., H. Daniels, and M. Holsheimer, Methodological and practical aspects of data mining, *Information & Management*, DOI= http://dx.doi.org/10.1016/S0378-7206(99)00051-8, *37*(5), 271-281,2000.

Gibert, K., J. Izquierdo, G. Holmes, I. Athanasiadis, J. Comas, and M. Sànchez-Marrè, On the role of pre and post-processing in environmental data mining, International Congress on Environmental Modelling and Software - 4th Biennial Meeting, pp. 1937-1958, Barcelona, Spain, 2008.

Zhang, S., C. Zhang, Q. Yang, Data Preparation for Data Mining*, Applied artificial intelligence, 17*, 375-381, 2003.