2010 13th International IEEE
Annual Conference on Intelligent Transportation Systems
Madeira Island, Portugal, September 19-22, 2010

Validation of both number and coverage of bus Schedules using AVL data

Luís Matias, João Gama, João Mendes-Moreira, Jorge Freire de Sousa

Abstract—**It is well known that the definition of bus schedules is critical for the service reliability of public transports. Several proposals have been suggested, using data from Automatic Vehicle Location (AVL) systems, in order to enhance the reliability of public transports. In this paper we study the optimum number of schedules and the days covered by each one of them, in order to increase reliability. We use the Dynamic Time Warping distance in order to calculate the similarities between two different dimensioned irregularly spaced data sequences before the use of data clustering techniques. The application of this methodology with the K-Means for a specific bus route demonstrated that a new schedule for the weekends in non-scholar periods could be considered due to its distinct profile from the remaining days. For future work, we propose to apply this methodology to larger data sets in time and in number, corresponding to different bus routes, in order to find a consensual cluster between all the routes.**

I. INTRODUCTION

THERE are several problems involved in public transport management. One of them, indeed a critical one, is to understand the factors that can change the transport schedule and, consequently, the transport's reliability. Unreliable schedules may reduce the number of clients and lead to a critical loss of revenue [15]. In order to increase schedules adherence and, consequently, the quality of the service as understood by the clients, several schedules are defined covering each one a different day type. However, if the number of schedules is too large, it is more difficult for the passengers to memorize trips' offer. Consequently, the choice of this number is necessarily a compromise between the way the trips offer is understood by the passengers and the need the companies have to increase schedules adherence in order to reduce costs and increase clients' satisfaction.

The problem presented and solved in this paper is the choice of the number of different bus schedules and the setting of days covered by each of these schedules for a given bus route. In order to accomplish this objective, we have studied data from the trip durations of one bus route. We did so by applying clustering techniques in order to find similar weekday's profiles. Clustering guarantees that the created groups have intra-groups similarity and, simultaneously, inter-groups dissimilarity. This guarantees, as well, the accomplishment of the concerns previously pointed out: the use of a low number of schedules that are able to maintain a satisfactory level of adherence.

The schedule plan defined for this bus route has four different schedules: working days on scholar periods, working days on non-scholar periods, Sundays and holidays, and Saturdays. So, we have used actual data from its trips to validate it and eventually to propose some changes.

Using the trip times of one bus route during the first eight months of 2004, we tried to identify these profiles by means of data clustering techniques. In order to accomplish this goal, one of the major problems we faced has been the differences verified in the number of daily trips: the irregularly spaced data sequences (ISDS) have values differently defined and also diverse lengths. Therefore, the Euclidean distance will not properly calculate the distances [7].

We have used Dynamic Time Warping (DTW) to calculate a quadratic matrix with the distances between all the ISDS. Based on this matrix, we have used the K-Means

Manuscript submitted April 19, 2010. Revised July 21, 2010.

Luís Matias is with the Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal and with LIAAD-INESC Porto L.A. Rua de Ceuta, 118, 6º; 4050-190 Porto - Portugal (corresponding author to provide phone: 00351-91-4221647; e-mail: luis.matias@fe.up.pt).

João Gama is with the Faculdade de Economia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal and with LIAAD-INESC Porto L.A. Rua de Ceuta, 118, 6º; 4050-190 Porto - Portugal (e-mail: jgama@fep.up.pt).

J. Mendes-Moreira is with the Departamento de Engenharia Informática, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal and with LIAAD-INESC Porto L.A. Rua de Ceuta, 118, 6º; 4050-190 Porto - Portugal (e-mail: jmoreira@fe.up.pt).

J. F. Sousa is with the Departamento de Engenharia Industrial e Gestão, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, s/n 4200-465 Porto - Portugal (e-mail: jfsousa@fe.up.pt).

978-1-4244-7659-6/10/$26.00 ©2010 IEEE

algorithm to create logical partitions (clusters) with the different days present on the dataset. We have tested different values for the K parameter. We have also compared the profiles of two clusters in order to understand how different the profiles could be. We have concluded that the weekends in a non-scholar period have a profile totally distinct from the remaining days and this probably can justify a fifth schedule definition in the schedule plan just for this group of days.

This paper is structured as follows: Section 2, states a brief description of the problem we want to solve. The used techniques are presented as well as some related work. Finally, it is described how these techniques are used to solve the problem. Section 3 presents summarily the dataset used, its main characteristics and some statistics about it. In section 4, we present: (1) the results obtained through the clustering of the data, using different values of K; (2) a mean profile of some relevant clusters; and (3) a discussion about these results. Section 5 relates synthetically the conclusions stated from the work described in this paper as well as the future work we intend to carry out on this subject.

II. Problem and Methods

The problem was to realize if and how we can group the different weekdays into logical groups corresponding to different bus schedules for different periods. If you could group the Saturdays with the weekdays of the non-scholar period, you can, for instance, assign the same schedule to it.

These groups allow the understanding of bus behavior and the adjustment of its schedules in order to achieve its best performance. This grouping (or clustering) was made based on data about the daily bus trips, namely, its duration. This information gave us ISDS, with different sizes, where each sequence has the durations of all trips from a given day, making the traditional method to calculate the distances between the sequences (i.e. Euclidean) inefficient [7].

In this section, we will present data clustering algorithms, as well as the Dynamic Time Warping distance to compare ISDS. We will state other approaches to this problem and other applications of these techniques. Finally, we will present the methodology we have used to solve it.

A. Data Clustering

The goal of data clustering, also known as cluster analysis, is to discover the natural groupings of a set of patterns, points, or objects [1,2]. With this, we can create logical groups from the original dataset. The most popular clustering algorithm is K-means. It has been chosen due to its simplicity, efficiency and efficacy [4,5,6]. This subsection follows subsection 2.3 from [1].

K-Means Algorithm

The K-Means finds a partition of K clusters minimizing the squared error between the mean of cluster C_i and the points of cluster C_k. It starts by selecting K random points (centroids). Then, it classifies each data point according to the minimum distance for the existing centroids, creating partitions (the clusters). The new centroids of the K clusters are calculated minimizing the sum of the distances between each centroid and the data points from its partition, according to the previous classification. Based on the new centroids, new partitions are generated, and the process goes on iteratively [3] until the centroids don't change anymore between iterations.

There are mainly three parameters to K-Means algorithm: the number of clusters to create (K); the seed number to select the first K centroids; the distance metric to be used by the algorithm. Usually, the distances between the points and the centers are calculated using the Euclidean distance, but there are many formulas to calculate the distances between the points like the Manhattan distance, for instance.

Despite the Euclidean distance is used to discover the differences between two ISDS, it is very sensitive to variations in the depth and in the granularity of the time axis [7]. In the next subsection we present another distance metric that surpasses these problems.

B. Dynamic Time Warping

The DTW algorithm calculates the distance between two ISDS that could have, or have not, different dimensions. Despite the fact that the Euclidean distance is the most used distance metric for data clustering [7,8,9,10], the DTW is equally used with different Data Mining techniques with better results [7] for time-varying sequences.

This description follows the subsection 2.2 in [7]. If we have two sequences as P_n and Q_m and we want to align them using DTW, we need to construct an *n-by-m* matrix containing the distances between all the points of the two series. Then, a warping path is defined. This warping path is a contiguous set of matrix elements that defines a possible and optimized mapping between P and Q. The k_{th} element of the warping path is defined as $w_k = (i,j)$, so we have:

$$W = w_1, w_2, \dots, w_k, \dots, w_K \quad (1)$$

$$\max(m,n) \leq K \leq m + n - 1 \quad (2)$$

This path is subjected to three main constraints:

- **Boundary conditions:** $w_1 = (1,1)$ and $w_k = (m,n)$. This requires that the warping starts and ends in the diagonally opposite cells of the matrix.

- **Continuity**: Let $w_k = (a, b)$. Then $w_{k-1} = (a', b')$, where $a - a' \leq 1$ and $b - b' \leq 1$. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).

- **Monotonicity**: Let $w_k = (a, b)$. Then $w_{k-1} = (a', b')$, where $a - a' \geq 0$ and $b - b' \geq 0$. This forces the points in W to be monotonically spaced in time.

In order to build an optimized path satisfying the conditions above, it is necessary to minimize the warping cost:

$$DTW\ (P, Q) = \min\left\{\frac{\sqrt{\sum_{k=1}^{K} Wk}}{K}\right\} \quad (3)$$

C. Related Work

As far as we know, in the transportation field, there is no implementation or related work of DTW applied to data in order to cluster it to understand the weekday's profiles as we present in this paper. However, some other works exist that are somehow related to our research. For instance, in transportation, we can also find some related work using data mining techniques like clustering in order to evaluate and identify cycles in the temporal variability of transit use [17]. In the electricity area, Martinez Alvarez *et. al* [16] used cluster algorithms in order to discover patterns in ISDS. The main difference to our work is that the ISDS are equally sized. Patnaik *et. al* [18] used data clustering to develop bus scheduling plans with real data.

The Dynamic Time Warping (DTW) algorithm is commonly used to calculate the distances between two ISDS in several research areas such as Euler and Riedel do in speech recognition [12], Allon and Sclaroff in video clustering [14], Ashraf and Wong in Motion Editing [13] or Cheng and Fu on pattern matching [11]. Again, it is not a new technology in the research field of transportation: Shibuhisa *et. al* [19] used it in order to align transports positions in a map, according to its routes.

D. Methodologies

In order to cluster the weekdays, we extracted from the initial data 243 ISDS with different dimensions. These distances were used to fill a quadratic and symmetric matrix containing these distances. The main contribution here was to consider that we can characterize the profile of the weekdays using the distance between its ISDS and the other we can compare to. We have clustered the relative distances between the sequences and not the sequences it selves. We have used the DTW to calculate these distances because the ISDS had different lengths.

Then, it was applied to that matrix a clustering algorithm (in this case, the K – Means) with different values of K in order to understand if we could discover some patterns between the weekday's behaviors. Finally, we drew a representative curve of the travel-days belonging to that cluster, by segmenting the time in 15 minutes length stretches and averaging the round-trip times for each one of these stretches. The results of an application of this methodology are described in section 4.

III. Dataset

In this section, it is summarily described the dataset and the preprocessing applied to it. It was considered the source of the data, its organization and the first transformations applied to it in order to group the information into sensible partitions.

The source of this data was STCP, the Public Transport Operator of Porto, Portugal. The dataset was obtained through a bus dispatch system that integrates an Automatic Vehicle Location (AVL) system. The data captured through this system contains data about the trips made on the same route in the first eight months of 2004. This dataset has one entry for each trip of the bus with the following information: the date of the start of the trip; the departure time; the model of the bus; the code of the driver; the code of the route's service; the number of the day of the year (i.e.: two for the second, twenty for the twentieth); the type of the day (normal day, holiday and consecutive holidays) and the duration of the trip. In table 1 some statistics of the data are presented.

TABLE I
DATASET STATISTICS

Total of Trips	7206,00
Longest Trip Time	13200,00
Shortest Trip Time	2655,00
Minimum of Trips per day	8,00
Maximum of Trips per day	58,00
Mean of trip's time	4322,95
Median of trip's time	4279,00
Standard Deviation of trip's time	899,61

Round-Trip Statistics (seconds).

The schedule defined for this route is splitted in four big clusters: Saturdays, Sundays and holidays, working days in a scholar period and working days in a non-scholar period.

Preprocessing

Using the initial dataset, a new dataset was constructed. This dataset has an entry for each day present on the initial one, with several variables for the different durations of the trips. The variables (or columns) of the new processed dataset are: the number of the day in the year; the date; the

weekday name; the type of the day (normal day, holiday and consecutive holidays) and the duration of the trips.

Each entry will have as many missing values as the difference between the number of trips of this day and the number of duration variables (58). The days with one trip or less were automatically deleted. After this preprocessing, the data was ready to be manipulated. In figure 1 a summary of the characteristics of the data is presented. We can see that there is a huge difference between the profiles of the days: they have large variations in both number of trips and its duration.

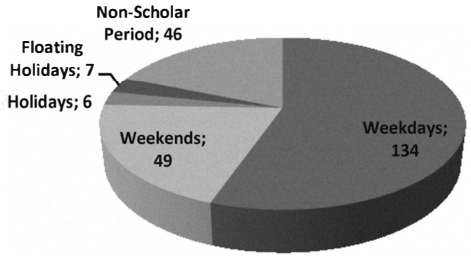

Fig. 1. Type and percentage of days contained in the data. The total number is 242.

IV. RESULTS

We applied the methodologies presented in section two to the dataset described in section three. It is expected to compare the results obtained with the schedules defined for the bus route belonging to this data set.

A. The Clusters Obtained

It was expected to see the logical partitions of the weekdays for the clusters and to understand the meaning of those partitions: weekend, holidays or vacancies, for instance. Therefore, the K-Means clustering algorithm was applied to the DTW quadratic matrix using the R Software [20]. In figures from 2 and 3 are presented the charts corresponding to the clusters using K=2, K=3, K=4 and K=7. The holidays are clearly highlighted with a F-mark. The horizontal axis is the time related one and it has the month's start day highlighted. The vertical one is cluster related (integer values naming the assigned cluster). The points correspond to the days, divided by the clusters.

B. Centroids Analysis

In order to understand the typical behavior of each cluster along the day it is presented, in figure 4, two charts corresponding to the round-trip times in each cluster's days. The expected day profile is calculated based on the round-trip time's mean. The mean was calculated for the round-trip times in each 15-minute length stretch of the variable departure time. This experiment was made only with the clusters generated for K=2.

C. Discussion

As it can be seen in the previous two subsections, there is a nearly perfect clustering for K=2: the weekends, holidays and the non-scholar period in cluster one, and the remaining days in cluster two. There were some points escaping to this logical partitioning that must be analyzed: some are floating holydays and weekends, other are just days with a similar profile.

In the clustering charts for K=3 and K=4, we can see a refinement of the first two clusters, with some weekends running faster than others in K=3 chart and some weekdays failing the schedule in K=4. The chart with the clustering results for K=7 demonstrate groups mainly for the Saturdays, Sundays and Holidays, Fridays and other days of the week, being the meaningless chart of the total set of four.

Considering this route's schedule, we consider that it is good but it can, probably, be optimized (the ineffective schedule plans must be always improved [18]). Ignoring the chart with the clustering results for K=7 (Patnaik et al. [18] states that the schedules must be not too fragmented), we can conclude three different facts: as it is possible to see in Figure 4, the weekdays and the weekends in the scholar period have distinct profiles – so, the definition of two or more different schedules for this two periods is correct. The holidays are, in more than 90% of the cases, grouped with the Sundays in the different clusters – they appear to have similar profiles, what is correct. In other hand, despite the different schedules defined for Saturdays and for Sundays, they are almost always in the same cluster, indicating that may be not needed to have two schedules for them. Finally, we can see in the chart with the clustering results for K=3 and K=4 that the weekends in a scholar period are always grouped in a different cluster than the weekends in a non-scholar period (some isolated days are included in this cluster too). These differences justify, in our opinion, a deeper study about the need to create a fifth schedule plan to the weekends in a non scholar period for its unique behavior and to study the hypothesis to group the weekends and holidays in the same schedule. Nevertheless, it is well known that we must develop a schedule to a route network and not to a specific one. A larger dataset is needed, with different routes, in order to justify this need, but this work can be a good start for it.

The creation and validation of schedule plans for bus routes networks using data clustering techniques is not a new thing (in [18], this kind of techniques is used to see and understand what is happening). But we proved that the DTW can be really useful to compare the ISDS of the round-trip times. The clusters obtained using the DTW distance can be used to validate the existing schedules plan and extract information in order to improve it, if necessary.

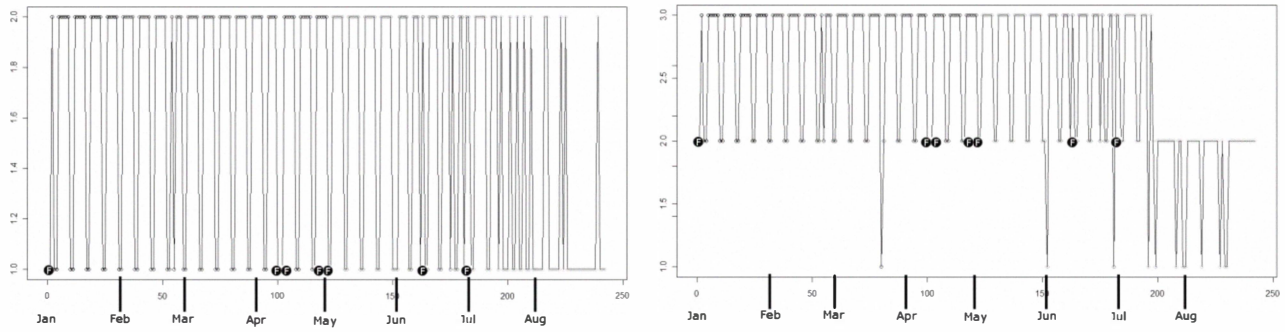

Fig. 2. Clustering results for K=2 and for K=3, respectively.

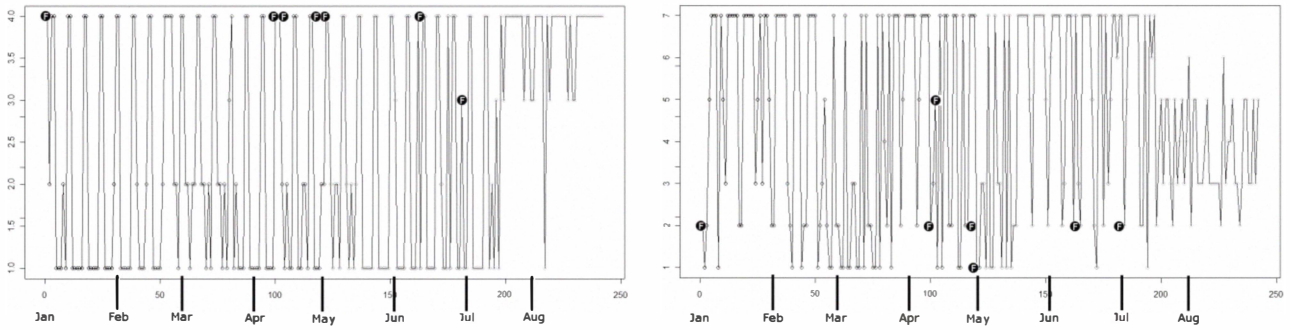

Fig. 3. Clustering results for K=4 and for K=7, respectively.

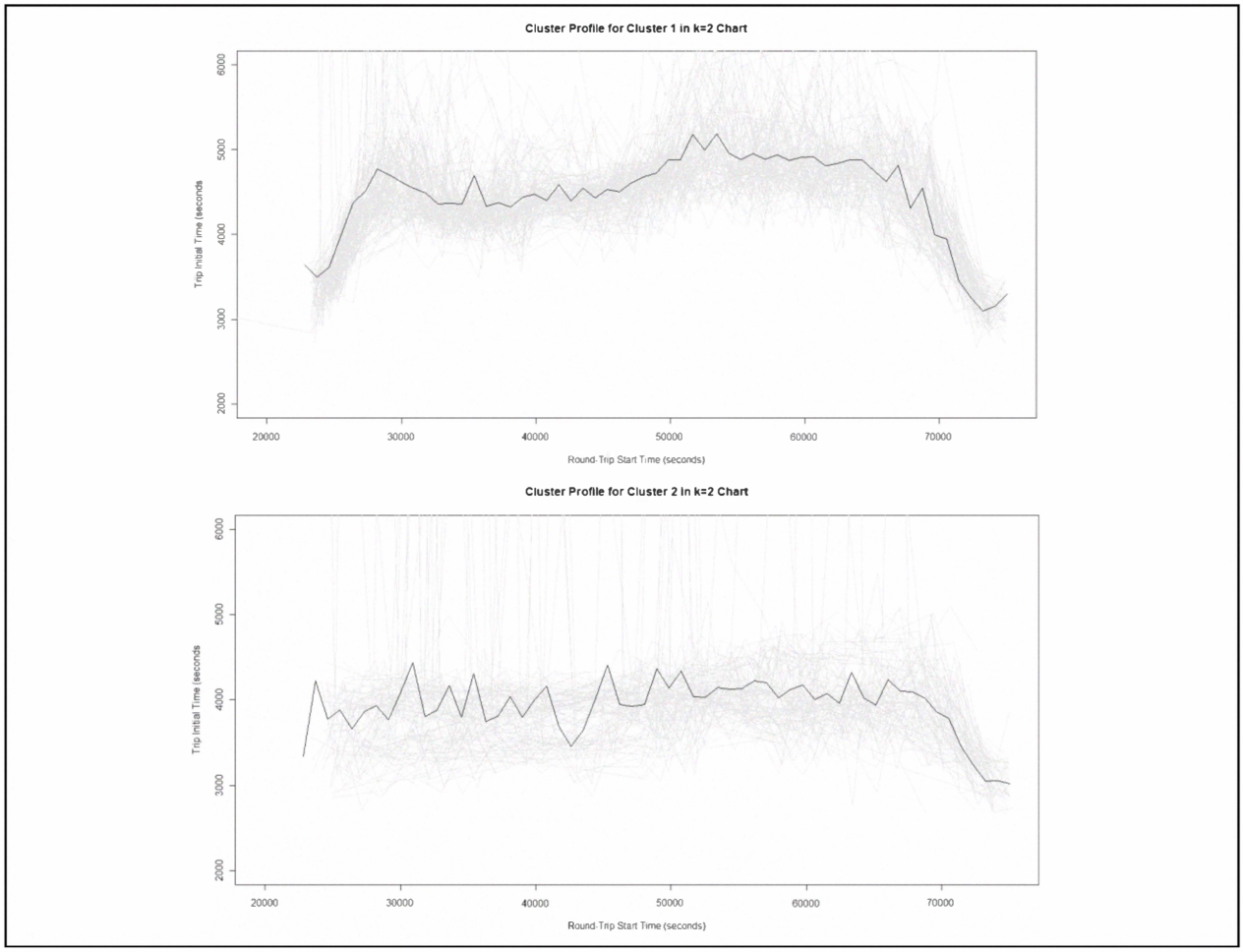

Fig. 4. Comparison between cluster's mean profiles for k=2.

V. Conclusions and Future Work

In this paper, we have presented a new methodology to validate the number and the coverage of bus schedules using AVL data transformed with DTW. The biggest advantage of DTW is that we do not need to have data with the same number of trips for each day of each route in order to compare it properly.

Analyzing the experiments done with several values of K, we concluded that the clusters found correspond, mostly, to the schedule assigned to the bus route considered. This can be better observed in figure 4, with totally different curves to the weekdays and weekends clusters (k=2). According to the previous section, the holidays could be considered in the same schedule of the weekends but it is known that the holidays have unique behaviors from country to country and, sometimes, from city to city. So, we do not propose this kind of change.

However, a fifth group is clearly identified in the several clusters generated: the group of weekends in a non-scholar period. In other hand, Saturdays are almost always grouped with Sundays. The current schedule plan (working days on scholar periods, working days on non-scholar periods, Sundays and holidays, and Saturdays) may not be the optimal solution for this route.

This conclusion supports, in our opinion, that this is a valid methodology in order to validate schedule plans based on real bus route trips data. The biggest contribution of this methodology is the use of DTW to calculate the distances between ISDS existing in AVL data, calculating an optimized distance between sequences that have not the same length.

Future Work

The schedule plan must be developed to an entire route network and not only to an isolated one [18]. So, to totally support the change proposed above, a deeper study with a bigger dataset and several routes is needed. The data should be from, at least, one complete year. The work presented in this paper should be repeated for all the routes of the network and a consensual cluster should be calculated (to know more about consensual clustering see, for instance, section 3 from [21]).

References

[1] Jain, A.. "Data Clustering: 50 Years Beyond K-Means." *Pattern Recognition Letters*, 2009.

[2] Meriem-Webster. *Cluster Analysis.* 1948. http://www.merriam-webster.com/dictionary/cluster%20analysis (accessed in 2010).

[3] Jain A., Dubes R.. *Algorithms for Clustering Data.* Prentice Hall, 1988.

[4] Lloyd, S. "Least squares quantization in PCM" *IEEE Transactions on Information Theory, 28*, 1982: 129-137.

[5] Ball G., Hall d.. "ISODATA, a novel method of data analysis and classification." *Tech. Rep.. Stanford University, Stanford, CA.*, 1965.

[6] Macqueen, J. "Some methods for classification and analysis of multivariate observations." *Fifth Berkeley Symposium on Mathematics, Statistics and Probability.* California: University of California Press, 1967. 281-297.

[7] Chu S., Keogh E., Hart D., Pazzani M. "Iterative deepening dynamic time warping for time series." *Proceedings of the 2nd SIAM international conference on data mining.* 2002.

[8] Agrawal R., Lin I., Sawhney H., Shim K.. "Fast similarity search in the presence of noise, scaling, and translation in times-series databases." *Proc. of 21st International Conference on Very Large Data Bases.* 1995.

[9] Das G., Lin K., Mannila H., Renganathan G., Smyth P.. "Rule Discovery form Time Series." *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining.* 1998. 16-22.

[10] Debregeas A., Hebrail G.. "Interactive Interpretation of Kohonen Maps Applied to Curves." *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining.* 1998. 179-183.

[11] Cheng H., Fu K.. "VLSI architectures for string matching and pattern matching." *Pattern Recognition, 20.* 1987. 125-141.

[12] Euler S., Riedel K.. "Design and Implementation of a Speech Server for Unix Based Multimedia Applications." *EUROSPEECH '93 - Third European Conference on Speech Communication and Technology.* 1993. 1963-1966.

[13] Golam A., Wong K.. "Dynamic time warp based framespace interpolation for motion editing." *Graphics Interface*, Maio de 2002: 45-52.

[14] Alon J., Sclaroff S., Kollios G., Pavlovic V.. "Discovering clusters in motion time-series data." *IEEE CVPR.* 2003.

[15] Strathman J., Dueker K., Kimpel T., Gerhart R., Turner K., Taylor P., Callas S., Grifin D., Hopper J.. "Automated bus dispatching, operations control and service reliability: analysis of tri-met baseline service date." *Technical report, University of Washington - U.S.A.*, 1998.

[16] Martínez-Álvarez F., Troncoso A., Riquelme J., Riquelme J. M.. "Discovering Patterns in Electricity Price Using Clustering Techniques." ICREPQ – International Conference on Renewable Energy and Power Quality. Sevilla, Spain, 2007

[17] Morency C., Trépanier M., Agard B.. "Measuring Transit Use Variability With Smart-Card Data." Transport Policy, 14, 2007: 193-203.

[18] Patnaik J., Chien S, Bladikas A.. "Using Data Mining Techniques on APC Data to Develop Effective Bus Scheduling." Journal of Systemics, Cybernetics and Informatics - volume 4, number 1, 2006: 86-90.

[19] Shibuhisa N., Sato J., Takahashi T., Ide I., Murase H., Kojima Y., Takahashi A.. "Accurate Vehicle Localization using DTW between Range Data Map and Laser Scanner Data Sequences." Proceedings of IEEE 2007 Intelligent Vehicles Symposium. Istanbul: Turkey, 2007. 975-980.

[20] R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2005.

[21] Campedel M., Kyrgyzov I., Maıtre H.. "Unsupervised feature selection applied to spot5 satellite images indexing." JMLR Workshop and Conference Proceedings Volume 4 : New challenges for feature selection in data mining and knowledge discovery (FSDM), volume 4. Anvers, Belgique, 2008.