ISBN 978 -987-633-574-4

Vehicular Flow Analysis Using Clusters

Gary Reyes[1] , Laura Lanzarini[2], César Estrebou[2] , Victor Maquilón[1]

[1]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil,
Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador
{gary.reyesz,victor.maquilonc}@ug.edu.ec
[2]Universidad Nacional de La Plata, Facultad de Informática,
Instituto de Investigación en Informática LIDI (Centro CICPBA) 1900 La Plata,
Buenos Aires, Argentina
{laural,cesarest}@lidi.info.unlp.edu.ar

Abstract. The volume of vehicular traffic in large cities has increased in recent years, causing mobility problems, which is why the analysis of vehicular flow data becomes important for researchers. Intelligent Transportation Systems perform vehicle monitoring and control by collecting GPS trajectories, information that provides real-time geographic location of vehicles, which allows the identification of patterns on vehicle flow using clustering techniques. This paper presents a methodology capable of analyzing vehicular flow in a given area, identifying speed ranges and maintaining an updated interactive map that facilitates the identification of areas of possible traffic jams. The results obtained on a dataset from the city of Guayaquil-Ecuador are satisfactory and clearly represent the speed of vehicle displacement by automatically identifying the most representative ranges for each instant of time.

Keywords: vehicular flow, cluster, GPS trajectory

1 Introduction

Nowadays the constant increase in the volume of traffic in large cities causes problems in the vehicular flow, so the analysis of the data generated by the vehicle monitoring and control systems becomes relevant. Its study through descriptive techniques allows to identify relationships between vehicle trajectories facilitating the analysis of the flow of vehicles. They currently provide solutions in a variety of areas, such as health, finance, telecommunications, agriculture and transport, among others [1].

Data clustering is a technique widely used to identify common characteristics between instances of the same problem [2]. Over time, researchers have proposed improvements to the limitations identified in some techniques such as [3] where a correct initialization of the algorithm is achieved in a much shorter time. In other cases, techniques have been adapted to work in a specific context, such as for spatial data mining [4] [5] [6] or for GPS trajectory analysis [7]. A GPS trajectory is defined by a set of geographic locations each of which is represented by

ISBN 978 -987-633-574-4

its latitude and longitude, in an instant of time. This paper proposes a methodology for the analysis of vehicular flow in traffic through the analysis of GPS trajectories. For this purpose, each zone within an area of interest is characterized according to the average speed and the number of vehicles it contains, in a given period of time. The zones are delimited at the beginning of the process and their size depends on the precision with which the analysis is to be carried out. Then, using a variation of the dynamic clustering algorithm for data flows named Dyclee, originally defined by Barbosa et al. [8], the zones with similar characteristics are identified and an interactive map is constructed on which the ranges of speeds corresponding to the current vehicular flow and the zones where they occur can be observed. This methodology can be used, together with other tools, by traffic managers in a city to plan urban roads, detect critical points in traffic flow, identify anomalous situations, predict future mobility behavior, analyze vehicular flow, among others. The proposed methodology was used to characterize the data corresponding to GPS trajectories generated by a group of students from the University of Guayaquil, Ecuador. The obtained results allow to identify at different time instants, sectors of the city where vehicles have common speeds.

This article is organized as follows: section 2 analyzes some related work that were identified in the literature and present various solutions to the problem, section 3 describes the proposed methodology, section 4 presents the obtained results and finally, section 5 contains the conclusions and lines of future work.

2 Related work

Clustering techniques have been used in trajectory analysis for several years. They are usually adaptations of conventional algorithms using similarity metrics specially designed for trajectories [9] [10]. Such is the case of the Improved DBScan algorithm [11] which improves the traditional DBScan algorithm using its own density measurement method that suggests the new concept of motion capability and the introduction of data field theory. Another example is the Tra-DBScan algorithm [12], which uses the DBScan [13] algorithm adding a trajectory segmentation phase, in which it partitions the trajectories into sections and uses the Hausdorff distance as a similarity measure.

Yu et al. [14], an improved trajectory model is proposed and a new clustering algorithm is presented, with a similarity measure that calculates the distance between two trajectories based on multiple features of the data, achieving maximization of the similarity between them. On the other hand, Ferreira et al. [15] presents a new trajectory clustering technique that uses vector fields to represent the centers of the clusters and propose a definition of similarity between trajectories. A GPS vehicular trajectory clustering method using angular information to segment trajectories and a pivot-guided similarity function is presented by Reyes et al. [16]. Research efforts in this area continue today [17] [18].

In summary, it can be stated that clustering techniques have proven to perform well in the analysis of vehicular trajectories although their parameterization

ISBN 978 -987-633-574-4

remains an interesting challenge. This is related to the fact that they are unsupervised techniques that generally combine distance and density metrics to control the construction of the clusters.

This article used a dynamic clustering algorithm for data streams. This type of algorithms process data flows managing to overcome some of the limitations of traditional clustering algorithms, which usually iterate over the dataset more than once, causing greater memory usage and increasing execution time [19] [20]. As the distribution of the data in each stream changes continuously, it is important that these clustering algorithms that process data flows generate dynamic groups, where the number of groups depends on the distribution of the data of the flow [21] [22].

In particular, in this article a variation to the DyClee algorithm has been made, originally designed by Barbosa et al. [8]. It is a dynamic clustering algorithm for tracking evolving environments capable of adapting the clustering structure as the data is processed.

Dyclee uses a two-stage clustering approach [23]. The first stage consists of clustering the examples based on their similarity and density. The groups obtained as a result of this stage are called microclusters. Then, in the second stage, the microclusters are clustered starting with the densest ones and taking into account their overlapping and similarity in terms of their density.

3 Methodology

This paper proposes a methodology for the analysis of vehicular flow in traffic. This methodology is represented in the figure 1 and consists of three steps: the first step is to properly represent the data of the trajectories within the area of interest; the second step uses an adaptation of a dynamic clustering algorithm to identify relationships and the third step consists of creating interactive resources for the visualization of the results. Each step mentioned is described below.

Representation of vehicular flow

The first step is to provide an adequate representation of the data that make up the trajectories. To do this, first of all, the area of interest must be established. Here, it should be indicated which is the geographical area to which the trajectories to be analyzed belong. Once the area is established, it is partitioned into cells, or smaller zones, in a uniform manner. The size of each cell will depend on the precision with which it is desired to analyze the vehicular flow. In this work, 200 m2 cells were used. This is important data to take into account since the information to be analyzed corresponds to a summary of what is happening in each cell in a given period of time. The methodology proposed here consists of analyzing what is happening in each cell as a whole instead of considering each vehicle trajectory separately. This facilitates analysis and visualization.

In particular, in this article, the data corresponding to vehicular flow, represented in each cell, were analyzed in batch mode in 3-minute periods. However,

ISBN 978 -987-633-574-4

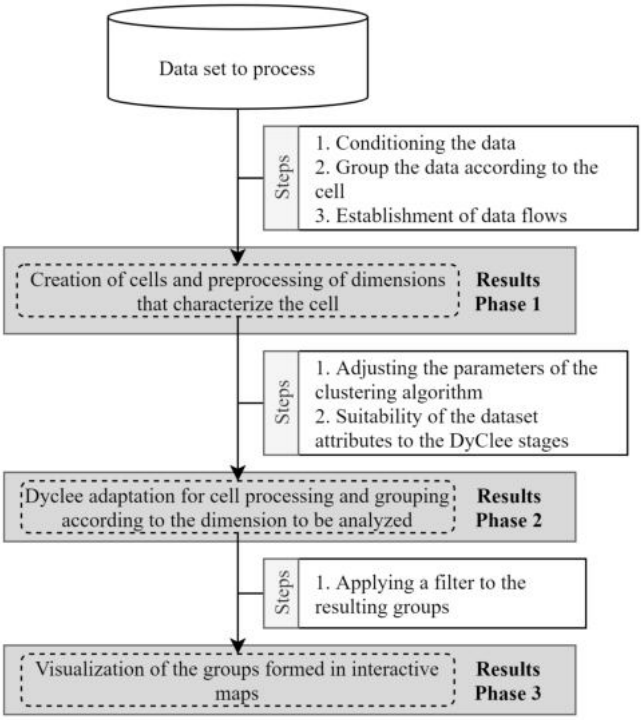

Fig. 1. Proposed methodology

if a specific analysis is desired, these periods can be shorter, for example one minute. Each period is considered an evolution since the DyClee algorithm updates the clustering with the incorporation of each block of data sequentially. In each evolution, a data flow will be entered to perform the respective calculations for each cell and to obtain characteristic information of these cells.

Adapted Dyclee algorithm

The second step uses an adaptation of the DyClee algorithm, defined in this article, to process the trajectories within each cell. The table 1 identifies the basic elements and parameters that were used in this adaptation.

As previously mentioned, the first of the two stages of the DyClee algorithm refers to the construction of microclusters. In this article, instead of using directly the GPS locations and their density, as proposed in the original article, the velocities of the trajectory sections included in each cell were considered. Thus, for a given time period, each cell will be represented by the average velocity of the trajectory sections it contains. This implies trimming the trajectories appropriately considering that speed variations may lead to a vehicle traveling at very high speed not being recorded (or having very few GPS locations) when

ISBN 978 -987-633-574-4

Table 1. Concepts and parameters associated with DyClee processing

Element	Definition
Microcluster	Represents the dataset with similar characteristics.
Hyperbox	Determines the area of the microcluster.
Relative size	Size of the microclusters with respect to the processing area.

passing through a cell. In addition, it is important to consider that speeds should be averaged considering the vehicles and not the number of locations recorded. Regarding the size of the microclusters, the value of the "relative size" parameter specifies the relative size of the "hyperbox" parameter concerning to the area to be processed. That is, as its value decreases, the number of microclusters increases and vice versa.

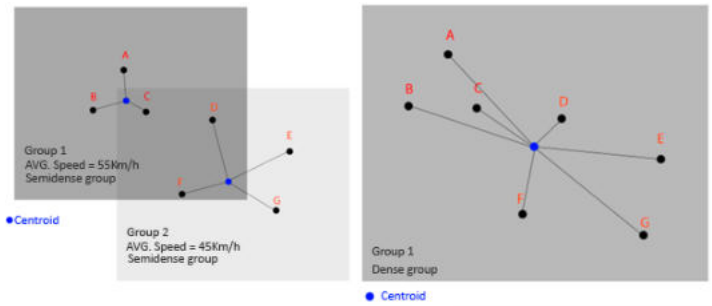

Fig. 2. Second stage operation of Dyclee algorithm. (A) Identification of directly connected microclusters and (B) Resulting expansion

In its second stage, the algorithm analyzes the densities of the microclusters formed and classifies them into two categories: dense and semi-dense. From the dense microclusters it starts to join those that are directly connected. Two microclusters will be considered directly connected if the maximum distance at which the centroids of two microclusters with similar densities can meet does not exceed the value of the "hyperbox" parameter. That is, as the value of the "hyperbox" parameter increases, the number of clusters will be smaller and therefore, the velocity ranges in the area of interest will have greater amplitude. Figure 2 illustrates the operation of this stage.

For each cluster completed in the log, a record is generated for each cell used in the clusters, which contains trajectory data within a period of analyzed time. Although each result is reflected in an interactive map as the analysis progresses, its registration allows reconstructing previous situations.

ISBN 978 -987-633-574-4

Visualization of grouping

With the result of each grouping, an interactive map is created in which the relevant information of each cell can be analyzed graphically and dynamically. Also, using the log mentioned in the previous section, it is possible to reconstruct all the maps from the beginning of the vehicular flow analysis. This provides a quick visualization of the traffic state. In the interface of each map there are layer selection controls and reference legends to interpret the results represented on the map. Two types of maps are generated: a map of the last evolution or time period analyzed and a map with all the evolutions.

In the map corresponding to a particular evolution, each cluster has been represented in a different layer and the user can select one or several layers of the map to filter the information of interest. To facilitate the visualization, a different color has been used in each layer; in this way, if more than one cluster is displayed simultaneously, it will be possible to distinguish to which of them the marked cells belong. The map also has the possibility to select the display of markers that show information of both the group and the selected cell. Figure 3 (B) illustrates the latter.

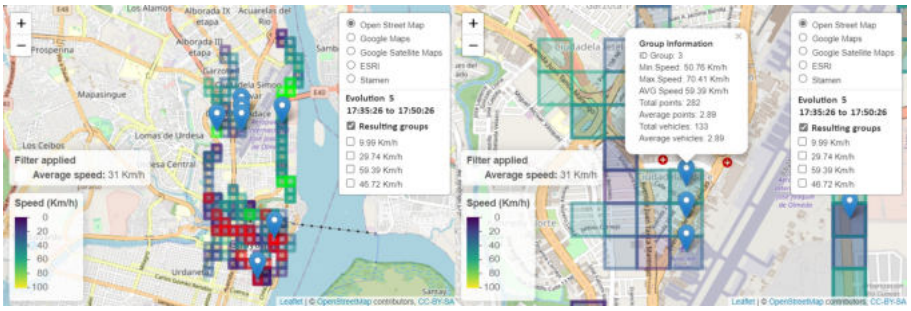

Fig. 3. Maps by evolution. (A) Layer containing the delimitation of the corresponding areas. (B) Layer on which certain markers are activated.

In the map of all the evolutions are visualized, in a single map all the evolutions made with a different color for each evolution, being able at some point to choose one or several evolutions according to the analysis to be carried out.

4 Results and Discussion

To test the performance of the methodology proposed in this article, we work with own data collected in the city of Guayaquil, Ecuador, on October 28, 2017. These data correspond to 218 trajectories made by university students traveling in some means of transportation such as cab, motorcycle and metrovia. The locations in this dataset were collected by smartphones with an average time interval

ISBN 978 -987-633-574-4

between two consecutive locations of 5 seconds. Each record contains trajectory id, latitude, longitude, time, user name, email and type of transportation.

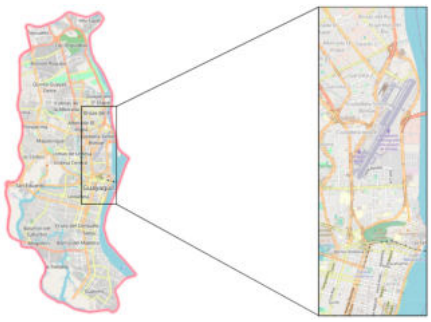

Fig. 4. Area representing the dataset for the city of Guayaquil

Given that this is a small set of trajectories, the analysis was carried out between 16:30 and 18:30 hours, as this is considered to be the time of greatest concentration of records. As a result of this filtering process, 30557 records were obtained, representing 206 trajectories of the entire data set. The area representing the selected data set is shown in figure 4.

The configuration of DyClee is done by means of the necessary parameters. The value for the "relative size" parameter was 0.2, from this value and based on the processing area, the value of the "hyperbox" dimension was defined. The 30557 records were divided into 8 blocks of 15 minutes each and analyzed consecutively. This 15-minute time period could be considered excessive, but its duration is in relation to the volume of data collected. It is important to consider that only the vehicular flow corresponding to the trajectories followed by the students who collaborated with the data collection is being analyzed. To know the vehicular flow of the city at that time, it is necessary to add the information of the rest of the vehicles that circulate in the area in that time range.

The information represented for each cell consists of the average speed of the vehicles registered inside the cell during the period of analyzed time. In Figure 5 the maps corresponding to the 8 evolutions carried out can be observed. Below each map, the information corresponding to the carried out grouping with the adaptation of the DyClee algorithm is indicated. For each group, the minimum, maximum and average speeds are indicated, as well as the deviation of the speeds belonging to that group. These values show the low overlap between them.

Figure 6 illustrates the evolution of the average speeds of the groups over time. There it can be seen that, although some change from one evolution to another, if analyzed in order, they form six common velocity ranges identified as velocity 0, 1, 2, 2, 3, 4 and 5.

ISBN 978 -987-633-574-4

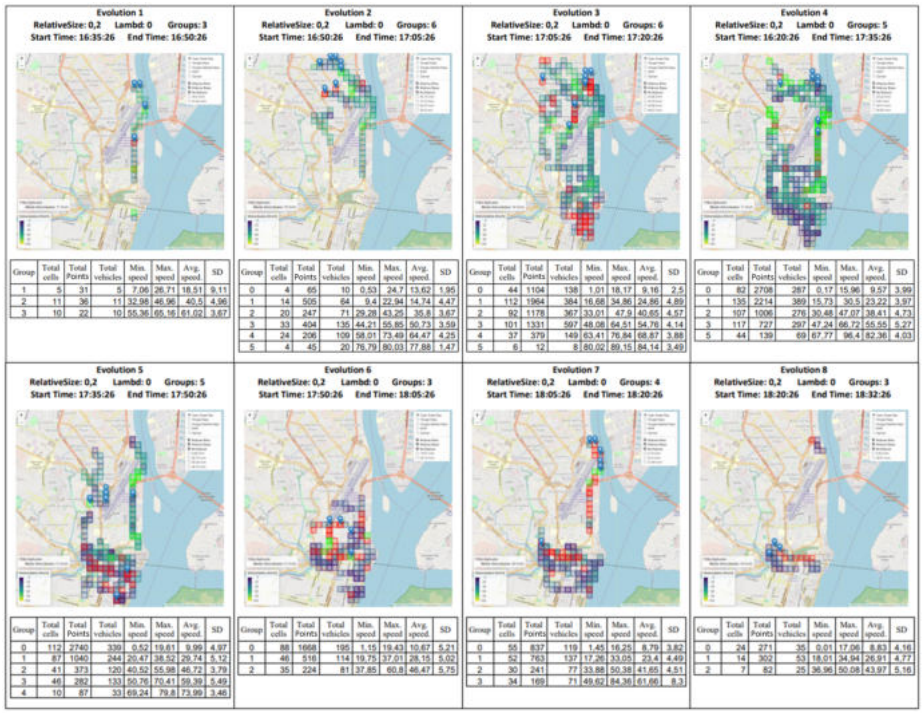

Fig. 5. Results obtained from the 8 experiments in consecutive time periods

In the case of speed rank 3 it only appears in evolutions 1, 2, 3, 4, 5 and 7 which evidences the dynamic characteristic of group conformation of the used algorithm. In addition, from evolution 3 onwards the 3 lowest speed ranges are maintained while the highest speeds are mainly concentrated between evolutions 2 to 5. This shows the speed variation that occurs in the vehicular flow over time identifying that in the first hour of the analysis the traffic reaches high speeds while in the last 45 minutes the trips are made at lower speeds. Figure 6 shows the average speeds in each group identifying the six speed ranges.

5 Conclusions

This article has proposed a methodology to identify, dynamically, the characteristics of the vehicular flow in a period of time. In order to do this, the information of the trajectories has been represented in cells and processed using an adaptation of the DyClee algorithm. As a result of the clustering, different groups that dynamically change from one evolution to another were obtained, identifying common speeds at different time instants, which allows making decisions regarding the city traffic.

ISBN 978 -987-633-574-4

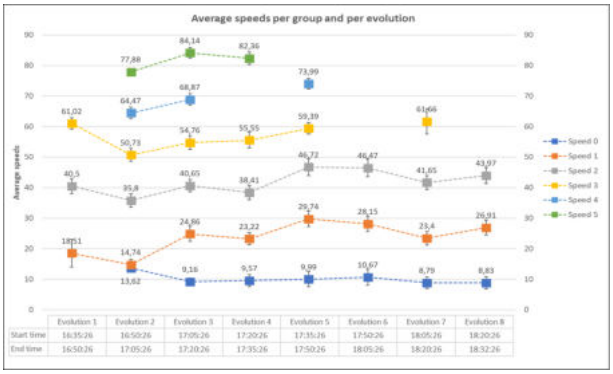

Fig. 6. Average speeds per group and evolution

Interactive maps as part of the methodology are an extremely useful tool when it comes to visualizing, according to the study area, the cells belonging to the different groupings. Through it, is possible to observe particular characteristics of each group and analyze the flow of traffic in specific sectors of the city.

As lines of future work, it is proposed to analyze the incremental incorporation of the data within the clustering together with the concept of forgetting. In this way, it will be sought to give the clustering the possibility of detecting changes in the behavior of vehicular traffic that will help to identify congestion in a more efficient manner.

References

1. A. Jain, "Data clustering: 50 years beyond k-means. 2009," *Pattern Recognition Letters*, 2009.
2. T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.
3. B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," 2012.
4. H. F. Tork, "Spatio-temporal clustering methods classification," in *Doctoral Symposium on Informatics Engineering*, vol. 1, pp. 199–209, Faculdade de Engenharia da Universidade do Porto Porto, Portugal, 2012.
5. J. Han, M. Kamber, and A. K. Tung, "Spatial clustering methods in data mining," *Geographic data mining and knowledge discovery*, pp. 188–217, 2001.
6. B. M. Varghese, A. Unnikrishnan, and K. Jacob, "Spatial clustering algorithms-an overview," *Asian Journal of Computer Science and Information Technology*, vol. 3, no. 1, pp. 1–8, 2013.
7. J. D. Mazimpaka and S. Timpf, "Trajectory data mining: A review of methods and applications," *Journal of Spatial Information Science*, vol. 2016, no. 13, pp. 61–99, 2016.
8. N. Barbosa Roa, L. Travé-Massuyès, and V. H. Grisales-Palacio, "Dyclee: Dynamic clustering for tracking evolving environments," *Pattern Recognition*, vol. 94, pp. 162–186, 2019.

9. M. Y. Choong, R. K. Y. Chin, K. B. Yeo, and K. T. K. Teo, "Trajectory pattern mining via clustering based on similarity function for transportation surveillance," *International Journal of Simulation-Systems, Science & Technology*, vol. 17, no. 34, pp. 19–1, 2016.
10. J. Kim and H. S. Mahmassani, "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories," *Transportation Research Procedia*, vol. 9, pp. 164–184, 2015.
11. T. Luo, X. Zheng, G. Xu, K. Fu, and W. Ren, "An improved dbscan algorithm to detect stops in individual trajectories," *ISPRS International Journal of Geo-Information*, vol. 6, no. 3, 2017.
12. L. X. Liu, J. T. Song, B. Guan, Z. X. Wu, and K. J. He, "Tra-dbscan: A algorithm of clustering trajectories," in *Frontiers of Manufacturing and Design Science II*, vol. 121 of *Applied Mechanics and Materials*, pp. 4875–4879, Trans Tech Publications Ltd, 1 2012.
13. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.
14. Q. Yu, Y. Luo, C. Chen, and S. Chen, "Trajectory similarity clustering based on multi-feature distance measurement," *Applied Intelligence*, pp. 2315–2338, 2019.
15. N. Ferreira, J. T. Klosowski, C. Scheidegger, and C. Silva, "Vector field k-means: Clustering trajectories by fitting multiple vector fields," 2012.
16. G. Reyes-Zambrano, L. Lanzarini, W. Hasperué, and A. F. Bariviera, "GPS trajectory clustering method for decision making on intelligent transportation systems," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5529–5535, 2020.
17. H. Hu, G. Lee, J. H. Kim, and H. Shin, "Estimating Micro-Level On-Road Vehicle Emissions Using the K-Means Clustering Method with GPS Big Data," *Electronics*, 2020.
18. J. Lou and A. Cheng, "Behavior from Vehicle GPS / GNSS Data," *Sensors*, 2020.
19. B. Babcock and J. Widom, "Models and Issues in Data Stream Systems." 2002.
20. M. Garofalakis, J. Gehrke, and R. Rastogi, *Data Stream Management.* 2016.
21. M. R. Ackermann, C. Lammersen, C. Sohler, K. Swierkot, and C. Raupach, "StreamKM++: A clustering Algorithm for Data Streams," *ACM Journal of Experimental Algorithmics*, vol. 17, pp. 173–187, 2012.
22. C. C. Aggarwal, *Data Streams : An Overview and Scientific Applications.* 2010.
23. C. C. Aggarwal, P. S. Yu, J. Han, and J. Wang, "- a framework for clustering evolving data streams," in *Proceedings 2003 VLDB Conference* (J.-C. Freytag, P. Lockemann, S. Abiteboul, M. Carey, P. Selinger, and A. Heuer, eds.), pp. 81–92, San Francisco: Morgan Kaufmann, 2003.