Using Data Mining Techniques on APC Data to Develop Effective Bus Scheduling Plans

Jayakrishna PATNAIK
Department of Industrial and Manufacturing Engineering, NJIT
Newark, NJ 07102, USA

Steven CHIEN
Department of Civil and Environmental Engineering, NJIT
Newark, NJ 07102, USA

and

Athanassios BLADIKAS
Department of Industrial and Manufacturing Engineering, NJIT
Newark, NJ 07102, USA

ABSTRACT

Various trip generators (e.g., buildings, shopping malls, recreational centers) continually influence travel demand in urban and suburban areas. As a result, the headway regularity that should be kept among transit vehicles is difficult to maintain, specifically during peak hours. The variation of headways lengthens the average wait times and deteriorates service quality. Providing a tool to monitor and maintain most up-to-date information through Advanced Traveler Information Systems (ATIS) can assist effective system planning and scheduling, while reducing the door-to-door travel time. This paper develops a methodology for clustering the state variables (number served passengers and halting stations in each vehicle trip) and using that for service planning. The data used to develop the models were collected by Automatic Passenger Counters (APC) on buses operated by a transit agency in the northeast region of the United States. The results illustrate that the developed tool can provide suggestions for improving systems performance as well as future planning.

Keywords: Bus Transit, Headway, Cluster, Classification, CART, APC, and ATIS

1. INTRODUCTION

A passenger's decision to use transit depends on how well transit service quality compares with that of competing modes. Many previous studies have indicated that transit ridership is lost when service is perceived to be unreliable and of poor quality. Unreliable service causes longer waiting times, resulting in passenger anxiety and annoyance. London Transport estimated that the elasticity of demand due to unplanned service losses (i.e., scheduled vehicle-miles not operated) is 33 % larger than the elasticity related to planned service cuts [1]. An analysis of automatic vehicle location systems (AVLS) and Automatic Passenger Counter (APC) data in Portland, Oregon found that a 10% reduction in headway delay variation on radial bus routes during the morning peak hour led to an increase of 0.17 passengers per trip per time-point [2]. Passengers perceive extra waiting time to be even longer than it actually is and this perception can reduce the attractiveness of transit [3]. If agencies can monitor ridership changes along the route and adjust their headway plans periodically, the quality of service can be improved and the mode can become more attractive.

With the advent of Intelligent Transportation Systems (ITS), real time information plays an important role in assisting transit agencies to make critical decisions quickly, and intelligently. Through ITS, huge amounts of data can be collected and processed with high performance computing systems that can help transportation engineers and planners do their jobs more efficiently [5]. Transportation planners who can convincingly present their plans with all the associated facts to policy makers and stakeholders can be more effective and productive. Modern automatic passenger counter systems relying on sensors are highly complex and sophisticated. These systems are capable of counting the number of boarding and alighting passengers and the associated door open and close times at all bus stops.

Headway is the time period between two successive buses dispatched from the origin of a bus route. From a transit supplier's point of view, a way to provide the best level of service with the least operating cost is to operate optimal headways to cater to different demand levels at different times-of-day. This in effect will reduce passenger-waiting time, if vehicles can be frequently dispatched during peak periods and at a lesser frequency during off peak periods. This will also reduce passengers' anxiety and impact positively on their perception about transit service.

The data collected by APC could be mined to provide useful information on whether the agency utilizes optimum operational headways. A methodology is outlined here to develop, implement and monitor effective bus scheduling plans (i.e., optimum operational headways) using available data mining tools. The procedure can be extended to the entire routing network to enable transit agencies to efficiently run their fleet and service various segments of all routes experiencing varied transit ridership at different times of the day. This study determines operational headways for a transit agency and demonstrates a prototype application of Hierarchical Clustering, and Classification and Regression Trees (CART) [7,8,9].

2. OBJECTIVES AND SCOPE

Transportation planners and transit schedulers need to develop effective vehicle scheduling and routing plans to improve transit service quality. There is always a need to improve ineffective schedule plans that may have been established quite some time ago and do not reflect the current land use and patronage distribution a transit system serves. In this study, transit ridership demand was evaluated by using clustering techniques on data collected by an Automatic Passenger Counting (APC) system.

3. DATA COLLECTION AND PREPARAION

This study was carried out using one-year (2002) data collected from the archive data system (ADS) of a transit agency in the northeastern region of United States. The ADS archives entire trip data captured by APC units installed on buses serving a 30-mile urban bus route on daily basis. The transit agency is one of the largest and reputable public transportation corporations in the nation covering more than 5000 square miles of service area. The agency owns around 2,100 buses, 610 trains and 45 light rail vehicles serving more than 380,500 customers each day and consequently providing nearly 224 million passenger trips every year. The ADS established by the transit agency emphasizes on improving transit operation in connection with improved reliability, fleet management, increased passenger safety and security, reduced response time, performance monitoring, and improved communication

The transit agency provides services along the studied route over different time periods with a number of different "patterns" in the inbound and outbound direction. Patterns differ in terms of their routing plans. All route patterns serve fourteen important stops at identical physical locations (known as "Time Points") that are listed on the agency timetables. These "Time Points" are significant trip generators [6]. Data collected from outbound weekday trips for a specific pattern were used in the study to develop and calibrate the proposed model of operational headway plans. The original APC data were processed by converting arrival times at stops to inter-stop travel times, and extracting the actual number of stops, the total dwell time, and the number of alighting and boarding passengers for every bus trip. It is worth noting that occasionally the APC system records erroneous data. For example, the door open time at a stop may be earlier than the door close time at the immediately preceding stop. Obviously erroneous data were excluded. The training datasets were then subjected to data mining. Existing data mining tools such as CART (Classification and Regression Trees) for decision trees as well as SAS (Version 8.02) for Hierarchical Clustering were used to analyze the data and extract information.

4. METHODOLOGY

The methodology consists of creating a training dataset by accumulating passengers boarding at every single stop as well as the number of times the bus halted during a trip between any pair of points along the route. This generates key information of the state variables, "Cumulative number of Passengers on board" and "Count of total number of stops". The total number of people served (i.e., number of people who got on the bus) and the total number of intermediate stops between origin and destination were identified as the state variables. They were chosen because they are the only statistically significant variables (with a probability < 0.001) that the APC system is capable of collecting at this time. The clustering technique is applied to the training dataset to identify similar groups of data samples and therefore specific sets of headway plans. The optimal number of clusters can be determined based on the cubic clustering criterion (CCC) stopping rule [11].

$$CCC = \ln\left\{\left[1-E(R^2)\right]/(1-R^2)\right\} * \left\{\left[(np/2)^{0.5}\right]/\left[0.001+E(R^2)\right]^{1.2}\right\} \quad (1)$$

Where:
R^2 – proportion of variance accounted for by the clusters
p - estimate of between cluster variation dimensionality
n – sample size

Then, a classification technique is used on the dataset and a check is made on whether or not all data points are sampled into one of the headway plans derived by clustering. The classification tool generates rule-based decision trees to apply headway plans generated by clustering depending on the values of the state variables. These decision trees are known as regression trees, and are based on continuous variables (e.g., time). In the event of misclassification of a set of new data to the classes defined already, the scheduler may be advised to redo the clustering, and develop new clusters and update the headway plans. The process of classification is repeated iteratively to see if the misclassified points fall into one of the new sets of headway plans that are derived by clustering at every repetition.

5. CASE STUDY

Based on historic data and ridership at different time periods, transit agencies set their operational headways and bus scheduling plans. A major shortcoming of the timing plans is that the attributes considered in developing a plan are subject to change as the trip generating characteristics along the route segments change [4,6,7]. The research effort reported here used a small dataset to demonstrate how the technique can help provide an effective plan by forming clusters and validating the plans against any new ridership information that may be available as time progresses. The case study utilized data for the outbound weekday trips and for a specific pattern ID.

Clustering

The statistical technique of Cluster Analysis is used to divide (n) observations into (k) groups or clusters. The members of the same cluster are more similar to each other than the members of different clusters [10]. Cluster analysis was used to identify data points with similar ridership characteristics and to establish optimum headway plans for dispatching buses. In this study, Hierarchical Clustering is done on the training dataset using the centroid method. The sample data points are grouped into clusters in sequence. The algorithm initially starts with 'n' clusters, each containing one observation. Any two clusters can be joined based on the measure of dissimilarity (d), which is the Euclidean distance between the centers of the clusters as in Equation (2).

$$d_{i,j} = \sqrt{\sum_{k=1,n} (x_i^k - y_j^k)^2} \quad (2)$$

The distance measures dissimilarity among new clusters as they get formed. The procedure continues iteratively until all observations are included in some cluster. In hierarchical clustering it is important to determine the optimal number of clusters based on “stopping rules”. Otherwise, if one requires extreme closeness of all observations within a cluster, no observations will be grouped and the final number of clusters will be ‘n’, or in the other extreme, only one cluster may be formed if one does no care about how close the observations should be to each other. The cubic clustering criterion (CCC) used here was given a value of four, and the clusters produced were as shown in Figure 1. This indicates that the transit agency should have four different operational headways. The tree structure is shown in Figure 2.

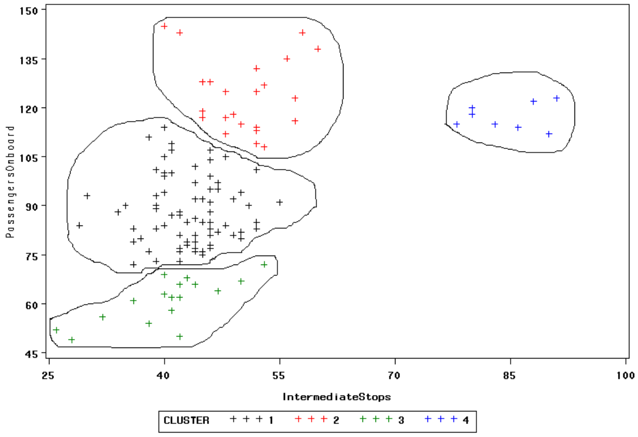

Figure 1: Four Optimal Headway Plans Result of Hierarchical Clustering

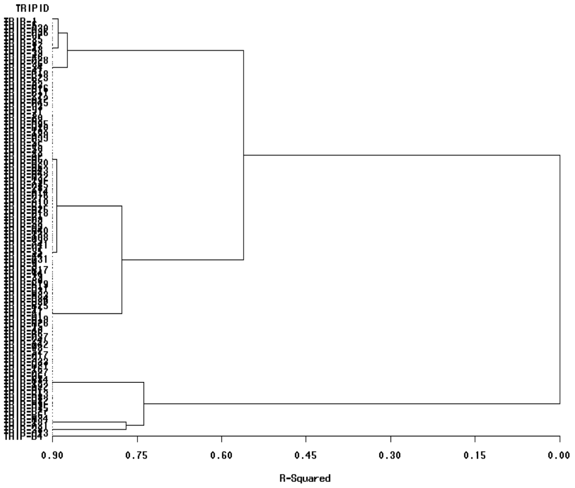

Figure 2: Tree structure of Hierarchical Clustering

Classification

Urban areas continually change, with new industrial, commercial, and residential complexes that are significant trip generators evolving over the years. Hence there is a need to monitor transit ridership demand and change the headway plans when required. This can be achieved by using clustering results in conjunction with classification systems. An urban transportation planner can check each new case of data sample collected from the field and determine if it “classifies” into the cluster that is associated with one of the optimum bus headway plans currently in use. If a number of misclassified states are obtained, the planner can conclude that conditions of ridership have reached a point where current headway plans may no longer be applicable, and that it is time to re-cluster and develop a new set of headway plans and interval break points.

The predictive modeling statistical tool “Classification and Regression Trees” (CART) was used to develop the classification rules. CART of Salford Systems Inc., is a powerful data-mining tool that searches for important patterns and relationships in the training dataset and generates optimal tree structures using the “no-stopping” rule and enforcing a binary recursive split on the data. In this study CART used 80% of the data as the learn set to train the model and the remaining 20% of the data to test the model’s validity. CART tool uses the training dataset to develop classification models. The classification results from the CART tool and the associated headway plan are shown in Figure 3. There are 9, 12, 15, and 25-minute headways according to the current operating schedule, and the R^2 is 0.75. Optimal headway plans can also be obtained by specifically transforming the ridership of the clusters using the following Equation (3) derived in a previous study [12].

$$h^* = \sqrt{\frac{2Lu_b}{VQv_u}} \qquad (3)$$

Where

h^* : Optimum headway that minimizes total cost.
L: Route Length (miles).
u_b : Bus operating cost ($/bus-hour), assumed as $50.
V: Average bus-operating speed (mph).
Q: Demand (bus-passengers/hour).
v_u : User’s value of time ($/passenger-hour), assumed as $10.

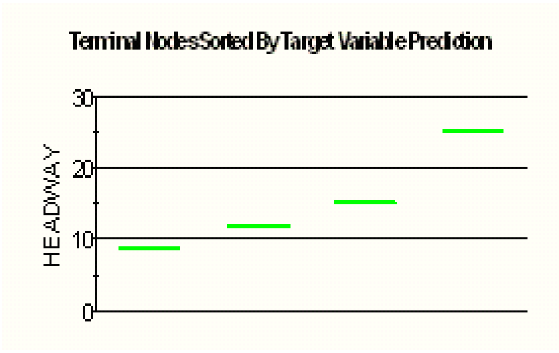

Figure 3: Headway Plans for Four Leaf Nodes

Urban public transportation planners may wish to use this data-mining tool to continually monitor any deterioration of effectiveness of the optimum operational headways and repeated violations of bus schedules, and suggest changes if needed, to ensure higher quality of passengers service. To illustrate how the tool would work if varying ridership along the route were to be recorded; it is assumed that there is a 15 percent increase in both state variables during a weekday afternoon time period. It can be seen in Table 1 that there are 8 out of 43 (18.60%) class type 15 cases (i.e. 15-minutes headway plan) that are misclassified. The four classes shown in the Table are the current scheduled headways.

Table 1: Misclassified Cases of Sample Test Dataset

Classes	Cases	Misclassified	% Error
9	18	0	0.00
12	57	0	0.00
15	43	8	18.60
25	8	0	0.00

This warrants a review of the existing scheduling plans and the development of a new set of clusters for headway plans as shown in Figure 4. The process of revalidation shows the cases that were misclassified earlier are eliminated and confirm the validity of the 5-clustered optimum operational headway plan as depicted in Figure 5. The plan consists of 9, 12, 13, 15, and 25-minute headways and the R^2 is 0.75.

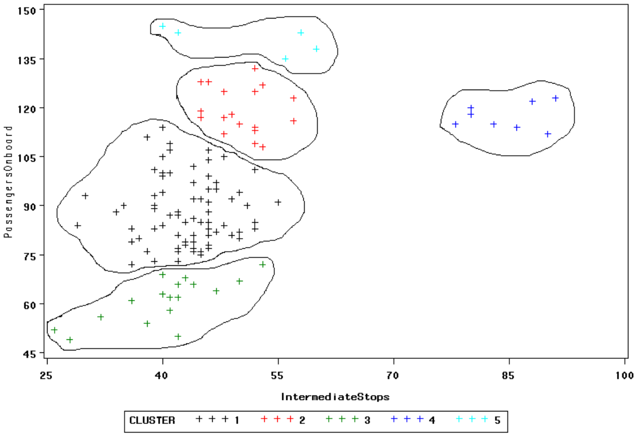

Figure 4: Five Optimal Headway Plans Result of Hierarchical Clustering

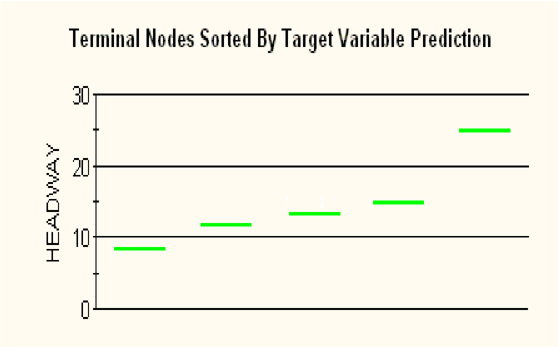

Figure 5: Headway Plans for Five Leaf Nodes

6. DECISION TREE RULES TO APPLY ON VARIOUS HEADWAY PLANS

The attributes necessary for tree formation are the state variables (Intermediate Stops and Demand/Passenger Service Rate).

The 5-cluster (9, 12, 13, 15, and 25-minute) headway plan is developed on the basis of the following rules:

If the intermediate stops are less than or equal to 30 and the demand/passenger service rate is less than or equal to 90 passengers per hour, then use a 25-minute headway plan.

If the intermediate stops are greater than 30 and the passenger service rate is less than or equal to 90 passengers per hour, then use a 12-minute headway plan.

If the intermediate stops are greater than 30 but less than or equal to 51 and the passenger service rate is greater than 90 but less than or equal to 96 passengers per hour, then use a 9-minute headway plan.

If the intermediate stops are greater than 51 and the passenger service rate is greater than 90 but less than or equal to 96 passengers per hour, then use a 15-minute headway plan.

If the intermediate stops are greater than 51 and the passenger service rate is greater than 96 passengers per hour, then use a 13-minute headway plan.

The decision tree associated with the above rules is shown in Figure 6.

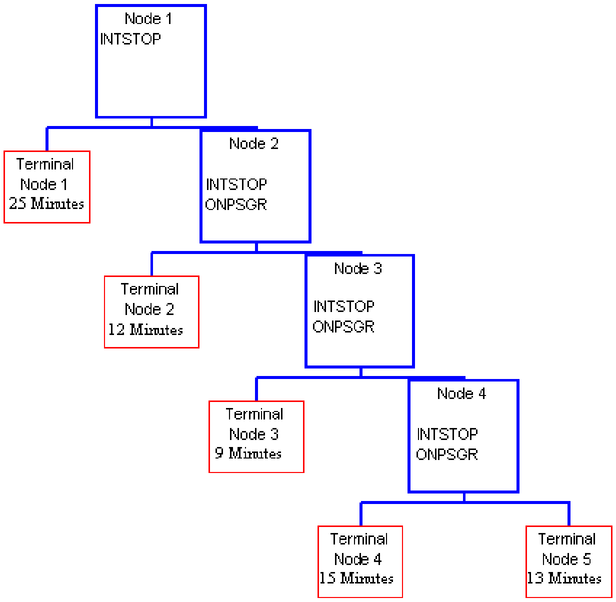
Node 1
INTSTOP
Terminal Node 1 25 Minutes
Node 2
INTSTOP
ONPSGR
Terminal Node 2 12 Minutes
Node 3
INTSTOP
ONPSGR
Terminal Node 3 9 Minutes
Node 4
INTSTOP
ONPSGR
Terminal Node 4 15 Minutes
Terminal Node 5 13 Minutes

Figure 6: Decision Tree Diagram

7. CONCLUSION

APC systems collect plentiful data that when skillfully used can identify bus ridership demand variations at various segments/links along the transit route at different times of day. Transit providers can use this information to operate buses at optimum headways, and adjust those headways periodically, as demand fluctuates over time. However, it is not easy to process the data efficiently and interpret the underlying meaning of their values without using advanced data manipulation tools. Thus, the use of data mining tools is important to interpret such vast amounts of data. Clustering and classification techniques can help planners to improve operational bus headway plans and thus, mitigate the undesirable impact of passenger-waiting times and increase the level of service offered by the transit providers. This research demonstrated that data mining tools could provide transit planners with adequate information on the behavior of the system, and enable them to monitor performance of bus transit plans under changing demand patterns.

8. FUTURE RESEARCH

Future research on the subject can take several directions. One issue to be investigated could be how often a transit agency should run the test described in this paper to try to change the already existing bus scheduling plans. Also, the issues of whether or not the scheduling plans that are suggested by the tool could be implemented in practice, and if not, how to alter them so that they could be implemented in a real world scenario. Two state variables were used in what was presented here. In the future, additional state variables, such as time-of-day or day-of-the-week may be added to or substitute the present variables. This approach of applying data mining techniques to data collected by APC for scheduling purposes is relatively new. Therefore, not many problems have been identified and solved to date. Identifying a pragmatic approach and putting it to practice is the only way to assess what real world issues might arise while implementing such systems.

9. REFERENCES

[1] F.V. Webster, and P.H. Bly, "The Demand for Public Transport", **Transport and Road Research Laboratory**, Crowthorne, Berkshire, England, 1980.

[2] T. J. Kimpel, J. G. Strathman, K. J. Dueker, D. Griffin, R. L. Gerhart, and K. Turner, "Time Point Level Analysis of Passenger Demand and Transit Service Reliability", **Report TNW**, TransNow, Seattle, WA, 2000-03.

[3] R.G. Mishalani, M. M. McCord, and J. Wirtz, "Passenger Waiting Time Perceptions at Bus Stops", **Transportation Research Board**, CD-ROM, 2005.

[4] D. M. Meyer, and E. J. Miller, **Urban Transportation Planning,** New York: McGraw Hill Inc., 2nd Edition, 2001.

[5] S. D. Maclean, and D. J. Dailey, "Wireless Internet Access to Real-Time Transit Information", **Transportation Research Record 1791**, 2002, pp. 92-98.

[6] J. Patnaik, S. Chien, A. Bladikas, "Estimation of Vehicle Arrival Times using APC data", **Journal of Public Transportation**, Vol. 7, No. 1, 2004, pp. 1-20.

[7] B. L. Smith, W. T. Scherer, and T. A. Hauser, "Data-Mining Tools for the Support of Signal-Timing Plan Development", **Transportation Research Record 1768**, 2001, pp. 141-147.

[8] S. Chien, and Y. Ding, "A Dynamic Headway Control Strategy for Transit Operations", **Conference Proceedings (CD-ROM), 6th World Congress on ITS**, 1999, Toronto, ITS Canada.

[9] W. Mendenhall, and T. Sincich, **A Second Course in Statistics: Regression Analysis**, New Jersey: Prentice Hall, 5th edition, 1996.

[10] B. Ripley, **Pattern Recognition and Neural Networks**, Cambridge, United Kingdom: Cambridge University Press, 1999.

[11] C. Milligan, "An Examination of Procedures for Determining the Number of Cluster in a Data Set", **Psychometrika**, Vol. 50, No. 2, 1985, pp. 159 – 179.

[12] S. Chien, L. Spasovic, R. Chhonkar, and E. Elefsiniotis, "Evaluation of Feeder Bus Systems with Probabilistic Time–Varying Demands and Non-Additive Time Costs", **Journal of the Transportation Research Board TRR No. 1760**, 2001, pp. 47-55.