BUS Public Transportation System Fuel Efficiency Patterns

José de Almeida, and João C Ferreira

Abstract—This paper provides an overview of a set of OLAP and Data Mining based feedback tools for management of a public BUS transportation system, providing data to continually educate drivers and network planners in regards to Fuel Usage Efficiency. Details on collection, extraction, transformation, loading and analysis of data into a data warehouse and multi-dimensional cube are provided. Data Warehouse based Data Mining Patterns discovery processes are presented.

Keywords— Data Warehouse, Data Mining, Decision Systems, Fuel Efficiency, Information Retrieval, Knowledge Discovery, Public Transportation.

I. INTRODUCTION

As per [1] the transport sector will increase its share of the global energy usage from 53% to 60% on the 2009-2035 period. This report also states road transportation to be the largest energy user sub-sector, and further shows the importance of continually providing driver's behavior feedback on best practices usage to maximize fuel efficiency in the long term adoption promotion. Furthermore, the selection, fine tuning and planning of BUS public networks should take into consideration the selection of most fuel efficient drivers and equipments to operate each route, as well as periodic route analysis to identify improvement opportunities, e.g. prioritizing the introduction of new BUS lanes. In this scope, the authors have established a Key Performance Indicators tool set, allowing measures from data collected by an event driven recording system from vehicle CAN Bus of a fleet of city buses, crossing those variables with publicly available historical meteorological data, as to answer the most fuel efficient:

- vehicle for a given route;
- driver's driving characteristics;
- meteorological conditions;
- group of drivers;
- gain driving characteristics worth changing.

The authors elected to build a decision system prototype to allow slicing and dicing by any selected dimension, to apply Data Mining techniques as to find the hidden patterns in order to allow assessment of driver's, vehicles, routes, periods of the day and meteorological conditions in the selection of the most efficient entities on any given scenario. The selected platform and methodology used were centered on Microsoft SQL Server 2008R2 with SSIS/SSAS and SEMMA.

The present work starting point is a SQL Server database of collected event driven totalizing variables, e.g., any time a driver starts and stops a voyage, an on-board recording system records the accumulated distances traveled, fuel consumption, time traveled per engine rotation band, etc. From these variables the following ratios were defined:

- Average Fuel Consumption (L/100 Km);
- Time Percentage of Engine Rotation in Green Band;
- Time Percentage of Engine Rotation in Yellow Band;
- Time Percentage of Engine Rotation in Red Band;
- Time Percentage of Engine Idle Rotation;
- Excessive Accelerator Events per 100 Km;
- Excessive Braking Events per 100 Km;
- Inertial Time Traveled Percentage;
- Time Percentage Inertial Movement;
- Brake Usage per 100 Km;
- Clutch Usage per 100 Km;
- Accelerator Usage per 100 Km.

Additionally historical meteorological data was obtained from Weather Underground, assuming a unique point as representative of the city's regions. Collected variables examples are Mean Sea Level Pressure, Temperature, Visibility, Wind Speed, Rainfall, and weather Events (storm, heavy rain, etc...).

II. RELATED WORK

Discovering patterns in sequences is an important knowledge discovery and data-mining research area. CAN Bus generate a huge date about engine events and the interpretation from raw data of real human drivers has not yet been tackled and will be an area of future research [2]. Driving style, seen as "the attitude, orientation and way of thinking for daily driving", based on questionnaire's' surveys [3]-[4]. More recent works use a virtual driving simulator to collect realistic driving data from human drivers and to model human driving behavior [5], or classify driving style by combining objective rank method [6].

Within the context of energy efficiency, much has been achieved on the level of performance of engines and vehicles, obtaining substantial improvement and energy savings. Little...
attention has been focused on driving quality and on methods for continuously promoting energy-efficient driving. This is due in part to the difficulty of objectively evaluating driver's performance.

The driver controls the speed and the position of the vehicle on the street, in an environment characterized by determined conditions of traffic, itinerary, load, weather conditions, amongst others parameters. In controlling, the driver acts directly not only on the speed and position of the vehicle but also on parameters such as acceleration, deceleration, engine speed rotation and gear engaged [7]-[8]. The mode in which the driver acts and controls these parameters in relation to the environment determines his or her driving style. Different driving styles result in different instantaneous and average fuel consumption and more generally determines driving quality. However, the environment also exerts an influence on consumption, interacting with and conditioning the driver's decisions.

II. THE PROTOTYPE

The typical Extraction, Transformation and Loading of the Data Warehouse was achieved by a SQL SSIS package to load a star shaped data repository as shown on Fig. 1.

![Fig. 1 ETL SSIS package.](image)

A. Data Warehouse

The developed Data Warehouse follows the star shape design principle proposed by [9]

B. Grain Declaration

The analysis will be conducted per Driver, Route, Bus, Date and Time. Historical meteorological data is used as a characteristic of the day and a single meteorological data collection station is assumed to be representative of the geographic area covered.

C. Analysis Dimensions

The Analysis dimensions were derived from the grain declaration and as so the authors implemented dimensions for:
- Driver;
- Bus;
- Route;
- Date;
- Time.

Metrics

The metrics collected are to be grouped by its origin in two classes:
- Operational (and collected) data;
- Meteorological data.

Operational variables derived ratios were implemented as cube calculated members, to always obtain ratio of the sums instead of the typical trap of obtaining the sum of ratios, as data is being rolled up, down or sliced.

Fig. 2 presents basic statistics for the obtained data consisting of 397,261 facts.

![Fig. 2 Data Warehouse fact variables basic statistics.](image)

D. Typical OLAP Questions

What is the most efficient Bus for a given Route? Who is the most efficient driver for a Route or Bus?

E. Typical Data Mining Questions

What are the fuel average consumption key influences? Whom and how to educate first? What routes to analyze?

IV. RESULTS

By slicing and dicing with the multidimensional cube, the typical OLAP questions are readily answered, e.g. the answer to the question “What are the most efficient vehicles to operate a given route?”. As sown by Fig. 3 it is rather preferable to select bus 886 for route 1021 than to use it on route 1024.

The major objective was the analysis of fuel consumption (in liters per 100Km) per driver, per bus, per route, per day and per time. By discretization of the continuous Average Fuel Consumption variable, imposing it the DISCETIZED SSAS type in the data mining structure (using the automatic discretization functionality that tries first an equal areas
approach and then an Expectation Maximization Clustering approach), five consumption classes were created:

1. Below 44.74l;
2. From 44.74l to 55.73l;
3. From 55.73l to 67.025l;
4. From 67.025l to 79.53l;
5. Above 79.53l.

Using Naive-Bayes Data Mining models, relationships between variables from the perspective (case) of the Driver, the Bus and the Route were assessed.

A. Driver Case Data Mining

Taking into account the lower consumption class (below 44.74 liters per 100Km), the key influencing factors on fuel efficiency for the driver case are shown in Fig. 4. From this analysis, the top four key influences highlighted are the utilization of clutch, the observation of optimal engine rotation, the minimum engine idling and inertial movement.

It is similarly trivial to determine which driver behaviors to promote, as to maximize fuel efficiency, e.g. to educate a "class (2) efficient driver into a class (1) driver", emphasis should be placed in the importance of optimal clutch, engine rotation and averting engine running in idle, as shown by Fig. 5.

From the data available and contrary to common believe, no evidence of meteorological conditions influence on the fuel efficiency arising from driver's driving style adaptation was found, as shown by Fig. 6.

B. Route Case Data Mining

In this case, the analysis showed an influence of meteorological events (like Thunderstorms or Heavy Rain) on the Average Fuel Consumption of some routes, as demonstrated on Fig. 7.

C. Bus Case Data Mining

This analysis, as shown by Fig. 8, presents only a relatively weak link of accelerator usage to fuel efficiency as being the most influencing factor of altering a Bus fuel efficiency.

D. Prediction of Driver Average Fuel Class

As an early attempt of extending the application of the implemented Data Mining models to prediction tasks, the Naive-Bayes models were trained with ten percent of the available data to investigate their performance. Fig. 9 and Fig 10 respectively show the lift chart and the classification matrices of predicting Average Fuel Consumption.

V. CONCLUSION

Implementing a Data Warehouse is instrumental to allow key business users better decisions and actuation upon the daily running of bus fleet. As a by-product of the ETL data warehouse activities, the application of data mining models to describe data and detect key influencing factors on fuel efficiency is facilitated and allows better insights from the collected data.

The principal acquisition of knowledge was derived from the case of the Driver. This reinforces the prototype derived insights value on the driver's continuous education and motivation cycle to long term fuel efficiency maximization, contributing to Eco-responsible driving behaviors adoption retention.

Meteorological conditions are more linked to specific routes than to drivers behavior adaptation to environmental conditions, thus suggesting the need of improving affected
routes road quality.

The prototype built although using data collected from a fleet of buses can be applied to other types of vehicles, like privately owned cars, provided with similar data collection systems.

The predictive accuracy and capability of the models built is still to be optimized and compared with other Data Mining algorithms to be applied to the same mining structures, to improve prediction accuracy, such as classification and regression trees, KNN Clustering and Neural Networks as to overcome the known frailties of linear classifiers.

**ACKNOWLEDGMENT**

The first author thanks the support of Professor Nuno Datia.

**REFERENCES**


José de Almeida was born in Lisbon, in August 1968. He is a student on the Master Degree on Informatics Engineering at Instituto Superior de Engenharia de Lisboa.

João Ferreira was born in Angola, in June 1967. He is Professor of Mathematics on Informatics and Telecommunication degree courses in Polytechnic Institute of Lisbon (IPL/ISEL). He is also consultant with different companies and Portuguese institutions. He graduated in Physics at Technical University of Lisbon (UTL/IST), Portugal, received an MSC in Telecommunication and a PhD degree in Computer Science Engineering from UTL/IST. Currently, he teaches subjects in the area of mathematics and information systems and he supervises several year end, MSc and PhD students. His professional and research interests are in retrieval information, data mining, electric vehicle, intelligent transportation (ITS) and sustainable mobility systems. As well as UML, MDE Approach, automobile business process improvements and CAD systems. He is author or co-author of more than 100 peer-reviewed scientific papers of several international conferences and workshops in different areas of computer science.
Fig. 6 Driver case links to Average Fuel Consumption.

Fig. 7 Route case links to Average Fuel Consumption.