Information Sciences 293 (2015) 299–313

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Validating the coverage of bus schedules: A Machine Learning approach

João Mendes-Moreira [a,b], Luís Moreira-Matias [c,a,b,*], João Gama [a,d], Jorge Freire de Sousa [e,f]

[a] LIAAD-INESC TEC, 4200-465 Porto, Portugal
[b] Dep. de Eng. Informática, Fac. de Engenharia, U. Porto, 4200-465 Porto, Portugal
[c] Instituto de Telecomunicações, 4200-465 Porto, Portugal
[d] Faculdade de Economia, U. Porto, 4200-465 Porto, Portugal
[e] UGEI-INESC TEC, 4200-465 Porto, Portugal
[f] DGEI, Fac. de Engenharia, U. Porto, 4200-465 Porto, Portugal

ARTICLE INFO

Article history:
Received 4 April 2013
Received in revised form 14 August 2014
Accepted 7 September 2014
Available online 27 September 2014

Keywords:
Public transportation
Operational planning
Schedule Coverage validation
Clustering
Consensual cluster
Rule induction

ABSTRACT

Nowadays, every public transportation company uses Automatic Vehicle Location (AVL) systems to track the services provided by each vehicle. Such information can be used to improve operational planning. This paper describes an AVL-based evaluation framework to test whether the actual Schedule Plan fits, in terms of days covered by each schedule, the network's operational conditions.

Firstly, clustering is employed to group days with similar profiles in terms of travel times (this is done for each different route). Secondly, consensus clustering is used to obtain a unique set of clusters for all routes. Finally, a set of rules about the groups content is drawn based on appropriate decision variables. Each group will correspond to a different schedule and the rules identify the days covered by each schedule.

This methodology is simultaneously an evaluator of the schedules that are offered by the company (regarding its coverage) and an advisor on possible changes to such offer. It was tested by using data collected for one year in a company running in Porto, Portugal. The results are sound.

The main contribution of this paper is that it proposes a way to combine Machine Learning techniques to add a novel dimension to the Schedule Plan evaluation methods: the day coverage. Such approach meets no parallel in the current literature.

1. Introduction

The bus has become a key player in highly populated urban areas. Inner-city transportation networks are becoming larger, and thus good operational planning is crucial. However, many vehicles are still failing to meet their schedules even when the defined plan is suitable. The *arriving as planned* factor is the most important achievement of both transit planners and passengers [21,27].

Throughout the 1990s, many mass transit companies started to install new computer-aided Bus Dispatch Systems. A vital part of these systems are the Automatic Vehicle Location (AVL) technologies. The installation consisted of equipping

* Corresponding author at: Dep. de Eng. Informática, Fac. de Engenharia, U. Porto, 4200-465 Porto, Portugal. Tel.: +351 22 508 2142.
E-mail address: lmatias@fe.up.pt (L. Moreira-Matias).

http://dx.doi.org/10.1016/j.ins.2014.09.005
0020-0255/

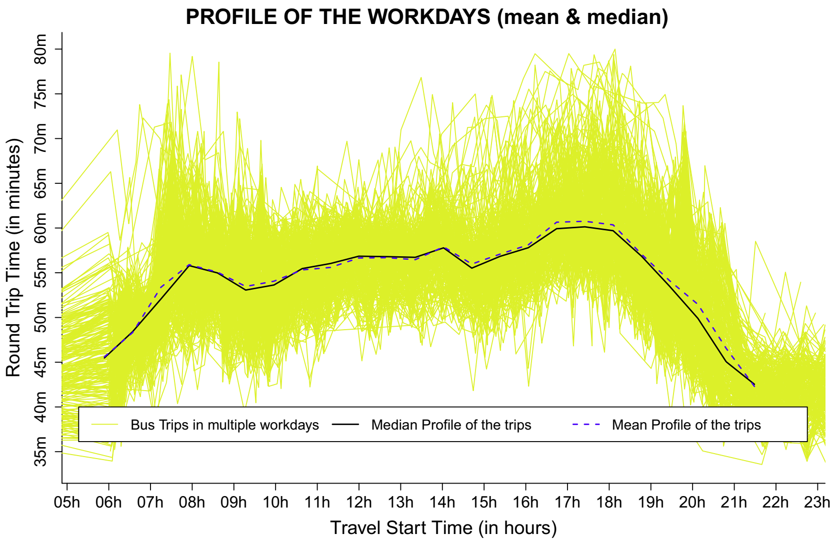

Fig. 1. Daily profiles of the behavior of a given route on the working days during a one year-period. The black line represents the median of those profiles while the blue one represent the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fleets with GPS-based communication systems capable of transmitting information on the **status** and the **location** of each vehicle to a data server with a certain but short periodicity. Many researchers understood the hidden potential of the stored AVL data to provide insights on how to evaluate (and then improve) the Operational Planning (OP) [26,12,27,2]. As a consequence, many research projects to collect and mine this massive amount of data arose in many cities around the world. Some examples include New Jersey, Chicago, Minneapolis, Seattle (United States); Ottawa and Montreal (Canada) or Eindhoven and The Hague (Netherlands). Promising insights on those case studies were reported in the survey presented by Furth et al. [10]. This paper focuses on improving the operational planning by mining AVL historical data.

Typically, a Schedule Planning process for a given route relies on three distinct steps: (1) the first step is defining the number k of schedules and their individual coverage, C_i. Secondly, (2) the *schedule time points* are chosen among all the route bus stops, and finally, (3) timetables t_i for each route schedule S_i are defined containing the time the buses pass at each schedule time point (per trip).[1] This process is done for all routes. It should be guaranteed that he number k of schedules and the coverage C_i are the same for all routes to facilitate the SP memorization by the passengers.

1.1. Literature review

The traditional Schedule Plan (SP) evaluation is highly focused on the time-tabling task. Specifically, this evaluation uses three key indicators: (1) on-time arrival performance, (2) headway adherence and (3) cruising time adherence [21,29,2]. A statistical framework to evaluate the schedule reliability levels – such as the running time adherence to the scheduled time or the headway regularity – is presented in Lin et al. [14]. In Patnaik et al. [23], the authors proposed a method to evaluate the defined timetable by clustering AVL data. The trips within the same cluster should have the same headway regularity defined throughout the day. One of the most common planning-related problems is the Travel Time Prediction (TTP) (or Arrival Time Estimation) and relevant research has been conducted to solve this problem [1,6,4,22]. Frameworks to propose changes on the timetables are using such TTP models. A good example is the work proposed by El-Geneidy et al. [9] – it uses multiple regression models to measure the variation between the scheduled times and (1) the actual headway and/or (2) the *round-trip times* (that is, bus cruising time, bus travel time, time elapsed between a vehicle departure and its arrival at the destination stop). The predictive models were employed to (re) define the round-trip times in the schedule points along the route. Studies focused on headway deviation effects – such as Bus Bunching[2] – were also conducted to evaluate schedule reliability. The research presented in Moreira-Matias et al. [20] introduces a novel framework to find frequent headway deviation sequences which explain this phenomenon. Despite the framework's relevance to evaluate schedule reliability, it is clear that every timetable definition is largely **dependent** on the k number of existing schedules, and moreover, on **its day coverage** C_i. However, at the best of our knowledge, there

[1] The step 2 just changes the way the schedules are presented to the public. It does not change neither the definition of trips nor the assignment of duties.

[2] This phenomenon occurs when two or more buses of the same line are running in a *platoon* or with a very short headway [20,15,5].