Transportation Research Part C 93 (2018) 211–227

Contents lists available at ScienceDirect

Transportation Research Part C

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

Prediction of arrival times of freight traffic on US railroads using support vector regression

William Barbour[a,c], Juan Carlos Martinez Mori[a], Shankara Kuppa[b], Daniel B. Work[a,c,*]

[a] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave, Urbana, IL 61801, USA
[b] CSX Transportation, Jacksonville, FL 32202, USA
[c] Department of Civil and Environmental Engineering & Institute for Software Integrated Systems, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37235-1826, USA

ARTICLE INFO

Keywords:
Freight rail
Machine learning
Data driven estimation

ABSTRACT

Variability of travel times on the United States freight rail network is high due to large network demands relative to infrastructure capacity, especially when traffic is heterogeneous. Variable runtimes pose significant operational challenges if the nature of runtime variability is not predictable. To address this issue, this article proposes a data-driven approach to predict *estimated times of arrival* (ETAs) of individual freight trains, based on the properties of the train, the properties of the network, and the properties of potentially conflicting traffic on the network. The ETA prediction problem from an origin to a destination is posed as a machine learning regression problem and solved using support vector regression trained and cross validated on over two years of detailed historical data for a 140 mile section of track located primarily in Tennessee, USA. The article presents the data used in this problem and details on feature engineering and construction for predictions made across the full route. It also highlights findings on the dominant sources of runtime variability and the most predictive factors for ETA. Improvement results for ETA exceed 21% over a baseline prediction method at some locations and average 14% across the study area.

1. Introduction

1.1. Motivation

The rail network in the United States has significant infrastructure capacity limitations that cause congestion of the rail traffic. Few rail corridors contain exclusively double (or more) track that allows simultaneous bi-directional traffic (Murali et al., 2010). In comparison, the double and triple track railroads in Europe provide for double the train density of US rail networks (Wyman, 2016). Many US corridors contain a single track with short sections of double track known as *sidings*, where trains may meet or pass each other. These *movements* (i.e., meets, passes) are implemented in the railroad signaling system, but are directed by human dispatchers. Dispatchers are experienced with working on specific track corridors, but movements on sidings require planning and precise timing in order to achieve efficient operations (Vromans et al., 2006; Kecman and Goverde, 2013). Freight volume is expected to increase in the US, so either infrastructure capacity must be increased or operational improvements must be made to increase capacity (Systematics, 2007; Weatherford et al., 2008; Association, 2013).

In addition to the track infrastructure constraints, there are numerous other factors that can contribute to variability of the

* Corresponding author at: Department of Civil and Environmental Engineering & Institute for Software Integrated Systems, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37235-1826, USA.
E-mail address: dan.work@vanderbilt.edu (D.B. Work).

https://doi.org/10.1016/j.trc.2018.05.019
Received 28 June 2017; Received in revised form 24 April 2018; Accepted 20 May 2018

runtime on a track segment. Traffic heterogeneity and the train priority differences directly influence both the runtime of trains and also the variability in the runtime (Dingler et al., 2009, 2010). Physical characteristics of trains such as the length, tonnage, and power further influence the runtime due to track grade, track curvature, and siding lengths (Dingler et al., 2009). The ability of a train to complete a trip and exit the *line of road* (i.e., the track segments connecting distant terminals) is also influenced by the degree of congestion in the arrival terminal. This is compounded by the possible actions required for the train in the terminal, such as refueling, inspection, switching of cars, or crew change (Dingler et al., 2009; Higgins et al., 1995). Railroad operating strategies such as dynamically scheduled trains and maximizing train length are particularly vulnerable to delay (Lu et al., 2004; Mu and Dessouky, 2011).

In the presence of runtime variability, ETAs are necessary in order to improve real-time decision making and the efficiency of the network (Hertenstein and Kaplan, 1991; Hallowell and Harker, 1998). For example, future train schedules can be continually updated to provide new train plans to allow traffic to flow smoothly between terminals on the network Kraay and Harker (1995). Although there are many techniques available to derive optimal schedules (see Goverde, 2005 for a thorough review), the schedule may be very sensitive to delays when the network is near capacity. High capacity utilization leads to more complex dispatching where small delays are created, leading to larger deviations from the train plan (Khoshniyat and Peterson, 2017); this is referred to as *knock-on delay* (Vromans et al., 2006; Murali et al., 2010; Goverde and Meng, 2011).

Highly variable runtimes increase operational uncertainty for the railroad and for other transportation systems that depend on them. On the rail network, propagation of delay to other trains is significant (D'Ariano and Pranzo, 2009), and there are large direct costs incurred due to additional operating time alone (Lovett et al., 2015). Delays on the rail network can also influence non-rail transportation services. For example, surface street traffic and emergency vehicles, which conflict with rail freight traffic at grade crossings (Estes and Rilett, 2000), can be significantly delayed if a grade crossing is occupied by a train for an extended period of time. If accurate, real-time ETAs are made available, revisions to the operating plan can be implemented, and surface street transportation services can be re-routed.

1.2. Problem statement and related work

The main focus of the present article is the prediction problem for ETAs on US freight railroads using real-time data. The estimation problem requires new ETAs to be produced as time elapses and the train progresses down the line of road. Each time the train reaches one of a number of fixed locations on the track, data is collected and a new estimated travel time to the destination is produced.

To produce the ETA estimate, a variety of routinely collected and maintained data sources available to freight railroads are used. This includes track geometry data (containing grade and curvature information, single and multi-track territory, length of sidings, etc.), historical runtime profiles of all trains, properties of all trains (such as length and tonnage), and crew records.

Several methodologies to produce ETAs are available, including microscopic simulation (Petersen and Taylor, 1982; Şahin, 1999; Marinov and Viegas, 2011), analytical approaches (Assad, 1980), and data-driven techniques (Bonsra and Harbolovic, 2012; Wang and Work, 2015). Due to the complexity of the freight rail network (which limits the accuracy of analytical abstractions) and the difficulty to capture all delay inducing factors in a simulation based model (e.g., decisions made by human dispatchers, special cases involving priority elevation, unplanned maintenance, and weather), a data-driven approach is proposed in this article (Li et al., 2014). This approach is made possible through access to a large and comprehensive freight rail dataset also described in the article. Well designed data-driven techniques are able to generalize to similar but unseen scenarios to those represented in the training dataset, making them useful for prediction of ETAs during typical operations (Marković et al., 2015). Note however that the methods may not extrapolate well to rare and extreme events such as heavy network disruptions, especially when few or no examples exist in the training data.

Many ETA prediction methods for buses and cars also rely on data-driven algorithms (Altinkaya and Zontul, 2013; Mori et al., 2015) similar to those discussed for freight rail. There are, however, fundamental differences in operations between buses and cars, and the rail freight traffic considered in our work. Bus operations are characterized by frequent stops where delays occur due to the passenger boarding and alighting process (Chien et al., 2002). Buses are also delayed en-route between stations due to traffic signals and other vehicular traffic, which are delay factors for cars as well. Importantly, in the bus system, the vehicular traffic represents an external disturbance to the system. Finally, buses and cars are generally physically homogeneous (e.g., they have similar performance characteristics and consequently similar dynamics) to other buses and cars, respectively. These properties are in contrast to freight rail traffic, where the trains are quite heterogeneous with respect to tonnage, power, length, and priority, all of which affect centralized dispatching decisions and, ultimately, delays.

Several lines of research are related to the problem of freight rail ETA prediction. We briefly summarize the most closely related works, and direct the interested reader to the comprehensive reviews available in the works by Bonsra and Harbolovic (2012) and Gorman (2009). The majority of freight trains operate according to a schedule that is constructed in an offline manner and robust to some random unplanned disturbances (Mu and Dessouky, 2011; Vromans et al., 2006; Khoshniyat and Peterson, 2017). When extreme disturbances cause the original schedule to deteriorate, online rescheduling measures must be implemented to account for the delay and to maintain robustness to further delay (D'Ariano et al., 2007; Hallowell and Harker, 1998; Khadilkar et al., 2017). Numerous efforts are aimed at understanding and quantifying the causes of delay that influence scheduling, rescheduling, and predictability (Murali et al., 2010; Chen and Harker, 1990; Dingler et al., 2010). Delay is typically formulated in terms of deviation from a train schedule or historical performance, but it can be extended to arrival time prediction for individual trains (Bonsra and Harbolovic, 2012).