Supporting automated operations with improved arrival time predictions on US freight railroads

William Barbour*[1], Shankara Kuppa[2], and Daniel Work[1]

[1]Dept. of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, USA.
[2]CSX Transportation, Jacksonville, FL 32202, USA.

1 Research Objective

Freight rail traffic in the United States is operating on an increasingly congested rail network, and the freight volume is expected to continue to increase in the foreseeable future [1, 2]. Additional trains on the network influence each other due to infrastructure capacity limitations that are pervasive in the US freight rail network. In single track territory, conflicting and overtaking trains may pass only at short (i.e., on the order of the train length) segments of double track known as *sidings*, which require precise scheduling by human dispatchers during high traffic volumes if delays are to be avoided [3, 4].

The high variability of runtimes is detrimental both to the railroads and to other interconnected transportation systems. The magnitude of delay for non-priority freight rail can be on the order of hours and propagation of delay to other trains is significant [5]. Real-time revisions to the operating plan can be implemented, which are currently performed by humans but will soon be understood and handled by an automated system.

The primary focus of this work is to address the problem of ETA prediction on freight railroads in the US. Specifically, we explore the ETA prediction problem for individual freight trains in an *online* setting, where ETAs are continuously produced as new information becomes available. These efforts will serve to forward the goal for the railroads of automating operational decisions and, eventually, actual train and freight movement.

2 Methodology

To produce the ETA estimate, a variety of routinely collected and maintained data sources available to freight railroads are used. This work uses a series of datasets describing the rail network and operations from January 1, 2015 through December 31, 2015, inclusive. It

*Presenting and corresponding author. Contact: wwbarbo2@illinois.edu, +1 423 429 8078

consists of freight train movement, yard work, crew, and locomotive data in the CSX Transportation Nashville division, extracted from dispatching and mainframe data; the territory is primarily located in Tennessee, USA. This is supplemented by track geometry data detailing grade and curvature information, single and multi-track territory, length of sidings, and other attributes.

Several methodologies to produce ETAs are available, including microscopic simulation, analytical approaches, and data-driven techniques. 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 work. Several works, such as Kecman and Goverde [4] and Wang and Work [6], have proposed to empirically produce delay or runtime estimates using historical data for passenger rail networks. The most closely related estimation works on freight trains are the works of Gorman [7] and Bonsra and Harbolovic [8].

In this work, the problem of predicting an estimated time of arrival for a train from an origin point to a destination point on the rail network is posed as a supervised machine learning regression problem. The goal of the regression problem is to predict the true runtime $y(i) \in \mathbb{R}^1$ of a train i given the properties of train i, the network, and other traffic on the network, which are contained in the feature vector $x(i) \in \mathbb{R}^n$.

The central difficulty of posing the online ETA prediction problem into the standard machine learning framework above stems from the fact that many of the features used for prediction change in time and in space as the train moves towards the destination. If a single model is used for all origin-destination predictions, it may be difficult to predict area-specific delays (e.g., due to local dispatching decisions or route characteristics) that may not occur throughout the network.

To address these difficulties, we propose to build a distinct regression model for each origin-destination pair for which predictions are required. Because the models are independent, each model can be trained using all trips that pass between the corresponding origin-destination pair by constructing features according to the state of the train and network at the time the train reaches the origin node. Localized and geography-specific performance characteristics may be captured in the individual models without explicitly constructing them in the feature vector. The regression problem of predicting ETAs from a vector of features is solved with a *support vector regression* (SVR) machine.

3 Results and Future Work

Performance of each model is compared to that of a historical median predictor by *mean average error* (MAE) for each of the 41 prediction sites on the territory under study. In general, the improvement of the SVRs compared to the baseline decreases as the origin point becomes closer to the destination point. Across the 41 origin-destination predictions, all SVR-based algorithms show an improvement over the historical median baseline. The best performing algorithm is the SVR model combined with a non-linear *radial basis function kernel* and using the full feature set, which achieves a 7% average improvement on the MAE relative to the historical median benchmark. The inclusion of features to quantify the traffic

on the line of road resulted in the best performance, and suggests additional improvements are likely possible with further refinements on the congestion measures leading to meets, passes, and siding utilization.

In the process of investigating the modest improvement of the SVR algorithms over the naive predictors, a dominant source of runtime variability was discovered that overshadows the predictive improvements achieved by the models. Specifically, it was discovered that *recrewed* trains (i.e., a train that did not reach its destination before the crew reached its maximum on-duty time and needed a relief crew) define the dominant source of variability of runtimes on the track segment.

To further investigate the impact of recrews on train variability, all trains were ex post facto labeled as either recrewed or non-recrewed. Less than 10% of the trains on the route were recrewed. Next, the two classes (recrewed and non-recrewed trains) were separated and descriptive statistics were calculated for each class on each of the 41 origin-destination pairs. The standard deviation of runtimes was used to quantify the runtime variability of trains in each class as well as the variability of all trains in the dataset (not separated on recrew). The runtime variability of the recrewed trains is several times larger than that of the non-recrewed trains across all origin-destination pairs. Even though the recrewed trains represent less than 10% of the trips, they represent 53% of the variance within the dataset of all trains, when averaged across the full route. The single fact that a train was recrewed explains more variability in the runtime than all of the other features we explored. Due to the large variance caused by recrews, we are interested to develop a data-driven classifier to preemptively and automatically classify trips that are likely to be recrewed in future work.

References

[1] Brian A Weatherford, Henry H Willis, David S Ortiz, Louis T Mariano, J Enrique Froemel, and Sara A Daly. *The State of US Railroads: A Review of Capacity and Performance Data*. Rand Corporation, 2008.

[2] Association of American Railroads. Class I railroad statistics, February 2013.

[3] Michiel J.C.M. Vromans, Rommert Dekker, and Leo G Kroon. Reliability and heterogeneity of railway services. *European Journal of Operational Research*, 172(2):647–665, 2006.

[4] Pavle Kecman and Rob MP Goverde. An online railway traffic prediction model. In *RailCopenhagen2013: 5th International Conference on Railway Operations Modelling and Analysis, Copenhagen, Denmark, 13-15 May 2013*. International Association of Railway Operations Research (IAROR), 2013.

[5] Andrea D'Ariano and Marco Pranzo. An advanced real-time train dispatching system for minimizing the propagation of delays in a dispatching area under severe disturbances. *Networks and Spatial Economics*, 9(1):63–84, 2009.

[6] Ren Wang and Daniel B Work. Data driven approaches for passenger train delay estimation. In *IEEE 18th International Conference on Intelligent Transportation Systems*, pages 535–540. IEEE, 2015.

[7] Michael F Gorman. Statistical estimation of railroad congestion delay. *Transportation Research Part E: Logistics and Transportation Review*, 45(3):446–456, 2009.

[8] Kunal Baldev Bonsra and Joseph Harbolovic. Estimation of run times in a freight rail transportation network. Master's thesis, Massachusetts Institute of Technology, 2012.