Optimization methods for analysis of empirical rail dispatching relative to train plans

William Barbour*[1] and Daniel B. Work[1]

[1]Department of Civil and Environmental Engineering, Vanderbilt University, Institute for Software Integrated Systems; 1025 16th Ave S, Suite 102; Nashville, TN 37212; United States

June 2021

Abstract

A major factor in railroad operational efficiency and punctuality is the quality of train planning and dispatching. Schedules or dispatching plans may also not be actualized for a variety of reasons. This work proposes a methodological tool set, called the dispatch analysis problem, that can analyze recent, empirical train dispatching data against an optimal dispatching plan. A multitude of questions can be answered using the dispatch analysis methods and we address three: 1) At what times did dispatching actions reduce the optimality of future replanning? 2) What corrective actions could have mitigated the negative impacts of past dispatching actions? 3) Which trains introduced secondary effects to other train plans? We demonstrate the application of the dispatch analysis methods to these questions using illustrative case studies from a North American freight railroad and find: specific periods of time can be isolated that demonstrate a significant deterioration in replanning ability; small modifications to past actions are identified that could improve replanning outcomes; certain trains can exhibit small delays that lead to large secondary consequences for neighboring trains. The results of the case study demonstrate the types of actionable findings on real railroad data that are possible with the dispatch analysis methods.

1 Introduction

1.1 Motivation

The challenges of rail network congestion and efficiency motivate the need to answer critical questions about how trains are dispatched and how dispatching decisions evolve the railroad state through time. This analysis on railroads

*william.w.barbour@vanderbilt.edu

during operation is made possible by increasing availability of railroad data. The trends toward better automation and forecasting are continuing, but delays and deviations from the operating plan will persist due to realities that include weather, mechanical failures, and train heterogeneity. There remains the need to improve schedules and dispatching to reduce sources of delay and the variability they cause. Particularly in single-track territories, maintaining a schedule or operating plan during dispatching is difficult with capacity pressure on the network.

Many prior works have addressed specific questions about railroad operational practices such as propagation of train delay, impact of disturbances, robust scheduling, and replanning in the presence of delays and disturbances, to name a few (Hansen, Goverde and van der Meer, 2010; Milinković, Marković, Vesković, Ivić and Pavlović, 2013; Lusby, Larsen and Bull, 2018; Fang, Yang and Yao, 2015; Boroun, Ramezani, Vasheghani Farahani, Hassannayebi, Abolmaali and Shakibayifar, 2020). Overall, optimization is most commonly used in deriving a train schedule, deriving a detailed train movement plan, and during online replanning.

Some analyses of rail dispatching can be performed using micro-simulation (Dick and Mussanov, 2016; Mussanov, Nishio and Dick, 2017). A simulation environment can emulate dispatching and train movements given a network state and schedule, even incorporating random delay and robustness in some cases. Simulation, by comparison to optimization, is used more in the analytical sense by spending less computational time on schedule and dispatching refinement and more computational time on detailed train dynamics, yard operations, and track conditions. Some of the limitations or challenges with simulation-based analysis are optimality conditions, scenario exploration, and cost quantification. Most simulation environments do not run a routine that guarantees global optimization and, therefore, can not compare to an optimal baseline and the minimum cost associated with the optimal scenario.

The aim of our work is to provide optimization-based methods for analytics. Since optimization is key to scheduling and dispatching before and during the train's run, it is a natural means by which to evaluate post-facto the empirical train movements in order to determine the causes and possible remedies of delays and operational constraints. But to date it has not been used as such.

Deploying tools to analyze specific dispatching and scheduling practices has the potential to reveal more detailed findings. Certain periods of time, such as particularly problematic dispatching scenarios or instances when operations ran better than normal, are useful to analyze in detail. What went right or what went wrong can be revealed and this insight can inform future strategies.

Consider the following examples that illustrate the possible impact of a small train schedule deviation. The difference between the train interactions is only in the length of the trains. In Figure 1, two short trains traveling in opposite directions are scheduled to meet on the middle siding. A small delay (red trajectory segment) for train 2 forces replanning of this meet event, which can instead occur on an alternate siding and mitigate the impact of the disruption to train 1. The space-time trajectories of the trains are represented by the blue

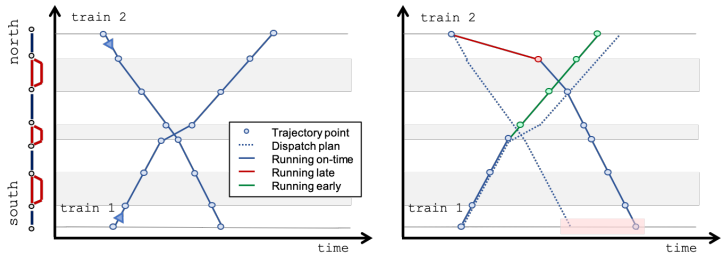


Figure 1: **Minor delay, no secondary impact.** Space-time plot of a hypothetical train meet event, with original plan shown on the left and the resulting train trajectories on the right. A small delay for train 2, shown by the red trajectory section, necessitates replanning. The two trains are able to replan their meet location with low impact to overall runtime.

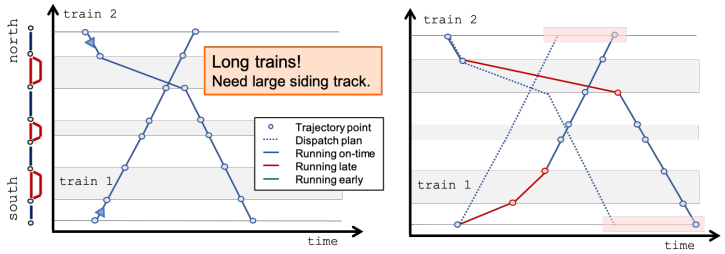


Figure 2: **Minor delay, plus impact on other train.** Space-time plot of a meet event similar to Figure 1, but with longer trains that may not use the middle siding. The original plan is shown on the left and the resulting train trajectories on the right. An initial delay for either train results in a delay for the other train, because the meet event cannot be replanned to the middle siding.

lines and points and the siding tracks are shown by the grey shaded areas. The same interaction, if the length of the trains prohibits them from using the middle siding, is shown in Figure 2. In this case, a small delay for either train would result in a delay for the other because the meet event can not be replanned for the middle siding.

1.2 Problem statement

In this article, we introduce an optimization methodology that performs analysis on empirical dispatching data. We use the term *dispatching* to mean the direction and supervision of real-time train movements and interactions. Real-time dispatching is commonly based on a pre-conceived train plan or schedule, which in many cases is generated in whole or part from optimization. We re-

fer to *optimal train planning* or *scheduling* as the pre-dispatch creation of a time-space train operating plan. The methods in our work analyze post-facto empirical train operations (how the trains actually ran) in the context of the original schedule, be that a schedule generated by an optimization model or a custom schedule used for a given period.

The methods we develop allow one to analyze the decisions that occurred in practice and evaluate how consequential these decisions were. Small train delays and sub-optimal decisions on the railroad are virtually inevitable, and this methodology solves a set of problems that reveal these actions and their effects on the short-term dispatch plan as a whole. We refer to this method as the *dispatch analysis problem* because it can be used to answer a broad set of empirical dispatch analysis questions. In this work we define and address three such problems.

Problem 1: impact of dispatch decisions.

The first question we answer is the following: What is the overall dispatching cost, given the current network state? What events occurred in the evolution of the network state that diminished the ability to continue running the optimal dispatch plan? And how costly is the current network state in terms of its impact on the future best dispatching performance, assuming optimal replanning?

To answer this question, we solve an optimal replanning problem across a period of dispatch time. The beginning of the period is progressed according to the empirical data and we optimally replan for the remainder of the period. The difference in performance between optimal replanning after some empirical decisions (i.e., trains have moved beyond the initial network state) versus optimal planning from the beginning is referred to as the *empirical optimality gap*.

Problem 2: alterations to dispatch decisions.

Given a negative effect that the current (sub-optimal) network state has had on the ability to replan, which specific alterations could have been made to the network state in order to reduce past and future runtime and by what amount?

In light of the first question, in the case that costly decisions have already been made, it is useful to know whether small changes to decisions in the past could have significantly reduced any negative impacts on core performance criteria. For example, a suboptimal meet location could have not just immediate delay for one of the trains, but also impacts on future meet events that are delayed in turn.

Problem 3: impact of individual trains.

The third addresses specific trains in a dispatch: Which trains, in particular, have the largest impact on the ability to run to schedule? And to what degree were these effects caused by the train's own performance or caused by its secondary impact on other trains?

Train volume on many network sections fluctuates over the course of a day or week. As such, trains running during less congested periods could experience large delays, but have very little impact on other trains. Conversely, a train in highly congested periods, or a train with which many others interact, can have a large impact on overall dispatch performance even if its own deviation from

an optimal dispatch plan is small.

1.3 Contributions

The three application questions and key findings from the analysis are: 1) Determine immediate future impact of dispatch decisions that have been made; a temporal analysis shows the periods during which empirical optimality gap grew most substantially, degrading potential performance. 2) Find possible remedies to past decisions that would be particularly impactful; during certain periods, a short list of changes of small magnitude to empirical data was found to have a magnifying effect on potential future performance. 3) Assess the impact that individual trains had on other trains and the dispatch plan as a whole; certain trains exhibit unique behavior in terms of propagation of their own delay onto others and their effects on dispatch decision making.

We are ultimately interested in a variety of dispatch analysis questions. With this methodology, we show that each of these three problems can be posed as a data-constrained optimal dispatch problem and each is a sub-problem of the methodology's general form. Importantly, this methodology (and the three sub-problems) are in the same class of problem as the optimal dispatch problem. In most cases, the dispatch analysis methods could follow directly from the constraint set used in an optimal dispatch model. This dispatch analysis problem methodology delivers the tools to critically evaluate empirical dispatching decision making and performance questions.

The remainder of this article is organized as follows. Section 2 discusses existing literature related to this work. The formulation of the dispatch analysis methodology is explained in Section 3. Problem 1, the impact of dispatch decisions, is introduced in Section 4.1; Problem 2, alterations to dispatch decisions, is introduced in Section 4.2; Problem 3, impact of individual trains, is introduced in Section 4.3. Section 5 discusses data used in the case studies for each problem. Sections 6, 7, and 8 present and discuss results for the three problems. Finally, Section 9 concludes the article and discusses approaches to future work on the topic.

2 Literature review

In this section we discuss a variety of works related to rail dispatching from both passenger and freight rail. Specifically, we address scheduling, replanning schedules in the presence of deviations, and train delay dynamics.

2.1 Scheduling and replanning

Unexpected delays, to say nothing of the many potential sources of delay, are inevitable. The ability to limit the size and impact of these delays maintains performance of the railroad and reduces their associated cost. Delays costs the railroads and their customers money and reduces the competitiveness of shipping

(and transporting passengers) by rail compared to other modes (Lovett, Dick and Barkan, 2015).

Schedules can be made somewhat resilient to small, unexpected delays by designing them to absorb these events for certain trains without necessitating a change to other trains (Salido, Barber and Ingolotti, 2008). If railroads endeavor to adhere to a schedule, it is also desireable for such a schedule to be realistic and repeatable under small, inevitable deviations (Hallowell and Harker, 1998). A survey of such work can be found in Lusby et al. (2018).

Disruption management literature is well-summarized by Fang et al. (2015) and is generally most relevant at the real-time level (Narayanaswami and Rangaraj, 2011). *Rescheduling* or *replanning* is a process by which the schedule is modified during the relevant period of dispatching (and perhaps immediately before a train's route begins) because of deviations from the expected conditions on which the schedule relies. Corman, Quaglietta and Goverde (2018) evaluate various rescheduling approaches in an automated control context for their ability to maintain rail traffic that is resilient to delays. Fast algorithms are required if computer aided dispatching and rescheduling is to be useful; Törnquist (2007) develops a heuristic approach for use in replanning under disturbances. Rescheduling can occur with various objectives and these are assessed on how well they meet performance measures. Minimizing total final delay tends to impose delay on fewer trains and a short planning horizon was shown to be sufficient for longer-term results. Shakibayifar, Sheikholeslami, Corman and Hassannayebi (2020) develop a disruption rescheduling approach using a basic scheduling model as input and solve using a two-stage heuristic method. An alternative to heuristic methods is an exact decomposition, which is presented in (Lamorgese and Mannino, 2015), and in (Luan, De Schutter, Meng and Corman, 2020) for real-time management of large-scale networks.

Large disruptions can easily make the timetable infeasible, even with some small adjustment. Corman and D'ariano (2012) construct alternative graphs to serve as decision support in cases of large disruptions and estimate performance metrics of various alternatives. Gestrelius, Aronsson, Forsgren and Dahlberg (2012) use optimization to develop schedules that proactively consider schedule alternatives should trains experience delay and require replanning. Pellegrini, Marlière and Rodriguez (2016) perform experiments quantifying the differences in replanning using optimization algorithms and the actual manner in which professional dispatchers handle disruption scenarios; they find potential gains for implementing optimal replanning over current practice.

Railroad operating philosophies differ in North American freight with respect to *schedule flexibility*, the allowance for trains to depart or modify their departure time based on their, status as opposed to a strict, pre-conceived schedule. High schedule flexibility generally makes it more difficult for railroads to establish and run an optimal meet-pass plan. In order to allow railroads to operate in real time with greater schedule flexibility, (Sehitoglu, Mussanov and Dick, 2018) explore the viability of compensating for this effect with greater allowable operating speed, which allows some delays to be mitigated and maintain certain more beneficial schedules.

2.2 Train delay dynamics

As mentioned earlier, delays are inevitable and costly for railroads. The effects of a delay or disruption extend beyond the direct cost incurred by the event because they can influence other trains. This interdependence is generally referred to as delay/disturbance propagation, secondary delay, or knock-on delay. *Knock-on delay* occurs when a schedule deviation in one train has a delaying effect on another train. Isolation and prediction of knock-on delays is difficult because of multiple factors including primary delay magnitude and location, multiple sources of primary delay, timetable of trains, and infrastructure configuration. The time dynamics of delays are complex: a delay at one point in time may not have its impacts felt until considerably later and the effects may be felt amongst multiple other trains. Our dispatch analysis work addresses some of these points.

Three common methods in literature for determining delays are analytical methods, simulation, and empirical statistics (Milinković et al., 2013).

Daamen, Goverde and Hansen (2009) identify knock-on delays using continually-updated blocking time graphs to determine train conflicts, where the logical and temporal interactions are based in colored Petri nets. Milinković et al. (2013) construct fuzzy Petri nets to isolate train delays using historical data or railroad expertise, when data is not available. Hansen et al. (2010) use a timed event graph, informed by historical data, to model delay propagation and predict arrival times. Lovett, Dick and Barkan (2017) analytically quantify the cascading effects and operational costs of slow orders on the railroad, which introduce disturbances to schedules.

Carey and Kwieciński (1994) model relationship between scheduled train headway and knock-on delay using stochastic simulation. Murali, Dessouky, Ordóñez and Palmer (2010) predict delays on aggregated sections of network using results of simulation by Lu, Dessouky and Leachman (2004), which allows the assessment of scheduling across large networks. Hwang and Liu (2009) uses micro-simulation for modeling interactions between trains and measuring delay. Disturbances are introduced to the simulation model of existing scheduled timetable and effects in terms of track occupancy and arrival at stations are measured. They assume that some amount of schedule recovery is available to trains. Quaglietta, Corman and Goverde (2013) evaluate schedule stability under a stochastic environment using a dispatching tool and rail simulation, finding that propagating disturbances result in unstable rescheduling plans. Diaz de Rivera, Dick and Evans (2020) found that moving blocks and train fleets can reduce delay caused by train meets on single track corridors.

The manner in which a railroad is constructed and operated influences its ability to handle delay events. Mussanov et al. (2017) look at the impacts of schedule flexibility on rail line performance and find that introducing rigid scheduling of trains does little to affect overall performance until a large portion of operations are run in this manner. However, schedule flexibility imposes additional infrastructure requirements to maintain the same level of service compared to more rigidly operated schedules (Dick and Mussanov, 2016). Dingler,

Koenig, Sogin and Barkan (2010) used Rail Traffic Controller (RTC) to simulate traffic scenarios and found, as one delay effect, that opposing direction traffic causing meet events had a much larger effect on delay compared with same direction traffic causing reduced speeds. Gorman (2009) used an econometric model to determine causes of delay and the marginal delay impact of adding trains; results showed that train interactions – meets, passes, and overtakes – contributed most to delay. Yuan and Hansen (2007) analyze the relationship between capacity utilization and the sensitivity of the schedule to disturbances and find that schedule buffer time decrease is exponentially related to knock-on delay.

3 Methodology

In this section, we explain the general form of the dispatch analysis problem mathematical model for single-track rail lines with passing sidings. It is derived from the general form of the optimal dispatching problem, which is discussed first. We use a specific dispatch formulation for illustration on empirical data in later sections, but emphasize the strategy could be applicable to other dispatch formulations. Assumptions made for both of these models are also discussed.

3.1 Preliminaries: optimal dispatch problem

Here we review the optimal dispatch problem and the constraints representing physical and legal operating rules, which are common to optimal dispatching as well as our dispatch analysis work. The model detailed here is a form of Petersen, Taylor and Martland (1986), detailed completely in Barbour, Samal, Kuppa, Dubey and Work (2018). We enumerate some of the important details to illustrate the complexity of the dispatching rules. However, we would emphasize that the dispatch analysis methods are easily extensible beyond the specific optimal dispatch problem discussed here and could, in principle, be used with another optimization-based dispatching formulation. Additionally, the measures of train and network performance (e.g., the objective function) used here could be made specific to a given rail operator with additional knowledge about their dispatching practices.

In this work, we consider the time values at which trains passed fixed locations on the network. These fixed locations are called *OS-points* and delineate the endpoints of track segments. In this manner, we work with the times at which each train reaches the end of each segment of track. Track segments belong to the ordered set M, where $M : 0, 1, 2, 3, ...$ and is indexed by $m \in M$. Each section of the network has trains running in two directions: directions 1 and 2. Trains in direction 1 are the set I, indexed by $i \in I$, and trains in direction 2 are the set J, indexed $j \in J$. We therefore refer to each timing value, the time at which a train $i \in I$ completed track segment $m \in M$, as $x_{i,m}$. Likewise, for trains $j \in J$, we have $x_{j,m}$. Passing sidings are track segments belonging to the set $S \subset M$ and indexed by $s \in S$. This formulation deals only with two

directions along a single-route network segment (e.g., a rail subdivision) and cannot deal with network topologies that contain cycles or alternative routes.

It is worth noting that when referring to the times at which a train $i \in I$ in direction 1 and a train $j \in J$ cross the same track segment $m \in M$, these times are actually referring to the two separate endpoints of that track segment. Because they operate in opposite directions, the completion points of the segment are at opposite ends. Additionally, note that not all trains operate across every track segment. When relevant, we denote M_i as the subset of track segments on which train $i \in I$ has timing values, where $M_i \subseteq M$; this is the same for trains $j \in J$, where $M_j \subseteq M$.

The collection of time values at OS-points for a train is referred to a train's *trajectory*. Symbolically, we denote the trajectory for train $i \in I$ as $x_{i,m} \forall m \in M_i$; likewise for train $j \in J$: $x_{j,m} \forall m \in M_j$. These trajectories for each train in a dispatch problem are assembled into the vector of decision variables, which we denote x.

Recall that an optimal dispatch problem finds the train trajectories that minimize some measure of dispatching cost. It may be posed in the general form:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & f(x,z) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b, \end{aligned} \tag{1}$$

where the decision variables are $x \in \mathbb{R}_+^p$ and $z \in \mathbb{Z}^q$. In a common formulation Petersen et al. (1986) and in this work, the decision variables x encode times at which trains reach various points on the network, while the integer decision variables z encode dispatching logic that indicates if and where meets and overtakes occur on the network and track assignment for trains. The function f quantifies some measure of dispatching cost, which is to be minimized. The physical and operational constraints, such as the permissible locations of meet and overtake events, headway constraints, and train travel times, are encoded in the inequality constraints $A_1 x + A_2 z \leq b$ and assumed to be mixed integer linear.

There are many constraints required for a functioning optimal dispatch model, by which we mean the model produces logically and operationally feasible train trajectories (additional operational rules and practices could be considered). A full discussion of the constraints for this particular model formulation can be found in Barbour et al. (2018). The subset discussed here, briefly, are for convenience and illustration of the problem complexity.

The most fundamental constraint defines the minimum runtime of each train on its assigned segments, shown for a train $i \in I$:

$$x_{i,m} \geq x_{i,m-1} + T_{i,m} \tag{2}$$

The difference in segment endpoint arrival times must be greater than or equal to the minimum possible main line runtime for the train on that segment, given by $T_{i,m}$. This applies for each segment that train i crosses: $m \in M_i$.

For pairs of opposing-direction trains $i \in I$ and $j \in J$, we state that one must never enter a track segment before the other has cleared the segment, plus a safety clearance headway time $H_{i,j}$:

$$\texttt{IF} \;\; \pi_{i,j,m} = 1, \texttt{THEN} \;\; x_{i,m} + H_{i,j} \leq x_{j,m+1}, \texttt{ELSE} \;\; x_{j,m} + H_{i,j} \leq x_{i,m-1}. \quad (3)$$

By defining this constraint for all single-track segments $m \in (M \setminus S)$ that they share, we ensure that their trajectories may only cross in time and space on siding track segments. The train's timing point to which the safety headway is applied is determined based on which train crossed the segment first; the binary variable $\pi_{i,j,m} = 1$ if train i crossed single-track segment $m \in (M \setminus S)$ first. Note that the conditional logic is abbreviated here in IF/THEN/ELSE notation for clarity, whereas in programming the logic is encoded in mixed integer linear constraints.

Meet events occurring between opposing-direction trains i and j on siding track $s \in S$ are denoted by the binary variable $\mu_{i,j,s}$. For each of these events that occurs ($\mu_{i,j,s} = 1$), we force one of the trains to take the siding track, as opposed to staying on the main line. The binary variable $\sigma_{i,s} = 1$ indicates that train i took siding track s, and likewise for train j. Therefore the sum of these variables during a meet event must equal one:

$$\texttt{IF} \;\; \mu_{i,j,s} = 1, \texttt{THEN} \;\; \sigma_{i,s} + \sigma_{j,s} = 1. \quad (4)$$

In cases where a train $i \in I$ took a siding $s \in S$ because of a meet or pass event, indicated by the binary variable $\sigma_{i,s} = 1$, its runtime across the segment $m \in M$ is subject to a different minimum, $U_{i,s}$, which is greater than or equal to the main line runtime $T_{i,m}$. This constraint is written as:

$$\texttt{IF} \;\; \sigma_{i,s} = 1, \texttt{THEN} \;\; x_{i,s} \geq x_{i,s-1} + U_{i,s}. \quad (5)$$

The complete list of constraints, parameters, and variables for both directions of trains are enumerated in detail in Barbour et al. (2018). The additional constraints refer to overtake events (trains traveling in the same direction), overtake siding assignment, and simultaneous meet/overtake events. For a dispatching window of 24 hours on 190-mile (305 km) section of track containing 37 segments, the problem formulation generates approximately 5,000 variables and 20,000 constraints. We solve all problems in this work to global optimality using CPLEX 12.8. For this 24-hour problem size, solve time on a 16-core CPU is on the order of 10 minutes. A longer time horizon problem is possible to solve, but solve time increases super-linearly.

3.2 Dispatch analysis problem general form

The overall concept behind dispatch analysis is that it performs optimization of a modified optimal dispatch problem, in the presence of some empirical/historical data from the actual railroad operations. We see in the optimal dispatch problem from (1) that it finds train trajectories, the time at which each train reaches

the end of each track segment, which are denoted x. In the cases where the optimal dispatch problem corresponds to a real scenario that occurred in historical data, each of these optimization variables also has a corresponding empirical value from the historical data. The decision variable $x_{i,m}$ has a corresponding value that is the true time at which train $i \in I$ reached the end of track segment $m \in M_i$; this empirical value we refer to as $\tilde{x}_{i,m}$.

We delineate a finite time horizon and finite track area over which to solve the dispatch analysis problem and refer to this as a *dispatch scenario*. The time bounds for the dispatch scenario are denoted $t_{\min}$ and $t_{\max}$, where each are real clock times, and the track area is defined in the set M for all tracks and subset S for passing siding tracks. Let $\tilde{X}$ be a set of all empirical timing points $\tilde{x}_{i,m}$ and $\tilde{x}_{j,m}$ for all trains with values in the interval $[t_{\min}, t_{\max}]$. Therefore, let X be the set of decision variables in an optimal dispatch problem corresponding to these empirical values in $\tilde{X}$. This defines decision variables corresponding to each train and each timing point that is observed in the empirical data. Likewise, let Z and $\tilde{Z}$ be the integer decision variables and their corresponding empirical values, respectively, for the dispatch scenario.

Both of the sets of decision variables, X and Z, are rewritten in their vector form as x and z by ordering variables according to train and track segment.

One can analyze empirical dispatching data in the context of optimal dispatching with added use of the function $g(x - \tilde{x}, z - \tilde{z})$ to match empirical data with decision variables.

We can write the most generic form of the dispatch analysis optimization problem to minimize an objective function that is the combination of dispatching cost, $f(x, z)$ and deviation of dispatched train trajectories from empirical data, $g(x - \tilde{x}, z - \tilde{z})$. The value λ determines the tradeoff between f and g and allows either function to be disregarded by setting $\lambda = 0$ or $\lambda = 1$. Both f and g may also be active constraints in the dispatch analysis problem, which requires the solution to meet a dispatching cost limit or a maximum deviation from empirical data, respectively. The constraint limits may be selectively activated by setting the values α and β to be finite. This general form may be written:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & \lambda f(x, z) + (1 - \lambda) g(x - \tilde{x}, z - \tilde{z}) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b \\ & f(x, z) \leq \alpha \\ & g(x - \tilde{x}, z - \tilde{z}) \leq \beta. \end{aligned} \tag{6}$$

The potential constraints on f and g, depending on the values for α and β, are in addition to the feasibility constraints for the trajectory values and integer variables: $A_1 x + A_2 z \leq b$. Note that the form of (6) is somewhat non-standard, but is intended to reflect the similarities in form between the specific dispatch analysis problems described later.

3.3 Dispatch analysis at a point in time

In the process of dispatch analysis, we are often looking at decisions that have already been made in the context of schedule replanning options and cost for the future. Formulating these questions requires the separation of subsets of decision variables and empirical data based on their relationship to a specific moment in time. We introduce a time parameter τ, which is a clock time within the interval $[t_{\min}, t_{\max}]$. The parameter is used to emulate the delineation of a portion of data which has already occurred (before τ) and a portion that has yet to occur (after τ). Note that τ can also be set to equal $t_{\min}$ or $t_{\max}$. All time variables used in a problem realization are all minimum-regularized according to the lower bound of the dispatch scenario, $t_{\min}$. All timing points are measured as their difference to $t_{\min}$, in seconds, and are therefore greater than or equal to zero.

Let $\tilde{X}_{\tau-}$ denote the subset of $\tilde{X}$, the empirical data, that occurred at or before time τ (i.e., on the interval $[t_{\min}, \tau]$); and let $\tilde{X}_{\tau+}$ be the subset of $\tilde{X}$ that occurred after time τ (i.e., on the interval $(\tau, t_{\max}]$). The separation of the decision variables, X, is based on the time value of each corresponding empirical point. A decision variable $x_{i,m}$ or $x_{j,m}$ is contained in $X_{\tau-} \subseteq X$ if its corresponding empirical value, $\tilde{x}_{i,m}$ or $\tilde{x}_{j,m}$ has a value less than or equal to τ (i.e., $\tilde{x}_{i,m} \in \tilde{X}_{\tau-}$ or $\tilde{x}_{j,m} \in \tilde{X}_{\tau-}$). Likewise, a decision variable is contained in $X_{\tau+} \subseteq X$ if its corresponding empirical value is greater than τ (i.e., $\tilde{x}_{i,m} \in \tilde{X}_{\tau+}$ or $\tilde{x}_{j,m} \in \tilde{X}_{\tau+}$). This separation based on τ is performed based on the empirical data, $\tilde{X}$, because it has known time values.

The separation of the set of integer variables, Z, corresponds to sets $X_{\tau-}$ and $X_{\tau+}$. Additionally, we define a division of the sets of trains, I and J, based on τ. Trains that have any portion of their empirical trajectory (i.e., any variable $\tilde{x}_{i,m}$ or $\tilde{x}_{j,m}$), included in the set $\tilde{X}_{\tau-}$ are in sets $I_{\tau-}$ and $J_{\tau-}$. Likewise, trains that have any portion of their empirical trajectory in the set $\tilde{X}_{\tau+}$ are in sets $I_{\tau+}$ and $J_{\tau+}$. It is possible for trains to be included in both sets $\tau-$ and $\tau+$, respective of their direction. This notation or empirical data and decision variables that will be used in subsequent explanations is summarized in Table 1.

The general form of the dispatch analysis problem described in (6) can be refined to consider only values in $\tilde{X}_{\tau-}$ or $\tilde{X}_{\tau-}$. When this separation of values is used in functions f and g, we denote those functions $f_{\tau-}$, $f_{\tau+}$, $g_{\tau-}$, and $g_{\tau+}$. The interpretation of each function is then:

- $f_{\tau-}(x_{\tau-}, z_{\tau-})$: dispatch performance metric evaluated on decision variables in $x_{\tau-}$ and $z_{\tau-}$, at or before time τ.
- $f_{\tau+}(x_{\tau+}, z_{\tau+})$: dispatch performance metric evaluated on decision variables in $x_{\tau+}$ and $z_{\tau+}$, after time τ.
- $g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-})$: function quantifying the difference between empirical data, $\tilde{x}_{\tau-}$ and $\tilde{z}_{\tau-}$, and corresponding decision variables, $x_{\tau-}$ and $z_{\tau-}$, at or before time τ.

Quantity	**Description**
$\tilde{x}_{i,m}$, $\tilde{x}_{j,m}$	Empirical values for the actual time at which train $i \in I$ or $j \in J$ completed track segment $m \in M$.
$x_{i,m}$, $x_{j,m}$	Individual optimization decision variables for the time at which a train, $i \in I$ or $j \in J$, completed track segment $m \in M$.
$t_{\min}$, $t_{\max}$	Lower and upper time bound of dispatch interval, respectively.
τ	Time value in interval $[t_{\min}, t_{\max}]$ that delineates a point at which a specific analysis occurs.
$\tilde{X}$	Set of all empirical timing points that fall within the interval $[t_{\min}, t_{\max}]$.
$\tilde{Z}$	Set of integer values corresponding to $\tilde{X}$.
X, Z	Sets of all optimization decision variables corresponding to the empirical values $\tilde{X}$ and $\tilde{Z}$, respectively, for the interval $[t_{\min}, t_{\max}]$.
$\tilde{X}_{\tau-}$, $\tilde{Z}_{\tau-}$	Subsets of $\tilde{X}$ and $\tilde{Z}$ where values of $\tilde{x}_{i,m}$ or $\tilde{x}_{j,m}$ are on the interval $[t_{\min}, \tau]$.
$\tilde{X}_{\tau+}$, $\tilde{Z}_{\tau+}$	Subsets of $\tilde{X}$ and $\tilde{Z}$ where values of $\tilde{x}_{i,m}$ or $\tilde{x}_{j,m}$ are on the interval $(\tau, t_{\max}]$.
$X_{\tau-}$, $Z_{\tau-}$, $X_{\tau+}$, $Z_{\tau+}$	Subsets of X and Z corresponding to values in $\tilde{X}_{\tau-}$, $\tilde{Z}_{\tau-}$, $\tilde{X}_{\tau+}$ and $\tilde{Z}_{\tau+}$, respectively.
$x_{\tau-}$, $z_{\tau-}$	Ordered vectors of decision variable sets $X_{\tau-}$ and $Z_{\tau-}$ for each train and for each track segment.
$x_{\tau+}$, $z_{\tau+}$	Ordered vectors of decision variable sets $X_{\tau+}$ and $Z_{\tau+}$.
$\tilde{x}_{\tau-}$, $\tilde{z}_{\tau-}$, $\tilde{x}_{\tau+}$, $\tilde{z}_{\tau+}$	Ordered vectors of empirical value sets $\tilde{X}_{\tau-}$, $\tilde{Z}_{\tau-}$, $\tilde{X}_{\tau+}$ and $\tilde{Z}_{\tau+}$, respectively.
$I_{\tau-}$, $J_{\tau-}$	Subsets of trains I and J, which contain trains that have any portion of their empirical trajectory in the set $\tilde{X}_{\tau-}$ (i.e., before or at time τ).
$I_{\tau+}$, $J_{\tau+}$	Subsets of trains I and J, which contain trains that have any portion of their empirical trajectory in the set $\tilde{X}_{\tau+}$ (i.e., after time τ).

Table 1: Summary of general notation for optimization variable sets and empirical data.

- $g_{\tau+}(x_{\tau+} - \tilde{x}_{\tau+}, z_{\tau+} - \tilde{z}_{\tau+})$: function quantifying the difference between empirical data, $\tilde{x}_{\tau-}$ and $\tilde{z}_{\tau-}$, and corresponding decision variables, $x_{\tau+}$ and $z_{\tau+}$, after τ.

When functions $f_{\tau-}$, $f_{\tau+}$, $g_{\tau-}$, and $g_{\tau+}$ are used in binding constraints (as in (6)), the limit values α and β corresponding to f and g will be denoted $\alpha_{\tau-}$, $\alpha_{\tau+}$, $\beta_{\tau-}$, and $\beta_{\tau+}$, respectively. They are thus interpreted as the upper limit values on each of these functions – the required dispatch performance and the allowable difference between decision variables and empirical data, before or after time τ.

3.4 Assumptions and limitations

We briefly summarize key assumptions that are made in the dispatch analysis methodology and comment on their importance, as well as the limitations of these methods.

First and foremost, the methods rely on a pre-existing optimization-based scheduling model that produces realistic train plans. We present one such model in this article with some assumptions and simplifications, which we document below, but other models must be used on other network topologies (e.g., double/triple track) or operating paradigms (e.g., passenger rail). The analysis produced is dependent on the model that is used and can capture only the detail with which the underlying optimization model is constructed, so some delays may be explainable by other necessary operations activities on the railroad (e.g., set out of rail cars on auxiliary tracks).

The true root cause of a delay – be that a dispatch decision, mechanical failure, train power limitation, etc. – is not directly produced by the analysis, but rather it is a strong indicator that requires interpretation of the results.

The dispatch analysis methodology considers data at the track segment level. Therefore, the feasibility of a train trajectory is only determined in timing at ends of the segment and minimum train-specific free run times, and not by train performance capabilities in the middle of the segment. This assumption is critical to the particular optimization model presented here, but the methods can be generalized to other higher-fidelity data streams by modifying the model. We must also assume that atypical train movements, such as reversing, did not and can not occur. This is a simplifying assumption made to reduce model complexity, and it is reasonable given the rarity of these movements.

A fundamental assumption with respect to train schedules is that trains were intended to depart at the time where they registered their first OS-point. This is related to a larger point about dependency: there are events outside of the dispatch window, both temporally and spatially, that impacted the trajectories of trains inside the window. In this work we assume that no information past the spatial boundary of the dispatch scenario is known. A train entering the boundary is assumed be pre-determined at that exact time when it enters, even though it is dependent on other trains within/exiting the area. Events at the beginning or end of a dispatch window should be further analyzed by shifting

the window in the relevant direction to capture more spatial and/or temporal context. Given that an optimal dispatching problem across multiple days is computationally difficult, some care must be taken with the dispatch window.

Finally, we use in this work a measure of optimality based solely on train runtime. For the sake of generality and simplicity in this work, we choose this basic measure. Many other formulations could be chosen in practice that better capture the operational strategy of a given railroad, particularly if such a function was known to be the basis of scheduling and dispatching decisions. For example, priority weighted runtime is a frequent aspect of decision making; this requires, though, the known priority mapping to be informative and valuable.

4 Formulations of specific dispatch analyses

In this section we instantiate specific forms of the dispatch analysis problem to solve the three problems identified in this work. Each follows from the general form given in (6).

4.1 Problem 1: impact of dispatch decisions

The first application of the dispatch analysis problem is to quantify the impact of the current network state on the ability of the schedule to continue to run at or near optimality. As events happen on the network and trains deviate from the original optimal schedule, the plan must be re-optimized to take into account deviations.

We allow the network to evolve up to time τ, and then replanning is initiated based on the positions of trains at this time. This combination gives us the best possible dispatch achievable given the decisions that have already been made, assuming we make optimal decisions moving forward. It is therefore a lower limit on future dispatching cost.

The best possible scenario as time progresses is for trains to maintain the optimal schedule. If deviations do occur, it is the best case that they do not impact the future schedule or other trains. That is, if it is possible that a train deviates from its optimal schedule, but does so in a manner that does not reduce its ability to catch back up at a later point (i.e., the optimal schedule is non-unique). Deviations that do impact the future schedule will result in a replanned future schedule that has a larger objective value than the original, optimal schedule (e.g., by delaying other trains). That is, adding additional constraints to the optimal dispatch problem in the form of fixing decision variables to empirical data, can only serve to increase a minimized objective value (assuming the optimal schedule is unique). As time progresses and the schedule is fixed further to the empirical data, the objective value of the replanned schedule will converge on the objective value of the wholly empirical data (i.e., when τ reaches the end of the data window $[t_{\min}, t_{\max}]$). The specific steps in this analysis are as follows:

1. For a time window $[t_{\min}, t_{\max}]$, solve the pure optimal dispatch problem (1) given only the departure location (i.e., starting at the network boundary or at any network segment therein) and corresponding departure time (within $[t_{\min}, t_{\max}]$) of each train and no other empirical data. This establishes the **baseline** dispatch where we refer to the total runtime of all trains as r_0.

2. Assemble the empirical trajectories of all trains within the interval $[t_{\min}, \tau]$, which is $\tilde{x}_{\tau-}$.

3. Fix the decision variables $x_{\tau-}$ to their empirical values, $\tilde{x}_{\tau-}$ using a constraint on the function $g(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-})$.

4. Solve an optimal dispatch problem for the remainder of the time window (minimize $f_{\tau+}(x_{\tau+}, z_{\tau+})$ with the added constraints on $g_{\tau-}$. *This formulation is given below in* (7).

5. The new objective value is the best dispatch achievable given the decisions made up to τ. The total runtime of all trains for this problem is denoted r_τ, where $r_\tau \geq r_0$, and the increase in runtime is denoted $\Delta = r_\tau - r_0$.

In this formulation we wish to find the best dispatch plan, while before time τ holding the difference between each empirical data point and corresponding timing variable, $g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-})$, to zero. This fits into the general dispatch analysis formulation as:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & f(x,z) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b \\ & g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-}) = 0. \end{aligned} \tag{7}$$

In this problem, we take the objective function f, quantifying the dispatch performance, to be the sum of the runtime of all trains. The distribution of any delay on the route is not of concern, so long as the train reaches its destination at the earliest possible time. The specific form of f is therefore:

$$f(x,z) = \sum_{i \in I} x_{i,q_i} + \sum_{j \in J} x_{j,q_j}, \tag{8}$$

where q_i and q_j are the final track segments of the trajectories for trains $i \in I$ and $j \in J$, respectively. This means that x_{i,q_i} and x_{j,q_j} denote the completion time of the final track segment for trains i and j.

In order to hold the value of each timing variable in $x_{\tau-}$ to its empirical value in $\tilde{x}_{\tau-}$, we impose the $\mathcal{L}_1$ norm on the difference between decision the variables and constraint the value of $g_{\tau-}$ to be zero:

$$g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-}) = ||x_{\tau-} - \tilde{x}_{\tau-}||_1 = 0, \tag{9}$$

where $||\cdot||_1$ is the $\mathcal{L}_1$ norm defined as the absolute difference between timing points, in seconds.

4.2 Problem 2: alterations to dispatch decisions

We presented in Section 4.1 a method to study the impact of empirical decision making on the baseline optimal dispatch plan. This reveals the temporal manner in which deviations from the baseline plan impacted total train runtime. We are now interested in isolating more specific instances of empirical performance that could have been *changed* in order to improve future dispatch performance. At a given time τ, where the empirical data before τ, $\tilde{x}_{\tau-}$, has introduced Δ minutes of additional runtime to the dispatch in excess of the baseline optimal dispatch, we find the minimal changes to the empirical data in $\tilde{x}_{\tau-}$ that could be made which would decrease Δ to a desired level.

Alterations to the empirical data are measured in absolute minutes of a train's segment runtime (i.e., adding or subtracting runtime). An alteration to one segment runtime for a train affects all of the subsequent timing points for that train. Altered segment runtimes must still obey constraints, but no other limits on train performance are placed beyond the existing constraint set.

The steps to address this problem are as follows:

1. For a time window $[t_{\min}, t_{\max}]$, solve the optimal dispatch problem(1) given only the initial condition of each train. This establishes the **baseline** dispatch where we refer to the total runtime of all trains as r_0.

2. Assemble the empirical trajectories of all trains within the interval $[t_{\min}, \tau]$: $\tilde{x}_{\tau-}$.

3. Fix the decision variables $x_{\tau-}$ to their empirical values, $\tilde{x}_{\tau-}$ and dispatch optimally for the remainder of the time period, as described in Problem 1 by equation (7). This represents the best dispatch achievable given the decisions made up to τ. The total runtime of all trains for this problem is denoted r_τ, where $r_\tau \geq r_0$.

4. Calculate $\Delta = r_\tau - r_0$, the runtime in excess of the baseline dispatch that was added because of the empirical decisions up to τ. Determine a reduced value of Δ, which is to be achieved by modifying the empirical data; we denote this value Δ' and calculate the desired total runtime as $r' = r_0 + \Delta'$.

5. Impose a constraint on the total runtime, represented by $f(x, z)$, which must be less than or equal to r'.

6. Solve the empirical improvement optimization problem, which minimizes the alteration to the empirical data before τ, $\tilde{x}_{\tau-}$, while achieving a total runtime of all trains less than or equal to r'. *This problem is given below in* (10).

This problem minimizes the alteration to the empirical data, $g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-})$, before τ that is required to reduce the objective value of the

replanned dispatch, $f(x, z)$, at τ to a desired level:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-}) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b \\ & f(x, z) \leq r', \end{aligned} \tag{10}$$

where r' is the value of runtime of all trains that is to be achieved by alteration of empirical data; it is described above in step 4.

We again define $f(x, z)$, the performance function for dispatched trajectories that is constrained to reduce the lower bound of the replanned dispatch, to be the total runtime of all trains:

$$f(x, z) = \sum_{i \in I} x_{i,q_i} + \sum_{j \in J} x_{j,q_j}, \tag{11}$$

where q_i and q_j are the final track segments of the trajectories for trains $i \in I$ and $j \in J$, respectively, and x_{i,q_i} and x_{j,q_j} denote the completion time of the final track segment for trains i and j.

We define the objective function $g_{\tau-}$, the alteration to empirical trajectories before τ, as:

$$\begin{aligned} g_{\tau-}(x_{\tau-} - \tilde{x}_{\tau-}, z_{\tau-} - \tilde{z}_{\tau-}) = & \sum_{i \in I_{\tau-}} \sum_{n=p_i+1}^{q_i} |(x_{i,n} - x_{i,n-1}) - (\tilde{x}_{i,n} - \tilde{x}_{i,n-1})| \\ & + \sum_{j \in J_{\tau-}} \sum_{n=p_j-1}^{q_i} |(x_{j,n} - x_{j,n+1}) - (\tilde{x}_{j,n} - \tilde{x}_{j,n+1})|, \end{aligned} \tag{12}$$

where $I_{\tau-}$ and $J_{\tau-}$ are the subsets of I and J for which some portion of the trajectory of $i \in I$ or $j \in J$ are included in $X_{\tau-}$; p_i and p_j are the first track segments in the trajectories of i and j that are included in $X_{\tau-}$; q_i and q_j are the final track segments in the trajectories of i and j that are included in $x_{\tau-}$. This function computes the summation of the differences in segment runtimes between the decision variables $x_{\tau-}$ and their empirical values $\tilde{x}_{\tau-}$. The differences in segment runtimes are used (as opposed to the difference in timing points) because any alteration applied to the difference between these values will affect each successive timing point in the trajectory, effectively shifting the remainder of the train's trajectory in time. This is analogous to altering a train's empirical trajectory so that it would have run faster on a particular track segment, and thus arrived at successive OS-points sooner.

4.3 Problem 3: impact of individual trains on dispatch plan

Quantifying the impact of empirical decisions run up to a given time, described in Problem 1 in Section 4.1, provided information on a temporal basis. It can

also be informative to assess the empirical data with respect to specific trains. Beyond the impact of runtime incurred by a train itself, what impact did that train have on others in the dispatch plan?

In this problem, we find not the alterations to a train's trajectory, but instead the impact that fixing such train's trajectory has on the rest of the schedule. Small deviations of a primary train's trajectory from the optimal schedule have the potential to secondarily impact other trains significantly if the primary train was tightly integrated, indicating that the schedule is very sensitive to that train. Conversely, a train may have very little secondary effect on the schedule, even if it experiences large deviations from its optimal trajectory.

This problem is similar to the quantification of knock-on delay, but more precisely, it measures, relative to the optimal schedule, the secondary changes to other train schedules that are necessitated by a primary train's deviation from its trajectory in the optimal schedule. The steps to evaluating a specific train, denoted w, in this application are as follows:

1. For a time window $[t_{\min}, t_{\max}]$, solve the optimal dispatch problem (1) given only the departure location and departure time of each train. This establishes the **baseline** dispatch, where we refer to the total runtime of all trains as r_0.

2. Let the runtime of train w in the baseline dispatch be γ_w.

3. Assemble the empirical trajectory of train w from all empirical data, $\tilde{X}$, on the interval $[t_{\min}, t_{\max}]$. Let the empirical runtime of train w be γ'_w.

4. Fix the decision variables $x_{w,m}$ to their empirical values $\tilde{x}_{w,m}$ for all tracks segments in the trajectory of train w, by constraining $g(x_w - \tilde{x}_w, z_w - \tilde{z}_w)$.

5. Solve an optimal dispatch problem for, effectively, all trains except w by minimizing $f(x, z)$. This represents the best dispatch achievable given fixed trajectory of train w. *This formulation is given below in* (13).

   (a) The total runtime of all trains with train w fixed is denoted r_w, where $r_w \geq r_0$.

   (b) The difference between the empirical runtime of train w and its baseline dispatch value, $\gamma'_w - \gamma$, is the *primary added runtime* for this train.

   (c) The difference between the runtime with train w fixed and the baseline runtime, for all trains except w, is the *secondary added runtime*. This can be calculated by subtracting the runtime difference for w (primary added runtime) from the overall runtime difference: $(r_w - r_0) - (\gamma'_w - \gamma_w)$.

In this problem, we wish to find the best dispatch of all trains, $f(x, z)$, but with the variables for train w fixed to their empirical values by holding function

$g(x_w - \tilde{x}_w, z_w - \tilde{z}_w)$ equal to zero:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & f(x,z) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b \\ & g(x_w - \tilde{x}_w, z_w - \tilde{z}_w) = 0, \end{aligned} \tag{13}$$

where w is a specific train in the dataset that is being assessed, x_w and z_w are the decision variables for train w, and $\tilde{x}_w$ and $\tilde{z}_w$ are the empirical values for train w.

We again define $f(x, z)$, the objective value for dispatched trajectories, to be the total runtime of all trains:

$$f(x,z) = \sum_{i \in I} x_{i,q_i} + \sum_{j \in J} x_{j,q_j}, \tag{14}$$

where q_i and q_j are the final track segments of the trajectories for trains $i \in I$ and $j \in J$, respectively, and x_{i,q_i} and x_{j,q_j} denote the completion time of the final track segment for trains i and j. Note that minimizing (14) is an equivalent optimization to minimizing the runtime of all trains excluding train w when the trajectory for w is fixed, but all trains are included in $f(x, z)$ for simplicity. This just requires a post-solve separation of runtime by train from the value of f.

In order to fix the empirical trajectory of train w, we impose for function g an $\mathcal{L}_1$ norm on the difference between empirical and solved trajectory timing values of train w, and constrain the value of g to be zero:

$$g(x_w - \tilde{x}_w, z_w - \tilde{z}_w) = ||x_w - \tilde{x}_w||_1 = 0, \tag{15}$$

where $||\cdot||_1$ is the $\mathcal{L}_1$ norm defined as the absolute difference between timing points, in seconds, x_w refers to the trajectory timing variables for train w, and $\tilde{x}_w$ refers to its empirical trajectory timing values.

5 Case study data preparation

We now present a description of freight rail dispatch data that is used as a case study to answer the three dispatch analysis questions in this work.

The historical dispatch dataset is collected from a rail network section of a U.S. Class-I railroad, between two major cities. The network section is single track with 17 passing sidings of various lengths and a total of 37 track segments. It is approximately 190 miles (305 km) in length and has a consistently high volume-to-capacity ratio. The specifics of the corridor and its operating railroad are not discussed for data confidentiality. Note, again, that the methods presented in this article are not specific to a railroad and the results are meant only to be illustrative of the methods' usefulness.

Case study analyses are given for various ranges of data. A single window of dispatching data for discussion of each of the three dispatch analysis problems

is taken from nine hours of data during the week of January 4, 2016. Aggregated analysis for problems 1 and 3 are performed across multiple windows of dispatching data from January 1, 2016, to January 31, 2016. All dispatch data is reconciled according to the process developed and discussed in Barbour et al. (2018) in order to remove any small errors or omissions and ensure feasibility prior to use in the dispatch analysis problem.

Values are derived via historical data mining for the following optimization parameters: directional minimum main line track runtime ($T_{i,m}$ and $T_{j,m}$), directional minimum siding track runtime ($U_{i,s}$ and $U_{j,s}$), clearance headway for meet events specific to each end of each siding ($H_{m,i,j}$), and follow headway for same direction trains specific to each direction and each track segment (H_{m,i_1,i_2} and H_{m,j_1,j_2}). These values, as formulated, allow differentiation on a per-train basis. In this work, we assume all trains have the same value for each track segment; but in practice, rail operators can define these values with differentiation based on train dynamics. The use of identical values for all trains can lead to overestimating performance in some cases, so higher-confidence values are desireable to more accurately identify a feasible trajectory for each train.

Runtime distributions for each track segment in each direction were mined from two years of historical data: January 2014 through December 2015. The minimum main line free run traversal time of each track segment in each direction, $T_{i,m}$ and $T_{j,m}$ (as in constraint (2)), was taken to be the 90th percentile lowest observed value for each direction. This choice was made based on inspection of runtime distributions. Historical timing at OS-points in the dataset is given to the nearest minute, which results in unreasonably low runtime value on short track segments if the pure minimum value is used (e.g., 1 minute on a 1.5-mile track segment, implying 90mph travel speed). Applying the 90th percentile rule mitigates this rounding effect without requiring deviation of a significant number of trains (or a large magnitude of deviation) when data reconciliation is performed, due to their lower runtime values on segments.

In order to ascertain siding track traversal times, the runtime values for trains during meet events were isolated. Within each meet event, the main line runtime was assumed to be the lower of the two values and the siding runtime was assume to be the higher of the two, separated by direction. Figure 3 shows these lower and higher runtime value distributions for direction 1 trains taken from meet events, across a subset of sidings on the network section. Three plots, grouped by color for each siding, show the set of lower runtime values (left), the set of higher runtime values (right), and the comparison set of all observed runtimes (middle) regardless of whether a meet occurred. Each violin plot is similar to a histogram, with the widest area indicating a higher frequency of values than the thinner tails. The median value is marked with the solid colored dash and one standard deviation to each side is shown by the dotted line segments. The scale for exact runtime values is not given for data confidentiality, but clear distinction in the lower/higher runtime values for meet events can be seen for each siding.

The 90th percentile lowest runtime value is taken from the set of higher runtimes for each direction to be the minimum siding runtime for the optimiza-

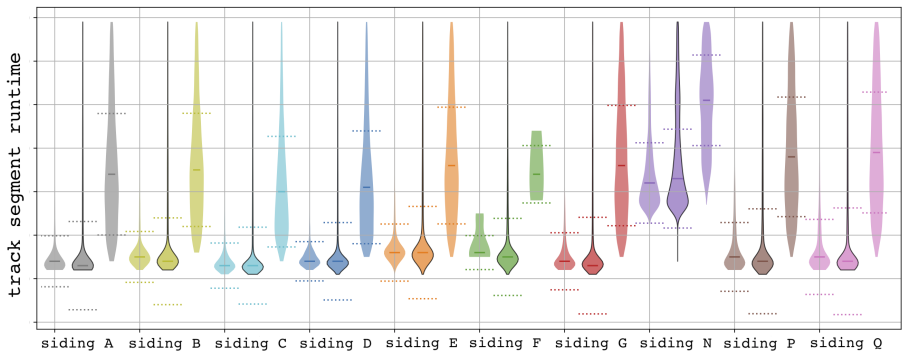


Figure 3: Distribution of runtimes on select sidings during meet events. Runtimes are normalized by the minimum and maximum values from all sidings. For each siding, grouped by color, the left violin plot is the distribution of runtimes for the faster of the two trains in the meet event and the right plot is the distribution of runtimes for the slower of the two trains. The middle plot in each group, outlined in black, is the distribution of all trains on the siding, not just those in meet events.

tion parameters $U_{i,s}$ and $U_{j,s}$ (direction 1 and direction 2, respectively). As a reminder, these values are used in siding runtime constraints, as in equation (5).

Headway values for trains in opposing directions, relevant for meet event clearance time, is similarly taken from the set of meet events observed to occur at each siding. Each endpoint of the siding track is considered separately by computing the difference in arrival time at this point by the two opposing-direction trains. The 80th percentile value is taken as the headway value in order to reduce rounding effects that result in very small headway values at certain sidings.

For same-direction headway, we consider all trains in the two-year mined dataset at the end of each track segment, independently. These trains are sorted by their arrival time at the end of the track segment and the different between each successive pair of trains is computed. Those with separation times of less than 30 minutes are assumed to be roughly following each other, and from this filtered set of separation times we take the 95th percentile minimum value as the minimum follow headway.

6 Results, Problem 1: impact of dispatch decisions

The analysis question we set out to answer in problem 1 is how to quantify the cost of past dispatch decisions on replanning in the future. As discussed in Section 4.1, we fix a set of empirical dispatching data on the time interval $[t_{\min}, \tau]$ and then, considering the true locations of trains at time τ, optimally

replan into the future, $(\tau, t_{\max}]$. The objective function used for dispatching is the minimization of overall train runtime. The interpretation of an objective function value at time τ is the lower bound on total train runtime if optimal decisions are made from τ forward. We first show results on applying this analysis to a single dispatch window, and then expand the analysis to cover two weeks of dispatching.

Consider the time-space diagram in Figure 4, which we refer to as a *stringline* (also known in other works as a *time distance plot*). Siding tracks, where trains may meet and pass each other, are shaded grey, with all other single track segments in white. The train's speed profile is assumed to be linear between timing points. It should be noted that in cases where a train stops on a siding track, the linear depiction of the train's speed profile does not visually show a stoppage. This diagram shows a portion of the optimal dispatch plan in green alongside the empirical dispatch data for this time interval, in blue. Many of the empirical trajectories can be seen diverging from their optimal trajectories. When we fix empirical data up to a time τ, we then assume that trains are at their empirical locations at this time and must be dispatched from there. Because many of the empirical trajectories diverge, significant replanning must be performed in order to make a new optimal plan.

Figure 5 shows the replanning process at time $\tau = 300$, which is marked on the stringline diagram by the dashed blue line. Empirical trajectories (dark blue) are run up to their last observed timing point before $\tau = 300$, which is why some stop short of the dashed blue demarcation. At this point, they have diverged significantly from what could have been their optimal trajectories up to this point, shown in green. After $\tau = 300$, replanning must be performed to develop the new plan; these replanned trajectories are shown in red. The total runtime of the empirical trajectories (blue) plus the replanned trajectories (red) make up the total train runtime for this replanned dispatch at time $\tau = 300$.

Each time more empirical data is introduced into the optimal dispatch problem the value of the objective function (total train runtime) must either remain the same (if optimal decisions were made) or increase (if any sub-optimal decisions were made). The empirical data adds not only its own sub-optimality with respect to the baseline optimal dispatch plan (i.e., in the form of train delay or alterations to the optimal plan), but it also has the potential secondary effect of requiring a change to the future of the optimal plan. The replanned future is optimal given the constraints, but introduces cost because of the sub-optimal positions of trains at the replanning point. The separation of these primary and secondary effects is not addressed in this work.

6.1 Analysis on a single dispatch period

We simulate the effect of dispatching moving forward in time by gradually increasing the τ parameter in 30-minute increments across a 9-hour dispatch time window, from 0 minutes to 540 minutes. The dispatch time window is described in Section 5. Indeed, increasing the τ parameter increases the overall objective function value as can be seen in Figure 6a. The green line shows the lower

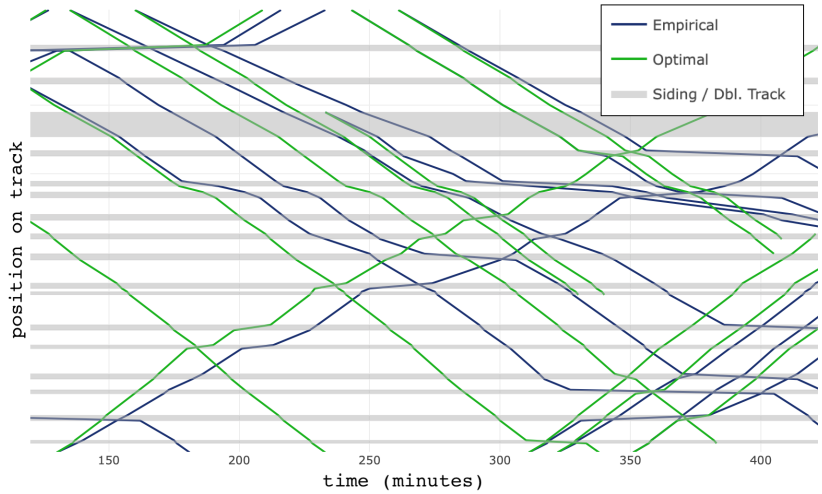


Figure 4: Stringline diagram of optimal baseline dispatch plan (shown in green) versus empirical dispatch (dark blue).

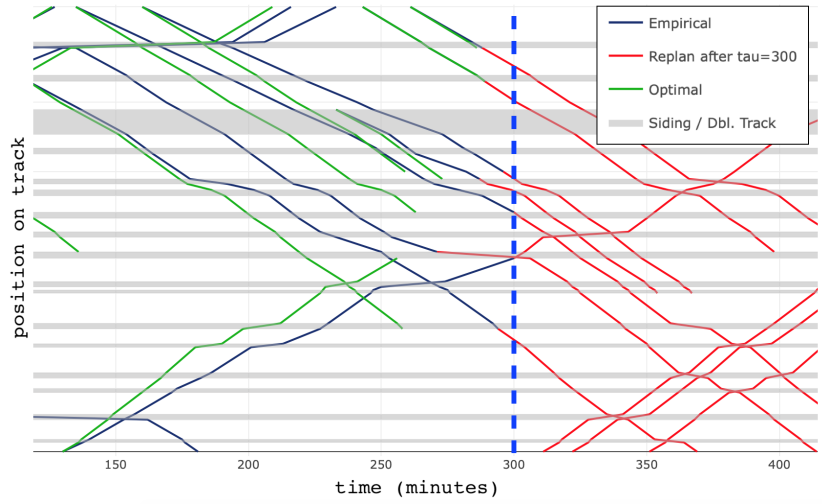


Figure 5: Stringline diagram for the same dispatch scenario as Figure 4, with dispatch replanning at $\tau = 300$. Empirical data is shown by the dark blue trajectories up to time $\tau = 300$ (marked by the dashed blue line). The trajectories of trains under the baseline (optimal) plan are shown by the green trajectories, for comparison of the empirical versus optimal locations of the trains at time τ. The replanned trajectories moving forward from the empirical locations at $\tau = 300$ are in red. The sections of empirical (blue) plus the replanned (red) trajectories constitute the total runtime value that is evaluated in this section.

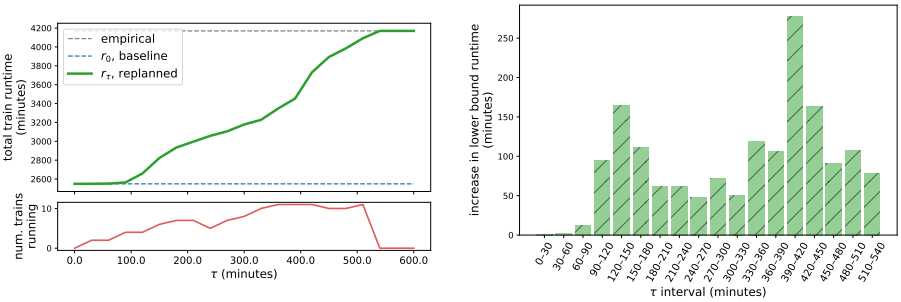
bound objective value (total train runtime), considering the empirical data and the replanned future. As previously mentioned for Figure 5, the total runtime of empirical trajectories (blue) plus the replanned future (red) constitute this total runtime lower bound at a specific τ value (a point on the green curve in Figure 6a). At $\tau = 0$, no dispatching decisions have yet occurred that will introduce sub-optimality; therefore, no cost is incurred by replanning and the dispatch is effectively the baseline plan. This baseline runtime is marked by the light blue dashed line in Figure 6a. At $\tau = 540$, the end of the dispatch window, all empirical dispatch decisions have occurred and no replanning is performed; therefore, the cost of replanning will be the empirical total runtime. This runtime is marked by the grey dashed line.

At the point $\tau = 300$ in Figure 6a, the plan for which we visualized in Figure 5, the lower bound runtime now has a value of $r_{300} = 3178$ minutes, compared with the optimal value of $r_0 = 2549$ minutes. However, the increase in lower bound runtime from the previous point, $\tau = 270$, is modest: $r_{300} - r_{270} = 72$ minutes. Compare this to the increase in lower bound from $\tau = 120$ to $\tau = 150$ or the increase from $\tau = 390$ to $\tau = 420$, which are much more severe: $r_{150} - r_{120} = 164$ minutes and $r_{420} - r_{390} = 286$ minutes. This indicates that decisions made on these time intervals, $[120, 150]$ and $[390, 420]$, were much more costly to the dispatch plan. Indeed, Figure 6b shows the amount of increase in the lower bound runtime over each successive step of τ; the latter intervals mentioned experience the largest increases in lower bound runtime over this dispatch window. Again, the separation of whether this was due to primary or secondary effects is a separate question, but it indicates where, temporally, runtime is being introduced in excess of what is possible in the optimal case.

6.2 Results across multiple periods

The trend in how the lower bound runtime increases across each particular window of dispatch data is expected to differ. The trend will be affected not only by the dispatching decisions but also by the distribution and volume of trains present on the network segment. Figure 7 shows the accumulation patterns of runtime across two weeks of data in 9-hour windows shifted by 3 hours (thereby overlapping by 6 hours for successive windows). For comparison purposes, the replanned lower bound values were min-max normalized to [0, 1] using the baseline and empirical runtime values. As is to be expected, there is fairly wide variation in how each accumulates delay within the 9 hours. The mean trend is shown with the black dashed line.

Two particular 9-hour dispatch time windows from different days are highlighted in purple, labeled "morning A" and "morning B". These two are highlighted because they demonstrate very opposing trends in how their lower bound runtimes evolved. By $\tau = 90$ minutes, the lower bound runtime on "morning A" has already increased 50% of the way from its baseline to empirical values; in contrast, "morning B" has reached only around 2%. The large increases for "morning B" occur between $\tau = 300$ and $\tau = 540$, where its lower bound runtime increases around 90% of its baseline-empirical range.



(a) (b)

Figure 6: (a) Lower bound on train runtime (green curve), r_τ, at time τ, when empirical decisions are run from $t_{\min}$ to τ and the optimally-replanned dispatch is executed from τ to $t_{\max}$. As additional empirical decisions are taken into account, the lower bound runtime increases from the baseline optimal value ($r_0 = 2549$ minutes) when $\tau = 0$ to the empirical value (4170 minutes) after $\tau = 540$. The number of trains running during this period is shown by the red curve. (b) Increase in lower bound runtime caused by each timestep of τ. Larger bars indicate larger increases in the lower bound and, thus, a costly change in the network state over the respective time interval.

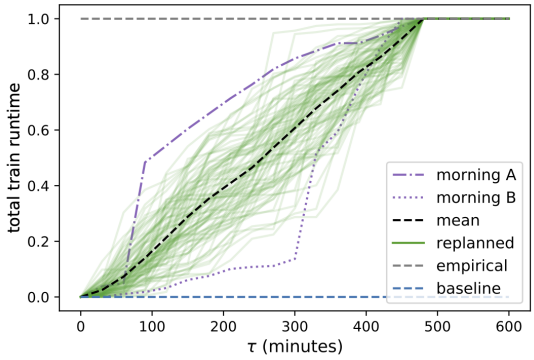


Figure 7: Accumulation of additional runtime due to replanning, shown for 90 windows of data, each of length 9 hours and overlapping by 6 hours. Each runtime value was min-max (baseline-empirical) normalized to [0, 1] for consistency. Different temporal patterns in the accumulation of delay can be seen in the green curves and the mean is shown by the black dashed line. Two dispatch windows, "morning A" and "morning B" are highlighted to demonstrate the vastly different patterns that occur.

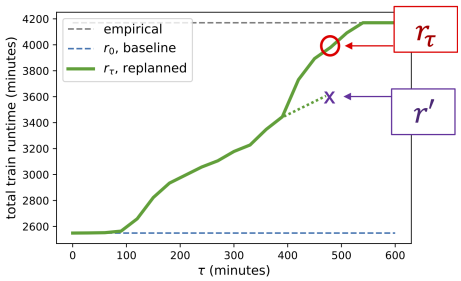


Figure 8: Representation of reducing lower bound runtime under replanning from its initial value of r_τ to a lower value r'.

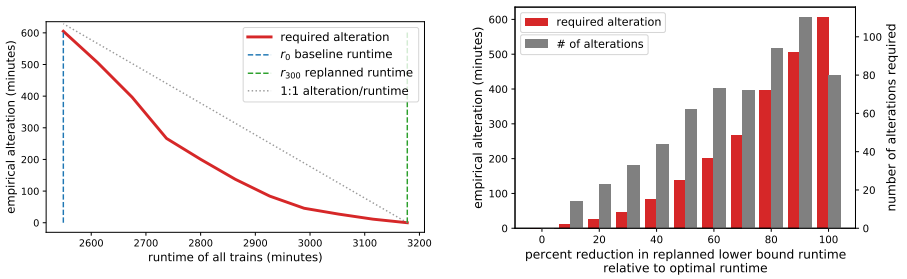
7 Results, Problem 2: alterations to dispatch decisions

The analysis question at hand in problem 2 is: which alterations could have been made to the current network state in order to reduce the lower bound runtime under future replanning? Given that an amount of time has elapsed and empirical decisions have been made up to time τ that increase the lower bound runtime by $\Delta = r_\tau - r_0$, what could have been done differently that would have reduced Δ? Figure 8 shows, graphically, the effect of reducing runtime under replanning from r_τ to a lower value, denoted r', achieved by altering the network state at τ while replanning into the future. We first address this question at a specific value of τ and then show results for other values of τ in the same dispatch window.

The alteration of empirical decisions to reduce the overall runtime value from the replanned dispatch lower bound presents two bounding cases: 1) if no reduction in the lower bound runtime, r_τ, is desired, then no alteration of the empirical decisions is required; 2) if a reduction in the objective value back to its baseline optimal value, r_0, is desired, then the empirical data must be changed all the way back to the baseline dispatch plan (assuming the baseline plan was uniquely optimal).

For reductions in the overall runtime value between these two cases, we minimize the amount of *alteration* of the empirical data that is required. As a reminder, an "alteration" is defined in (12) of the formulation of problem 2 as a change in a train's segment runtime, compared to its empirical value. Changing a segment runtime naturally shifts the timing points for all successive track segments by the same amount. The $\mathcal{L}_1$ norm on alterations in (12) promotes sparsity in the alterations that are found in the solution. That is, the objective function favors fewer alterations as opposed to the larger number of small alterations that would result from an $\mathcal{L}_2$ norm, for instance. The amount of *empirical alteration* is thus defined as the sum of these alterations and measured in minutes.

We first analyze the same 9-hour dispatch time period discussed in Section 6.



(a) (b)

Figure 9: (a) Required amount of alteration to $\tau = 300$ empirical data to reduce replanned runtime (approximately 3150 minutes, green dashed line) back towards its optimal value (approximately 2550 minutes, blue dashed line). (b) Comparison of the number of alterations and the magnitude of alterations required to reduce runtime at $\tau = 300$ from its replanned lower bound by the given percentage, relative to the optimal runtime for the dispatch scenario.

At time $\tau = 300$, we solve the minimum empirical alteration problem for runtime reduction values of 0% to 100% in increments of 10%. Figure 9a shows the amount of alteration required (red curve) to reduce the overall runtime from its replanned (at $\tau = 300$) lower bound value of 3178 minutes (green dashed line) to the baseline ($\tau = 0$) optimal value of 2549 minutes (blue dashed line). Take the point on the graph at 2990 minutes of runtime. This represents a 20% reduction in replanned runtime, and this would require an alteration to empirical segment runtimes of 45 minutes (for decisions up to $\tau = 300$).

The low slope of this curve at higher values of runtime (lower values of reduction) indicates that alterations to empirical data are yielding large effects on overall runtime. At the point where runtime is reduced to 2990 minutes, this reduction of 188 minutes was achieved by an alteration of 45 minutes, less than 25% of the magnitude. In this regime where the amount of empirical alteration produces a larger effect on the overall runtime, it must necessarily be affecting the secondary delay. Presuming that, in the 20% reduction case, the 45 minutes of alteration were all used to directly decrease (and not increase) runtime in the empirical trajectories, the alterations produced an additional 135-minute reduction in other runtimes that was the result of an improved ability to replan.

In Figure 9a, a 1:1 line for empirical alteration (y-axis) to runtime reduction (x-axis) is also shown. A set of empirical alterations which produce a decrease in overall runtime only as large as the alterations themselves would produce a slope on this plot the same as this 1:1 line. Therefore, we can interpret slopes of the alteration curve lower than this line as more effective, producing magnifying effects on runtime reduction. Slopes greater than this 1:1 line producing inefficient effects on runtime reduction – a minute of alteration produces less than one minute of overall runtime reduction.

An example of one of these magnifying changes is shown in Figure 10. This

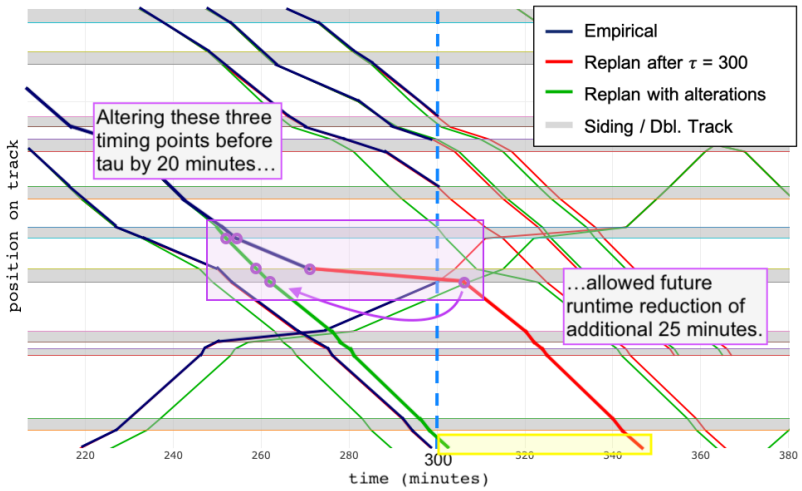


Figure 10: Example stringline diagram for reduction in replanning lower bound by 20%, compared to baseline plan, using empirical alteration. Lower bound replanning at $\tau = 300$ (blue dashed line) is shown by the red trajectories. The reduced lower bound with empirical alteration is shown in green. The purple box highlights alterations to the empirical data that were applied to a particular train (bold trajectory lines). These alterations (a total of 20 minutes change) would allow a meet event to occur earlier in time at a downstream siding, removing an additional delay of 25 minutes shown by the purple arrow. This train's runtime was reduced 45 minutes (yellow box) as a consequence.

stringline diagram shows two sets of trajectories: red trajectories are the replanned dispatch at $\tau = 300$, which accounts for empirical decisions made up to this time (marked by the blue dashed line); green trajectories are the replanned dispatch with some changes made to empirical data before $\tau = 300$ that decrease the overall runtime. In general, we should see green trajectories complete before red ones since the overall runtime is lower, but in some cases tradeoffs can be made where some trains experience increased runtime but the total decreases. Two purple boxes highlight changes to the empirical trajectories of one train that reduced its runtime before $\tau = 300$ and allowed it to make a more efficient meet event with an opposing train at siding further downstream, instead of waiting a long time at a closer siding. This change that was allowed is highlighted by the purple circle and arrow. Notice the downstream location where the meet between these two trains could occur, instead. This train's runtime is allowed to decrease over 40 minutes as a result (change shown in yellow).

The amount of empirical alteration, as described earlier, is measured by the magnitude of the $\mathcal{L}_1$ norm of changes made to track segment runtimes. This is the quantity that is minimized in the formulation for problem 2; the $\mathcal{L}_1$ norm also promotes sparsity in the alterations. The alterations highlighted

in problem 2 are perhaps most useful if they are small, unique decisions that could have significantly changed dispatching. For example, a sub-optimal meet location which introduced unnecessary delay is a decision that could be easily investigated and perhaps corrected. From the same test that was described above, which generated the required alteration curve in Figure 9a, we extract not only the magnitude of alteration, but the number of track segment runtimes that were changed. For each percentage reduction in lower bound runtime relative to optimal runtime (0% to 100%, increments of 10%), Figure 9b shows the number of distinct alterations (grey hatched bars) alongside the magnitude of alteration (red bars). A 100% reduction in this context corresponds to altering the empirical data such that the optimal plan and optimal overall runtime are feasible. We see that a runtime reduction of 20% can be achieved by changing 23 segment runtimes in the empirical data. While this is a modest number, the trend does not exhibit the same degree of positive effects that are seen with the magnitude of alteration; the number of alterations is more linear with respect to runtime reduction than the magnitude of alteration. We believe this reflects the fact that, in this instance under analysis, there are a significant number of smaller alterations that must be made to enable the schedule to return closer to optimal. This would tend to indicate that the suboptimality encountered in this instance is complex and interdependent, as opposed to the result of a single decision.

Thus far, we have presented correction results only for $\tau = 300$ minutes. Figure 11 shows the same curves for empirical alteration required to produce runtime reduction at values of $\tau = 120$, 180, 240, 300, 360, 420. The baseline lower bound is denoted by the blue dashed line. A similar trend is observed for values of τ on this same window of dispatch data. Large initial gains in overall runtime reduction (e.g., 10%, 20%) are possible with a very small degree of alteration to empirical data, after which returns diminish. The runtime reduction for $\tau = 420$ is slightly more aggressive for small amounts of empirical alteration. This could be due to a larger set of decisions that can be altered or greater secondary effects of those decisions that have propagated to other trains.

Figure 12 shows the magnitude of empirical alteration and the number of unique alterations for percentage reduction values in this same dispatch window. Interestingly, the 20% runtime reduction for $\tau = 420$ requires fewer alterations than the equivalent reduction for $\tau = 240, 300, 360$. This supports the idea that at $\tau = 420$ there are a larger set of impactful decisions that could be mitigated, or dispatch decisions have caused greater secondary effects that can be reduced by altering the initial decision.

8 Results, Problem 3: impact of individual trains

Problem 3 addresses the effects that a single train in the dispatch plan can have on other trains in the same plan. Specifically, what is the cost to the dispatch plan of fixing the trajectory of a single train to its empirical value? We first discuss results on a single window of dispatching data and analyze a

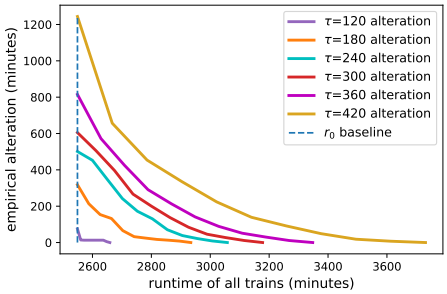


Figure 11: Required alteration to empirical data at various values of τ to shift replanned runtime (not shown) toward the optimal plan value of approximately 2500 (blue dashed line).

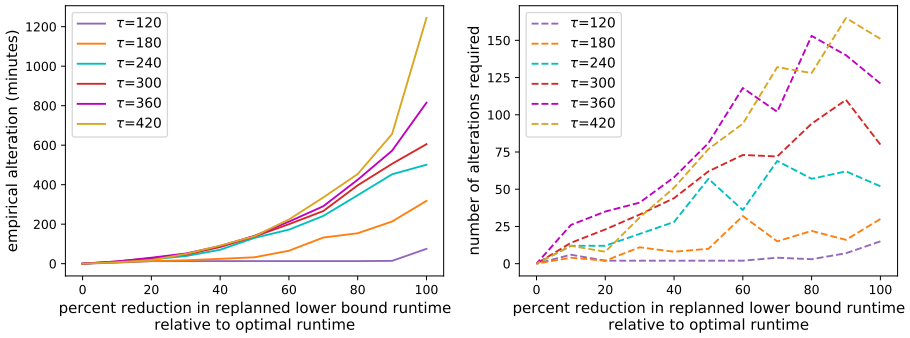


Figure 12: Comparison of the number of alterations (bottom) and the magnitude of alteration (top) required to reduce replanned runtime for this dispatch window by a given percentage, relative to the optimal runtime of the dispatch scenario. Values of τ from 120 to 420 minutes are shown.

specific train in detail within this time period. Then we look at broader trends exhibited by trains on this network segment within a month of dispatching.

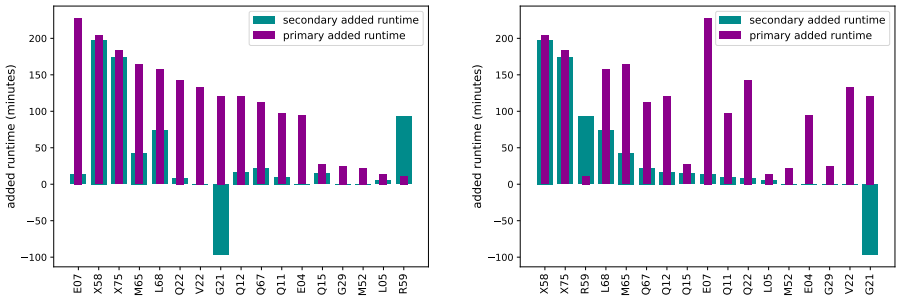
As discussed in Section 4.3, we define primary added runtime for a given fixed train w to be the difference between its empirical runtime, γ'_w and its baseline dispatch value, γ. Secondary added runtime is computed as the increase in runtime of all trains, minus the primary added runtime of train w: $(r_w - r_0) - (\gamma'_w - \gamma_w)$. It is possible for either the primary or secondary values to be negative, but not both; the sum of primary and secondary added runtime must be positive because the baseline dispatch plan was globally optimal and no possible trajectory for the fixed train may decrease the overall runtime value. In the case that a train runs faster than its optimal trajectory, it can do so only at the expense of increasing runtime for other trains by at least as much. In the case that secondary added runtime is negative, it is possible only because the fixed train experienced increased runtime.

8.1 Analysis on a single dispatch period

We run the fixed-train dispatch for each train in the 9-hour dispatch window during the week of January 4, 2016, as described in Section 5. Figure 13 shows the primary and secondary added runtime caused by each train when it was fixed in the dispatch. Figure 13a sorts the set of trains by primary added runtime. Two of the top three trains in terms of their own added runtime were the top two contributors to the added runtime of others, when fixed: X58 and X75. However, the train experiencing the greatest primary added runtime caused almost zero secondary effect. At the other end of the spectrum, train R59 experienced very little primary added runtime, but the secondary effect was much larger. This train will be analyzed in more detail, later.

In Figure 13b, the same set of trains are sorted according to secondary added runtime. At the opposite end of this spectrum, fixing train G21 caused a substantial decrease in runtime for other trains, but at the cost of its own runtime. Note that the net effect is still added runtime, but it shows that running the schedule around the departure time of train G21 slowed down the rest of the trains, significantly. This effect indicates that placing G21 on the network at a less sensitive moment could likely have improved overall performance.

Let us now analyze the case of train R59 in more detail. This train experienced a very small amount of added runtime with respect to the optimal dispatch plan, but this small primary effect caused an outsize secondary effect, over 5x larger. Consider the partial stringline diagram for train R59 in Figure 14, which shows a magnified area of the diagram in the vicinity of the train. Train R59 completes only 5 network segments and incurs 11 minutes of added runtime, shown by the red trajectory, relative to its optimal trajectory, which is shown in blue to its left. This shift in the trajectory causes the two trains following in the same direction, Q11 and G28, to run slower in order to maintain the required headway. Q11 was supposed to complete its run immediately after R59, before a train in the opposite direction departed. However, the delay required Q11 to wait out on the final siding for two trains in the opposite direction



(a) (b)

Figure 13: Comparison of primary added runtime incurred by trains and their minimum induced secondary added runtime on the optimal trajectories of other trains. Trains in the same window of data are sorted by primary added runtime (a), and by secondary added runtime (b).

to clear, before it could complete its route.

8.2 Results across multiple periods

The findings presented thus far concern only a single dispatch window with 17 trains, but can also be analyzed for a longer range of data. We now evaluate a shifting dispatch window of 9 hours over a month of data, January 2016, with overlap of 6 hours between windows. Since a train can be observed in multiple dispatch windows due to overlap, the primary and secondary added runtime effects are selected for each train from the dispatch window where secondary effect reaches its observed maximum absolute value. The distribution of secondary added runtime effects for all trains observed in the month is shown in Figure 15a, clipped to the upper limit of 800 minutes. A large number of trains cause over 3 hours (180 minutes) of secondary added runtime and a few, approximately 20, cause over 5 hours (300 minutes). These large secondary added runtime values are observed even in the presence of optimal replanning, caused just by the fixing of a single other train's trajectory. The trains causing these would merit further investigation into the configuration that caused such large secondary effects.

Secondary effects are highly dependent on the primary effect that induced them. We therefore consider the ratio of secondary to primary added runtime effects in Figure 15b. The vast majority of trains have less than a 2.0 ratio, meaning that secondary effects were less than twice the value of primary effects, but approximately 40 exceeded a 2.0 ratio value and a few were past a 3.0 ratio value, again indicating high relative impact of these trains on the schedule.

Finally, we observe in Figure 16 these primary-secondary pair values on a scatter plot. Primary added runtime is on the x-axis and secondary on the y-axis. A linear trendline for the dataset is shown by the dotted black line, which

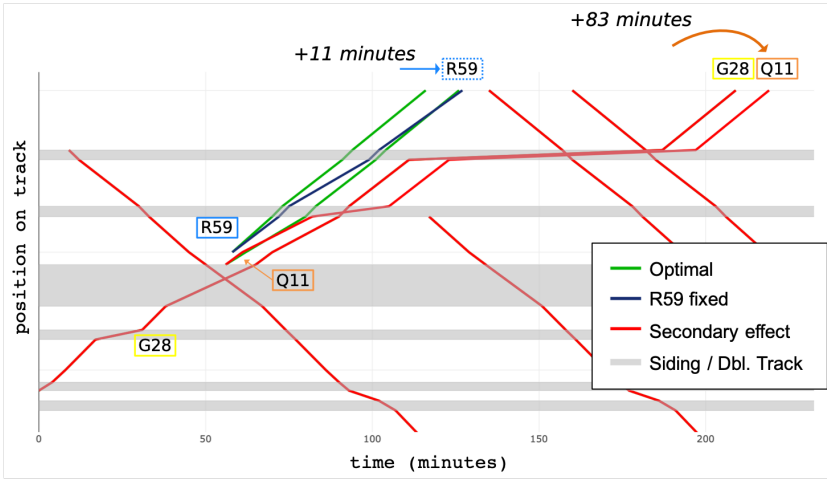


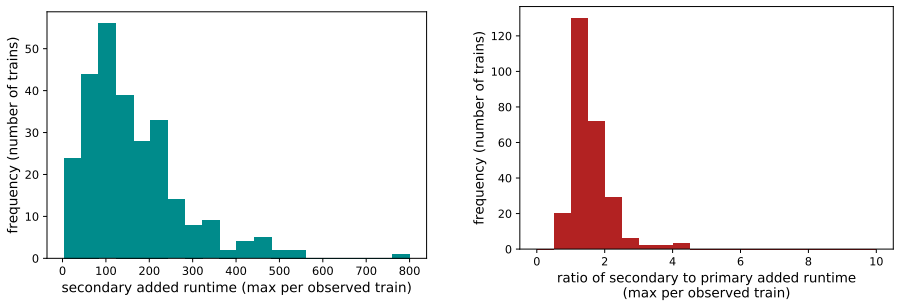
Figure 14: Portion of a stringline diagram with one train, R59, shown in red, fixed to its empirical data. The optimal plan is shown in blue, behind the replanned green trajectories that consider the fixing of train R59. An added 11 minutes of runtime for R59 resulted in a delay of at least 83 minutes for successive train Q11.

has a slope of 1.4 minutes/minute. This means that on average, a train adds 1.4 minutes of secondary runtime to other trains for every minute of its own runtime in excess of its optimal trajectory. The primary/secondary delay relationship will be affected by the volume and distribution of trains on the network at a given point in time and the particular dynamics or configuration of the network segment, itself, but is valuable for identifying problematic dependencies and train interactions and investigating those instances further.

9 Conclusion

In this article, we present a methodology by which to analyze empirical dispatch performance with regard to its optimal dispatch plan. This methodology is applied to answer three principal questions: How did a current network state contribute to a deterioration in the optimal dispatch plan? Which specific changes could have hypothetically been made to the network state that would have reduced the cost of replanned dispatch? Which trains' performance caused an effect on others in the schedule in terms of inducing additional runtime?

These three questions are addressed using the proposed *dispatch analysis problem*, which follows from a common form of optimization-based dispatching. The general form of the dispatch analysis problem accommodates each of these three questions by changing only the objective function and a few key constraints.



(a) (b)

Figure 15: Distribution of the secondary added runtime effects of trains (a) and the distribution of ratios of secondary/primary added runtime effects for each train (b).

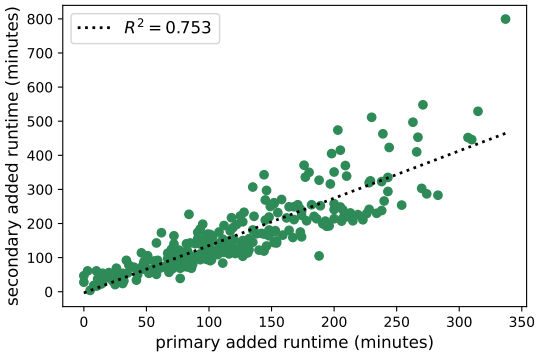


Figure 16: Scatter plot of primary (x-axis) versus secondary (y-axis) added runtime for a month of trains. Each point is a train's primary/secondary values in the dispatch window for which it exhibits the largest secondary added runtime value. The dotted black line is a linear trendline with $R^2 = 0.753$. Its slope is 1.4 minutes/minute, indicating 1.4 minutes of secondary added runtime for every minute of primary added runtime by each train.

We apply the methodology to the three questions identified and show, for question 1, that in identifying the deterioration of the optimal dispatch plan, it can construct a timeline of how the lower bound dispatch performance under optimal replanning increases. Sharp increases in this lower bound indicate decisions that were costly to the dispatch plan immediately and into the future. Each window of dispatching data exhibits a unique pattern around which this deterioration manifests, an indicator of which portions of the schedule or dispatching merit reevaluation and possible improvement. In application to question 2, for alteration of the empirical data to reduce sub-optimality of the replanned dispatch, it demonstrates that small changes to empirical data can have a magnifying effect on the reduction of total runtime. In one particular case, a magnifying effect of 4:1 was observed, meaning that, effectively, each minute saved in the past would have saved 4 minutes in the overall plan. Not all of these modifications would be strictly feasible due to features outside of the model, but reflect critical train interactions and portions of train performance that imposed outsized downstream consequences. Finally, the question 3 analysis of specific trains in a dispatch window revealed that trains have highly non-uniform effects on the schedule with regard to their impact or dependency on other trains. A small number of trains have secondary effects on the schedule that far exceed the effect of their own deviation from the optimal schedule. Addressing the performance or scheduling of these trains could free up possibilities in the schedule to reduce the runtime of other trains or make the schedule more resilient to delay.

Overall, we believe this methodology can serve to become a powerful analysis and examination engine for empirical dispatching practices. Being a natural extension of optimization-based models that exist in practice, it is generalizable to a variety of contexts that rely on these models and extensible to additional details and operating features. In future work, adding fidelity to the dispatching model could help create more realistic train trajectories via optimal dispatch. Looking at the relationship between train volume and dispatch performance relative to an optimal schedule could be valuable for capacity and infrastructure planning. Also, extending the application to other railroads and territories and analyzing aggregate effects could reveal larger trends or overall dispatching performance.

References

Barbour, W., Samal, C., Kuppa, S., Dubey, A., Work, D.B., 2018. On the data-driven prediction of arrival times for freight trains on us railroads, in: Proceedings of the IEEE 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2289–2296.

Boroun, M., Ramezani, S., Vasheghani Farahani, N., Hassannayebi, E., Abolmaali, S., Shakibayifar, M., 2020. An efficient heuristic method for joint optimization of train scheduling and stop planning on double-track railway

systems. INFOR: Information Systems and Operational Research 58, 652–679.

Carey, M., Kwieciński, A., 1994. Stochastic approximation to the effects of headways on knock-on delays of trains. Transportation Research Part B: Methodological 28, 251–267.

Corman, F., D'ariano, A., 2012. Assessment of advanced dispatching measures for recovering disrupted railway traffic situations. Transportation Research Record 2289, 1–9.

Corman, F., Quaglietta, E., Goverde, R.M., 2018. Automated real-time railway traffic control: an experimental analysis of reliability, resilience and robustness. Transportation planning and technology 41, 421–447.

Daamen, W., Goverde, R.M., Hansen, I.A., 2009. Non-discriminatory automatic registration of knock-on train delays. Networks and Spatial Economics 9, 47–61.

Dick, C.T., Mussanov, D., 2016. Operational schedule flexibility and infrastructure investment: capacity trade-off on single-track railways. Transportation Research Record 2546, 1–8.

Dingler, M., Koenig, A., Sogin, S., Barkan, C.P., 2010. Determining the causes of train delay, in: AREMA Annual Conference Proceedings.

Fang, W., Yang, S., Yao, X., 2015. A survey on problem models and solution approaches to rescheduling in railway networks. Transactions on Intelligent Transportation Systems 16, 2997–3016.

Gestrelius, S., Aronsson, M., Forsgren, M., Dahlberg, H., 2012. On the delivery robustness of train timetables with respect to production replanning possibilities .

Gorman, M.F., 2009. Statistical estimation of railroad congestion delay. Transportation Research Part E: Logistics and Transportation Review 45, 446–456.

Hallowell, S.F., Harker, P.T., 1998. Predicting on-time performance in scheduled railroad operations: Methodology and application to train scheduling. Transportation Research Part A: Policy and Practice 32, 279–295.

Hansen, I.A., Goverde, R.M., van der Meer, D.J., 2010. Online train delay recognition and running time prediction, in: 13th International IEEE Conference on Intelligent Transportation Systems, IEEE. pp. 1783–1788.

Hwang, C.C., Liu, J.R., 2009. A simulation model for estimating knock-on delay of taiwan regional railway, in: Proceedings of the Eastern Asia Society for Transportation Studies Vol. 7 (The 8th International Conference of Eastern Asia Society for Transportation Studies, 2009), Eastern Asia Society for Transportation Studies. pp. 213–213.

Lamorgese, L., Mannino, C., 2015. An exact decomposition approach for the real-time train dispatching problem. Operations Research 63, 48–64.

Lovett, A.H., Dick, C.T., Barkan, C.P., 2015. Determining freight train delay costs on railroad lines in North America, in: Proceedings of RailTokyo2015: 5th International Conference on Railway Operations Modelling and Analysis, International Association of Railway Operations Research (IAROR).

Lovett, A.H., Dick, C.T., Barkan, C.P., 2017. Predicting the cost and operational impacts of slow orders on rail lines in north america, in: Proceeding of the 7th International Conference on Railway Operations Modelling and Analysis (RailLille2017), Lille, France, 4-7 April 2017.

Lu, Q., Dessouky, M., Leachman, R.C., 2004. Modeling train movements through complex rail networks. ACM Transactions on Modeling and Computer Simulation (TOMACS) 14, 48–75.

Luan, X., De Schutter, B., Meng, L., Corman, F., 2020. Decomposition and distributed optimization of real-time traffic management for large-scale railway networks. Transportation Research Part B: Methodological 141, 72–97.

Lusby, R.M., Larsen, J., Bull, S., 2018. A survey on robustness in railway planning. European Journal of Operational Research 266, 1–15.

Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N., 2013. A fuzzy petri net model to estimate train delays. Simulation Modelling Practice and Theory 33, 144–157.

Murali, P., Dessouky, M., Ordóñez, F., Palmer, K., 2010. A delay estimation technique for single and double-track railroads. Transportation Research Part E: Logistics and Transportation Review 46, 483–495.

Mussanov, D., Nishio, N., Dick, C.T., 2017. Delay performance of different train types under combinations of structured and flexible operations on single-track railway lines in north america, in: Proceedings of the International Association of Railway Operations Research (IAROR) 7th International Seminar on Railway Operations Modelling and Analysis.

Narayanaswami, S., Rangaraj, N., 2011. Scheduling and rescheduling of railway operations: A review and expository analysis. Technology Operation Management 2, 102–122.

Pellegrini, P., Marlière, G., Rodriguez, J., 2016. A detailed analysis of the actual impact of real-time railway traffic management optimization. Journal of Rail Transport Planning & Management 6, 13–31.

Petersen, E., Taylor, A., Martland, C., 1986. An introduction to computer-assisted train dispatch. Journal of Advanced Transportation 20, 63–72.

Quaglietta, E., Corman, F., Goverde, R.M., 2013. Stability analysis of railway dispatching plans in a stochastic and dynamic environment. Journal of Rail Transport Planning & Management 3, 137–149.

Diaz de Rivera, A., Dick, C.T., Evans, L.E., 2020. Potential for moving blocks and train fleets to enable faster train meets on single-track rail corridors. Journal of Transportation Engineering, Part A: Systems 146, 04020077.

Salido, M.A., Barber, F., Ingolotti, L., 2008. Robustness in railway transportation scheduling, in: 2008 7th World Congress on Intelligent Control and Automation, IEEE. pp. 2880–2885.

Sehitoglu, T., Mussanov, D., Dick, C.T., 2018. Operational schedule flexibility, train velocity and the performance reliability of single-track railways, in: Proceedings of the transportation research board 97th annual conference.

Shakibayifar, M., Sheikholeslami, A., Corman, F., Hassannayebi, E., 2020. An integrated rescheduling model for minimizing train delays in the case of line blockage. Operational Research 20, 59–87.

Törnquist, J., 2007. Railway traffic disturbance management—an experimental analysis of disturbance complexity, management objectives and limitations in planning horizon. Transportation Research Part A: Policy and Practice 41, 249–266.

Yuan, J., Hansen, I.A., 2007. Optimizing capacity utilization of stations by estimating knock-on train delays. Transportation Research Part B: Methodological 41, 202–217.