This is the accepted version of a paper presented at Transportation Research Board 86th Annual Meeting.

Citation for the original published paper:

Qin, Y., Ma, X., Jiang, S. (2017)
A stochastic process approach for modeling arrival delay in train operations.
In:

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-201030
A STOCHASTIC PROCESS APPROACH FOR MODELING ARRIVAL DELAY IN TRAIN OPERATIONS

Yuanqi Qin
System Simulation & Control,
Department of Transport Science,
KTH Royal Institute of Technology,
Teknikringen 10, Stockholm, Sweden.

Xiaoliang Ma*
System Simulation & Control,
Department of Transport Science
KTH Royal Institute of Technology
Teknikringen 10, Stockholm, Sweden
Email: liang@kth.se

Sida Jiang
WSP Sverige Analysis & Strategy,
Arenavägen 7, Stockholm, Sweden.

*Corresponding author.

Word Count: 5085 words + 4 figure(s) + 5 table(s) = 7335 words

Submission Date: November 15, 2016
ABSTRACT

Compared with air and road transport, railway transport nowadays plays an essential role in catering for continuously increasing travel demand in Sweden given its high capacity and convenience. However, as the demand approaches the maximum capacity, issues like delay propagations undermine the punctuality and reliability of railway systems. Nevertheless, punctuality in train operation is usually considered as the major indicator when passengers evaluate the service performance of railway transport. This study, therefore, attempts to quantify the relationship between the arrival delays of passenger trains and their influential factors in operation and weather condition. A stochastic modeling approach based on Wiener process is introduced to analyze the train operation. The model parameters are estimated by the maximum likelihood estimation approach. A case study is carried out using data of two rail paths connecting two main railway stations, Stockholm Center and Örebro Center over a one-year period in 2009. Three data sources, consisting of train movement information, rail failure data, and weather data, are applied in this study. Models are specified and then estimated using part of the available data. Finally, several goodness-of-fit measures are evaluated to validate and compare different model specifications using an independent dataset. As a result, it is found that the predicted arrival delays have a similar tendency to the actual arrival delays. The estimation results of the model taking weather condition into account fit the actual arrival delays more precisely, largely supporting the hypotheses proposed.
INTRODUCTION

Disruptions in railway transport often lead to significant delays, not only disorganizing passengers’ travel plan but also leading to additional social economic cost. Apart from comfort and ticket price, rail passengers usually consider train delays as an important quantitative indicator on service performance. Moreover, facing the challenge of augmented travel demand, the railway operation business has to utilize the infrastructure more efficiently and create additional capacity by rescheduling traffic. This measurement is usually implemented with a trade-off which negatively impacts the punctuality and reliability of train operation.

There are different factors that may delay train arrival times. Infrastructure malfunction is among the most common problems. In this study, rail malfunction consists of two components, infrastructure failures, and non-infrastructure failures. The former one can be further categorized into different failure types such as signal, electricity, communication etc. This paper defines train departure delay at a station as the difference between the actual departure time and the scheduled departure time at a station. Correspondingly, train arrival delay is defined as the difference between the actual arrival time and the scheduled arrival time at a station.

This work focuses on analyzing late arrival delay whereas early arrival is not considered. The train arrival delay is calculated by

\[ \Delta T^a = t^a - t^s - \Delta T^d_p \] (1)

where \( \Delta T^a \) is the arrival delay at a station; \( t^a \) is the actual arrival time at the station; \( t^s \) is the scheduled arrival time in train operation timetable. In addition to the deviation of actual arrival time from schedule, the departure delay from previous station, \( \Delta T^d_p \), is also introduced in the definition of the arrival delay.

The objective of this study is to analyze and model train arrival delays using rail operational data and weather information. The result may indicate essential factors that affect the reliability of train schedule. Further research may lead to better practice on train operations concerning e.g. delay management and so on. A stochastic process based modeling approach is introduced in the paper after relevant literature is reviewed in the next section. A case study based on Swedish rail operation data and corresponding weather data is then conducted to analyze the influential factors for train arrival delay. The final section summarizes the results and discuss the indication and limitation of the study.

RELATED WORK

As mentioned before, train delays perturb passengers’ travel plan, but also affect subsequent train schedule in the related railway network. Moreover, according to Taylor (1), a severe delay could significantly undermine transport service performance and increase passengers’ negative emotions. Berger et al. (2) state that passengers often consider train delay as an indicator of poor performance, which becomes the main reason that leads to the public transport complaints.

Due to the emphasis of train punctuality, investments in rail infrastructure are mostly spent for reducing travel time and improving reliability. Moreover, the study of travelers’ valuation of travel time reliability is beneficial for the cost-benefit analysis of rail investment. Several studies investigated the valuation of travel time saving including Hensher (3) and Jara-Díaz (4). Mackie et al. (5) focus on whether and how the value of travel time should be used within the evaluation. Börjesson and Eliasson (6) investigate passengers’ valuation of unexpected delays about scheduled travel time and travel costs.
Furthermore, research topics related to punctuality factors are also attractive to scholars. Harris \cite{harris1981} uses least-squares multiple linear regression to study how a number of factors impact punctuality in the UK and the Netherlands. Influencing factors taken into account consist of train length, distance covered, previous number of stops, age of motive power unit and track occupation. As a result, only train length and distance covered are statistically significant in affecting punctuality. The Swedish National Audit Office (Riksrevisionsverket \cite{swedish2002}) analyzes the monthly punctuality statistics with a finding that precipitation, average temperature and the number of travelers are correlated to around half of train delays. They also make a conclusion that the dwell time in stations should depend on the time needed during low-traffic periods, not peak hours. Veiseth et al. \cite{veiseth2007} estimate the effects of delay-causing incidents and analyze the possibility of identifying root causes for delays, especially due to infrastructure failures, by combining infrastructure data and operational data. The outputs of their research could be used not only to prioritize improvement and maintenance activities but to quantify the effects of improvement measures.

In literature, delay is involved only if trains do not arrive at a specific station within a certain time. For example, it was mentioned by Yuan \cite{yuan2000} that trains arrive less than 5 minutes late are normally not considered as a delay in many European countries. The study proposed to distinguish two types of delays: primary delay and secondary delay (also called knock-on delay). The primary delays are often caused by technical failures, bad weather conditions, and hence running at a lower speed compared with train schedule. The secondary delay refers to the condition that the scheduled train arrival and departure can not be accomplished on time because of internal factors. For instance, the track may be occupied by another train. However, the secondary delay is usually excluded since it is hard to be measured in reality.

Approaches used in predicting and reducing train delay are many in literature. For instance, Weston et al. \cite{weston2004} studied in a congested part of the U.K. rail network, with both passenger trains and freight trains sharing the same track. They target examining if the same strategy is appropriate for distinct classes of the train that operate on the common track resources. Their proposed simulation is used to evaluate the reduction in various train delay algorithm. Keyhani et al. \cite{keyhani2005} present an effective probabilistic approach by carrying out experiments on real customer queries and timetables of all trains in Germany. This approach is used to estimate the reliability of train connections in a large rail network. Furthermore, the reliability ratings of their study are computed by validating the predicted delays with actual delays from German Railways.

Modern railway signaling and train control technologies make it possible to implement advanced real-time railway management. Optimization algorithms can be utilized to minimize train delays. Fan \cite{fan2006} compares some optimization approaches and then, based on ant colony optimization, develops a new hybrid algorithm which has a better performance concerning optimality and computational speed. Peters et al. \cite{peters2008} develop a model for real-time timetable monitoring and optimization. The system they proposed can process existing delays in the network to generate predicted delays shortly. Furthermore, this rule-based system is also used as a comparison to the specially developed neural network so as to evaluate the accuracy of an artificially intelligent component.

Considering factors that probably affect train arrival delay, some related studies have been accomplished. Heinz \cite{heinz2009} studies on passengers’ boarding and alighting time at railway stations. An important finding from his study is that variation of train type makes differences on the time that trains spend at stations. Hence, a high variation of boarding and alighting time could, to some extent, lead to a high risk of train delay. It is well recognized that high utilization of rail infrastruc-
ture capacity could reduce the punctuality. Gibson et al. (16) investigated the relationship between
capacity utilization of rail infrastructure and train delays in the UK. Although, in their study, no
significant relationship between capacity utilization and train delays is found they state that there
exists an exponential correlation between the level of congestion and the level of performance on
the railway network.

Weather factors, particularly some extreme weather, have been considered in the study of
train arrival delay. Leviäkangas et al. (17) investigated how the extreme weather affects transport
systems. In their work report, they state that extreme weather may lead to consequences that impact
main quality dimensions for rail transport. Besides, they hold the view that, in winter, long-lasting
low temperatures, snowfall, and strong wind gusts are considered as the most harmful weather
phenomena for rail transport. Some other kinds of harmful weather, such as thunderstorms, winter
storms/snow, and cold waves are also involved in their report. Also, they mention that the incre-
ment of precipitation may increase runoff and flooding, which could do harm to low-lying railways.
In the study of Juntti (18), they drew a conclusion that bad weather affects the performance of rail-
way infrastructure. Furthermore, they also stated that switches and crossings and road crossing are
vulnerable to heavy snow and ice. Similarly, Stenström et al. (19) successfully conducted a case
study of a railway in Sweden. They concluded that the cold Nordic climate has a direct impact on
the reliability of railway infrastructure, which, in turn, influences the quality of service of assets
and capacity of railway transport system.

Although the causes of train delay have been extensively studied in literatures, there is
little effort in modeling train delay using train operation and other data. This might be due to the
limited availability of train operation data. This study focuses on developing a stochastic approach
for modeling train delays. The diffusion models, especially Brownian motion, has gained wide
application in finance, economics, and engineering fields. For example, Gerstein and Mandelbrot
(20) propose the first simple diffusion model of spike activity of the neuron. Later on, Gluss
(21) considers an exponential membrane decay in the model with a diffusion and a linear drift.
In fact, Wiener diffusion process (WDP) can be implemented in rail transportation as well. The
departure and destination stations can be considered respectively as the start point and end point
of a continuous Wiener process. The average train speed and detailed train positions can also be
involved in the process formulation.

**METHODOLOGY**

This section proposes to model train arrival delays using a Wiener process based approach. We
will start with the introduction of Wiener process and modeling of first hitting time in the context
of train operations.

**First hitting time model**

In stochastic process modeling, the first hitting time refers to the required time for a random pro-
cess, starting from an initial state until encountering a threshold for the first time. When the
threshold is firstly reached an event is usually triggered. For instance, the first hitting time of
Wiener diffusion process is frequently applied to model the probability of financial ruin in finance
and economics.

Let’s consider a Wiener diffusion process, and it starts with a location \( c > 0 \) and ends at 0
after the first hitting time \( T \). Indeed, such a stochastic process can be mathematically represented
by

\[ X(t) = c - \mu t + \sigma W(t) \]  \hspace{1cm} (2)

where \( X(t) \) in this study is the location of mass (train in our case) at time \( t \); \( c \) is the initial value. In our case, \( X(0) = c \) represents the distance between the origin station and destination station; \( \mu \) is the drift coefficient which is regarded as the average speed of train in each segmentation (\( \mu > 0 \)); \( \sigma \) is the diffusion coefficient which can be considered as the uncertainty of train location \( \sigma > 0 \), i.e.,

\[ \text{Var}[X(t)] = \sigma^2 t. \]  \hspace{1cm} (3)

In the model, \( W(t) \) is the random term and represented by a standard Wiener process (often called Brownian motion), being continuous on \( t \in [0, T] \) and satisfying the following properties:

- \( W(0) = 0 \);
- \( W(t) - W(s) \sim \sqrt{t-s} \in N(0, 1) \), for \( 0 \leq s < t \leq T \) where \( N(0, 1) \) is a zero-mean normal distribution with unit variance;
- \( W(t) - W(s) \) and \( W(v) - W(u) \) are independent, for \( 0 \leq s < t < u < v \leq T \).

The first hitting time is analytically defined by

\[ T = \inf \{ T : X(T) = 0 \} \]  \hspace{1cm} (4)

which is often called "the first hitting time to absorption to 0" in literature (22). The distribution of the process defined by Equation [2] is shown to fulfill an inverse Gaussian distribution with the probabilistic density function as follow:

\[ P(T) = \frac{c}{\sqrt{2\pi\sigma^2 T^3}} \exp \left\{ \frac{-(c - \mu T)^2}{2\sigma^2 T} \right\}. \]  \hspace{1cm} (5)

But the standard probability density function of inverse Gaussian distribution has a form of

\[ f(T; \mu_h, \lambda_h) = \left[ \frac{\lambda_h}{2\pi T^3} \right]^{\frac{1}{2}} \exp \left\{ \frac{-\lambda_h(T - \mu_h)^2}{2\mu_h^2 T} \right\} \]  \hspace{1cm} (6)

with mean \( \mu_h \) and shape parameter \( \lambda_h \). Hence, it can be shown that the probability characteristics of the first hitting time can be derived as follows:

\[ \mathbb{E}[T] = \mu_h = \frac{c}{\mu} \]  \hspace{1cm} (7)

\[ \lambda_h = \left( \frac{c}{\sigma} \right)^2 \]  \hspace{1cm} (8)

\[ \text{Var}[T] = \frac{\mu_h^3}{\lambda_h} = \frac{c\sigma^2}{\mu^3}. \]  \hspace{1cm} (9)
FIGURE 1: Illustration of the Wiener diffusion processes in train operation.

TABLE 1: Classification of railways according to their features.

<table>
<thead>
<tr>
<th>Rail class</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>Information of the rail class is not available.</td>
</tr>
<tr>
<td>1</td>
<td>Transportation center with many trains, heavy traffic.</td>
</tr>
<tr>
<td>2</td>
<td>Around transportation center and major freight routes.</td>
</tr>
<tr>
<td>3</td>
<td>Major freight routes and areas of important trains flows.</td>
</tr>
<tr>
<td>4</td>
<td>Freight routes plus parts of major freight flows.</td>
</tr>
<tr>
<td>5</td>
<td>Low volume at very low density lines.</td>
</tr>
</tbody>
</table>

**Train operation problem**

The research problem can be implied using a train operation process described in Figure 1. The process starts when a passenger train travels from its origin station, and ends when it arrives its destination via $n$ intermediate stations. The entire journey is divided into $n+1$ segments, each of which is modeled as an elementary Wiener diffusion process with first hitting time as travel time on a segment.

The general idea is then to model such a process with observable potential factors. The observations for each segment consist of track length, rail class, occurred malfunctions, departure delay and arrival time. The first four variables can be considered as model input factors, whereas arrival time can be used to calculate travel time (first hitting time modeled by Equation 2), and $j$ is the observation index. The classification of railways is explained in Table 1, whereas the involved railway classes in this study only include class 1, class 2, class 3 and class NA. Different rail classes refer to railways that are distributed in different areas with distinct features. For instance, railways that are located around transportation centers usually have heavier traffic and higher occupancy,  

---

1Source: Swedish Transport Administration, http://www.trafikverket.se.
which means the occurrence probability of arrival delay should be higher than that of less busy railways in practice.

In addition, train type is considered as an essential factor since X2000 and regional train operate in different speed levels and have a different priority to pass when these two kinds of trains meet during the operation process. Rail malfunction directly leads to additional operation time due to the need of reparation. So it is another key variable that should be mainly explored in this study. Besides, since the departure delay may potentially affect the train operation, it should not be neglected in model building.

Summarizing these potential factors, the model of average speed $\mu$ and location uncertainty $\sigma$ in the corresponding observation can be regressed as

$$\hat{\mu} = \alpha \cdot z \quad (10)$$

$$\hat{\sigma} = \exp \{ \beta \cdot z \}$$

where $\alpha$ and $\beta$ are vectors of model parameters and should be estimated by real operation data. $z$ is a vector of model inputs. It should be mentioned that exponential transform is applied for the model of $\sigma$ to assure non-negative prediction value. Consequently, models of $\mu$ and $\sigma$ can be applied to predict the arrival time and then arrival delay based on Equation 1.

15 Model identification approach

Given that the stochastic distribution of train travel time is modeled, it is natural to resort to the maximum likelihood estimation (MLE) approach to identify the model parameters. The set of observations in this study is represented by $z_j$ including rail classes, rail malfunction, train type and other factors (such as weather) as well as corresponding train departure and arrival times.

For each process between two stations, the likelihood for observing the data given the model form of Equation 5 and parameter set $\theta = \{ \alpha, \beta \}$ can be analytically represented by

$$P(z_j | \theta) = \frac{c_j}{\sqrt{2\pi \sigma_j^2 T_j^3}} \exp \left\{ -\frac{(c_j - \mu_j T_j)^2}{2\sigma_j^2 T_j} \right\}. \quad (11)$$

Both $\mu_j$ and $\sigma_j$ can be obtained using models of Equation 10. $T_j$ is the train travel time that can be calculated by train operation times i.e.

$$T_j = t_j^a - t_j^d \quad (12)$$

where $t_j^d$ is actually departure time from initial position $c_j$.

As all the observations are assumed identically and independently distributed (i.i.d) the likelihood function of the whole dataset is, therefore, the joint product of the likelihood of each observation

$$P(z_1, z_2, ..., z_M | \theta) = \prod_{j=1}^{M} P(z_j | \theta) \quad (13)$$

$$= \prod_{j=1}^{M} \frac{c_j}{\sqrt{2\pi \sigma_j^2 T_j^3}} \exp \left\{ -\frac{(c_j - \mu_j T_j)^2}{2\sigma_j^2 T_j} \right\}.$$
where $M$ is the size of the model estimation dataset. So the final log-likelihood function is written as follows

$$LL(\theta) = -\frac{M}{2} \log(2\pi) + \sum_{j=1}^{M} \log c_j - \frac{1}{2} \sum_{j=1}^{M} \log(T_j \sigma_j^2) - \sum_{j=1}^{M} \frac{(c_j - \mu_j T_j)^2}{2\sigma_j^2 T_j}. \quad (14)$$

Finally, the parameter estimator is obtained by

$$\hat{\theta} = \arg \max_{\theta} LL(\theta). \quad (15)$$

**Performance measures**

When the model parameters are identified using a dataset, an independent dataset should be used to evaluate the developed model. So statistical measures are applied for such evaluation. The predicted train travel time ($\hat{T}$) is compared with the actual operation time ($\tilde{T}$) using two goodness-of-fit measures, the mean absolute percentage error (MAPE) and the root-mean-square error (RMSE).

The equations of MAPE and RMSE are shown as follow

$$\text{MAPE} = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\tilde{T}_n - \hat{T}_n}{\tilde{T}_n} \right| \quad (16)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\tilde{T}_n - \hat{T}_n)^2} \quad (17)$$

where $N$ is the size of the validation dataset. These two measures describe the deviation of the measured values from prediction by models.

Additionally, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are usually calculated respectively when comparing models. Also, AIC can deal with the trade-off between the goodness of fit of models. Similar to AIC, BIC is based on the likelihood function. Both AIC and BIC can identify the over-fitting cases when the increased number of model parameters results in increased likelihood value. AIC and BIC can be calculated by

$$\text{AIC} = 2k - 2 \cdot LL(\hat{\theta}) \quad (18)$$

$$\text{BIC} = -2 \cdot LL(\hat{\theta}) + k \cdot \log(N) \quad (19)$$

where $LL(\hat{\theta})$ is the maximum value of the likelihood function; $k$ is the number of estimated parameters in the model; $N$ is the number of validation observations.

**CASE STUDY**

**Data and basic setup**

A case study aiming to model the relationship between train arrival delay and rail operation factors is carried out. Three related data sources provide relevant data including

- Train movement record database - TFÖR
- Failure record database - 0FELIA
- Weather record database - VViS.
FIGURE 2: Train operation data from two railway paths between Stockholm and Örebro are used in our case study.

TFÖR provides the train information, such as train number, train type, rail class, departure delays, segment length etc. 0FELIA records the number of occurred failures in each segment along the train trip. VViS is the resource of weather information, which contains precipitation type, road surface temperature, air temperature etc.

It has been reported that the Western Main Line (WML) in Sweden connecting Stockholm and Göteborg is heavily used for mixed rail traffic. Punctuality is known to be influenced especially because high-speed trains (X2000) operates on this line mixing with regional passenger trains, freight trains, etc. On the other hand, train arrival delays due to local congestion situations also provide interesting data in other lines. Therefore, data from two railway paths between Stockholm Center and Örebro Center are selected in this case study.

Figure 2 depicts the two railway paths on Google map. The north path starts from Stockholm Center station to Örebro Center station via Västerås station. The south path includes one segment of WML line, from Stockholm Center station to Hallberg Station, and the rest connects Hallsberg Station to Örebro Center station. It should be mentioned that trains using the north path run between Stockholm and Örebro without any transfer process, even though there are 16 intermediate stations in between. However, passengers who choose the south path have to transfer to another train to Örebro station in Hallsberg.

In our data processing, trains running on the north path are all included due to shorter travel time and better convenience. The average travel time, from Stockholm Center station to Örebro Center station, of all the trains on the northern path is computed as 113 min.

As for trains operating on the south path, a transfer process at Hallsberg Station should
be included in the trip. Hence, the scheduled travel time of those trains is the summation of
train operation time and transfer time. The "train-pair" concept is then introduced to illustrate the
corresponding trains that can combine for an integrated trip between Stockholm Center station and
Örebro Center station. Several rules are then applied in our data processing:

• trains with a scheduled travel time longer than 160 minutes are not taken into account;
• the transfer time at Hallsberg Station should not be less than 15 minutes;
• If a train pair A has earlier departure time and later arrival time than the train pair B, the
data of train pair A are removed for our modeling work.

The data selection procedures mentioned above are conducted using R, a statistical programming
language.

Modeling
Although the train position \( X(t) \) between origin and destination can be considered as a general
stochastic process, each segment between intermediate stations is treated as a basic Wiener process.
This indeed assumes that train operations at individual segments are independent. In our studies,
two models are developed for predicting train arrival delay.

The first model is developed with basic train operation data mentioned in the previous
section. This includes railway class information, infrastructure failure data, train type and finally
departure delay. As discussed before, different rail classes are distributed in different segments.
For instance, rail class 1 is often near the transportation center with heavier traffic than other rail
classes. Infrastructure failures are usually considered as a negative factor that gives rise to extra
travel time. Hence, several assumptions are introduced as follows:

• Trains near transportation center are more probably suffer from delays due to the busy
traffic;
• Rail malfunctions reduce the average train speed and increase the train arrival delay;
• Trains with long departure delays usually disorganize operation plan and thus suffer from
significant impact on their speed.

To describe the model more concisely, all the involved variables are summarized in Table 2.
Model 1 regresses train average speed and location uncertainty by the following equations:

\[
\mu_j = \alpha_0 + \alpha_1 RC_1j + \alpha_2 RC_2j + \alpha_3 RC_3j + \alpha_4 FA_j + \alpha_5 FT_j + \alpha_6 DD_j \\
\sigma_j = \exp \{ \beta_0 + \beta_1 RC_1j + \beta_2 RC_2j + \beta_3 RC_3j + \beta_4 FA_j + \beta_5 FT_j + \beta_6 DD_j \} 
\]

where \( FA_j \) is a dummy variable indicating whether infrastructure failure occurs or not and \( j \) is data
index; \( FT_j \) is dummy variable and refers to high speed trains (X2000 trains) that run with a top
speed of 210 km/h, i.e.,

\[
FT_j = \begin{cases} 
1, & \text{if the train is an X2000 train;} \\
0, & \text{if the train is a regional train.}
\end{cases}
\]
TABLE 2: List of Variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable name</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC₁</td>
<td>Rail class 1</td>
</tr>
<tr>
<td>RC₂</td>
<td>Rail class 2</td>
</tr>
<tr>
<td>RC₃</td>
<td>Rail class 3</td>
</tr>
<tr>
<td>FA</td>
<td>Rail failures</td>
</tr>
<tr>
<td>FT</td>
<td>X2000 train</td>
</tr>
<tr>
<td>DD</td>
<td>Departure delay</td>
</tr>
<tr>
<td>W₁</td>
<td>No precipitation</td>
</tr>
<tr>
<td>W₂</td>
<td>Snowy weather</td>
</tr>
<tr>
<td>W₃</td>
<td>Rainy weather</td>
</tr>
<tr>
<td>RT₁</td>
<td>Road surface temperature range [-20°C, -10°C]</td>
</tr>
<tr>
<td>RT₂</td>
<td>Road surface temperature range [-10°C, 0°C]</td>
</tr>
<tr>
<td>RT₃</td>
<td>Road surface temperature range [0°C, 10°C]</td>
</tr>
<tr>
<td>RT₄</td>
<td>Road surface temperature range [10°C, 20°C]</td>
</tr>
<tr>
<td>RT₅</td>
<td>Road surface temperature range [20°C, 30°C]</td>
</tr>
<tr>
<td>RT₆</td>
<td>Road surface temperature range [30°C, 40°C]</td>
</tr>
<tr>
<td>RT₇</td>
<td>Road surface temperature range [40°C, 50°C]</td>
</tr>
</tbody>
</table>

1. RCₖ represents railway class, denoted as dummy variables

\[
RC_k = \begin{cases} 
1, & \text{if railway class } = k; \\
0, & \text{otherwise}. 
\end{cases}
\]  

(23)

where \(k\) is 1 or 2 or 3. \(\alpha\) and \(\beta\) are model parameters.

2. The second model proposed considers not only operation factors but also impacts of weather.

3. Several weather-related variables, such as precipitation type and road surface temperature, are added in the model specification as follows:

\[
\mu_j = \alpha_0 + \sum_{k=1}^{3} \alpha_k RC_{kj} + \alpha_4 FA_j + \alpha_5 FT_j + \alpha_6 DD_j + \sum_{k=1, k \neq 4}^{6} \alpha_{k+6} RT_{kj} + \sum_{k=1}^{3} \alpha_{k+11} W_{kj} 
\]  

(24)

\[
\sigma_j = \exp \left\{ \beta_0 + \sum_{k=1}^{3} \beta_k RC_{kj} + \beta_4 FA_j + \beta_5 FT_j + \beta_6 DD_j + \sum_{k=1}^{3} \beta_{k+6} RT_{kj} + \sum_{k=1}^{3} \beta_{k+9} W_{kj} \right\} 
\]  

(25)

4. where \(ST_k\) refers to seven categories of road-surface temperature (degree of centigrade). Basically, the categories are created every 10 degree. They are all dummy variables. \(W_1\) is a dummy variable indicating if there is any precipitation. Similarly, \(W_2\) and \(W_3\) are dummy variables that indicate snowy weather and rainy weather with value either 1 or 0 respectively.

Parameter estimation

The parameter estimation is conducted using R. Table 3 illustrates the estimation results of Model 1. All the model input variables have a significant impact on train delay. Rail class 1 and 2 have positive impact on train speed level whereas class 3 has negative impact. The estimation results in Table 3 and Table 4 show the value of \(\alpha_2\) is consistently larger than that of \(\alpha_1\), which indicates, if
TABLE 3: Estimation results for Model 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variable</th>
<th>Estimate</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>Constant</td>
<td>1.275</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>RC$_1$</td>
<td>0.395</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>RC$_2$</td>
<td>0.731</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>RC$_3$</td>
<td>-5.890</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>FA</td>
<td>-0.173</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>FT</td>
<td>0.066</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>DD</td>
<td>-0.006</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>Constant</td>
<td>-1.183</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>RC$_1$</td>
<td>1.272</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>RC$_2$</td>
<td>1.608</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>RC$_3$</td>
<td>6.009</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>FA</td>
<td>0.187</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>FT</td>
<td>-0.224</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>DD</td>
<td>0.0008</td>
<td>***</td>
</tr>
</tbody>
</table>

*** 99% of significance level.

Other factors are fixed, trains operating on rail class 2 have higher average speed than trains on rail class 1. This is mainly because rail class 1 is located at the transportation center, in other words, with heavier traffic.

Among all three classes, class 3 shows largest impact on location uncertainty since $\beta_3$ is much larger than $\beta_1$ and $\beta_2$. The negative sign of $\alpha_4$ gives a reasonable indication that the occurrence of failure reduces the average speed of the train. Meanwhile, the positive sign of $\beta_4$ means failure increases the variation of train location. Hence, the findings in this specification are consistent with the assumptions made above. Besides, the dummy variable FT refers to the high-speed train in Sweden. The sign of the corresponding coefficient, $\alpha_5$, for speed is positive, as expected. More than that, the presence of high-speed train reduces the uncertainty of train location given that $\beta_5$ is negative.

Since the assumption made in this study is that the departure delay is known in advance, DD is considered as a variable in the model. The sign of $\alpha_6$ is negative so that the increment of DD reduces the average speed of the train. This is in parallel with our assumption that long departure delay demotes average train speed. The finding can be explained by that the departure delay leads to the chaos of train operation plan. Delayed trains usually give ways to trains on time. In the meanwhile, the magnitude of $\alpha_6$ and $\beta_6$ is rather low, indicating that the increment of variable DD affects $\mu$ and $\sigma$ only slightly.

Table 4 summarizes the estimation results of model 2. This model specification is enhanced by excluding insignificant terms. When looking into the parameters of the same variables to model 1, the values fluctuate without new features found in this specification. A new finding is when the road-surface temperature is within the range $[-20^\circ C, 10^\circ C]$, the average train speed is lower than the situation with higher road-surface temperature in the range $[20^\circ C, 40^\circ C]$. Consistently, the sign of $\beta_7$, $\beta_8$ and $\beta_9$ are all positive, revealing that low temperature leads to larger uncertainty on train location.
### TABLE 4: Estimation results for Model 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variables</th>
<th>Estimate</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>Constant</td>
<td>1.606</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>$RC_1$</td>
<td>0.371</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>$RC_2$</td>
<td>0.699</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>$RC_3$</td>
<td>-1.282</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>$FA$</td>
<td>-0.178</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_5$</td>
<td>$FT$</td>
<td>0.669</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_6$</td>
<td>$DD$</td>
<td>-0.005</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_7$</td>
<td>$RT_1$</td>
<td>-0.069</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_8$</td>
<td>$RT_2$</td>
<td>-0.066</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_9$</td>
<td>$RT_3$</td>
<td>-0.087</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>$RT_4$</td>
<td>0.044</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_{11}$</td>
<td>$RT_5$</td>
<td>0.031</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_{12}$</td>
<td>$W_1$</td>
<td>-0.273</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_{13}$</td>
<td>$W_2$</td>
<td>-0.289</td>
<td>***</td>
</tr>
<tr>
<td>$\alpha_{14}$</td>
<td>$W_3$</td>
<td>-0.309</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>Constant</td>
<td>-1.668</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$RC_1$</td>
<td>1.261</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$RC_2$</td>
<td>1.605</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$RC_3$</td>
<td>2.850</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>$FA$</td>
<td>0.189</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>$FT$</td>
<td>-0.229</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>$DD$</td>
<td>0.0007</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>$RT_1$</td>
<td>0.053</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>$RT_2$</td>
<td>0.054</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>$RT_3$</td>
<td>0.064</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>$W_1$</td>
<td>0.465</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$W_2$</td>
<td>0.451</td>
<td>***</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$W_3$</td>
<td>0.469</td>
<td>***</td>
</tr>
</tbody>
</table>

*** 99% of significance level.

The value of $\alpha_{13}$ and $\alpha_{14}$, parameters of two newly added weather factors "Snowy" and "Rainy", both get a negative sign, proving that the train speed level is degraded in snowy and rainy weather conditions. The weather with no precipitation is, in principle, good for train operation, but it actually reduces average train speed and increases the uncertainty of train location. One probable explanation is that some other factors not explicitly modeled, such as wind speed and wind direction, are not taken into account in this specification but they make differences on train operation as well.

### Validation

In order to evaluate and compare the performance of different model specifications, an independent dataset, not used in the model identification, is applied for estimating model performance indexes introduced earlier. This includes comparing the value of MAPE, RMSE, AIC and BIC for two
FIGURE 3: Illustration of the actual and model predicted train trajectory and arrival delays at stations for the north rail path.

TABLE 5: Statistical validation of two model specifications

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>27.2%</td>
<td>25.7%</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.145</td>
<td>1.125</td>
</tr>
<tr>
<td>AIC</td>
<td>333552</td>
<td>332854</td>
</tr>
<tr>
<td>BIC</td>
<td>333592</td>
<td>332862</td>
</tr>
</tbody>
</table>
FIGURE 4: Illustration of the actual and model predicted train trajectory and arrival delays at stations for the south rail path.

Model setups. Ten specific train trips are randomly chosen to compose the independent dataset. In predicting the ten chosen trips, the value of MAPE for the models varies from 12% to 35% with a mean of 27.2% and 25.7% for Model 1 and Model 2 respectively. Obviously, Model 2 shows better performance than Model 1 because of the inclusion of weather factors. It is necessary to mention that train travel time is used for estimating MAPE mainly because the actual value can be zero when arrival delay is applied. The other performance measures for the ten trips is summarized in Table 5. Both AIC and BIC are used in model evaluation since these statistical measures are useful for identifying the over-fitting case. By considering the values of AIC and BIC, the criterion is that the model specification with lower values of AIC and BIC is recommended, which suggests that model 2 is preferred. Meanwhile, no over-fitting problem has been observed, which may indicate that some other potential factors can be added to refine our prediction.
In order to intuitively compare the performances of the two models, two randomly selected train trips, one for the north rail line and one for the south line, and their arrival delays are portrayed in Figure 3 and Figure 4. In both figures, we compare the actual train trajectory with the predicted trajectories by showing a part of the train trips. Train arrival delays are also used to directly compare the predictable results of two models. It is notable that, for both cases, the predicted trajectories have similar movements to the actual trajectory, although different from the scheduled trajectory. Model 2 always gives a better approximation on the actual trajectory in comparison to model 1. Furthermore, the value of the actual arrival delays is in general higher than predicted arrival ones, revealing that other potential factors may contribute to the actual delay.

**SUMMARY AND CONCLUSION**

This paper introduces a stochastic process approach for modeling train arrival delays. The train operation is described by a Wiener diffusion process that starts when the train leaves the origin, and ends when the train arrives at the destination. The first hitting time distribution of Wiener process is used to model the train travel time between two connected stations. The model identification approach based on Maximum Likelihood method is then derived with train average speed and location variance modeled by both operational factors and weather impacts.

A case study, regarding two selected rail paths connecting Stockholm Center station and Örebro Center station, is carried out using one-year data in 2009. The datasets include not only train operation records but also weather information. The baseline model only considers train operation factor whereas the extended model includes also weather factors. Both models are estimated using part of the whole data and validation is carried out using an independent dataset. Several performance measures are established to compare the model specifications. As a result, both models can forecast arrival delay with a certain level of confidence, though Model 2 fits the actual arrival delays more precisely than the baseline model.

This study established a quantitative relationship between train arrival delays and selected train operation factors. The modeling results reveal that those factors, such as rail classes, train type, train departure delay, malfunctions and weather conditions, significantly affect train travel time and arrival delay. However, there remain limitations concerning data and methodological assumptions, which could affect the precision of estimation results, therefore requiring further study and methodology refinement. For example, detailed train location data measured by GPS during operation could improve the model accuracy as similar approaches have been successfully applied for modeling road travel time.

Furthermore, this study is considered as the first step of realizing the real-time train delay prediction. Based on the output of this study, some further steps can also be envisioned, such as developing passenger travel-path choice model given the predicted train arrival delays and enabling the real-time train arrival delay prediction for a transport network. More than that, accurate prediction of train delay may facilitate the operational schedule development by finding bottlenecks and providing valuable suggestions concerning potential impact factors.
REFERENCES


