Formal punctuality analysis of frequent bus services using headway data

Daniël Reijsbergen* and Stephen Gilmore

Laboratory for Foundations of Computer Science
The University of Edinburgh
Edinburgh, Scotland

Abstract. We evaluate the performance of frequent bus services in Edinburgh using the punctuality metrics identified by the Scottish Government. We describe a methodology for evaluating each of these metrics that only requires measurements of bus 'headways' — the time between subsequent bus arrivals. Our methodology includes Monte Carlo simulation and time series analysis. Since one metric is given in ambiguous language, we provide a formal description of the two most plausible interpretations. The automated nature of our method allows public transport operators to continuously assess whether the performance of their network meets the targets set by government regulators. We carry out a case study using Automatic Vehicle Location (AVL) data involving two frequent services, including the AirLink service to and from Edinburgh airport.

Keywords: Public transportation, punctuality, headways.

1 Introduction

A key feature of a sustainable city is a well-run public transportation network. This is witnessed, among other reasons, by the fact that satisfaction with public transport quality is included as an indicator for a 'smart' city [4]. One important measure for the performance of a public transport network is its *punctuality*, as this has been observed to be a major factor in passenger satisfaction and perceived service quality [3]. However, a formal definition of punctuality is not straightforward to give, partially because passenger perception of punctuality may depend on the nature of the service. In particular, for a non-frequent service (e.g., one bus departure every 30 minutes) strict timetable adherence is the main factor for punctuality. However, strict timetable adherence is less relevant for frequent services, which are defined as those with one bus departure every ten minutes or less. Punctuality metrics for frequent services are primarily dependent on the probability distribution of the times between departures — the so-called 'headways'. In general, less headway variance means better punctuality.

* This work is supported by the EU project QUANTICOL, 600708. Corresponding author: dreijsbe@inf.ed.ac.uk

Several punctuality metrics have been proposed in the scientific literature; [9] and [11] are two recent papers that present an overview. In this paper, we focus on the three punctuality metrics for frequent services identified in the guidance document on Bus Punctuality Improvement Partnerships by the Scottish Government [12]; all of these depend on the headways. Two of these metrics coincide with the metrics identified in [9] and [11].[1] The third metric does not; furthermore, it is ambiguously worded, so we formalise the two most plausible interpretations, resulting in a total of four metrics. We then provide a formal methodology for the evaluation of the four metrics that only requires headway measurements. The methodology is statistical in nature, so we particularly focus on providing approximate confidence intervals for the estimates of the metrics. This is a challenge because the probability distributions of some of the quantities under consideration are unclear. The evaluation of the two new metrics in particular is non-trivial, and we apply a range of statistical techniques including time series analysis, bootstrapping [6] and Monte Carlo simulation. Finally, we apply our methodology to a real-world set of headway measurements obtained using low-frequency Automatic Vehicle Location (AVL) data provided to us by the Lothian Buses company, based in Scotland and operating an extensive bus network in Edinburgh.

The outline of the paper is as follows. In Section 2, we discuss the routes considered and the datasets used. In Section 3, we formally define the three bus punctuality metrics used by the Scottish Government. We discuss a time series model for sequences of headway measurements in Section 4, and discuss the bootstrapping method for constructing approximate confidence intervals in Section 5. In Section 6, we evaluate the performance of two services operated by Lothian Buses using the punctuality metrics of Section 3. Section 7 concludes the paper.

2 Description and Visualisation of Routes and Data

In this section we explain the data processing that was applied to the raw AVL data before using it to compute the punctuality measures of interest. We had six datasets available: three for Route 100 (the AirLink service) and three for Route 31. For Route 100, three bus stops are of interest: the airport, the zoo, and George Street in Edinburgh city centre. For Route 31, the bus stops of interest are East Craigs, the zoo, and the Scott Monument on Princes Street in the city centre. The number of observations in each dataset is specified in Table 1.

The AVL data records the position of each bus in the fleet. Each bus has a unique identifier called a fleet number, and the assignment of buses to routes is captured in a schedule which is drawn up before the bus service begins for the day. The bus schedule maps buses by fleet number to routes but it can

[1] In particular, the metric of Section 6.1 is also mentioned in [9], while the related notion of the headway coefficient of variation is preferred in [11]. The metric of Section 6.2 is related to what is identified as an "*Extreme-Value based*" waiting time measure in [9]; it is also related to the Earliness Index of [11].

Table 1: Overview of the dataset sizes $n_1, \ldots, n_6$. Measurements were collected between 28th January 2014 at 11:31:14 and 30th January 2014 at 12:38:31.

Route	100 (AirLink)			31		
Stop	Airport	Zoo	George St	East Craigs	Zoo	Princes St
# measurements	127	126	128	102	102	105

change dynamically during the day in response to unpredictable problems such as mechanical failures of vehicles, or unexpectedly high or low levels of passenger demand. Thus the bus schedule serves as a guide for interpreting the AVL data but is not always accurate because it has not always been updated to record all unexpected events which occurred during the day. To address this problem, we use our custom visualisation tool [16] to plot buses on a map of Edinburgh. This allows us to check that they are serving the routes which we believe they are. If this was not done, incorrectly assigned buses would invalidate the computation of headway on routes. One Route 31 bus had to be identified manually. Furthermore, we suspect that one or two of the measurements in the Princes Street dataset correspond to wrongly assigned buses, but we have no evidence of this.

The schedule changes which make headway computation more difficult tend to occur at the start and the end of the day, when bus services have low frequency and the same bus is being used to serve several different routes. To eliminate this potential source of error in our interpretation of the data we restricted our observations to lie only between 9:00 and 17:00, when buses are frequent and rarely subject to route reassignments.

We linearly interpolate the AVL measurement data down to a granularity of one second between data points, and we detect departures from stops by dividing bus routes into zones and counting a departure as occurring when a bus moves from a zone containing a stop to the subsequent zone, using interpolated data. The bus stop zones were chosen such that they did not contain traffic lights.

3 Punctuality Measures

As mentioned in the introduction, we focus on the punctuality metrics set out by the Scottish government in [12]; we formalise these metrics in this section. Since buses are subject to a variety of unpredictable influences such as the number of passengers at bus stops and road congestion, the requirements are inherently *stochastic*. The randomness of the system is modelled through the headway, denoted by a random variable Y which takes values from $\mathbb{R}^+$ and is measured in seconds. The kth dataset, $k = 1, \ldots, 6$, is then a sequence $(y_{k1}, y_{k2}, \ldots, y_{kn_k})$ of realisations of Y_k, where n_k is as given in Table 1 (in the paper, we often leave out the dataset index k for brevity). Let $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathrm{Var}(Y)$. The requirements are then as follows.

In §2.13 of [12], which is under the header "*Starting point of the journey*", we find the following. "*For frequent services it is expected that on at least 95% of occasions:*

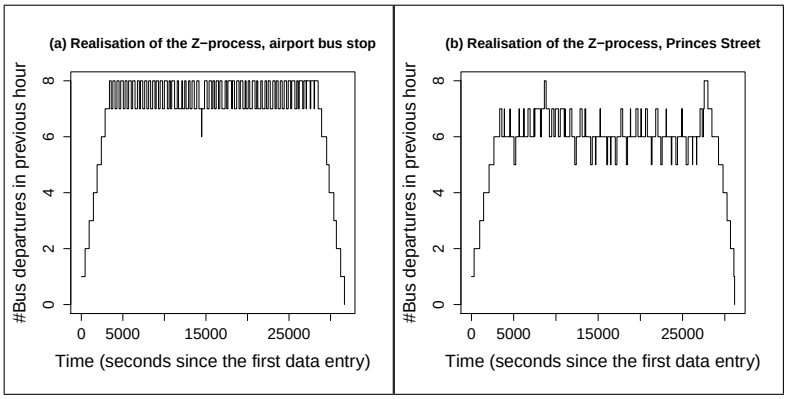

Fig. 1: Realisations of the process Z on 29th January 2014 for the airport bus stop dataset (left) and the Princes Street dataset (right).

– *Six or more buses will depart within any period of 60 minutes; and*
– *The interval between consecutive buses will not exceed 15 minutes.*"

Using seconds as the granularity, the latter requirement can be expressed as $\mathbb{P}(Y \leq 900) \geq 95\%$. We will call the probability $\mathbb{P}(Y \leq 900)$ the Extreme-Value Waiting Time (EVWT).

The former requirement is more intricate: given a sequence of headway measurements $(y_1, y_2, \ldots, y_{n_k})$, define for $t \in \mathbb{R}^+$

$$u(t) = \max\left\{j : \sum_{i=1}^{j} y_i \leq t\right\} \quad \text{and} \quad d(t) = \max\left\{j : \sum_{i=1}^{j} y_i \leq t - 3600\right\}.$$

Let $z(t) = u(t) - d(t)$, then $z(t)$ denotes the number of buses that departed in the hour prior to t. By construction, $u(t) \geq d(t)$ for all t so z is defined on $\mathbb{R}^+$. Figure 1 depicts the evolution of z for two of the six datasets.

The requirement that on 95% of "occasions" there must be six or more buses departures "within any period of 60 minutes" is slightly ambiguous; we will consider two interpretations. First, if we focus on the word "any", we could say that an "occasion" represents a time interval $[a, b]$ (a reasonable assumption would be that a denotes an hour after the departure of the first bus and b the departure of the last bus in a single day), and that $z(t)$ needs to be at least 6 at "any" point $t \in [a, b]$. The full requirement can then be expressed as: $\mathbb{P}(\forall t \in [a, b] : z(t) \geq 6) \geq 0.95$. We will call this requirement the Day-Long Buses-per-Hour Requirement (DLBHR).

The second interpretation is in terms of steady-state probabilities: for any measurement (assumed to be conducted when z is in steady-state) the probability that z is at least 6 needs to be at least 95%. To express this formally, define for any Boolean expression A

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\pi_z(j) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbf{1}(z(\tau) = j)d\tau, \quad \forall j \in \mathbb{N},$$

assuming that this distribution exists and is independent from $z(0)$. Then the latter requirement could be read as $\sum_{j=6}^{\infty} \pi_z(j) \geq 95\%$. We will call this requirement the Steady-State Buses-per-Hour Requirement (SSBHR).

Both the DLBHR and the SSBHR are hard to evaluate numerically, so we use simulation. To do this, we assume that the observations of z are realisations of a stochastic process Z; we then draw samples from Z to get probability estimates.

The final requirement is in §2.14 of [12], which is under the header '*Subsequent timing points*'. It reads as follows. "*For frequent services, measurement will be based upon Transport for London's concept of Excess Waiting Time (EWT). This is the difference between the average waiting time expected from the timetable, and what is actually experienced by passengers on the street. TC standards specify that EWT should not exceed 1.25 minutes.*"

The "*average waiting time expected from the timetable*" is assumed to be $\frac{1}{2}\mu$,[2] while the average waiting that "*is actually experienced by passengers on the street*" is given by $\frac{1}{2}(\mu^2 + \sigma^2)/\mu$ (see [7] or [14]). Hence, the Excess Waiting Time equals $\frac{1}{2}\sigma^2/\mu$. The maximum, according to the standards of the Traffic Commissioner (TC), is 1.25 minutes, or, equivalently, 75 seconds.

In summary, the EVWT, SSBHR and DLBHR are relevant only at journey starting points (which in our case study refers to the airport for Route 100 and to East Craigs for Route 31), while the EWT is relevant at all subsequent timing points. However, the former three metrics are also evaluated at the other timing points in Section 6 ; this is done for illustrative purposes.

4 Time Series Modelling

As we discussed in the previous section, the stochasticity of the headway between frequent buses is modelled using the random variable Y. To investigate whether the four requirements are satisfied, varying degrees of knowledge of the distribution of Y are needed. To calculate the EWT, we only need to know the expectation and variance of Y. To calculate whether the requirement on the time between subsequent bus departures is valid, we need to know the 95th percentile (i.e., the value x such that $\mathbb{P}(Y < x) = 95\%$; the requirement is satisfied if x is below 900). This is known if we know the quantiles and, hence, the entire probability distribution. To evaluate the requirements on the Z-process, we need to know the distribution of vectors $(Y_1, Y_2, \ldots)$ of measurements. This would be as hard as knowing the distribution of individual samples from Y if the samples that make up the vector were mutually independent. We will argue later in this section that they are not.

[2] There is no official scheduled headway for the AirLink service because the timetable for this service only says: "*at least every 10 minutes*". For Route 31, the difference between the timetabled and average observed headway is negligible (600 vs. roughly 580 seconds).

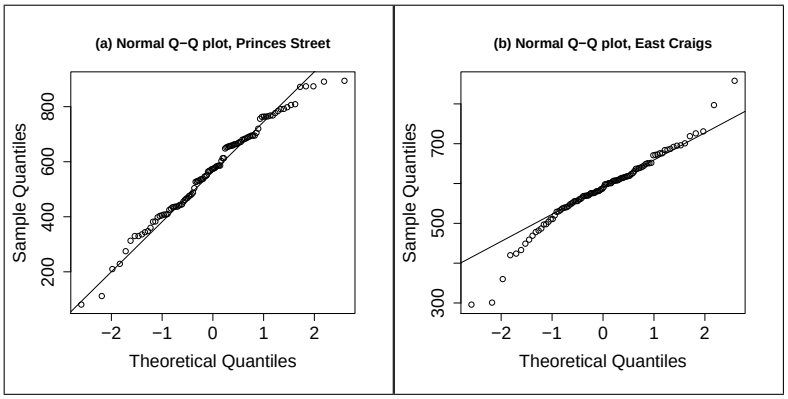

Fig. 2: Normal distribution Q-Q plots for the Princes Street dataset (left) and the East Craigs dataset (right).

In this paper, we assume that individual samples drawn from Y have a normal (Gaussian) distribution. To visualise whether the normality assumption is valid, we use normal Q-Q plots, i.e., plots of the quantiles of the empirical distribution against those of the normal distribution. This is done in Figure 2 for two datasets. A formal measure of the resemblance of an empirical distribution to the normal distribution is the Shapiro-Wilk test statistic, which assigns a value between 0 and 1 to an empirical distribution such that high values of the statistic represent close resemblance to the normal distribution. As the name suggests, it is used for the Shapiro-Wilk test [5] in which the null hypothesis that the sample has been drawn from the normal distribution is evaluated. The key result of the test is its p-value, which is the confidence in the validity of the null hypothesis. Values below 5% imply that the null hypothesis can be rejected at the 95% level. As we can see in Table 2, this only happens for the East Craigs dataset because of its fat (compared to the normal distribution) tails, especially on the left. Assuming that Y is normally distributed, the probability of interest can be computed using the normal cumulative distribution function, which is implemented in the statistical package R [13].

To generate a sample $(y_1, y_2, \ldots)$, we need to incorporate the correlation between measurements. Figure 3(a) depicts an Autocorrelation Function (ACF) plot for the dataset for the airport bus stop. The lag one autocorrelation is especially visible. The correlation is due to at least three sources:

1. Dependence by construction. Consider three buses with μ time units between departures; if the second is ϵ time units late, then the first headway will be $\mu + \epsilon$ and the second headway $\mu - \epsilon$ (negative correlation). This affects the lag one AC.
2. If a bus is late, then the number of passengers at the stop will be greater than normal, causing an additional delay. The next bus will need to pick up fewer passengers and may start to run early (negative correlation). This phenomenon is also mentioned in [1], and affects the lag one AC.

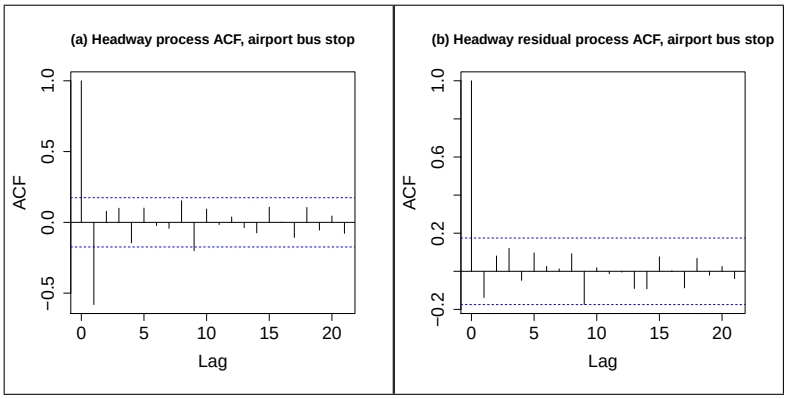

Fig. 3: ACF plots for the headway process $(y_1, \ldots, y_n)$ of the airport bus stop dataset (left) and its residual process $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$ after MA(1) fitting (right). The blue lines represent the levels at which the ACs are significant at the 5% level. Note that with 20 ACs plotted, the expected number of false positives at the 5% level is 1 (this could explain the seemingly significant lag 9 AC in both graphs.)

3. Factors that cause headway variation may persist: if one bus is late due to heavy traffic, then this traffic may have an influence on the next bus as well (positive correlation). This affects all ACs, in a decreasing fashion as the lag becomes bigger.

The lag one autocorrelation between the headways can be modelled using a Moving Average MA(1) time series model:

$$Y_i = \mu + \epsilon_i + \theta\epsilon_{i-1}, \quad \text{where } \epsilon_i \sim N(0, \sigma_\epsilon^2) \ \ \forall i = 1, \ldots, n.$$

The parameter θ captures the three forms of lag one correlation. The estimates of θ — denoted by $\hat{\theta}$ — for each of the six datasets are displayed in Table 2. Since the values $\hat{\theta}$ are estimates, we include the confidence in the null hypothesis that the true value θ equals 0. In each case, this is below 5%. Low values indicate that the time series model has a low explanatory power.

The ϵ's are commonly termed the *error terms*; the estimates $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ of these error terms based on a time series fit are called the *residuals*. We display an ACF plot for the residuals of the MA(1) model in Figure 3(b); as we can see, the lag one autocorrelation observed in Figure 3(a) is not present here. MA(1) models are part of the wider class of Autoregressive Moving Average ARMA(p,q) models. In general, these can be expressed as

$$Y_i = \mu + \sum_{j=1}^{p} \phi_j Y_{i-j} + \sum_{j=0}^{q} \theta_j \epsilon_{i-j}, \quad \text{where } \epsilon_i \sim N(0, \sigma_\epsilon^2) \ \ \forall i = 1, \ldots, n,$$

with $\theta_0 = 1$. The autoregressive terms (i.e., the ones that depend on the ϕ_j's) are well-suited to capture the effect of the third source of correlation on the higher

lag ACs. However, if the higher lag ACs are small then little explanatory power is added while the effect of the MA-terms becomes harder to isolate, so the net effect of adding these terms need not be positive. We have observed that MA(1) provides the best fit for all datasets, except the Princes Street dataset for Route 31. In this case the Autoregressive AR(1) process is best, although the difference is rather small (in terms of the Akaike Infomation Criterion). The MA(1) parameters θ and σ_ϵ can be estimated using methods of the `tseries` package in R. Given estimates of θ and σ_ϵ, we can draw realisations of $(Y_1, Y_2, \ldots)$ by drawing realisations of ϵ; methods for drawing standard normal random variables are implemented in R and the `SSJ` package in Java.[3] Realisations of Z can then be drawn analogously.

Table 2: Estimates of θ. Note that the values for $\hat{\theta}$ are negative because the first two sources of correlation mentioned in Section 4 apparently outweigh the third. The p-values of the t-test for $\theta = 0$ and the Shapiro-Wilk normality test for Y are given in the final two columns.

#	Stop	$\hat{\theta}$	t-test p-value	SW p-value
	Airport	-0.74903	$< 2{\cdot}10^{-16}$	0.1615
100	Zoo	-0.68416	$< 2{\cdot}10^{-16}$	0.0912
	George St	-0.56441	$2.22{\cdot}10^{-16}$	0.4436
	East Craigs	-0.21045	$2.07{\cdot}10^{-2}$	0.0047
31	Zoo	-0.36521	$6.24{\cdot}10^{-5}$	0.3439
	Princes St	-0.24497	$2.62{\cdot}10^{-3}$	0.1783

5 The bootstrapping method for confidence intervals

In the previous section, we described how to estimate the punctuality metrics used by the Scottish government. The estimates are based on realisations of $(Y_1, \ldots, Y_n)$, and will typically be different when the experiment is repeated. To account for the variation in the estimates, the estimates are given in the form of an interval estimate called the *confidence interval*. The interpretation of a $(1-\alpha)$ confidence interval for a statistic is as follows: if the experiment is repeated N times, then the number of confidence intervals that do not contain the true value of the statistic is expected to be αN. Throughout this paper we use $\alpha = 5\%$.

Whether a confidence interval for a statistic can be computed analytically (or approximated numerically) depends on how easy it is to express the probability distribution of the statistic. For some commonly-used test statistics their distribution is known explicitly, which means that confidence intervals can be constructed using methods implemented in common statistical tools such as R. However, even when nothing is known about the probability distribution of the

[3] `www.iro.umontreal.ca/~simardr/ssj/indexe.html`

test statistic, one can construct approximate confidence intervals using the *bootstrapping* method. A broad variety of bootstrapping methods exist; we use two of them, namely the non-parametric method of *case resampling* (through Monte Carlo simulation) and *parametric* bootstrapping.

5.1 Case resampling (Monte Carlo)

Case resampling is one of the most general forms of bootstrapping; given a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of independent and identically distributed realisations of some random variable X, and a test statistic f that is a function of $\boldsymbol{x}$, the approach is as follows. Let b be some positive integer. For each $j \in \{1, \ldots, b\}$, randomly draw n elements of $\boldsymbol{x}$ with replacement; let the new sample be $\boldsymbol{x}_j$ and $f_j = f(\boldsymbol{x}_j)$. Let $f^{(i)}$ be the ith smallest element obtained this way; an $(1 - \alpha)$ confidence interval is then given by

$$[f^{(b\alpha/2)+1}, f^{(b(1-\alpha/2))}], \tag{1}$$

assuming that $b\alpha/2$ and $b(1 - \alpha/2)$ are integers (we would use the floor and ceiling functions otherwise). The approximation gets better when the sample (empirical) distribution more closely resembles the true distribution of X.

5.2 Parametric Bootstrap

Parametric bootstrapping works the same way as case resampling, with the exception that we now have a stochastic model that allows us to draw random samples from X. The bootstrapping samples $\boldsymbol{x}_j$, $j = 1, \ldots, n$, are then obtained directly from the distribution of X. We still use (1) as the confidence interval.

6 Results

In this section, we discuss the results for the four performance metrics and requirements discussed in the previous sections: the Excess Waiting Time (EWT), the Extreme-Value Waiting Time (EVWT), the Steady-State Buses-per-Hour Requirement (SSBHR) and the Day-Long Buses-per-Hour Requirement (DLBHR). Each of these has its own subsection.

6.1 Excess Waiting Time

The Excess Waiting Time is relatively easy to evaluate: its computation only requires knowledge of μ and σ. These basic statistics are displayed for each of the six datasets in Table 3. Note that AirLink buses depart roughly every eight minutes, whereas Route 31 buses depart roughly every 10 minutes. We note that despite difference in means, the headway variance in related stops is very close. Consequently, the EWT is higher for the AirLink service than for Route 31. We further observe that the variance increases as the buses complete a larger part of

Table 3: EWT for each dataset, together with estimates of μ and σ. Confidence intervals were obtained using a parametric bootstrap with $b = 10\,000$.

#	Stop	$\hat{\mu}$	$\hat{\sigma}$	EWT	EWT c.i.	
100	Airport	477.2047	87.2564	7.9773	[5.182,	9.417]
	Zoo	475.2381	129.0810	17.5301	[12.261,	22.226]
	George St	473.5781	183.5627	35.5752	[26.087,	46.722]
31	East Craigs	585.7255	88.2538	6.6488	[5.261,	9.350]
	Zoo	588.7157	119.5863	12.1458	[9.231,	16.954]
	Princes St	568.8952	170.6736	25.6018	[19.156,	34.388]

the route, and with it the excess waiting time, which is what one would expect (as a bus completes its route it is increasingly subjected to sources of journey time variation, e.g., passenger numbers at stops).

In each dataset, the EWT is below the 75 second threshold. However, as mentioned in Section 5, the EWT estimates are subject to uncertainty because they are based on random samples. The empirical EWT can be computed from the sample variance and sample mean of a set of headway measurements $(Y_1, \ldots, Y_n)$. However, to generate a bootstrapped confidence interval for this statistic, we cannot use case resampling, as the measurements are correlated (although the effect of the correlation would vanish in larger samples). To remedy this, we conduct a parametric bootstrap using the time series model. The error terms ϵ are assumed not to have autocorrelation, so we could either use case resampling using the empirical dataset or draw samples directly from the normal distribution. The confidence intervals in Table 3 were obtained using the latter approach. In all datasets, the EWT is well below the 75 second mark with 95% confidence.

6.2 Extreme-Value Waiting Time

Table 4 summarises the results for the EVWT, i.e., the probability of a headway of over 900 seconds. As with the EWT, an estimate for the EVWT is easy to obtain; we only have to count the number of times this event occurred in the empirical dataset. Since this number is approximately binomially distributed, we can construct Clopper-Pearson confidence intervals [6] for the true probability. These intervals are very broad, owing to the small number of samples. For example, the upper bound of the confidence interval for the airport dataset is 2.863%, even though one would expect a much smaller probability based on the fact that the variance of Y is very small (as can be seen in Table 3). Using the assumption that Y is normally distributed, we can construct an estimate of the EVWT by using the normal distribution function combined with the estimates for μ and σ of Table 3. The results are in the '$\mathbb{P}(Y > 900)$' column of Table 4. Note that for East Craigs, the probability will be underestimated because this dataset differs so much from one with a normal distribution because it contains many more extreme values than one would expect.

Table 4: Probability of over 15 minute headway for each dataset. The first two numerical columns contain empirical estimates of these probabilities and exact (binomial / Clopper-Pearson) confidence intervals. In the final column we display the exact probabilities based on the normal distribution.

#	Stop	EVWT	EVWT c.i.	$\mathbb{P}(Y > 900)$
100	Airport	0	$[0, 2.863] \cdot 10^{-2}$	$6.3166 \cdot 10^{-7}$
	Zoo	0	$[0, 2.885] \cdot 10^{-2}$	$4.9976 \cdot 10^{-4}$
	George St	$2.344 \cdot 10^{-2}$	$[0.486, 6.697] \cdot 10^{-2}$	$1.0089 \cdot 10^{-2}$
31	East Craigs	0	$[0, 3.552] \cdot 10^{-2}$	$1.8470 \cdot 10^{-4}$
	Zoo	0	$[0, 3.552] \cdot 10^{-2}$	$4.6205 \cdot 10^{-3}$
	Princes St	0	$[0, 3.452] \cdot 10^{-2}$	$2.6191 \cdot 10^{-2}$

Based on the binomial confidence intervals, we can conclude that for all datasets except George Street, the requirement on the EVWT is met with more than 95% confidence. Based on the assumption of normality, the requirement is met for all datasets.

6.3 Steady-State Buses-per-Hour Requirement

An empirical estimate of the SSBHR for a given service is easy to obtain; in the realisation of z in this dataset (as is visualised in Figure 1), count the amount of time that z is lower than 6 and divide this by the total time. Formally, given a realisation of z on $[0, t]$, this means that the estimate $\hat{\pi}_z(k)$ for the steady-state probability of being in state k can be computed as

$$\hat{\pi}_z(k) = \frac{1}{t} \int_0^t \mathbf{1}(z(\tau) = k) d\tau. \tag{2}$$

The results are given in Table 5; instead of just the percentage of time that z spends below 6, the entire empirical steady-state distribution is given. For a given day, we start observing z one hour after the first bus departure (we assume that the process has approximately reached steady-state by then), and stop at the final bus departure.

Table 5: Empirical steady-state distributions of z for each of the six datasets.

#	Stop	$\hat{\pi}_z(5)$	$\hat{\pi}_z(6)$	$\hat{\pi}_z(7)$	$\hat{\pi}_z(8)$	$\hat{\pi}_z(9)$
100	Airport	0	$3.431 \cdot 10^{-4}$	0.460	0.540	0
	Zoo	0	$1.292 \cdot 10^{-3}$	0.472	0.525	$1.702 \cdot 10^{-3}$
	George St	0	$3.336 \cdot 10^{-2}$	0.483	0.454	$2.927 \cdot 10^{-2}$
31	East Craigs	$3.001 \cdot 10^{-2}$	0.753	0.215	$1.573 \cdot 10^{-3}$	0
	Zoo	$5.540 \cdot 10^{-2}$	0.705	0.240	0	0
	Princes St	$8.765 \cdot 10^{-2}$	0.675	0.233	$5.167 \cdot 10^{-3}$	0

Again, the empirical steady-state distribution is subject to variation, so we want to construct confidence intervals for these values. We have two options. First, we can use the time series model to generate long-run realisations of Z and use this to construct a parametric bootstrapping interval. Unfortunately, we have found that the parametric model is not well-suited for estimating the relatively small steady-state probabilities of low values of Z. Reaching the lower values of Z is particularly influenced by tail behaviour that is not captured well by the time series model, which tries to capture the dependence between subsequent headways in a single parameter (even though this dependence may vary throughout the day).

Therefore, we aim to use case resampling to construct a bootstrapping confidence interval. The question is what values to resample; obviously, sampling from the realisations $(y_1, \ldots, y_n)$ directly cannot be expected to work well because this would ignore the correlation between these realisations and the behaviour of Z depends on the behaviour of sequences of realisations of Y. Hence, we apply a renewal-like argument to estimate $\pi_z(j)$. Given a realisation of Z on $[0,t]$ and $m, l \in \mathbb{N}$, we partition $[0,t]$ into intervals $(I_{j1}, \ldots, I_{jm}) = ([\underline{I}_{j1}, \overline{I}_{j1}], \ldots, [\underline{I}_{jm}, \overline{I}_{jm}])$, when z equals j, and intervals $[\underline{J}_{j1}, \overline{J}_{j1}], \ldots, [\underline{J}_{jl}, \overline{J}_{jl}]$, when z does not equal j. It follows trivially from (2) that

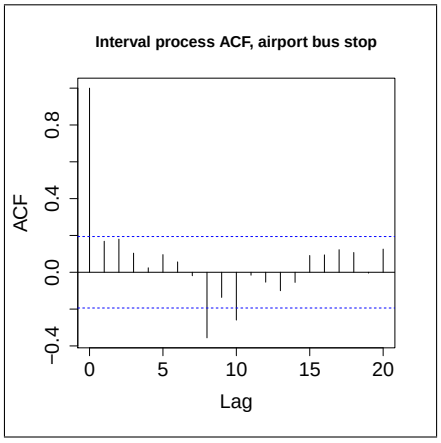

Fig. 4: ACF plot (see also Figure 3) for the series $(I_{j1}, \ldots, I_{jm})$ for $j = 7$ for the airport bus stop.

$$\hat{\pi}_z(j) = \frac{1}{t} \sum_{i=1}^{m} (\overline{I}_{ji} - \underline{I}_{ji}).$$

The idea is then to resample from the vector $(I_{j1}, \ldots, I_{jm})$ to construct a bootstrapping confidence interval. The key observation is that the intervals $(I_{j1}, \ldots, I_{jm})$ do have autocorrelation, but 1) that this is significantly less than for the headways and that 2) the effect of the correlation vanishes in large samples while the autocorrelation between the headways has an impact on the probability distribution of Z. The autocorrelation between the intervals for $j = 7$ for the airport dataset is displayed in Figure 4.

The results of the bootstrapping procedure are displayed in Table 6. Note that we have also resampled the values for $[\underline{J}_{j1}, \overline{J}_{j1}], \ldots, [\underline{J}_{jl}, \overline{J}_{jl}]$, to obtain the total times; again the correlation between measurements of I and J vanishes asymptotically. The confidence intervals for different values of j are not independent because they are implicitly based on the same samples. We observe that the AirLink seems to always satisfy the SSBHR. However, based on the confidence intervals for $z = 5$, we cannot conclude with 95% confidence for the Route 31 that the requirement will be satisfied. In fact, for Princes Street we

can conclude with 95% confidence that the requirement will *not* be satisfied. We note that this is not necessarily a problem, as the SSBHR is only imposed on the starting points of the routes.

Table 6: Confidence intervals for steady-state distributions of z for each of the six datasets, generated using bootstrapping with case resampling with $b = 1\,000\,000$.

#	Stop	$\hat{\pi}_z(5)$	$\hat{\pi}_z(6)$	$\hat{\pi}_z(7)$	$\hat{\pi}_z(8)$	$\hat{\pi}_z(9)$
100	Airport	0	$\begin{bmatrix} 2.875{\cdot}10^{-4}, \\ 4.253{\cdot}10^{-4} \end{bmatrix}$	$\begin{bmatrix} 0.434, \\ 0.485 \end{bmatrix}$	$\begin{bmatrix} 0.513, \\ 0.566 \end{bmatrix}$	0
	Zoo	0	$\begin{bmatrix} 7.660{\cdot}10^{-4}, \\ 2.921{\cdot}10^{-3} \end{bmatrix}$	$\begin{bmatrix} 0.439, \\ 0.505 \end{bmatrix}$	$\begin{bmatrix} 0.493, \\ 0.557 \end{bmatrix}$	$\begin{bmatrix} 6.200{\cdot}10^{-4}, \\ 4.298{\cdot}10^{-3} \end{bmatrix}$
	George St	0	$\begin{bmatrix} 1.524{\cdot}10^{-2}, \\ 8.018{\cdot}10^{-2} \end{bmatrix}$	$\begin{bmatrix} 0.437, \\ 0.53 \end{bmatrix}$	$\begin{bmatrix} 0.374, \\ 0.528 \end{bmatrix}$	$\begin{bmatrix} 1.450{\cdot}10^{-2}, \\ 9.070{\cdot}10^{-2} \end{bmatrix}$
31	East Craigs	$\begin{bmatrix} 1.819{\cdot}10^{-2}, \\ 5.060{\cdot}10^{-2} \end{bmatrix}$	$\begin{bmatrix} 0.698, \\ 0.807 \end{bmatrix}$	$\begin{bmatrix} 0.157, \\ 0.288 \end{bmatrix}$	$\begin{bmatrix} 4.511{\cdot}10^{-4}, \\ 3.502{\cdot}10^{-3} \end{bmatrix}$	0
	Zoo	$\begin{bmatrix} 3.294{\cdot}10^{-2}, \\ 0.101 \end{bmatrix}$	$\begin{bmatrix} 0.662, \\ 0.747 \end{bmatrix}$	$\begin{bmatrix} 0.186, \\ 0.32 \end{bmatrix}$	0	0
	Princes St	$\begin{bmatrix} 6.044{\cdot}10^{-2}, \\ 0.13 \end{bmatrix}$	$\begin{bmatrix} 0.631, \\ 0.717 \end{bmatrix}$	$\begin{bmatrix} 0.184, \\ 0.291 \end{bmatrix}$	$\begin{bmatrix} 1.547{\cdot}10^{-3}, \\ 1.580{\cdot}10^{-2} \end{bmatrix}$	0

6.4 Day-Long Buses-per-Hour Requirement

Table 7 summarises the results for the DLBHR. Note that it is impossible to estimate this measure from our current dataset without assuming an underlying time series model, as we only have a single measurement of an entire day (29th January). Hence, we draw realisations from Z by simulating the underlying time series model despite the weaknesses of this approach discussed in Section 6.3.

Confidence intervals are easy to generate because in a sample of day-long executions of Z, the number of days in which the process did not drop below 6 is binomially distributed. Hence, we can construct Clopper-Pearson confidence intervals using R. We conclude that, assuming that our time series model is correct, the requirement is met for all the AirLink stops with over 95% confidence (although it just barely holds for George Street), and the requirement is met at none of the Route 31 stops. This is not surprising; because the Route 31 service operates slightly over six buses per hour, even small deviations from the schedule cause a violation. The AirLink service, which runs about 7.5 buses per hour, is much more robust in terms of the DLBHR.

To construct the Clopper-Pearson confidence intervals for Table 7, it is necessary to fix a sample size beforehand. We have used 100 000 samples for the

Table 7: Whole day probability estimates, based on 100 000 samples. We also include Clopper-Pearson confidence intervals and the sample sizes N needed by the SPRT to reach a conclusion (which was correct in all of our experiments).

#	Stop	$\hat{p}$	95% C.I.			SPRT N
100	Airport	1	[0.9999631,	1]		1399
	Zoo	0.99999	[0.9999443,	0.9999997]		1399
	George St	0.95256	[0.9512243,	0.9538694]		12701
31	East Craigs	0.00159	[1.352615,	1.857001]	$\cdot 10^{-3}$	75
	Zoo	0.00022	[1.378778,	3.330638]	$\cdot 10^{-4}$	74
	Princes St	0.00229	[2.003263,	2.606186]	$\cdot 10^{-3}$	75

results in Table 7, but this choice is typically non-trivial; to evaluate whether the true probability is smaller than or greater than 95%, larger sample sizes are needed when the true probability is closer to 95%. A solution is to use the conceptual framework of hypothesis testing for *statistical model checking* [8]. In particular, we can use sequential tests that are able to terminate the simulation procedure as soon as enough evidence has been collected to make a statement about whether the requirement has been satisfied. We use the Sequential Probability Ratio Test (SPRT) [15] with indifference level $\delta = 0.001$ and $\beta = \alpha = 5\%$.[4] As we can see in the table, the requirement is the hardest to check for George Street; in all other cases, fewer than 10 000 samples were needed.

7 Conclusions

In this paper we have formalised the bus punctuality metrics used by the Scottish government. We investigated the performance of two services operated by Lothian Buses using these metrics. To do this, we have applied a number of statistical techniques such as time series modelling, bootstrapping and the sequential testing framework that is also employed in statistical model checking. Route 31, which operates six buses per hour, does better in terms of Excess Waiting Time, while the AirLink service, which runs over seven buses per hour, does better in terms of the Steady-State and Day-Long Buses-per-Hour Requirement.

A key feature of our methodology is its automated nature. Currently, when bus networks are subjected to a formal review by traffic regulators, the headway data is gathered manually by inspectors who are physically sent to the selected bus stops. Since AVL data is gathered systematically to aid live bus arrival time prediction at bus stops, an automated methodology for using the data to evaluate punctuality allows the traffic operator to detect potential shortcomings prior to the review, meaning that a fine can be avoided. This is of particular interest to Lothian Buses, which has been fined by regulators in the past [2].

[4] Note that the SPRT's assumptions with $\delta = 0.001$ are only just valid for the George Street dataset; for a discussion of the effect of the parameter choice on the test's output, see [10].

As part of further research, we aim to improve our model by incorporating the non-Gaussian tail behaviour of the East Craigs dataset. We also hope to investigate possible time dependence (within the day) of θ and the error terms.

Acknowledgements: This work is supported by the EU project *QUANTICOL*, 600708. The authors thank Bill Johnston of Lothian Buses and Stuart Lowrie of the City of Edinburgh council for providing access to the data which was used for the case study. We would also like to thank Allan Clark and Mirco Tribastone for their helpful comments on a draft version of this paper.

References

1. Mathew Berkow, John Chee, Robert L Bertini, and Christopher Monsere. Transit performance measurement and arterial travel time estimation using archived AVL data. *ITE District 6 Annual Meeting*, 2007.
2. Lothian Buses faced fine for bad service. *Edinburgh Evening News*, 21 June 2010. Web.
3. Margareta Friman. Implementing quality improvements in public transport. *Journal of Public Transportation*, 7(4), 2004.
4. Rudolf Giffinger, Christian Fertner, Hans Kramar, Robert Kalasek, Nataša Pichler-Milanović, and Evert Meijers. Smart cities: Ranking of European medium-sized cities. Technical report, Vienna University of Technology, 2007.
5. T. Hill and P. Lewicki. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, 2006.
6. Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric statistical methods*. John Wiley & Sons, 2013.
7. E.M. Holroyd and D.A. Scraggs. *Waiting times for buses in central London*. Printerhall, 1966.
8. A. Legay, B. Delahaye, and S. Bensalem. Statistical model checking: an overview. In *Runtime Verification*, pages 122–135. Springer, 2010.
9. Zhenliang Ma, Luis Ferreira, and Mahmoud Mesbah. A framework for the development of bus service reliability measures. In *Proceedings of the 36th Australasian Transport Research Forum (ATRF), Brisbane, Australia*, 2013.
10. D. Reijsbergen. *Efficient simulation techniques for stochastic model checking*. PhD thesis, University of Twente, Enschede, December 2013.
11. Meead Saberi, Ali Zockaie, Wei Feng, and Ahmed El-Geneidy. Definition and properties of alternative bus service reliability measures at the stop level. *Journal of Public Transportation*, 16(1):97–122, 2013.
12. The Scottish government. Bus Punctuality Improvement Partnerships (BPIPs) guidance. 2009.
13. R Development Core Team. R: A language and environment for statistical computing. 2005.
14. Y. Yang, D. Gerstle, P. Widhalm, D. Bauer, and M. Gonzalez. The potential of low-frequency AVL data for the monitoring and control of bus performance. In *Proceedings of the 92nd Annual Transportation Research Board Meeting*, 2013.
15. H.L.S. Younes and R.G. Simmons. Statistical probabilistic model checking with a focus on time-bounded properties. *Information and Computation*, 204(9):1368–1409, 2006.
16. Shao Yuan. Simulating Edinburgh buses. Master's thesis, The University of Edinburgh, 2013.