

Quantifying the Impacts of Transit Reliability on User Costs

Jeffrey M. Casello
Assistant Professor
School of Planning and Department of Civil and Environmental Engineering
University of Waterloo
200 University Avenue West
Waterloo, ON Canada N2L 3G1
jcasello@fes.uwaterloo.ca
519 888 4567 ext 37538

Akram Nour
Masters Candidate
Department of Civil and Environmental Engineering
University of Waterloo
200 University Ave West
Waterloo, ON Canada N2L 3G1
anour@engmail.uwaterloo.ca

Bruce Hellinga
Associate Professor
Department of Civil and Environmental Engineering
University of Waterloo
200 University Ave West
Waterloo, ON Canada N2L 3G1
bhellinga@uwaterloo.ca

Word count:
Words: 3619
Figures: 6
Tables: 2
Equivalent word count: 5619

ABSTRACT

Transportation modeling frameworks assume that travelers are economically rational; they choose the lowest cost alternative to complete a desired trip. Reliability of travel time is of critical importance to travelers. Being able to quantify reliability allows planners to estimate more accurately how system performance influences local travel behavior and to evaluate more appropriately potential investments in transportation system infrastructure.

In this paper, we present a methodology that makes use of automatic vehicle location (AVL) data from the Regional Municipality of Waterloo to estimate the reliability of transit service. Based on these data, we quantify the impacts of unreliable service on generalized transit user costs through a simulation model of bus arrivals and passengers' desired arrival times. We show that increasing reliability of arrivals at a station can decrease transit users generalized costs significantly, by as much as 15% in a reasonably reliable network. We further posit that including uncertainty in the calculation of generalized costs may provide better estimates for mode split in travel forecasting models. We conclude by describing future applications of the model.

INTRODUCTION

From a user's perspective, reliability in the transit network involves departing from the origin station on time; having reasonable limits on in-vehicle time; and most importantly, arriving at the destination station within a time frame that allows the traveler to reach his final destination without being late. Being able to quantify the degree of unreliability of a service allows transit planners to better estimate mode splits through travel forecasting models. Further, quantitative assessments of reliability provide estimates of tangible user benefits (through travel savings) which can be compared to investment costs in infrastructure to upgrade reliability such as queue jumpers, transit signal priority, or other means.

In this paper, we present a methodology that makes use of automatic vehicle location (AVL) data to estimate transit reliability at all stops along an express bus line service. We then utilize these data with various assumptions of passenger behavior to estimate the impacts of unreliable service on generalized transit user costs through a simulation model. The results of our model suggest that contemporary planning techniques may underestimate transit generalized cost in unreliable networks. We also provide an estimate of user benefits as a result of improved reliability in our network.

LITERATURE REVIEW

The issue of reliability in transit networks has been modeled for some time. Some of the earliest work was done by Osuna and Newell (1) as well as Wilson *et al.* (2). With the introduction of AVL technology, the opportunity arose to capture vast quantities of reliability data (3). Maximizing the benefit of this data collection is a major effort for many transit agencies.

The work presented here draws heavily from the formulation presented by Furth and Mueller (4). These authors quantify the expected and excess waiting time as a stochastic function of possible headways. They assumed passengers choose lines independent of headways and that passenger arrivals were not dependent on headways (i.e. uniform arrivals). They quantify the waiting times based on extreme cases of reliability, for example, the 95th percentile wait time. They suggest, and we concur, that mean waiting time is a poor indicator of wait time penalties in an unreliable network.

In this case, we take a similar approach but we concentrate on arrival times at the traveler's destination and the likelihood that an arrival will satisfy the traveler's trip objectives. Further, we extend the analysis to include a quantification of generalized cost, using a linear weighting proposed by Kittelson *et al.* (5) and used in most travel forecasting models. In our generalized cost model, we explicitly treat the impacts of early arrivals, late arrivals and departure time shifting as required by unreliability. For the impacts of late and early arrivals, we base our model on the seminal work of Bates *et al.* (6).

METHODOLOGY

We begin by defining the quality of service experienced at a given station. The Regional Municipality of Waterloo operates the iXpress service – a limited-stop, express service that travels between Waterloo, Kitchener and Cambridge. The alignment, shown in Figure 1, is approximately 33 km in length and consists of 13 stops. Along the route there are four downtowns (two in Cambridge), two universities, office complexes, major hospitals and regional shopping centers. iXpress operates throughout the day with 15 minute headways.

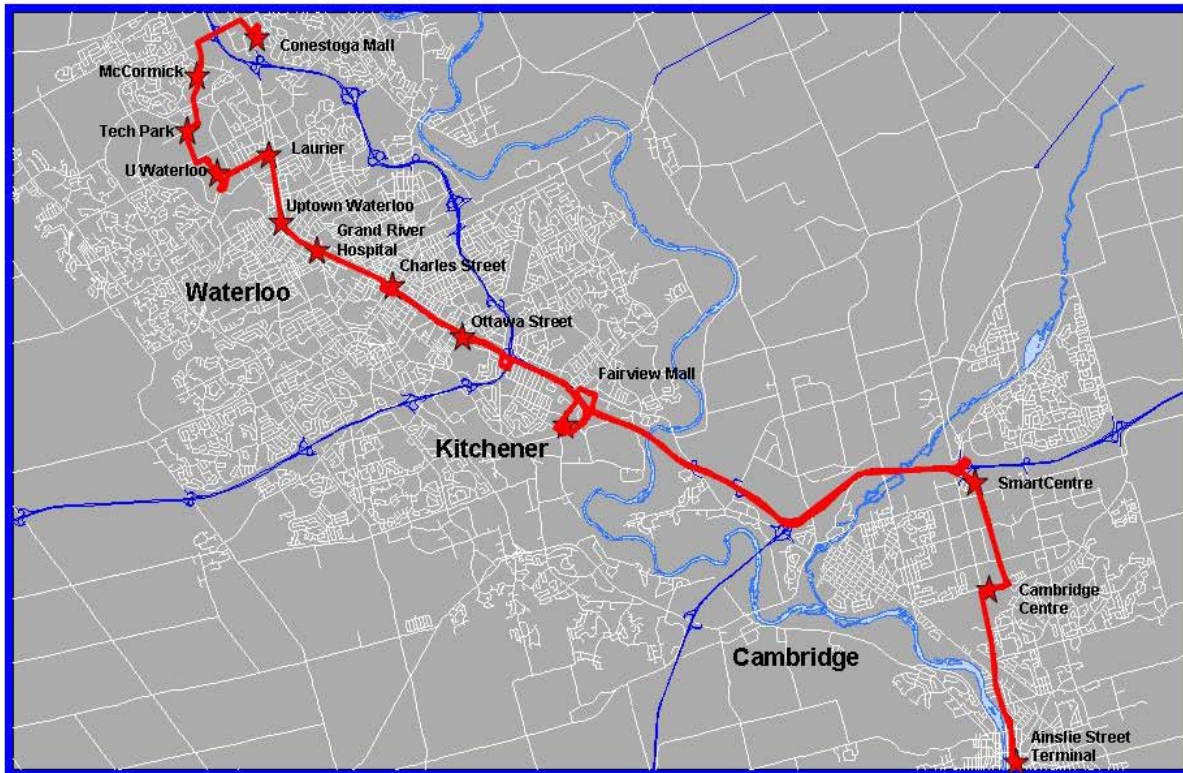


FIGURE 1 iXpress Route Serving Waterloo, Kitchener and Cambridge

Each of the iXpress vehicles is equipped with AVL technology. We collected real time arrival information for every stop, during a.m. and p.m. peak periods, for a week. In total, we gathered approximately 95 observations at each station. From these we defined service reliability as the difference between actual arrival time (AAT) and scheduled arrival time (SAT). For each station, we generate histograms of service reliability, shown in Figure 2, on which goodness of fit tests are completed to suggest an appropriate distribution. In each case, the PDF is a log-normal distribution¹ which meets *a priori* expectations: there exist a few early arrivals, many arrivals around the scheduled time, and a longer range of arrivals at times much later than the scheduled time.

¹ The log-normal distribution requires that all observations be greater than zero. To accommodate early arrivals, we add a constant to the test statistic approximately equal to the minimum observation.

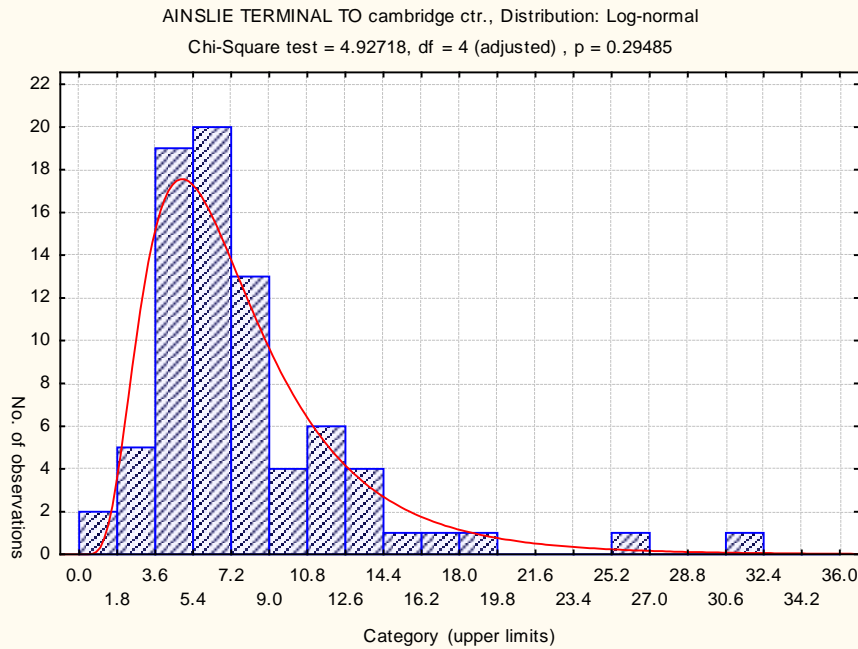


FIGURE 2 Observed frequency of service reliability

For the 13 stations, we observe a wide range of standard deviations (a measure of variance in service reliability), from 0.12 minutes to 1.06 minutes, with the average deviation being 0.44 minutes. In all, the iXpress is a relatively reliable system. From this empirical data, we define three station types – high reliability, medium reliability and low reliability – each with similar a mean but with varying deviations.

We next define three groups of travelers with varying risk tolerances which may represent the travelers’ personalities or trip purpose. One subset of travelers is very risk averse (RA), only choosing a transit departure if the likelihood of arriving late with that departure is less than 10%. This group may represent those commuters for whom work start times are fixed and highly inflexible. A second group of travelers is moderately risk averse (MRA), selecting a transit departure if the risk of arriving late to the destination is less than 30%. Finally, we define a risk neutral RN (group) who select a transit departure if the probability of a late arrival is less than 50%. This risk neutral group may be considered recreational travelers for whom arrival times have some flexibility.

With these definitions in place, we create travel behavior rules for each of the travelers to each of the stations. Let us assume that the necessary arrival time (NAT) – the latest time a traveler can arrive without being “late” – is a random variable that is uniformly distributed between two subsequent bus arrivals. We can define Δ to be the difference of the bus’ scheduled arrival time (SAT) and actual arrival time (AAT). Based on the cumulative distribution function (CDF) of the service reliability statistic and the traveler, there exists some Δ^* for which the probability of an arrival prior to $SAT + \Delta^*$ is equal to the traveler’s risk threshold. If the traveler’s NAT is later than the station’s Δ^* , then that traveler will choose the first transit arrival prior to his NAT. The relationship is shown graphically in Figure 3. Mathematically, we can express this relationship as follows:

$$\begin{aligned}
 & \text{Travelers choose } A_i \text{ if } NAT \geq \Delta^* \\
 & \text{where } \Delta^* \text{ is given by } \Pr(AAT - SAT \geq \Delta^*) \leq \text{risk tolerance} \quad (1)
 \end{aligned}$$

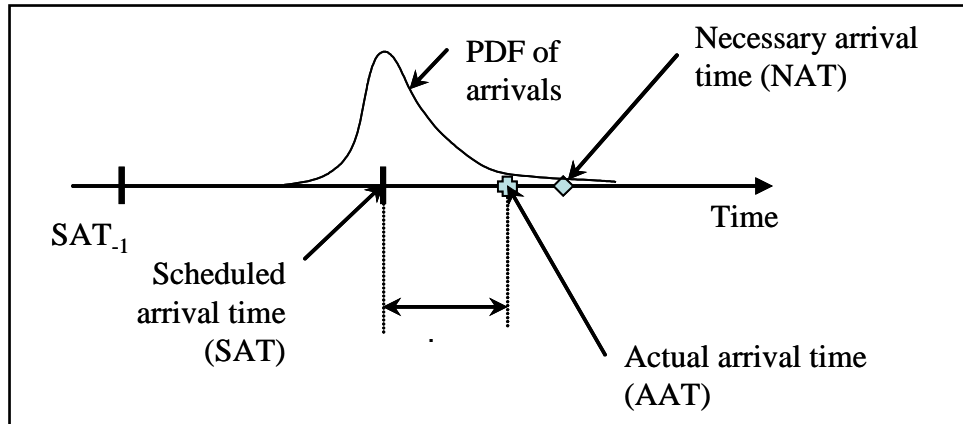


FIGURE 3 Graphical representation of bus and passenger arrival times

Suppose a traveler must reach his destination stop at 14:00. The closest scheduled bus arrival time is 13:55. If arrivals at that stop are sufficiently reliable that the scheduled 13:55 bus actually arrives prior to 14:00 90% of the time, then even the most risk averse traveler in our model will choose that departure time. If, alternatively, the scheduled 13:55 bus actually arrives prior to 14:00 only 40% of the time, then none of our travelers have sufficient risk tolerance to choose the scheduled 13:55 arrival.

Because we have the PDF of arrival times at each of our stations, we can compute the cumulative distribution function and from that determine the number of minutes after the scheduled arrival time which satisfies the risk aversion threshold for each of the model travelers. This is shown in Table 1 below.

TABLE 1 Station arrival distributions and critical arrival times

Station Type	:	Φ) * for each station and traveler type		
			risk averse	risk averse moderately	risk neutral
High reliability	1.245	0.2621	1.9	1.0	0.5
Medium reliability	1.228	0.4416	4.0	2.3	1.4
Low reliability	1.062	0.7792	7.0	3.4	1.9

One can interpret Table 1 as follows. A risk-averse traveler, traveling to a station with known, low-reliability service, will only choose the scheduled bus arrival immediately prior to his appointment if his appointment occurs later than seven minutes after the scheduled arrival time. This is because the actual arrival data at this station suggest that there is only a 10% chance, the risk-averse traveler’s threshold, of an arrival later than seven minutes after the scheduled arrival time.

Because we have assumed that the NATs are uniformly distributed, we can also estimate the probability of NAT occurring after)*. Mathematically, this is given by:

$$\Pr(NAT \geq \Delta^*) = 1 - \frac{\Delta^*}{h} \quad (2)$$

where h is the line's headway, in this case 15 minutes.

An appropriate question to ask here is what happens to those trips for which the NAT falls too near to the SAT to allow the traveler to choose that arrival. We assume that the traveler then elects to travel on an earlier bus. He then arrives at the destination one headway earlier than his original scheduled arrival time plus any unreliability he may experience on this bus. Mathematically, his scheduled arrival time on the previous bus, SAT_{-1} , becomes:

$$SAT_0 - \text{headway} + \text{unreliability}.$$

The impacts of this unreliability are explored in the next section.

Perceptions of travel time

Typically, in measuring travel time, modelers employ a generalized cost formulation which quantifies a linearly weighted sum of travel time components. A common example is as follows:

$$GC_T = (\alpha_0 AT + \alpha_1 WT + \alpha_2 IVT) VOT + \text{fare} \quad (3)$$

where GC_T is the generalized cost of a trip by transit (\$);

AT is the access time to the line (minutes);

WT is the waiting time, modeled as half the headway for short headways (minutes);

IVT is the in-vehicle time (minutes);

VOT is the value of time (\$/minute);

fare is the transit fare (\$);

\forall_i is the relative importance of that component.

While this formulation adequately measures the actual time and out of pocket costs from the time of departure from an origin (other than the transit stop) to the time of departure at the transit stop, it fails to account for two additional costs borne by transit travelers. First, because transit has discrete departure and arrival times, there is an inherent early arrival penalty which is not typically counted. Second, in light of reliability, travelers may experience a late arrival penalty, or may make travel choices to avoid being late (as described above) and therefore incur greater early arrival penalties.

Let us return to our example. If a traveler's NAT is after Δ^* , then the traveler will choose the scheduled arrival time nearest to his NAT. The bus' AAT, however, is stochastic which means the traveler may arrive very early (relative to his NAT) or may arrive after his NAT. If a traveler's NAT is before Δ^* , then the traveler chooses an earlier bus to minimize the potential for being late. In doing so, he increases the cost associated leaving earlier and arriving well before his NAT. This range of possibilities carries with it an inconvenience and as such an additional, quantifiable generalized cost. We attempt to model these costs following the example of Bates *et al.* (6).

Bates *et al.* suggest the following costs associated with early and late arrivals. For early arrivals, the cost decreases as the AAT moves towards the NAT. If the AAT equals the NAT then zero cost is experienced. If the AAT is later than the NAT by any amount, then the traveler

experiences a fixed cost, representative of failure to be on time. The late penalty also increases with increasingly late AATs. These cost functions are shown in Figure 4.

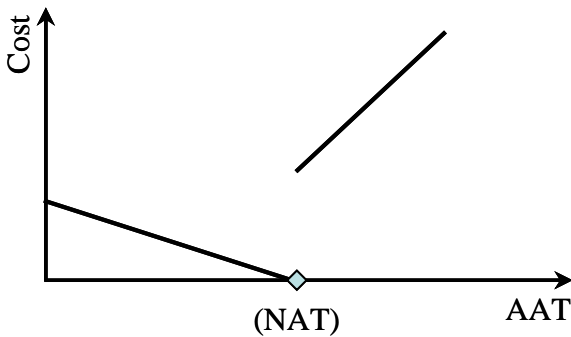


FIGURE 4 Early and late arrival penalties

In the case where NAT is after t^* , we define three separate responses to early and late arrivals that are representative of our traveler’s characteristics. For the risk-averse traveler, we assume the cost structure from Bates with very low early arrival penalties (because such risk-aversion likely produces frequent early arrivals). For the moderately risk-averse traveler, we assume a slightly higher penalty function for early arrivals, but a slightly lower penalty function for late arrivals. Finally, for the risk neutral traveler, we assume equal early and late arrival penalties. These multi-class cost functions are shown graphically in Figure 5 and quantitatively in Table 2. For the case where NAT is before t^* , we quantify the early departure penalty as one headway.

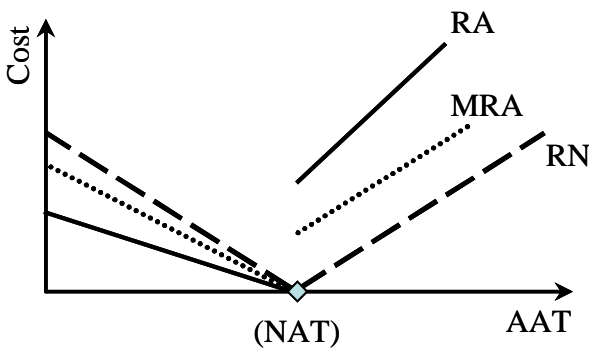


FIGURE 5 Graphical representation of multi-class penalty functions

TABLE 2 Penalties for early and late arrivals

Traveler Type	Early Arrival Penalty (EAP) minutes	Late Arrival Penalty (LAP) minutes
Risk-averse	$0.25*(NAT - AAT)$	$0.5h + (AAT - NAT)$
Moderately risk-averse	$0.5*(NAT - AAT)$	$0.25h + 0.5(AAT - NAT)$
Risk-neutral	$0.6*(NAT - AAT)$	$0.6(AAT - NAT)$

We now have four cases:

1. NAT is after)^{*} and the bus' AAT is prior to NAT – early arrival penalty incurred;
2. NAT is after)^{*} but the bus' AAT is after to NAT – late arrival penalty incurred;
3. NAT is before)^{*} and the bus' AAT is prior to NAT – early departure penalty and early arrival penalty incurred;
4. NAT is before)^{*} and the bus' AAT is after to NAT – early departure penalty and late arrival penalty incurred.

We can rewrite the generalized cost equation for each case. We assume for simplicity that access time is negligible (equal to 0) and waiting time is equal to one-half the headway (0.5 h) or 7.5 minutes. (Furth and Mueller (4) treat the impacts of reliability on waiting time). We also assume that VOT and fare are equal in all cases and therefore can be eliminated. This results in the following four generalized cost equations.

$$\text{Case 1: } GC_T = (2.5 \cdot WT + SIVT + 1.25 \cdot Late + EAP)$$

$$\text{Case 2: } GC_T = (2.5 \cdot WT + SIVT + 2.0 \cdot Late + LAP)$$

$$\text{Case 3: } GC_T = (EDP + 2.5 \cdot WT + SIVT + 1.25 \cdot Late + EAP)$$

$$\text{Case 4: } GC_T = (EDP + 2.5 \cdot WT + SIVT + 2.0 \cdot Late + LAP)$$

where $SIVT$ is the Scheduled in-vehicle travel time (minutes);
 $Late$ is AAT-SAT (minutes)

These generalized cost functions disaggregate the travel time components with weighting for each based on various sources. It is standard practice to assign in-vehicle travel time a value of 1.0 and rank all other time components as more or less important. In this case, however, we further disaggregate the in-vehicle travel time into two components: the scheduled in-vehicle travel time ($SIVT$) and the duration of the trip that is longer than expected. The $SIVT$ component is given the standard weight of 1.0 while the longer than expected portion of the trip is given a higher weight which varies depending on whether the bus' actual arrival is later than the necessary arrival time. The weighting of $Late$ is lower in cases 1 and 3 to represent a passenger's tolerance of behind schedule operation that still results in an early arrival. The weighting of $Late$ is much higher in cases 2 and 4 because the extra travel time causes the passenger to arrive after the NAT. The wait time weighting of 2.5 is derived from Kittelson's (5) average perception of wait time and the late and early penalties are derived from the previous equations.

Modeling Necessary and Actual Arrival Times

To account for both the discrete arrivals and reliability factors, it is necessary to predict the difference of AAT and NAT. An analytic solution to this problem requires the convolution of the lognormal pdf of arrival times and the uniform pdf of NAT (7). Mathematically, this is quite complex. Instead, we elect to simulate the results with appropriately distributed arrival and NAT events.

RESULTS

We create 10,000 travelers who are equally likely to have each risk-aversion characteristic and are equally likely to have a destination of each reliability category. We assume a scheduled in-

vehicle time of 20 minutes which results in a “traditional” generalized cost of 38.75 minutes. We employ the AVL derived arrival distributions presented above for our stations. The model predicts a frequency of 78.4%, 4.5%, 17.1%, and 0.1% for cases one through four defined above. This is to say that 78.4% of the time a passenger will choose the bus which is scheduled to arrive nearest to his NAT and actually arrive prior to the NAT. Only 4.5% of the time will a passenger choose the bus which is scheduled to arrive nearest to his NAT and arrive late. Just over 17% of travelers will choose an earlier bus to avoid the possibility of being late with nearly all of them arriving on time. The model predicts a probability of 0.1% of electing to take an earlier bus and still arriving late.

To investigate the impacts of discrete arrival times and unreliability on generalized cost, the model’s generalized costs results are shown in Figure 6. The dashed line represents the traditional generalized cost.

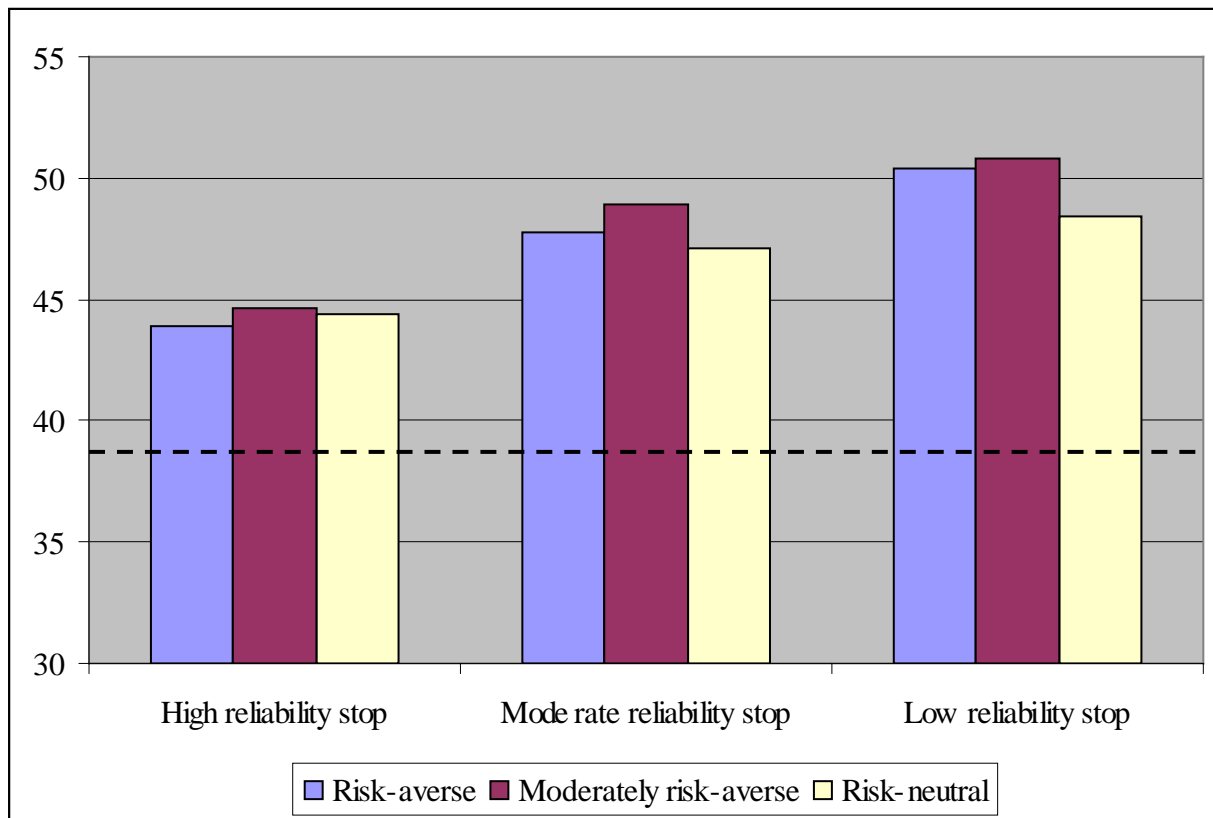


FIGURE 6 The model's generalized costs (minutes) results

The range of costs goes from 43.9 minutes for risk-averse travelers traveling to a high reliability stop to 50.8 minutes for moderately risk averse travelers to a low reliability stop. Note that for each class of traveler, the generalized cost increases with decreased reliability. The least reliable stop has generalized costs that exceed the most reliable stop by about 7 minutes or more than 15%. It is also interesting to note that the moderately risk-averse traveler has the highest costs in each case. This traveler has a slightly higher threshold for choosing the first arriving bus which results in the traveler experiencing late penalties more frequently than the risk-averse

traveler. So, while the most risk-averse traveler is more likely to experience an early departure penalty, he is much less likely to experience a late arrival penalty.

We also point out that the model provides intuitive results for the risk-neutral traveler. As reliability decreases for the destination stop, the risk-neutral traveler is least impacted of all. This is indicative of his overall risk tolerance and equal (and generally smaller) perception of penalty associated with either result.

We conclude with two points from Figure 6. Modelers who are predicting mode split based on traditional generalized costs are likely to be underestimating the generalized cost of transit. In our example, the underestimation is approximately 20-30%. This systematic underestimation may partially help to explain the need for the so-called transit bias coefficient which is often used to calibrate predicted mode splits to observed values. Second, we can employ this formulation to quantify the actual costs of unreliable transit service. In this example, the average costs are 44.3, 48.0 and 49.9 minutes for high, moderate and low reliability stops respectively. If all stops were upgraded to high reliability, the model suggests that savings of 3.1 minutes (approximately 7%) per passenger are possible. Multiplying this time savings per capita times ridership and value of time provides a financial estimate of the benefits accrued. This value can be directly compared to potential infrastructure investments such as queue jumpers, TSP, etc.

CONCLUSIONS AND FUTURE WORK

This model is based on data from the Region of Waterloo. It demonstrates a clear methodology to assess the impacts on reliability in a reasonably reliable network. We show that increasing reliability of arrivals at a station can decrease transit users generalized costs. We further posit that including uncertainty in the calculation of generalized costs may provide better estimates for mode split in travel forecasting models.

Given that this model has been created, it is a relatively straightforward exercise to test different transit systems for which AVL data exists. This allows for the comparison of overall reliability impacts on users from across networks. The model formulation also allows us to assess the impacts of longer headways on early arrival penalties. Perhaps most importantly, the formulation presented provides an opportunity to calibrate a model of user perceptions for disaggregate travel times. We may be able to calibrate a generalized cost model which includes separate weightings for deviations from expected travel times, as well as early and late arrivals.

ACKNOWLEDGEMENTS

This research was sponsored by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Regional Municipality of Waterloo.

REFERENCES

1. Osuna, E. E., and G. F. Newell. Control Strategies for an Idealized Public Transportation System. *Transportation Science*, Vol. 6, 1972, pp. 52-72.
2. Wilson, N. H. M., D. Nelson, A. Palmere, T. H. Grayson, and C. Cederquist. Service-Quality Monitoring for High-Frequency Transit Lines. In *Transportation Research Record 1349*, TRB, National Research Council, Washington, D.C., 1992, pp. 3-11.
3. Furth, P. G., B. Hemily, T. H. J. Muller, and J. G. Strathman. *Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Review and Potential*. TCRP Web Document 23 (Project H-28), 2003.

4. Furth, P., and T.H.J. Muller. Service Reliability and Hidden Waiting Time: Insights from Automatic Vehicle Location Data. *Journal of the Transportation Research Board* No. 1955, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 79-87.
5. Kittelson and Associates; KFH Group, Inc.; Parsons Brinckerhoff Quade and Douglass, Inc.; and K. Hunter-Zaworski. *TCRP Report 100: Transit Capacity and Quality of Service Manual*, 2nd ed. Transportation Research Board of the National Academies, Washington, D.C., 2003.
6. Bates, J., J. Polak, P. Jones, A. Cook. The Valuation of Reliability for Personal Travel. *Transportation Research Part E* No. 37, 2001, pp. 191-229.
7. Meyer, P. L. *Introductory Probability and Statistical Applications*. Reading, Addison-Wesley Publishing Company, 1965.