Evaluation of Improved Pushback Forecasts Derived from Airline Ground Operations Data

Francis Carr, Georg Theis, Eric Feron and John-Paul Clarke
International Center for Air Transportation
Massachusetts Institute of Technology
Cambridge MA 02139

July 21, 2003

Contents

Abstract

Accurate and timely predictions of airline pushbacks can potentially lead to improved performance of automated decision-support tools for airport surface traffic, thus reducing the variability and average duration of costly airline delays. One factor which affects the realization of these benefits is the level of uncertainty inherent in the turn processes. To characterize this inherent uncertainty, three techniques are developed for predicting time-to-go until pushback as a function of available ground-time; elapsed ground-time; and the status (not-started/in-progress/completed) of individual turn processes (cleaning, fueling, etc.). These techniques are tested against a large and detailed dataset covering approximately 10^4 real-world turn operations obtained through collaboration with Deutsche Lufthansa AG. Even after the dataset is filtered to obtain a sample of turn operations with minimal uncertainty, the standard deviation of forecast error for all three techniques is lower-bounded away from zero, indicating that turn operations have a significant stochastic component. This lower-bound result shows that decision-support tools must be designed to incorporate robust mechanisms for coping with pushback demand stochasticity, rather than treating the pushback demand process as a known deterministic input.

1 Introduction

Since deregulation in 1978, steadily growing demand for air transportation has exposed bottlenecks in the National Airspace System where traffic can easily outstrip capacity. Widespread criticism of the resulting delays and instability has driven the development of procedures and automation to accommodate the increased demand. From the viewpoint of air traffic control (ATC) and much of the flying public, this accommodation has focused on handling an increasing number of *flights*. There is a wide-ranging and substantial research literature devoted to proposed improvements for all phases of flight, from optimal routing during the taxi-out process to alerting systems for runway incursion after landing.

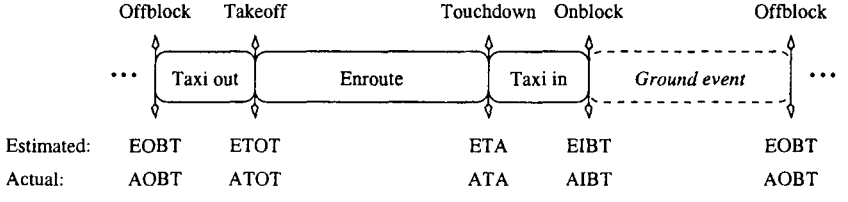

Figure 1: Flight-centric observability of air transportation.

For airlines the increasing traffic volume has led to larger and more complex problems of resource scheduling and synchronization in their ground operations. The interval from onblock to offblock is the proprietary business of these privately-held airline corporations, and published research on ground operations is quite sparse. Unlike "flights", there is not even a single terminology; "ramp operations", "turn process", "ground event", and "ground handling" are all in common use. One debilitating consequence of this paucity of research and collaboration is the lack of standardized efficient procedures and automation to smooth the transitions between inbound flights, ground operations, and subsequent outbound flights.

In analogy with handoff procedures in air traffic control, the natural design of such transitions requires that each agent (ATC, airport authorities, and/or the airline) controlling a particular process must estimate the time when control authority will transition to the next agent. A critical design/performance constraint for such handoff procedures is the level of uncertainty inherent in the turn process, or equivalently, the quality of available real-time observations of the turn process. Any handoff procedure (and associated decision-aiding tools and automation) must robustly cope with this inherent uncertainty. It is also the case that this inherent uncertainty affects the potential benefits from investing in improved transitions, and is thus an important factor in developing the business case to support such investments.

This report presents some of the first analyses of ground operations to support robust solutions for the ground-flight transition problem. Motivation for this research is presented in Section 2, including a survey of technical projects in airport surface traffic planning where pushback forecasts play an important role, and an approximate model of the value to airlines of a well-maintained schedule. Section 3 describes the real-world dataset used in these analyses. Based on these data, Section 4 analyzes the performance of several techniques for forecasting pushback times. The results show that, even when using the best currently-available observations in situations of minimum uncertainty, there is still a significant lower bound on the remaining uncertainty inherent to the turn process, and thus an upper bound

on the predictive power of any technique for forecasting departure demand. Conclusions are discussed in Section 5.

2 Motivation

2.1 Decision-support tools for airport surface traffic

One goal of the FAA's Free Flight Phase II program is the development of decision-support tools (DSTs) for airport surface traffic [15]. The role of these DSTs is to automate some of the monitoring, prediction, control and management tasks currently performed by air traffic controllers responsible for airport surface traffic. The proposed benefits include increased airport throughput, higher efficiency of taxi operations, and improved economic performance for air carriers. These benefits must be achieved without increasing controller workload or sacrificing system safety.

At the present time several such DSTs are deployed and/or undergoing active research and development. NASA Ames Research Center in cooperation with the FAA has developed the Surface Movement Advisor (SMA) currently in use at ATL and partially deployed at several other major airports [9]. The Surface Management System (SMS) is a newer Ames/FAA cooperative project which is presently being developed and field-tested at MEM [3]. Both SMA and SMS have been implemented for use in ATC towers and airline stations to provide real-time status information and shared awareness on airport surface traffic. In particular both systems make significant contributions to maintaining controller situational awareness with respect to expected future departure demand, runway queue lengths, taxi-out delays and airport departure rates for multiple possible tactical scenarios. The Center for Advanced Aviation System Development (CAASD) at MITRE is developing the Departure Enhanced Planning and Runway/Taxiway Assignment System (DEPARTS), an optimization-based tool which incorporates current airport conditions, departure demand, taxiing aircraft status, downstream traffic flow restrictions and user preferences to optimally assign and sequence

traffic to taxi routes, runways and departure fixes [7]. The Ground-Operation Situation Awareness and Flow Efficiency (GO-SAFE) concept is currently under development at Optimal Synthesis Inc. [5]. This tool has a similar technical focus and optimization-based approach as DEPARTS. However rather than being integrated into the ATC towers and airline stations, GO-SAFE is geared more towards use on the flight-deck to provide precision guidance, navigation and clearance delivery during the taxi process. Additional tools may also be under development; these four demonstrate the range of technical approaches and human-factors integration issues for DSTs for airport surface traffic.

The published literature on these DSTs states two common engineering assumptions related both to feasibility and performance. The first assumption is the availability of surface surveillance data, e.g. from multiple-sensor systems such as ASDE-X. For example, selection of MEM as a development site for SMS was influenced by the presence of existing infrastructure for the FAA's Safe Flight 21 program which duplicates ASDE-X performance [2]. Similarly initial modeling and site-adaptation of DEPARTS focused on ATL due to the availability of infrastructure from the SMA program [6]. The FAA is currently pursuing advanced surface surveillance through the ASDE-X program, and it is not unreasonable to expect that many airports will have the necessary infrastructure within the roll-out timeframe of current DST projects.

The second assumption is the availability of accurate and timely departure demand forecasts, where departure demand is interpreted as air carrier pushbacks at airports where movement on the ramp is under FAA control, or arrival of aircraft to ramp/taxiway transfer points at airports where ramp movement is under airline control. For example, in the SMS proposal one of the primary reasons cited for the selection of MEM as an initial site was the availability of partial pushback information from the two major carriers (Northwest Airlines and Federal Express) [18]. Similarly the selection of ATL for DEPARTS was influenced by the existing SMA infrastructure [6].

There is published research on the sensitivity of DST performance benefits (for DEPARTS

in particular [6]) with respect to pushback time forecast uncertainty and forecast horizon. In that study, a probability distribution was experimentally derived for the pushback time "error", defined as the difference between a flight's scheduled ready-for-pushback as per the flight plan, and the actual ready-for-pushback. This probability distribution was then scaled and shifted (under the side-constraint of preserving the coefficient of variation) to produce errors with mean absolute deviation of 0%, 20%, ..., 140% of the observed mean absolute deviation. Preliminary results showed that reduction from 100% (the baseline observed case) to 0% (perfect prediction of ready-for-pushback over a 10min horizon) coupled with the DEPARTS optimization engine reduced the average taxi-out time of each flight by approximately 1/3min; roughly half of this benefit occurred in the reduction from 20% to 0%. In an additional set of experiments, when DEPARTS was given perfect predictions of ready-for-pushback time over a finite time-horizon, the decrease in average taxi-out time per flight varied linearly with the length of the time-horizon. These preliminary results were later replicated and extended; see [7].

The problem of producing higher-quality forecasts of ready-for-pushback times, with reduced uncertainty over longer horizons, has been considered in the literature. For comparison, the DEPARTS study [6] was based on operations recorded at ATL in August 2000 including 18,586 observations of pushback time errors with magnitude less than one hour. This sample had mean 6.8min, standard deviation 12.6min, and mean absolute deviation 8.5min. It was noted in that study that pushback times at ATL have relatively low uncertainty compared to many other US hub airports.

Pushback forecasts at forecast horizons of zero to six hours from the Collaborative Decision Making (CDM) program frequently show errors with mean absolute deviation in excess of 20min even under the best weather and traffic conditions, while the mean absolute deviation can exceed 1.5hrs under poor conditions [17]. Based on an heuristic model of the internal airline decision-making processes of cancellations, swaps, intentional delays and hastening, Vanderson was able to reduce this prediction error by 0 to 30% [17]. In related work, the

Aircraft Sequencing Model (ASM) constructed in [1] was an optimization-based model which minimized passenger delay by modifying aircraft pushback times given constraints on the arrival sequence and timing, departure schedule, and gate and crew resources. Andersson selected a hub airport and used actual operations data over the 16:00-19:15 time-period on twelve successive days to show that the ASM yielded pushback predictions with standard deviation of 9 to 19min. An important caveat is that the ASM model assumed deterministic inputs including a perfect forecast of landing times over each 3.25hr period; landing times often show variability on the order of $\pm$10min as enroute aircraft approach the terminal area and encounter congestion and/or holding stacks [9]. Note that the forecasts developed in these studies have not shown uncertainty or horizons significantly better than the "raw" results observed in the DEPARTS study.

The continued development of DSTs is justified by the proposed benefits that will accrue to both ATC and air carriers. Current DST designs assume the availability of forecasts for upcoming pushbacks. This assumption is important enough to significantly affect which airports are selected for initial modeling, site-adaptation and integration. Furthermore the estimated benefits are known to be sensitive to both the uncertainty and horizon over which such forecasts are available. However only a handful of airports and airlines possess the necessary infrastructure to provide high-quality forecasts, while improved forecasts based on modeling the internal airline decision processes have not yet improved on simple "raw" forecasts (i.e. the ready-for-pushback time filed in the flight plan). We are not aware of research explicitly aimed at overcoming the technical hurdle of building such DSTs with only "raw" forecasts, nor of developing the business case to support airline and/or airport investment in the necessary infrastructure. There is a pressing need for further work in these areas.

2.2 The value of a well-maintained schedule

Much of the current R&D invested in DSTs for airport surface traffic is motivated by the needs of air traffic controllers. However, DSTs can also have significant benefits for air carriers. A reduction in surface traffic delays and uncertainty at one of an air carrier's hub airports can be leveraged to increase market share, reduce direct operating costs, etc. In addition, air carriers which invest in the necessary infrastructure to produce improved pushback forecasts may also see concomitant internal benefits such as improved situational awareness in airline operations centers and an increased ability to monitor, review and streamline internal processes. Note that these internal benefits are difficult to quantify without information on proprietary airline operations, and thus this report will focus on the potential for reduced delays and uncertainty.

The proper definition of "delay" is much debated in the research literature on air transportation. The US Department of Transportation (US DOT) collects and publishes statistics derived from the general rule-of-thumb that a flight which arrives onblock no more than 15min after scheduled arrival is "on-time". The corresponding Association of European Airlines (AEA) punctuality metric is adapted to the high variability of European enroute air traffic congestion: any flight which goes offblock no more than 15min after scheduled departure is considered "on-time". While these statistics have partially increased the transparency of the air transportation product for the traveling public, these definitions of delay do not capture the full complexity of maintaining a schedule of operations in an uncertain operating environment.

There are three main groups which benefit when an airline's schedule of operations is well-maintained: the traveling public, the ATC system, and the airlines themselves. In this general division of concerns, the ATC system is intended to include both national enroute ATC and local airport authorities, and similarly when referring to the airline, the term is intended to additionally include all of the aircraft servicing contractors involved in ground

operations. If one were simply to examine the US DOT or AEA delay statistics, it would be natural to assume that the costs experienced by these groups are roughly similar and are strictly linked to delayed flights. On closer examination, each of these groups incurs several different types of costs which do not scale in proportion to delay. Furthermore there can also be significant costs associated with flights which arrive or depart substantially *earlier* than scheduled; this justifies consideration of the larger problem of "schedule maintenance", of which minimizing delays is a significant component. Note that it is common practice to speak of delayed events and the corresponding delay, but a suitable antonym is lacking. Hence in this report an event which occurs prior to its expected or planned occurrence will be referred to as a *hastened* event with some corresponding *haste.*

The traveling public's direct valuation of delay and haste is difficult to measure. From the viewpoint of the ATC system and the airlines, passengers' valuation of their time is easiest to measure by proxy. As noted in the Introduction, the ATC system has come under increasing criticism as easily observed by the number and stated cause of passenger complaints. In high-demand markets where increased delays on one itinerary may drive passengers to substitute alternate itineraries or other forms of transportation, Januszewski [11] derives a marginal price-change of approximately $1USD per fare per minute of additional delay. For example, if passengers had the perception of an average five-minute delay on an hourly shuttle flight taking 6-7 trips per day and typically carrying 100 passengers, a marginal price-change of that magnitude could easily cost an airline $100,000USD in missed profits over the course of a month. A hastened flight could also cut into revenues if it arrived sufficiently early so that no gates were available; passengers are equally discomfited waiting to taxi in to the gate as they are waiting to depart.

A hastened flight wastes passengers' time, while flight delays lead to missed connections, cancellation of important meetings, and even unplanned overnight stays. However for an airline the effect of delays on a single flight-leg can also be magnified due to cascades of disruptions as later operations which share resources (crew, airframe, pax, etc.) with the

first disrupted flight-leg are affected. This propagation effect can be modeled as a *delay multiplier* by which the initial flight delay is scaled to account for the total delay incurred in the entire cascade [4]. While the exact values of the delay multipliers found in that study cannot be directly translated into other situations (e.g. different schedules, operating conditions, disruption recovery procedures, etc.), two important effects were observed: the additional delay due to the cascade effect (corresponding to the portion of the delay multiplier in excess of 1) tends to increase linearly with both the initial delay and the time-to-go until the end of the operational day. The former effect suggests that even if delays have a constant marginal cost as suggested by Januszewski, the total cost to the airline increases no slower than the *square* of the initial delay. The latter effect then suggests that delay costs should be linearly discounted as the operational day progresses. In a minor abuse of the standard "big-O" notation of computer science,

$$\text{delay cost} = \Omega(T_{\text{minus}} \cdot \text{delay}^2)$$

where T_{minus} is the time remaining until the end of the operational day.

Given this general cost-structure on delays, it is natural to consider what penalties or benefits could accrue from hastened flights. As noted above, flights which arrive or depart sufficiently ahead of schedule waste passengers' time and can artificially create gate shortages and congestion. However it is also apparent that arriving or departing just a little early is generally useful since it provides a small buffer against possible future delays and gives passengers the impression that everything is running smoothly. To incorporate these observations, this report assumes an airline cost structure of the form

$$\text{cost} = \alpha \cdot T_{\text{minus}} \cdot \text{deviation} \cdot (\text{deviation} - \beta). \tag{1}$$

where α and β are nonnegative constants and deviation is defined as the scheduled minus

the actual time of an event. For negative or large positive deviations, the cost is positive as expected. However for small positive deviations, when the actual time occurs before scheduled but the deviation is smaller than β, the cost is negative indicating a desirable outcome.

This cost structure can be used to produce "optimal" forecasts of pushback times, by minimizing the expected cost of deviations between the actual and forecast time-to-go over the duration of the forecast. This topic is treated in greater detail in Section 4. In addition there is an important conceptual message: cost does not scale in proportion to average delay or haste, and cost increases as the *dispersion* of deviations increases[1]. In concrete terms, twice the delay can cost four times as much, and a flight which is delayed by 5 ± 1min will always be more expensive than a flight which is delayed by 5min. It is natural for air carriers to include buffers and deliberate slack-time in their schedules in order to stay on-time and robust in the face of uncertain operating condtions. However there are significant reasons to keep the length of these buffers and all controllable sources of uncertainty under careful scrutiny to avoid rapidly ballooning costs.

3 Observing the turn process

While detailed real-time and historical data are available for the different stages of flight, until recently very few data have been automatically collected during ground operations. This lack of observability has blunted efforts to understand and improve these processes. However, as part of the Operational Excellence on-time initiative by Lufthansa Airlines, the ALLEGRO project was brought into operation: a unique real-time control and analysis system for airline ground operations [16]. Special focus was laid on the correct definition

[1]In fact this notion can be made precise. Suppose two processes with respective deviations $\mathbf{T}_1$ and $\mathbf{T}_2$ have the same average deviation $\mathrm{E}[\mathbf{T}_1] = \mathrm{E}[\mathbf{T}_2]$, but $T_1 \leq_{\mathrm{disp}} T_2$ using the definition of stochastic dispersion order $\leq_{\mathrm{disp}}$ from [13, p. 40]. Given this order on the deviations, it immediately follows from [13, Corr. 1.5.4(a), Thm. 1.7.6(b)] and the convexity of the cost structure that the mean, mean square, mean absolute deviation and variance of the costs must be similarly ordered.

of target times for each process of the turn. During ground operations on a single aircraft, ALLEGRO's monitoring function collects timestamps for up to 80 distinct types of events and compares them against the respective target times. The use of ALLEGRO increases the transparency of ground events, supports the hub control center in coordinating ground events, simplifies analyses of weaknesses in each process, enables continuous validation of the baseline schedule for ground events, and provides a foundation for internal and external performance agreements.

The analyses in this paper are based on a dataset of 17,344 turnaround operations obtained from the Lufthansa OBELISK data-warehouse, of which ALLEGRO is a subsystem. These turns were continental operations on Lufthansa which transited Frankfurt International Airport (FRA) between 1 Feb 2003 and 30 Apr 2003. For each transit through FRA, several observations were captured in OBELISK:

- Inbound and outbound flight parameters.

  These included aircraft type; flight numbers; connecting airports; scheduled, estimated and actual times for gate-arrival and pushback; scheduled and actual en-route and block times; and standardized delay codes.

- Ground operations data (from the ALLEGRO system).

  Both scheduled and actual ground time intervals were specified, and further subdivided into scheduled and actual start/end epochs for deplaning, cleaning, catering, fueling, and boarding. The type of pax-loading equipment (either bus or jetway) was also specified.

The timing data for inbound and outbound flights, and scheduled intervals for ground events, were reported with one-minute precision. The timing data for actual ground events were reported with one-minute or one-second precision depending on equippage. The accuracy of these data has been validated to the same order of magnitude as the reported precision [16].

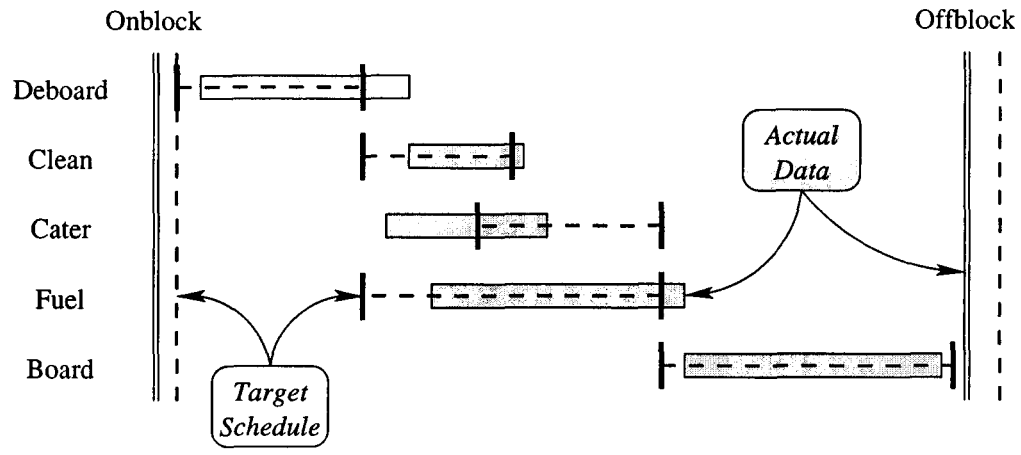

Figure 2: Illustration of ground operations data.

4 Inherent uncertainty and robustness

The airline decision processes involved in maintaining a schedule are substantially more complex than can be observed externally; to date modeling efforts aimed at mimicking these decisions have not significantly improved the quality of pushback forecasts. Uncertainty in pushback forecasts is partly due to the complexity of these decision processes (which may be amenable to improved models), and partly due to the natural inherent stochasticity of real-world operations (thus placing inherent lower bounds on the quality of any such forecasts). The ALLEGRO system enables novel analyses of this complexity/stochasticity factorization. Airline decision processes related to cancellations, swaps of crews or airframces, and intentional delay or hastening can be accounted for by examining those turns which were actually operated and the corresponding ALLEGRO target-times. A wide variety of exogenous sources of uncertainty can be filtered out using the delay codes attached to each turn. Thus it is possible to focus attention on the inherent uncertainty of the airline turn processes. This minimum inherent uncertainty is an important limiting factor for any forecast of expected offblock (EOBT) since it implies a corresponding upper bound on the achievable forecast performance.

To derive this upper bound, the dataset was filtered for *simple turns*, so-called because

	Board: Bus	Board: Jetway
Deboard: Bus	3820	842
Deboard: Jetway	157	6324

Table 1: Sample sizes for predictability analysis.

they occurred under the following set of conditions:

- *Similar target-times for the turn processes.* No towing occurred; the aircraft departed from the gate where it arrived. Operations occurred on similar aircraft types which require the same scheduled time for each process in the turn. The same type of pax-loading equipment was used. Changes in target-times due to different *available* ground-time (scheduled offblock minus actual onblock) were accounted for.

- *Only relevant delay-codes.* Turns with delays due to late inbound crews or loads; local or downstream weather; and control imposed by ATC or the airport authority were excluded. Also delays due to abnormal aircraft maintenance requirements were excluded, since those delays did not directly impact the normal turn processes and hence were not directly observable in the ALLEGRO data. The remaining delay codes are specific to aircraft servicing processes.

Some of these conditions may be anticipative, i.e. impossible to verify while a turn is actually taking place rather than after the fact. Hence this subset of turns yields a conservative approximation of the actual minimum uncertainty; real-world performance is guaranteed to be noisier. The first three conditions yielded a sample of 11,143 turn operations (64.2% of the total sample). Table 1 shows how these turns were further subdivided by the type of pax-loading equipment used on arrival and departure. Note that it is sometimes necessary due to customs or security measures to use different types of pax-loading equipment on the arrival and departure of the same aircraft; this does not indicate the aircraft was towed or otherwise changed gates.

Two caveats should be noted for the jetway-jetway turns. First, only limited data were

available on the end of deboarding for turns using jetways, since the method for measuring that epoch was not finalized in ALLEGRO at the time these data were collected. Second, ALLEGRO measures the end of boarding for jetway-jetway turns as the moment when the cabin doors are closed, an event which typically is only indirectly related to the actual end of boarding. For example, a flight delayed by ATC might need to have the crew swapped; or if a maintenance delay occured, the captain and maintenance personnel might re-open the cabin doors to perform inspections. However, for these and many other possible sources of error, there are corresponding delay codes and hence the affected turns have been filtered out and do not affect our results. In this report we focus on the 3820 bus-bus simple turns, which yield results based on the best available observations. The results for the other simple turns are similar and have been elided for brevity.

Given the same set of simple turns, a variety of statistical techniques can be used to forecast pushback times. The simplest forecasts are updated only once: once the available ground-time is known, the turn duration is forecast using descriptive statistics such as the mean, median or some percentile of observed turn durations. Any system with monitoring capabilities equivalent to ALLEGRO enables more sophisticated techniques. *Age-based* forecasts use the elapsed duration of a turn to compute a Bayesian estimate of the remaining time-to-go. A conceptually orthogonal *status-based* approach depends on the updated status (not-started/in-progress/completed) of the different processes comprising a turn. In the remainder of this section, age- and status-based forecasts are formalized and their actual performance against the ALLEGRO dataset is compared.

It is important to note that combined forecasts using both process status and the elapsed duration of each process cannot be derived *explicitly* from the available data. Even a discretized state-space for such a combined forecast would be many orders of magnitude larger than the number of real-world turn operations available for calibrating the forecast statistics. The most common approach proposed in the DST research literature is the use of simple descriptive statistics, a plan which is immediately workable, but ultimately limited

in performance by this "curse of dimensionality".

4.1 Forecasts using simple descriptive statistics

A simple analysis of the data shows that predicting the actual ground-time based on the scheduled ground-time has several drawbacks. In Figure 3 the bus-bus simple turns have been binned according to scheduled ground-time. The sample of turns in each bin is then described by a boxplot of the corresponding sample of actual ground-times[2]. From the plot, the minimum ground-time is reasonably apparent. The median of actual ground-time tends to track the scheduled ground-time, indicating that onblock and pushback typically occur on schedule. However the variability of actual ground-time as a function of scheduled ground-time is very high. This variability leads to poor predictions of actual ground-time given scheduled ground-time.

Much of this variability can be compensated for by replacing scheduled ground-time with available ground-time. In particular the effect of early or late inbounds can be accounted for. The resulting substantial decrease in variability is shown in Figure 4. It is possible to predict actual ground-time with much reduced uncertainty given the available ground-time; even this simple information may thus be very useful for developing DSTs with significant benefits. However it is also important to note that, due to the possible need for cancellations or swaps, the available ground-time cannot always be determined based on the arrival time of an inbound flight and the *scheduled* assignment of this inbound aircraft to an outbound flight. Since our dataset only describes turns which actually occurred, this problem does not affect our analyses. It is an interesting open question to determine how soon an airline station could accurately determine the available ground-time, i.e. how soon an airline station can accurately claim that a particular outbound flight will not be cancelled or swapped.

[2]Boxplots are a standard statistical method for robust visualization of scalar data. The "box" in a boxplot covers the interquartile range, with a line through the middle of the box to denote the median. This gives a robust estimate of the central tendency and dispersion of the data. The box has "whiskers" extending to the farthest datapoints within 1.5 times the interquartile range of the median. Observations outside the whiskers are marked individually; these are often treated as possible outliers.

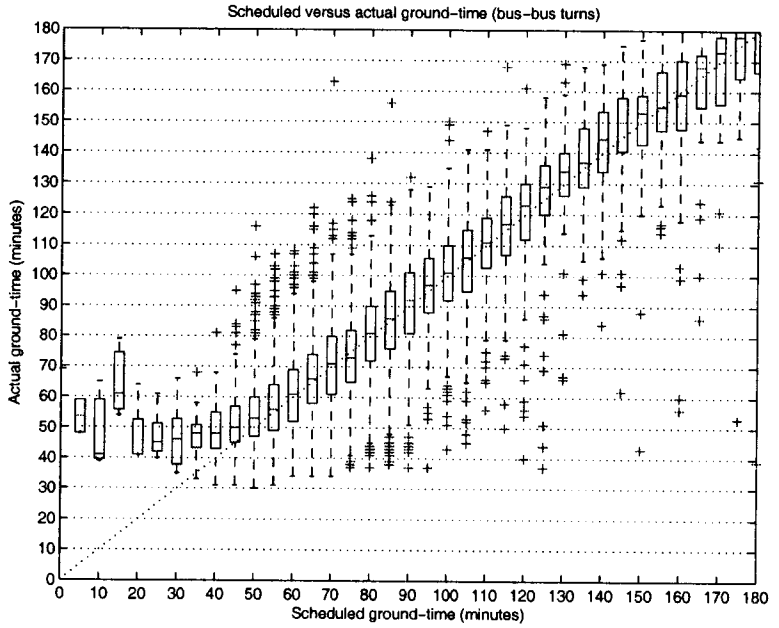

Figure 3: Actual ground-time as a function of scheduled ground-time.

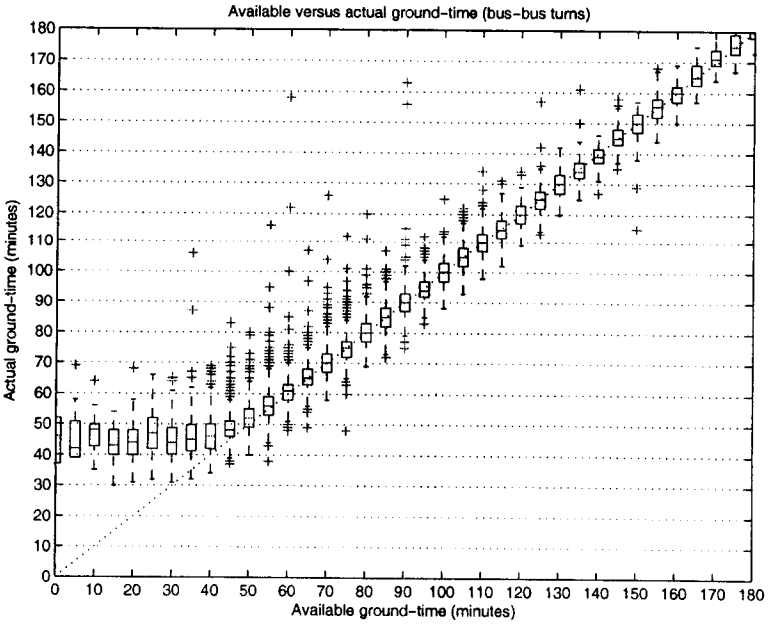

Figure 4: Actual ground-time as a function of available ground-time.

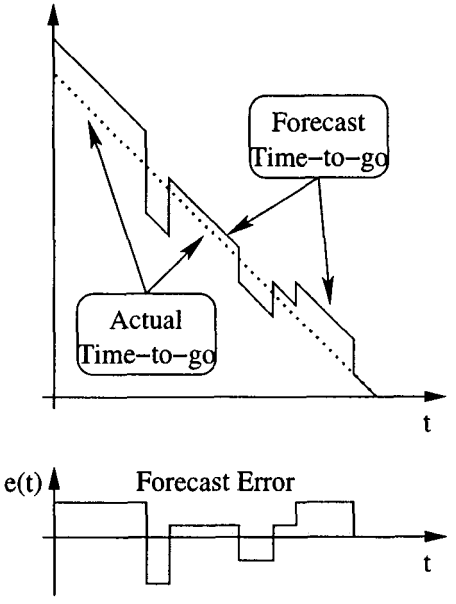

Figure 5: Instantaneous forecast error $e(t)$.

4.2 Bayesian forecasts using elapsed ground-time

The previous section demonstrated that, given the available ground-time, it is reasonable to forecast pushback using the average actual ground-time. This initial forecast can be significantly improved by using the *elapsed* ground-time to compute updates, especially for turns which are running unusually late and have exceeded the average actual ground-time.

For a turn with some given available ground-time, let the random variable $\mathbf{X}$ denote the actual ground-time. Characterize $\mathbf{X}$ by its complementary density function $G(t) \doteq \Pr(\mathbf{X} > t)$; for convenience the parametrization by the available ground-time is elided. At elapsed time t since onblock, the "perfect" forecast of time-to-go is simply $\mathbf{X} - t$. To approximate this perfect forecast, consider deterministic functions $\tilde{f}(t)$. The instantaneous forecast error $e(t)$ is then defined as the predicted time-to-go minus the actual time-to-go $\tilde{f}(t) - (\mathbf{X} - t)$ as illustrated in Figure 5.

The optimal $f(t)$ can then be constructed so that the expected total cost of these errors

according to the approximate quadratic cost-function of Equation (1) is minimized:

$$f \doteq \arg\min_{\{\tilde{f}\}} \mathrm{E}_{\mathbf{X}} \left[\int_0^{\mathbf{X}} \alpha(\tilde{f}(t) - (\mathbf{X} - t)) \cdot (\tilde{f}(t) - (\mathbf{X} - t) - \beta)\, dt \right].$$

In Appendix A this is solved to obtain

$$f(t) = \mathrm{E}[\mathbf{X} - t \mid \mathbf{X} > t] + \frac{\beta}{2}.$$

Note that while f is defined by an integral over a random-length interval, the final result depends on a pointwise-optimal Bayesian estimate of the *remaining life* $\mathbf{L}_t \doteq [\mathbf{X} - t \,|\, \mathbf{X} > t]$. This form is doubly appealing since it both naturally incorporates observations of the elapsed turn-time $\{\mathbf{X} > t\}$ and minimizes a reasonable cost-structure on the forecast error.

The following theorem is useful for characterizing the remaining life:

Theorem 1 (Remaining Life) *The moments of* $\mathbf{L}_t$ *are given by*

$$\mathrm{E}[\mathbf{L}_t^n] = \int_t^\infty \frac{G(v)}{G(t)} n(v-t)^{n-1}\, dv. \tag{2}$$

A derivation is presented in Appendix A. It then remains to approximate G given a set of samples of $\mathbf{X}$.

One approach is to estimate G nonparametrically. Given N iid samples $(x_1, \ldots, x_N)$, the standard histogram estimate of G is $\hat{G}(t) \doteq \#\{x_i > t\}/N$. This estimate can be substituted for G to approximate the moments of $\mathbf{L}_t$. In particular the first two moments can be used to approximate the Bayesian estimate $\mathrm{E}[\mathbf{L}_t]$ and its associated mean-square error $\mathrm{Var}(\mathbf{L}_t)$.

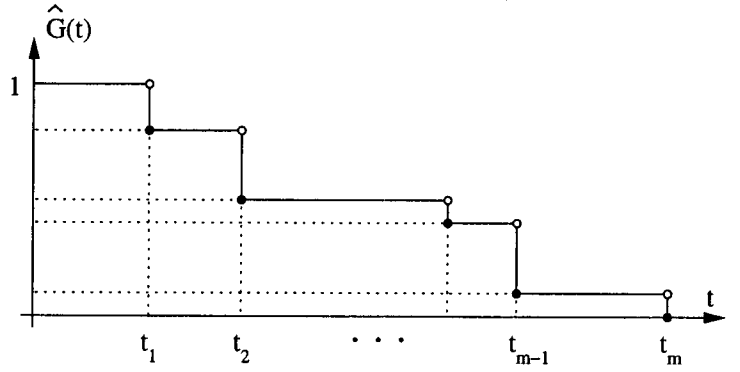

Figure 6: Illustration of $\hat{G}(t)$.

From Equation (2),

$$\begin{aligned}
\mathrm{E}[\mathbf{L}_t] &= \int_t^\infty \frac{G(v)}{G(t)}\,dv \\
&\approx \frac{1}{\#\{x_i > t\}} \int_t^\infty \#\{x_i > v\}\,dv \\
\mathrm{E}[\mathbf{L}_t^2] &= \int_t^\infty \frac{G(v)}{G(t)} 2(v-t)\,dv \\
&\approx \frac{2}{\#\{x_i > t\}} \int_t^\infty \#\{x_i > v\} v\,dv - \frac{2t}{\#\{x_i > t\}} \int_t^\infty \#\{x_i > v\}\,dv
\end{aligned}$$

The integrals can be further simplified because $\hat{G}$ has zero derivative except at a finite sequence of times $\min\{x_i\} = t_1 < \cdots < t_m = \max\{x_i\}$ as illustrated in Figure 6. Integrating over the intervals $[t_k, t_{k+1}]$ yields the identity

$$\int_t^\infty \#\{x_i > v\} v^n\,dv = \sum_{k=1}^{m} \#\{x_i \geq t_k\} \max\left\{0, \frac{t_k^{n+1} - \max\{t_{k-1}, t\}^{n+1}}{n+1}\right\} \tag{3}$$

where $t_0 \doteq 0$. Note that the term $\#\{x_i \geq t_k\}$ on the right-hand side is not a typo, as can be seen by examining the discontinuities in Figure 6.

An alternative approach is to fit some known parametric distribution to the data and directly compute the integrals in Theorem 1. Note that many analytically defined random variables have smooth complementary distribution functions. In this case the *hazard rate*

$r(t) \doteq -\frac{d}{dt} \ln G(t)$ is well-behaved. Remarkably one can then derive a simple closed-form recursion on the *dynamics* of $\mathrm{E}[\mathbf{L}_t^n]$ as a function of t:

Theorem 2 (Hazard-rate Remaining Life Recursion)

For a nonnegative random variable $\mathbf{X}$ *with bounded hazard rate* $r(t)$*, the moments of the remaining life* $\mathbf{L}_t \doteq [\mathbf{X} - t \,|\, \mathbf{X} > t]$ *follow the recursion*

$$\frac{\partial}{\partial t} \mathrm{E}[\mathbf{L}_t^n(t)] = -n \,\mathrm{E}[\mathbf{L}_t^{n-1}] + r(t) \,\mathrm{E}[\mathbf{L}_t^n]. \tag{4}$$

A derivation is given in Appendix A. The dynamics of the mean $\mu_L \doteq \mathrm{E}[\mathbf{L}_t]$ and variance $\lambda_L \doteq \mathrm{Var}(\mathbf{L}_t) = \mathrm{E}[\mathbf{L}_t^2] - \mathrm{E}[\mathbf{L}_t]^2$ are of particular interest:

$$\begin{aligned}
\frac{d}{dt}\mu_L(t) &= -1 + r(t)\,\mathrm{E}[\mathbf{L}_t] = -1 + r(t)\mu_L(t) \\
\frac{d}{dt}\lambda_L(t) &= \{-2\,\mathrm{E}[\mathbf{L}_t] + r(t)\,\mathrm{E}[\mathbf{L}_t^2]\} - 2\,\mathrm{E}[\mathbf{L}_t]\,\{-1 + r(t)\,\mathrm{E}[\mathbf{L}_t]\} \\
&= \{-2\mu_L(t) + r(t)[\lambda_L(t) + \mu_L^2(t)]\} - 2\mu_L(t)\,\{-1 + r(t)\mu_L(t)\} \\
&= r(t)[\lambda_L(t) - \mu_L^2(t)]
\end{aligned}$$

This yields a convenient ODE for computing the mean and variance:

$$y(t) \doteq \begin{pmatrix} \mu_L(t) \\ \lambda_L(t) \end{pmatrix}, \quad y(0) = \begin{pmatrix} \mathrm{E}[\mathbf{X}] \\ \mathrm{Var}(\mathbf{X}) \end{pmatrix}$$

$$\dot{y} = F(t, y) = \begin{pmatrix} -1 + r(t)\mu_L \\ r(t)(\lambda_L - \mu_L^2) \end{pmatrix}$$

$$\frac{\partial F}{\partial y} = \begin{pmatrix} r(t) & 0 \\ -2\mu_L r(t) & r(t) \end{pmatrix}$$

This ODE can be easily solved numerically, e.g. using *Matlab*.

To illustrate these approaches for approximating G, both approaches were applied against

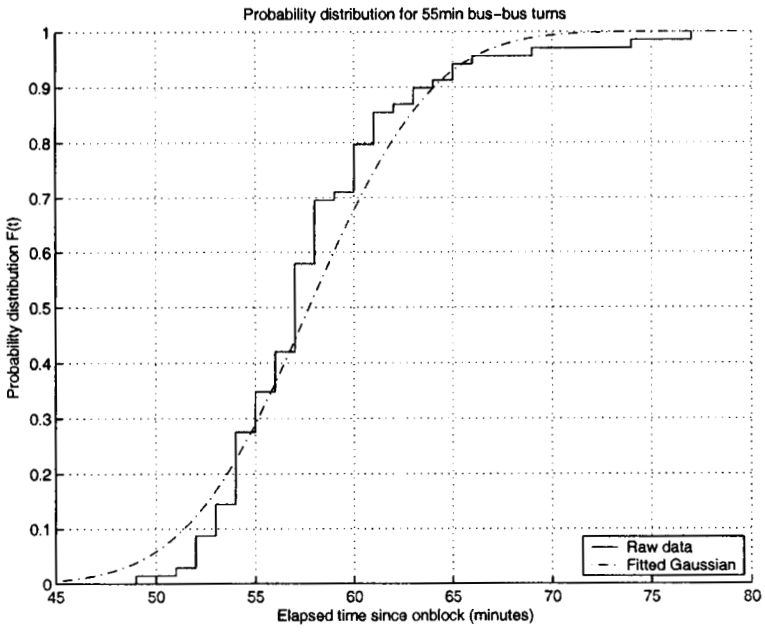

Figure 7: Probability distribution of 55min bus-bus turns.

the sample of simple bus-bus turns with 55min of available ground-time. These turns were selected because their available ground-time is close to the minimum available ground-time: they are neither guaranteed to be late nor guaranteed to have an excess of slack-time. The parametric (Gaussian) and nonparametric complementary distribution functions are shown in Figure 7. These complementary distributions are then used to derive a pair of forecasts for $\beta = 0$ as shown in Figure 8. For turn durations where a large number of datapoints are available there is little difference between the two forecast functions. However, for exceedingly long turns which are correspondingly rare, the nonparametric forecast must depend on only a handful of datapoints and some deviations are apparent.

For a specific sample of N turns, it is informative to consider the average instantaneous

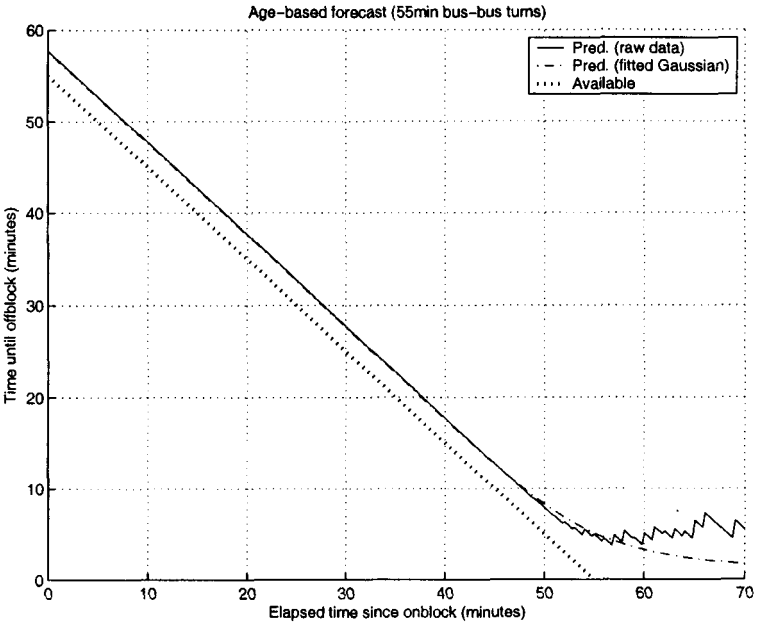

Figure 8: Age-based predictor for 55min bus-bus turns.

accuracy $\sigma(t)$ as a function of elapsed time:

$$N(t) \doteq \{\text{set of turns with duration} \geq t\}$$

$$\bar{e}(t) \doteq \frac{1}{|N(t)|} \sum_{i \in N(t)} e_i(t)$$

$$\sigma(t) \doteq \sqrt{\frac{1}{|N(t)| - 1} \sum_{i \in N(t)} \left(e_i(t) - \bar{e}(t)\right)^2}$$

The corresponding forecast accuracies are shown in Figure 9. While substantially better than the forecasts previously published in the literature, there is still a significant lower bound on the uncertainty of the age-based forecasts throughout most of the turn. Note that the apparent accuracy of the parametric forecast may be somewhat misleading, since it is derived under the assumption that the underlying data is in fact drawn from a Gaussian distribution.

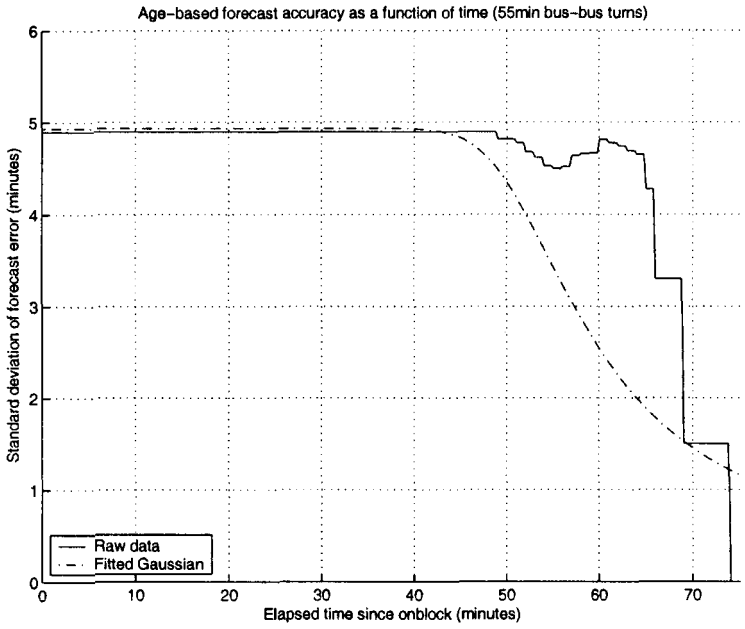

Figure 9: Age-based predictor: uncertainty as a function of time.

4.3 Forecasts using coupled updates of process status

Another reasonable approach for forecasting pushback time is to track the status (not-started/in-progress/completed) of the different processes composing a turn. Each process in a turn (e.g. catering) has a characteristic start-time and duration based on the available ground-time, and typically must occur in sequence with some predecessors (e.g. deplaning) and successors (e.g. boarding). If a process was running unusually late, one would expect this lateness to be transmitted to the successors and thus pushback to be correspondingly delayed. This assumption, that the processes can be divided into a sequence of phases, can be encoded into statistical models of varying complexity. It is expected that airline and air traffic controllers often use mental models of this form where the cause of an unusually late pushback is ascribed to a particular phase running late [12].

A turn can be approximately divided into three phases with stochastically independent durations: deplaning followed by "servicing" (catering, cleaning and fueling) followed by boarding. A turn which is running late is of greater importance operationally (early turns

are usually easy to delay), and thus the servicing phase is defined in such a way as to focus on whichever process is limiting. In particular the start of servicing is defined as the actual start-time of the subprocess which is scheduled to end last; the end of servicing is defined as the time when all three processes have completed. Under these assumptions the expected time-to-go until pushback is solely dependent on the available ground-time, the most recent status update, and the time elapsed since that update.

It is reasonable to expect that a turn with more available ground-time would not have shorter process durations. Furthermore, since the data were measured under real-world operating conditions where the deviation in each process was being monitored and controlled in real-time to adhere to the ALLEGRO target-times, for each process the amount of variation in time-to-go should be roughly constant with respect to the available ground-time, and should represent the minimum level of operationally acceptable (achievable) deviation around target-times. This expected behavior can be formalized by assuming that in a turn with available ground-time x, each process k is normally distributed with mean $\mu_k(x)$ and variance $\lambda_k(x)$ where $\mu_k(x)$ is non-decreasing and $\lambda_k(x)$ is constant.

Under these conditions the optimal regression of $\mu(x)$ is given by the PAVA algorithm [14] which essentially smoothes the usual estimated averages $\mu_{\text{est}}(x)$ to enforce the non-decreasing constraint. Note that using a standard regression would implicitly enforce the contrary assumption that turns with different available ground-time have no relationship at all. The smoothing effect can be seen by comparing Figures 10 and 11. The monotonic regression is able to pool information among turns of similar available ground-time, resulting in a much larger effective sample size and reduced noise.

The status-based forecast is constructed from the monotonic regressions as follows. When a turn of scheduled duration x arrives onblock, the initial forecast of time-to-go until offblock is just the average duration. As the turn progresses, the forecast counts downwards at a constant rate of -1sec/sec. When a phase of the turn changes status (starts or stops), the forecast is updated to the average time-to-go for that particular status change, and again

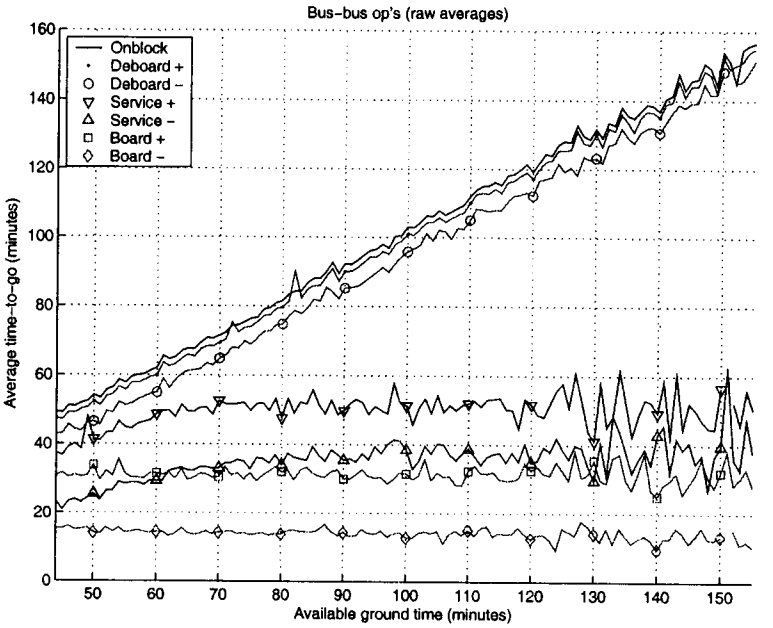

Figure 10: Average time-to-go as a function of process status.

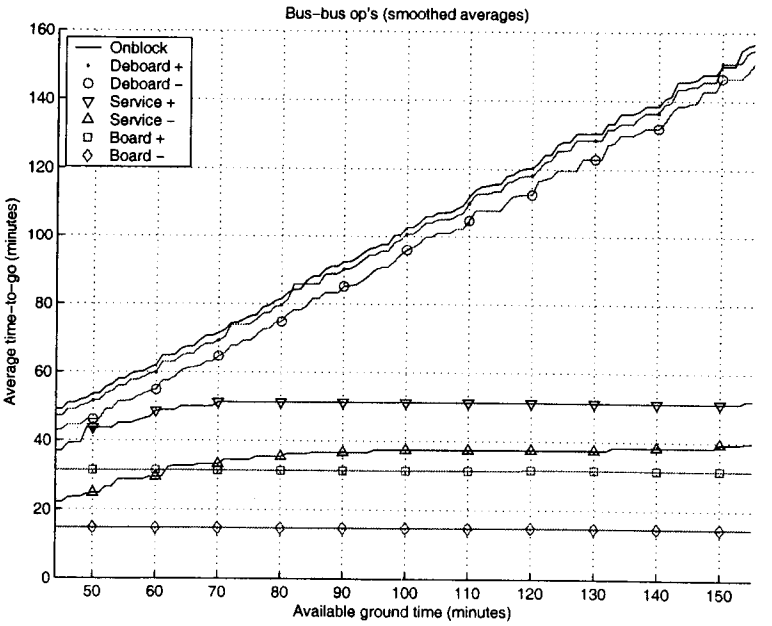

Figure 11: Average time-to-go, smoothed via PAVA algorithm.

the forecast again counts downwards at a constant rate of -1sec/sec. This straightfoward method of updating the forecast is sufficient for the majority of turns in the dataset. There are two conditions in which the update method is modified. If the predicted time-to-go until offblock is less than the actual time-to-go until the next change in process status, the forecast is not allowed to become negative but instead is held at zero. Also, if some process starts or finishes unusually early, the straightforward update method may produce a forecast which indicates the turn will take less than the minimum ground-time. When such a case arises, the forecast is instead updated to the minimum *feasible* remaining ground-time (i.e. the minimum ground-time minus the elapsed time).

Four examples are shown in Figures 12 through 15. To help characterize the quality of each forecast, an average forecast accuracy σ is computed for each turn:

$$\sigma \doteq \sqrt{\frac{1}{X}\int_0^X e^2(t)\,dt}$$

Forecasts for bus-bus flights which had punctual departures are shown in the first two figures; the following two figures correspond to delayed flights. For comparison each plot also shows the forecast based on available ground-time, and the "perfect" forecast based on the actual duration. Forecasts with good performance are shown in Figures 12 and 14. In particular for the late turn in Figure 14 the forecast successfully adapts to haste and delays in the turn processes and thus minimizes the forecast inaccuracy. In contrast, possible problems with status-based forecasts are shown in Figures 13 and 15, when phases deviate significantly from their expected times.

The average instantaneous forecast accuracy for a subset of the simple bus-bus turns is shown in Figure 16. Again a lower bound on the forecast accuracy throughout the turn is plainly apparent. In Figure 11 the average time-to-go clearly decreases as successive phases start and finish. However the standard deviation of the forecast accuracy in Figure 16 never drops below ±5min. For planning purposes, while at first glance it may appear desirable

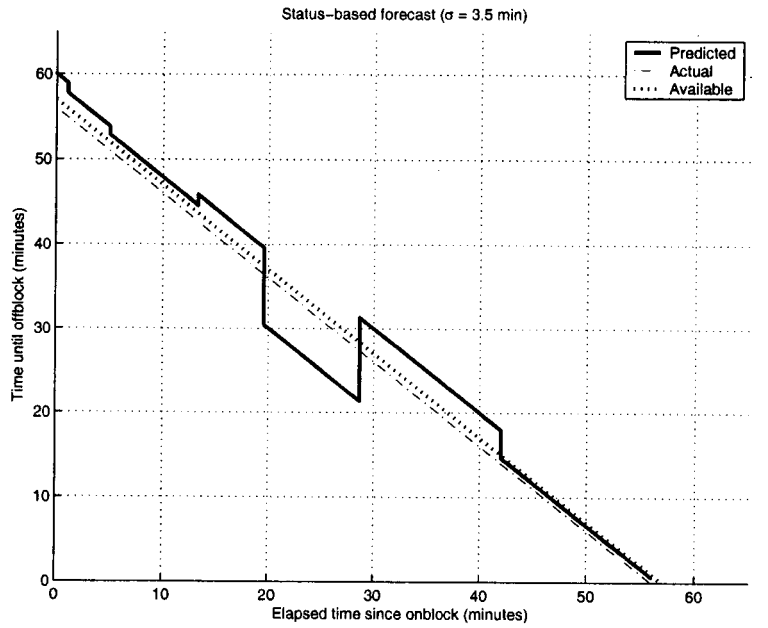

Figure 12: Status-based predictor: good performance on on-time flight.

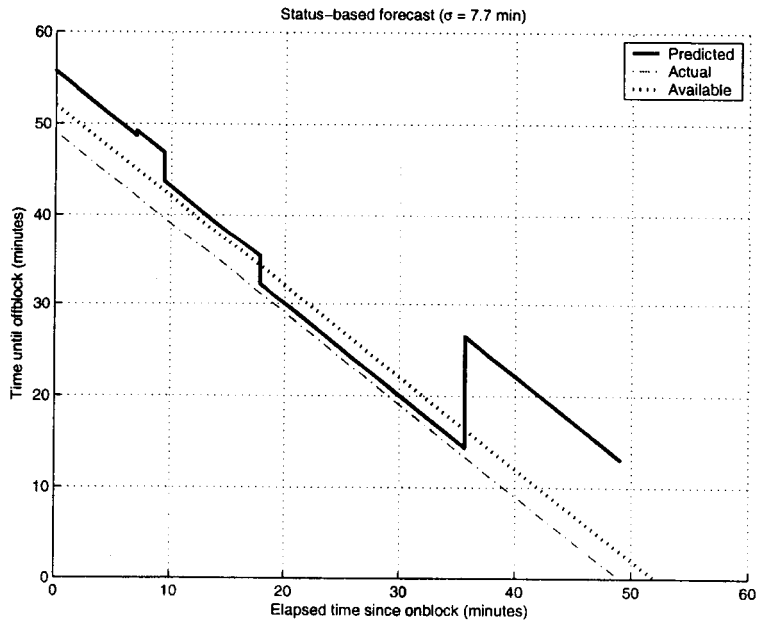

Figure 13: Status-based predictor: poor performance on on-time flight.

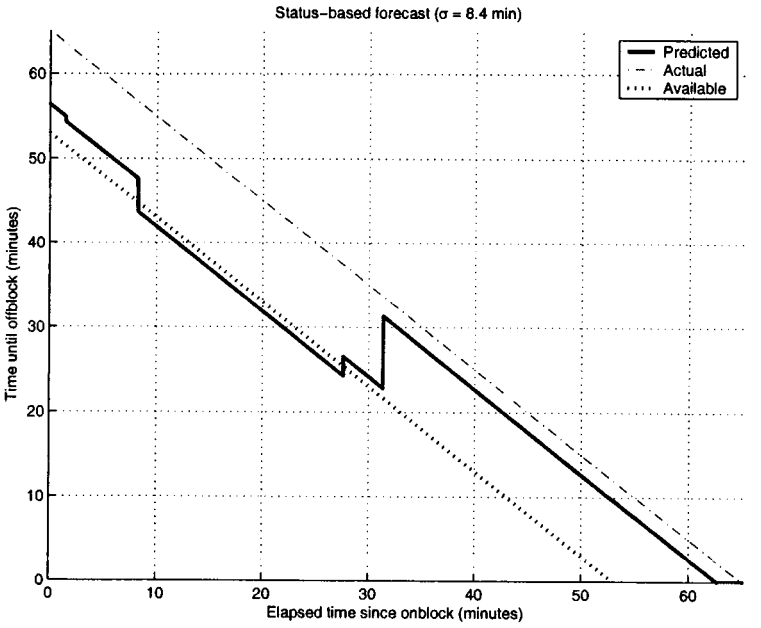

Figure 14: Status-based predictor: good performance on late flight.

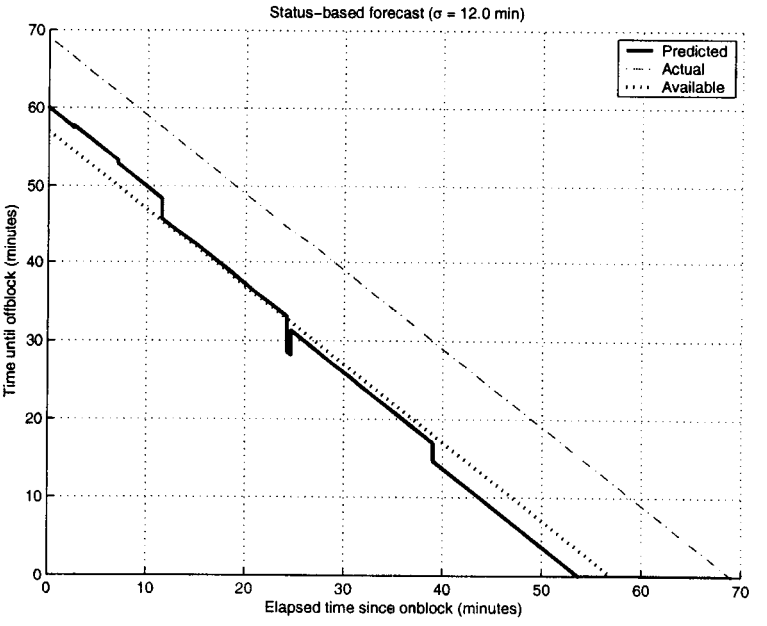

Figure 15: Status-based predictor: poor performance on late flight.

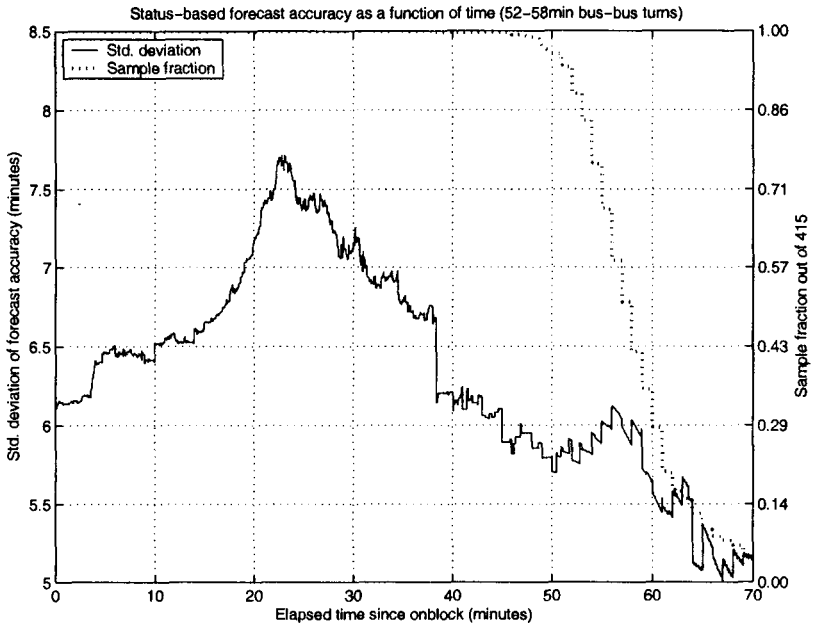

Figure 16: Status-based predictor: uncertainty as a function of elapsed time.

for airlines to supply downstream agents (airport and/or ATC) with every possible status update, the actual data only provide ambivalent support for this proposal.

A third metric for the forecast accuracy is the average instantaneous accuracy as a function of time-to-go (rather than elapsed time). It is natural to expect that the forecast accuracy should steadily improve for forecasts closer to the actual pushback. This expected behavior is observed in practice as seen from Figure 17. This result confirms that the status-based predictor is indeed behaving correctly; the lower bound observed from Figure 16 is due to the fact that the actual time-to-go is relatively uncertain and of course cannot be observed directly.

4.4 Combined pushback forecasts using status and age

At first glance, combining status-based and age-based forecasts should result in higher model fidelity and improvements in forecast accuracy. The obvious approach is to interpolate between status updates using the age-based equations. However the state-space over which

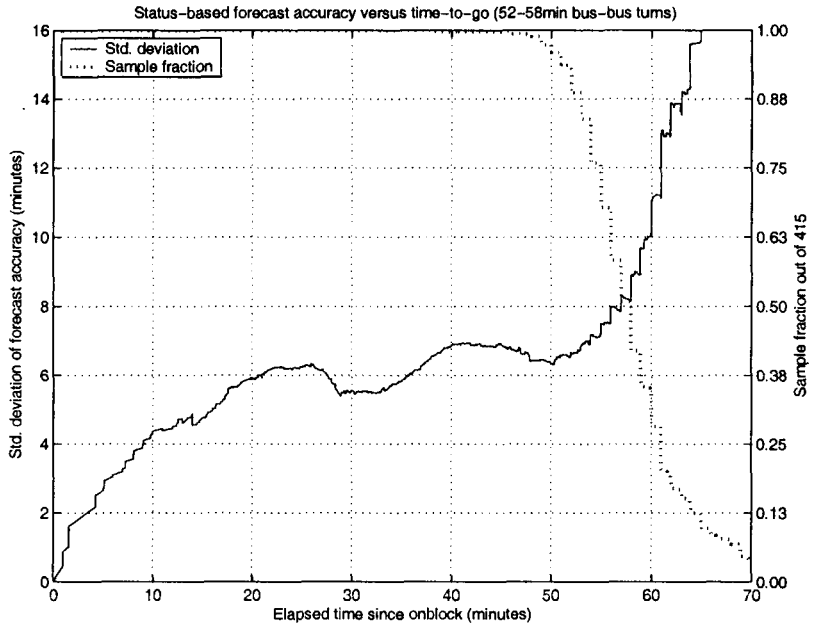

Figure 17: Status-based predictor: uncertainty as a function of time-to-go.

such a combined forecast is defined rapidly becomes enormous. Assuming there are three phases in a turn (deboard, service and board), and that the start/end epochs for each phase have only three possible observations (delayed, on-time or hastened), there are $3^6 = 729$ possible states to consider. Even a small improvement to this model such as adding a new epoch (actual onblock) and more accurate observations (very delayed or very hastened) increases the number of states to $5^7 = 78,125$; this is already 4.5 times larger than our current dataset covering 3 months of operations at a major hub airport, and nonparametric calibration of such a model is plainly infeasible.

The status-based forecast avoids this problem by assuming the phases have stochastically independent durations, while the age-based forecast has only a single continuous state-variable. Similar simplifying assumptions on the stochastic dependencies between processes must be made to yield a tractable combined forecast. For this purpose we propose a *bounding* technique. Under this bounding technique, the probabilistic distribution of each state variable is upper- and lower-bounded with a pair of Gaussian random variables. The "lower"

Gaussian has lower variance and is stochastically smaller; the "upper" Gaussian has higher variance and is stochastically larger. Dependencies among the state variables are encoded as correlations among the Gaussians: first the linear correlation coefficients of the raw data are computed, and then the upper/lower Gaussian bounds are scaled to leave the variances unchanged but the covariances fitted to produce identical correlation coefficients. This idea currently appears promising and is being systematically applied and validated using the ALLEGRO data.

5 Conclusions

Given their importance in maintaining an efficient and reliable air transportation system, it is remarkable that ground operations are not more transparent to both air traffic controllers and airline stations. Several decision-support tools for airport surface traffic are now in development, and there is published research linking the potential ATC benefits of these tools to the availability of accurate and timely pushback forecasts. Under weak assumptions on the structure of revenue loss due to deviations from schedule, airlines also stand to benefit financially from these ATC improvements through reductions in either the variability and/or average duration of ground delays.

However, airlines with the necessary infrastructure to provide such forecasts are the exception rather than the rule. To date only a few carriers have gone ahead and internally justified the business case to support infrastructure investment. It is worth noting that while current DST development and deployment has been heavily dependent on these well-equipped carriers, the extension of these DSTs to sites beyond the initial prototype airports may be significantly handicapped by a lack of high-quality pushback forecasts, an issue which has received little treatment in the literature.

Through collaboration with one of the best-equipped carriers, Deutsche Lufthansa AG, we have been able to perform several analyses supporting the development of the necessary high-

quality pushback forecasts. The simplest age-based forecasts only need measurements of the available and elapsed ground time for a given aircraft. The required investment is minimal since most major air carriers already record the onblock epoch automatically, and then the available ground time can be reported as soon as the onblock aircraft is paired operationally (i.e. barring any equipment swaps or mechanical cancellations) with an outbound flight. A more advanced status-based forecast integrates measurements of the many sub-processes in a turn. Automatic measurement of the start/completion times for all turn processes requires a larger airline investment, although there are concomitant benefits since each airline station gains the ability to continuously monitor, analyze and streamline its operations. For example, by comparing Figures 11 and 17, it appears in this case that forecast accuracy is better around the expected completion-times of each phase, and tends to be worse around the expected start-times. One possible interpretation is that individual phases are relatively well-controlled but that the gaps between phases are not as tightly regulated. This type of insight may help airline stations to optimize their internal processes. Finally, a proposal for a combined forecast based on both age and process status is now being tested.

Even after carefully filtering out a sample of real-world turn operations expected to exhibit minimal uncertainty, the standard deviation of forecast error for all of the forecast techniques is lower-bounded away from zero, indicating that turn operations have a substantial stochastic component. This intrinsic stochasticity imposes design and performance constraints on any automation or decision-aiding tool intended to smooth ground-flight handoffs. Such systems must be ready to cope with at least as much uncertainty in forecast pushback times and departure demand as reported above.

Rather then sending intra-ground event timestamps (e.g. cleaning ends, boarding begins, etc.) to all agents, the efficient strategy is for each air carrier to inform succeeding agents with only relevant but precise information: the predicted time-to-go until offblock for each flight, including the expected accuracy of each prediction. Ramp and ground controllers can then use methods such as those developed in [10] to predict the airborne time based

on various historic data (e.g. taxi-out times based on parking position and runway) and real-time data (e.g. number of aircraft on apron heading to take-off position), launching the flight into the ATC system. Accurate pushback predictions can lead to accurate estimates of departure demand, useful both for ATC planning purposes and for propagation downstream to provide improved predictability of arrival times to downstream airline stations.

6 Acknowledgements

The ground event timing data is remarkable, and enables an exceptional opportunity for systematic analyses of the turn process using accurate observations collected over a long period of time. We are indebted to Deutsche Lufthansa AG for this opportunity, and sincerely hope for a good working relationship between the International Center for Air Transportation at MIT, and the Traffic Flow Management group at Lufthansa. We would like to thank Arno Thon (Senior Manager ALLEGRO) and Manfred Rosenthal (Manager ALLEGRO) for their hospitality at Frankfurt International Airport, and for their professional and able assistance in obtaining and interpreting the ALLEGRO data.

References

[1] K. Andersson. Potential benefits of information sharing during the arrival process at hub airports. Master's thesis, Massachusetts Institute of Technology, 2000. Also available as Technical Report CSDL-T-1374, The Charles Stark Draper Laboratory, Inc., Cambridge MA.

[2] S. Atkins and C. Brinton. Concept description and development plan for the Surface Management System. *Journal of Air Traffic Control*, 44(1), January-March 2002.

[3] S. Atkins, C. Brinton, and D. Walton. Functionalities, displays and concept of use for the Surface Management System. In *Proceedings of the 21st DASC*, Irvine CA, Oct 2002. IEEE. IEEE Catalog Number 02CH37325C.

[4] R. Beatty, R. Hsu, L. Berry, and J. Rome. Preliminary evaluation of flight delay propagation through an airline schedule. In *Proceedings of the Second USA/Europe Air Traffic*

Management Seminar ATM-1998, Orlando FL, Dec 1998. EUROCONTROL and the US Federal Aviation Administration.

[5] V. H. L. Cheng. Collaborative automation systems for enhancing airport surface traffic efficiency and safety. In *Proceedings of the 21st DASC*, Irvine CA, Oct 2002. IEEE. IEEE Catalog Number 02CH37325C.

[6] W. Cooper, Jr., D. E. A. Cherniavsky, J. S. DeArmon, J. G. Foster, D. M. J. Mills, D. S. C. Mohleji, and F. Z. Zhu. Determination of minimum push-back time predictability needed for near-term departure scheduling using DEPARTS. In *Proceedings of the Fourth USA/Europe Air Traffic Management Seminar ATM-2001*, Santa Fe NM, Dec 2001. EUROCONTROL and the US Federal Aviation Administration.

[7] W. W. Cooper, Jr., R. H. Cormier, J. G. Foster, D. M. J. Mills, and D. S. Mohleji. Use of the Departure Enhanced Planning and Runway/Taxiway Assignment System (DEPARTS) for optimal departure scheduling. In *Proceedings of the 21st DASC*, Irvine CA, Oct 2002. IEEE. IEEE Catalog Number 02CH37325C.

[8] R. G. Gallager. *Discrete Stochastic Processes.* The Kluwer International Series in Engineering and Computer Science. Communications and information theory. Kluwer Academic Publishers, Boston, 1998.

[9] B. J. Glass and Y. Gawdiak. Integrated navigation applications in airport surface traffic management. In *4th St. Petersburg International Conference on Integrated Navigation Systems*, pages 105–113, St. Petersburg, Russia, May 26-28 1997. AIAA A97-30869 07-35.

[10] H. Idris, J.-P. Clarke, R. Bhuva, and L. Kang. Queueing model for taxi-out time estimation. *Air Traffic Control Quarterly*, 10(1):1–22, 2002.

[11] S. I. Januszewski. The effect of air traffic delays on airline prices. Job market paper, Economics Department, Massachusetts Institute of Technology, 2003. Currently available online at **http://econ-www.mit.edu/graduate/candidates/research.htm?athenan=silke**. After Sept. 2003, available from MIT Libraries as part of Januszewski's thesis.

[12] P. Martin, A. Hudgell, S. Vial, N. Bouge, N. Dubois, H. de Jonge, and O. Delain. Potential applications of Collaborative Decision Making. EEC Note No. 9/99, Flight Data Research, Eurocontrol Experimental Centre, Jul 1999.

[13] A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., 2002.

[14] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference*, chapter 1, pages 1–58. Wiley series in probability and mathematical statistics. John Wiley and Sons Ltd., GB, 1998.

[15] RTCA Select Committee for Free Flight Implementation. *National Airspace System Concept of Operations Addendum 4: Free Flight Phase 2*, 2000.

[16] G. Theis. Telematik Anwendungen im Luftverkehr. *Internationales Verkehrswesen*, 54(5):225–8, 2002.

[17] W. W. Vanderson. Improving aircraft departure time predictability. Master's thesis, Massachusetts Institute of Technology, September 2000.

[18] J. D. Welch and S. R. Bussolari. Initial Surface Management System preliminary operational concept. ATC Project Memorandum 92PM-AATT-0008, Massachusetts Institute of Technology Lincoln Laboratory, Lexington MA, Aug 31 2000.

A Derivations for Bayesian age-based forecasts

A.1 Age-Based Forecast

Consider a nonnegative random variable $\mathbf{X}$. When $\mathbf{X}$ is the lifetime of some object or process, $\mathbf{L}_t \doteq [\mathbf{X} - t \,|\, \mathbf{X} > t]$ is the *remaining life* given that the lifetime has exceeded some threshold t. It is of interest to forecast the remaining life via a deterministic function $f(t)$ to minimize the expected integrated quadratic cost

$$J(f) \doteq \mathrm{E}_{\mathbf{X}} \left[\int_0^{\mathbf{X}} \alpha(f(t) - (\mathbf{X} - t)) \cdot (f(t) - (\mathbf{X} - t) - \beta) \, dt \right]$$

where α and β are fixed nonnegative location/scale parameters.

Note that the expectation operator and the integral cannot be directly exchanged since both the integrand and the integration interval depend on $\mathbf{X}$. Let $F(x) \doteq \Pr(\mathbf{X} \leq t)$ be the cumulative distribution of $\mathbf{X}$. Then the arg-min of J can be explicitly expanded as

$$\arg\min_{\{f\}} J(f) = \arg\min \int_0^{\infty} \int_0^{x} \alpha(f(t) - (x - t)) \cdot (f(t) - (x - t) - \beta) \, dt \, dF(x).$$

Assuming the integrals converge and can be exchanged, this yields

$$= \arg\min \int_0^\infty \int_t^\infty \alpha(f(t) - (x-t)) \cdot (f(t) - (x-t) - \beta)\, dF(x)\, dt$$

Expanding the inner integral (and neglecting $\alpha \geq 0$) yields

$$\begin{aligned}(f^2(t) - \beta f(t)) \int_t^\infty 1\, dF(x) - 2f(t) \int_t^\infty (x-t)\, dF(x) \\ + \int_t^\infty (x-t)^2\, dF(x) + \beta \int_t^\infty (x-t)\, dF(x).\end{aligned}$$

Assuming the last two terms converge, they do not affect the arg-min and can be ignored. The remaining terms can be simplified by inspection and substituted into the outer integral to yield

$$\arg\min_{\{f\}} J(f) = \arg\min \int_0^\infty (f^2(t) - \beta f(t))G(t) - 2f(t)G(t)\,\mathrm{E}\big[\mathbf{X} - t | \mathbf{X} > t\big]\, dt.$$

In this case f does not need to be minimized as a whole function but only pointwise. For fixed time t, $f(t)$ can be treated as an unconstrained free parameter:

$$\begin{aligned}\frac{\partial}{\partial f}\left\{f^2 - (2\,\mathrm{E}\big[\mathbf{X} - t | \mathbf{X} > t\big] + \beta)f\right\} &= 0 \\ 2f - (2\,\mathrm{E}\big[\mathbf{X} - t | \mathbf{X} > t\big] + \beta) &= 0 \\ \longrightarrow f(t) &= \mathrm{E}\big[\mathbf{X} - t | \mathbf{X} > t\big] + \frac{\beta}{2}\end{aligned}$$

A.2 Remaining Life Theorem

It is thus of interest to compute the moments of $\mathbf{L}_t$. In this case it is more convenient to characterize $\mathbf{X}$ by its complementary distribution $G(t) \doteq \Pr(\mathbf{X} > t)$. For any nonnegative

random variable $\mathbf{Z}$ one has the identity [8, p. 8]

$$\mathrm{E}[\mathbf{Z}] = \int_0^\infty G_{\mathbf{Z}}(t)\,dt. \tag{5}$$

For $n \in \mathbb{N}^+$ and $t, \tau \geq 0$, $\mathbf{L}_t^n$ is nonnegative with complementary distribution function

$$\begin{aligned}
\Pr(\mathbf{L}_t^n > \tau) &= \Pr\big((\mathbf{X} - t)^n > \tau \mid \mathbf{X} > t\big) \\
&= \Pr(\mathbf{X} > \tau^{1/n} + t \mid \mathbf{X} > t) \\
&= \frac{G(\tau^{1/n} + t)}{G(t)}
\end{aligned}$$

Applying (5) yields the desired result:

$$\begin{aligned}
\mathrm{E}[\mathbf{L}_t^n] &= \int_0^\infty \frac{G(\tau^{1/n} + t)}{G(t)}\,d\tau \\
&\cdots \begin{cases} v = \tau^{1/n} + t, & dv = n^{-1}\tau^{(1-n)/n}\,d\tau, \\ \tau = (v-t)^n, & d\tau = n^{+1}\tau^{(n-1)/n}\,dv = n(v-t)^{n-1}\,dv \end{cases} \\
&= \int_t^\infty \frac{G(v)}{G(t)} n(v-t)^{n-1}\,dv
\end{aligned}$$

A.3 Hazard-Rate Remaining Life Recursion

Recall Liebniz' integral rule:

$$\frac{\partial}{\partial z}\int_{a(z)}^{b(z)} f(x,z)\,dx = \int_{a(z)}^{b(z)} \frac{\partial f}{\partial z}\,dx + f(b(z),z)\frac{\partial b}{\partial z} - f(a(z),z)\frac{\partial a}{\partial z}$$

For the case $n = 1$,

$$\begin{aligned}
\frac{\partial}{\partial t}\mathrm{E}[\mathbf{L}_t] &= -1 - \frac{\dot{G}(t)}{G(t)}\int_t^\infty \frac{G(v)}{G(t)}\,dv \\
&= -1 + r(t)\,\mathrm{E}[\mathbf{L}(t)]
\end{aligned}$$

while for $n > 1$,

$$\begin{aligned}\frac{\partial}{\partial t}\,\mathrm{E}\big[\mathbf{L}_t^n\big] &= -n\int_t^\infty \frac{G(v)}{G(t)}(n-1)(v-t)^{n-2}\,dv - \frac{\dot{G}(t)}{G(t)}\int_t^\infty \frac{G(v)}{G(t)}n(v-t)^{n-1}\,dv \\ &= -n\,\mathrm{E}\big[\mathbf{L}_t^{n-1}\big] + r(t)\,\mathrm{E}\big[\mathbf{L}_t^n\big]\end{aligned}$$

It is necessary to make two separate derivations since different terms from the right-hand side of Liebniz' integral rule are contributing in each case. Except in the trivial condition $\mathbf{L}_0 = 0$, however, $\mathrm{E}\big[\mathbf{L}_t^0\big] = \mathrm{E}\big[1\big] = 1$ and a single formula suffices:

$$\frac{\partial}{\partial t}\,\mathrm{E}\big[\mathbf{L}_t^n(t)\big] = -n\,\mathrm{E}\big[\mathbf{L}_t^{n-1}\big] + r(t)\,\mathrm{E}\big[\mathbf{L}_t^n\big].$$