International Journal of Wireless and Ad Hoc Communication

♦ ISSN: 2692-4056 Vol: 11, No. 01, 2027 ♦ PP. 22–33

Article DOI: https://doi.org/10.54216/IJWAC.110103



FULL LENGTH ARTICLE

Dwell Time and Fallback of 5G New Radio Connections on the Road: A Survival Analysis of Device-Side Traces with Early Warning

Lobna Osman[1,*]

[1] Delta Higher Institute for Engineering & Technology, Department of Electronics and Communications Engineering, Egypt

Email: englobna20@yahoo.com

Received: July 08, 2026 Revised: September 20, 2026 Accepted: November 09, 2026 ⋆ Corresponding author

ABSTRACT

A phone that shows a 5G icon in a moving car does not keep it for long, and how long it keeps it has rarely been measured as a duration. This paper treats every stretch of New Radio service in public device-side traces from a commercial network as a survival-time observation, with trace ends handled as censoring. Kaplan–Meier curves, five parametric families, a Cox model and a discrete-time hazard regression describe the dwell time; logistic and gradient-boosted classifiers, validated with whole traces held out, then try to announce a fallback ten seconds ahead. On the road the median dwell is 34 s, against more than an hour and a half when the phone is still. A two-component exponential mixture fits best: a fresh connection is about eighteen times more likely to drop in the next second than one that has lasted a minute. Signal level, speed and recent instability move the hazard; the neighbour-cell gap does not. At an alarm duty of 10% the boosted model warns of 83% of fallbacks, and a map of earlier fallbacks improves it further, since they recur at the same places. Mean download rate halves across a fallback while the median barely moves.

Keywords: 5G New Radio ▪ Dwell time ▪ Survival analysis ▪ Fallback prediction ▪ Vehicular mobility

1. INTRODUCTION

Commercial 5G arrived as an overlay. In the non-standalone architecture that every early launch used, a phone stays anchored to an LTE cell and receives a New Radio (NR) carrier as a secondary leg when one is within reach [1, 2]. The secondary leg can be released at any moment, by the network or by a radio-link failure, and the phone is then back on LTE alone, or on 3G where LTE is thin. Users see a 5G icon that comes and goes. Applications see a capacity that changes by an order of magnitude without notice, the hardest case there is for the throughput predictors on which adaptive video players depend [3].

Measurement campaigns have documented the effect many times. Drive tests in the first deployments found NR coverage in patches [4, 5], later campaigns counted handovers and classified them [6, 7], and an operator-side study of a whole country showed where handover failures concentrate [8]. What these studies report are shares and rates. The fraction of time or distance on NR, handovers per kilometre and failures per thousand attempts all leave open whether the time on NR is made of many short visits or of a few long ones, and the two cases call for different behaviour from a player, a scheduler or an edge server.

The question we ask is about durations. Let T be the time a connection spends on NR from the second it is established to the second it falls back. Three things about T matter in practice: its distribution, in particular the lower tail; the

© 2027 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

hazard, that is, the chance that a connection which has lasted t seconds ends in the next one; and whether measurements available on the device can announce the end some seconds ahead. Durations in traces are censored, since a log that stops while the phone is on NR shows only that T exceeded what was observed, and averaging the completed stretches alone would bias the answer downward. Survival analysis exists for this situation [9]. To our knowledge it has not been applied to NR dwell time in a public commercial dataset.

We apply it to the driving and static traces of Raca et al. [10]. The paper makes four contributions. It estimates the dwell-time and outage-time distributions with censoring handled, and compares the road with a parked phone. It shows that the fallback hazard falls steeply with elapsed time, that a two-component exponential mixture describes it better than the Weibull and log-normal forms, and what the mixture implies for the remaining life of a connection that has already lasted a given number of seconds. It quantifies, per standard deviation, how signal level, speed, load and recent instability move the hazard. And it measures how far ahead a fallback can be announced when whole traces are held out, how far that ability transfers across applications and what a map of earlier fallbacks adds, together with what a fallback costs in download rate. Section 2 reviews related work, Section 3 sets out the data and the estimators, Section 4 gives the results, Section 5 discusses their use and limits, and Section 6 concludes.

2. RELATED WORK

2.1 Measurements of commercial 5G

Narayanan et al. [4] published the first independent look at commercial 5G on smartphones, and Xu et al. [5] followed with coverage, throughput and energy measurements on an operational mid-band network. A second round of studies widened the scope to power and application quality [11] and to throughput prediction from location and context [12], and mobility came into focus with Hassan et al. [6], who dissected the handover procedures of non-standalone and standalone deployments from driving data and built a predictor for them. Others followed. Kousias et al. [7] examined handover events in non-standalone networks in Rome; Ghoshal et al. [13] drove across the United States to compare operators; Fezeu et al. [14] measured mid-band deployments on two continents. On the operator side, Kalntis et al. [8] analysed handovers countrywide and found that handovers to 3G are about 6% of all handovers and about 75% of the failures. The dataset used here [10] belongs to the first generation of these campaigns and has the merit of being public, per-second and annotated with the serving mode.

2.2 Dwell time in cellular models

Time spent in a cell is among the oldest quantities in teletraffic theory, older than the handset. Hong and Rappaport [15] derived channel holding times for handoff analysis under simple mobility assumptions, Zonoozi and Dassanayake [16] fitted cell residence times from simulated mobility with a generalised gamma law, and Fang and Chlamtac [17] showed how far performance results depart from the exponential case once residence times are allowed a general distribution. Lin et al. [18] later obtained sojourn times in closed form for a random-waypoint model on a tessellated plane. These are models of geometry and motion. NR dwell time under a non-standalone anchor depends in addition on the operator's measurement configuration and release timers [19, 20], which no mobility model contains, and so it has to be estimated from data.

2.3 Predicting mobility events

Learning-based handover management has a large literature, surveyed by Mollel et al. [21]. Most of it works on simulated channels. Among the studies on measured data, Hassan et al. [6] predict handover type from radio measurements, Chou et al. [22] train sequence models to warn of link breakdown on a non-standalone metro railway and report an F1 of 0.85 at a three-second horizon from 23 events, and Kabeer et al. [23] evaluate sequence-based handover detection and avoidance using a released multi-operator drive-test dataset. Two habits in this literature concern us. Scores are often computed on windows drawn at random from the same drives that supplied the training windows, which lets a model recognise a road, and results are reported as a single F1 at one threshold, which hides the trade between warnings and false alarms. We hold out whole traces. The trade is reported in full.

3. DATA AND METHODS

3.1 Traces and episodes

The dataset [10] holds 83 traces logged on a single handset on the network of one Irish operator between 20 November 2019 and 27 February 2020. Each row carries a timestamp, GPS position and speed, the serving cell, the network mode reported by the logging application, the reference-signal received power and quality (RSRP, RSRQ), signal-to-noise ratio (SNR), channel quality indicator (CQI), the strongest neighbour's RSRP, and the downlink and uplink bitrates. Rows are nominally one second apart: 85.5% of consecutive rows differ by exactly 1 s, 10.0% by 2 s, and 4.2% share a timestamp. Table 1 summarises the collection. Sentinel values (an RSRP below −150 dBm, 73 rows for the serving cell and 91 for the neighbour) were set to missing. Nothing else was altered.

Table 1. The traces by mobility pattern

	Driving	Static
Traces	60	23
file download	16	5
video streaming	44	18
Logged seconds	103,545	85,166
Share on NR (%)	56.3	91.8
Share on LTE (%)	29.6	5.3
Share on 3G (%)	14.0	2.9
NR episodes	377	35
ending in a fallback	345	12
open at trace start	42	21
Outage episodes	363	14
Median speed (km/h)	32	0

Write $m_t \in \{0, 1\}$ for the indicator that the mode at row t is NR. An NR episode is a maximal run of rows with $m_t = 1$

inside one trace, and an outage episode a maximal run with $m_t = 0$. For episode i with first timestamp a_i, the observed duration and the event indicator are

$$T_i = \begin{cases} b_i - a_i, & \delta_i = 1, \\ c_i - a_i + 1, & \delta_i = 0, \end{cases} \tag{1}$$

where b_i is the timestamp of the first row of the following episode, c_i the last timestamp of the trace, and $\delta_i = 0$ marks an episode that the end of the trace cut short (right censoring). An episode already under way when the trace begins is left-truncated at an unknown origin, so the main analysis excludes those 42 driving episodes and keeps 335, of which 305 end in a fallback, and the estimate with all 377 is reported beside it. Of the 345 driving fallbacks, 337 land on LTE and 8 on 3G. Figure 1 shows one trace and the mode shares.

3.2 Survival estimators

The survival function, the hazard and their relation are

$$S(t) = \Pr(T > t), \qquad h(t) = -\frac{d}{dt} \log S(t). \tag{2}$$

With distinct event times $t_{(1)} < \cdots < t_{(J)}$, d_j events at $t_{(j)}$ and n_j episodes still at risk just before it, the Kaplan–Meier estimator [24] is

$$\hat{S}(t) = \prod_{j : t_{(j)} \le t} \Big(1 - \frac{d_j}{n_j}\Big). \tag{3}$$

A censored episode leaves the risk set without counting as an event. Two groups are compared with the log-rank statistic [25],

$$\chi^2 = \frac{\big[\sum_j (d_{1j} - n_{1j} d_j / n_j)\big]^2}{\sum_j \dfrac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}}, \tag{4}$$

referred to a chi-squared distribution with one degree of freedom. Because medians are unstable when a curve is flat near 0.5, we also report the restricted mean survival time [26],

$$\mu(\tau) = \int_0^{\tau} \hat{S}(t)\, dt, \qquad \tau = 600 \text{ s}. \tag{5}$$

Episodes from one drive share a route, a day and a cell plan, so they are not independent, and intervals are therefore percentile intervals from a bootstrap that resamples whole traces [27], 500 replicates each, in place of the Greenwood formula, which would treat every episode as a separate draw. Estimation used the lifelines package [28].

3.3 Parametric dwell-time models

A parametric family with density f_θ and survival S_θ is fitted by maximising the censored likelihood

$$\ell(\theta) = \sum_i \big[\delta_i \log f_\theta(T_i) + (1 - \delta_i) \log S_\theta(T_i)\big]. \tag{6}$$

Five families are tried. The exponential, $S(t) = e^{-t/\lambda}$, is the memoryless benchmark that classical teletraffic analysis assumes [15]. The Weibull,

$$S(t) = \exp\big[-(t/\lambda)^k\big], \qquad h(t) = \frac{k}{\lambda}\Big(\frac{t}{\lambda}\Big)^{k-1}, \tag{7}$$

allows a hazard that falls ($k < 1$) or rises ($k > 1$) as a power of time. The log-normal and log-logistic have $S(t) = 1 - \Phi\big((\log t - \mu)/\sigma\big)$ and $S(t) = [1 + (t/\alpha)^\beta]^{-1}$. The fifth is a mixture of two exponentials [29],

$$S(t) = p\, e^{-t/\lambda_1} + (1 - p)\, e^{-t/\lambda_2}, \qquad \lambda_1 < \lambda_2, \tag{8}$$

which reads the episodes as two populations: a fraction p of unstable attachments with mean life λ_1 and a remainder of settled ones with mean life λ_2. Its hazard is

$$h(t) = \frac{\frac{p}{\lambda_1} e^{-t/\lambda_1} + \frac{1-p}{\lambda_2} e^{-t/\lambda_2}}{p\, e^{-t/\lambda_1} + (1 - p)\, e^{-t/\lambda_2}}, \tag{9}$$

which starts at $p/\lambda_1 + (1-p)/\lambda_2$ and decreases to the floor $1/\lambda_2$ as the unstable population dies out. Two further quantities follow in closed form and are of direct use to an application. Given that a connection has lasted t seconds, the probability that it belongs to the settled population and its expected remaining life are

$$w(t) = \frac{(1 - p)\, e^{-t/\lambda_2}}{S(t)}, \qquad m(t) = \mathbb{E}[T - t \mid T > t] = \frac{p\lambda_1 e^{-t/\lambda_1} + (1 - p)\lambda_2 e^{-t/\lambda_2}}{S(t)}. \tag{10}$$

Models are ranked by the Akaike criterion [30], $\text{AIC} = 2q - 2\ell(\hat{\theta})$ with q the number of parameters.

3.4 Hazard regression

Two regressions relate the hazard to covariates. The first is a Cox model [31] on conditions at the start of the episode,

$$h(t \mid \mathbf{z}_i) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{z}_i), \tag{11}$$

estimated from the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i : \delta_i = 1} \frac{\exp(\boldsymbol{\beta}^\top \mathbf{z}_i)}{\sum_{l \in \mathcal{R}(T_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_l)}, \tag{12}$$

where $\mathcal{R}(t)$ is the set of episodes at risk at t. The covariates $\mathbf{z}_i$ are speed, RSRP, RSRQ and SNR in the first second, the logarithm of the length of the preceding outage, and an indicator for download traces; continuous ones are standardised, so that e^β is a hazard ratio per standard deviation, and discrimination is summarised by Harrell's concordance [32].

Entry conditions go stale. The second regression therefore uses every second on NR as an observation with its current covariates $\mathbf{x}_{it}$ and models the probability that the episode ends before the next row, with a complementary log-log link [33, 34]:

$$\log\big[-\log(1 - h_{it})\big] = \alpha + \gamma \log(1 + e_{it}) + \boldsymbol{\beta}^\top \mathbf{x}_{it}, \tag{13}$$

where e_{it} is the elapsed time in the episode. Under this link e^β is again a hazard ratio, now for a time-varying covariate. The vector $\mathbf{x}_{it}$ holds RSRP, RSRQ, SNR, CQI, speed, the neighbour gap (neighbour RSRP minus serving RSRP), the number of fallbacks in the preceding 120 s and the download indicator. Standard errors in both regressions are of the sandwich form clustered on the trace [35]; the second model was

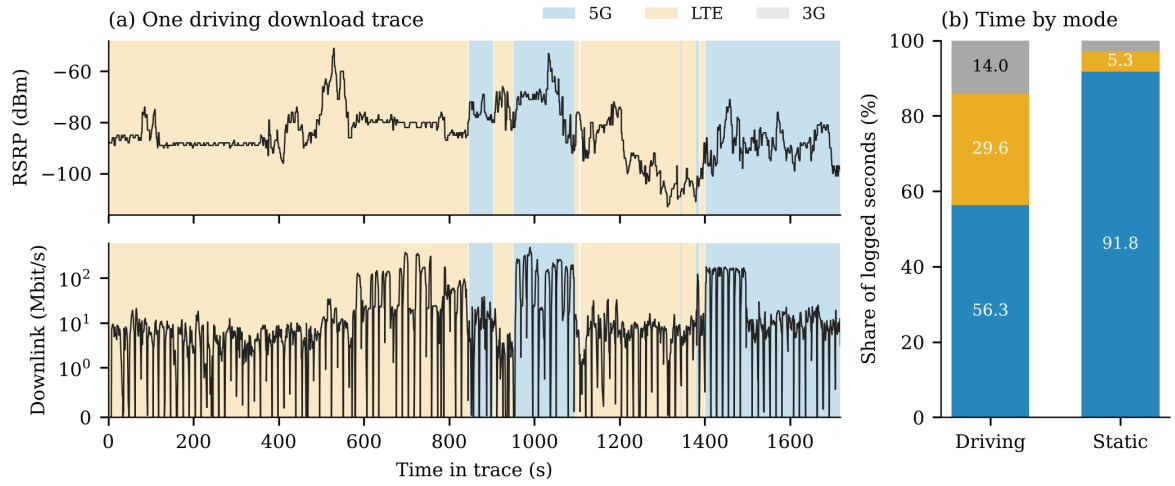


Figure 1. A driving download trace with the serving mode shaded behind the signal level and the downlink rate (a), and the share of logged seconds by mode for the two mobility patterns (b).

fitted with statsmodels [36].

3.5 Early warning

For a horizon of h seconds, each second t on NR receives the label

$$y_t^{(h)} = \mathbf{1}\left[\delta_i = 1 \text{ and } 0 < a_i + T_i - t \le h\right], \tag{14}$$

and seconds within h of a censored end are dropped, their label being unknown. Horizons of 5, 10 and 20 s are examined. The feature vector is built only from rows up to t and has three families: radio (the five measurements above, the neighbour gap, and for RSRP, RSRQ, SNR and CQI the mean over the last five rows and the change across them, plus the five-row minimum and ten-row standard deviation of RSRP); context (speed, seconds since the last change of serving cell, logarithm of the downlink rate); and history (logarithm of elapsed time on NR and the count of fallbacks in the last 120 s).

Six scorers are compared. The baseline is a threshold on the five-row mean RSRP; three are logistic regressions on standardised, median-imputed inputs,

$$\Pr\left(y_t^{(h)} = 1 \mid \mathbf{u}_t\right) = \frac{1}{1 + \exp(-w_0 - \mathbf{w}^\top \mathbf{u}_t)}, \tag{15}$$

with $\mathbf{u}_t$ the history, the radio, or all features. Two are gradient-boosted trees [37] on the radio or on all features, the additive score

$$F_M(\mathbf{u}) = F_0 + \nu \sum_{m=1}^{M} g_m(\mathbf{u}) \tag{16}$$

being built from $M = 200$ depth-four trees g_m with learning rate $\nu = 0.05$ on histogram-binned inputs [38]. No setting was tuned. Scores are out-of-fold predictions from five folds that keep every trace whole [39], so that no model is tested on a drive it has seen.

With fewer than one positive second in ten, the area under the precision-recall curve is the informative summary [40]. We report it as average precision,

$$\mathrm{AP} = \sum_k (R_k - R_{k-1})\, P_k, \tag{17}$$

with P_k and R_k the precision and recall at the k-th threshold, beside the area under the ROC curve (AUROC) and the Brier score [41], $N^{-1}\sum_t (\hat{p}_t - y_t)^2$. A per-second score is still one step from a usable alarm, so performance is also counted per fallback. For a threshold θ, the alarm duty and the warned fraction are

$$\begin{aligned} D(\theta) &= \frac{\sum_t \mathbf{1}[\hat{p}_t \ge \theta]\,(1 - y_t)}{\sum_t (1 - y_t)}, \\ W(\theta) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \max_{t \in \mathscr{W}_i} \mathbf{1}[\hat{p}_t \ge \theta], \end{aligned} \tag{18}$$

where $\mathscr{W}_i$ is the set of seconds between 10 s and 2 s before fallback i and n_e the number of fallbacks. A fallback whose episode is too short to contain such a second counts as a miss.

One further input is tested separately because it needs earlier drives over the same roads. With $\mathscr{F}$ the fallback positions and $\mathscr{N}$ the NR positions of the training drives other than the one being scored, the location prior at position $\mathbf{g}$ is

$$\pi(\mathbf{g}) = \frac{|\{f \in \mathscr{F} : \|f - \mathbf{g}\| \le r\}| + 0.5}{|\{u \in \mathscr{N} : \|u - \mathbf{g}\| \le r\}| + 50}, \qquad r = 150 \text{ m}, \tag{19}$$

with great-circle distances and additive smoothing.

3.6 Throughput and availability

The cost of a fallback is read from the download traces, where the application asks for as much as the link gives. Streaming traces are no use for this. A video player idles between segments, and its rate says little about the link. For every mode change with at least 30 rows on either side, the downlink rates $b_{i,\tau}$ at offsets $\tau = -30, \ldots, 30$ are aligned and summarised by

$$\bar{b}(\tau) = \frac{1}{n} \sum_{i=1}^{n} b_{i,\tau} \tag{20}$$

and by the median and quartiles across events. Finally, if NR and outage episodes alternate as a renewal process, the long-run share of time on NR is [42]

$$A = \frac{\mathbb{E}[T_{\mathrm{on}}]}{\mathbb{E}[T_{\mathrm{on}}] + \mathbb{E}[T_{\mathrm{off}}]}, \tag{21}$$

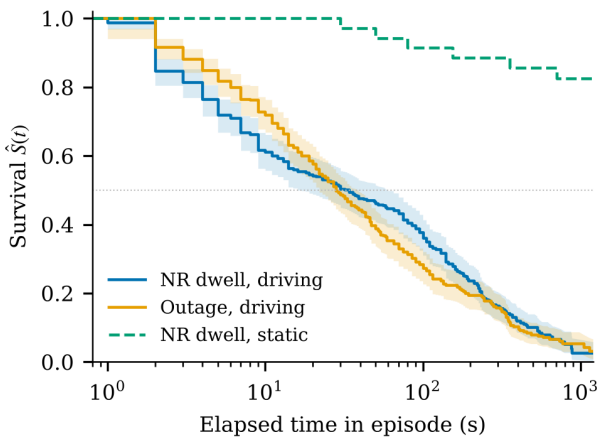


Figure 2. Kaplan–Meier survival of NR dwell time and of outage time on the road, with 95% bands, and of NR dwell time for the static traces.

which we evaluate with restricted means from the two Kaplan–Meier curves and compare with the share observed.

4. RESULTS

4.1 How long New Radio lasts

Table 2 and Figure 2 give the answer to the first question. On the road the median NR dwell is 34 s, with a trace-bootstrap interval from 14 s to 77 s. A quarter of the episodes are over within 5 s. By the tenth second 39% have ended, and only 16% last five minutes. The restricted mean over ten minutes is 131 s. Counting the episodes that were already open when logging began raises the median to 64 s, as it must, since those episodes are long by selection; the lower tail hardly changes, with 65% past ten seconds against 61%.

Outages are of the same order. Their median is 28 s (22 to 42), a quarter end within 9 s, and 16% are still running after five minutes. How deep the phone falls decides how long it stays down: the 303 outages spent entirely on LTE have a median of 22 s, whereas the 42 that touch 3G have a median of 349 s (log-rank $p < 0.001$) and, at 12% of the outages, hold 45% of all outage time. A phone on the road therefore alternates between the two states on a timescale of half a minute each, with heavy tails on both sides.

The static traces are a different regime. Only 12 of 35 NR episodes end in a fallback at all, the estimated median is 5,817 s, and 86% of episodes are intact after ten minutes. With so few events the static figures are indicative and we draw nothing finer from them. The contrast is plain enough. Mobility, far more than load or interference at a fixed spot, is what ends NR service on this network.

4.2 The shape of the hazard

No single rate describes these durations. Table 3 ranks the five families: the exponential is worst by a wide margin for both episode types, and for NR dwell the two-exponential mixture is best, 41 AIC units ahead of the log-normal and 80 ahead of the Weibull. Its estimates are $\hat{p} = 0.44$, $\hat{\lambda}_1 = 6.5$ s and $\hat{\lambda}_2 = 264$ s. On this reading 44% of NR attachments on the road belong to a population whose mean life is six and a half seconds, and the rest to one that lasts over four minutes on average.

Table 3. Parametric fits to driving episodes, ranked by AIC for NR dwell (lower is better)

	Estimates, NR dwell	AIC	
Family		NR	Outage
Mixture	$p = 0.44$, $\lambda_1 = 6.5$, $\lambda_2 = 264$	3,282.6	3,446.0
Log-normal	$\mu = 3.50$, $\sigma = 2.11$	3,323.7	3,438.7
Log-logistic	$\alpha = 31.8$, $\beta = 0.77$	3,349.0	3,453.3
Weibull	$\lambda = 89.1$, $k = 0.53$	3,362.3	3,502.8
Exponential	$\lambda = 140.2$	3,627.4	3,738.4

Figure 3 shows what this means for the hazard. Counted directly from person-seconds, the chance of a fallback in the next second is 0.079 during the first two seconds of an episode, 0.039 between the fifth and the tenth, 0.012 between the tenth and the twentieth, and settles between 0.002 and 0.005 from 40 s on. The mixture hazard of Eq. (9) follows these points from 0.070 at the origin to its floor of $1/\hat{\lambda}_2 = 0.0038$ per second, a ratio of about 18; the Weibull, with $\hat{k} = 0.53$, also declines but as a straight line on these axes, and it misses both the early plateau and the late floor.

Table 4 evaluates Eq. (10) at the fitted values. A connection that has just been made has a 56% chance of being of the settled kind, an expected life of 151 s and a 37% chance of dropping within ten seconds. Ten seconds later, if it is still there, the odds have changed: 85% settled, 226 s of expected life remaining and a 15% chance of dropping in the next ten. By 30 s the question is closed, with 99% settled. The longer a connection has lasted, the longer it will last. This is the reverse of wear, and it is what a memoryless model cannot express.

Table 4. Implications of the fitted mixture for a connection that has already spent t seconds on NR

t (s)	$w(t)$	$h(t)$ (s^{-1})	Drop within 10 s	$m(t)$ (s)
0	0.56	0.070	0.37	151
5	0.73	0.044	0.24	195
10	0.85	0.026	0.15	226
20	0.96	0.009	0.07	254
30	0.99	0.005	0.04	261
60	1.00	0.004	0.04	264

For outages the ranking changes at the top. The log-normal is first and the mixture 7 units behind, with a short component of 22 s for 65% of outages and a long one of 373 s. Recovery, then, has no sharp split into two kinds; it has a single broad scale.

A falling hazard has two possible sources, which survival data alone cannot fully separate: each connection may become safer as it ages, or connections may differ from the start and the fragile ones leave early. The mixture is the second story in its purest form. Section 4.3 gives some evidence for it, since conditions at entry do predict dwell time, but the experiment cannot exclude a contribution from the first.

Table 2. Kaplan–Meier summaries. Brackets are 95% percentile intervals from 500 trace-level bootstrap replicates; Q_1 and Q_3 are the quartiles and $\mu(600)$ the restricted mean over the first 600 s. For the static row the three survival columns are evaluated at 60, 300 and 600 s

Episode type	n	Events	Q_1 (s)	Median (s)	Q_3 (s)	$\hat{S}(10)$	$\hat{S}(60)$	$\hat{S}(300)$	$\mu(600)$ (s)
NR dwell, driving	335	305	5	34 [14, 77]	183	0.61 [0.55, 0.69]	0.45 [0.38, 0.54]	0.16 [0.11, 0.24]	131 [106, 168]
NR dwell, driving, all	377	345	6	64	259	0.65	0.50	0.23	157
Outage, driving	345	319	9	28 [22, 42]	112	0.72 [0.67, 0.77]	0.35 [0.30, 0.41]	0.16 [0.12, 0.21]	114 [94, 141]
NR dwell, static	35	12	1,883	5,817	–	0.94	0.89	0.86	533

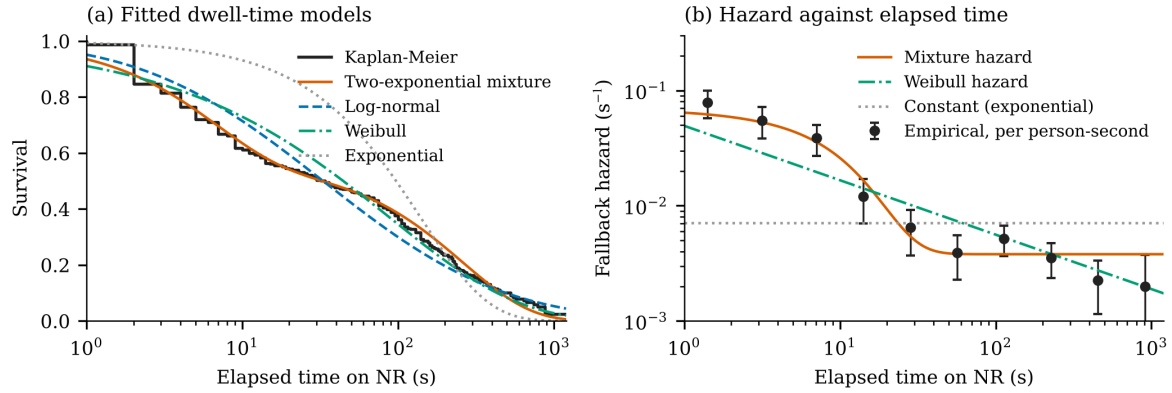


Figure 3. Fitted survival functions against the Kaplan–Meier estimate for NR dwell on the road (a), and the fallback hazard against elapsed time, with empirical rates per person-second and their 95% intervals (b).

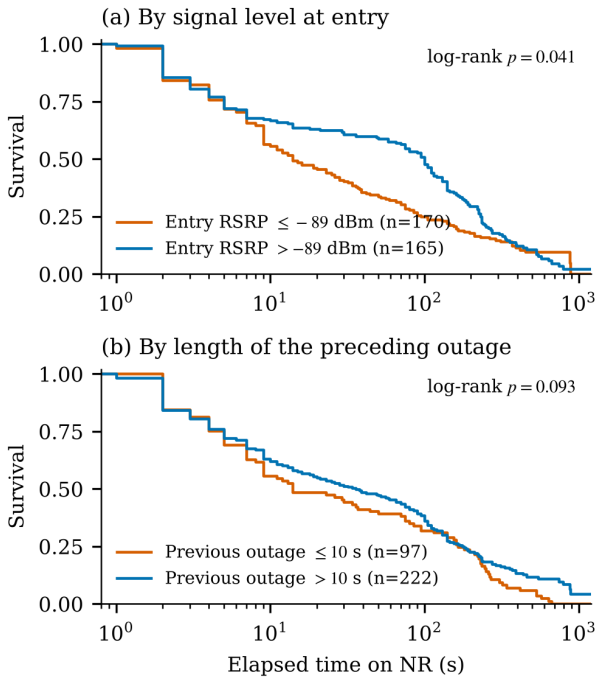


Figure 4. NR dwell on the road stratified by RSRP in the first second (a) and by the length of the outage that preceded the episode (b).

4.3 What moves the hazard

Stratified curves come first (Figure 4). Episodes that begin with an RSRP of −89 dBm or less, the sample median, have a median dwell of 14 s; those that begin above it have 96 s (log-rank $\chi^2 = 4.17$, $p = 0.041$), and episodes that follow an outage of 10 s or less are also shorter, 14 s against 36 s, though at $p = 0.093$ this split alone is not conclusive. Speed at entry, split at its median of 38 km/h, separates nothing ($p = 0.77$): the curves cross.

The Cox model of Table 5 sharpens the picture by adjusting the entry conditions for each other. A standard deviation of entry SNR (6.9 dB) lowers the hazard by 20% and one of RSRP (12.7 dB) by 15%. RSRQ adds nothing once these are in. The length of the preceding outage matters as much as the signal: each standard deviation on the log scale lowers the hazard by 17%, so an attachment made after a long absence is safer than one made seconds after the last fallback. Successive NR dwell times in a trace are positively correlated as well (Spearman 0.21, $p < 0.001$, 284 pairs). Instability comes in runs. Download traces carry a hazard 50% above streaming traces, an effect we did not anticipate and return to in Sections 4.9 and 5.3. The concordance is 0.64, which is modest: what is known in the first second orders episodes better than chance and not much better.

Table 5. Cox model of NR dwell on conditions at entry (335 episodes, 305 events). Hazard ratios per standard deviation, trace-clustered

Covariate (SD)	Ratio	95% interval	p
Speed (15.6 km/h)	1.07	[0.95, 1.21]	0.248
RSRP (12.7 dB)	0.85	[0.74, 0.97]	0.014
RSRQ (2.9 dB)	0.96	[0.83, 1.11]	0.592
SNR (6.9 dB)	0.80	[0.70, 0.91]	0.001
log prev. outage (1.67)	0.83	[0.75, 0.92]	< 0.001
Download trace	1.50	[1.11, 2.03]	0.008

Current conditions say more than entry conditions. Figure 5 plots the hazard ratios of Eq. (13), fitted to all 58,255 person-seconds on NR in the driving traces. CQI is the strongest radio covariate, a standard deviation (2.9 units) cutting the hazard by 44% (ratio 0.56, interval 0.49 to 0.64), followed by SNR (0.71) and RSRP (0.72). Speed now counts: 18 km/h more raises the hazard by 34% (1.34, 1.18 to 1.53), although speed in the first second did not. Elapsed time keeps a ratio of 0.62 per standard deviation of its logarithm after all of these

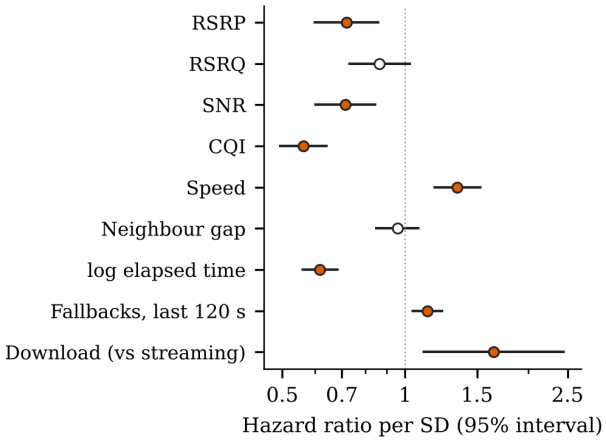


Figure 5. Hazard ratios from the discrete-time model with time-varying covariates. Filled markers: interval excludes one.

are in, which means that the declining hazard of Figure 3 is not explained away by the signal improving as an episode goes on. Each additional fallback in the last two minutes adds to the risk (1.13 per 0.68 fallbacks). Table 6 shows the speed effect without a model, as fallbacks per hour of NR time: 5 when the car is nearly at rest and 37 above 50 km/h. One null result deserves a line. The neighbour gap, the quantity on which textbook handover triggers are built [20], has a ratio of 0.96 with an interval that spans one. Whatever decides the release of the NR leg here, a stronger neighbour as reported in these logs is no sign of it.

Table 6. Fallback rate by current speed in the driving traces, with trace-bootstrap intervals

Speed (km/h)	NR seconds	Fallbacks	Per hour on NR
Below 10	8,352	12	5.2 [1.4, 13.3]
10 to 30	21,208	103	17.5 [13.3, 23.2]
30 to 50	21,570	157	26.2 [18.6, 34.3]
50 or more	7,125	73	36.9 [26.3, 51.1]

4.4 Early warning

Can the phone see a fallback coming? Table 7 and Figure 6 report the six scorers under trace-level cross-validation. At the ten-second horizon 4.6% of NR seconds are positive. The boosted model on all features reaches an average precision of 0.352 (interval 0.309 to 0.405) and an AUROC of 0.881; the same learner on radio features alone reaches 0.287, the logistic model on all features 0.269, and the RSRP threshold 0.203. Two paired comparisons are clear of zero under the trace bootstrap: all features over history alone in the logistic model, by 0.112 (0.083 to 0.150), and trees over the logistic model on the same inputs, by 0.083 (0.051 to 0.125). A third is not. Logistic regression on sixteen radio features is no better than a threshold on one of them (−0.004, interval −0.053 to 0.035), so the gain from the radio measurements comes only with a learner that can use their interactions.

Removing one feature family at a time from the boosted model tells the same story from the other side: without radio features its average precision falls from 0.352 to 0.163, without history to 0.306, and without context to 0.330. The radio family carries most of the signal (Figure 7). The ordering of the scorers is also stable across folds: the full boosted model is first in each of the five, with average precision between 0.304 and 0.447.

History behaves strangely, and the reason is instructive. On its own it scores 0.157, above the prevalence, yet almost all of that comes from knowing that young connections are fragile. Among seconds less than 30 s into an episode, where 13.2% are positive, the history-only model has an average precision of 0.322. Among the rest, where 3.3% are positive, it has 0.038, which is the prevalence. Elapsed time says who is at risk. It cannot say when.

Per-fallback counts in Table 8 translate the scores into alarms. If the boosted model on all features may keep its alarm raised during 10% of the seconds that are not followed by a fallback, it warns of 83% of the 345 fallbacks between two and ten seconds ahead; at a duty of 5% it warns of 73%. The history-only model looks respectable on all fallbacks (49% at 10%) and collapses to 6% on the 181 fallbacks that end episodes of 30 s or more, the ones a simple probation timer would not have covered anyway. For those, the radio features do the work: the boosted radio model warns of 74% at a 10% duty and 88% at 20%, about as many as the full model. Probabilities from the boosted model are also well calibrated out of fold over more than two orders of magnitude, the top decile predicting 0.298 against an observed 0.269 (Figure 7).

Table 8. Fallbacks warned 2 to 10 s ahead (%) at three alarm duties: all 345 fallbacks, and the 181 that end episodes of 30 s or more

	All			≥ 30 s		
Scorer	5%	10%	20%	5%	10%	20%
RSRP threshold	35.4	41.7	49.3	30.9	39.2	47.5
LR, history	43.8	48.7	55.4	0.6	6.1	17.1
LR, radio	48.7	63.5	77.1	44.8	59.7	74.6
LR, all features	61.4	75.1	87.2	42.5	59.1	78.5
GBT, radio	60.3	78.8	91.3	60.8	73.5	88.4
GBT, all features	72.8	83.2	92.8	56.9	71.3	89.0

Longer horizons are easier in absolute terms and harder relative to chance. Average precision of the full boosted model rises from 0.262 at 5 s to 0.412 at 20 s, while its ratio to prevalence falls from 10.3 to 5.1 and the AUROC from 0.901 to 0.847.

4.5 Transfer across applications

Grouped folds mix streaming and download drives. A harder test trains on one kind and scores the other (Table 9). From streaming to download nothing is lost: the full boosted model scores 0.363 on the download traces against 0.379 when they take part in cross-validation. The reverse direction fails for the full model, which drops from 0.339 to 0.222 on the streaming traces and ends below the radio-only model (0.267). The downlink rate is the likely culprit. It sits among the context features, its distribution under a saturating download is unlike that under a video player, and a model fitted on 16 download drives has learned thresholds on it that mean nothing elsewhere. Radio features travel better than traffic features.

Table 7. Early warning of a fallback within h seconds, out-of-fold over five trace-grouped folds (positive seconds: 2.5%, 4.6%, 8.0%). GBT: boosted trees; LR: logistic regression

	$h = 5$ s			$h = 10$ s			$h = 20$ s		
Scorer	AUROC	AP	Brier	AUROC	AP	Brier	AUROC	AP	Brier
RSRP threshold	0.663	0.134	–	0.662	0.203	–	0.643	0.261	–
LR, history	0.683	0.115	0.0236	0.652	0.157	0.0414	0.611	0.175	0.0711
LR, radio	0.805	0.142	0.0232	0.790	0.203	0.0399	0.754	0.280	0.0658
LR, all features	0.864	0.187	0.0224	0.845	0.269	0.0379	0.803	0.333	0.0631
GBT, radio	0.858	0.198	0.0224	0.851	0.287	0.0375	0.819	0.390	0.0602
GBT, all features	0.901	0.262	0.0212	0.881	0.352	0.0355	0.847	0.412	0.0589

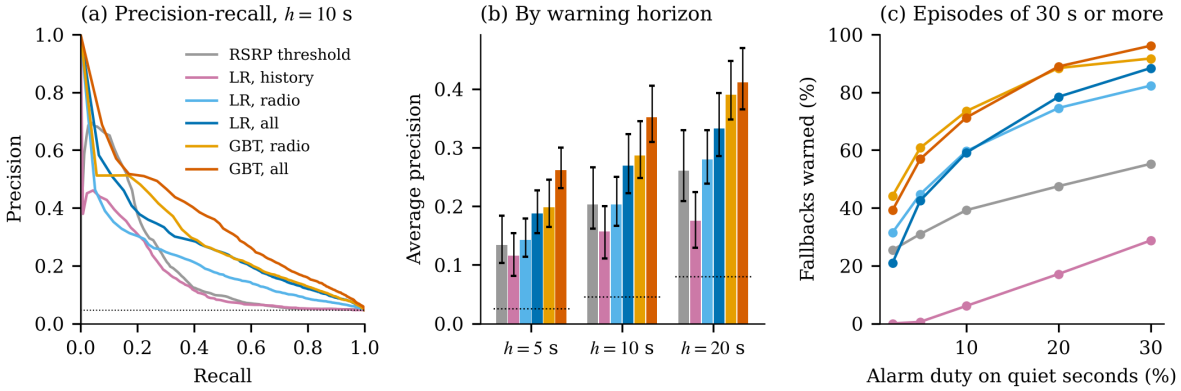


Figure 6. Precision-recall curves at the ten-second horizon (a); average precision by horizon with trace-bootstrap intervals, dotted lines marking prevalence (b); fallbacks warned against alarm duty for episodes of 30 s or more (c).

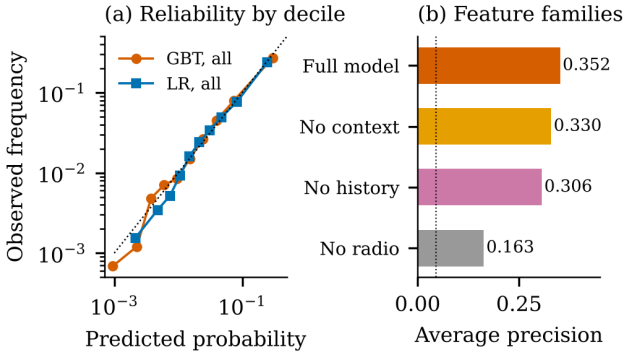


Figure 7. Out-of-fold reliability of two scorers at the ten-second horizon, by decile of predicted probability (a), and average precision of the boosted model with one feature family removed (b).

Table 9. Average precision at $h = 10$ s when training on one application type and testing on the other, against trace-grouped cross-validation on the same test traces

	Test: download		Test: streaming	
Scorer	Transfer	CV	Transfer	CV
GBT, radio	0.303	0.291	0.267	0.301
GBT, all features	0.363	0.379	0.222	0.339
RSRP threshold	0.233	–	0.204	–

4.6 Where fallbacks happen

The drives cover the same city repeatedly, which allows a question the radio features cannot answer: do fallbacks recur at the same places? They do (Figure 8). Of the 345 fallback positions, 72% lie within 50 m of a fallback recorded on a different drive, whereas only 32% of randomly drawn NR seconds do; at 100 m the figures are 79% and 48%. A location prior built from this, the smoothed ratio of other drives' fallbacks to other drives' NR seconds within 150 m, reaches an average precision of 0.273 at the ten-second horizon without a single radio measurement, about what the boosted radio model achieves (0.287). The two sources are complementary. Adding the prior to the full boosted model lifts its average precision from 0.352 to 0.409, a gain of 0.059 with a trace-bootstrap interval from 0.034 to 0.087. The interval excludes zero. In every fold the prior for a test drive was computed from training drives only, and for a training drive from the other training drives.

4.7 What a fallback costs

Table 10 gives downlink rates in the driving download traces by mode, and Figure 9 aligns the rate on the mode changes. The mean rate over the 25 s that end five seconds before a fallback is 27.5 Mbit/s and over the 25 s that begin five seconds after it 15.0 Mbit/s; the return is the mirror image, 15.6 Mbit/s before and 29.4 Mbit/s after. Means roughly halve and double.

Table 10. Downlink rate in the 16 driving download traces by serving mode (Mbit/s)

Mode	Seconds	Median	Mean	90th pct.	Below 1 (%)
NR	13,959	15.7	46.2	156.9	15.1
LTE	9,204	6.7	11.1	21.0	16.4
3G	4,428	2.4	3.3	7.3	28.5

Medians tell a quieter story: 9.1 Mbit/s before a fallback and 6.5 Mbit/s after, 7.1 and 9.0 Mbit/s around a return. Half of

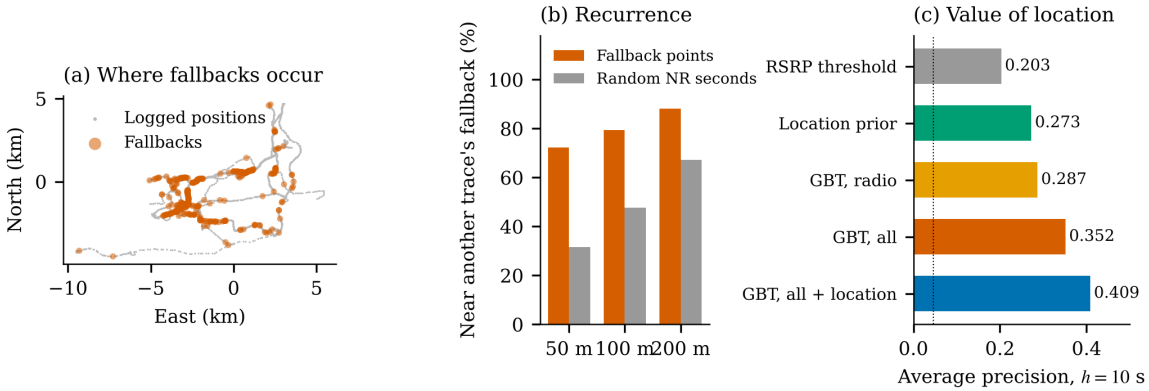


Figure 8. Positions of the driving traces and of the fallbacks, in kilometres from the centroid (a); share of points with a fallback from another drive within a given radius (b); average precision at the ten-second horizon with and without the location prior (c).

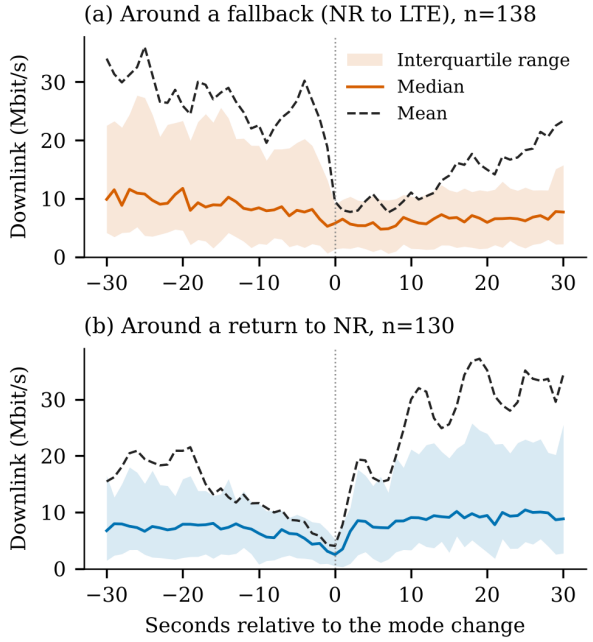


Figure 9. Downlink rate in driving download traces aligned on a fallback (a) and on a return to NR (b).

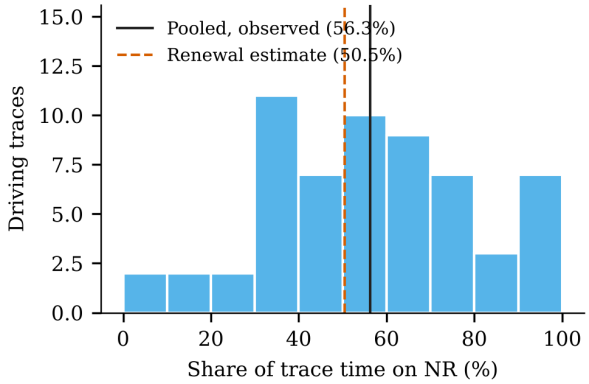


Figure 10. Share of time on NR per driving trace, with the pooled share and the renewal estimate from the main analysis.

all seconds near a mode change, on either side of it, carry less than 10 Mbit/s, and the share of seconds below 1 Mbit/s is the same on NR (15.1%) as on LTE (16.4%). What NR adds is an upper tail, with a 90th percentile of 157 Mbit/s against 21 Mbit/s on LTE. A fallback removes that tail. It does not take a connection from good to bad so much as from occasionally excellent to ordinary, and the seconds just before it are already weak, as the mean of 12.6 Mbit/s in the six seconds around the change shows.

4.8 Availability

Figure 10 closes the loop between durations and shares. Pooled over the driving traces the phone is on NR for 56.3% of the time, but individual traces range from none to all of it, with a standard deviation of 24 percentage points. Eq. (21) with the restricted means of the main analysis (158 s on, 155 s off) gives 50.5%. With the mean of all NR episodes (192 s) it gives 55.3%, one point from the observed share. The difference between the two estimates is the left-truncated episodes: leaving them out is right for the lower tail of the dwell distribution and biases the mean downward, and both facts should be kept in view when the numbers of Table 2 are reused.

4.9 Sensitivity to definitions

Several choices in Section 3.1 could have been made otherwise, and Table 11 repeats the dwell analysis under four alternatives. Treating an outage of 2 s or less as a flicker inside one NR episode changes nothing of substance: the median stays at 34 s and the mixture keeps 44% of episodes in a short component of 6.5 s. Extending the tolerance to 5 s removes 63 episodes and lifts the median to 58 s, yet the short component survives with 39% of the mass and a mean of 7.3 s, so the early hazard is no artefact of one-second flicker in the mode field. Dropping the four episodes shorter than 2 s moves the median to 36 s.

The split by application is another matter. NR episodes in download traces have a median of 10 s and those in streaming traces 73 s (log-rank $p < 0.001$), with 55% and 35% of episodes in the short component. The short component's mean life is the same in both, about six seconds. What differs between the applications is how often an attachment turns out to be of the fragile kind, and the settled ones also last less long under download (160 s against 311 s).

Table 11. NR dwell on the road under alternative definitions: Kaplan–Meier median and mixture estimates

Variant	n	Median (s)	$\hat{p}$	$\hat{\lambda}_1$ (s)	$\hat{\lambda}_2$ (s)
Main analysis	335	34	0.44	6.5	264
Outages ≤ 2 s merged	306	34	0.44	6.5	290
Outages ≤ 5 s merged	272	58	0.39	7.3	304
Episodes ≥ 2 s only	331	36	0.43	6.7	264
Download traces only	131	10	0.55	6.1	160
Streaming traces only	204	73	0.35	6.7	311

5. DISCUSSION

5.1 Using the results

Three uses follow directly. The first is a probation rule. An application that has just seen the mode change to NR should not plan on NR capacity yet, because for the next ten seconds the connection is ten to twenty times as likely to drop in any given second as it will be later; after 20 to 40 s the hazard has reached its floor and the connection can be treated as settled (Table 4). A video player can implement this with a timer and no model at all. The second is a warning. With the features a logging application already exposes, and a model small enough to run on the handset, four fallbacks in five can be announced a few seconds ahead at the price of an alarm that is raised needlessly a tenth of the time. That is enough to fetch a segment early or to hold a bitrate switch. Where a route is travelled often, a map of past fallbacks is worth as much again as the radio features, and a navigation application already knows the route. The third concerns simulation. Studies that draw NR on and off periods from exponential laws, as the classical holding-time literature would suggest [15, 17], understate the number of very short visits by a wide margin, and a two-population mixture with the parameters of Table 3 is a better default.

The null result on the neighbour gap has a practical side too. Predictors that are designed around the A3-style comparison of serving and neighbour power may be looking at the wrong quantity for this event, since the release of a secondary leg is governed by NR measurements and timers that an LTE-side neighbour report does not reflect [2, 19].

Algorithm 1 puts the first two uses together in the form an application could run once per second.

Algorithm 1. Probation and warning for an application on NR

1. **Input:** mode m_t, feature vector $\mathbf{u}_t$, probation time t_p (20 s here), threshold θ chosen for the alarm duty that can be afforded.
2. If $m_t = 0$: plan for LTE capacity; set $e \leftarrow 0$; stop.
3. $e \leftarrow e + 1$ (seconds on NR since attachment).
4. If $e < t_p$: plan for LTE capacity, since $w(e) < 0.96$; stop.
5. Compute $\hat{p}_t = F_M(\mathbf{u}_t)$ by Eq. (16).
6. If $\hat{p}_t \ge \theta$: raise the alarm, prefetch at the current quality and hold any switch upward.
7. Otherwise plan for NR capacity.

5.2 Relation to earlier measurements

Three points of contact with earlier work are worth stating. Kalntis et al. [8] found that handovers to 3G are rare and fail often, and the device-side view here supplies the consequence: an outage that reaches 3G lasts sixteen times as long at the median as one that stays on LTE. Hassan et al. [6] showed that handovers can be predicted from radio measurements, which agrees with the weight the radio family carries in our ablation, although their event is a handover between cells and ours the loss of the NR leg. The railway study of Chou et al. [22] reports a higher score than any in Table 7. The two are not comparable. Its F1 comes from 23 events on one line at a three-second horizon, ours from 345 events with whole drives held out, and we would expect any score to fall under that protocol.

5.3 Limitations

One handset, one operator and one city over three months of an early deployment are the whole sample, and the network has presumably been reconfigured since. The absolute numbers belong to that rollout. The mode field is what the logging application reported; the dataset does not document how the application decides between LTE and NR, nor whether the radio measurements in a row refer to the anchor or to the NR carrier, and this bears on the interpretation of every radio covariate. Sampling is at one second, so attachments shorter than that are invisible and the short component of the mixture is, if anything, underestimated. Elapsed time is undercounted for the 42 episodes already open at trace start, which enter the discrete-time regression and the warning experiment but not the dwell estimates. The higher hazard in download traces is confounded with route and day, because applications were not randomised over drives; whether heavy traffic itself provokes release of the NR leg is a question for a controlled test. The location prior, lastly, assumes drives over known roads and says nothing about a first visit. No hyperparameter was tuned, and no sequence model was tried. Both choices keep the comparison clean and leave the best attainable warning score unknown.

6. CONCLUSION

On the road, in this network, an NR connection is a short-lived thing with a long-lived minority. Half of the episodes end within 34 s and a quarter within 5 s, while a parked phone keeps NR for an hour and more. The durations follow a mixture in which 44% of attachments have a mean life of 6.5 s and the rest one of 264 s, so that the fallback hazard obeys

$$\frac{h(0)}{h(\infty)} = \frac{\hat{p}/\hat{\lambda}_1 + (1-\hat{p})/\hat{\lambda}_2}{1/\hat{\lambda}_2} \approx 18. \tag{22}$$

After ten seconds on NR the chance that a connection is of the settled kind has risen from 56% to 85%, and after thirty it is 99%. Signal quality, speed and recent fallbacks raise or lower the hazard by 13% to 44% per standard deviation. The neighbour gap leaves it where it was. Elapsed time identifies the fragile connections and radio measurements time the fallback of the settled ones, and a boosted model on both warns of 83% of fallbacks at a 10% alarm duty when tested on drives it has not seen; fallbacks also recur at the

same places, 72% of them within 50 m of one seen on another drive, so that a location prior raises average precision from 0.352 to 0.409. The cost is in the tail. Across a fallback the mean download rate halves and the median falls by less than a third.

Two things would be needed to carry these numbers beyond their source. The first is the same analysis on standalone deployments, where there is no anchor to fall back to and the event of interest becomes a change of band or of carrier aggregation state. The second is a log that records the release cause, without which the two populations of the mixture remain a statistical description. We have no physical account of them.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Sköld, 5G NR: The Next Generation Wireless Access Technology, 2nd ed. London, U.K.: Academic Press, 2020.

[2] 3GPP, "NR; Multi-connectivity; Overall description; Stage-2," 3rd Generation Partnership Project, TS 37.340, Release 15, 2019.

[3] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in Proc. ACM SIGCOMM, pp. 325–338, 2015.

[4] A. Narayanan et al., "A first look at commercial 5G performance on smartphones," in Proc. The Web Conference (WWW), pp. 894–905, 2020.

[5] D. Xu et al., "Understanding operational 5G: a first measurement study on its coverage, performance and energy consumption," in Proc. ACM SIGCOMM, pp. 479–494, 2020.

[6] A. Hassan et al., "Vivisecting mobility management in 5G cellular networks," in Proc. ACM SIGCOMM, pp. 86–100, 2022.

[7] K. Kousias et al., "Implications of handover events in commercial 5G non-standalone deployments in Rome," in Proc. ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases (5G-MeMU), 2022.

[8] M. Kalntis et al., "Through the telco lens: a countrywide empirical study of cellular handovers," in Proc. ACM Internet Measurement Conference (IMC), 2024.

[9] J. P. Klein and M. L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed. New York, NY, USA: Springer, 2003.

[10] D. Raca, D. Leahy, C. J. Sreenan, and J. J. Quinlan, "Beyond throughput, the next generation: a 5G dataset with channel and context metrics," in Proc. 11th ACM Multimedia Systems Conference (MMSys), pp. 303–308, 2020.

[11] A. Narayanan et al., "A variegated look at 5G in the wild: performance, power, and QoE implications," in Proc. ACM SIGCOMM, pp. 610–625, 2021.

[12] A. Narayanan et al., "Lumos5G: mapping and predicting commercial mmWave 5G throughput," in Proc. ACM Internet Measurement Conference (IMC), pp. 176–193, 2020.

[13] M. Ghoshal et al., "Performance of cellular networks on the wheels," in Proc. ACM Internet Measurement Conference (IMC), 2023.

[14] R. A. K. Fezeu et al., "Unveiling the 5G mid-band landscape: from network deployment to performance and application QoE," in Proc. ACM SIGCOMM, 2024.

[15] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," IEEE Transactions on Vehicular Technology, vol. 35, no. 3, pp. 77–92, 1986.

[16] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility patterns," IEEE Journal on Selected Areas in Communications, vol. 15, no. 7, pp. 1239–1252, 1997.

[17] Y. Fang and I. Chlamtac, "Teletraffic analysis and mobility modeling of PCS networks," IEEE Transactions on Communications, vol. 47, no. 7, pp. 1062–1072, 1999.

[18] X. Lin, R. K. Ganti, P. J. Fleming, and J. G. Andrews, "Towards understanding the fundamentals of mobility in cellular networks," IEEE Transactions on Wireless Communications, vol. 12, no. 4, pp. 1686–1698, 2013.

[19] 3GPP, "NR; Radio Resource Control (RRC); Protocol specification," 3rd Generation Partnership Project, TS 38.331, Release 15, 2019.

[20] M. Tayyab, X. Gelabert, and R. Jäntti, "A survey on handover management: from LTE to NR," IEEE Access, vol. 7, pp. 118907–118930, 2019.

[21] M. S. Mollel et al., "A survey of machine learning applications to handover management in 5G and beyond," IEEE Access, vol. 9, pp. 45770–45802, 2021.

[22] P.-H. Chou, D.-C. Lin, H.-Y. Wei, W. Saad, and Y. Tsao, "Measurement-driven early warning for resilient 5G NSA railway networks," arXiv preprint arXiv:2511.08851, 2025.

[23] M. Kabeer, R. Nordin, M. Behjati, and L. S. Lun, "Sequence-Based Deep Learning for Handover Optimization in Dense Urban Cellular Network," arXiv preprint arXiv:2510.02958, 2025.

[24] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," Journal of the American Statistical Association, vol. 53, no. 282, pp. 457–481, 1958.

[25] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," Cancer Chemotherapy Reports, vol. 50, no. 3, pp. 163–170, 1966.

[26] P. Royston and M. K. B. Parmar, "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome," BMC Medical Research Methodology, vol. 13, Art. no. 152, 2013.

[27] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

[28] C. Davidson-Pilon, "lifelines: survival analysis in Python," Journal of Open Source Software, vol. 4, no. 40, Art. no. 1317, 2019.

[29] G. J. McLachlan and D. Peel, Finite Mixture Models. New York, NY, USA: Wiley, 2000.

[30] H. Akaike, "A new look at the statistical model identifi-

cation," IEEE Transactions on Automatic Control, vol. 19, no. 6, pp. 716–723, 1974.

[31] D. R. Cox, "Regression models and life-tables," Journal of the Royal Statistical Society: Series B, vol. 34, no. 2, pp. 187–220, 1972.

[32] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," JAMA, vol. 247, no. 18, pp. 2543–2546, 1982.

[33] R. L. Prentice and L. A. Gloeckler, "Regression analysis of grouped survival data with application to breast cancer data," Biometrics, vol. 34, no. 1, pp. 57–67, 1978.

[34] J. D. Singer and J. B. Willett, "It's about time: using discrete-time survival analysis to study duration and the timing of events," Journal of Educational Statistics, vol. 18, no. 2, pp. 155–195, 1993.

[35] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," Biometrika, vol. 73, no. 1, pp. 13–22, 1986.

[36] S. Seabold and J. Perktold, "Statsmodels: econometric and statistical modeling with Python," in Proc. 9th Python in Science Conference, pp. 92–96, 2010.

[37] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[38] F. Pedregosa et al., "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[39] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017.

[40] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, Art. no. e0118432, 2015.

[41] G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, no. 1, pp. 1–3, 1950.

[42] S. M. Ross, Stochastic Processes, 2nd ed. New York, NY, USA: Wiley, 1996.