International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 3, No. 3, 2026

Improving Cross-Regional Generalizability of Bridge Deck Deterioration Prediction through Transfer Learning and Explainable Discrete-Time Survival Models

Irene Mitchell, Lucas Lam *

Department of Computer Science, College of Engineering, University of California, Davis, CA, USA

* **Corresponding author**: (Email: lam201802@ucdavis.edu)

Abstract: Data-driven deterioration models for bridge decks are trained and validated on pooled national data but are used by individual agencies whose climate, materials and inspection practice differ. On the nationwide National Bridge Inventory panel of 150,144 concrete decks, the published discrete-time gradient-boosted hazard model (DT-GBH) reaches a C-index of 0.769 on a random split but only 0.647 on states withheld from training, no better than Cox regression, while its 5-year calibration error grows from 0.009 to 0.050. We show that this failure combines a shift in the level and shape of the hazard with a shift in covariate effects, and that covariate-shift reweighting and state-blocked model selection do not repair it. We propose TL-DTGBH, which adapts a national DT-GBH to a target state from a short local history in two stages on the hazard-logit scale: a five-parameter, ridge-shrunk recalibration of level, slope and period-specific baseline (HLR), followed by residual boosting whose size is chosen by spell-grouped cross-validation and may be zero. The adapted logit splits exactly into transferred national evidence, a local baseline correction and local covariate corrections, each attributed to features by TreeSHAP. In leave-one-state-out experiments on 14 states, HLR halves the calibration error, from 0.103 to 0.054 with only 50 local sojourns, and the full method raises the C-index from 0.666 to 0.684 with 100 and 0.732 with 500 local sojourns, outperforming local-only and Cox-updating models when local data are scarce and matching a pooled model that must be retrained on national data for every state. Under the original unseen-state protocol, a 20% local history lifts the pooled C-index from 0.646 to 0.757.

Keywords: Bridge management, deterioration modelling, transfer learning, discrete-time survival analysis, gradient boosting, SHAP, National Bridge Inventory, domain shift.

1. Introduction

Bridge decks deteriorate faster than any other bridge component and absorb a large share of preservation budgets [1]. Under the U.S. National Bridge Inventory (NBI) every deck receives an integer condition rating (CR) from 9 to 0 at each inspection [2], and the quantity that matters for preservation planning is not the current rating but how long a deck will stay in it. Time-in-rating models have therefore moved from Markov chains and parametric sojourn models [3–5] to Bayesian survival analysis at national scale [6] and to survival machine learning [7].

Almost all of this work rests on an assumption that is rarely tested: that a model estimated on one population of bridges is valid for another. The question arises constantly in practice. A national model is offered to a state agency; a state model is reused by a county or toll authority with a short inspection history; a jurisdiction changes its rating guidance and its old model ages. States differ in climate, de-icing practice, deck protection, construction standards and in how inspectors assign ratings, so the conditional distribution of sojourn times given NBI attributes need not be shared [8]. Zhang and Qiu [9] made this concrete. Their discrete-time gradient-boosted hazard model (DT-GBH), which learns the annual downgrade probability of a deck from person-period records with a leakage-controlled cohort design, reached a C-index of 0.766 on a random split of the national panel; when a quarter of the states was withheld entirely, every model fell to about 0.66 and the advantage of the boosted hazard over a Cox model disappeared.

That result raises two questions which this paper answers. What exactly breaks when a deterioration model crosses a state line: the ranking of decks within the state, the level and shape of the hazard, or both? And can the gap be closed with the data an agency actually holds, namely a short local history of a few dozen to a few hundred observed sojourns, without discarding the national model? Transfer learning offers a vocabulary for both [10], but its standard tools fit the problem poorly. Importance weighting for covariate shift [11] assumes that the conditional model is shared, which is what fails here. Boosting-based transfer such as TrAdaBoost [12], applied to pavement roughness in [13], reweights instances of a regression problem rather than adapting a censored-data model. Transfer learning for survival analysis has mostly been developed for Cox regression [14]. Model updating in clinical prediction, which re-estimates an intercept and a calibration slope on new data [15, 16], is cheap and robust, but it cannot learn interactions that are specific to the new population.

We propose TL-DTGBH, a two-stage adaptation of a national DT-GBH model that works entirely on the hazard-logit scale. The first stage, hazard-logit recalibration (HLR), fits a ridge-shrunk intercept, slope and period-specific offsets on the local person-period records, which corrects the level and the shape of the baseline hazard with five parameters. The second stage, residual boosting (RB), grows a small tree ensemble that starts from the recalibrated national logit; its size is chosen by spell-grouped cross-validation and is allowed to be zero, so the method falls back to recalibration when the local sample carries no further signal. Because both

stages act additively on the logit, every adapted prediction splits exactly into transferred national evidence, a local baseline correction and local covariate corrections, each of which can be attributed to features with TreeSHAP [17].

The contributions are fourfold. (i) A diagnosis of cross-state failure on the national NBI panel of [6]: we reproduce the DT-GBH benchmark, separate within-state discrimination from calibration, and show that covariate-shift reweighting and state-blocked model selection do not repair the gap. (ii) TL-DTGBH, a transfer method for discrete-time survival ensembles that combines shrinkage recalibration with cross-validated residual boosting. (iii) An exact three-part explanation of adapted predictions, which identifies the deterioration drivers that transfer across states and those that must be relearned locally, including their dependence on time already spent in the rating. (iv) A leave-one-state-out evaluation on 14 states with 50 to 500 local sojourns, a replication of the unseen-state protocol with local adaptation, and a decision-level analysis with a preservation prioritization index.

2. Related Work

2.1. Deterioration Models Estimated from Inspection Records

Markov chains with transition probabilities estimated from condition ratings remain the backbone of bridge management systems [3], although their memoryless property conflicts with hazards that rise with time in state. Semi-Markov and Weibull sojourn models relax this assumption [4], and time-based accelerated-failure-time models quantify the effect of protection, continuity and overlays on deck deterioration [5]. Fleischhacker et al. [6] fitted a Bayesian Weibull time-in-rating model to more than 150,000 decks observed over 23 years and released the panel publicly [18]. Machine learning entered the field through classification of the rating itself [19] and, more recently, through survival learners such as random survival forests [7] and neural survival models, whose accuracy was shown to depend on the heterogeneity of the bridge population [8]. Discrete-time hazard models [20, 21] match the annual NBI reporting cycle because a downgrade is only known to have occurred within a one-year interval; gradient boosting [22] turns the resulting binary likelihood into a flexible, non-proportional hazard learner. Almost all of these models are fitted and validated within one population, and geographic validity is rarely examined.

2.2. Transfer Learning and Model Updating

Transfer learning is usually organised into instance-, feature- and parameter-based approaches [10]. Instance-based methods reweight source data, either by density ratios under covariate shift [11] or iteratively by boosting [12]; the latter has been used to build pavement performance models for road agencies with little data [13]. For survival outcomes, Transfer-Cox couples source and target Cox models through an L2,1 penalty [14]. A separate and highly practical literature on updating clinical prediction models distinguishes recalibration in the large, logistic recalibration of the linear predictor and full model revision, and shows that shrinkage is essential when the new sample is small [15]; the same ideas were developed for prognostic survival models [16]. TL-DTGBH borrows the updating logic and transposes it to tree ensembles in discrete time: recalibration acts on the conditional hazard of every period, and revision is carried out by boosting from the recalibrated model, with the amount of revision chosen by cross-validation instead of being fixed in advance.

2.3. Explaining Tree-Based Hazard Models

SHAP values provide additive feature attributions with uniqueness guarantees [23], and TreeSHAP computes them exactly in polynomial time for tree ensembles [17]. Because attributions are linear in the model output, a model built as a sum of components can be explained component by component. We use this property to make the adaptation itself interpretable rather than only the final prediction.

3. Data and Cross-Regional Setting

3.1. NBI Deck Panel and Sojourn Cohort

We use the nationwide concrete-deck panel released with [6] through its public repository [18]. For 150,144 concrete highway bridge decks it records the deck CR reported in every year from 1992 to 2014 [2], time-invariant inventory attributes (year built and reconstructed, deck structure type, structural material, design type, wearing surface, membrane, deck protection, deck area, lanes, designated inspection interval, average daily truck traffic (ADTT), distance from seawater, functional class, maintenance responsibility and a nine-level climatic region) and an anonymised state code, reported here as S01–S52. All condition data are observed records.

Sojourn spells are built with the leakage-controlled procedure of the DT-GBH framework [9], which we adopt together with its learner and hyper-parameters, its time-resolved attribution profile and its preservation prioritization index. Reusing these components unchanged makes our cross-regional results directly comparable with the national benchmark and lets the transfer stages be judged against a strong, well-specified source model. A spell in state s starts in the first year $t_0 \in [1997, 2009]$ with $c_i(t_0) = s$ and $c_i(t_0 - 1) > s$, so that entry is observed and five complete years of prior ratings exist. The sojourn Ti is the number of years to the first downgrade; a spell is right-censored when the panel ends, when the rating becomes unavailable, or when it increases, which signals rehabilitation. One spell per deck is kept. Seven history features (rating before entry, rating five years before entry, five-year drop, numbers of downgrades and upgrades, mean rating and consecutive years above s in the window) are computed inside the same five-year window, and calendar-position variables are excluded. We add one safeguard: a reconstruction is counted only if its year precedes t_0, because the inventory stores the latest reconstruction year and a later value would leak post-entry information. Categorical attributes are one-hot encoded with levels below 1% pooled, which gives 59 design columns.

Table I. Sojourn cohorts rebuilt from the national panel

Entry state	Spells (ours / reported)	Downgrades (ours / reported)	States
CR 8	11,109 / 11,109	6,452 / 6,452	41
CR 7	48,854 / 48,901	16,165 / 16,182	47
CR 6	52,699 / 53,055	14,326 / 14,394	47

Reported: counts published for the DT-GBH framework on the same panel.

Table I shows that the construction reproduces the published cohorts: the CR 8 cohort matches exactly and the

CR 7 and CR 6 cohorts agree to within 0.7%, the remaining difference being consistent with record-level filtering details (for example, records with missing or inconsistent year built). The primary cohort, entry into CR 7 ("good"), contains 48,854 spells from 47 states, 33.1% of which end in an observed downgrade; the median observed sojourn is 7 years and the median deck age at entry 30 years.

3.2. Cross-Regional Setting

We treat each state as a domain. For a target state t, the source data contain all spells of the other 46 states, and the target supplies a local labelled sample of n spells, its inspection history, together with a disjoint test set. The 14 states with at least 1,000 CR 7 spells serve as targets; together they hold 36,187 spells (74% of the cohort), with downgrade proportions between 12.4% and 61.6%. Fig. 1(a) illustrates how different these domains are: Kaplan–Meier curves [24] of time in CR 7 for the six largest states separate from the second year onwards, and the probability of still being rated 7 after ten years ranges from 0.45 to 0.76. Fig. 1(b) shows the consequence for a national model. Averaged over the test decks of each state, the zero-shot DT-GBH misses the observed 5-year downgrade probability by up to 0.39 (absolute) and by 0.103 on average, whereas the adapted model introduced below reduces the mean deviation to 0.027 with 200 local spells.

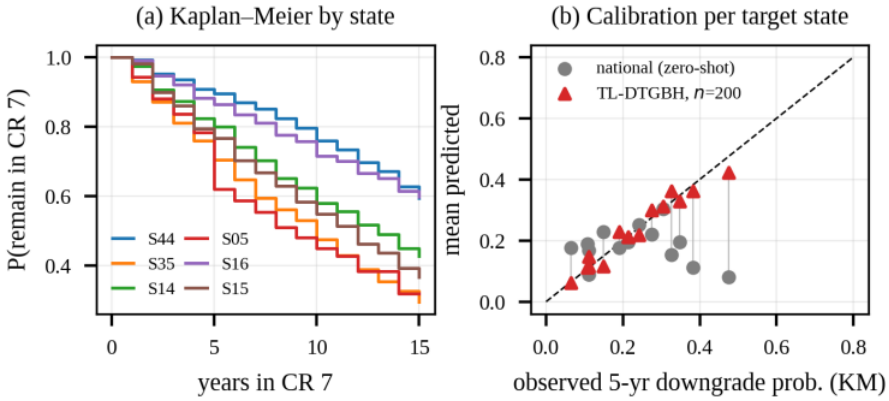


Figure 1. Cross-state heterogeneity. (a) Kaplan–Meier probability of remaining in CR 7 for the six largest states. (b) Mean predicted versus observed 5-year downgrade probability in each of the 14 target states, for the national DT-GBH (zero-shot) and for TL-DTGBH adapted with n = 200 local spells (first draw).

4. Method

4.1. Discrete-Time Hazard and the Source Model

NBI ratings are reported once a year, so a downgrade first observed in year $t_0 + k$ is only known to have occurred in $(t_0 + k - 1, t_0 + k]$. Following [20], [21], the hazard is modelled in discrete time,

$$h(k|x) = \Pr(T = k \mid T \geq k, x), \qquad S(t|x) = \Pi_{k=1}^{t} [1 - h(k|x)] \quad (1)$$

Expanding spell i into T_i person-period records with label $y_{ik} = 1$ only at $k = T_i$ for an observed downgrade turns the censored likelihood into a Bernoulli log-likelihood,

$$\ell = \Sigma_i \Sigma_{k \leq T_i} [y_{ik} \log h(k|x_i) + (1 - y_{ik}) \log(1 - h(k|x_i))] \quad (2)$$

The source model is DT-GBH, a LightGBM ensemble [22] on the logit scale in which the elapsed period k is an ordinary input,

$$h_S(k|x) = \sigma(F_S(x, k)), \qquad F_S(x, k) = \Sigma_m \eta\, f_m(x, k) \quad (3)$$

So the baseline hazard is non-parametric and covariate effects may change with k. We keep the published configuration (learning rate 0.05, 63 leaves, at least 200 records per leaf, row and feature subsampling 0.8, L2 penalty 1, early stopping on 15% of the source decks). Decks are ranked by the negative restricted mean sojourn $r(x) = -\Sigma_{t \leq 15} S(t|x)$. Fig. 2 summarises how the source model is adapted to a target state.

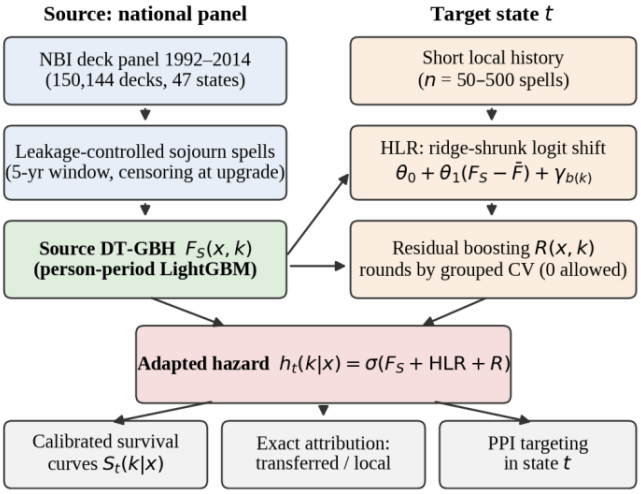


Figure 2. TL-DTGBH. A national DT-GBH is trained on the source states; a short local history of the target state drives a five-parameter hazard-logit recalibration (HLR) and a cross-validated residual ensemble (RB). The adapted hazard yields calibrated survival curves, an exact split of every prediction into transferred and local evidence, and inputs for preservation targeting.

4.2. Stage 1: Hazard-Logit Recalibration

Let $\bar{F}S$ be the mean source logit over the local person-periods and b(k) a coarse period bin (1–2, 3–5, 6–9 and at least 10 years). HLR keeps the national logit as an offset and adds five parameters,

$$F_H(x, k) = F_S + \theta_0 + \theta_1(F_S - \bar{F}_S) + \gamma_{b(k)}, \qquad \gamma_1 = 0 \quad (4)$$

Estimated on the local records by

$$\hat{\theta} = \arg\min_\theta \{-\ell_t(\theta) + (\lambda/2)\|\theta\|^2\} \quad (5)$$

Where ℓ_t is the log-likelihood (2) of the local person-

periods. The penalty shrinks the adapted model toward the national one, so $\theta = 0$ reproduces the zero-shot model; $\lambda \in \{0.3, 1, 3, 10, 30, 100\}$ is chosen by five-fold cross-validated log-likelihood with folds formed by spells, never by person-periods. θ_0 is the familiar recalibration in the large and θ_1 a calibration slope that expands or compresses the national risk contrast [15]; the period offsets γ correct the shape of the baseline hazard, which a proportional-hazards update [16] cannot do. The problem is a five-parameter logistic regression with an offset and is solved by Newton–Raphson in milliseconds.

4.3. Stage 2: Residual Boosting

HLR leaves the national ordering of decks almost unchanged. RB lets the local data revise it,

$$h_t(k|x) = \sigma(F_H + R), \quad R(x, k) = \Sigma_{m \le M} \nu\, g_m(x, k) \qquad (6)$$

Where the trees g_m are fitted by LightGBM to the local person-periods with F_H supplied as the initial score. The trees are deliberately small (7 leaves, at least 20 records per leaf, L2 penalty 10, $\nu = 0.05$, subsampling 0.8). The number of rounds $M \in \{0, \ldots, 400\}$ is selected by five-fold spell-grouped cross-validation in which HLR is refitted inside every fold, so that each candidate M is scored on the complete two-stage procedure; $M = 0$ recovers HLR. Unlike instance reweighting [12], RB never revisits the national data: the national model enters only through F_H. Adaptation is therefore cheap enough to run for every target state, and the local trees are confined to the structure that the national model misses.

4.4. Exact Explanation of the Adapted Model

Let $\varphi^S_j(x, k)$ and $\varphi^R_j(x, k)$ be the TreeSHAP values of F_S and R [17], [23]. SHAP values are linear in the model output and the HLR offsets depend on k alone, so the adapted logit decomposes exactly as

$$F_t(x, k) = \varphi_0 + \Sigma_j [(1 + \theta_1)\, \varphi^S_j + \varphi^R_j] + \Delta_{b(k)} \qquad (7)$$

Where φ_0 collects the constants and $\Delta_{b(k)}$ is the centred period offset, attributed to the time-in-state input. The three terms are national evidence transferred to the state (rescaled by the local calibration slope), local covariate corrections and local baseline recalibration, all on the log-odds scale of the annual downgrade probability. We summarise them by mean absolute attributions over test person-periods and, as in the DT-GBH framework, by the time-resolved profile of the local part,

$$\Phi^R_j(k) = E_x\, |\varphi^R_j(x, k)| \qquad (8)$$

Which shows whether a state-specific correction acts early or late in the sojourn.

4.5. Competing Approaches

Zero-shot methods use no target outcomes: Kaplan–Meier [24] and ridge Cox [25] fitted on the source states; the source DT-GBH; DT-GBH with state-blocked early stopping, whose validation set consists of whole held-out source states, a model-selection device borrowed from domain generalization; and importance-weighted DT-GBH, in which source spells are weighted by the density ratio estimated with a logistic domain classifier on unlabelled target covariates [11], clipped to [0.1, 10]. Few-shot competitors use the same n local spells: a local Kaplan–Meier curve; local ridge Cox with the penalty chosen by cross-validation; Cox updating, which re-estimates the slope of the national linear predictor and a local Breslow baseline [16]; DT-GBH trained on the local spells only, with the small trees and cross-validated rounds of RB; and pooled DT-GBH, trained on the national data plus the local spells with a target-indicator column and the national number of rounds. HLR only and RB only (trees started from the unrecalibrated national logit) are ablations of TL-DTGBH.

5. Experimental Setup

Three protocols are used. (1) Reproduction: the random 70/15/15 split over decks, and the unseen-state protocol in which 25% of the states are withheld entirely, repeated for five random draws of states. (2) Leave-one-state-out (LOSO) transfer: for each of the 14 target states the source model is trained on all other states; the target spells are split once into a test half and an adaptation pool, from which $n \in \{50, 100, 200, 500\}$ spells are drawn three times (twice for the pooled model, which retrains the national ensemble every time). With a typical downgrade proportion of one third, $n = 50$ corresponds to about 15 observed downgrades. (3) Unseen states with local adaptation: the five draws of protocol (1) are repeated, but each withheld state contributes a random 20% of its spells as local history and is evaluated on the remaining 80%.

Discrimination is measured by Harrell's C [26] on r(x), by Uno's C [27] truncated at 10 years and by the time-dependent AUC at 5 years; probabilistic accuracy by the integrated Brier score (IBS) over 2–10 years [28]; calibration by $CITL_5$, the absolute difference between the mean predicted 5-year downgrade probability and its Kaplan–Meier estimate. Censoring weights are estimated within each evaluated target set because censoring patterns differ between states, and under protocol (3) we also report the within-state (stratified) C, which only compares decks of the same state. Methods are compared with two-sided Wilcoxon signed-rank tests on state-level means (14 pairs). The code uses LightGBM [22], whose exact TreeSHAP routine provides all attributions, scikit-survival [29] and XGBoost [30]; all runs used one CPU core and fixed seeds. A complete source-plus-adaptation run for one target state takes about three minutes, and adapting an already trained national model takes seconds.

6. Results

6.1. Reproducing the Benchmark and Locating the Gap

Table II. Reproduction of the national benchmark (CR 7)

Model	Random C	Random IBS	Unseen C	Unseen IBS	Unseen C_w
Cox PH, ours	0.683	0.163	0.639	0.183	0.632
Cox PH, reported	0.687	0.168	0.658	0.157	–
XGBoost-Cox, ours	0.762	0.140	0.651	0.184	0.653
XGBoost-Cox, reported	0.759	0.147	0.657	0.161	–
DT-GBH, ours	0.769	0.136	0.647	0.185	0.654
DT-GBH, reported	0.766	0.143	0.656	0.163	–

Unseen: 25% of states withheld; ours = mean of five draws of states, reported = one draw. C_w: within-state (stratified) C. Reported values are those published for the DT-GBH framework.

Table II compares our implementation with the values published for DT-GBH and its baselines [9]. On the random split the reproduced models agree with the published values to within 0.005 in C and 0.008 in IBS, and the ordering is the same: DT-GBH 0.769, XGBoost-Cox 0.762, Cox 0.683. The cross-state collapse is reproduced as well. Over five draws of withheld states the boosted hazard reaches C = 0.647 (standard deviation 0.041 across draws), essentially the level of Cox (0.639), and its IBS (0.185) is barely better than that of the covariate-free Kaplan–Meier curve (0.192).

Two statistics that the original protocol does not report locate the failure. The within-state concordance of DT-GBH, which compares only decks of the same state, drops from 0.733 on the random split to 0.654 on withheld states, so the national model does not merely lose the ordering between states; it also ranks decks inside a new state less well. At the same time its 5-year calibration error rises from 0.009 to 0.050. Cross-state failure is therefore a mixture of a shift in covariate effects and a shift in the level and shape of the hazard, and both require local outcome information.

6.2. Zero-Shot Remedies

Table III. Zero-shot transfer to the 14 target states

Model (no local outcomes)	C	Uno C	IBS	$CITL_5$
Kaplan–Meier	0.500	0.500	0.209	0.107
Cox PH	0.642	0.646	0.197	0.084
DT-GBH	0.666	0.672	0.202	0.103
DT-GBH, state-blocked ES	0.663	0.670	0.194	0.095
DT-GBH, importance-weighted	0.648	0.656	0.230	0.141

Means over target states (LOSO). ES: early stopping on whole held-out source states.

Table III gives the zero-shot results of the leave-one-state-out protocol. The national DT-GBH reaches C = 0.666 on average, with a calibration error of 0.103 and an IBS (0.202) close to that of the national Kaplan–Meier curve (0.209). Neither remedy that avoids local outcomes changes this. State-blocked early stopping selects far smaller ensembles (median 38 instead of 624 rounds) but leaves discrimination where it was (C = 0.663). Importance weighting is harmful (C = 0.648, $CITL_5$ = 0.141): moving the source sample toward the covariate profile of the target cannot correct a shift that lies in the conditional hazard, and the weights inflate the variance of the fit.

6.3. Few-Shot Transfer

Table IV. Few-shot transfer: mean over 14 target states

Method	n = 50	100	200	500
(a) C-index				
Cox, local only	0.629*	0.645*	0.670*	0.692*
Cox updating	0.634*	0.640*	0.642*	0.642*
DT-GBH, local only	0.639*	0.666*	0.706	0.729
DT-GBH, pooled + indicator	0.678	0.687	0.706	0.731
HLR only	0.666	0.665	0.667*	0.667*
RB only	0.673	0.683	0.709	0.730
TL-DTGBH (HLR+RB)	0.670	0.684	0.711	0.732
(b) Integrated Brier score				
Cox, local only	0.209*	0.187*	0.177*	0.171*
Cox updating	0.188	0.181	0.179*	0.179*
DT-GBH, local only	0.188*	0.180	0.167	0.159
DT-GBH, pooled + indicator	0.188	0.181	0.166	0.158
HLR only	0.183	0.180	0.176*	0.177*
RB only	0.186	0.180*	0.169*	0.161*
TL-DTGBH (HLR+RB)	0.184	0.177	0.166	0.158

Bold: best in column. * TL-DTGBH significantly better, † significantly worse (two-sided Wilcoxon signed-rank test over states, $p < 0.05$).

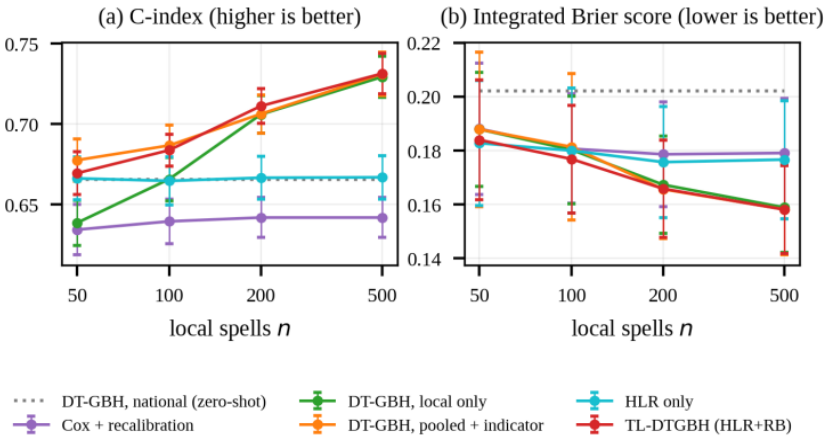


Figure 3. Leave-one-state-out transfer as a function of the number of local spells n (mean over 14 target states; bars: standard error across states). The dotted line is the national DT-GBH without adaptation.

Table IV and Fig. 3 show that the two stages of TL-DTGBH repair different parts of the error. HLR alone halves the calibration error with 50 local spells: $CITL_5$ falls from 0.103 to 0.054 and the IBS from 0.202 to 0.183, the lowest IBS of all methods at n = 50, while the C-index stays at the national level (0.666), as expected from a recalibration that is monotone in the national logit. The fitted slopes are negative in every state (θ_1 between −0.86 and −0.002 at n = 500): the national model is systematically over-confident outside its training states, and its log-odds contrasts must be shrunk by a state-specific factor $1 + \theta_1$ between 0.14 and 1.00. The level corrections θ_0 span −0.62 to 1.84 log-odds units.

Residual boosting recovers discrimination once the local sample holds a few dozen downgrades. TL-DTGBH raises C from 0.666 to 0.684 with 100 local spells and to 0.732 with 500, it beats the national model in 11 of 14 states at n = 100 (p = 0.049) and in 14 of 14 at n = 500. When local data are scarce it clearly outperforms the alternatives that ignore national information: at n = 100 it exceeds DT-GBH trained on the local spells alone by 0.018 in C (p = 0.042), Cox updating by 0.044 and local Cox by 0.039, and it has the lowest IBS of all methods. The pooled model, which retrains the national ensemble with the local spells and an indicator, is the strongest competitor: it is slightly ahead in C at n = 50 and 100 (by 0.008 and 0.003; p = 0.119 and = 0.153), behind at n = 200 and level at n = 500.

As the local history grows, the data-driven methods converge: at n = 500 TL-DTGBH, local-only DT-GBH and the pooled model differ by at most 0.002 in C and 0.001 in IBS, and TL-DTGBH has the lowest calibration error (0.023). Two practical conclusions follow. Below about 100 local spells (roughly 30 downgrades) the residual stage is not yet reliable: at n = 50, RB only and the pooled model reach a slightly higher C than the full pipeline, which is itself not significantly better than the national model, while HLR only has the lowest IBS; recalibration alone is the safer choice there. From about 100 spells onward the two-stage adaptation is statistically indistinguishable from the best alternative on every metric without touching the national data; adapting a trained national model to a state takes seconds, whereas the pooled model has to be retrained for every state and every update of its history.

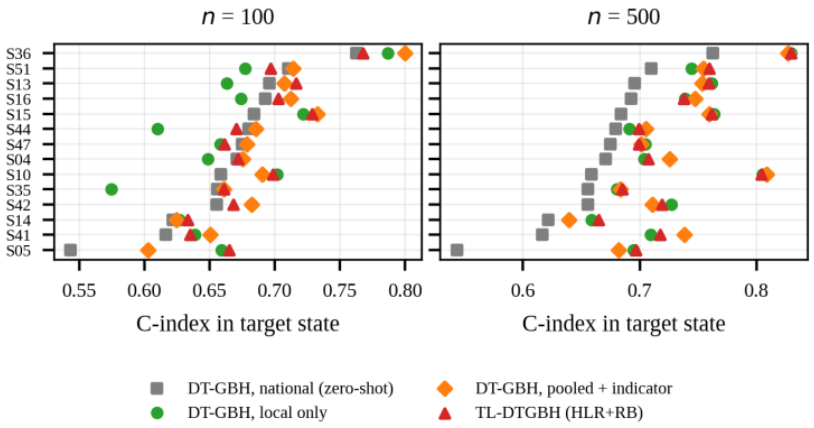


Figure 4. C-index in each target state (sorted by the national model) for the national DT-GBH and three ways of using n local spells; mean over draws.

Fig. 4 shows the same comparison state by state. Without adaptation, C ranges from 0.543 in state S05 to 0.763 in state S36, and the largest gain occurs where the national model fails most: in S05, TL-DTGBH reaches C = 0.665 with 100 and 0.697 with 500 local spells, a gain of 0.154. Across the 14 states, however, the gain at n = 500 is only weakly related to the national C (Spearman $\rho = -0.22$); states that the national model already ranks well also benefit from local relearning.

6.4. Unseen-State Protocol with Local History

Table V. Unseen-state protocol with 20% local history

Model	C	C_w	IBS	$CITL_5$
Cox PH, national	0.639	0.631	0.179	0.074
DT-GBH, national	0.646	0.654	0.181	0.096
Cox updating	0.688	0.631	0.166	0.022
HLR only	0.708	0.654	0.160	0.022
TL-DTGBH (HLR+RB)	0.757	0.723	0.147	0.023

Means over five draws of withheld states; C pools all withheld states, C_w compares decks within a state, $CITL_5$ is averaged over states weighted by test size.

Table V returns to the protocol of the national benchmark, now giving each withheld state 20% of its spells as local history (about 2544 spells in total per draw) and evaluating on the remaining 80%. TL-DTGBH raises the pooled C over all withheld states from 0.646 to 0.757, close to the 0.769 obtained on a random split in which every state is seen during training, and the within-state C from 0.654 to 0.723, and it reduces the state-level calibration error from 0.096 to 0.023. HLR alone delivers about half of the pooled gain (C = 0.708) because withheld states differ strongly in their baseline hazard, but it leaves the within-state C unchanged (0.654); that requires the residual stage. Classical Cox updating calibrates as well as HLR ($CITL_5$ = 0.022) but remains limited by the proportional-hazards national model (within-state C = 0.631).

6.5. What Transfers and What Must Be Relearned

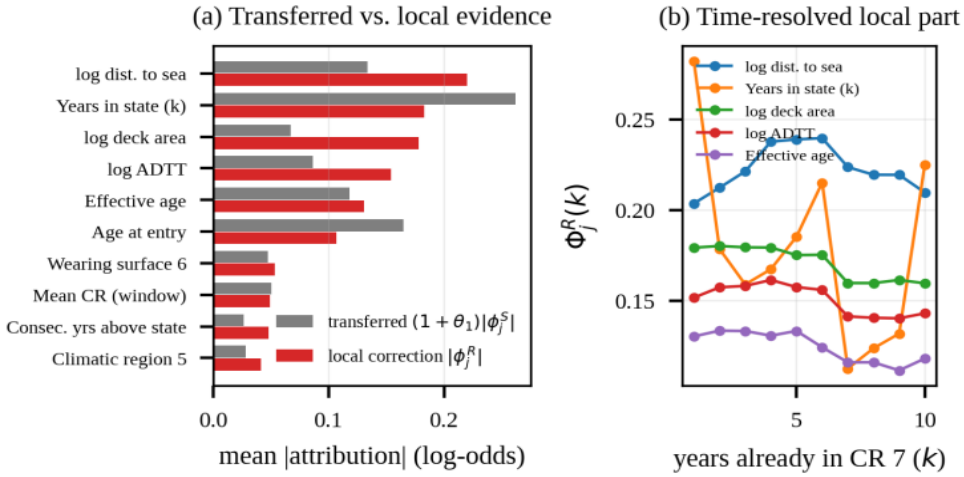


Figure 5. Exact decomposition of the adapted models (n = 500, first draw, 14 states). (a) Mean absolute attribution of the ten features with the largest local corrections: transferred national evidence versus local residual ensemble. (b) Time-resolved profiles $\Phi^R_j(k)$ of the five largest local corrections.

The decomposition (7) is exact in practice: re-fitting the stored HLR and reloading the residual ensembles reproduces the stored predictions to within 3×10^{-8}, and the three components add up to the model logit to within 4×10^{-14}. Local relearning is substantial. Averaged over the 14 states the local ensembles carry 45% of the total attribution mass, from 16% to 83% across states, and cross-validation kept a non-empty residual ensemble in every state (median 146 rounds). Fig. 5(a) shows that the corrections are concentrated. Distance from the sea, deck area and ADTT receive local attributions larger than their transferred ones (0.220 versus 0.134, 0.178 versus 0.067 and 0.154 versus 0.086), and the period index k, through which the shape of the baseline hazard is expressed, is also corrected substantially (0.183 versus 0.262). The recent rating history, in contrast, is almost entirely transferred: the rating five years before entry carries a transferred attribution of 0.200 but a local one of only 0.034. The variables that must be relearned are those through which state practice plausibly enters the NBI record, namely chloride and de-icing exposure, traffic loading and the size-dependent use of overlays and protection systems, whereas the meaning of a deck's rating trajectory is common to all states.

Fig. 5(b) applies the time-resolved profile (8) to the local part. Most covariate corrections act almost uniformly over the sojourn (distance from the sea between 0.20 and 0.24, deck area between 0.16 and 0.18 across k = 1–10), so they shift the hazard level of particular decks. The correction carried by the period index itself is strongly time-dependent: it is largest in the first year after entry (0.28), smallest in the middle of the sojourn (0.11 at k = 7) and rises again after ten years (0.23). States therefore differ most in how quickly decks leave CR 7 right after entering it and in the hazard of long sojourns, a change in the shape of the baseline hazard that no proportional-hazards update can represent. Engineers reviewing an adapted model can therefore read, deck by deck and year by year, how much of a prediction rests on national experience and how much on the state's own history.

6.6. Decision Impact

Table VI. Targeting efficiency inside the target states

Risk input to the ranking	TE (PPI)	TE (risk only)
Exposure / cost only	0.706	–
DT-GBH, national	0.701	0.457
DT-GBH, local (n = 100)	0.748	0.496
TL-DTGBH (n = 100)	0.729	0.473
DT-GBH, local (n = 500)	0.779	0.592
Pooled + indicator (n = 500)	0.792	0.587
TL-DTGBH (n = 500)	0.786	0.586
Random selection	0.237	0.294

TE: benefit captured relative to a clairvoyant ranking, averaged over budgets of 5–50% and over states (first draw).

Table VI repeats the prioritization experiment of the DT-GBH framework inside each target state: decks are ranked by PPI = $\hat{p}_5 \cdot e/c$, with exposure e proportional to ADTT times deck area and normalised cost c with a fixed share of 0.3, and the selection is scored against downgrades observed within five years in the test half. With the national model the PPI is no better than ranking by exposure and cost alone (TE = 0.701 versus 0.706): a miscalibrated risk input adds nothing to the decision. With 500 local spells every adapted model lifts TE to about 0.78–0.79 (TL-DTGBH 0.786), and ranking by risk alone shows the same pattern (0.457 for the national model, 0.586 for TL-DTGBH). At n = 100 the local-only model attains a higher TE than TL-DTGBH (0.748 versus 0.729) despite its lower C, a reminder that targeting depends on the very top of the ranking, which the C-index weighs only lightly.

7. Discussion and Limitations

The results suggest a simple operating procedure for agencies that inherit a national model: recalibrate it on the hazard-logit scale as soon as about 50 local sojourns are available, switch on residual boosting at about 100, and repeat the adaptation, which takes seconds, whenever new inspections arrive. The exact decomposition makes each update auditable, since it shows which predictions changed because of the state's own history.

Several limitations bound these conclusions. State identifiers in the public panel are anonymised, so local corrections cannot be linked to documented state policies. Target states were restricted to those with at least 1,000 CR 7 spells to keep test sets large; smaller states enter only through protocol (3). Local samples were drawn at random from a state's record, whereas an agency's history is a contiguous period, which adds temporal shift. Rating increases are treated as independent censoring, a competing-risks formulation would remove this assumption, and the rating is a partly subjective index, so inspector effects and physical deterioration cannot be separated with these data. Finally, the decision analysis assumes perfectly effective treatment.

8. Conclusion

A national discrete-time boosted hazard model transfers poorly across states because both its covariate effects and the level and shape of its hazard are state-specific, and covariate-shift corrections cannot fix either. TL-DTGBH adapts such a model from a short local history by ridge-shrunk recalibration of the hazard logit followed by cross-validated residual boosting. On 14 target states it halves the calibration error with 50 local sojourns and lifts the C-index from 0.666 to 0.732 with 500. It outperforms local-only and Cox-updating models when local data are scarce and matches a pooled model that must be retrained for every state. Under the unseen-state protocol, a 20% local history restores a pooled C-index of 0.757, close to the random-split level. An exact attribution identifies the drivers that each state must relearn. Future work will add competing risks for rehabilitation, adaptation from contiguous inspection periods and hierarchical priors for the recalibration parameters.

References

[1] American Society of Civil Engineers. (2025). 2025 Report Card for America's Infrastructure. ASCE.

[2] Federal Highway Administration. (1995). Recording and coding guide for the structure inventory and appraisal of the nation's bridges (Rep. FHWA-PD-96-001). U.S. Department of Transportation.

[3] Madanat, S., Mishalani, R., & Ibrahim, W. H. W. (1995). Estimation of infrastructure transition probabilities from condition rating data. Journal of Infrastructure Systems, 1(2), 120–125. https://doi.org/10.1061/(ASCE)1076-0342(1995)1:2(120)

[4] Sobanjo, J. O. (2011). State transition probabilities in bridge deterioration based on Weibull sojourn times. Structure and Infrastructure Engineering, 7(10), 747–764. https://doi.org/10.1080/15732470902917028

[5] Manafpour, A., Guler, I., Radlińska, A., Rajabipour, F., & Warn, G. (2018). Stochastic analysis and time-based modeling of concrete bridge deck deterioration. Journal of Bridge Engineering, 23(9), 04018066. https://doi.org/10.1061/(ASCE)BE.1943-5592.0001285

[6] Fleischhacker, A., Ghonima, O., & Schumacher, T. (2020). Bayesian survival analysis for US concrete highway bridge decks. Journal of Infrastructure Systems, 26(1). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000511

[7] Lu, M., & Guler, S. I. (2022). Comparison of random survival forest with accelerated failure time-Weibull model for bridge deck deterioration. Transportation Research Record, 2676(7). https://doi.org/10.1177/03611981221078281

[8] Valkonen, A., & Glisic, B. (2023). Effect of bridge data heterogeneity on neural network survival predictive performance. In Proceedings of the 14th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP14).

[9] Zhang, S., & Qiu, L. (2026). Explainable discrete-time survival learning for bridge deck deterioration and risk-based preservation prioritization. In Proceedings of AMMEE.

[10] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[11] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2), 227–244. https://doi.org/10.1016/S0378-3758(00)00115-4

[12] Dai, W., Yang, Q., Xue, G.-R., & Yu, Y. (2007). Boosting for transfer learning. In Proceedings of the 24th International Conference on Machine Learning (ICML) (pp. 193–200). https://doi.org/10.1145/1273496.1273521

[13] Marcelino, P., Antunes, M. de L., Fortunato, E., & Gomes, M. C. (2020). Transfer learning for pavement performance prediction. International Journal of Pavement Research and Technology, 13, 154–167.

[14] Li, Y., Wang, L., Wang, J., Ye, J., & Reddy, C. K. (2016). Transfer learning for survival analysis via efficient L2,1-norm regularized Cox regression. In Proceedings of the IEEE 16th International Conference on Data Mining (ICDM) (pp. 231–240). https://doi.org/10.1109/ICDM.2016.0034

[15] Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., & Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. Statistics in Medicine, 23(16), 2567–2586. https://doi.org/10.1002/sim.1844

[16] van Houwelingen, H. C. (2000). Validation, calibration, revision and combination of prognostic survival models. Statistics in Medicine, 19(24), 3401–3415.

[17] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[18] Fleischhacker, A., Ghonima, O., & Schumacher, T. (2018). Bridge data and code to accompany Bayesian survival analysis for US concrete highway bridge decks [GitHub repository]. https://github.com/flyaflya/nbiData

[19] Bektas, B. A., Carriquiry, A., & Smadi, O. (2013). Using classification trees for predicting national bridge inventory condition ratings. Journal of Infrastructure Systems, 19(4), 425–433.

[20] Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. Sociological Methods & Research, 13, 61–98. https://doi.org/10.2307/270718

[21] Tutz, G., & Schmid, M. (2016). Modeling discrete time-to-event data. Springer.

[22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

[23] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774).

[24] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.

[25] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220.

[26] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. JAMA, 247(18), 2543–2546.

[27] Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30(10), 1105–1117. https://doi.org/10.1002/sim.4154

[28] Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18(17–18), 2529–2545.

[29] Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research, 21(212), 1–6.

[30] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785