applied system innovation

MDPI

Article

Entity Overlap and Generalisation in Last-Mile Route-Deviation Prediction: Evidence from Driver Holdout, Address Suppression, and Cross-Country Transfer

Sara Atef [1,2,*] and Amr B. Eltawil [3,4]

[1] Department of Industrial Engineering, College of Engineering, University of Business and Technology, Jeddah 23435, Saudi Arabia
[2] Department of Industrial Engineering, Faculty of Engineering, Zagazig University, Zagazig 44519, Egypt
[3] Department of Industrial and Manufacturing Engineering, Egypt-Japan University of Science and Technology (E-JUST), New Borg Elarab City 21934, Egypt; eltawil@ejust.edu.eg
[4] Production Engineering Department, Faculty of Engineering, Alexandria University, Alexandria 21544, Egypt
* Correspondence: sara.a@ubt.edu.sa

Abstract

Planned last-mile routes often change during execution, making pre-dispatch identification of likely deviations useful. Performance measured when test routes involve drivers and addresses represented in training may not extend to unseen entities. This study audits training–test entity overlap in Norwegian–Danish route data, using ridge-regularised linear models as instruments and gradient-boosted trees as a nonlinear check. Rolling-origin, leave-driver-out, paired known-versus-unseen driver, graph-blocked leave-community-out, and bidirectional cross-country evaluations share a strict cohort of 18,723 routes. In the paired design, identical test routes were scored under matched training size, deviation prevalence, and country mix. The driver-identity gain in the area under the receiver operating characteristic curve (AUC) was 0.0837 for drivers observed in training and −0.0007 for unseen drivers, as indicator coding implies, and a within-dataset experience descriptor added little. Address identity added 0.0426 for unseen drivers with the linear baseline; this increment was not reproduced in the tested gradient-boosting specification with target encoding or under graph-blocked evaluation with reduced address overlap. Under cross-country transfer, AUC exceeded chance modestly in both directions, every unadjusted Brier skill score was negative, and identity-fitted specifications ranked routes better than structural features alone only into Norway. Findings are limited to one operator and two countries.

Keywords: last-mile delivery; route-deviation prediction; training–test entity overlap; model evaluation; generalisation; cross-country transfer; probability calibration

Check for updates

Academic Editor: Bo Lu

Received: 4 October 2026
Revised: 16 October 2026
Accepted: 21 October 2026
Published: 23 October 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Last-mile delivery is the stage at which a planned route is tested in day-to-day operation. A route that appears workable before dispatch may still be altered during execution, changing the order in which stops are served. Predicting this departure from plan could help dispatchers direct limited review capacity towards routes most likely to change. Its operational value, however, depends on whether the reported predictive performance persists when future routes involve drivers, addresses, or markets that were not represented during model training. This study examines that evaluation question. The following sections

define the prediction problem, position it within the related literature, and present the research questions and contributions.

1.1. Planned Routes, Executed Routes, and the Scope of Prediction

A planned route specifies an intended stop sequence, yet the sequence executed by the driver often differs. In a 1-year study of a beverage distributor operating in Mexico and the United States, 75% of 55,525 planned route instances were not executed as planned. On average, 55.3% of the segments in the planned sequence were not followed, although the accompanying increase in straight-line distance was only 2.6% [1]. The contrast suggests that deviation is common and is expressed mainly through changes in stop order rather than through large increases in travel distance. Konovalenko et al. [2] reported a similar pattern for a commercial operation spanning Norway and Denmark: only about one route in four was executed in exactly the planned order.

The difference between a planned route and its execution can partly be attributed to information available to drivers but absent from the planning representation. Repeated service gives drivers operational knowledge of access constraints, customer availability, and traffic conditions in the areas they serve [2]. Because such knowledge is experience-based and difficult to formalise, it can remain outside explicit planning rules [3]. Controlled vehicle-routing experiments have likewise shown that routes violating hidden time-window, capacity, and precedence constraints can be distinguished from compliant routes using the observed stop sequence alone [4]. Such knowledge is not uniform across drivers: couriers serving the same work area can exhibit distinct routing profiles [5], and driver-specific behaviour has also been incorporated into routing models [6]. These findings make driver and address identities plausible predictors. A driver identifier may summarise recurring behavioural patterns, while an address identifier may capture persistent local conditions. The same features, however, create a generalisation risk: the history represented by an identifier is available only when the corresponding driver or address has previously appeared in training.

This tension is particularly important for compliance prediction. Rather than attempting to reproduce an executed tour, the task asks before dispatch whether a planned route is likely to be followed. It has been formalised on a commercial dataset covering operations in Norway and Denmark, where a positional route-quality score, a normalised edit-distance score, and a relative distance-deviation score were combined to classify routes as Deviated (D) or Not Deviated (ND) [2]. The publicly released data include the planned and executed route pairs together with driver and address identifiers [7]. Such prediction has a direct operational use: when only a limited number of routes can be reviewed, a dispatcher can prioritise those most likely to depart from the plan. Assessing that operational value therefore requires an evaluation regime aligned with the intended deployment population.

In the published temporal design, the data were ordered chronologically and stratified so that each Driver-identity $\times$ Day combination was represented across the training, validation, and test subsets [2]. The resulting test set contained drivers with training history, and the present audit also identified substantial training–test recurrence of delivery addresses. Such overlap does not inherently constitute data leakage; it is appropriate when deployment remains within the same recurring operation [8]. Grouped or blocked evaluation is instead required when prediction is intended for new individuals, locations, or operational groups [9]. The published split estimates within-operation performance, but it cannot by itself establish performance for a newly hired driver, a newly onboarded address, or a new national market.

This distinction can materially change reported performance. In wearable-sensor activity recognition, for example, a random forest achieved 89% accuracy under subject-

mixed cross-validation but 76% under leave-one-subject-out evaluation, where each test subject was entirely unseen during training [10]. A related limitation has been observed in last-mile routing: a label-guided transition matrix learned from observed routes deteriorated by 48.8% relative to a rule-based alternative on held-out routes because previously unobserved transitions carried no learned information [11]. Together, these examples show that predictive performance can depend on entity-specific information that is available only when the evaluation preserves operational continuity. The relevant question is not whether a within-operation design is legitimate, but how the conclusions change when the continuity of drivers, addresses, and national markets is progressively removed.

1.2. Related Work and the Evaluation Gap

Research closest to this problem has focused primarily on route generation, particularly following the release of the Amazon Last Mile Routing Research Challenge dataset [12]. The resulting methods fall into three broad families. Optimisation-based approaches extract structural properties from historical routes and incorporate them into constrained local search or hierarchical travelling-salesman formulations [13,14]. Learning-augmented approaches guide sequence construction or selection using a learned route-quality classifier, regularised route-score models, or learned transition weights passed to a solver [11,15,16]. Neural approaches use attention-based pointer networks [17] and zone-based graph and pointer architectures trained separately by spatial zone [18]. These families reflect the broader distinction between methods in which learning constructs a routing solution and those in which it guides an optimisation procedure [19]. Despite their methodological diversity, they share the objective of constructing a route that resembles an observed executed sequence.

These methods are adjacent to, but distinct from, route-deviation prediction. Route generation receives a set of stops and constructs an order, whereas deviation prediction receives a completed plan and asks whether that plan will survive execution; its output supports dispatch review rather than route construction. The data requirements differ accordingly. The challenge dataset provides executed routes without their corresponding planned sequences, whereas the dataset examined here contains both the plan and its execution [2]. Because the latter represents recurring service areas, drivers and addresses can reappear across routes. The resulting identity information may be predictive, but its contribution cannot be understood without measuring which entities recur across training and test.

Most reported evaluation designs do not manipulate that recurrence. Random or partially described partitions include the 4889/1223-route split used in [13,17], the 85/15 city-stratified split in [15], and a 5112/1000-sample partition without a reported entity-disjointness condition in [6]. Other studies evaluate within a single metropolitan area [18] or preserve entity continuity through time: the design in [5] trains on 2 months and tests on a 3rd using separate per-courier models, while [20] uses a chronological 6:2:2 split and separate models by city. Such designs permit couriers, locations, or cities to remain represented in both the training and test sets. They therefore assess new route instances within an existing operational setting but do not directly test whether identity-related gains persist when the corresponding entities are unseen. More broadly, inconsistent evaluation conventions already limit comparisons across routing studies [19].

Under these designs, transferable route structure and recurring entity identity are not measured separately. Location-related information has nevertheless proved predictive: Li and Phillips reported a generalised R-squared value of 0.155 for route-centroid postal-code information alone in their United States logistic model [1], while Konovalenko et al. found the ordered planned-route representation particularly informative for regression [2].

Neither analysis isolated address identity from route order, structural characteristics, or regional context. It therefore remains unclear whether the observed gains reflect transferable route regularities, repeated entity history, or a combination of the two.

Evaluation methods for unseen entities are well established outside the immediate task. Grouped and blocked validation can target prediction for new individuals, groups, or locations [9]. Within last-mile routing, however, direct precedents remain limited. The closest geographical-transfer exercise trained a sequencing method on United States challenge routes and tested it on 66 routes from Uruguay, where zone sequencing was reported as the principal source of error because the learned information came from non-local routes [21]. That study addressed route generation and examined only one transfer direction. Cross-city generalisation has likewise been identified as an open challenge on industrial delivery data, but the available evidence consists of performance differences between separately trained per-city models rather than direct transfer from one city to another [20].

Taken together, this literature leaves an evaluation gap rather than a modelling gap. No closely related route-deviation study identified here combines a paired known-versus-unseen driver comparison on identical test routes, controlled suppression of the address-identity channel, and bidirectional transfer between national markets. Feature admissibility against the dataset's own time-reference construction has likewise not been audited. The present study therefore treats training–test entity overlap and feature admissibility, rather than a new predictive architecture, as the primary experimental objects.

1.3. Objectives, Research Questions, and Contributions

Against this background, the aim is not to propose a more accurate route-deviation predictor, but to determine what reported predictive performance means when training–test entity overlap changes. Four research questions, denoted RQ1–RQ4, organise the audit. RQ1: How much of the in-operation predictive value attributed to driver identity depends on the test driver having been observed in training, and does a driver-level descriptor that remains defined for drivers not used in training carry transferable information? RQ2: Under the same unseen-driver condition, how much of the predictive value attributed to address identity remains, and are the two identity channels complementary or substitutable? RQ3: How does predictive gain respond as the address-identity vocabulary available to the model is progressively withdrawn? RQ4: Under cross-country transfer, where drivers and addresses are jointly unseen, can the model still distinguish deviated from non-deviated routes, produce reliable probabilities, and prioritise useful cases within a fixed review budget?

These questions are addressed through a controlled multi-regime evaluation. The five primary regimes—rolling-origin evaluation, leave-driver-out (LDO) evaluation, the paired Known/Unseen comparison, leave-community-out evaluation on the address co-occurrence graph, and bidirectional cross-country transfer—use the same strict cohort of 18,723 routes with regime-specific training and test partitions, while a separate 18,741-route temporal cohort is retained as a descriptive reference. Before the confirmatory analysis, a time-reference invariance audit establishes the admissible pre-route feature set by excluding level variables tied to an outcome-informed time origin. The central paired comparison scores identical test routes under matched training size, training deviation prevalence, and country mix, both with and without the test drivers' history; uncertainty in the resulting attenuation estimates is quantified using a repeat-aware driver-cluster bootstrap. The address channel is examined separately through graded suppression of the training address vocabulary. Bidirectional cross-country transfer then evaluates discrimination, probability scoring, and ranking value under fixed review budgets. Together, these elements turn

training–test entity overlap from an incidental property of a data split into an explicit experimental factor.

The resulting claims are deliberately bounded. The evidence comes from one logistics operator in two national markets and is established using regularised linear models with a bag-of-addresses representation, with gradient-boosted trees and out-of-fold target encoding as a nonlinear check of the paired design. The reported driver and address effects are predictive associations rather than causal effects, and their magnitudes are not assumed to transfer unchanged to sequence models or learned embeddings. Direct numerical comparison with the published neural-model results is avoided because the cohorts, representations, and evaluation regimes differ.

2. Materials and Methods

2.1. Dataset and Outcome Definitions

The public dataset of planned and driven last-mile routes described by Konovalenko et al. [7] and released in the accompanying repository [22] was used in this study. It contains stop-level records for routes executed by a commercial operator in two Nordic countries, coded 0 = Norway and 1 = Denmark. Each stop carries a route identifier, a driver identifier, a stop identifier, an address identifier corresponding to a unique physical location, planned and actual sequence indices, planned and actual geodesic distances to the preceding stop, time-window bounds, an arrival time, a day-of-week field, and binary depot and delivery indicators. Multiple stop identifiers may map to the same address identifier. Routes are ordered chronologically by first arrival, and a route may span 2 calendar days when it crosses midnight. To illustrate this stop-level structure, Table 1 presents five consecutive records from a single route using a selected subset of the available variables. The route, driver, stop, and address identifiers have been replaced with short labels for readability, while the sequence indices and stop-type indicators retain their original values.

Table 1. Anonymised extract of five consecutive stop-level records from a single route.

Route	Driver	Stop	Address	Day of Week	IndexP	IndexA	Depot	Delivery
R1	D1	S1	A1	Monday	0	0	1	0
R1	D1	S2	A2	Tuesday	1	4	0	1
R1	D1	S3	A3	Tuesday	2	5	0	1
R1	D1	S4	A4	Monday	3	2	0	1
R1	D1	S5	A4	Monday	4	1	0	1

Note: The extract corresponds to a route from Denmark (Country = 1). IndexP and IndexA denote the planned and actual sequence positions, respectively. Depot and Delivery are binary indicators. Identifiers were relabelled in order of appearance; consequently, S4 and S5 show that two distinct stop identifiers may refer to the same physical address, A4.

Table 2 summarises the dataset as distributed together with the two cohorts derived from it; their construction is described in Section 2.2.

All time fields are expressed in minutes relative to a route-level reference

$$t_{\text{ref}} = \min\{\, t_a(s),\ t_e(s),\ t_l(s)\ :\ s \in S \,\} \tag{1}$$

where S is the set of stops on the route and t_a, t_e, and t_l denote arrival time, earliest window bound, and latest window bound. This construction is central to Section 2.3: because t_{ref} is a minimum taken over a set that includes realised arrival times, any feature whose value is a level on this time axis is partly a function of the outcome, whereas differences between two times on the same axis are invariant to the reference and are admissible.

Table 2. Dataset composition and construction of the reference temporal and strict audit cohorts.

Quantity	Value
Stop-level records	249,231
Routes	19,647
Drivers (country-compounded keys)	401
Addresses (country-compounded keys)	10,872
Weeks	32
Countries	2 (0 = Norway, 1 = Denmark)
Routes in Driver–Day groups of three or fewer routes	906
Excluded: routes with no non-depot stop	18
Reference temporal cohort (14,472/3026/1243)	18,741
Strict audit cohort	18,723

The distributed file contains 19,647 route identifiers, matching the count reported by Konovalenko et al. [2], whereas the Data in Brief abstract reports 19,497 route pairs. The file-derived route count is therefore used and reported.

The published outcome formulas and decision thresholds were implemented directly from the released file [2]. For each route, three deviation measures compare the planned route with its execution: the Route Quality Score (RQS), a continuous score in [0, 1] derived from normalised positional displacement, where 1 indicates perfect adherence; the Edit Distance Score (EDS), the Levenshtein edit distance between the two sequences normalised to [0, 1], following the permutation-representation distance framework of Sörensen [23]; and the Distance Deviation Score (DDS), the relative difference between actual and planned travelled distance. The central RQS and EDS summaries were closely reproduced. Repeated address identifiers were tagged by occurrence number before positional matching for RQS, whereas EDS was calculated on the original address sequence, including repeated tasks at the same physical location. Minor count differences remained, and the file-based DDS maximum was 5.227 compared with 2.00 reported in the reference study. Because the publication does not document a DDS cap, none was imposed; the file-based recomputation was retained and the discrepancy reported transparently. A route is labelled ND if and only if

$$RQS \geq 0.85 \ \wedge \ EDS \leq 0.30 \ \wedge \ DDS \leq 0.03 \tag{2}$$

Here, $\wedge$ denotes logical conjunction. A route is labelled D otherwise; failure on any single criterion is sufficient for the D label. Regression targets RQS, whereas classification targets D as the positive class.

Deviation is the majority outcome: 8141 of 19,647 routes (41.44%) are labelled ND, and only 5269 routes (26.82%) are executed with an exactly matching address sequence. Across all routes, mean RQS is 0.8211 (standard deviation (SD) = 0.1897; median = 0.8750) and mean EDS is 0.3255 (SD = 0.2619). DDS has a median of exactly zero and a slightly negative mean (-0.0012; SD = 0.1909 across the 19,646 routes with a finite value). The mean DDS was close to zero; thus, route-level sequence deviation was not associated with a positive mean distance difference in this file. This descriptive result does not identify why drivers deviated. Because RQS is strongly left-skewed with a mass at 1, its mean-squared error (MSE) should not be interpreted as if the target were symmetric.

2.2. Common Eligible Cohort

The five primary audit regimes apply the same eligibility rule to a strict cohort of 18,723 routes. A separate reference temporal cohort is retained for descriptive comparison and is not pooled with those analyses. Two exclusions are applied to the 19,647 routes, in order: 906 routes are removed by the Driver-Day group-size rule, which excludes

groups with three or fewer observations, matching the published eligibility filter [2]; a further 18 routes are removed because they contain no non-depot stop and their outcome is degenerate. No minimum number of routes per driver is imposed, and the filter is applied once rather than iterated.

The two cohorts serve different purposes: the 18,741-route cohort supports the descriptive reference split, whereas the 18,723-route strict cohort supports the five primary audit regimes. The strict audit cohort contains 18,723 routes (95.30% of all routes) and carries both exclusions; every cross-regime comparison in Section 3 uses it. Within it, 8013 routes are Norwegian and 10,710 Danish. The reference temporal cohort contains 18,741 routes, applies only the Driver-Day exclusion, and is partitioned chronologically into 14,472 training, 3026 validation, and 1243 test routes; it supports the reference results reported in Section 3.2 and is never pooled with the strict-cohort regimes.

The Driver-Day exclusion altered outcome prevalence: the excluded routes had a deviation rate of 0.4415, compared with 0.5932 in the retained strict cohort and 0.5856 across all routes. All regime-specific deviation rates are therefore conditional on this eligibility rule.

2.3. Time-Shift Invariance Audit

Because the dataset's time axis is defined relative to a route-level reference that is itself computed from a set containing realised arrival times (Section 2.1), feature admissibility cannot be settled by inspection. A mechanical time-shift invariance test was therefore implemented, specified before the final confirmatory run following pilot inspection. For each candidate route-level feature, the raw stop records were shifted and the feature was recomputed under two schemes: (i) a single global shift of +1000 min applied to all routes and (ii) one independently sampled route-specific shift drawn from a uniform distribution over 0–5000 min for each route. A feature is admissible only if its recomputed value is invariant under both schemes to within a feature-specific numerical tolerance. The per-route scheme is the binding test, since a feature normalised against a dataset-wide time statistic would pass a global shift but fail per-route shifts.

Fourteen candidate features were audited. Twelve are invariant under both schemes and were retained. Two—the route minimum earliest-window bound and the route maximum latest-window bound—changed by up to 4999.92 min under per-route shifting and were rejected as time-reference level features. This is not a defect of the dataset but a direct consequence of the documented reference construction, and it means that time-window information may enter a model only as differences (window width, route span), never as levels. The audit establishes that no retained feature depends on the route's time origin; it does not by itself certify the absence of every other form of leakage, which is addressed by the feature-family classification below. To assess whether this exclusion affected predictive performance, the rolling-origin models were also refitted with the two rejected level features added to the Structure set. These refits are reported in Section 3.2 for comparison only and are not used in any other analysis.

2.4. Feature Families and Feature Sets

Every variable is assigned to exactly one family carrying an explicit admissibility decision. Twelve planned-structure features are knowable before execution and survived the invariance audit: country, stop and address counts, delivery counts and share, depot count, planned distance and leg statistics, mean time-window width, and route span. Delivery share is the number of delivery stops divided by the total number of stops, including depot stops. The planned-leg statistics are the mean and maximum of the planned distance to the preceding stop, calculated across all stops on the route. Mean time-window width is the average of the per-stop differences between the latest and

earliest allowed arrival times. Route span is the maximum latest-window bound minus the minimum earliest-window bound across all stops on the route. Both time-window features are computed from window bounds only; no arrival time enters either feature. Driver identity and address identity are represented by country-compounded keys.

In the main feature sets, the Driver block represents identity only. A driver-level descriptor that remains defined for drivers not observed in training is evaluated separately (Section 2.6.4). This Experience feature is the number of routes completed by the same driver in the dataset during the 2 dataset weeks preceding the week of the route, counted from route dates only; it therefore measures experience observed within the dataset rather than total professional experience. Experience uses no outcome and no time field of the route itself, so it is unaffected by the time-reference construction in Section 2.3. Because it is computed from the complete route record, a route has the same Experience value in the Known and Unseen conditions. The week of a route is taken from its first stop; using the first planned stop instead changes the week of two routes and the Experience value of eight routes. The preceding history is incomplete for the 926 strict-cohort routes in the first two dataset weeks; these routes are retained, their Experience value is set to zero, and a binary indicator marks the incomplete history. In the strict cohort, the median Experience is six routes (7.65% of routes with zero); among the 17,797 routes with complete history, it is seven routes (5.42% with zero).

Outcome variables—actual sequence index, actual distance, arrival time, RQS, EDS, DDS, and the D label—are never predictors. Two features were rejected outright by the invariance audit.

One constant baseline and four nested feature sets were evaluated. The Global Historical Average (GHA) predicted the training-set mean RQS for regression and the training-set prevalence of D for classification. Because it assigned the same classification score to every route, it provided chance-level discrimination. The fitted feature sets were Structure, Structure + Driver, Structure + Address, and Structure + Driver + Address. Driver and address identities were encoded using country-compounded indicator variables. The Address block was a sparse bag-of-addresses representation over 10,872 keys and recorded whether an address occurred on the planned route, not its planned position.

Before each model was fitted, indicator columns absent from its training data were removed from both the training and test design matrices. A driver or address appearing only in the test set was therefore represented by an all-zero identity block and received no fitted identity-specific coefficient. Structure columns with zero variance in the corresponding training split were handled in the same way.

2.5. Diagnostic Models and Regularisation Tuning

The primary diagnostic models were ridge-regularised linear models—least-squares models for RQS regression and logistic models for D/ND classification—fitted using MATLAB's fitrlinear and fitclinear functions with the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) solver and a fitted bias term. Within each split, non-finite Structure values were replaced using the corresponding training-set median, and the Structure columns were standardised using training-set means and standard deviations. Driver and Address indicators were left unscaled. RQS predictions were clipped to the valid RQS range from 0 to 1. The object of study is how much predictive signal flows through each feature channel under each entity-overlap regime. A linear model attributes that signal to channels transparently and remains computationally tractable across the factorial design and the 10,872-column sparse identity block. Regularisation is tuned within each training split, whereas uncertainty is estimated separately using the cluster-bootstrap

procedures described in Section 2.7. Accordingly, the reported magnitudes are specific to this model class.

Regularisation strength (λ) is the one hyperparameter that must not be fixed across feature sets. The Structure block has 12 columns and the Structure + Address block has more than 10,000, so a single shared penalty would compare them on incommensurable effective scales; a preliminary fixed-penalty run exhibited this artefact as a spuriously large address advantage. In the reported analysis, λ is selected inside the training portion for every feature set and every fold from a grid of 21 values spanning 10^{-8} to 10^{2} in half-decade steps, applied directly to the penalty term without rescaling by sample size. Selection uses validation MSE for regression and validation Brier score (BS) for classification; the model is then refitted on training plus validation at the selected λ, and the outer test set is never used for selection.

Across the 664 fitted models, 18 distinct grid values were selected, spanning 1.0×10^{-8} to 3.16. The selected value sat at the lower grid boundary in eight cases (1.20%), all of them Structure-only models; no Driver or Address model selected a boundary value. In the additional analyses, the leave-community-out fold in which validation selected the strongest penalties used the upper grid boundary (λ = 100) for RQS regression in all four fitted feature sets, and three Structure + Experience fits used the lower boundary; no other additional fit selected a boundary value.

The inner validation construction followed the outer evaluation regime. Rolling-origin evaluation used the adjacent chronological 10% block. LDO and paired Unseen conditions reserved complete training drivers until approximately 20% of the outer training routes were assigned to validation. The paired Known condition used the latest approximately 20% of each training driver's routes, and cross-country transfer used the latest 20% of the source-country routes. After λ was selected, the model was refitted on the complete outer training set. Within the address-suppression experiment, λ was selected at 0% suppression and then held fixed at the 20%, 40%, 60%, and 80% levels.

To test whether the conclusions depend on the linear structural baseline, gradient boosting machine (GBM) models with decision-tree learners [24] were also run through the paired design (Section 2.6.4) at 0% address suppression. Classification used LogitBoost [25] and RQS regression used least-squares boosting, implemented with MATLAB's fitcensemble and fitrensemble functions. The inputs were the 12 Structure features, either alone or with out-of-fold target encodings of driver identity, address identity, or both [26]. Each encoding replaced an entity by its smoothed training mean of the target, with smoothing equivalent to 10 routes, computed out of fold over five folds within the training data; an entity absent from the training data received the training mean, and the address encoding of a route was the mean over its distinct non-depot addresses. The learning rate was 0.1, and the minimum leaf size was 20 routes. Tree size (4 or 16 splits) and the number of boosting cycles (50, 100, 200, or 400) were selected on the same inner validation split as λ, by Brier score for classification and MSE for regression. The boosted models were evaluated primarily using AUC and RQS MSE; their Brier scores are also reported in Table S1. Classification probabilities were not recalibrated.

2.6. Evaluation Regimes

The six evaluation regimes are summarised in Table 3. Regime 1 uses the separate reference temporal cohort; Regimes 2–6 use the strict common cohort.

Table 3 reports realised training–test entity overlap before address suppression. Driver overlap is zero under LDO, in the Unseen arm of the paired comparison, and in both cross-country directions, but not in the Known arm or under leave-community-out evaluation (41.4–66.2%). Address overlap remains substantial in the other within-country evalua-

tions, ranging from 0.915 to 0.978, is reduced to 0.229–0.371 under leave-community-out evaluation, and falls to zero in cross-country transfer. The six evaluation regimes are described below.

Table 3. Evaluation regimes, test-set sizes, and realised training–test entity overlap.

Regime	Cohort	Test Routes	Driver Seen	Address Seen
Reference temporal split	Reference temporal	1243	100%	96.8%
Rolling-origin temporal	Strict	1872 per origin	94.1–98.1%	91.5–96.4%
LDO	Strict	≈3744 per fold	0%	92.9–96.5%
Paired Known/Unseen	Strict	≈1885 per fold	100%/0%	97.8%/95.1%
Leave-community-out	Strict	3699–3842 per fold	41.4–66.2%	22.9–37.1%
Cross-country	Strict	Denmark: 10,710; Norway: 8013	0%	0%

Note: Approximate test counts summarise folds with slightly varying sizes. Driver seen is the percentage of test routes, whereas Address seen is the percentage of test stops. Paired address-seen rates are measured at 0% suppression.

2.6.1. Reference Temporal Split

A single chronological split of the reference temporal cohort assigns 14,472 routes to training, 3026 to validation, and 1243 to testing. This split reuses the published outcome definitions and eligibility rule for descriptive comparison; it does not replicate the published predictive model, neural architectures, or full input set. Neither the experience covariate nor the day proxy is used as a predictor. Results are reported as point estimates without confidence intervals (CIs) and are not pooled with results from the strict-cohort regimes.

2.6.2. Rolling Origin Temporal Evaluation

Five rolling origins were evaluated, with cutoffs defined by route counts rather than calendar windows. The cohort is ordered chronologically; training grows from 40% to 80% of it in 10-point steps (7489; 9361; 11,233; 13,106; 14,978 routes), the adjacent 10% block serves as the inner validation set for selecting λ, and the following 10% block (1872 routes at every origin) is the test set. Validation and test blocks each contain approximately 1872 routes.

2.6.3. Leave-Driver-Out Evaluation

Twenty-five LDO folds (five repeats of five folds) were evaluated. All routes of the held-out drivers were removed from training in each fold. Training sizes are 14,977–14,981 and test sizes 3742–3746. Driver overlap in the test set is exactly zero, while 92.9–96.5% of test stops still occur at addresses seen in training, which is why this regime alone cannot separate the driver channel from the address channel.

2.6.4. Paired Known and Unseen Driver Comparison

The paired Known/Unseen comparison was the primary inferential design. Twenty-five outer cells were formed from five repeated five-fold driver partitions. Within each cell, candidate drivers assigned to the held-out fold were retained if they had at least four eligible routes. Each retained driver's routes were ordered chronologically; the earlier half formed the history available in the Known condition, and the latter half formed the common test set, with at least two routes required in each part. This construction yielded 63–64 eligible test drivers and 1882–1888 identical test routes per cell.

In the Unseen condition, all routes belonging to the eligible test drivers were excluded from training. In the Known condition, the earlier routes of those drivers were added to training, and an equal number of routes from other drivers were removed within each

Country × D cell. The replacement routes were sampled without replacement within each Country × D stratum using a fixed seed specific to the repeat and fold. Training size, training D prevalence, and country mix were therefore matched exactly between the two conditions in all 25 cells. Because the same test routes were scored in both conditions, their differences could be evaluated as paired contrasts. Address overlap remained close but not identical (97.80% in the Known condition and 95.08% in the Unseen condition); the comparison is therefore interpreted as a controlled predictive contrast rather than a causal effect of driver history. Removing the test drivers' routes from training also removes part of the training coverage of the areas they serve, which accounts for the lower address overlap in the Unseen condition; driver and address effects therefore cannot be separated in this design. The full calendar-time distribution of the training routes was not explicitly matched between conditions.

Two further feature sets, Structure + Experience and Structure + Experience + Address, were evaluated in the same 25 cells and both conditions at 0% address suppression, with λ selected in the same way. Because Experience is defined for every driver, its gain over Structure in the Unseen condition tests whether this descriptor adds predictive information for drivers absent from model training.

Address-identity suppression was crossed with the driver condition at five levels: 0%, 20%, 40%, 60%, and 80%. Within each of the five repetitions, addresses were randomly ordered separately by country, and the suppressed sets were nested so that the 20% set was contained within the 40% set, the 40% set within the 60% set, and so forth. Suppression set the selected address-indicator columns to zero in both training and test data while retaining the corresponding routes, structural features, and outcomes. For each outcome and address-bearing feature set, λ was selected at 0% suppression and then held fixed at the higher suppression levels. This manipulation measures signal carried by the address-identity channel and does not simulate entry into a new geographical service area.

2.6.5. Bidirectional Cross-Country Transfer

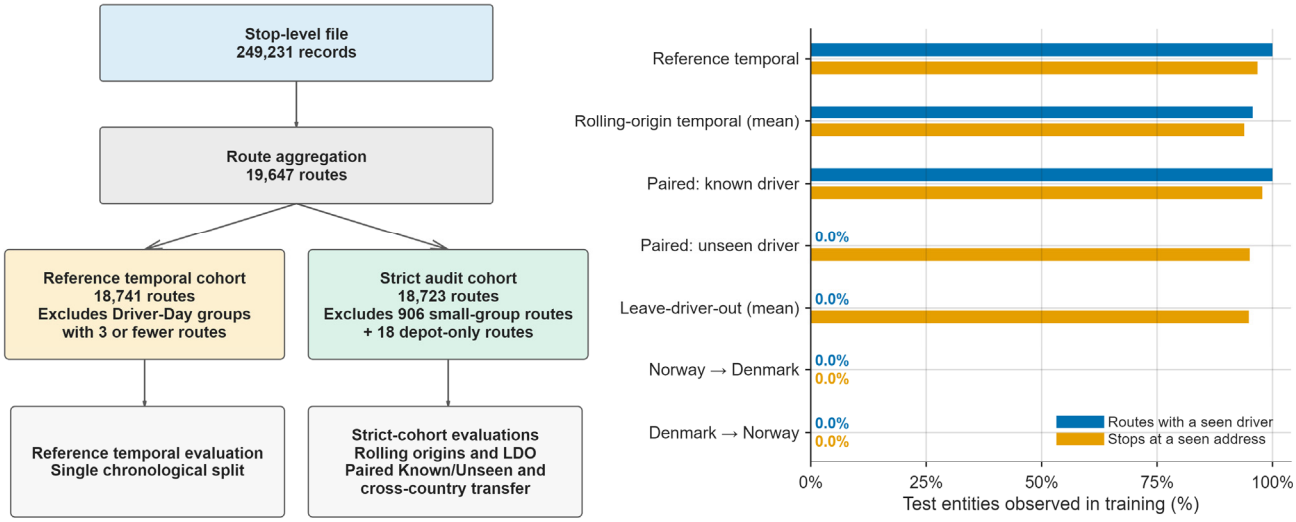
In one direction, all Norwegian routes (8013) were used for training and all Danish routes (10,710) were used for testing; the countries were reversed in the other direction. Measured driver and address overlap between source and target is zero in both directions, making this the only regime in which drivers and addresses are jointly unseen. Outcome prevalence differs sharply between the markets—Norway 0.3972, Denmark 0.7398—so this regime also stresses the probability scale, not only discrimination.

Figure 1 consolidates the data construction and evaluation design. Figure 1a traces the aggregation of the stop-level records into the reference temporal cohort and the strict audit cohort and links each cohort to its corresponding evaluation procedures. Figure 1b reports the realised training–test entity overlap for selected evaluation regimes: the proportion of test routes whose driver was observed in training and the proportion of test stops whose address was observed in training. Zero values in the cross-country regimes indicate complete absence of driver and address overlap rather than missing data.

2.6.6. Leave-Community-Out Evaluation

The public release contains no coordinates [7], so spatial blocking was not possible. A graph-blocked evaluation with reduced address overlap was therefore constructed on the address co-occurrence graph. Within each country, addresses are nodes, and edges connect addresses served on the same route, weighted by the number of such routes. Communities were detected by Louvain modularity maximisation [27] at resolution 1, retaining the partition with the highest modularity across five random seeds; this gave 43 communities (modularity 0.80) in Norway and 32 communities (modularity 0.91) in Denmark. Each

route was assigned to the community containing the majority of its distinct non-depot addresses, which for the median route contained all of them. Whole communities were allocated to five folds balanced by route count, giving 3699–3842 test routes per fold, and each fold was held out in turn, with the remaining strict-cohort routes used for training. The inner validation set for λ was also blocked by community.



(a) Data and evaluation workflow

(b) Training-test entity overlap by evaluation regime

Figure 1. Data construction and training–test entity overlap. (**a**) Route aggregation, cohort construction, and associated evaluation procedures. (**b**) Proportions of test routes with a driver observed in training and test stops with an address observed in training; cross-country zero values indicate no entity overlap.

Across the five folds, 22.9–37.1% of test stops occurred at addresses seen in training (22.9–36.0% when each distinct address of a route is counted once), 41.4–66.2% of test routes belonged to drivers seen in training, and 3576 test routes (19.1%) had no address seen in training. The communities are defined by co-service rather than by geography, so this regime is a graph-blocked evaluation with reduced address overlap rather than a spatial holdout or a holdout of entirely unseen addresses, and it reduces driver and address familiarity together. Performance is also reported separately for the test routes with no address seen in training. The deviation rate also differed between folds (0.25–0.51 in Norway and 0.68–0.83 in Denmark), so results are reported with each fold's PR-AUC baseline.

2.7. Metrics and Bootstrap Inference

Regression was scored by MSE on RQS. Classification was scored by the area under the receiver operating characteristic curve (AUC), the area under the precision–recall curve (PR-AUC), and BS; PR-AUC was interpreted against the corresponding positive-class baseline. Regression and classification metrics were not combined across tasks because they concern different targets fitted by different models.

Feature-block gains were oriented so that positive values indicated improvement on both tasks. For classification, AUC gain was calculated as the AUC of the augmented model minus the AUC of its baseline. For regression, MSE reduction was calculated as the baseline MSE minus the augmented-model MSE. Driver gain compared Structure + Driver with Structure; Address gain compared Structure + Address with Structure; Combined gain compared Structure + Driver + Address with Structure; and Driver increment compared Structure + Driver + Address with Structure + Address. Where the Address increment over Driver is reported, Structure + Driver + Address was compared with Structure + Driver.

For paired comparisons, attenuation was defined as the Known-condition gain minus the corresponding Unseen-condition gain.

For cross-country transfer, raw BS values are not comparable between directions because target prevalence differs. The Brier skill score (BSS) is therefore reported against a constant forecaster fixed at the target country's prevalence,

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{model}}}{\mathrm{BS}_{\pi_{\mathrm{target}}}} \tag{3}$$

where $\mathrm{BS}_{\mathrm{model}}$ is the model's BS, $\mathrm{BS}_{\pi_{\mathrm{target}}}$ is the BS of the constant forecast at the target-country prevalence, and π_{target} denotes the target-country prevalence. BSS equals zero for the reference forecaster and is positive only if the transferred model beats it. A negative BSS means that the transferred probability predictions score worse than the target-prevalence constant forecast. Calibration was assessed separately using expected calibration error (ECE) over 10 equal-count bins, the calibration slope, and reliability diagrams.

An oracle prior-adjusted variant is additionally reported in which predicted probabilities are recalibrated using the observed target-country prevalence. The adjustment was implemented as $\mathrm{logit}(\mathrm{p}_{\mathrm{adj}}) = \mathrm{logit}(\mathrm{p}) + \mathrm{logit}(\pi_{\mathrm{target}}) - \mathrm{logit}(\pi_{\mathrm{source}})$, where p is the original transferred probability and π_{source} and π_{target} are the source- and target-country deviation prevalences, respectively. All probabilities were clipped to $[10^{-6}, 1 - 10^{-6}]$ before the logit transformation. Because π_{target} is the realised target-country prevalence, this information would not be available before deployment. The adjustment is therefore diagnostic rather than deployable; its role is to distinguish whether transfer failure arises primarily from the probability level or from the ranking itself. A deployable alternative estimates the target prior from the unlabelled target routes by the expectation–maximisation (EM) procedure of Saerens et al. [28] and applies the same logit adjustment with the estimated prior in place of π_{target}. Because EM increases a log-likelihood that is concave in the prior, its limit is the maximiser of that likelihood on $[10^{-6}, 1 - 10^{-6}]$, which was computed directly. The procedure assumes that the class-conditional distributions are unchanged between countries and that the source probabilities are calibrated, so it is evaluated here rather than assumed to improve transfer.

Operational value is reported as Precision@Budget, Recall@Budget, and Lift@Budget at review budgets of 5%, 10%, 20%, and 30% of routes, where Lift@b is defined as:

$$\mathrm{Lift@}b = \frac{\mathrm{Precision@}b}{\pi_{\mathrm{target}}} \tag{4}$$

with b denoting the review budget expressed as a fraction of routes. Because lift is not prevalence-free, the two transfer directions are not ranked by raw lift. Where a prevalence-free comparison of discrimination is required, AUC is used. Tied scores at a budget boundary were allocated fractionally; consequently, the constant-score GHA policy equals the random-order expectation and has lift equal to 1. Random ordering gives Precision@b = π_{target}, and perfect ranking gives min(1, π_{target}/b). Because every budget evaluated is below both target prevalences, the perfect-ranking ceiling of Precision@b is 1 and the maximum attainable lift is $1/\pi_{target}$, that is, 1.352 into Denmark and 2.518 into Norway. Precision@Budget is therefore reported alongside both references.

Unless otherwise stated, CIs are 95% percentile intervals based on 2000 bootstrap replicates. For the paired analyses, point estimates are calculated from pooled test predictions within each repetition and then averaged over the five repetitions. A repeat-aware paired driver-cluster bootstrap resamples 317 driver clusters together with the five repetitions, keeping the Known and Unseen conditions locked to the same resampled routes. This procedure is used for the identity-channel, Experience, and gradient-boosting comparisons.

For graph-blocked evaluation, a driver-cluster bootstrap computes the metrics within each fold and then averages them over the five folds.

For cross-country transfer, the original target-country driver-cluster bootstrap resamples the target-country drivers (175 for Norway → Denmark and 144 for Denmark → Norway) while holding the fitted source model fixed. Feature-set differences are computed as paired contrasts within each replicate using the same resampled drivers for every specification. This avoids inferring a difference merely because one interval excludes a reference value and another does not [29].

A source-refit bootstrap with 500 replicates additionally captures training-side variability. In each replicate, source-country drivers are resampled as clusters, each specification is refitted at the λ selected on the full source data, and target-country drivers are resampled independently. These intervals do not include uncertainty from re-selecting λ. Cross-country comparisons that include both training-side and test-side variability use this procedure. For the EM-adjusted probabilities, the prior is re-estimated within each target-country bootstrap replicate, with the fitted source model held fixed.

Because rolling-origin and LDO partitions share observations, independence-based *p*-values across folds are not computed [30].

All model-fitting and split randomisation used fixed seed 20260905, and bootstrap post-processing used seed 20260906.

Model fitting and prediction were performed in MATLAB R2025b (The MathWorks, Inc., Natick, MA, USA). Community detection used the Louvain implementation in NetworkX 3.6.1 (Python 3.13; seeds 0–4). Bootstrap post-processing for the Experience, gradient-boosting, and graph-blocked analyses was performed in Python using NumPy 2.5.3 and the saved MATLAB predictions, with the resampling schemes described above and seed 20260906.

3. Results

3.1. Cohort and Entity-Overlap Audit

The realised overlap pattern summarised in Table 3 and Figure 1b sets the scope of the predictive results. Under the reference temporal regime, every test route belonged to a driver observed in training, and 96.8% of test stops occurred at an address observed in training. Even under LDO evaluation, 92.9–96.5% of test stops remained linked to addresses observed in training. This persistence reflects the connected address structure: in Norway, 99.1% of routes and 98.8% of addresses fell within a single connected component, while the three largest Danish components covered 99.8% of routes. Blocking whole communities of the address co-occurrence graph reduced this overlap: under leave-community-out evaluation, 22.9–37.1% of test stops occurred at addresses seen in training, and 41.4–66.2% of test routes belonged to drivers seen in training. By contrast, both entity-overlap rates were zero in cross-country transfer, the only evaluated regime in which drivers and addresses were jointly unseen.

3.2. Temporal and Leave-Driver-Out Results

On the reference single chronological split (17,498 training-plus-validation routes; 1243 test routes), adding identity features improved both outcomes. For classification, AUC increased from 0.6833 with Structure to 0.8364 with Structure + Driver + Address. For RQS regression, the corresponding MSE reduction relative to Structure was 0.01148. No CI is reported for this descriptive single-split result.

The strict-cohort evaluations showed a different pattern by identity channel. Across the five rolling origins, mean AUC increased from 0.7165 for Structure to 0.8302 for Structure + Driver + Address. Under leave-driver-out evaluation, adding Driver to Structure changed

mean AUC only from 0.7240 to 0.7246, whereas adding Address increased it to 0.7555. The full model reached 0.7554, essentially the same as Structure + Address, indicating that Driver added no meaningful discrimination when the test drivers were unseen.

Fold-to-fold variation is summarised in Table 4. Across the five rolling origins, the AUC standard deviation was 0.022 for Structure and 0.010–0.015 for the identity-bearing feature sets, and every identity-bearing set exceeded Structure at all five origins. Across the 25 LDO folds, the AUC standard deviation was 0.031–0.035. Structure + Address exceeded Structure in all 25 folds (mean difference 0.032, range 0.004–0.066), whereas the Structure + Driver difference averaged 0.0006 (range −0.0013 to 0.0054) and was positive in 10 of 25 folds. Because the folds share training data, these are descriptive measures rather than inferential intervals.

Table 4. Fold-level dispersion of AUC and MSE across rolling origins and leave-driver-out folds.

Regime	Feature Set	AUC Mean	AUC SD	AUC Range	MSE Mean	MSE SD	MSE Range
Rolling origin (5)	GHA	0.5000	0.0000	0.5000–0.5000	0.03496	0.00064	0.03436–0.03605
	Structure	0.7165	0.0219	0.6891–0.7442	0.03237	0.00071	0.03142–0.03321
	Structure + Driver	0.8143	0.0101	0.8001–0.8278	0.02381	0.00107	0.02254–0.02519
	Structure + Address	0.8180	0.0149	0.7973–0.8320	0.02253	0.00096	0.02144–0.02355
	Structure + Driver + Address	0.8302	0.0121	0.8114–0.8411	0.02180	0.00079	0.02114–0.02284
Leave-driver-out (25)	GHA	0.5000	0.0000	0.5000–0.5000	0.03590	0.00411	0.02761–0.04650
	Structure	0.7240	0.0307	0.6794–0.7840	0.03306	0.00359	0.02718–0.04237
	Structure + Driver	0.7246	0.0312	0.6794–0.7848	0.03309	0.00367	0.02702–0.04302
	Structure + Address	0.7555	0.0345	0.6836–0.8174	0.02821	0.00319	0.02258–0.03693
	Structure + Driver + Address	0.7554	0.0348	0.6870–0.8158	0.02831	0.00314	0.02275–0.03693

Note: SD and range are descriptive; folds share training data, so no inferential interval is reported.

Adding the two level features rejected by the invariance audit changed the rolling-origin AUC by −0.00005 on average for Structure (range −0.00023 to +0.00001 across origins) and by −0.00006 for Structure + Driver + Address (range −0.00020 to +0.00008), and every MSE change was below 0.00004 in absolute value. In this model class, their exclusion therefore did not remove an apparent gain; it is retained because their values depend on the outcome-informed time origin.

3.3. Paired Driver-Holdout Attenuation

Table 5 reports four paired attenuation contrasts on identical test routes. Training size, D prevalence, and country mix are matched between the Known and Unseen conditions, although address overlap differs slightly (97.80% versus 95.08%) because removing the test drivers' routes also removes coverage of their service areas (Section 2.6.4). All eight 95% bootstrap intervals reported in Table 5 exclude zero.

Table 5. Paired known-versus-unseen driver attenuation in AUC gain and MSE reduction, with 95% bootstrap intervals.

Contrast	AUC Gain [95% CI]	MSE Reduction [95% CI]
Driver-gain attenuation	0.0844 [0.0600, 0.1089]	0.00644 [0.00269, 0.01016]
Address-gain attenuation	0.0450 [0.0299, 0.0620]	0.00301 [0.00101, 0.00511]
Combined-gain attenuation	0.0583 [0.0402, 0.0775]	0.00377 [0.00129, 0.00628]
Driver-increment attenuation	0.0132 [0.0074, 0.0189]	0.00076 [0.00014, 0.00131]

In the Known condition, adding driver identity to Structure increased AUC by 0.0837 [0.0587, 0.1085], compared with −0.0007 [−0.0038, 0.0023] in the Unseen condition. The directly estimated Known–Unseen attenuation was 0.0844 [0.0600, 0.1089]. For RQS regression with unseen drivers, the MSE reduction from adding Driver to Structure + Address was −0.00039 [−0.00075, −0.00006], indicating a small but supported increase in squared error. Under indicator coding, an unseen driver receives an all-zero Driver block, so the Unseen-condition Driver gain cannot carry driver-specific information; the value of −0.0007 reflects only the re-estimation of the intercept and structural coefficients when the Driver block is included in training. The attenuation of 0.0844 is therefore numerically close to the Known-condition gain and measures how much of the in-operation driver gain depends on the driver having been observed in training.

Adding address identity to Structure increased AUC by 0.0876 [0.0637, 0.1116] in the Known condition and 0.0426 [0.0270, 0.0589] in the Unseen condition. The directly estimated attenuation was 0.0450 [0.0299, 0.0620]. The Unseen gain was 48.6% of the Known-condition point estimate; this ratio is reported descriptively, without a CI.

The combined identity gain over Structure was 0.1002 [0.0761, 0.1252] in the Known condition and 0.0419 [0.0271, 0.0576] in the Unseen condition. For unseen drivers, the Driver increment over Address was −0.0007 [−0.0033, 0.0019] on AUC and −0.00039 [−0.00075, −0.00006] on MSE. For known drivers, the corresponding AUC increment was 0.0125 [0.0065, 0.0185], whereas the corresponding MSE reduction was 0.00037 [−0.00014, 0.00082] and its interval included zero. The driver-increment AUC attenuation was 0.0132 [0.0074, 0.0189].

Figure 2 summarises the four AUC attenuation contrasts for the ridge models.

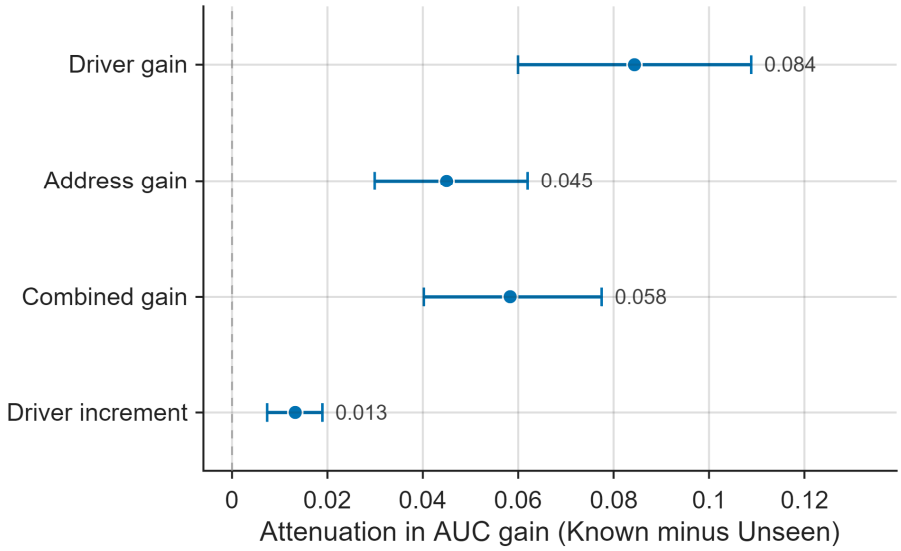


Figure 2. Known–Unseen attenuation in driver, address, combined, and incremental AUC gains; error bars show 95% paired driver-cluster bootstrap intervals.

Table 6 reports the Experience arm. Adding Experience to Structure changed AUC by 0.0015 [−0.0002, 0.0033] in the Known condition and 0.0018 [−0.0006, 0.0042] in the Unseen condition, and the Known–Unseen difference was −0.0003 [−0.0015, 0.0009]. Added to Structure + Address, Experience changed AUC by 0.0004 [−0.0006, 0.0013] for known drivers and 0.0017 [0.0002, 0.0035] for unseen drivers. The MSE reductions were below 0.0002 in every case. Experience is available for drivers excluded from model training, but its gain over Structure was small and its intervals included zero in both conditions; the Known–Unseen comparison does not establish that the two gains are equal. A small positive gain appeared only when Experience was added to the Address model for drivers not used

in training. Experience did not approach the driver-identity gain of 0.0837 observed for known drivers.

Table 6. Experience arm in the paired design: gains with 95% repeat-aware paired driver-cluster bootstrap intervals.

Contrast	AUC Gain [95% CI]	MSE Reduction [95% CI]
Experience gain, Known	0.0015 [−0.0002, 0.0033]	0.00011 [−0.00001, 0.00024]
Experience gain, Unseen	0.0018 [−0.0006, 0.0042]	0.00013 [−0.00002, 0.00030]
Experience-gain difference (Known–Unseen)	−0.0003 [−0.0015, 0.0009]	−0.00002 [−0.00012, 0.00005]
Experience increment over Address, Known	0.0004 [−0.0006, 0.0013]	0.000065 [0.0000002, 0.000134]
Experience increment over Address, Unseen	0.0017 [0.0002, 0.0035]	0.000086 [−0.000003, 0.000174]
Driver-identity gain, Known (for comparison)	0.0837 [0.0587, 0.1085]	0.00660

Note: The last row repeats, for comparison, the Known-condition driver-identity gain reported in Section 3.3; its MSE reduction is shown as a point estimate.

Table 7 reports the nonlinear check. On the 12 Structure features, gradient-boosted trees exceeded the ridge Structure model by 0.066 [0.046, 0.088] AUC in the Known condition and 0.050 [0.032, 0.070] in the Unseen condition. Over this nonlinear baseline, target-encoded driver identity added 0.027 [0.010, 0.043] AUC for known drivers, compared with 0.0837 for driver indicators over the ridge baseline, and reduced AUC by 0.065 [0.038, 0.092] for drivers not used in training, giving an attenuation of 0.091 [0.064, 0.119]. Target-encoded address identity added 0.017 [0.005, 0.031] for known drivers and −0.007 [−0.024, 0.011] for drivers not used in training, so the address increment observed with ridge indicators was not reproduced in the tested GBM specification with target encoding. Because this specification changed both the model and the representation of identity, the result does not show that nonlinear interactions alone absorbed the address information. For drivers not used in training, Structure + AddressTE with boosting and Structure + Address with ridge regression had almost the same AUC (difference 0.001 [−0.010, 0.012]).

Table 7. Gradient-boosted trees in the paired design: AUC gain and MSE reduction with 95% repeat-aware paired driver-cluster bootstrap intervals.

Contrast	AUC Gain [95% CI]	MSE Reduction [95% CI]
GBM Structure–ridge Structure, Known	0.066 [0.046, 0.088]	0.0083 [0.0055, 0.0115]
GBM Structure–ridge Structure, Unseen	0.050 [0.032, 0.070]	0.0073 [0.0050, 0.0098]
Driver TE over GBM Structure, Known	0.027 [0.010, 0.043]	0.0010 [−0.0002, 0.0021]
Driver TE over GBM Structure, Unseen	−0.065 [−0.092, −0.038]	−0.0057 [−0.0085, −0.0032]
Driver TE attenuation (Known–Unseen)	0.091 [0.064, 0.119]	0.0066 [0.0039, 0.0099]
Address TE over GBM Structure, Known	0.017 [0.005, 0.031]	0.0009 [0.0001, 0.0017]
Address TE over GBM Structure, Unseen	−0.007 [−0.024, 0.011]	−0.0006 [−0.0018, 0.0005]
Address TE attenuation (Known–Unseen)	0.024 [0.011, 0.037]	0.0015 [0.0004, 0.0028]
Driver TE + Address TE over GBM Structure, Known	0.034 [0.019, 0.051]	0.0018 [0.0008, 0.0028]
Driver TE + Address TE over GBM Structure, Unseen	−0.010 [−0.028, 0.009]	−0.0017 [−0.0033, −0.0003]

Note: TE = out-of-fold target encoding. Entities absent from the training data receive the training mean.

Mean predicted deviation probabilities of the boosted models ranged from 0.562 to 0.638, compared with an observed deviation rate of approximately 0.56. Higher discrimination did not consistently translate into better probability scores (Table S1).

3.4. Response to Address Suppression

Figure 3 traces the Address gain as the training address vocabulary is progressively suppressed. Gains declined in both driver conditions and on both outcomes.

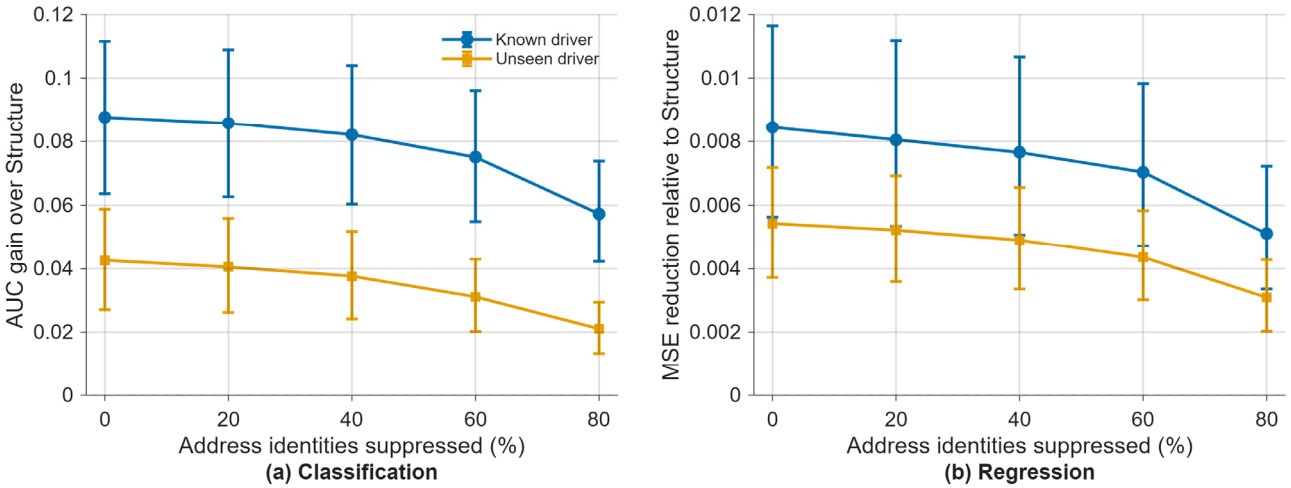


Figure 3. Address-identity gains under nested suppression for known and unseen drivers; error bars show 95% paired driver-cluster bootstrap intervals.

On AUC, the Known-condition gain fell from 0.0876 [0.0637, 0.1116] at 0% suppression to 0.0573 [0.0423, 0.0739] at 80%, while the Unseen-condition gain fell from 0.0426 [0.0270, 0.0589] to 0.0211 [0.0133, 0.0294]; the regression panel showed the same pattern on MSE. The decline was markedly sub-linear: suppressing 80% of the training address vocabulary removed about 35% of the Known-condition gain and about 50% of the Unseen-condition gain. At 80% suppression, the residual Address increment over Driver was 0.0196 [0.0128, 0.0270] AUC for unseen drivers, compared with 0.0049 [−0.0016, 0.0119] for known drivers.

The sub-linear decline should be read against the number of addresses that remain visible to the model. Test routes contained a median of 10 distinct non-depot addresses (interquartile range (IQR) 7–14). An address is counted as visible if it is not suppressed and appears in the training data. In the Unseen condition, the median number of visible addresses per test route was 10 (IQR 6–14), 8 (5–11), 6 (3–8), 4 (2–6), and 2 (1–3) at 0%, 20%, 40%, 60%, and 80% suppression, and 81.5% of test routes retained at least one visible address at 80%. In the Known condition, the medians were the same (10, 8, 6, 4, and 2), 82.6% of test routes retained at least one visible address at 80%, and the means were slightly higher (10.29 versus 9.99 at 0% and 2.03 versus 1.97 at 80%) because its training data include the earlier routes of the held-out drivers. Random suppression therefore left most routes with a few visible addresses.

3.5. Leave-Community-Out Results

Table 8 reports the leave-community-out evaluation. Averaged over the five folds, AUC was 0.725 [0.695, 0.749] for Structure, 0.744 [0.715, 0.768] for Structure + Driver, 0.723 [0.694, 0.748] for Structure + Address, and 0.736 [0.707, 0.760] for Structure + Driver + Address. The Address gain over Structure was −0.002 [−0.008, 0.005], compared with a mean of 0.032 under LDO evaluation, in which 92.9–96.5% of test stops occurred at addresses seen in training. The Driver gain was 0.020 [0.008, 0.033], consistent with 41.4–66.2% of test routes belonging to drivers seen in training. Among the 3576 test routes with no address seen in training, the Address gain was −0.0002 [−0.006, 0.006]; for these routes, the address block is all zero at test time, so this value reflects only re-estimated structural coefficients. In fold 1, validation blocked by community selected $\lambda = 100$ for regression and $\lambda = 1$ for classification in every feature set, so the identity contributions in that fold were close to zero; excluding fold 1, the mean Address gain was −0.002. The address-identity gain observed under LDO evaluation was therefore not reproduced under

graph blocking with reduced address overlap. Because this split also changed driver overlap and route composition, the difference is not attributed to the address channel alone.

Table 8. Leave-community-out evaluation by fold.

Fold	Test Routes	D Rate	Address Seen *	Driver Seen	AUC S	AUC S + D	AUC S + A	AUC S + D + A
1	3842	0.582	37.1%	63.5%	0.709	0.709	0.709	0.709
2	3711	0.584	29.7%	41.4%	0.794	0.815	0.771	0.792
3	3700	0.507	28.0%	66.2%	0.742	0.769	0.743	0.744
4	3699	0.652	32.4%	62.8%	0.716	0.723	0.726	0.741
5	3771	0.641	22.9%	60.8%	0.663	0.705	0.664	0.693
Mean [95% CI]	18,723	0.593			0.725 [0.695, 0.749]	0.744 [0.715, 0.768]	0.723 [0.694, 0.748]	0.736 [0.707, 0.760]

Note: * Percentage of test stops at addresses seen in training. D rate equals the PR-AUC baseline of each fold. S = Structure, D = Driver, A = Address. Intervals: driver-cluster bootstrap (2000 replicates) of the mean over folds.

3.6. Cross-Country Transfer Results

The cross-country analysis then examined the setting in which both identity channels were unavailable. Table 9 reports classification performance in both directions. Because target-country deviation prevalence differed substantially between Norway (0.3972) and Denmark (0.7398), direction-specific PR-AUC baselines and BSS references are reported.

Table 9. Cross-country classification performance and BSS values before and after oracle prior adjustment.

Direction	Feature Set	AUC [95% CI]	AUC, Source-Refit CI	PR-AUC [95% CI]	BSS Original	BSS Oracle	BSS EM [95% CI]
Norway → Denmark	GHA	0.5000	—	0.7398 [0.7036, 0.7729]	−0.6095	0.0000	—
	Structure	0.5822 [0.5443, 0.6219]	[0.5413, 0.6223]	0.7933 [0.7555, 0.8294]	−0.3179	−0.0061	−0.3518 [−0.4184, −0.2937]
	Structure + Driver	0.5883 [0.5517, 0.6261]	[0.5485, 0.6241]	0.8008 [0.7645, 0.8346]	−0.3832	0.0064	−0.3518 [−0.4184, −0.2937]
	Structure + Address	0.5857 [0.5486, 0.6238]	[0.5450, 0.6241]	0.7979 [0.7617, 0.8327]	−0.2719	−0.0119	−0.3518 [−0.4184, −0.2937]
	Structure + Driver + Address	0.5859 [0.5488, 0.6240]	[0.5457, 0.6240]	0.7986 [0.7623, 0.8332]	−0.2979	−0.0062	−0.3518 [−0.4184, −0.2937]
Denmark → Norway	GHA	0.5000	—	0.3972 [0.3481, 0.4454]	−0.4901	0.0000	—
	Structure	0.5750 [0.5366, 0.6160]	[0.5136, 0.6207]	0.4600 [0.3958, 0.5289]	−0.4775	−0.0440	−0.0454 [−0.2461, −0.0086]
	Structure + Driver	0.6419 [0.6026, 0.6815]	[0.5879, 0.6838]	0.5337 [0.4622, 0.6068]	−0.3071	0.0306	−0.6590 [−0.7683, −0.2497]
	Structure + Address	0.6441 [0.6059, 0.6829]	[0.5906, 0.6798]	0.5385 [0.4660, 0.6144]	−0.3111	0.0295	−0.6590 [−0.7494, −0.1787]
	Structure + Driver + Address	0.6448 [0.6064, 0.6837]	[0.5923, 0.6808]	0.5393 [0.4665, 0.6154]	−0.3182	0.0319	−0.4936 [−0.7150, −0.0862]

Note: The PR-AUC baseline equals the target-country prevalence, whereas GHA forecasts the source-country training prevalence. Unless labelled otherwise, intervals are 95% target-country driver-cluster bootstrap intervals that hold the fitted source model fixed; the source-refit column also resamples and refits on source-country drivers (500 replicates). For the EM-adjusted BSS, the prior was re-estimated in each replicate. Intervals for the original and oracle-adjusted BSS are shown in Figure 4; every oracle-adjusted interval for a trained feature set includes zero. Target-country identity blocks are all zero, so the rows labelled Structure + Driver, Structure + Address and Structure + Driver + Address denote specifications whose source-country fit included identity effects.

All four trained feature sets had AUC values above 0.5 in both directions, and every 95% interval excluded 0.5 (Table 9); the lowest lower limits were 0.544 for Structure into Denmark and 0.537 for Structure into Norway. These intervals hold the fitted source model fixed (Section 2.7). When source-country drivers were also resampled and the models refitted, every AUC interval still excluded 0.5, but the lower limit for Structure into Norway

fell to 0.514. The discrimination achieved was therefore modest and close to chance for Structure into Norway. PR-AUC point estimates were above the target-country positive-class baseline in both directions. GHA remained the constant-score reference, with an AUC of 0.5 and PR-AUC equal to target prevalence. The full Structure + Driver + Address model reached AUC values of 0.6448 into Norway and 0.5859 into Denmark. Models fitted with identity blocks showed the larger AUC difference relative to Structure into Norway, where AUC rose from 0.5750 for Structure to 0.6448 for the full model; the corresponding difference into Denmark was from 0.5822 to 0.5859.

The unadjusted BSS was negative for every trained feature set in both directions. A constant forecast set to the target-country deviation prevalence therefore achieved a lower BS than the transferred probability forecasts. For the Structure model transferred from Norway into Denmark, mean predicted probability was 0.5020 against a target prevalence of 0.7398, ECE was 0.2378, and the calibration slope was 0.4869. Together, these values show substantial underprediction and poor probability calibration in this direction.

Oracle prior adjustment reduced ECE for the Structure model transferred from Norway into Denmark from 0.2378 to 0.0616. Every 95% interval for the oracle-adjusted BSS nevertheless included zero in both directions; no trained model therefore had an interval wholly above the zero-skill comparator. Original and oracle-adjusted BSS values are compared in Figure 4.

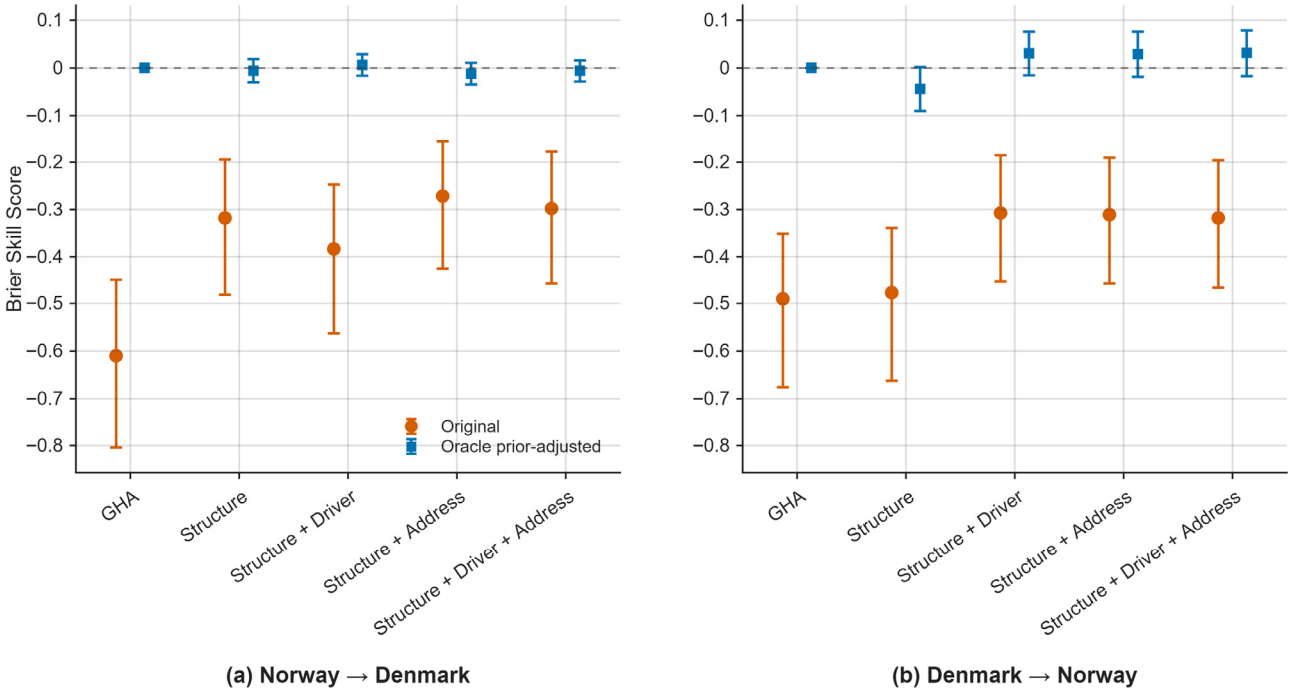


Figure 4. Cross-country BSS values before and after oracle prior adjustment; points show estimates and error bars show 95% target-country driver-cluster bootstrap intervals.

The EM-estimated prior was not consistently reliable (Table 9). Into Denmark, the estimate converged to the upper boundary for all four feature sets (realised prevalence 0.740), so the adjusted probabilities approached 1 and BSS became −0.352, the value obtained by predicting a deviation for every route. Into Norway, the Structure model gave an estimated prior of 0.343 [0.166, 0.536] (realised 0.397), and its BSS improved from −0.478 to −0.045 [−0.246, −0.009], close to the oracle-adjusted value but still negative. The identity-fitted specifications gave estimated priors of 0.000–0.044 and BSS values of −0.659 to −0.494, lower than without adjustment.

Figure 5 presents 10-bin reliability diagrams for GHA, Structure and Structure + Driver + Address using original and oracle-prior-adjusted probabilities. For the original Structure + Driver + Address probabilities transferred into Denmark, observed deviation rates increased from 0.606 in the lowest bin to 0.859 in the highest, while mean predicted probabilities increased from 0.273 to 0.767. The probability level was incorrect, but the ordering retained information. Into Norway, the direction of the average level error reversed, and the observed rates in the first three bins were 0.345, 0.212, and 0.268. The persistence of ranking information despite poor probability calibration motivates the fixed-budget analysis in Section 3.7.

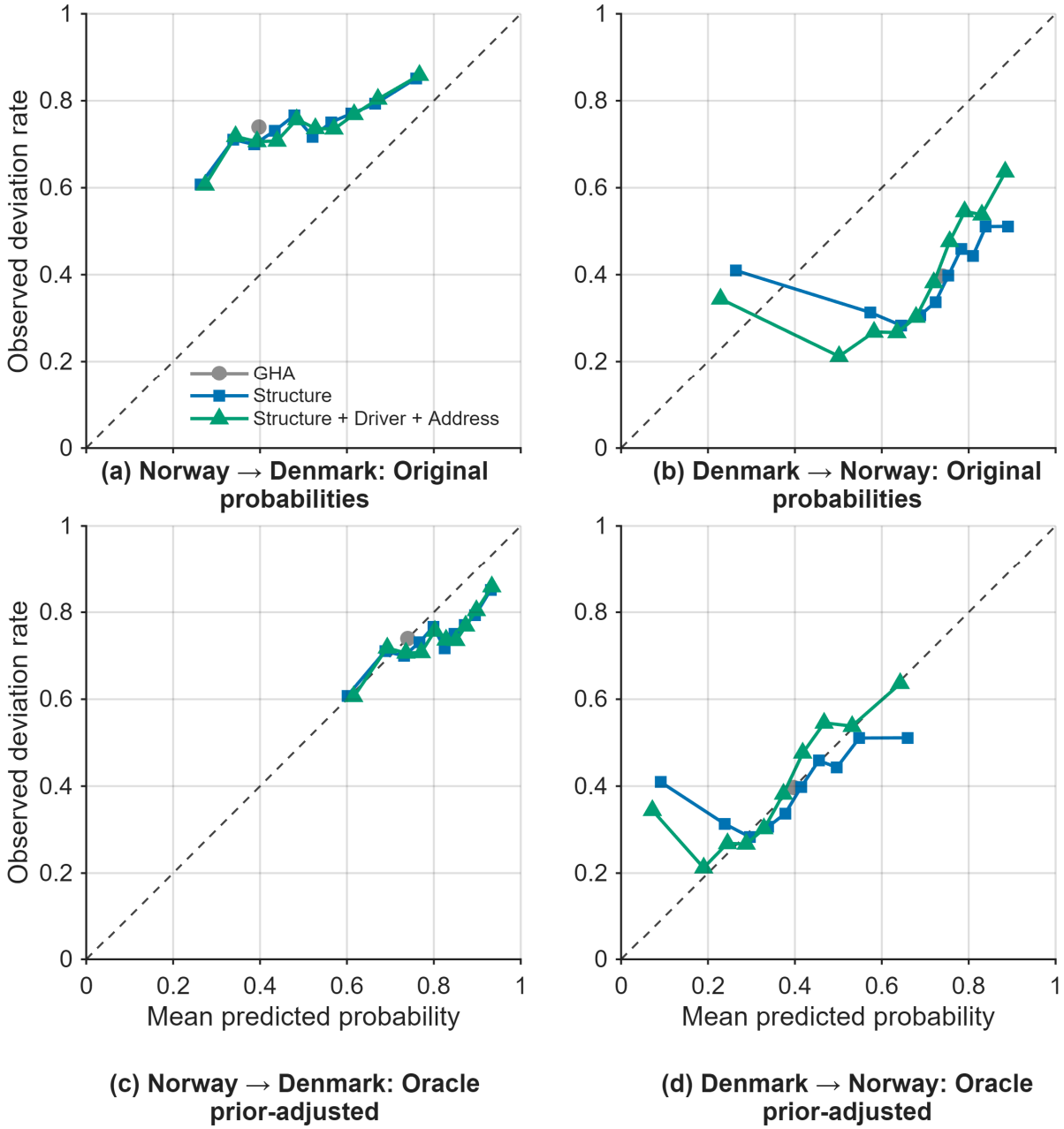


Figure 5. Reliability diagrams for GHA, Structure, and Structure + Driver + Address using original and oracle prior-adjusted cross-country probabilities; the diagonal denotes perfect calibration.

For the separate RQS regression target, transfer was asymmetric. Into Denmark, Structure + Driver + Address produced an MSE reduction relative to GHA of 0.0034 [0.0023, 0.0045]. Into Norway, the corresponding MSE reduction was −0.0089 [−0.0124, −0.0059],

while Structure-only had the largest negative MSE reduction, −0.0375 [−0.0490, −0.0276]. All four trained feature sets transferred into Norway had negative MSE reductions relative to GHA, with 95% intervals entirely below zero.

3.7. Operational Route-Review Policy

The transfer results translate into the decision a dispatcher faces: with capacity to review a fixed fraction of tomorrow's planned routes, which ones should be reviewed? Figure 6 shows recall and lift for every feature set across review budgets of 5%, 10%, 20%, and 30%.

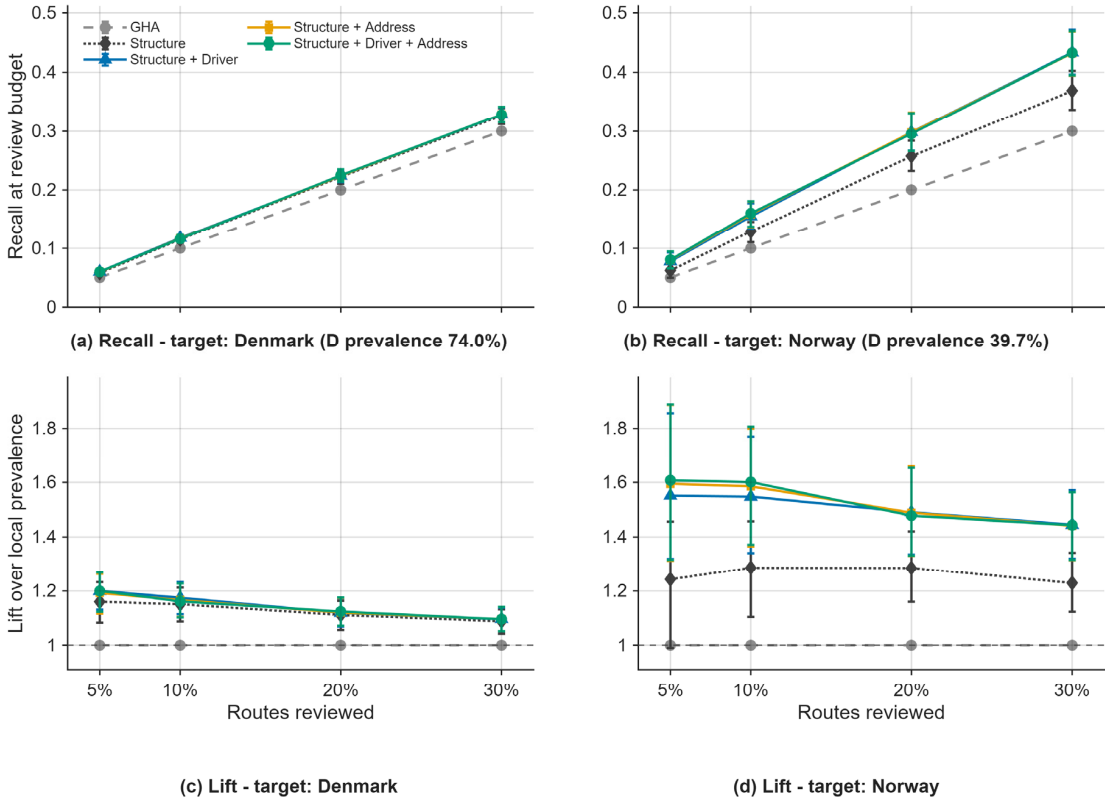


Figure 6. Recall and lift of cross-country route-review policies across review budgets; lift is interpreted relative to the target-country prevalence.

For the full Structure + Driver + Address model, transferring from Norway into Denmark yields a recall of 0.0601 and a lift of 1.200 at a 5% review budget, and a recall of 0.1161 and a lift of 1.161 at 10%. Transferring from Denmark into Norway, the same model yields a recall of 0.0804 and a lift of 1.607 at a 5% budget, and a recall of 0.1602 and a lift of 1.601 at 10%. Because lift is precision divided by the target-country prevalence, these lift values are interpreted only within a direction and are not compared between the two markets.

Table 10 reports Precision@Budget with 95% intervals. At the 5% budget, the full model reached a precision of 0.888 [0.831, 0.930] into Denmark and 0.638 [0.505, 0.770] into Norway against random baselines of 0.740 and 0.397 and a perfect-ranking ceiling of 1 in both directions; it therefore realised about 57% and 40%, respectively, of the headroom above random ordering. Comparison with random ordering uses lift, which is computed within each replicate and so accounts for the replicate-specific prevalence. With the source model held fixed, the Structure lift into Norway was 1.243 [0.990, 1.456] at the 5% budget, and its intervals excluded 1 at 10–30%. With source refitting, its interval included 1 at the 5% and 10% budgets, had a lower limit of 1.0001 at 20%, which is borderline, and excluded

1 at 30% (1.229 [1.063, 1.400]). Every other lift interval, for every specification into Denmark and for the identity-fitted specifications into Norway, excluded 1 under both procedures.

Table 10. Precision@Budget [95% CI] under cross-country transfer (random baseline = target prevalence; perfect-ranking ceiling = 1 at every budget).

Direction	Feature Set	5%	10%	20%	30%
Norway → Denmark (random 0.740)	Structure	0.858 [0.799, 0.912]	0.852 [0.798, 0.898]	0.822 [0.772, 0.864]	0.805 [0.759, 0.844]
	Structure + Driver	0.888 [0.838, 0.931]	0.869 [0.820, 0.911]	0.828 [0.783, 0.872]	0.810 [0.767, 0.849]
	Structure + Address	0.882 [0.826, 0.928]	0.863 [0.814, 0.907]	0.829 [0.784, 0.873]	0.811 [0.767, 0.851]
	Structure + Driver + Address	0.888 [0.831, 0.930]	0.859 [0.811, 0.907]	0.831 [0.784, 0.874]	0.810 [0.767, 0.850]
Denmark → Norway (random 0.397)	Structure	0.494 [0.381, 0.592]	0.511 [0.420, 0.601]	0.511 [0.432, 0.595]	0.488 [0.416, 0.565]
	Structure + Driver	0.616 [0.508, 0.742]	0.615 [0.518, 0.718]	0.592 [0.509, 0.676]	0.574 [0.496, 0.647]
	Structure + Address	0.633 [0.502, 0.765]	0.630 [0.522, 0.732]	0.591 [0.505, 0.679]	0.572 [0.494, 0.647]
	Structure + Driver + Address	0.638 [0.505, 0.770]	0.636 [0.526, 0.737]	0.587 [0.504, 0.678]	0.573 [0.495, 0.647]

Paired contrasts with Structure, computed within each replicate of the source-refit bootstrap [29], showed a direction-specific pattern (Table 11). Into Norway, every identity-fitted specification ranked routes better than Structure: AUC differences were 0.067–0.070 (lower limits 0.022–0.027) and PR-AUC differences were 0.074–0.079. Precision differences excluded zero at the 20% and 30% budgets for all three specifications and at 10% for Structure + Address and Structure + Driver + Address, but not at the 5% budget (differences 0.122–0.145, lower limits −0.020 to −0.007). Into Denmark, no AUC, PR-AUC, or precision contrast excluded zero (AUC differences 0.004–0.006). With the source model held fixed, the intervals were narrower and further contrasts excluded zero; these are not used for inference.

Table 11. Paired contrasts with Structure under cross-country transfer (difference [95% source-refit bootstrap interval], 500 replicates).

(**A**) Discrimination

Direction	Specification	AUC Difference	PR-AUC Difference
Denmark → Norway	Structure + Driver	0.0669 [0.0218, 0.1223]	0.0737 [0.0130, 0.1347]
	Structure + Address	0.0691 [0.0255, 0.1238]	0.0785 [0.0184, 0.1334]
	Structure + Driver + Address	0.0699 [0.0271, 0.1274]	0.0793 [0.0198, 0.1380]
Norway → Denmark	Structure + Driver	0.0062 [−0.0043, 0.0215]	0.0075 [−0.0039, 0.0256]
	Structure + Address	0.0035 [−0.0072, 0.0192]	0.0046 [−0.0062, 0.0237]
	Structure + Driver + Address	0.0038 [−0.0072, 0.0210]	0.0053 [−0.0053, 0.0253]

(**B**) Precision differences

Direction	Specification	Precision@5%	Precision@10%	Precision@20%	Precision@30%
Denmark → Norway	Structure + Driver	0.1222 [−0.0198, 0.3864]	0.1035 [−0.0092, 0.3008]	0.0811 [0.0035, 0.1846]	0.0857 [0.0162, 0.1473]
	Structure + Address	0.1397 [−0.0085, 0.3832]	0.1185 [0.0062, 0.2913]	0.0805 [0.0018, 0.1924]	0.0840 [0.0189, 0.1418]
	Structure + Driver + Address	0.1446 [−0.0073, 0.4010]	0.1247 [0.0078, 0.3046]	0.0761 [0.0003, 0.1945]	0.0849 [0.0206, 0.1473]
Norway → Denmark	Structure + Driver	0.0299 [−0.0194, 0.1074]	0.0177 [−0.0143, 0.0815]	0.0056 [−0.0104, 0.0422]	0.0056 [−0.0056, 0.0259]
	Structure + Address	0.0243 [−0.0241, 0.1050]	0.0112 [−0.0228, 0.0708]	0.0065 [−0.0097, 0.0426]	0.0059 [−0.0067, 0.0254]
	Structure + Driver + Address	0.0299 [−0.0219, 0.1078]	0.0075 [−0.0207, 0.0786]	0.0093 [−0.0092, 0.0436]	0.0056 [−0.0066, 0.0272]

Note: Positive values favour the identity-fitted specification. Lift is precision divided by the target-country prevalence.

4. Discussion

The findings form a consistent progression across increasingly demanding deployment populations. In the temporal evaluations, most test drivers and addresses remained represented in training, and models containing identity blocks improved discrimination. Once driver continuity was removed, the driver-specific benefit of identity indicators disappeared, while recurring addresses continued to support prediction in the ridge analysis with high training–test address overlap. That address increment was not reproduced under graph blocking or in the tested GBM specification with target encoding. Once both forms of continuity were removed in cross-country transfer, some ranking information remained, but reliable probability scoring did not. The regimes therefore provide answers to different deployment questions rather than competing estimates of a single model's accuracy. This distinction is consistent with guidance that validation blocks should reflect whether future observations arise from known or new individuals, groups, or locations [9], and with evidence from other prediction domains that subject-disjoint evaluation can materially alter reported performance [10].

The paired comparison identifies driver history as the main source of this regime sensitivity. Driver-identity AUC gain fell from 0.0837 when the test drivers had training history to −0.0007 when they did not, producing an attenuation of 0.0844 [0.0600, 0.1089]. Because the same test routes were used and training size, deviation prevalence and country mix were matched, this change is not explained by a different test-route population or training-set size. It follows directly from the representation: an unseen driver is assigned an all-zero identity block and therefore has no fitted driver-specific coefficient. The attenuation is therefore close to the Known-condition gain by construction, and the paired design measures how much of the driver gain depends on prior observation rather than whether driver-level information can transfer to drivers not used in training. The result does not imply that driver behaviour or tacit operational knowledge is unimportant. Rather, it shows that an indicator representation can use such history only for drivers already observed during fitting. This qualifies the deployment scope of courier-specific modelling [5] and parallels the inability of unobserved route transitions to carry learned information [11]. The small increase in regression error when Driver was added to Structure + Address for unseen drivers further shows that a high-dimensional identity block can add estimation cost after its transferable information has disappeared.

Causal-loop models of human behaviour in socio-technical systems map the performance-shaping factors that influence human performance [31]. Descriptors of this kind are defined for any driver, unlike identity indicators, and can therefore be evaluated for drivers not observed in training; Experience is examined in this way in Section 3.3. In the paired design, adding Experience to Structure changed AUC by 0.0015 for known drivers and 0.0018 for drivers not used in training, with both intervals including zero and a Known–Unseen difference of −0.0003 [−0.0015, 0.0009]. A small positive increment appeared when Experience was added to the Address model for unseen drivers. Recent route counts therefore did not reproduce the large Known-condition identity gain, although this result does not rule out other transferable driver descriptors.

The nonlinear check changes the size, but not the direction, of the driver result. Boosted trees on the Structure features raised AUC by 0.05–0.07 over the ridge Structure model, showing that the same Structure features supported higher discrimination under boosting. Target-encoded driver identity still added AUC for known drivers, whereas for drivers not used in training, whose encoding takes the training mean, it reduced AUC. The benefit of the encoding therefore did not carry over to such drivers under this specification; the result concerns the trained model and its handling of drivers absent from training, and it does not show that driver information in general impairs prediction.

Address identity followed a different pattern in the ridge analysis. Its AUC gain decreased from 0.0876 for known drivers to 0.0426 for unseen drivers but remained positive. For unseen drivers, the combined identity gain was 0.0419, almost identical to the Address gain alone, while the Driver increment over Address was −0.0007. For known drivers, Driver added 0.0125 AUC beyond Address. Persistent access conditions, customer constraints or local service patterns could support address-related prediction across drivers, consistent with earlier findings on location and planned-route information [1,2]. The paired comparison does not establish that mechanism: removing the test drivers' routes also reduced address coverage from 97.80% to 95.08%, so the address-gain attenuation reflects an entangled change in driver history and territory coverage.

Random suppression showed a gradual decline in the address channel. Removing 80% of the vocabulary reduced the Known-condition AUC gain by about 35% and the Unseen-condition gain by about half, while most test routes still retained a few visible addresses. This response is compatible with both information spread across addresses and a low-dimensional area-level factor represented through correlated addresses; it does not distinguish them. At 80% suppression, an Address increment over Driver was supported for unseen drivers but was not established for known drivers. Because the associated routes, outcomes, and structural features remained in training, suppression does not represent deployment in a new service area.

Graph-blocked evaluation provided a separate test with reduced address overlap. Only 22.9–37.1% of test stops occurred at addresses seen in training, and the Address gain was −0.002 [−0.008, 0.005], so the gain observed under LDO evaluation was not reproduced. This comparison also changed driver overlap and route composition. Moreover, the communities reflect co-service rather than geographical boundaries. The result therefore limits the portability of the observed address gain without isolating an address effect or establishing performance in a new geographical area.

The address increment observed with ridge indicators was also not reproduced in the tested GBM specification with target encoding for unseen drivers. That check changed both the model and the identity representation: the address encoding summarises a route's addresses in a single value rather than retaining separate indicators. It therefore does not identify which change accounts for the result. The observed address gain should be interpreted as conditional on overlap, model class, and identity representation, rather than as general evidence that address information cannot transfer.

Cross-country transfer completes the progression by removing both identity channels while also changing outcome prevalence from 0.3972 in Norway to 0.7398 in Denmark. All trained feature sets retained AUC point estimates above 0.5 and PR-AUC values above the target-country baseline in both directions, indicating that some ordering information survived. Probability performance did not: every unadjusted BSS was negative relative to a constant forecast fixed at target prevalence. For Structure transferred into Denmark, the gap between the mean prediction of 0.5020 and the observed prevalence of 0.7398 illustrates the scale of the level error. Oracle prior adjustment substantially reduced ECE in that example, but no oracle-adjusted BSS interval for a trained model lay wholly above zero. The prevalence shift therefore explains part of the failure, but correcting the class prior alone does not establish probability-level transfer.

The deployable EM estimate of the target prior was not consistently reliable: it recovered the Norwegian prior approximately for the Structure model but reached the boundary or moved away from the realised prevalence for the other specifications. Its assumptions cannot be verified from target data alone, and the low calibration slopes of the transferred probabilities are compatible with a departure from them.

The cross-country feature-set comparisons require an additional qualification. Target-country driver and address indicators were all zero because none of those entities occurred in source-country training. Consequently, the higher AUC of models fitted with identity blocks into Norway does not represent direct transfer of source-country identity coefficients. Any difference from Structure must arise from the way inclusion of those blocks altered estimation of the intercept and structural coefficients during source-country fitting. These comparisons are informative about complete model specifications, but they do not demonstrate that driver or address identities themselves transfer between markets. In this sense, the identity indicators act as fixed effects that absorb source-country heterogeneity, and the Structure + Driver + Address model applied to the target country is the structural component of a model fitted jointly with source-country identity effects.

This distinction between ordering and probability level explains why the route-review results are more encouraging than the BSS results. A fixed-budget policy uses the ranking of routes rather than treating predicted probabilities as literal risk estimates. Paired contrasts that include training-side variability showed that specifications fitted with source-country identity effects ranked routes better than Structure when transferring into Norway, whereas into Denmark, the differences were not distinguishable from zero; at the 5% budget, the advantage was not distinguishable from zero in either direction. Even the full model, however, captured only 6.0% of deviations into Denmark and 8.0% into Norway at the 5% budget, so the operational gain remained modest in absolute coverage. Lift is also bounded differently in the two markets: the maximum attainable lift is 1.352 into Denmark and 2.518 into Norway, so the full model's 5% lifts of 1.200 and 1.607 correspond to precisions of 0.888 and 0.638 and realise about 57% and 40% of the available headroom. Ranking may therefore support targeted review even when probability-based thresholds would be unreliable. That use remains provisional: performance must be verified against outcomes in the target market, and probability-based deployment would require target-specific recalibration. The asymmetric RQS regression results reinforce the same caution, since transfer improved on GHA into Denmark but deteriorated into Norway. Discrimination, calibration, continuous-score prediction, and policy value are related but distinct properties and should not be inferred from one another.

The broader implication is that realised entity overlap should be reported alongside predictive performance. A temporal split with recurring drivers and addresses is appropriate for continued operation within an established network; LDO evaluation addresses drivers excluded from model training who serve largely familiar locations; graph-blocked leave-community-out evaluation reduces address overlap while also changing driver overlap; and cross-country transfer addresses the setting in which both drivers and addresses are unseen. No single regime replaces the others. Their value lies in showing which parts of reported performance depend on operational continuity, an important distinction given the inconsistent evaluation conventions already noted across routing studies [19].

Several limitations define the scope of these conclusions. The evidence concerns one operator and one pair of national markets. The main models are ridge-regularised linear models using an unordered bag-of-addresses representation. The gradient-boosting check used out-of-fold target encodings rather than identity indicators, so it changed the model and the representation together, and its probabilities were not recalibrated. Sequence models, learned embeddings, and representations designed to share information across unseen entities may behave differently. The within-country evaluation with reduced address overlap was graph-blocked rather than spatial, because coordinates are not available in the public release; 22.9–37.1% of its test stops still occurred at addresses seen in training, and it reduced driver and address familiarity together. The paired conditions matched training size, deviation prevalence, and country mix, but they did not explicitly match the

complete calendar-time distribution of the training routes, and a small address-overlap imbalance remained. Finally, the experiments identify predictive channels rather than behavioural mechanisms. Broader validation should therefore include additional operators and markets, geographically blocked within-country tests where feasible, alternative representations with explicit unseen-entity handling, and prospective evaluation of recalibration and review policies.

5. Conclusions

Reported performance in last-mile route-deviation prediction is conditional on the deployment population represented by the evaluation design. In the paired comparison, the driver-identity AUC gain of 0.0837 for drivers with training history did not extend to drivers absent from training, as indicator coding implies, and a within-dataset experience descriptor available for those drivers added little.

Address identity retained a gain of 0.0426 for unseen drivers in the linear model, and random suppression of most of its vocabulary reduced but did not remove it. That gain was not reproduced under graph-blocked evaluation with reduced address overlap, and the address increment was not reproduced in the tested gradient-boosting specification with target encoding, whose structural baseline itself raised AUC by 0.05–0.07 over the linear one. The address result is therefore conditional on address overlap, the model, and the identity representation.

When both drivers and addresses were unseen in cross-country transfer, AUC remained modestly above chance, but every unadjusted Brier skill score was negative, and an EM estimate of the target prior did not provide a consistently reliable correction. Specifications fitted with source-country identity effects ranked routes better than structural features alone when transferring into Norway, whereas into Denmark the differences were not distinguishable from zero, and at a 5% review budget the advantage was not distinguishable from zero in either direction once training-side variability was included.

Within-operation evaluation remains valid for a continuing network, but it should be reported together with realised entity overlap and the intended deployment population. Cross-market use requires target-market validation, and probability-based deployment additionally requires recalibration. These conclusions are limited to one operator, two countries, and ridge-regularised linear models with a gradient-boosting check; broader evidence requires additional markets, geographically blocked evaluations, and models designed to handle unseen entities.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/asi9110219/s1, Table S1: Brier scores of the ridge and gradient-boosted specifications in the paired design, and Brier reduction with 95% repeat-aware paired driver-cluster bootstrap intervals.

Author Contributions: Conceptualization, S.A. and A.B.E.; methodology, S.A.; software, S.A.; validation, S.A.; formal analysis, S.A.; investigation, S.A. and A.B.E.; resources, S.A.; data curation, S.A.; writing—original draft preparation, S.A.; writing—review and editing, S.A. and A.B.E.; visualization, S.A.; supervision, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The source dataset analysed in this study is publicly available in Mendeley Data at https://doi.org/10.17632/kkwgfvmtxn.1. The original contributions presented in this study are included in the article and its Supplementary Materials. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Phillips, W. *Learning from Route Plan Deviation in Last-Mile Delivery*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2018.
2. Konovalenko, A.; Hvattum, L.M.; Iversen, K.A.H. Predicting Last-Mile Delivery Route Deviations Using Machine Learning. *Expert Syst. Appl.* **2026**, *298*, 129921. [CrossRef]
3. Stenmark, D. Leveraging Tacit Organizational Knowledge. *J. Manag. Inf. Syst.* **2000**, *17*, 9–24. [CrossRef]
4. Konovalenko, A.; Hvattum, L.M.; Msakni, M.K. Using Machine Learning to Identify Hidden Constraints in Vehicle Routing Problems. *Comput. Oper. Res.* **2025**, *179*, 107029. [CrossRef]
5. Pegado-Bardayo, A.; Lorenzo-Espejo, A.; Muñuzuri, J.; Onieva, L. A Predictive Framework for Last-Mile Delivery Routes Considering Couriers' Behavior Heterogeneity. *Comput. Ind. Eng.* **2024**, *198*, 110665. [CrossRef]
6. Dieter, P.; Caron, M.; Schryen, G. Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework. *Eur. J. Oper. Res.* **2023**, *311*, 283–300. [CrossRef]
7. Konovalenko, A.; Hvattum, L.M.; Iversen, K.A.H. Data Describing Planned Routes and Driven Routes from a Last-Mile Delivery Vehicle Routing Problem. *Data Brief* **2026**, *66*, 112744. [CrossRef] [PubMed]
8. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. [CrossRef] [PubMed]
9. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
10. Rehman, S.U.; Ali, A.; Khan, A.M.; Okpala, C. Human Activity Recognition: A Comparative Study of Validation Methods and Impact of Feature Extraction in Wearable Sensors. *Algorithms* **2024**, *17*, 556. [CrossRef]
11. Sun, H.J.; Arslan, O. A Contextual Framework for Learning Routing Experiences in Last-Mile Delivery. *Transp. Res. Part B Methodol.* **2025**, *194*, 103172. [CrossRef]
12. Merchán, D.; Arora, J.; Pachon, J.; Konduri, K.; Winkenbach, M.; Parks, S.; Noszek, J. 2021 Amazon Last Mile Routing Research Challenge: Data Set. *Transp. Sci.* **2024**, *58*, 8–11. [CrossRef]
13. Guo, X.; Mo, B.; Wang, Q. Amazon Last-Mile Delivery Trajectory Prediction Using Hierarchical TSP with Customized Cost Matrix. *arXiv* **2023**. [CrossRef]
14. Cook, W.; Held, S.; Helsgaun, K. Constrained Local Search for Last-Mile Routing. *Transp. Sci.* **2024**, *58*, 12–26. [CrossRef]
15. Rashidi, H.; Nourinejad, M.; Roorda, M. Generating Practical Last-Mile Delivery Routes Using a Data-Informed Insertion Heuristic. *Transp. Res. Part C Emerg. Technol.* **2025**, *179*, 105278. [CrossRef]
16. Özarık, S.S.; Da Costa, P.; Florio, A.M. Machine Learning for Data-Driven Last-Mile Delivery Optimization. *Transp. Sci.* **2024**, *58*, 27–44. [CrossRef]
17. Mo, B.; Wang, Q.; Guo, X.; Winkenbach, M.; Zhao, J. Predicting Drivers' Route Trajectories in Last-Mile Delivery Using a Pair-Wise Attention-Based Pointer Neural Network. *Transp. Res. Part E Logist. Transp. Rev.* **2023**, *175*, 103168. [CrossRef]
18. Ruiz-Fas, À.; Granell, C.; Ramos, J.F.; Huerta, J.; Trilles, S. A Zone-Based Training Approach for Last-Mile Routing Using Graph Neural Networks and Pointer Networks. *arXiv* **2026**. [CrossRef]
19. Zhou, F.; Lischka, A.; Kulcsár, B.; Wu, J.; Haghir Chehreghani, M.; Laporte, G. Learning for Routing: A Guided Review of Recent Developments and Future Directions. *Transp. Res. Part E Logist. Transp. Rev.* **2025**, *202*, 104278. [CrossRef]
20. Wu, L.; Wen, H.; Hu, H.; Mao, X.; Xia, Y.; Shan, E.; Zheng, J.; Lou, J.; Liang, Y.; Yang, L.; et al. LaDe: The First Comprehensive Last-Mile Express Dataset from Industry. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; ACM: Barcelona, Spain, 2024; pp. 5991–6002.
21. Hernandez, F.; Sotelo, R.; Forets, M. Optimization Algorithms for Adaptative Route Sequencing on Real-World Last-Mile Deliveries. *Ingenius* **2024**, *31*, 64–80. [CrossRef]
22. Konovalenko, A.; Hvattum, L.M.; Iversen, K.A.H. *Last-Mile Delivery Route Deviations Dataset: Planned vs. Actual Routes*; Elsevier: Amsterdam, The Netherlands, 2024. [CrossRef]
23. Sörensen, K. Distance Measures Based on the Edit Distance for Permutation-Type Representations. *J. Heuristics* **2007**, *13*, 35–47. [CrossRef]
24. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
25. Friedman, J.; Hastie, T.; Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). *Ann. Stat.* **2000**, *28*, 337–407. [CrossRef]
26. Micci-Barreca, D. A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems. *ACM SIGKDD Explor. Newsl.* **2001**, *3*, 27–32. [CrossRef]
27. Blondel, V.D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of Communities in Large Networks. *J. Stat. Mech.* **2008**, *2008*, P10008. [CrossRef]

28. Saerens, M.; Latinne, P.; Decaestecker, C. Adjusting the Outputs of a Classifier to New a *Priori* Probabilities: A Simple Procedure. *Neural Comput.* **2002**, *14*, 21–41. [CrossRef] [PubMed]
29. Gelman, A.; Stern, H. The Difference Between "Significant" and "Not Significant" Is Not Itself Statistically Significant. *Am. Stat.* **2006**, *60*, 328–331. [CrossRef]
30. Dietterich, T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]
31. Di Nardo, M.; Gallo, M.; Madonna, M.; Santillo, L.C. A Conceptual Model of Human Behaviour in Socio-Technical Systems. In *Intelligent Software Methodologies, Tools and Techniques*; Fujita, H., Guizzi, G., Eds.; Communications in Computer and Information Science; Springer International Publishing: Cham, Switzerland, 2015; Volume 532, pp. 598–609, ISBN 978-3-319-22688-0. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.