Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

Measuring Temporal Fluidity in Credit-Card Customer Personas: An Exposure-Aware Markov Framework with a Pre-Registered Audit of Predictive Value

Yong Hee Han [1], Bangwon Ko [2] and Sung-Seok Ko [3,*]

[1] Department of Entrepreneurship and Small Business, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Republic of Korea; amade@ssu.ac.kr
[2] Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-ro, Dongjak-gu, Seoul 06978, Republic of Korea; bko@ssu.ac.kr
[3] Department of Industrial Engineering, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
* Correspondence: ssko@konkuk.ac.kr

Abstract

Static e-commerce segmentation leaves category change in payment data unmeasured. We analyze 96.19 million credit-card transactions, including online-shopping categories but no channel flags, across 13 quarters for 229,586 customers, measuring transition exposure, set turnover, retention, and novelty. Temporal homogeneity was rejected ($\chi^2(132) = 57{,}222.2$), and a second-order Markov model outperformed a first-order model in held-out log loss (0.886 vs. 0.913). The hidden Markov benchmark matched within ± 0.005 log loss, but only two of five fold intervals excluded zero; the nonhomogeneous specification did not improve. A pre-registered audit found a mean-spending ΔR^2 of 0.0009 over the original baseline after recency correction; neither three primary decision targets nor the auxiliary adoption target reached the area under the receiver operating characteristic curve (AUC) threshold of ΔAUC 0.01 in any fold. One of three gates passed; its increment fell below the residual criterion with lagged category-change information. The re-analysis added nonlinear learners, recent-window and monthly resolutions, and spending-change and category-entropy targets; all increments stayed below 0.01. Equivalence tests placed 95% upper bounds at 0.0038, 0.0017, and 0.0032, bounding increments below materiality. We contribute an exposure-aware audit reporting this bound on payment-category data for e-commerce analytics.

Keywords: customer segmentation; longitudinal customer analytics; credit-card transactions; Markov models; hidden Markov model; pre-registration; decision utility; auditability

Academic Editor: Hua Pang

Received: 27 July 2026
Revised: 6 September 2026
Accepted: 17 September 2026
Published: 20 September 2026

1. Introduction

Customer segmentation commonly represents customers through a feature vector estimated in a fixed observation window [1,2]. Such summaries remain useful for describing a portfolio at one point in time, but they do not quantify whether the category set attached to one customer changes in the next period [3]. This omission matters when customer records are refreshed repeatedly, because a stable cross-sectional partition and a changing individual history answer different questions. Longitudinal analysis must, therefore, identify which observations form a valid transition before it summarizes a profile [4].

Segment-instability research treats customer segments as moving targets and requires a defined period pair for a temporal comparison [3].

Digital commerce makes that distinction visible in customer records. Omnichannel journeys distribute search, purchase, and service encounters across channels and periods [5,6]. Dynamic conversion behavior also shows that an observed relationship can change during the life of a customer model [7]. The present transaction schema includes mixed merchant categories, including online-shopping categories, but neither a transaction-level online or offline flag nor clickstream, search-log, or intent-level events. The study consequently evaluates temporal measurement in payment data connected to e-commerce analytics, rather than treating the data as an e-commerce-only panel.

This study uses 96.19 million credit-card transactions observed in 13 quarters. It constructs a primary cohort of 229,586 customers only when adjacent observations form a calendar-consecutive quarter pair. This gap-safe rule retains 926,871 observed transitions and excludes 40,717 nonconsecutive successive-quarter pairs. It prevents a missing interval from becoming an artificial behavioral transition.

Transition exposure is uneven in this cohort. Of the 229,586 customers with an eligible pair, 52,484 have one observed transition, whereas 177,102 have at least two and can enter the earlier adjusted Transition Entropy (TE) calculation. A score that ignores this distribution can combine a short observed sequence with a longer observed sequence as though their evidentiary support were the same. The primary profile consequently assigns Category Churn Rate (CCR) ranks within exact exposure strata and reports alternative profile definitions as sensitivities. This design makes exposure a reported condition of the measurement, rather than a nuisance variable removed after profile assignment.

The measurement problem has two related parts. First, customers have unequal numbers of observed transitions, so a profile label can reflect exposure rather than measured change. The primary profile, therefore, ranks CCR within exact transition-exposure strata. Second, turnover, persistence, and diversity are not interchangeable constructs. CCR is the primary turnover measure; category retention and category novelty report two directional set relations; Persona Vector Similarity (PVS) is auxiliary because it is nearly rank-redundant with CCR and adjusted TE is reported only as a diagnostic after no estimator met the pre-registered accuracy criterion for short histories. We refer to this measurement system as the Dynamic Persona Fluidity Framework (DPFF).

The empirical design also separates structural description from predictive and decision claims. The four-state category-breadth process rejects temporal homogeneity, $\chi^2(132) = 57{,}222.2$, and the held-out second-order log loss is lower than the first-order value, 0.886 versus 0.913. Yet the hidden Markov comparator matches the second-order benchmark within 0.005 held-out log loss, and nonhomogeneous Markov specifications do not improve on it. These comparisons retain the evidence for temporal heterogeneity while avoiding a pooled homogeneous summary as an explanation of all quarters.

The prospective evidence is equally constrained. After correction of the recency input, the mean incremental spending ΔR^2 is 0.0009 across five rolling-origin folds. The pre-registered decision-utility audit evaluates exit, online abandonment, spending loss, and online adoption against a strong comparator and finds no target that reaches an incremental area under the receiver operating characteristic curve (ΔAUC) of 0.01 in the required folds. The three-gate audit passes one gate only, and the apparent alignment result falls below the registered residual criterion once lagged category-change information is included. These results set the scope of the paper: the measures are operationally defined and descriptively characterized, but they do not show material incremental predictive value beyond the lagged-transaction baseline or decision value under the registered thresholds.

This framing addresses five risks in longitudinal customer analytics. A small prospective increment is reported against strengthened comparators rather than treated as practical value. Rejected temporal homogeneity is examined with order- and time-varying com-

parisons rather than with a homogeneous interpretation. CCR and PVS are separated as alternative set-overlap representations, and TE is excluded from scoring because its short-history estimators fail their accuracy criterion. A hidden Markov model (HMM) comparator is evaluated on the same held-out transitions. Finally, the exposition uses short definitions and reports the assumptions, comparison set, and failed gates needed to assess the evidence.

The present analysis builds on the same raw transaction records, the initial rule-based persona source, and the set-based fluidity measures (CCR, raw TE, PVS, Persona Persistence (PP), State Transition Frequency (STF), and Persona Volatility Index (PVI)) and three-profile scheme introduced in an earlier domestic study [8]. Section 3.2 documents the shared source and the cohort difference. The present analysis reports an exposure-aware audit of measurement, dynamic modeling, predictive value, and decision utility. The relationship is limited to these documented data and cohort facts; the present claims rest on the analyses reported here.

The contribution is methodological and empirical rather than theoretical: the paper does not propose a new theory of customer dynamics but a measurement design and an audit record that bound what such measures can claim.

This paper makes three contributions. First, it provides an exposure-aware measurement design whose elements, added in the present analysis, are calendar-consecutive pairing and exposure-conditioned profiles. It reports CCR, retention, novelty, and persistence without treating an observation gap as a transition. Second, it evaluates category-breadth dynamics with homogeneity tests and held-out order comparisons. The evidence rejects homogeneity, favors second-order dependence over first-order dependence, and does not show a consistent gain from hidden-state or nonhomogeneous alternatives: HMM was superior in 2024Q1, inferior in 2024Q2, and indistinguishable in the other three folds, whereas the nonhomogeneous specifications were inferior in all five folds (Section 4.4). Third, it prespecifies and discloses an audit of incremental prediction and decision utility. The audit records the comparator ladder, information-time symmetry, registered thresholds, and failed as well as passed gates.

Section 2 reviews static segmentation, dynamic state models, and evaluation designs. Section 3 defines the cohort, measures, models, and pre-registered audit. Section 4 reports the measurement, dynamic, and audit results. Section 5 interprets the results, and Section 6 concludes.

2. Literature Review

2.1. Static Segmentation and Its Temporal Limits

Customer segmentation partitions a customer population using features observed in a selected period [1]. Recency, frequency, and monetary value (RFM) analysis and related database-marketing systems likewise compress recency, frequency, and monetary behavior into a current customer representation [2]. Clustering, hierarchical procedures, and model-based segmentation can reveal structure in that representation, but they do not themselves identify whether one customer's subsequent profile differs from the earlier profile [9,10]. A temporal measurement must, therefore, define both the period pair and the behavioral object that is compared.

Segment instability research makes this temporal boundary explicit by treating customer value and segment membership as moving targets [3]. In digital and omnichannel settings, customer journeys supply further reasons to distinguish a current profile from a sequence of profiles [5,6]. The issue is not that static segmentation is invalid. Rather, a cross-sectional segment and a within-customer transition have different estimands. This study

uses static clustering as a descriptive benchmark and evaluates longitudinal measures only on calendar-consecutive quarter pairs.

Dynamic segmentation extends the unit of analysis from a fixed coordinate to a history. Studies have combined dynamic segmentation with portfolio logic, clustered time-indexed RFM measures, modeled evolving fuzzy membership in e-commerce, and learned preferences under sparse data [11–14]. These approaches motivate a longitudinal perspective, yet a profile change can still be defined in several non-equivalent ways. Replacement of categories, persistence of a dominant category, and diversity of observed state transitions describe different features of a history. The present study keeps those features separate and conditions the primary profile on observed transition exposure.

2.2. Dynamic Models of Customer States

Discrete-time Markov models represent customer movement by transition probabilities between observed states [4]. They have been used to characterize customer relationships when the current state is treated as the input to the next state [4]. Customer portfolios also display history dependence that a current-state summary can miss [15]. A first-order transition matrix is, therefore, a useful descriptive baseline, but it is an empirical assumption rather than a property of the observed data.

Higher-order models relax the first-order restriction by allowing recent state history to affect the next transition. Higher-order extensions of the Double Chain Markov Model provide one formal route for incorporating such dependence [16]. In the present setting, the relevant diagnostic is direct: compare first- and second-order models on held-out transitions that share the same state definition and information cutoff. This comparison assesses whether an additional observed lag improves prediction; it does not infer a latent behavioral mechanism from an order difference.

Hidden Markov models take a different route by distinguishing latent relationship states from observed sequences [17]. They can capture persistence or switching that is not represented by an observed-state matrix. Their contribution must nevertheless be assessed on the same forecast origin and outcome as an observed-state comparator. This study consequently treats the hidden Markov model as a benchmark alongside first- and second-order Markov models, rather than as an assumed superior representation.

Time-varying and regime-based Markov models address another possible limitation, temporal nonhomogeneity. A nonhomogeneous specification allows transition matrices to differ by period, while a regime representation pools periods that share estimated transition behavior. Such flexibility can describe structural variation, but it can also fail to improve a held-out comparison if lagged state information is more predictive. The appropriate diagnostic is thus two-part: test homogeneity and compare order and nonhomogeneous alternatives with the same rolling-origin evaluation. Section 4.4 reports fold-specific change-point candidates and full-sample regime blocks.

These model classes answer a different question from set-based fluidity measures. A Markov model summarizes transitions among category-breadth states. CCR, retention, novelty, and persistence summarize changes in the category sets and dominant categories that generate a customer's profile. Reporting both levels makes it possible to test whether dynamic structure is supported without presenting a pooled transition law as a complete account of individual profile change.

2.3. Incremental Value and Decision-Utility Audits

Feature evaluation requires a comparator that reflects the information already available at the forecast origin. Out-of-sample evaluation is designed to separate apparent in-sample fit from prospective performance [18]. For longitudinal customer data, a rolling-

origin design can refit preprocessing and models before each test quarter, then compare a common row set at the same information time. This study applies that discipline to a comparator ladder: an original transaction baseline, a strengthened lagged-transaction baseline with fixed effects, and dynamic-model benchmarks before additional fluidity features are considered.

Incremental R^2 and incremental AUC answer different questions. Incremental R^2 measures whether an added feature set improves a continuous-outcome prediction relative to its comparator. Incremental AUC measures discrimination for a specified binary target. Neither statistic establishes a business action or net benefit by itself; decision-analytic evaluation requires the relevant outcome, threshold, and consequences to be stated [19]. The present audit, therefore, uses pre-specified thresholds and reports no decision recommendation when the threshold is missed.

Comparator choice is especially important when a target shares a construction element with a feature. The alignment target in this study is a next-quarter category-set change, so lagged category-change features form a necessary comparator. Section 4.5 reports the corresponding audit outcomes. The result is interpreted as persistence in category-change history, not as evidence of an independent external outcome.

Pre-registration can make an observational study more auditable when it fixes the comparison, threshold, and reporting rule before the pertinent results are inspected [20]. It does not convert observational estimates into causal effects. Its value here is procedural: it requires the same treatment of favorable and unfavorable results. The audit registered five rolling-origin folds, information-time symmetry for all features, and separate uncertainty from fold refitting and paired bootstrap comparisons.

The decision-utility audit applies the same comparator discipline to exit, online abandonment, spending loss, and online adoption. Section 4.6 reports the corresponding audit outcomes after leakage was removed from the exit analysis. The audit, therefore, documents the null result instead of translating ranking differences into a deployment claim.

The metric sources also delimit what each measure represents. CCR is Jaccard distance [21], and PVS is cosine similarity [22]; they are alternative representations of set overlap. Raw TE is Shannon entropy [23], while the earlier adjusted form uses a Miller–Madow correction [24] discussed in entropy-estimation work [25]. Retention and novelty are set ratios, and PP, STF, and PVI are run-based descriptive statistics defined from the observed histories. Unlike CCR, PVS, and TE, these three run-based statistics have no established published counterpart that this study could adopt as a comparison measure, so they are reported as internally defined descriptors rather than as improvements on an existing index. The study reports raw TE as a descriptive quantity and excludes adjusted TE from profiles because no evaluated estimator met the pre-registered short-history accuracy rule.

Here, "auditable" has a narrow operational meaning. It refers to the five versioned records that Figure 1 attaches to the stages of this analysis: the source and output hashes, the aggregate cohort flow and exposure counts, the metric summaries, the exact transition counts and probabilities with a hashed held-out partition, and the frozen protocol commits and sensitivity results. Earlier dynamic-segmentation studies report portfolio dynamics, time-indexed RFM clustering, fuzzy membership, or sparse-preference segmentation [11–14]. To our reading of the cited designs [11–14], none of these four designs reports these five records together as a single measurement-and-evaluation record; we do not claim a systematic search. What follows is conditional rather than categorical: an analysis is less auditable along a given dimension when the corresponding record is absent from its report, and any of these four designs would be equally auditable if it supplied the same records. The distinction is, therefore, procedural and concerns what each report makes checkable, not which model reveals a truer customer state.

Table 1 organizes the streams above along the dimensions this study must fix before it can evaluate anything: the representation of customer change each stream adopts, the question each leaves open in the present setting, and how this study uses it, as a source of definitions, a comparator, or a benchmark. A cell-by-cell comparison of the cited articles on datasets, sample sizes, and reported predictive performance is not included, because the cited reports do not present those quantities in a form that permits comparable tabulation and assembling one from partial information would imply a ranking that the published record does not support. The table, therefore, contextualizes the design choices rather than scoring the cited designs against a scale defined by this study.

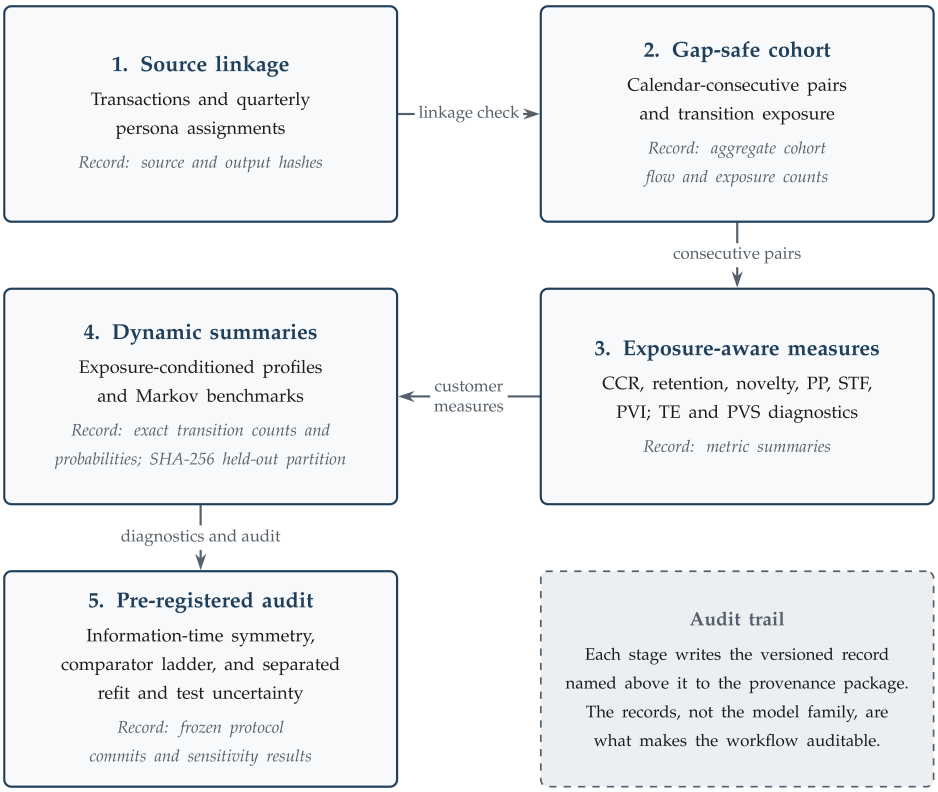

Figure 1. Conceptual framework for the exposure-aware measurement and audit workflow. Note: The italic line inside each stage names the versioned record that the stage writes to the provenance package. CCR: Category Churn Rate; TE: Transition Entropy; PVS: Persona Vector Similarity; PP: Persona Persistence; STF: State Transition Frequency; PVI: Persona Volatility Index.

Table 1. Literature streams and their roles in this study.

Stream (References)	Representation of Customer Change	What the Stream Leaves Open Here	How This Study Uses the Stream
Static segmentation and RFM [1,2,9,10]	Partition of a population using features observed in one selected window	Whether a customer's later profile differs from the earlier one, since the estimand is cross-sectional	Source of the cross-sectional benchmark against which the longitudinal measures are set
Segment instability and customer journeys [3,5,6]	Segment membership and customer value treated as moving targets	Which operational definition of a profile change is being compared, among set replacement, dominance, and transition diversity	Motivates keeping set replacement, dominant-category persistence, and transition diversity as separate measures

Table 1. *Cont.*

Stream (References)	Representation of Customer Change	What the Stream Leaves Open Here	How This Study Uses the Stream
Dynamic segmentation designs [11–14]	Portfolio dynamics, time-indexed RFM clustering, evolving fuzzy membership, sparse-preference segmentation	Whether the number of observed transitions per customer is held fixed when profiles are compared	Source of the longitudinal segmentation perspective adopted here
First-order observed-state Markov [4,15]	Transition probabilities between observed states, with the current state as the input to the next	Whether one lag suffices, which is an empirical assumption rather than a property of the data	Descriptive baseline in the model ladder
Higher-order Markov [16]	Recent state history affects the next transition	Whether the additional lag improves held-out prediction at a common information cutoff	Second-order comparator evaluated on the same folds
Hidden Markov models [17]	Latent relationship states behind the observed sequence	Whether the latent representation improves on an observed-state comparator at the same forecast origin	Benchmark, not an assumed superior representation
Time-varying and regime specifications [4,17]	Period-specific or regime-pooled transition matrices	Whether rejecting homogeneity implies a better held-out model	Source of the homogeneity test and of the nonhomogeneous comparator
Metric sources for set overlap and entropy [21–25]	Jaccard distance, cosine similarity, Shannon entropy and its Miller–Madow correction	Whether these estimators are accurate over the short exposure range observed here	Source of the CCR, PVS, and TE definitions
Out-of-sample and rolling-origin evaluation [18]	Separation of in-sample fit from prospective performance	Which comparator represents the information already available at the forecast origin	Source of the rolling-origin evaluation protocol
Decision-analytic evaluation [19]	Outcome, threshold, and consequences stated before a recommendation	Whether an incremental statistic by itself implies a business action	Basis for the pre-specified pass and fail thresholds in Section 3.7
Pre-registration of observational analyses [20]	Comparison, threshold, and reporting rule fixed before the results are inspected	Whether procedural auditability substitutes for causal identification, which it does not	Basis for the Stage A2 pre-registration protocol in Section 3.9

Note: Streams in the reviewed literature, the representation of customer change each adopts, what each leaves open in the present setting, and how this study uses each. The table organizes the reviewed work by stream; it is not a systematic search, not a scoring of individual articles, and not a ranking of design quality. RFM: recency, frequency, and monetary value; CCR: Category Churn Rate; PVS: Persona Vector Similarity; TE: Transition Entropy.

3. Materials and Methods

3.1. Overview

Figure 1 summarizes the analysis. We first linked de-identified transaction records to quarterly provider-assigned persona labels. We then retained only calendar-consecutive customer-quarter pairs. This gap-safe rule prevents an unobserved interval from becoming

an artificial transition. Each retained customer has an explicit transition exposure n, the number of eligible pairs used in customer-level summaries. We estimated set-change measures, profile assignments, and breadth-state dynamics from this common cohort. We then evaluated their descriptive and predictive roles through pre-specified diagnostics, dynamic comparators, and an audit with matched information sets. The framework is auditable in this restricted sense: each stage writes one versioned record to the provenance package, namely the source and output hashes, the aggregate cohort flow and exposure counts, the metric summaries, the exact transition counts and probabilities with a hashed held-out partition, and the frozen protocol commits and sensitivity results. Each record answers a specific audit question: which observations were eligible, how much transition evidence supports each customer summary, what was estimated, whether the comparators shared an information set and a held-out partition, and which thresholds were fixed before the results were inspected. Auditability in this sense is a property of the analysis record rather than of a model family. An implementation is less auditable along a given dimension when the corresponding record is absent, and a competing model supplied with the same records would be equally auditable. Section 2.3 reports which of these records the cited dynamic-segmentation designs provide.

3.2. Data Source and Cohort

An unnamed major financial institution in the Republic of Korea supplied de-identified transaction and quarterly persona-assignment sources from January 2022 through March 2025. The transaction source contained 96,189,818 records for 303,388 customers. The persona source contained 1,262,536 customer-quarter observations for 294,948 linked customers, 318 observed categories, and 746 observed persona labels. The transaction schema records provider-defined merchant categories, including several online-shopping categories, but it does not contain a transaction-level online or offline channel field. The corpus, therefore, represents mixed merchant-category card activity rather than an e-commerce-only population.

Table 2 describes the linked source. On average, a customer appeared in 4.281 quarters and had 13.968 active categories and 21.958 persona labels in an observed quarter. Labels use a category-plus-condition form. Examples observed in the source include “Mart: at least one visit”, “E-commerce: at least one visit”, “Convenience store: at least one visit”, “Mart: monthly average at least one visit”, and “Fixed expenditure: less than 30%” of total expenditure. These translations illustrate the observed label strings, not the provider’s rule specification. Provider-defined thresholds determined the labels, and the rule specification itself was not supplied.

Table 2. Descriptive statistics for the linked persona source.

Measure	Value	Measure	Value	Scope
Linked customers	294,948	Customer-quarters	1,262,536	2022Q1–2025Q1
Observed categories	318	Persona labels	746	2022Q1–2025Q1
Observed quarters per customer	4.281 mean	Median	3	customer level
Categories per customer-quarter	13.968 mean	Median	9	customer-quarter level
Labels per customer-quarter	21.958 mean	Median	14	customer-quarter level
CCR	0.638 mean	Transition exposure	4.037 mean	eligible cohort

Note: The linked source includes $N = 294{,}948$ linked customers and 1,262,536 customer-quarters. The persona-source summary files provide the 13-quarter range and source-wide counts but not quarter-by-quarter transaction counts. CCR: Category Churn Rate.

We converted periods to calendar-quarter ordinals and sorted observations within customer. We retained a pair only when adjacent observed quarters differed by one ordinal. Of 294,948 persona-linked customers, 235,965 had at least two observed persona quarters.

The rule retained 926,871 pairs for 229,586 customers and excluded 40,717 nonconsecutive successive pairs rather than bridging them. A further 6379 customers had multiple observations but no consecutive pair. Table 3 records the resulting flow. The sources cover the observed institutional product cohort, not a population sample. They remain subject to linkage, activity, attrition, product-selection, and unobserved spending on other cards or platforms. Selection and inverse-probability weighting diagnostics are reported with the results.

Table 4 reports raw customer-level and customer-quarter spending and transaction distributions for the calendar-consecutive-pair and exposure-eligible cohorts. Monetary values retain the source-data unit and are not converted. Both distributions are extremely right-skewed and include negative values, so their means are not representative central values. Among the 229,586 customers with a calendar-consecutive pair, total spending has a median of 5,049,934.5 against a mean of 32,155,082.63 and a maximum of 802,417,606,188, and its minimum is −120,517,428; the corresponding customer-quarter minimum is −69,081,710. Negative totals arise where refunds and cancellations recorded in a period exceeded purchases in that period. The table reports untransformed values so that these features remain visible. Every logarithmic transformation of spending in this paper is $\log(1 + x)$, and refunds are not truncated or shifted before it. The original audit and the decision-utility audit (Section 3.7) instead treat a negative quarterly total as a missing value rather than transforming it: a row whose next-quarter total is negative is excluded from the row universe of every target that depends on it, and a negative current-quarter total entering as a feature is imputed with the training-fold median. The confirmatory re-analysis (Section 3.8) inherits the original audit's row universe and applies the same rule to its own targets and features. The cohort-funnel weighting diagnostic (Table 5) enters untransformed spending totals as standardized covariates, so no logarithmic domain rule applies there.

The cohort funnel labels S0–S4 respectively denote raw transaction customers, customers with constructible quarterly records, customers with at least two observed quarters, customers with at least one calendar-consecutive pair, and customers with at least two transitions (Table 5). For the raw-to-analytic inclusion diagnostic, the inverse-probability-weight model used total spending, total transactions, active quarters, mean spending and transactions per active quarter, unique categories, first and last activity quarters, and 2022 activity; S4 weights were capped at their 99th percentile.

The expanded diagnostic reports unweighted and inverse-probability-weighted (IPW) standardized mean differences (SMDs) for every covariate, the trimmed-weight distribution from its minimum through the first, 50th, and 99th percentiles to its maximum, and effective sample size; it defines balance as $|\text{SMD}| < 0.10$.

The earlier domestic study used the same raw transaction records and initial rule-based persona source, as well as CCR, raw TE, PVS, PP, STF, PVI, and a three-profile scheme, for 70,833 customers with at least two quarters [8]. The present analysis uses 235,965 customers with at least two observed persona quarters and retains 229,586 customers with calendar-consecutive pairs. The two cohorts differ by inclusion rule: the earlier study applied its own rule to the same source, whereas the present rule (Section 3.2) requires calendar-consecutive observed quarters and reports the funnel in Table 5. The earlier study reported the three-profile description and a first-order homogeneous Markov summary for its cohort. The present analysis adds the calendar-consecutive cohort, exposure conditioning, the estimator audit, the hidden Markov and nonhomogeneous benchmarks, the selection analysis, and the pre-registered predictive and decision audit; the fluidity statistics reported here are computed on the present cohort. The static k-means benchmark instead uses the deterministic 20,000-customer sample described in Section 3.3, and the online-abandonment target uses the 152,333-customer online-active subset reported in Section 4.7.

The present revision additionally uses exposure-conditioned profiles, common-fold Markov comparators, and a pre-registered audit.

Table 3. Cohort construction and retained observations.

Stage or Pair Count	Count
Transaction customers	303,388
Persona-linked customers	294,948
At least two observed persona quarters	235,965
At least one calendar-consecutive pair	229,586
At least two calendar-consecutive pairs	177,102
Retained calendar-consecutive pairs	926,871
Excluded nonconsecutive successive-quarter pairs	40,717

Note: The flow begins with $N = 303{,}388$ transaction customers and ends with the listed retained customer cohorts and calendar-consecutive pairs.

Table 4. Raw descriptive statistics for the analysis cohorts.

Panel A. Customer level: calendar-consecutive-pair cohort

Statistic	Total spend	Total tx	Active quarters	Spend per active quarter	Tx per active quarter	Unique categories
N	229,586	229,586	229,586	229,586	229,586	229,586
Mean	32,155,083	413	5	4,855,825	60	35
SD	1,807,443,299	871	3	346,657,422	89	33
Min	−120,517,428	2	2	−17,216,775	1	1
Q1	1,326,522	32	3	351,870	8	10
Median	5,049,934	120	5	1,134,550	27	24
Q3	16,571,907	399	7	3,083,919	77	49
Max	802,417,606,188	43,113	13	160,483,521,238	3442	239

Panel B. Customer level: exposure-eligible cohort

Statistic	Total spend	Total tx	Active quarters	Spend per active quarter	Tx per active quarter	Unique categories
N	177,102	177,102	177,102	177,102	177,102	177,102
Mean	40,466,867	517	6	5,727,852	69	41
SD	2,057,481,637	966	3	394,234,863	95	34
Min	−120,517,428	3	3	−17,216,775	1	1
Q1	2,413,798	57	4	485,789	11	14
Median	7,678,978	185	6	1,413,112	34	31
Q3	22,365,967	546	8	3,606,532	91	58
Max	802,417,606,188	43,113	13	160,483,521,238	3442	239

Panel C. Customer-quarter level

Statistic	Calendar-pair spend	Calendar-pair tx	Exposure-eligible spend	Exposure-eligible tx
N	1,219,805	1,219,805	1,099,091	1,099,091
Mean	6,052,079	78	6,520,628	83
SD	743,662,399	122	783,298,038	126
Min	−69,081,710	1	−69,081,710	1
Q1	233,250	7	283,677	8
Median	1,149,365	29	1,317,920	34
Q3	3,793,460	104	4,104,418	114
Max	802,415,739,598	9788	802,415,739,598	9788

Note: Monetary values retain the raw source-data unit; transaction, quarter, and category measures are counts. Values are rounded to whole units. SD: standard deviation; Q1 and Q3: first and third quartiles.

Table 5. Cohort-selection funnel and inverse-probability-weighting diagnostics.

Stage	Definition	Customers
S0	Raw transaction customers	303,388
S1	Customers with constructible quarterly records	294,948
S2	At least two observed quarters	235,965
S3	At least one calendar-consecutive pair	229,586
S4	At least two transitions, adjusted TE estimable	177,102
Diagnostic	**Unweighted**	**IPW**
Mean CCR	0.6268	0.6269
Mean adjusted TE	0.6173	0.5997
Low state share	0.3969	0.4297
VeryHigh state share	0.2488	0.2283

Note: The CCR and TE means are customer-level means among the 177,102 S4 customers, with the second column reweighted by trimmed IPW. IPW: inverse-probability weighting; CCR: Category Churn Rate; TE: Transition Entropy.

3.3. Breadth States and Static Benchmark

For a static benchmark, we aggregated each customer's active categories over the observation period and formed a binary customer-by-category matrix. We fitted k-means for $k \in \{5, 8, 10, 12, 15\}$ to a deterministic 20,000-customer sample and evaluated silhouette scores on a deterministic 5000-customer subsample [9,10]. This benchmark concerns Euclidean separation in an aggregated binary representation. It does not test temporal change.

For longitudinal models, each observed quarter was assigned one of four breadth states labeled Low (1–5 categories), Medium (6–10), High (11–20), and VeryHigh (at least 21). The fixed 5/10/20 cutpoints provide interpretable breadth bands and avoid outcome-dependent tuning. They classify active categories, rather than persona-rule counts, because several rules may refer to the same category.

3.4. Fluidity Measure Definitions

All pair-level measures use only calendar-consecutive observations and are then averaged within each customer. The Category Churn Rate (CCR) is the Jaccard distance between category sets [21]. For category sets C_{t-1} and C_t,

$$\mathrm{CCR}_t = 1 - \frac{|C_t \cap C_{t-1}|}{|C_t \cup C_{t-1}|}. \tag{1}$$

Zero denotes identical category sets and one denotes no overlap. We also separated the overlap into retention and novelty. Retention is the share of the preceding set retained at t, whereas novelty is the share of the current set that was absent at $t-1$:

$$\mathrm{RET}_t = \frac{|C_{t-1} \cap C_t|}{|C_{t-1}|}, \tag{2}$$

$$\mathrm{NOV}_t = \frac{|C_t \setminus C_{t-1}|}{|C_t|}. \tag{3}$$

Customer features use expanding means through time t. A behavioral version replaces rule-mechanical additions and removals where the raw transaction mapping can adjudicate them; changes that cannot be adjudicated remain unchanged.

Persona Persistence (PP) is the mean duration of a run with the same dominant category. The dominant category is the one with the largest number of active provider rules in a quarter, with lexicographic tie-breaking. A calendar gap ends a run. State Transition Frequency (STF) is the proportion of eligible pairs with a change in breadth

state. Persona Volatility Index (PVI) is the within-customer standard deviation of active-category breadth. PVS is a supplementary cosine similarity of binary category vectors, not an independent construct:

$$\mathrm{PVS}_t = \frac{\mathbf{v}_t^{\top}\mathbf{v}_{t-1}}{\|\mathbf{v}_t\|_2\|\mathbf{v}_{t-1}\|_2}. \tag{4}$$

We report a construct-validity matrix of Pearson and Spearman correlations together with the pre-specified rank-redundancy rule: $|\rho| \geq 0.90$ is rank-redundant, $0.50 \leq |\rho| < 0.90$ is related but not redundant, and smaller correlations are weak.

Raw TE is the Shannon entropy of a customer's observed origin-destination transition types [23]. If n_{ij} is the count of transition $i \to j$, $n = \sum_{ij} n_{ij}$, and $\hat{p}_{ij} = n_{ij}/n$, then

$$\widehat{H}_{\mathrm{raw}} = -\sum_{i,j:n_{ij}>0} \hat{p}_{ij} \log_2 \hat{p}_{ij}. \tag{5}$$

The diagnostic adjusted score applies the Miller–Madow correction [24] and the maximum-observable normalization used in the earlier analysis [25]:

$$\widehat{H}_{\mathrm{adj}} = \min\left\{1, \frac{\min\{4, \widehat{H}_{\mathrm{raw}} + (k-1)/(2n\ln 2)\}}{\log_2\{\min(16, n)\}}\right\}, \tag{6}$$

where k is the number of observed transition types. The score is undefined at $n = 1$. We report raw and adjusted TE diagnostically but exclude adjusted TE from the primary score. In simulations with 154 empirical strata, $n = 2, \ldots, 12$, 2000 replications, seed 20260829, and both a true-support and a fixed-16 denominator, no estimator met the pre-registered accuracy rule. The candidate set comprised estimators E0 to E5: the current Miller–Madow-corrected, capped adjusted estimator (E0), jackknife (E1), Chao–Shen (E2) [26], Dirichlet posterior means (E3), Nemenman–Shafee–Bialek (NSB; E4) [27], and empirical-Bayes shrinkage (E5). Each candidate required Spearman $\rho \geq 0.5$, root mean squared error (RMSE) ≤ 0.20, and absolute bias ≤ 0.05 in every $n \geq 4$ stratum, plus an upper-boundary mass no more than 0.05 above the true mass at $n = 2, 3$. This design guards against retaining a score solely because it ranks some short histories well.

3.5. Fluidity Profiles

The primary exposure-conditioned fluidity profile uses the exact-n stratum. Within each stratum, we rank a customer's mean CCR by percentile and divide the deterministic rank ordering into equal thirds: Low, Intermediate, and High. Because TE failed the pre-registered estimator criterion, it is not included in this score. Customers with $n = 1$ remain in their exact-n stratum and, therefore, receive a CCR-only profile rather than being discarded or assigned a TE value. These profiles are relative descriptive positions, not latent types or treatment-response classes.

We assessed five alternative, report-only definitions: pooled-rank tertiles, quartiles collapsed to three labels, absolute CCR thresholds at one-third and two-thirds, a two-dimensional grid of CCR tertile by exposure tertile, and k-means with $k = 3$. The alternatives test dependence on a rank rule, threshold rule, two-dimensional partition, and clustering rule. None could replace the primary definition after outcomes were known.

Exact-n ranking makes the profile allocation insensitive to the marginal distribution of exposure. It does not make the underlying CCR estimates equally precise across strata. We, therefore, retain exposure in all profile summaries and use the alternatives as sensitivity descriptions rather than as competing estimands. The primary definition also separates a high CCR caused by a short history from a high rank among customers with the same

number of observed transitions. This choice responds to the finite-history problem without imputing unobserved transitions or treating a missing quarter as persistence.

3.6. Markov Model Specifications

We summarized breadth dynamics with a pooled first-order, four-state Markov matrix. If N_{ij} counts eligible transitions from state i to state j, the descriptive row probability is

$$\widehat{p}_{ij} = \frac{N_{ij}}{\sum_{k=1}^{4} N_{ik}}. \tag{7}$$

We tested temporal homogeneity with origin-state-specific Pearson contingency tests across origin quarters. Repeated transitions within customers mean that this is a descriptive diagnostic, so we also report origin-state effect sizes. First- and second-order Markov models were evaluated on a held-out customer partition defined by a SHA-256 customer hash. Both used add-0.5 smoothing and held-out log loss. We rebuilt all transitions under cutpoints 4/8/16, 5/10/20, 6/12/24, and pooled empirical quartiles for boundary sensitivity.

We used a categorical hidden Markov model (HMM) as a dynamic comparator over the four breadth-state observations. Within each rolling-origin fold, model size was selected from two through six hidden states by the Bayesian information criterion (BIC); parameters were estimated by expectation–maximization (EM) [28]. Each candidate used five random initializations, at most 100 EM iterations, and stopped when the log-likelihood increment was below 10^{-5}. Evaluation used a filtering posterior $P(S_t \mid \text{obs}_{1:t})$, never a smoothed posterior that uses future observations. Thus, HMM features and second-order transition features were available at the same information time as the remaining predictors.

For the Markov–HMM comparison, we stored each customer's held-out log loss and formed percentile 95% confidence intervals from 1000 paired customer bootstrap resamples with seed 20260830.

We also fitted a nonhomogeneous Markov specification. For quarter q, each row is partially pooled toward the pooled matrix:

$$P_q = \frac{N_q + \kappa P_{\text{pooled}}}{n_q + \kappa}, \qquad \kappa \in \{10, 100, 1000\}. \tag{8}$$

Here n_q is the row total, and κ is selected by the log-likelihood of the quarter immediately preceding the fold origin. We searched all 12 quarter boundaries in the complete descriptive series for at most three BIC-selected change points [29]. In fold-specific prediction, candidates were restricted to boundaries at or before the origin. We estimated separate matrices for the resulting regime blocks. Descriptive propagation multiplies the estimated matrices in chronological order over a finite horizon. We define the mixing horizon as the smallest number of quarters for which the largest total-variation distance between any two propagated state distributions falls below one-half of its initial value. We used 200 customer-block bootstrap replicates with seed 20260830 for held-out comparisons.

The pooled matrix is retained as a compact descriptive summary, while the order, HMM, and nonhomogeneous specifications are predictive comparators. The comparison does not permit later quarters to influence a fold's transition probabilities, change points, hidden-state fit, or feature construction. This timing rule matters because a model can appear stronger if it uses a regime boundary or posterior that would not have been available at the prediction origin. The four boundary schemes apply the same gap rule and held-out protocol. They, therefore, separate sensitivity to breadth discretization from sensitivity to observation continuity.

3.7. Pre-Registered Audit of Predictive Value

The audit used rolling origins with test quarters from 2023Q4 through 2024Q4. Every feature at quarter t used information available by t. The comparator ladder began with the original transaction baseline, then used the fixed-effects baseline, which adds seasonal indicators, trend, and lagged transaction blocks. It then added the domain comparators, namely HMM filtering posteriors and second-order Markov features, before evaluating the remaining fluidity features. Decision-utility models used the strengthened lagged-transaction baseline: the fixed-effects baseline, transaction frequency, monetary value, label breadth, target-specific lags, HMM posteriors, and second-order features.

Continuous targets were next-quarter $\log(1 + \text{spend})$ and the next-quarter set-change target. Binary targets were exit, online abandonment, spending loss, and online-label adoption. The six online labels were E-commerce, Shopping mall, Mobile and online games, Online education, Home-shopping and live-commerce spending-related, and Live commerce. Exit equals one when a customer is absent from the raw customer-quarter source at $t + 1$. Abandonment equals one when all six pre-specified online labels present at t are absent at $t + 1$; adoption reverses that condition. Spending loss equals one when the change in $\log(1 + \text{spend})$ is at or below the fold-training tenth percentile. The target universes follow the pre-specified observation requirements, so exit does not require $t + 1$ presence, whereas abandonment, adoption, and spending loss do. The original audit treats a negative quarterly spending total as a missing value before the logarithmic transformation: 1777 rows (1632 customers) whose next-quarter total was negative were removed before forming the eligible row universe of 527,609 rows and 138,720 customers used in the original audit (Section 4.5), and 919 rows whose current-quarter total was negative were imputed with the training-fold median as a feature. The domain-ceiling auditreuses that same row universe, and the alignment audit (both reported in Section 4.5) starts from it and retains only the rows on which its own target is defined, which is why the per-fold held-out counts in the notes to their respective tables differ. The decision-utility audit (Section 4.6) applies the identical rule within each target's universe, starting from a larger base panel of 646,587 rows (170,313 customers) because its exit target retains the 117,201 rows whose next quarter is not observed and which, therefore, have no next-quarter total. The remaining 529,386 rows are the pool from which the original audit removed its 1777 negative next-quarter rows, so both audits count the same 1777 rows. Its spending-loss target is undefined for rows whose current- or next-quarter total was negative (1488 and 1777 rows, respectively, counted within the base panel; 569 of the 1488 lie outside the original audit's universe, leaving the 919 rows reported above), so those rows leave the spending-loss universe, which, therefore, equals the original audit's universe less those 919 rows, 526,690 rows in all, while the exit, abandonment, and adoption targets, which do not depend on a spending total, retain them subject to their own observation requirements; a negative current-quarter total entering as a feature for the other targets is imputed with the training-fold median.

Ridge models used $\alpha = 1.0$, training-median imputation, training standardization, and an unpenalized intercept. Binary models used L2 logistic regression with $C = 1.0$. We separately estimated uncertainty from fold refitting and from 200 paired customer bootstrap draws within the test rows. Inverse-probability weights for the cohort funnel were estimated from the raw-to-analytic inclusion process using a 99th-percentile weight truncation [30]. We decomposed CCR into rule-mechanical and behavioral components, decomposed the 2023Q1 shift by entry vintage, and evaluated the online-active subset. The pre-specified discontinuity detector used an $n \geq 4$ pooled bucket for its High flag rather than exact-n strata; we disclose this implementation deviation and do not reinterpret gate (i).

The gates were fixed before results were inspected. Stage A denotes the protocol fixed before the submitted analysis, comprising gates (i) through (iii); Stage A2 denotes the post-review amendment protocol, comprising items P0 through P3 together with the descriptive amendments recorded in Section 3.9. Gate (i) required a 2023Q1 absolute z score of at least 3 for mean CCR or the High flag, no 2022 false positive, and no equally early simple occupancy signal. Gate (ii) required at least 50% behavioral churn and a same-sign standardized CCR coefficient of at least 0.10 in four of five folds. Gate (iii-a) required $\Delta R^2 \geq 0.01$ and a positive 95% confidence-interval lower bound in four folds; gate (iii-b) required TE beyond CCR to achieve $\Delta R^2 \geq 0.0005$ in four folds.

The alignment target is next-quarter pair CCR, so a lagged CCR block is a necessary comparator for (iii-a). That comparator and the residual increment computed after it were registered separately as Stage A2 item P0-a, whose fourth model is the full DPFF residual over the lag block, at the same 0.01 threshold deliberately carried over from (iii-a). The registration recorded before execution that this residual was unlikely to reach the threshold, because the Stage A CCR-only and full-DPFF levels differed by only 0.005 to 0.007. The gate result is, therefore, reported as met under its own criterion and qualified by the separately registered residual test, not downgraded by a rule introduced after the levels were seen. Gate (iii) is satisfied if either sub-criterion (iii-a) or (iii-b) is met.

The P0-b domain-ceiling rule classified the domain as capped when the maximum ΔR^2 among HMM filtering posterior, second-order Markov, and their combined features was below 0.005 in all five folds. P1 required the estimator criteria stated above. P2 was registered under the label decision utility; it measures incremental discrimination (ΔAUC) on decision-relevant targets and does not by itself establish a business action or net benefit. P2 required $\Delta\text{AUC} \geq 0.01$ with a positive 95% lower bound in four folds for at least one of exit, abandonment, or spending loss. P3 required a positive lower confidence bound for improvement in held-out log loss over the homogeneous second-order Markov model in four folds. These thresholds were fixed before the corresponding results were viewed.

The HMM override required both lower held-out log loss than the first-order Markov with a positive confidence-interval lower bound and nonpositive incremental DPFF ΔR^2 after HMM posterior probabilities for both spending and alignment targets. If both conditions held, the registered interpretation was restricted to measurement and audit rather than practical contribution.

The audit reports all attempted specifications within these pre-specified families. The original baseline is retained to reproduce the submitted comparison, but the stronger baselines determine whether an increment is interpreted as additional information. The HMM and second-order features define a domain-comparator ceiling rather than a target for model selection. We use the same test customers for each comparator in a fold, so the paired bootstrap evaluates a difference in performance rather than two unrelated sampling distributions. The cohort funnel and IPW analysis address selection into the analytic cohort; they do not recover behavior absent from the institutional source.

3.8. Stage C Confirmatory Re-Analysis

Stage C was a separate confirmatory re-analysis, registered after the first review round, that tested whether the fluidity feature family adds predictive value under specifications the original audit had not attempted. Its row universe was the 527,609 customer-quarter rows of the eligible panel of the original audit (Section 3.7), covering 138,720 customers. Estimation used five rolling-origin folds with test quarters from 2023Q4 through 2024Q4; training rows grew from 183,570 to 469,203 and test rows fell from 74,724 to 58,406 across those folds. Every Stage C feature respected the same information-time restriction used in Section 3.7.

Three confirmatory families were fixed before execution. C1 predicted the next-quarter level of $\log(1 + \text{spend})$, C2 predicted the next-quarter change in that quantity, and C3 predicted the Shannon entropy of next-quarter category spending shares in nats. C3 restricted the row universe to rows with positive next-quarter spending, which excluded 312 rows, or 0.059% of the panel; this exclusion removes rows with non-positive spending, for which the entropy is undefined, and is separate from the negative-spending missing-value rule stated below. Within the C3 universe, 8.73% of rows carried zero entropy because spending concentrated in a single category.

Each family was compared against a strengthened autoregressive baseline rather than the original transaction baseline. The C1 baseline held 44 features: the fixed-effects block, four quarterly lags of transaction count, category breadth, and log spending, amount dispersion, demographic indicators, three monthly aggregates, and the corresponding missingness indicators. The C2 baseline added the contemporaneous change in log spending, and the C3 baseline added three lags of category entropy together with second-order Markov state probabilities. These baselines are already strong. Pooled across folds, they explain 0.539 of the variance in next-quarter log spending and 0.573 of the variance in category-share entropy, while the change target is intrinsically harder at 0.159. Reporting this level matters for reading the increments in Section 4.5, because it bounds what any additional feature family can contribute.

The extension blocks were cumulative fluidity features, a recent-window block, and monthly aggregates, each with its own missingness indicators. The recent-window block held the pair CCR of the most recent transition, the mean of the two most recent pair CCRs, an exponentially weighted pair CCR with a smoothing weight of 0.5 per quarter, and the retention, novelty, and state-change indicators of the most recent transition. One of these features carried a prior-use history: the pair CCR had already been used in the Stage A2 alignment-target analysis, so a recent-window result would have to be read together with that earlier residual result. Missingness indicators entered the baseline and the extension identically. The fold-level training and test missingness rates of the four lags are reported in the registered artifacts and ranged from 0.3% to 2.7% at the first lag and from 21.1% to 63.5% at the fourth. A separate rule governs refund-dominated rows. The logarithmic transform used for the spending features and for the C2 target is undefined at or below -1, so rows whose net quarterly spending fell in that range were set to missing rather than truncated or shifted. This affected 919 of the 527,609 panel rows, or 0.17%. This panel reuses the eligible row universe of the original audit (Section 3.7), which already excludes rows with a negative next-quarter total, so next-quarter spending totals were never negative here. The rule, therefore, reached the C2 target only through its current-quarter term. Because that target is then undefined, these 919 rows were excluded from the C2 family; the C1 and C3 targets do not depend on the current-quarter total, so those families retained them, C3 subject to its own eligibility rule stated above. As a feature, the missing current-quarter term was imputed with the training-fold median, as in the original audit, and its missingness entered the baseline and the extension through the same indicators. The negative minima in Table 4 belong to the calendar-consecutive-pair and exposure-eligible cohorts of Section 3.2, whose customer-quarter rows (1,219,805 and 1,099,091) form broader universes than this 527,609-row panel.

The judgment learner was fixed before execution as a histogram gradient-boosted regressor with depth three, learning rate 0.05, 500 iterations, no early stopping, and a minimum of 20 samples per leaf. No hyperparameter search was performed, so the tuning degrees of freedom were zero. A ridge model with $\alpha = 1.0$ on the same design matrix was retained as a technical report for continuity with the earlier audit and had no role in the decision. Pooling the recorded fold-level ridge coefficients of determination with the same

equal weights gives increments of 0.0032 for C1, 0.0061 for C2, and 0.0053 for C3. These ridge values are descriptive: the bootstrap was not run for that specification, and all three remain below the 0.01 threshold that the registered rule applied to the judgment learner.

Uncertainty came from 2000 customer-clustered paired bootstrap draws that preserved the fold structure and applied the same resampling weights to every family and specification. The draws recomputed weighted coefficients of determination from the fitted predictions without refitting the models. That procedure excludes training variation, so the resulting intervals can be narrower than intervals from a procedure that refits on every draw; we disclose this rather than present them as a full account of estimation uncertainty. Simultaneous 95% intervals used a studentized max-T statistic computed jointly across the three families, with a critical value of 2.35, and multiplicity was controlled by a one-sided Westfall–Young stepdown procedure. Although C1, C2, and C3 are described above as three confirmatory families of models, they form a single multiplicity family for this inference. The critical value of 2.35 is the 95th percentile of $\max_j |T_j|$ across the three targets, giving 95% joint coverage for the three intervals, and the one-sided Westfall–Young stepdown p controls the family-wise Type I error across C1–C3 at 0.05. The equivalence test, registered in the Stage C protocol as a technical report, declares an increment confirmed below materiality when its one-sided 95% upper bound is below 0.01; that bound remains per-family and was not adjusted for simultaneity.

Four controls had to pass before the locked run, and all four passed. Under label permutation, the pooled increments were -0.0009, -0.00002, and 0.0005, inside the registered bound of 0.002 in absolute value. A sentinel run that deliberately injected a noisy copy of the next-quarter target produced a C1 increment of 0.3108, far above the 0.05 detection floor, confirming that the pipeline registers leakage of that magnitude. An information-time audit over the 97-column feature dictionary returned no feature whose source period extended beyond t. A monthly-to-quarterly reconciliation matched row counts, transaction counts, and spending in all 13 quarters with no mismatched rows; registration made any mismatch grounds for invalidating the run.

The decision rule was fixed before execution. It required all three of a pooled ΔR^2 of at least 0.01, a simultaneous confidence interval excluding zero, and at least four of the five folds meeting the same threshold with a positive percentile lower bound. Registration also fixed in advance the sentence to be reported if the rule failed, and it forbade a further analysis stage.

3.9. Pre-Registration and Disclosure

We froze the Stage A protocol before analyzing its revision outputs in the Stage A and Stage A2 pre-registration documents (versions `23a1c50` and `7442dd7` in the provenance package). Amendments were additive. Amendment A2-5 invalidated and reran an analysis after detecting outcome information encoded through an incomplete retention-novelty merge. Amendments A2-3 and A2-4 corrected a quarter-end parsing defect: Q2–Q4 recency values carried offsets of 91–184 days, while within-quarter ordering was preserved. We reran the affected analysis with correctly reconstructed quarter ends and retained the invalid run.

Amendment A2-7 clarified partial pooling, the full change-point search, origin-bounded fold candidates, and finite-horizon propagation before the replacement P3 run. Amendments A2-6 and A2-8 record a decision departure from the pre-registered withdrawal branch: the authors elected to resubmit with narrowed claims rather than withdraw. The amended protocol did not change thresholds, folds, or seeds. Invalid and superseded executions remain preserved with their provenance rather than being deleted.

After the 30 August 2026 decision to resubmit and before execution, Amendment A2-9 registered four additional descriptive analyses: a construct-validity correlation matrix, an expanded selection-balance diagnostic, a literature-comparison table, and customer-paired bootstrap confidence intervals for the Markov and HMM comparison. These analyses introduced no pass-fail rule and did not change thresholds, folds, seeds, or decision criteria.

Before execution on 30 August 2026, Amendment A2-10 registered a descriptive table of raw customer-level and customer-quarter spending and transaction distributions for the calendar-consecutive-pair and exposure-eligible cohorts; it introduced no decision rule and did not change thresholds, folds, seeds, or decision criteria.

A separate Stage C protocol was registered at commit `0e08f1a` on 30 August 2026, and its code was frozen at commit `b4e7e52` later the same day, before any Stage C model was fitted to a real target. It was a pre-specified post-review analysis on previously examined data, rather than a blind pre-registration: the authors had already seen the original prospective result and the reviewer's critique when the protocol was written. The fluidity feature family and the quarterly panel had already been used in the originally submitted analysis, and one extension feature carried a specific prior-use history, namely the recent-window pair CCR, which had been used in the Stage A2 alignment-target analysis. Stage C, therefore, tested new outcomes, resolutions, and learners on data whose feature construction was known. The protocol fixed three confirmatory families, a single decision threshold, the fold structure, and the seeds before execution, and four mandatory controls had to pass before the locked run, as Section 3.8 describes. Stage C introduced no change to any earlier registered threshold, fold, or seed.

The disclosure record distinguishes a corrected implementation from a changed decision rule. In particular, correcting the recency parser changed a data transformation, whereas the fold definitions and pass thresholds remained fixed. The retained invalid outputs permit readers to identify which result was replaced and why. Likewise, the decision recorded in Amendments A2-6 and A2-8, in which the authors elected to resubmit with narrowed claims rather than follow the registered withdrawal branch, does not convert a failed gate into a passed one. It preserves the originally specified result while documenting the departure itself.

This provenance practice also separates reproducibility from substantive support. A matching hash establishes the identity of an input or output artifact, but it does not itself establish that a measurement, comparator, or interpretation is appropriate. Each reported inference, therefore, remains tied to its estimand, information set, and stated diagnostic boundary.

The three protocol documents are supplied in full as Supplementary Documents S1–S3, so that every threshold, fold definition, seed, decision rule, and amendment cited in this section can be read in its registered form. They are English translations of the Korean originals frozen at the commits named above; a fidelity check confirmed that no numeric value, hash, or amendment was altered or omitted in translation.

3.10. Computational Environment and AI-Assisted Preparation

The originally submitted analysis ran locally with Python 3.14.6, pandas 3.0.3, NumPy 2.5.1, SciPy 1.18.0, scikit-learn 1.9.0, PyArrow 25.0.0, and DuckDB 1.5.5. The revision analyses ran locally with Python 3.14.6, pandas 3.0.5, NumPy 2.5.2, SciPy 1.18.1, scikit-learn 1.9.0, and PyArrow 25.0.1. The additional analyses were executed on 29–30 August 2026 with versioned inputs, code, output records, and fixed seeds stated above. OpenAI Codex CLI and Anthropic Claude Code assisted English revision, code drafting and review, structural checks, and citation-consistency checks. Their output was

not treated as empirical evidence. The authors checked proposed changes against analysis artifacts and cited sources and made all final decisions.

4. Results

4.1. Cohort, Descriptives, and Selection

The primary cohort contained 229,586 customers and 926,871 calendar-consecutive pairs. Transition exposure was heterogeneous: 52,484 customers had one pair, 40,089 had two, 54,477 had three or four, 63,740 had five to eight, and 18,796 had nine or more. Accordingly, adjusted TE was estimable for 177,102 customers. The deterministic static benchmark represented 20,000 sampled customers across 303 categories present in that sample. The highest silhouette score was 0.099 at $k = 5$; scores declined to 0.076, 0.053, 0.014, and 0.001 for $k = 8$, 10, 12, and 15, respectively. These values document weak separation under the specified aggregated binary representation.

Table 5 makes the selection path explicit. The final analytic subset is selected primarily for sustained observation rather than transaction amount: the standardized mean difference for active quarters was 0.64, whereas that for spending was no more than 0.01. Among the 177,102 S4 customers, inverse-probability weighting changed the customer-level mean of per-customer mean CCR by 0.0001, from 0.6268 to 0.6269. It shifted adjusted TE and the breadth-state composition modestly, so selection remains relevant for distributional descriptions even where the CCR mean is insensitive.

After 99th-percentile trimming, balance met $|\text{SMD}| < 0.10$ for total spending (0.007), mean spending per active quarter (0.003), and 2022 activity (0.079), whereas residual imbalance remained for active quarters (0.548) and unique categories (0.321); the trimmed-weight effective sample size was 168,456 and 1.0% of weights were capped. The remaining activity-level imbalance means that IPW is reported as a descriptive selection diagnostic, not as a correction that removes selection bias. Table 6 reports the unweighted and weighted standardized mean differences for all nine covariates.

Table 6. Covariate balance after inverse-probability weighting.

Covariate	Unweighted SMD	IPW SMD	Assessment
total spend	0.009	0.007	Balanced
total transactions	0.228	0.185	Residual imbalance
active quarters	0.637	0.548	Residual imbalance
mean spend per active quarter	0.004	0.003	Balanced
mean transactions per active quarter	0.224	0.173	Residual imbalance
unique categories	0.401	0.321	Residual imbalance
first activity quarter ordinal	−0.261	−0.213	Residual imbalance
last activity quarter ordinal	0.335	0.302	Residual imbalance
active in 2022	0.110	0.079	Balanced
Trimmed-IPW ESS = 168,456; trimmed weights = 1.0%			

Note: $N = 177{,}102$ S4 TE-estimable customers versus the $N = 303{,}388$ S0 raw-transaction reference set. IPW uses the pre-registered trimmed inverse-probability weights. TE: Transition Entropy; IPW: inverse-probability weighting; SMD: standardized mean difference; ESS: effective sample size.

Table 7 illustrates the quarterly category histories for two randomly selected, de-identified customers from the Low- and High-fluidity profiles. Each row reports the observed category breadth, category overlap and additions, CCR, breadth state, and three translated category examples. A missing quarter creates no transition pair, so CCR is reported only for immediately preceding calendar-consecutive observations. Profiles are assigned by CCR rank within exact transition-exposure strata (Section 3.5), so the Low-profile example can have a CCR above the pooled Low-profile mean of 0.424. The category names are provider-defined merchant-category labels reproduced from the source data

dictionary, not constructs of this study; "MBTI P consumption preference" is one such provider label.

Table 7. Illustrative de-identified quarterly histories.

Example ID	Quarter	$\|C_t\|$	Retained	New	CCR	Breadth State	Example Categories
8b35bf97	2022Q1	18	NA	NA	NA	High	IT services, furniture and interior, health supplements
	2022Q2	23	10	13	0.677	VeryHigh	snacks, health supplements, fixed expenditure
	2022Q3	17	10	7	0.667	High	home appliances, fixed expenditure, transit
	2022Q4	17	11	6	0.522	High	health supplements, fixed expenditure, road and parking
	2023Q1	9	4	5	0.818	Medium	transit, road and parking, household services
	2023Q2	6	1	5	0.929	Medium	fixed expenditure, supermarket, leisure facilities
7ef4dabe	2022Q1	31	NA	NA	NA	VeryHigh	gas utility, home appliances, meat
	2022Q2	35	21	14	0.533	VeryHigh	gas utility, fixed expenditure, public utilities
	2022Q3	27	19	8	0.558	VeryHigh	MBTI P consumption preference, snacks, fixed expenditure
	2022Q4	33	19	14	0.537	VeryHigh	MBTI P consumption preference, gas utility, home appliances
	2023Q1	38	25	13	0.457	VeryHigh	gas utility, home appliances, fixed expenditure
	2023Q2	32	20	12	0.600	VeryHigh	home appliances, fixed expenditure, transit

Note: Illustrative de-identified quarterly histories for $N = 2$ randomly selected customers, 8b35bf97 from the High-fluidity profile and 7ef4dabe from the Low-fluidity profile. IDs are SHA-256 eight-character pseudonyms. Quarters use YYYYQn notation. $|C_t|$: number of active categories in quarter t; CCR is the Category Churn Rate for the immediately preceding calendar-consecutive quarter; NA denotes that no preceding pair is available. Retained and New are category counts; the retention and novelty ratios of Section 3.4 divide these counts by the relevant set sizes. Profiles are assigned by CCR rank within exact transition-exposure strata (Section 3.5), so the Low-profile example can have a CCR above the pooled Low-profile mean of 0.424. Category names are provider-defined merchant-category labels reproduced as recorded in the source data dictionary and are not constructs of this study; "MBTI P consumption preference" is one such provider label.

4.2. Fluidity Measures

Table 8 reports customer-level distributions. CCR, retention, novelty, PVS, PP, STF, PVI, and raw TE are available for all customers with at least one consecutive pair. Adjusted TE is available only where two or more transitions permit its calculation. Retention and novelty were computed as cumulative customer measures from the observed pair history; their separate reporting preserves the direction of continuing categories and newly appearing categories.

Table 8. Customer-level distributional summary of fluidity measures.

Metric	N	Mean	SD	Median	Min	Max
CCR	229,586	0.638	0.208	0.648	0.000	1.000
RET	229,586	0.529	0.242	0.553	0.000	1.000
NOV	229,586	0.449	0.244	0.417	0.000	1.000
PVS (auxiliary)	229,586	0.508	0.218	0.532	0.000	1.000
PP (quarters)	229,586	1.702	1.243	1.333	1.000	13.000
STF	229,586	0.398	0.365	0.375	0.000	1.000
PVI	229,586	4.396	3.735	3.555	0.000	38.347
Raw TE (bits)	229,586	0.876	0.865	0.918	0.000	3.459
Adjusted TE (diagnostic)	177,102	0.617	0.411	0.766	0.000	1.000

Note: CCR and PVS are customer-level means of each customer's pair-level values. Values use $N = 229{,}586$ customers except adjusted TE, which uses $N = 177{,}102$ customers. CCR: Category Churn Rate; RET: category retention; NOV: category novelty; PVS: Persona Vector Similarity; PP: Persona Persistence; STF: State Transition Frequency; PVI: Persona Volatility Index; TE: Transition Entropy; SD: standard deviation.

CCR and PVS were nearly rank-redundant (Spearman $\rho = -0.984$ among the 229,586 customers with calendar-consecutive pairs and -0.982 among the 177,102 exposure-eligible customers). At the transition-pair level ($N = 926{,}871$), retention and novelty were related but not redundant (Pearson $\rho = -0.312$; Spearman $\rho = -0.208$). These pair-level

values differ in unit and cohort from the customer-level values in Table 9, where retention–novelty correlations were −0.727 (Pearson) and −0.648 (Spearman). CCR and adjusted TE were less strongly associated (Spearman $\rho = 0.385$). Figure 2, therefore, retains the original distributions while identifying adjusted TE as a diagnostic rather than a profile component.

For the construct-validity analysis, the pre-specified rule classifies $|\rho| \geq 0.90$ as rank-redundant, $0.50 \leq |\rho| < 0.90$ as related but not redundant, and smaller correlations as weak. In the exposure-eligible cohort, CCR–PVS was the only rank-redundant pair; related but nonredundant pairs were CCR–retention ($\rho = -0.858$), CCR–novelty ($\rho = 0.842$), adjusted TE–STF ($\rho = 0.881$), PVS–retention ($\rho = 0.889$), PVS–novelty ($\rho = -0.878$), and retention–novelty ($\rho = -0.648$), whereas CCR–adjusted TE was 0.385 and all remaining pairs were below 0.50 in absolute value. This result is consistent with treating PVS as a supplementary, non-independent measure: the High versus Low profile contrast was 2.216 pooled standard deviations for CCR and −2.221 for PVS. Table 9 reports the full correlation matrix.

Table 9. Construct-validity correlations among fluidity measures.

Metric	CCR	Adj. TE	PVS	PP	STF	PVI	RET	NOV
CCR	1.000	0.371	−0.964	−0.397	0.369	0.039	−0.870	0.853
Adj. TE	0.385	1.000	−0.263	−0.204	0.863	0.291	−0.208	0.168
PVS	−0.982	−0.340	1.000	0.376	−0.264	0.062	0.919	−0.911
PP	−0.451	−0.212	0.450	1.000	−0.184	−0.086	0.347	−0.337
STF	0.381	0.881	−0.330	−0.188	1.000	0.216	−0.192	0.170
PVI	−0.024	0.356	0.063	−0.080	0.335	1.000	0.068	−0.129
RET	−0.858	−0.274	0.889	0.415	−0.253	0.058	1.000	−0.727
NOV	0.842	0.238	−0.878	−0.412	0.235	−0.115	−0.648	1.000

Note: Lower-triangle entries are Spearman rank correlations; upper-triangle entries are Pearson correlations. All correlations use the 177,102 exposure-eligible customers (Section 3.2); pairwise complete observations were used, and every cell has $N = 177{,}102$. CCR: Category Churn Rate; Adj. TE: adjusted Transition Entropy; PVS: Persona Vector Similarity; PP: Persona Persistence; STF: State Transition Frequency; PVI: Persona Volatility Index; RET: category retention; NOV: category novelty.

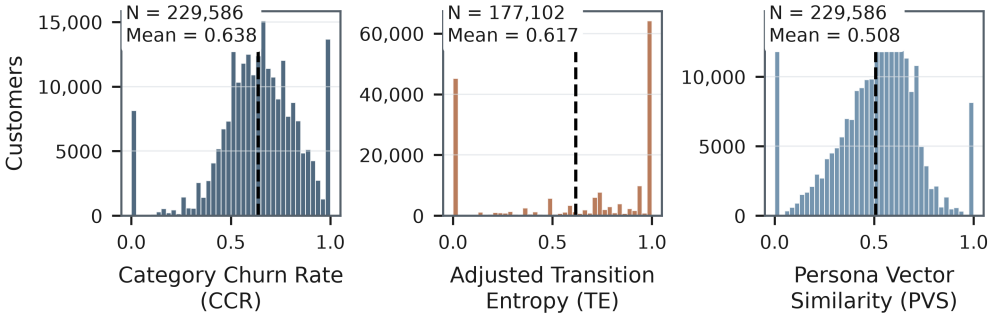

Figure 2. Customer-level distributions of fluidity measures. Note: Customer-level distributions of Category Churn Rate (CCR), adjusted Transition Entropy (TE), and Persona Vector Similarity (PVS). The dashed vertical line in each panel marks the mean of that measure. Adjusted TE is shown only for the 177,102 customers with at least two calendar-consecutive transitions and is retained as a diagnostic measure.

The estimator simulations explain that restriction. Table 10 reports that, under the empirical K_{true} design, the adjusted estimator had Spearman correlations of 0.803–0.843 and RMSE of 0.295–0.395 for $n = 4$–12. The fixed-16 sensitivity improved RMSE to 0.057–0.253, but still failed the retained-for-all-exposures rule. The independent E0–E5 comparison likewise had zero estimators satisfying every pre-registered criterion, including RMSE ≤ 0.20, across the relevant exposure range. Adjusted TE was consequently excluded from the score and from the primary profiles. In the originally submitted pooled-rank profiles, 81.9% of the High-fluidity group had adjusted TE at the upper boundary of 1.0;

this boundary mass motivated the estimator audit and the exclusion of adjusted TE from the score.

Table 10. Simulation performance of adjusted Transition Entropy estimators.

Design or Estimator	Spearman ρ	RMSE	Bias
M4 empirical K_{true}	0.803–0.843	0.295–0.395	−0.284–−0.164
M4 fixed 16	0.887–0.978	0.057–0.253	0.020–0.137
E0 current adjusted estimator	0.723–0.816	0.335–0.421	−0.321–−0.202
E1 jackknife	0.752–0.878	0.100–0.272	−0.093–−0.009
E2 Chao–Shen	0.748–0.856	0.104–0.274	−0.090–−0.016
E3 Dirichlet $\alpha = 1$	0.787–0.885	0.093–0.141	−0.085–−0.048
E4 NSB	0.792–0.883	0.102–0.224	−0.127–−0.045
E5 empirical Bayes	0.736–0.878	0.134–0.334	−0.191–−0.057
Pre-registered passes	0 estimators		

Note: Values are ranges across $n = 4$–12 evaluation strata, with 2000 simulation repetitions per stratum. RMSE: root mean squared error. M4 denotes the Stage A simulation of the adjusted estimator in Section 3.4, under the empirical true-support denominator (K_{true}) and under a fixed 16-category denominator; E0–E5 are the candidate estimators of the Stage A2 simulation.

4.3. Exposure-Conditioned Profiles

The primary profiles rank CCR within exact transition-exposure strata. By construction, the within-stratum tertile rule gives approximately equal Low, Intermediate, and High shares at each observed n, so the three profiles share the same mean exposure (4.037 in Table 11), rather than allowing short histories to determine a profile label. Table 11 shows that the three profiles have almost identical mean exposure, while their CCR, retention, novelty, PP, and STF values remain ordered. Figure 3 displays the corresponding profile means. The top primary group contains 17.65% of customers at the CCR upper boundary, which is reported as a descriptive feature rather than evidence of a latent type.

Table 11. Exposure-conditioned primary fluidity profiles.

Profile	N	CCR	RET	NOV	PP	STF	Mean n
Low	76,533	0.424	0.727	0.257	2.188	0.227	4.037
Intermediate	76,530	0.658	0.543	0.428	1.599	0.481	4.037
High	76,523	0.832	0.317	0.663	1.319	0.486	4.037

Note: $N = 229{,}586$ customers with at least one calendar-consecutive pair. RET: category retention; NOV: category novelty; CCR: Category Churn Rate; PP: Persona Persistence; STF: State Transition Frequency; n: transition exposure.

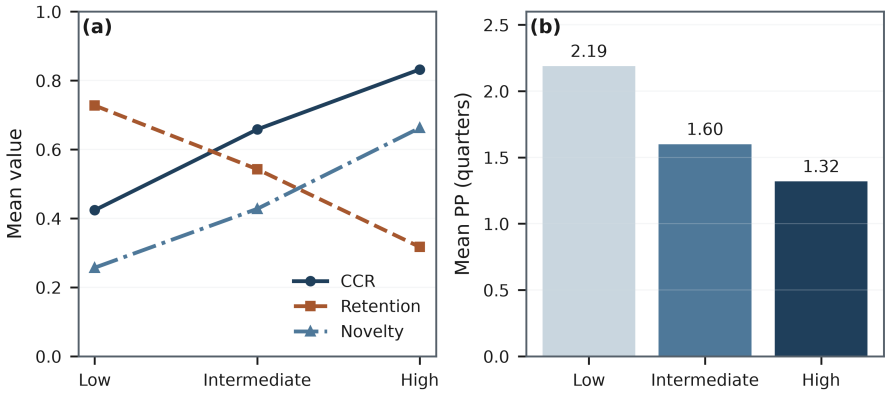

Figure 3. Fluidity-measure means across exposure-conditioned primary profiles. Note: Means of Category Churn Rate (CCR), category retention, category novelty, and Persona Persistence (PP) across exposure-conditioned primary profiles for $N = 229{,}586$ customers. Panel (**a**) uses the common zero-to-one scale for CCR, retention, and novelty; panel (**b**) reports PP in quarters.

For transparency, Table 12 retains the pooled-rank comparison from the originally submitted version, but labels it as a sensitivity analysis. The pooled and exposure-conditioned classifications agreed for 82.8% of the 177,102 customers with estimable adjusted TE (146,560/177,102). The same directions appeared after requiring at least four transitions ($N = 106{,}934$, 83.1% agreement) and at least six transitions ($N = 62{,}083$, 75.1% agreement). These comparisons show that the rank-based labels depend on the definition used, which is why the exposure-conditioned profile is primary.

Table 12. Pooled-rank versus exposure-conditioned profile classifications.

Pooled-Rank Profile	Exposure-Conditioned Primary Profile		
	Low	**Intermediate**	**High**
Low	53,051	4179	1804
Intermediate	5983	44,665	8386
High	0	10,190	48,844

Note: $N = 177{,}102$ customers with at least two transitions. Entries are customer counts.

Alternative definitions varied substantially (Table 13). The exact-n quartile partition had Cramér's $V = 0.817$ with the primary labels. Absolute thresholds and k-means had agreement of 0.557 and 0.525, with $\kappa = 0.336$ and 0.287, respectively. The two-dimensional grid reuses the primary CCR axis, so its agreement is mechanically 1.000. Most importantly, the absolute threshold High share declined from 0.595 in the first exposure tercile to 0.311 in the third, and its Low share from 0.092 to 0.026. Thus, the alternative definitions retain an exposure gradient that the primary definition makes explicit.

Table 13. Sensitivity of exposure-conditioned profiles to alternative definitions.

Alternative Definition	Agreement	κ	V
Exact-n quartiles	–	–	0.817
Absolute CCR thresholds	0.557	0.336	–
Two-dimensional CCR axis	1.000	1.000	–
k-means on CCR and $\log(1+n)$	0.525	0.287	–

Note: $N = 229{,}586$ customers with at least one calendar-consecutive pair. CCR: Category Churn Rate; κ: Cohen's kappa; V: Cramér's V; n: transition exposure.

4.4. Markov Dynamics

Figure 4 reports pooled transition probabilities from the 926,871 exact counts, and Figure 5 shows the corresponding count-weighted flow. Boundary-state persistence was highest: $p_{\text{Low,Low}} = 0.735$ and $p_{\text{VeryHigh,VeryHigh}} = 0.772$. Medium customers moved to Low with probability 0.364, and High customers moved to VeryHigh with probability 0.204. These transitions describe measured category breadth, not direct customer value.

Transition probabilities varied over origin quarters ($\chi^2(132) = 57{,}222.2$, $p < 0.001$), with origin-state Cramér's V from 0.115 to 0.167. This heterogeneity is detectable, but the SHA-256 held-out partition described in Section 3.6 favored the homogeneous second-order model: aggregate log loss was 0.886 versus 0.913 for first order across 137,736 test transitions. Table 14, in contrast, uses rolling-origin folds and shows the same ordering by fold. HMM matched the second-order model within ± 0.005 across folds, with lower log loss in 2023Q4 and 2024Q1 and higher log loss in the other three folds; the pre-registered second-order-minus-last-quarter improvement interval was negative in all five folds, and the regime specification's point log loss exceeded the second-order value in every fold, giving a 0/5 P3 decision. For the 2023Q4 and 2024Q3 rows that appear tied after three-decimal rounding, the paired-bootstrap point estimates in Table 15 determine the reported direction.

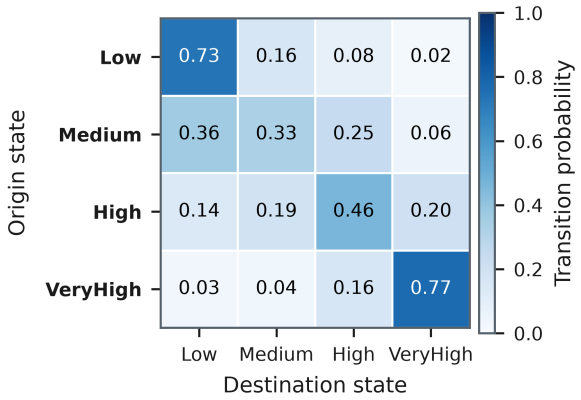

Figure 4. Pooled quarterly transition probabilities among Low, Medium, High, and VeryHigh category-breadth states. Values use 926,871 calendar-consecutive pairs.

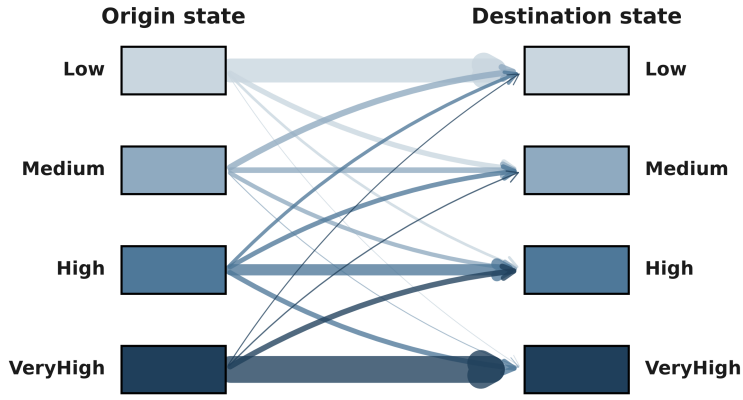

Figure 5. Observed full-cohort transition flow. Note: Observed full-cohort transition flow from 926,871 calendar-consecutive pairs. Edge widths are proportional to exact transition counts. Edge color follows the ordered state palette shared with Figure 4, so the origin state is identified by shade alone and the encoding is preserved in grayscale reproduction. The labels beside the left and right node columns name the origin and destination states, respectively.

Table 14. Held-out log-loss benchmarks for Markov specifications.

Test	First	Second	HMM	Last-Quarter	Regime	κ	CI
2023Q4	0.960	0.914	0.914	0.936	0.936	1000	[−0.023, −0.020]
2024Q1	1.012	0.971	0.966	0.986	0.986	10	[−0.017, −0.014]
2024Q2	1.007	0.969	0.972	0.976	0.983	10	[−0.008, −0.004]
2024Q3	0.927	0.893	0.893	0.913	0.911	10	[−0.021, −0.018]
2024Q4	0.940	0.904	0.905	0.924	0.924	10	[−0.022, −0.019]

Note: Five rolling-origin folds, with 84,158–110,642 held-out customers per fold. HMM: hidden Markov model; CI: 95% bootstrap confidence interval for second-order minus last-quarter log loss; κ: selected transition-smoothing parameter.

The paired customer bootstrap shows that the second-order-minus-HMM difference excluded zero only in 2024Q1 (0.0049 [0.0040, 0.0058], favoring HMM) and 2024Q2 (−0.0025 [−0.0034, −0.0014], favoring the second-order model); the other three fold intervals included zero, and every absolute difference was at most 0.005. First-order-minus-HMM differences ranged from 0.034 to 0.047 in all five folds and every interval excluded zero, consistently favoring HMM over the first-order model. Second-order-minus-regime differences ranged from −0.014 to −0.022 in all five folds and every interval excluded zero, favoring the second-order model over the nonhomogeneous regime model and agreeing with the 0/5 P3 decision.

Table 15. Bootstrap confidence intervals for held-out log-loss differences.

Fold	Second-Order−HMM [95% CI]	First-Order−HMM [95% CI]	Second-Order−Regime [95% CI]
2023Q4	0.0004 [−0.0005, 0.0012]	0.0462 [0.0445, 0.0478]	−0.0216 [−0.0230, −0.0201]
2024Q1	0.0049 [0.0040, 0.0058]	0.0465 [0.0446, 0.0482]	−0.0157 [−0.0171, −0.0143]
2024Q2	−0.0025 [−0.0034, −0.0014]	0.0349 [0.0331, 0.0366]	−0.0140 [−0.0155, −0.0125]
2024Q3	−0.0001 [−0.0010, 0.0010]	0.0337 [0.0318, 0.0355]	−0.0181 [−0.0198, −0.0166]
2024Q4	−0.0007 [−0.0019, 0.0004]	0.0357 [0.0339, 0.0375]	−0.0197 [−0.0212, −0.0181]

Note: Customer-paired bootstrap (1000 resamples) percentile confidence intervals across five folds, with 84,158–110,642 held-out customers per fold. Differences are the first-named model minus the second-named model, so negative values favor the first-named model. HMM: hidden Markov model; CI: 95% confidence interval; P3: nonhomogeneous Markov analysis; the regime model is the P3 nonhomogeneous Markov specification.

The full-sample regime blocks were 2022Q2 to 2023Q2, 2023Q3 to 2023Q4, 2024Q1 to 2024Q4, and 2025Q1. In the rolling-origin folds, the 2023Q4, 2024Q1, and 2024Q2 candidates were 2022Q4, 2023Q1, and 2023Q2; the 2024Q3 and 2024Q4 candidates were 2023Q1, 2023Q2, and 2023Q4. Finite-horizon propagation is displayed in Figure 6; its state occupancy changes over the first eight quarters, while the mixing horizon was three or four quarters across origin matrices. The temporal heterogeneity was observed consistently across boundary schemes, while the second-order specification gave lower held-out log loss than the first-order, last-quarter, and regime specifications in every fold and the hidden Markov comparator matched it within 0.005. Section 4.5 reports the associated vintage-composition evidence.

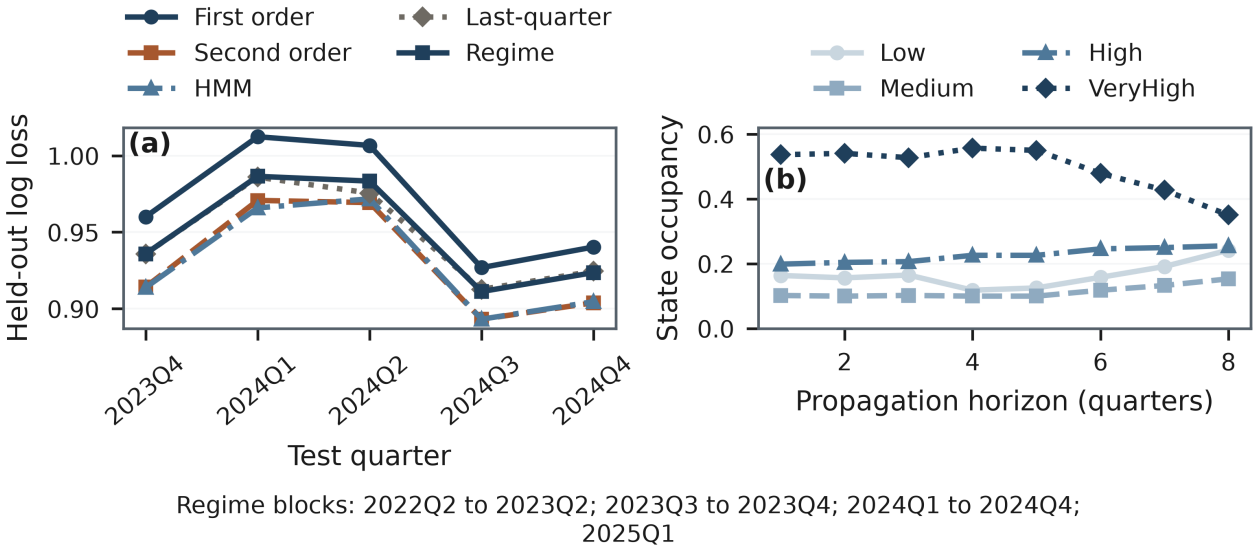

Figure 6. Nonhomogeneous Markov results. Note: Nonhomogeneous Markov results. Panel (**a**) compares five fold-level held-out log-loss results. Panel (**b**) shows full-sample regime blocks and finite-horizon state occupancy under the time-varying transition sequence.

The boundary-persistence pattern was retained under all evaluated cutpoints (Table 16). Temporal homogeneity was rejected for each discretization, but the table is limited to observed self-transition probabilities and the homogeneity diagnostic. It does not extrapolate a time-invariant process from the pooled matrix.

Table 16. Sensitivity of transition probabilities to category-breadth boundaries.

Boundaries	p_{LL}	p_{MM}	p_{HH}	p_{VV}	$\chi^2(132)$
4, 8, 16	0.702	0.301	0.414	0.792	57,737.6
5, 10, 20	0.735	0.330	0.463	0.772	57,222.2
6, 12, 24	0.760	0.356	0.507	0.752	55,769.2
Pooled quartiles (4, 12, 24)	0.702	0.473	0.507	0.752	56,877.0

Note: Values use 926,871 calendar-consecutive pairs. p_{LL}, p_{MM}, p_{HH}, and p_{VV}: observed probabilities of remaining in the Low, Medium, High, and VeryHigh states, respectively; $\chi^2(132)$: chi-square homogeneity statistic with 132 degrees of freedom.

4.5. Pre-Registered Audit: Predictive Value

The predictive audit used five rolling-origin test quarters and predictors available by the predictor quarter. Table 17 preserves the original baseline as submitted because its recency field had a quarter-end parsing defect. The corrected comparison is reported alongside it rather than replacing it. Correcting recency increased the baseline R^2 in every fold and reduced the mean incremental R^2 from 0.0011 to 0.0009 (0.00112 to 0.00088 at five decimals). Thus, the originally submitted baseline result is not interpreted as evidence of a practically meaningful increment. The strengthened lagged-transaction baseline yields separate full-block increments of 0.0003 to 0.0012, also below the relevant pre-registered threshold.

Table 17. Prospective log-spending comparison.

Test Quarter	Original Base R^2	Original ΔR^2	Corrected Base R^2	Corrected ΔR^2	Strengthened ΔR^2
2023Q4	0.5300	0.0009	0.5335	0.0006	0.0007
2024Q1	0.5028	0.0011	0.5042	0.0009	0.0012
2024Q2	0.5377	0.0015	0.5452	0.0011	0.0012
2024Q3	0.5157	0.0016	0.5180	0.0011	0.0011
2024Q4	0.4297	0.0005	0.4348	0.0007	0.0003
Mean increment		0.0011		0.0009	

Note: Before and after correction of the recency parsing defect, across five folds with 58,406–74,724 held-out customer-quarter rows per fold. The final column is the full DPFF-block increment beyond the strengthened lagged-transaction baseline. R^2: coefficient of determination; DPFF: Dynamic Persona Fluidity Framework.

The expenditure branch of gate (iii-b) required adjusted Transition Entropy (TE) to add at least 0.0005 in four of five folds beyond CCR. Its observed increments ranged from −0.000194 to 0.000350, so none of the five folds met the criterion. This distinguishes a small model-comparison difference from the pre-registered evidential threshold.

The alignment target met gate (iii-a) against the strengthened baseline, with all-DPFF increments from 0.1356 to 0.1724 and positive bootstrap lower bounds in all five folds (Table 18). This pass does not establish an independent predictive contribution. The target is the next-quarter Jaccard-set change, whereas CCR is also defined from successive category sets. The lag block raised R^2 to 0.237–0.283, and the residual contribution of the non-CCR DPFF features was only 0.0038 to 0.0093, below the 0.01 criterion in every fold. The alignment result is, therefore, reported as mechanical continuity in the CCR time series rather than as external validation.

The domain-ceiling analysis added fold-consistent HMM filtering posterior probabilities and second-order Markov features to the baseline before assessing remaining DPFF information. The maximum comparator increment was 0.0031, 0.0031, 0.0039, 0.0048, and 0.0018 across the five folds (Table 19); the residual DPFF increment after the combined comparator was 0.0004 to 0.0007. The second HMM override condition was not

met because the required nonpositive DPFF increment for both targets was not observed. The earlier origin-only posterior output was not used because it did not provide a valid fold-consistent comparator.

Table 18. Alignment-target comparison after lag adjustment.

Test Quarter	All-DPFF ΔR^2	CI Lower Bound	Lag-Block R^2	Residual DPFF Block after Lag Block
2023Q4	0.1715	0.1655	0.2828	0.0038
2024Q1	0.1356	0.1295	0.2367	0.0054
2024Q2	0.1397	0.1333	0.2746	0.0044
2024Q3	0.1656	0.1595	0.2665	0.0044
2024Q4	0.1724	0.1661	0.2483	0.0093

Note: Lag-block R^2 is the R^2 level of the base model after adding the lagged CCR block, not an increment. Base R^2 before the lag block ranged from 0.0331 to 0.0987. Across five folds with 56,525–73,622 held-out customer-quarter rows per fold. CCR: Category Churn Rate; CI: paired-bootstrap 95% confidence interval; DPFF: Dynamic Persona Fluidity Framework.

In 2023Q4, C (0.0029) was below A (0.0031) because the A, B, and C ridge specifications were fitted separately within each fold, with no constraint that the combined specification must weakly dominate either component under penalization.

Table 19. Domain-ceiling comparison for log spending.

Test Quarter	2023Q4	2024Q1	2024Q2	2024Q3	2024Q4
A: HMM posterior ΔR^2	0.0031	0.0031	0.0039	0.0038	0.0014
B: second-order Markov ΔR^2	0.0028	0.0026	0.0032	0.0031	0.0013
C: combined comparator ΔR^2	0.0029	0.0030	0.0035	0.0048	0.0018
Maximum comparator ΔR^2	0.0031	0.0031	0.0039	0.0048	0.0018
Residual DPFF ΔR^2 after C	0.0006	0.0007	0.0007	0.0004	0.0004

Note: Across five folds with 58,406–74,724 held-out customer-quarter rows per fold. A, B, and C are the respective increments over the strengthened baseline from the HMM filtering posterior, second-order Markov features, and their combined block; the maximum comparator increment is the maximum of A–C. Residual DPFF is the increment after C. HMM: hidden Markov model; ΔR^2: incremental coefficient of determination.

Rule decomposition identified 84.3% of mappable category disappearances as behavioral, exceeding the first condition of gate (ii). The behavioral and total CCR measures had a Pearson correlation of 0.975 across 926,871 calendar-consecutive pairs. However, the standardized $CCR_{behavioral}$ coefficients were -0.0145, -0.0075, -0.0009, 0.0005, and -0.0014; the same-sign coefficient threshold of 0.10 was met in zero folds. The gate, therefore, did not pass. The vintage decomposition further attributed the 2023Q1 mean-CCR movement to between-vintage composition: the within-vintage share was -0.175, while the balanced-cohort 2023Q1 statistic was 0.7744. Table 20 makes the limited mapping coverage explicit.

The break-detection result also did not pass gate (i). The quarterly pair-level mean CCR rose to 0.6210 in 2023Q1 ($z = 10.10$), but leave-one-out evaluation produced a baseline-period false positive. In addition, the primary High flag was defined operationally in an aggregated $n \geq 4$ exposure bucket, rather than the exact-n rule in the registered gate, and the simple state-change comparator signaled in the same quarter ($z = 10.89$). These observations are disclosed as an implementation deviation and do not support a claim of earlier detection.

Stage C tested three pre-specified outcomes with a strengthened autoregressive baseline, gradient-boosted trees in addition to ridge, and cumulative, recent-window, and monthly feature resolutions. The outcomes were spending level, spending change, and category-share

entropy. Its decision rule required all three of pooled $\Delta R^2 \geq 0.01$, a simultaneous confidence interval excluding zero, and at least four of five folds meeting the threshold.

Table 20. Rule-decomposition and vintage diagnostics.

Diagnostic	Measure	Value
Mappable disappearances	Behavioral proportion	84.3%
	Coverage of all disappearances	59.6%
$CCR_{behavioral}$	Coefficient range across folds	−0.0145 to 0.0005
	Folds meeting 0.10 criterion	0/5
2023Q1 vintage decomposition	Within-vintage share	−0.175
Balanced cohort	2023Q1 mean-CCR z	0.7744

Note: Category Churn Rate (CCR), using 926,871 calendar-consecutive pairs; coefficient results use five rolling-origin folds. The behavioral coefficient is standardized in the single-feature spending specification. z: standardized score.

Bootstrap p values printed as 0.000 denote $p < 1/2000$, the smallest value resolvable by the 2000 paired resamples, and not an exact zero probability.

For C1 spending level, the pooled ΔR^2 was 0.0036 (95% simultaneous CI, joint across C1–C3, [0.0034, 0.0039]; 1/5 qualifying folds). The single qualifying fold was 2024Q1 at 0.0143, while the other four ranged from 0.0003 to 0.0016, so the pooled value is not a summary of five similar folds; for C2 spending change, it was 0.0014 ([0.0010, 0.0018]; 0/5) and for C3 category-share entropy, it was 0.0029 ([0.0025, 0.0033]; 0/5). None of the three families met the rule. Although the increments were distinguishable from zero, all were below the registered 0.01 materiality threshold. The one-sided equivalence tests, which evaluate the sampling distribution of each pooled point estimate rather than the dispersion across the five folds, placed the three 95% upper bounds below 0.01, confirming that the increments were below materiality rather than merely lacking evidence. Those bounds are per-family (Section 3.8); the simultaneous intervals in Table 21, which do apply the joint critical value, have upper limits of 0.0039, 0.0018, and 0.0033, so a simultaneity-adjusted bound would likewise fall below 0.01. One C1 fold (2024Q1) exceeded 0.01, whereas the remaining four did not, and the protocol pre-specified that a single fold does not overturn the conclusion. The last column of Table 21 reports the residual error reduction (RER), defined as the pooled increment divided by the variance the strengthened baseline leaves unexplained, $\Delta R^2/(1-\bar{R}^2_{\text{base}})$, where $\bar{R}^2_{\text{base}}$ is the mean baseline coefficient of determination across the five folds. On that scale, the three families are 0.0079, 0.0017, and 0.0068, so rescaling the increments does not change the conclusion.

Table 21. Stage C pre-specified predictive-value results.

Family	ΔR^2	95% CI	Folds	p	Equivalence	RER
C1 spending level	0.0036	[0.0034, 0.0039]	1/5	0.000	0.0038	0.0079
C2 spend change	0.0014	[0.0010, 0.0018]	0/5	0.000	0.0017	0.0017
C3 category-share entropy	0.0029	[0.0025, 0.0033]	0/5	0.000	0.0032	0.0068

Note: ΔR^2: pooled incremental coefficient of determination over the strengthened autoregressive baseline; 95% CI: 95% studentized simultaneous confidence interval computed jointly across the three families; Folds: number of the five rolling-origin folds meeting the pre-registered 0.01 rule; p: one-sided Westfall–Young stepdown p-value; Equivalence: one-sided 95% upper bound of the equivalence test, computed per family and not adjusted for simultaneity; RER: residual error reduction. All three equivalence tests confirm the increment lies below the 0.01 materiality threshold.

The ablation results in Table 22 show that no single feature block carried a material increment in any family. The largest block mean was 0.0040 for the C1 monthly resolution, which remained below the threshold.

The label-permutation and sentinel controls both passed, so the null result was not an artifact of a broken pipeline. A sentinel target constructed to be predictable produced a pooled increment of 0.3108 under the same code path.

Table 22. Stage C block ablation results.

Family	Block	Mean ΔR^2	Features
C1 spending level	Cumulative DPFF	0.0035	50
C1 spending level	Recent window	0.0032	56
C1 spending level	Monthly	0.0040	53
C1 spending level	Indicators	0.0009	57
C2 spend change	Cumulative DPFF	0.0034	51
C2 spend change	Recent window	0.0009	57
C2 spend change	Monthly	0.0014	54
C2 spend change	Indicators	0.0020	58
C3 category-share entropy	Cumulative DPFF	0.0004	61
C3 category-share entropy	Recent window	0.0006	67
C3 category-share entropy	Monthly	0.0024	64
C3 category-share entropy	Indicators	−0.0000	65

Note: Each incremental R^2 is the mean across five rolling-origin test folds. Cumulative DPFF: cumulative DPFF features; Recent window: recent-window features; Monthly: monthly aggregate features; Indicators: indicator features. Features: number of features in the extension block for that family.

4.6. Pre-Registered Audit: Decision Utility

The decision-utility audit compared the extended feature set with the strong baseline. None of the three primary targets reached the pre-registered requirement of ΔAUC at least 0.01 with a positive confidence-limit lower bound in four of five folds (Table 23). The auxiliary adoption target had positive lower bounds in all folds, but its increments of 0.0081 to 0.0099 remained below 0.01. The results, therefore, did not establish incremental decision-relevant discrimination beyond the strengthened baseline. The limited increments are consistent with substantial information overlap between DPFF features and the lagged variables already included in the strong baseline.

Table 23. Decision-utility comparison against the strong baseline.

Target	Test Quarter	ΔAUC	95% CI
Exit	2023Q4	0.0003	[−0.0004, 0.0008]
	2024Q1	−0.0004	[−0.0008, 0.0000]
	2024Q2	−0.0005	[−0.0008, 0.0000]
	2024Q3	0.0003	[0.0000, 0.0007]
	2024Q4	0.0003	[−0.0001, 0.0008]
Abandonment	2023Q4	−0.0001	[−0.0008, 0.0007]
	2024Q1	−0.0001	[−0.0007, 0.0006]
	2024Q2	0.0001	[−0.0005, 0.0007]
	2024Q3	0.0014	[0.0008, 0.0020]
	2024Q4	−0.0004	[−0.0011, 0.0004]
Spending loss	2023Q4	0.0013	[0.0003, 0.0023]
	2024Q1	−0.0005	[−0.0017, 0.0008]
	2024Q2	0.0019	[0.0009, 0.0029]
	2024Q3	0.0045	[0.0033, 0.0056]
	2024Q4	−0.0001	[−0.0012, 0.0008]

Table 23. *Cont.*

Target	Test Quarter	ΔAUC	95% CI
Adoption (auxiliary)	2023Q4	0.0099	[0.0078, 0.0125]
	2024Q1	0.0081	[0.0057, 0.0107]
	2024Q2	0.0099	[0.0075, 0.0124]
	2024Q3	0.0083	[0.0056, 0.0110]
	2024Q4	0.0099	[0.0067, 0.0134]

Note: Across five folds. Held-out customer-quarter rows vary by target and fold from 23,773 to 91,326. Values are foldwise ΔAUC with paired-bootstrap 95% confidence intervals. AUC: area under the receiver operating characteristic curve.

An initially favorable exit result is retained only as an invalidated analysis. The retention–novelty feature file had been merged from a $t+1$ universe into the exit universe, leaving exit rows entirely missing; median replacement then encoded the label. The invalidated exit increments were 0.026 to 0.029. After amendment A2-5 regenerated retention–novelty features on the full history universe, the exit increments became -0.0005 to 0.0003. An independent recalculation reproduced the exit comparison within an absolute ΔAUC difference of 0.0004. The originally submitted specification with six DPFF features was also evaluated as a sensitivity analysis and did not alter the decision-utility conclusion.

4.7. Scorecard and Sensitivity

Table 24 consolidates the frozen criteria and observed outcomes. Of the three Stage A gates, only gate (iii) passed, through sub-criterion (iii-a). The pre-registered decision path for this outcome was withdrawal and repositioning; the authors instead elected to resubmit with narrowed claims (Section 5.4). All pre-registered criteria are reported unchanged.

Table 24. Pre-registered gate and amendment scorecard.

Test	Pre-Registered Criterion	Observed Result	Decision
Gate (i), break detection	2023Q1 $\lvert z\rvert \geq 3$, no baseline false positive, earlier than simple tracking	CCR $z = 10.10$; baseline false positive; comparator $z = 10.89$	Did not pass
Gate (ii), behavioral churn	Behavioral share ≥ 0.50 and standardized coefficient ≥ 0.10 in 4/5 folds	84.3%; coefficient criterion 0/5	Did not pass
Gate (iii-a), alignment	$\Delta R^2 \geq 0.01$ and CI lower bound > 0 in 4/5 folds	0.1356–0.1724; 5/5; residual 0.0038–0.0093 after lag block	Passed, with limit
Gate (iii-b), spending TE	$\Delta R^2 \geq 0.0005$ in 4/5 folds	-0.000194 to 0.000350; 0/5	Did not pass
P0-a, alignment lag baseline	$\Delta R^2 \geq 0.01$ and CI lower bound > 0 in 4/5 folds for each DPFF candidate above the lagged CCR block	Full DPFF residual 0.0038–0.0093; 0/5 for all four candidates	Did not pass
P0-b, domain ceiling	Maximum HMM/second-order/combined comparator $\Delta R^2 < 0.005$ in all 5 folds	0.0018–0.0048; residual DPFF 0.0004–0.0007	Capped
HMM override	HMM improvement over first order and nonpositive remaining DPFF increments for spending and alignment	Required nonpositive remaining increments not observed	Did not apply
P1, TE estimator	Spearman $\rho \geq 0.50$, RMSE ≤ 0.20, and absolute bias ≤ 0.05 in every $n \geq 4$ stratum; upper-boundary mass no more than 0.05 above the true mass at $n = 2, 3$	Absolute bias above the 0.05 rule in every candidate, with RMSE also above the 0.20 rule at low exposure in seven of the eight	Did not pass

Table 24. *Cont.*

Test	Pre-Registered Criterion	Observed Result	Decision
P2, decision utility	ΔAUC ≥ 0.01, CI lower bound > 0, in 4/5 folds	Primary targets 0/5; adoption 0.0081–0.0099	Did not pass
P3, nonhomogeneous Markov	Second-order log-loss improvement, CI lower bound > 0, in 4/5 folds	0/5 folds	Did not pass

Note: Scorecard for pre-registered gates and registered amendment analyses. Gate (iii) comprises two alternative sub-criteria, (iii-a) and (iii-b); (iii-a) met its own registered criterion. CCR: Category Churn Rate; TE: Transition Entropy; DPFF: Dynamic Persona Fluidity Framework; HMM: hidden Markov model; RMSE: root mean squared error; CI: 95% confidence interval; z: standardized score; ρ: Spearman correlation; ΔR^2: incremental coefficient of determination; ΔAUC: incremental area under the receiver operating characteristic curve; P0-a: alignment lag-baseline probe; P2: decision utility; P3: nonhomogeneous Markov analysis.

The auxiliary sensitivity measures in Table 25 are retained as diagnostics. Aggregate TE changed little when the minimum transition count increased from two to four, but this does not remedy the estimator failure at low exposure. PVS increased when rare categories were removed, so the full vocabulary remains the reference specification. Across 57 of the most prevalent categories, one within-customer contrast compared mean next-quarter spending when the category was present versus absent; seven had unadjusted $p < 0.05$, but none survived the Benjamini–Hochberg correction. These contrasts provide no multiplicity-controlled corroboration.

Table 25. Auxiliary-measure sensitivity analysis.

Minimum Consecutive Transitions	Customers	Mean Adjusted TE	Median Adjusted TE
2	177,102	0.617	0.766
3	137,013	0.623	0.731
4	106,934	0.619	0.715
6	62,083	0.583	0.658
Vocabulary	**Defined pairs**	**Excluded zero-vector pairs**	**Pair-level mean PVS**
Top 25	876,917	49,954	0.646
Top 50	907,947	18,924	0.596
Top 100	921,913	4958	0.557
All 318 categories	926,871	0	0.542

Note: The adjusted-TE panel reports the customer counts shown for each threshold; the PVS panel uses up to 926,871 calendar-consecutive pairs. TE: Transition Entropy; PVS: Persona Vector Similarity.

The online-active subset contained 152,333 customers (51.6% of 294,948). Its alignment increments were approximately one-third to one-half of the full-sample values (0.0420 to 0.0806 versus 0.1356 to 0.1724), and the spending-target CCR coefficient reversed sign. These are sensitivity observations rather than evidence of a separate effect. Together with the scorecard, they limit the results to the documented measurement and dynamic findings while leaving the predictive and decision-relevant discrimination criteria unmet.

5. Discussion

5.1. Summary of Audited Findings

This study established a gap-safe, exposure-aware description of observed category change in a cohort of 229,586 customers with calendar-consecutive quarter pairs. The 177,102 customers with estimable entropy remained a restricted diagnostic subset because no estimator met the pre-registered accuracy criterion across the relevant short sequences. Category Churn Rate (CCR), retention, novelty, Persona Persistence (PP), State Transition

Frequency (STF), and Persona Volatility Index (PVI) describe different operational aspects by construction; Table 9 tested every pairwise rank correlation and identified CCR–PVS as the only pair exceeding $|\rho| \geq 0.90$, although multivariate or factor-level redundancy was not assessed. Adjusted Transition Entropy (TE) remains a diagnostic quantity. The primary profiles condition CCR ranks on the exact number of observed transitions, which keeps the three profile groups balanced across transition-exposure tertiles. These measurement decisions address the concern that short histories can alter the scale of temporal summaries [25] rather than reveal a stable customer property.

The breadth-state results also establish two related empirical regularities. Time homogeneity was rejected for the four-state process, with $\chi^2(132) = 57{,}222.2$, and the second-order Markov model reduced held-out log loss relative to the first-order model in the original comparison (Table 14). The HMM benchmark matched the second-order model within ± 0.005 across the five folds, with paired 95% intervals excluding zero only in 2024Q1 and 2024Q2. The nonhomogeneous Markov models likewise captured quarter-specific change points, yet their log loss was worse than the second-order model in all five folds. Thus, the data contain both quarter-specific heterogeneity and second-order dependence, but the latter supplied the stronger held-out representation in this analysis [4,17].

The apparent 2023Q1 discontinuity requires a narrower interpretation. The within-vintage share of the CCR change was -0.175, and the balanced-cohort statistic was $z = 0.7744$, rather than the pre-specified evidence for a behavioral regime change. The result was consistent with a composition-led shift rather than a common change within continuing customers. This result limits any reading of the quarter-level pattern as a change in a latent customer attribute. It also shows why cohort composition must accompany time-series summaries of customer states [3].

Several propositions did not hold under the pre-registered audit. The corrected prospective and strengthened-baseline spending results are reported in Table 17, and the TE-over-CCR criterion passed in zero of five folds (Table 24). The alignment target initially passed, but its residual feature block after the lagged CCR block did not meet the registered criterion (Table 18). The alignment result, therefore, describes persistence in the CCR series, not an independent validation of the broader feature set.

The negative audit results extend beyond spending. The domain-ceiling comparison bounded the attainable spending increment below the pre-registered 0.005 rule for the registered comparators (Table 19), while a later pre-specified analysis found one quarterly fold above that level although the pooled increment remained below the 0.01 materiality threshold (Table 21), and the decision-utility, rule-based state-change, behavioral-decomposition, and TE-estimator analyses did not meet their respective pre-registered criteria (Table 24). These results delimit the framework as a measurement and audit exercise rather than evidence of incremental decision-relevant discrimination beyond the strengthened baseline.

Across the pre-specified quarterly outcomes (spending level, spending change, and category composition), learners (ridge and gradient boosting), feature resolutions (cumulative, recent-window, and monthly), and the $t + 1$ horizon, the fluidity feature family did not add a material increment ($\Delta R^2 \geq 0.01$) over a strengthened autoregressive baseline. The one-sided 95% upper bounds for the three families were 0.0038, 0.0017, and 0.0032 (Table 21). The largest of the three is 38% of the registered materiality threshold, so this is an equivalence result rather than an inconclusive one: the data are consistent with an incremental contribution that is bounded below materiality, not merely with a failure to detect one. Obtaining that bound required the registered comparator ladder, because the originally submitted specification produced a larger increment that survived neither the recency correction nor the strengthened baseline. The conclusion is limited to quarterly customer-level outcomes of a single card product; shorter resolutions, category-level out-

comes, and other issuers remain open. The audit tested transactional payment-category fluidity at quarterly resolution; it did not test, and, therefore, does not rule out, predictive utility for real-time clickstream data, search logs, or intent-level behavioral panels.

5.2. Implications for Dynamic-Segmentation Research

The audited null is reported because it compares observed estimates with thresholds fixed before the results were inspected. Section 4.5 reports the original and corrected spending increments, and Table 19 reports the comparator increments. In this data setting, the principal result is not that one dynamic representation defeated another, but that several representations shared a low incremental ceiling beyond the lagged-transaction baseline. This comparison is consistent with the need to assess dynamic customer models against explicit temporal comparators rather than a contemporaneous summary alone [4,17].

The alignment audit clarifies a second design issue. A target defined from the next-quarter category set naturally overlaps with lagged category-set change; Table 18 reports the lag-block R^2 values and residual increments. A positive association with an aligned target, therefore, does not by itself demonstrate information beyond the history already available at the prediction time. Dynamic-segmentation studies can, therefore, state the target, the information time, and the comparator ladder together. The distinction is especially important when set-overlap measures derive from related representations of the same transaction history [21,22].

The Markov comparisons further distinguish two questions that are often combined. The homogeneity test and the change-point analysis showed quarter-specific heterogeneity, while the held-out comparisons showed that the second-order model remained better than the nonhomogeneous first-order alternatives in five of five folds. This pattern supports evaluating time-varying specifications against higher-order models, not treating rejection of homogeneity as sufficient evidence for one modeling family. The fold-specific HMM comparison in Table 15 likewise distinguishes improvement over first-order prediction from the absence of a consistent gain over the second-order benchmark. Such model comparisons retain their descriptive value even when they do not produce an incremental spending result [4,17].

The rule decomposition adds a measurement boundary. Of the mappable category disappearances, 84.3% met the behavioral definition, but mapping coverage was 59.6% and the behavioral and total CCR measures correlated at 0.975. Consequently, a provider-defined label can retain rule-mechanical dependence even after a behavioral decomposition. The profile sensitivity analysis gives the same caution from another direction: absolute thresholds changed the High-profile share from 0.595 to 0.311 across exposure tertiles, whereas the exposure-conditioned primary definition held each group near one third. These findings support reporting the construction rule, exposure distribution, and alternative definitions as part of a dynamic-segmentation measurement claim [11–14].

The relevance of this evidence to electronic-commerce research is bounded by the same channel limitation. The corpus records provider-defined merchant categories, several of which are online-shopping categories, but it contains no transaction-level channel field, so an individual purchase cannot be assigned to an online or an offline channel. What transfers to an electronic-commerce setting is, therefore, the measurement and audit procedure rather than a channel-specific behavioral estimate. The gap-safe pair rule, the explicit transition exposure, the information-time restriction, and the pre-registered comparator ladder apply unchanged to clickstream, search-log, or order-level panels in which the channel is observed directly. The one part of the corpus that can be isolated on this axis behaves differently from the whole. Customers with activity in the online-shopping categories number 152,333 of 294,948 linked customers, or 51.6%, and within that subset the alignment increments

fall to approximately one-third to one-half of the full-sample values, 0.0420 to 0.0806 against 0.1356 to 0.1724, while the CCR coefficient on the spending target reverses sign. The subset, therefore, differs from the full cohort on this axis. Because it is defined by merchant category rather than by an observed channel field, a channel explanation and a composition explanation cannot be separated, so the contrast is reported as a reason to conduct a channel-observed replication rather than as an estimate of what such a replication would return. Section 5.4 states the corresponding boundary on interpretation.

No deployment recommendation follows from these results. The observed evidence supports bounded comparisons of measurement choices and information sets, but the decision-utility audit produced zero passing folds for each primary outcome. The framework can still serve as an audit report. For one portfolio, such a report can record the exposure-conditioned profile counts of 76,533, 76,530, and 76,523 in Table 11, the fold-specific benchmark log-loss evidence in Table 14, and the vintage-composition decomposition in Table 20; these entries describe measurement without authorizing targeting or intervention. The study, therefore, does not infer an intervention effect, profitability effect, or cross-product performance effect from the observed customer histories. Those distinctions follow the observational boundary of customer-relationship data and remain necessary even when a longitudinal measure is precisely defined [31,32].

5.3. Implications for Practice

A bounded null carries decision content that an inconclusive result does not. For an issuer weighing an investment in fluidity features for quarterly spending or attrition scoring on a single card product, the registered upper bounds place the attainable gain below the level at which the protocol had committed in advance to call a gain material. On this evidence, and at that registered threshold, the audit does not support funding that pipeline for these targets, and it states the comparator against which that assessment holds: a strengthened lagged-transaction baseline with fixed effects, not a contemporaneous summary. Whether to fund it remains a decision that also requires implementation costs, an operating cut-off, and realized consequences, none of which this study observes. Section 4.5 shows why that comparator matters, because the originally submitted specification produced a larger increment that did not survive the recency correction or the strengthened baseline.

The transferable deliverable is, therefore, the acceptance test rather than the feature set. A practitioner presented with a vendor or internal claim that dynamic segmentation improves a customer-level forecast can apply the same four steps: fix the target and the information time, require a comparator that already contains the lagged form of the target, register in advance the increment that would justify adoption, and evaluate on rolling origins with preprocessing refit inside each fold. This study reports what that test returns when it is applied to the features the study itself proposes, which is the condition under which such a test is informative to a reader who did not run it.

To make that test concrete, consider how it would enter a portfolio decision. An analyst scoring next-quarter attrition ranks customers first by the strengthened baseline alone and then by the baseline plus the fluidity block, and adopts the extended model only if the incremental discrimination clears the increment registered in advance. The registered criterion is stated on ΔAUC, a threshold-free summary of the whole ranking; a deploying team would additionally fix its own operating point and the costs attached to it, neither of which this study observes. The comparison is executable on the present data, and Table 23 reports what it returns: the auxiliary adoption target came closest, with fold increments of 0.0081 to 0.0099 and interval lower bounds above zero in all five folds, and it still did not reach the registered 0.01. The vignette, therefore, ends where the evidence ends. The

procedure is specified, the comparison is repeatable by a reader who did not run it, and on this portfolio the registered adoption threshold is not met.

The measurement layer retains an operational use that does not depend on the predictive result. Exposure-conditioned profiles hold the three groups near one third across transition-exposure tertiles, whereas absolute thresholds moved the High-profile share from 0.595 to 0.311 across those same tertiles. A portfolio-monitoring report that compares profile shares across quarters is, therefore, sensitive to which rule generated the labels, and the exposure-conditioned rule is the one that keeps the comparison interpretable across customers with different observation lengths. Tables 11 and 20 are the entries such a report would carry; they describe portfolio composition and do not authorize targeting or intervention.

This does not rebut the objection that an increment of 0.0011, or 0.0009 after the recency correction, is too small to matter. It concedes the objection and reports the bound at which the concession holds. The registered protocol fixed in advance that the increment would be reported whether or not it cleared the threshold, and that the failure branch would be disclosed. What the study adds under that commitment is the ceiling itself, the comparator that produced it, and the sensitivity of that ceiling to learner, feature resolution, and target definition. A study reporting only a within-sample association would not have established any of the three.

5.4. Limitations

The study measures provider-defined transaction profiles, not latent identity or total consumption. Data from one card product omit other cards, cash, and platforms, and the rule labels mechanically depend on observed transactions. An observed category disappearance may reflect cross-card or cross-platform spending reallocation rather than a change in underlying preference; within this boundary, the measures describe product-specific observed behavior. Selection into the product, activity, linkage, and survival (attrition) bias restrict external validity. The observational design does not identify the effect of a profile or an intervention on later spending.

After IPW trimming, substantial residual imbalance remained in activity-level covariates, including active quarters (SMD 0.548) and unique categories (SMD 0.321), so reweighting does not remove selection bias from the observed cohort.

The entropy results add a finite-sample limitation. None of the six estimators passed the pre-registered accuracy screens for the short exposure range, so adjusted TE was excluded from the score and primary profiles. The primary profiles also depend on a chosen definition: quartiles had Cramér's $V = 0.817$ with the exposure-conditioned grouping, whereas k-means had $\kappa = 0.287$. These results make profile labels conditional on the stated construction rather than fixed types. The rule decomposition has a parallel restriction because only 59.6% of disappearance events were mappable, even though the mappable behavioral proportion remained high.

The recency variable had a documented parsing defect in the earlier pipeline. The calculation used a quarter-end date that was offset by 91, 183, and 184 days in Q2, Q3, and Q4, respectively, while preserving ranks within a quarter. Correcting the dates raised the original baseline R^2 in all five folds and reduced the original-specification increment, as reported in Section 4.5. The strengthened baseline results were unchanged to the reported precision, but the original recency-based headline must be interpreted with this correction. The analysis also used a single product and a single institution, so external replication is required before claims about other payment settings.

The e-commerce boundary is partial rather than transaction-level. The data include categories associated with online shopping, but they contain no transaction-level on-

line/offline channel indicator. The online-active subset included 152,333 of 294,948 linked customers, or 51.6%, and its alignment increments were approximately one-third to one-half of the full-sample values (0.0420 to 0.0806 versus 0.1356 to 0.1724); its CCR coefficient also reversed sign relative to the full sample. These results do not establish that the full cohort represents electronic-commerce behavior. Because the corpus lacks a transaction-level channel indicator, this partial boundary neither confirms nor excludes the predictive utility of the audited measures in clickstream, search-log, or intent-level behavioral data, where the channel is observed directly. The discontinuity analysis has a separate implementation deviation because its primary High flag used an $n \geq 4$ pooled exposure bucket rather than exact-n strata. This deviation does not change the failed gate, but it limits comparison between that detector and the primary profile rule.

The pre-registration also requires a disclosure boundary. Its default P1-failure and P2-failure branches were withdrawal and repositioning, but the authors elected to resubmit with narrowed claims rather than withdraw. Reporting the complete pre-registered audit, including its failure criteria, for a large proprietary cohort remains informative to researchers and practitioners who might otherwise adopt fluidity features from within-sample associations alone. The resubmission follows the combined constraints of the P1-failure and P2-failure branches, including TE exclusion and no decision-utility claim. The data remain unavailable for public redistribution under the provider agreement, and reviewer access must remain compatible with that agreement. These restrictions limit independent end-to-end replication even though aggregate outputs, scripts, pre-registration records, and invalid executions are retained in the provenance materials.

6. Conclusions

This study set out to meet three objectives that mirror its stated contributions. The first was an exposure-aware measurement design that treats an observation gap and a behavioral transition as different events. Sections 4.1 and 4.3 report the calendar-consecutive pairing and exposure-conditioned profiles that meet that objective. The second was an evaluation of category-breadth dynamics through homogeneity tests and held-out order comparisons. Section 4.4 reports rejected homogeneity, a second-order advantage over first-order dependence, and no consistent gain from hidden-state or nonhomogeneous alternatives (Table 15). The third was a prespecified, disclosed audit of incremental predictive and decision value. Sections 4.5 and 4.6 report that audit in full, including its failed gates. All three objectives were met.

The headline finding is plain. The measurement system is documented and auditable, but the fluidity features examined here do not add material predictive or decision value beyond a strong lagged-transaction baseline. The audit bounds that increment rather than merely failing to detect it. The second-order advantage in category-breadth dynamics, and the heterogeneity behind it, are technical facts about the data-generating process, not a predictive edge. The hidden Markov and nonhomogeneous comparators did not consistently improve on the second-order benchmark across the held-out folds. The predictive increment is a bounded null: the confirmatory re-analysis kept its upper bound below the pre-registered materiality threshold in every family (Table 21), and the decision-utility and rule-based audits recorded the same outcome (Table 24).

These results carry different implications for five audiences. For card-issuer analytics and CRM teams, the registered acceptance test in Section 5.3 is the transferable tool, and on this evidence, at the registered materiality threshold and without the cost and operating-point data this study does not observe, a fluidity-feature pipeline for quarterly spending or attrition scoring on a single card product is not a supported investment. For portfolio and retention managers, exposure-conditioned profiles remain a monitoring descriptor;

their shares stay interpretable across customers with different observation lengths, but they do not authorize targeting or intervention. For analytics vendors and their buyers, a persona-value claim can be examined with the same acceptance test rather than accepted on a vendor's internal validation. For methodologists and dynamic-segmentation researchers, the comparator ladder, information-time discipline, and pre-registration protocol are reusable contributions independent of this dataset's null result. For e-commerce researchers and practitioners with channel-observed data, the measurement and audit procedure can be applied to search logs, clickstream records, and order-level panels, whereas the bounded-null estimates reported here do not transfer.

A decision-maker who receives a persona-value claim built on this framework should require the four-step acceptance test in Section 5.3 before acting on it, above all a comparator that already contains the lagged form of the target and an adoption increment registered before the result is inspected. The exposure distribution behind any profile label should be disclosed with that label. This evidence licenses no deployment decision on its own. A later comparison is interpretable against the present audit only if it retains the calendar-consecutive cohort rule, reports transition exposure, and distinguishes the full cohort from the smaller entropy-estimable subset used for adjusted TE. It should also retain the corrected recency construction and report an invalid analysis rather than replace it with a favorable estimate, as Section 4.6 documents.

These conclusions hold for transactional payment-category fluidity observed at a quarterly resolution. Whether clickstream data, search logs, or intent-level behavioral panels carry predictive utility was not tested here and remains open to a channel-observed replication. Future research should replicate the measurement and audit in independent institutions, card products, and data systems that identify transaction channels. It should also evaluate time-varying, higher-order, and latent-state models under the same rolling-origin folds and information-time restrictions used here, retaining a lagged-transaction comparator [4,17]. Future work should also compare rule definitions with independently documented behavioral categories and report mapping coverage and sensitivity to profile construction. A separate line of work should test decision utility in prespecified experiments with observed outcomes, costs, and intervention assignments, a utility that this observational study cannot identify [31,32].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jtaer21090336/s1, Document S1: Stage A pre-registration; Document S2: Stage A2 pre-registration and amendments; Document S3: Stage C registered re-analysis protocol. All three are English translations of the Korean protocol documents frozen in the authors' versioned provenance package.

Author Contributions: Conceptualization, Y.H.H., B.K. and S.-S.K.; methodology, Y.H.H., B.K. and S.-S.K.; investigation, Y.H.H., B.K. and S.-S.K.; formal analysis, Y.H.H., B.K. and S.-S.K.; writing—original draft preparation, Y.H.H., B.K. and S.-S.K.; writing—review and editing, Y.H.H., B.K. and S.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (RS-2021-NR060140), and by the ANCHOR program through the Seoul ANCHOR Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government, Republic of Korea (2026-ANCHOR-01-020-04).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable; the study used de-identified secondary records and involved no direct interaction with cardholders.

Data Availability Statement: Restrictions apply to the availability of these data. The de-identified transaction and persona records were obtained from a financial institution under a data-utilization competition agreement and are not publicly available. Requests for access should be directed to the corresponding author and remain subject to the data provider's permission and terms. Source and output hashes, aggregate cohort flow, exact transition counts and probabilities, metric summaries, sensitivity results, and figure source data are retained in a versioned provenance package and can be supplied to editors and reviewers under provider-compatible access conditions. The Stage A and Stage A2 pre-registration documents and the experiment ledger are retained in that provenance package, and English translations of the Stage A, Stage A2, and Stage C protocol documents are provided as Supplementary Documents S1–S3. Customer-level records and derived person-level metrics remain restricted. Independent end-to-end reproduction, therefore, requires authorized access to the source corpus.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI Codex CLI (version 0.150.1) and Anthropic Claude Code (version 2.1.251) for language editing, code review, and manuscript-format checks. The authors reviewed and edited all outputs and take full responsibility for the content of the publication.

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

References

1. Wedel, M.; Kamakura, W.A. *Market Segmentation: Conceptual and Methodological Foundations*, 2nd ed.; Springer: New York, NY, USA, 2000. [CrossRef]
2. Hughes, A.M. *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program*; Probus Publishing: Chicago, IL, USA, 1994.
3. Blocker, C.P.; Flint, D.J. Customer segments as moving targets: Integrating customer value dynamism into segment instability logic. *Ind. Mark. Manag.* **2007**, *36*, 810–822. [CrossRef]
4. Pfeifer, P.E.; Carraway, R.L. Modeling customer relationships as Markov chains. *J. Interact. Mark.* **2000**, *14*, 43–55. [CrossRef]
5. Verhoef, P.C.; Kannan, P.K.; Inman, J.J. From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *J. Retail.* **2015**, *91*, 174–181. [CrossRef]
6. Lemon, K.N.; Verhoef, P.C. Understanding customer experience throughout the customer journey. *J. Mark.* **2016**, *80*, 69–96. [CrossRef]
7. Moe, W.W.; Fader, P.S. Dynamic conversion behavior at e-commerce sites. *Manag. Sci.* **2004**, *50*, 326–335. [CrossRef]
8. Lee, M.S.; Han, Y.H. From static segmentation to dynamic profiling: Customer behavior pattern analysis through fluidity metrics. *J. Creat. Inf. Cult.* **2026**, *12*, 125–137. (In Korean) [CrossRef]
9. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; Le Cam, L.M., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
10. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
11. Viviani, J.L.; Komura, A.; Suzuki, K. Integrating dynamic segmentation and portfolio theories for better customer portfolio performance. *J. Strateg. Mark.* **2023**, *31*, 140–153. [CrossRef]
12. Abbasimehr, H.; Bahrini, A. An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. *Expert Syst. Appl.* **2022**, *192*, 116373. [CrossRef]
13. Wu, T.; Liu, X. A dynamic interval type-2 fuzzy customer segmentation model and its application in e-commerce. *Appl. Soft Comput.* **2020**, *94*, 106366. [CrossRef]
14. Zhu, T.; Liu, Y. Learning personalized preference: A segmentation strategy under consumer sparse data. *Expert Syst. Appl.* **2023**, *215*, 119333. [CrossRef]
15. Schweidel, D.A.; Bradlow, E.T.; Fader, P.S. Portfolio dynamics for customers of a multiservice provider. *Manag. Sci.* **2011**, *57*, 471–486. [CrossRef]
16. Berchtold, A. High-order extensions of the Double Chain Markov Model. *Stoch. Models* **2002**, *18*, 193–227. [CrossRef]
17. Netzer, O.; Lattin, J.M.; Srinivasan, V. A hidden Markov model of customer relationship dynamics. *Mark. Sci.* **2008**, *27*, 185–204. [CrossRef]
18. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]

19. Vickers, A.J.; Elkin, E.B. Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* **2006**, *26*, 565–574. [CrossRef] [PubMed]
20. Nosek, B.A.; Ebersole, C.R.; DeHaven, A.C.; Mellor, D.T. The preregistration revolution. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2600–2606. [CrossRef] [PubMed]
21. Real, R.; Vargas, J.M. The probabilistic basis of Jaccard's index of similarity. *Syst. Biol.* **1996**, *45*, 380–385. [CrossRef]
22. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
23. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
24. Miller, G.A. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*; Quastler, H., Ed.; Free Press: Glencoe, IL, USA, 1955; pp. 95–100.
25. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef]
26. Chao, A.; Shen, T.J. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* **2003**, *10*, 429–443. [CrossRef]
27. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems 14*; Dietterich, T.G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; pp. 471–478.
28. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1977**, *39*, 1–22. [CrossRef]
29. Yao, Y.C. Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.* **1988**, *6*, 181–189. [CrossRef]
30. Horvitz, D.G.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **1952**, *47*, 663–685. [CrossRef]
31. Fader, P.S.; Hardie, B.G.S.; Lee, K.L. RFM and CLV: Using iso-value curves for customer base analysis. *J. Mark. Res.* **2005**, *42*, 415–430. [CrossRef]
32. Ascarza, E.; Neslin, S.A.; Netzer, O.; Anderson, Z.; Fader, P.S.; Gupta, S.; Hardie, B.G.S.; Lemmens, A.; Libai, B.; Neal, D.; et al. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Cust. Needs Solut.* **2018**, *5*, 65–81. [CrossRef]