Perturbed Utility Based Flow Estimation from Link-Based Person-Time Measures

Jingran Su*[1] and Eiji Hato[1]

[1]Department of Civil Engineering, The University of Tokyo, Japan

Short summary

Indoor pedestrian route choice analysis in large three-dimensional spaces is difficult because Wi-Fi-based observations are noisy and path-level map matching is often unreliable. This paper proposes a unified framework based on a probabilistic link-based person-time measure derived from uncertain Wi-Fi observations. The model jointly estimates pseudo-OD-specific directed flow and perturbed-utility-based behavioral parameters, while LLM-derived semantic attributes are included as contextual covariates. Using the Shibuya Station pedestrian network, the framework closely reproduces the observed person-time field and yields stable, interpretable route choice parameters, indicating that person-time-based flow estimation provides a practical basis for indoor pedestrian behavior analysis under observational uncertainty.
Keywords: directed flow distribution; indoor pedestrian behavior; inverse flow estimation; route choice; Wi-Fi observation

1 Introduction

Route choice modeling in complex pedestrian environments is commonly built on a path-based logic: individual trajectories are first reconstructed from observations and then used to estimate behavioral parameters. This approach becomes fragile in large three-dimensional indoor–outdoor pedestrian environments because, although ITS sensing technologies such as Wi-Fi and BLE offer superior indoor availability relative to GPS, signal attenuation, multipath propagation, and vertical separation still introduce substantial observational uncertainty, complicating both localization and path-level map matching (Su & Hato, 2025).

Even when indoor localization is sufficiently accurate to identify plausible links, deterministic path reconstruction remains sensitive to short-term classification errors and missing observations (Mai et al., 2023; Xiong et al., 2021). This motivates a shift from path recovery to direct network-flow inference from observations. Related studies have inferred path flows or OD demand from aggregate link observations such as link counts or link flow densities (Bell, 1991; Lo & Chan, 2003; de Grange et al., 2017), while pedestrian studies using Wi-Fi probe or RSS data have mainly focused on flow monitoring, path identification, and occupancy prediction (Fukuzaki et al., 2014; Tsiamitros et al., 2023; Yang et al., 2022).

Another challenge is that route choice models often rely on a limited set of manually defined variables, which may not adequately capture latent contextual features or complex interaction effects in indoor environments. Recent studies have shown that large language models can extract semantic embeddings for geographic context and route choice analysis (He et al., 2024; Qiu et al., 2025), although these high-dimensional representations are not directly suitable as covariates in conventional route choice models.

This paper proposes a unified framework centered on a probabilistic link-based person-time measure from uncertain Wi-Fi observations. Rather than treating individual paths or observed link flows as the primary object, the framework uses this measure as a basis for directed flow estimation and perturbed-utility-based behavioral interpretation. The resulting approach provides a coherent link from uncertain observations to network-feasible flow and route-choice representations. LLM-derived semantic attributes are further incorporated as low-dimensional covariates to capture contextual information that is hard to represent with manually defined variables alone.

2 Methodology

The proposed pipeline has four stages: (i) link observation probabilities from Wi-Fi fingerprints, (ii) person-time measure construction, (iii) measure-based inverse flow modeling through alternating joint estimation of flow and behavioral coefficients, and (iv) behavioral specification enriched by LLM-derived semantic attributes.

Observation probabilities

Indoor pedestrian location is observed through Wi-Fi fingerprint measurements. At observation time t for trip i, the raw signal is represented by an RSSI fingerprint vector

$$\mathbf{r}_{it} = (r_{it1}, \ldots, r_{itM}),$$

where r_{itm} is the received signal strength from access point m and M is the number of observed access points. We adopt GConvLoc (Kim & Suh, 2023) as the observation model because it predicts link probabilities by exploiting similarity among reference-point fingerprints on a graph.
Given the RSSI fingerprint $\mathbf{r}_{it}$, the model outputs a posterior probability distribution over candidate links,

$$p_{it}(l) = P(l \mid \mathbf{r}_{it}), \qquad l \in \mathcal{L},$$

where $\mathcal{L}$ is the set of undirected links. These probabilities form the direct input to person-time measure construction.

Person-Time Measure Construction

The methodological core of the paper is the translation from soft link observation probabilities to a network-feasible directed flow distribution through link-level person-time measures. To obtain stable demand units, network nodes are first clustered into pseudo-zones, and each trip is assigned to a pseudo-origin-destination pair $k \in \mathcal{K}$ according to its origin and destination links. Let q_k denote the observed number of trips in pseudo-OD pair k,

$$q_k = \sum_i \mathbf{1}\{\text{trip } i \text{ belongs to pseudo-OD } k\}.$$

Because each pseudo-zone represents an area rather than a single gate, the OD demand is not injected at one node only. Instead, the demand of pair k is split equally across all nodes in the origin zone and absorbed equally across all nodes in the destination zone.
Using the observation interval Δt_{it}, the soft link probabilities are aggregated within each pseudo-OD pair into the **OD-specific person-time measure**

$$y_{kl} = \sum_{i \in \mathcal{T}_k} \sum_t \Delta t_{it}\, p_{it}(l), \qquad k \in \mathcal{K},\ l \in \mathcal{L},$$

where $\mathcal{T}_k$ denotes the set of trips belonging to pseudo-OD pair k. Since the link probabilities at each observation time sum to one, they represent one pedestrian distributed fractionally across candidate links; temporal aggregation therefore gives the expected link occupancy time on each link in person-seconds. The corresponding aggregate person-time measure is

$$y_l = \sum_{k \in \mathcal{K}} y_{kl}, \qquad l \in \mathcal{L}.$$

To relate these observed person-time measures to network flow, each physical link is duplicated into two directed arcs. Let $\mathcal{A}$ denote the directed arc set, let $x_{ka} \geq 0$ be the latent directed flow on arc a for pseudo-OD pair k, let P_{la} map directed arcs to their corresponding undirected links, and let τ_a denote the nominal traversal time of arc a,

$$\tau_a = \frac{\ell_a}{v_0},$$

where ℓ_a is the length of arc a, and $v_0 = 1.2$ m/s is a standard pedestrian walking-speed reference used to convert directed flow into person-seconds. The corresponding flow-implied OD-specific person-time measure is written as

$$\hat{y}_{kl} = \sum_{a \in \mathcal{A}} P_{la} \tau_a x_{ka}, \qquad k \in \mathcal{K},\ l \in \mathcal{L},$$

with aggregate measure

$$\hat{y}_l = \sum_{k \in \mathcal{K}} \hat{y}_{kl}, \qquad l \in \mathcal{L}.$$

When residual person-time is represented explicitly, the slack-augmented fitted person-time measure is defined as

$$\tilde{y}_{kl} = \hat{y}_{kl} + s_{kl}, \qquad \tilde{y}_l = \sum_{k \in \mathcal{K}} \tilde{y}_{kl}.$$

Measure-based inverse flow modeling

The estimation combines three elements: network-feasible flow conservation, person-time fit, and PURC-based behavioral regularization. Let B be the node-arc incidence matrix and b_k the OD supply vector. For each pseudo-OD pair, flow conservation is imposed as

$$Bx_k = q_k b_k, \qquad \forall k \in \mathcal{K}.$$

Here, $b_k(n) = -1/|O_k|$ for origin-zone nodes $n \in O_k$, $b_k(n) = +1/|D_k|$ for destination-zone nodes $n \in D_k$, and $b_k(n) = 0$ otherwise. To absorb residual mismatch between the directed-flow representation and the observed person-time measure, a nonnegative OD-specific slack term s_{kl} is introduced. The behavioral component follows the PURC transformation through the OD-specific directed flow density

$$\rho_{ka} = \frac{x_{ka}}{q_k \ell_a},$$

and the perturbed response

$$Y_{ka}(x) = \log\left(1 + \rho_{ka}\right) = \log\left(1 + \frac{x_{ka}}{q_k \ell_a}\right).$$

Let Z_a denote the vector of link attributes and let β denote the PURC coefficient vector. The corresponding systematic utility index is

$$\eta_a = Z_a^\top \beta.$$

At alternating iteration m, the behavioral response is linearized around the current flow $x^{(m)}$ as

$$Y_{ka}(x) \approx Y_{ka}^{(m)} + g_{ka}^{(m)} \left(x_{ka} - x_{ka}^{(m)}\right),$$

where

$$Y_{ka}^{(m)} = \log\left(1 + \frac{x_{ka}^{(m)}}{q_k \ell_a}\right), \qquad g_{ka}^{(m)} = \frac{1}{q_k \ell_a + x_{ka}^{(m)}}.$$

Given the current coefficient vector $\beta^{(m)}$, the flow update solves

$$\begin{aligned} \min_{x \geq 0,\, s \geq 0} \quad & \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}} \left(\hat{y}_{kl} + s_{kl} - y_{kl}\right)^2 \;+\; \lambda_x \sum_{k \in \mathcal{K}} \sum_{a \in \mathcal{A}} x_{ka}^2 \\ & +\; \lambda_s \sum_{k \in \mathcal{K}} \sum_{l \in \mathcal{L}} s_{kl}^2 \;+\; \lambda_b \sum_{k \in \mathcal{K}} \sum_{a \in \mathcal{A}} q_k \left(Y_{ka}^{(m)} + g_{ka}^{(m)} \left(x_{ka} - x_{ka}^{(m)}\right) - Z_a^\top \beta^{(m)}\right)^2. \end{aligned} \tag{1}$$

subject to

$$\begin{aligned} Bx_k &= q_k b_k, && \forall k \in \mathcal{K}. \\ x_{ka} &\geq 0, && \forall k \in \mathcal{K},\ a \in \mathcal{A}. \\ s_{kl} &\geq 0, && \forall k \in \mathcal{K},\ l \in \mathcal{L}. \end{aligned}$$

The first term matches the fitted OD-specific person-time measures $\tilde{y}_{kl}$ to the observed OD-specific person-time measures, the second term regularizes OD-specific directed flow, the third penalizes residual person-time, and the fourth encourages consistency with the PURC behavioral specification. Within the same alternating scheme, the coefficient vector is updated by weighted least squares,

$$\beta^{(m+1)} \in \arg\min_{\beta} \sum_{k \in \mathcal{K}} \sum_{a \in \mathcal{A}} q_k \left(\log\left(1 + \frac{x_{ka}^{(m+1)}}{q_k \ell_a}\right) - Z_a^\top \beta\right)^2.$$

where the weight is the pseudo-OD trip count q_k. These two updates are alternated until both the directed flow and the behavioral coefficients stabilize. The resulting formulation therefore estimates flow and coefficients jointly rather than sequentially.
Using the aggregate observed, fitted, and movement-only person-time measures defined above, the fitted person-time RMSE metrics are

$$\mathrm{RMSE}^{\mathrm{fit}}_{\mathrm{total}} = \sqrt{\frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}(\tilde{y}_l - y_l)^2},$$

$$\mathrm{RMSE}^{\mathrm{fit}}_{\mathrm{OD}} = \sqrt{\frac{1}{|\mathcal{K}||\mathcal{L}|}\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}}(\tilde{y}_{kl} - y_{kl})^2}.$$

To distinguish the explanatory power of movement flow from the total fitted person-time field, the movement-only RMSE metrics are defined as

$$\mathrm{RMSE}^{\mathrm{move}}_{\mathrm{total}} = \sqrt{\frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}(\hat{y}_l - y_l)^2},$$

$$\mathrm{RMSE}^{\mathrm{move}}_{\mathrm{OD}} = \sqrt{\frac{1}{|\mathcal{K}||\mathcal{L}|}\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}}(\hat{y}_{kl} - y_{kl})^2}.$$

The slack ratio is defined as

$$\text{Slack Ratio} = \frac{\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} s_{kl}}{\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} y_{kl}}.$$

The correlation measure is the Pearson correlation between the observed and fitted aggregate person-time measures,

$$\mathrm{Corr}_{\mathrm{fit}} = \mathrm{Corr}(\{y_l\}_{l\in\mathcal{L}}, \{\tilde{y}_l\}_{l\in\mathcal{L}}).$$

Perturbed utility specification

The PURC component of the model is specified through the link-level covariate vector Z_a, which enters the behavioral component of the joint estimation through the linear predictor $Z_a^\top \beta$. Here, Z_a denotes the explanatory attributes associated with directed arc a, and β is the corresponding coefficient vector. Under the soft joint formulation, these coefficients are not estimated in a separate post hoc step; instead, they are updated jointly with the OD-specific directed flow through the alternating estimation procedure described above.
Different perturbed utility specifications are defined by the choice of variables included in Z_a. In the empirical analysis, Model A uses structural variables only, Model B adds environmental variables, and Model C further augments the specification with residualized semantic factors derived from LLM embeddings. The three specifications are summarized in Table 1, and the detailed variable definitions are reported in Table 2; the corresponding attribute construction is described in the following subsection.

Table 1: Specification summary for the three PURC models.

Included parameter set	Model A	Model B	Model C
Structural variables	☺	☺	☺
Environmental variables	–	☺	☺
Semantic attributes	–	–	☺

Link attributes and semantic augmentation

The explanatory variables are defined at the link level and organized into three groups: structural, environmental, and semantic attributes. Structural attributes describe the geometric, topological, and infrastructural properties of pedestrian links derived from the 3D network. Environmental attributes describe the surrounding context of a link rather than the link geometry itself, including retail frontage, transport adjacency, and local disruption based on nearby OSM and polygon-based spatial data.

Table 2: Descriptions of each explanatory variable.

Variable	Description
Structural Variables	
log length	Natural logarithm of link length [log m].
underground	Underground location indicator.
stair	Stair connection indicator.
escalator	Escalator connection indicator.
indoor	Indoor location indicator.
sidewalk	Sidewalk or pedestrian-space score [-].
flow	Baseline link flow-potential score [-].
closeness	Link-level closeness-type centrality derived from endpoint node closeness [-].
betweenness	Link-level betweenness centrality [-].
Environmental Variables	
shop	Retail frontage or shop intensity score [-].
crossing-signal	Presence of nearby crossing or traffic signal.
disruption	Presence of nearby disruption or constrained condition.
transport	Presence of nearby transport facility or station interface.
LLM Latent Factors	
factor 1–20	Residualized latent factors derived from PCA-compressed LLM link embeddings.

Semantic attributes are introduced to capture contextual information that is difficult to represent with a small set of manually defined variables. Following the prompt-generation idea of LLMGeoVec, we construct a structured text prompt for each pedestrian link rather than for an individual point. The prompt structure incorporates:

- **Instruction:** a short task description that asks the model to identify the geographic and urban context of a pedestrian link.
- **Link description:** physical and infrastructural information from the 3D pedestrian network, including link type, length, floor transition, elevation change, and other link attributes.
- **Address:** endpoint coordinates and reverse-geocoded start and end addresses describing the location hierarchy of the link.
- **Nearby places:** nearby tagged map objects around the link midpoint, including amenities, shops, and transport-related places from historical OpenStreetMap queries.

Data sources include OpenStreetMap (Neis & Zipf, 2012), with addresses derived through Nominatim reverse geocoding (Serere et al., 2023) and nearby places queried through the Overpass API (Olbricht, 2011). We embed these prompts using the Alibaba NLP model `gte-Qwen2-1.5B-instruct`, a general-purpose multilingual text embedding model from the GTE family (Li et al., 2023; Alibaba-NLP, 2024). The embeddings are then compressed by principal component analysis into a smaller set of semantic factors, and the retained factors are residualized with respect to the structural and environmental variables so that they capture additional contextual variation beyond the observed covariates.

3 Results and discussion

Case Study and Experimental Setting

The case study is based on the Shibuya Station area in Tokyo using pedestrian observation data collected in 2017. The study area covers indoor and outdoor pedestrian passageways around the station across multiple levels, and the analysis network contains 185 pedestrian links. The experiment involved 40 volunteers aged 18 to 68 years with a 1:1 gender ratio, and all participants were familiar with the layout of Shibuya Station. Figure 1 shows the network topology used in this paper.

For the observation model, the Wi-Fi dataset comprises 18,101 labeled training samples and 4,525 test samples, and each sample is represented by 8,955 RSSI signals. Hyperparameters are selected

by 5-fold cross-validation and grid search. The adopted GConvLoc classifier uses a two-layer graph attention network trained with cross-entropy loss for 100 epochs, and outputs a probability vector over candidate pedestrian links for each observation record. On the test set, GConvLoc achieved 96.29% overall accuracy, with 96.65% indoor accuracy and 95.42% outdoor accuracy, and these soft link probabilities were used as the input to the subsequent flow estimation.

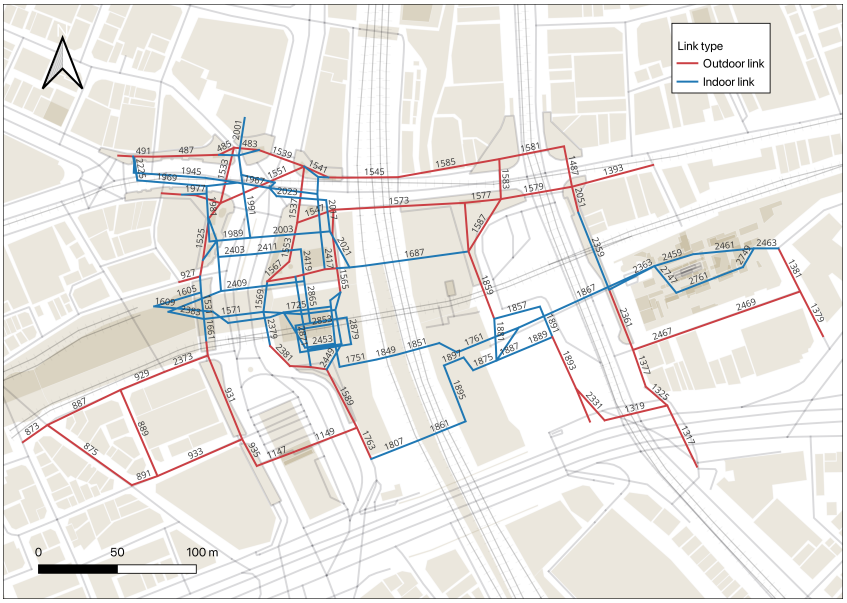


Figure 1: Pedestrian network topology in the Shibuya Station area. Blue links indicate indoor segments, red links indicate outdoor segments, and the background is from OpenStreetMap.

Joint Estimation Results

The proposed model is estimated under the OD-specific soft joint formulation with $\lambda_b = 300$, $\lambda_x = 10^{-3}$, and $\lambda_s = 10^{-2}$. Table 3 summarizes fitted person-time accuracy and movement-only explanatory accuracy for the three nested PURC specifications. The fitted person-time RMSE values are nearly unchanged across Models A–C, indicating stable reconstruction of the observed field under the common pseudo-OD structure, conservation constraints, and observation model. By contrast, movement-only RMSE decreases as richer covariate sets are introduced, while the adjusted PURC R^2 rises from 0.106 in Model A to 0.140 in Model B and 0.195 in Model C. The slack ratio remains below 2% in all cases.
Figure 2 illustrates representative pseudo-OD flow-share patterns under the soft joint Model C specification. The estimated flows concentrate on coherent corridor structures between pseudo-origin and pseudo-destination zones across different demand levels rather than diffuse local patches. This supports the interpretation that the framework recovers behaviorally plausible corridor-level movement while preserving the pseudo-OD structure implied by the observed person-time measures.

Table 3: Soft joint estimation results for the OD-specific model under $\lambda_b = 300$, $\lambda_x = 10^{-3}$, and $\lambda_s = 10^{-2}$. The total observed person-time measure is 201,874 person-seconds.

Model	Corr. fit	RMSE total fit	RMSE OD fit	RMSE total movement	RMSE OD movement	Adj. PURC R^2	Slack ratio
Model A	**0.9947**	**288.792**	11.3864	287.917	11.6437	0.1057	1.93%
Model B	0.9947	288.795	11.3864	287.888	11.6356	0.1398	1.86%
Model C	0.9947	288.795	**11.3864**	**287.005**	**11.6201**	**0.1947**	**1.78%**

Behavioral Interpretation

Table 4 reports the final converged PURC coefficients obtained under each joint specification. These coefficients are the stabilized β values from the alternating estimation procedure rather



Figure 2: Pseudo-OD flow-share examples from the soft joint Model C specification.

than post hoc parameters fitted to a fixed flow field. The semantic factors represent additional link context extracted from LLM-based text embeddings.

Table 4 shows that the estimated coefficients are stable and behaviorally interpretable across the nested PURC specifications. Log length and underground location remain strongly negative in all models, indicating a consistent preference for shorter and less enclosed routes. Stair links are positively associated with route choice, whereas escalator links are negative once contextual controls are introduced. Among the structural measures added in Models B and C, both closeness and betweenness are positive, suggesting that the jointly estimated flows concentrate on topologically important connectors.

The environmental and semantic variables further refine this pattern. Retail frontage is positively associated with route choice, while crossing-signal, disruption, and transport-adjacent links are negative. The explanatory power increases from Model A to Model C, indicating that environmental and semantic attributes provide additional information beyond the basic structural variables.

Sensitivity Analysis

Table 5 reports sensitivity to the behavioral regularization weight under the Model C specification. The main fit metrics are based on the fitted person-time measure, while the movement-only metrics isolate the contribution of directed flow itself. For $\lambda_b = 0$, the PURC coefficient update is retained only as a diagnostic baseline and does not affect the flow solution. As λ_b increases, PURC R^2 rises monotonically, but the slack ratio also grows. The no-slack values, shown in parentheses, indicate that removing slack keeps the fitted person-time RMSE slightly lower, but at large behavioral weights it weakens the movement contribution of the implied flow. In particular, when $\lambda_b = 10{,}000$, the fitted total and OD-level RMSE remain relatively stable, whereas the movement-only RMSE deteriorates sharply, showing that the slack term becomes important as a residual channel that preserves overall person-time fit when strong behavioral regularization would otherwise distort the flow field.

Table 6 evaluates sensitivity to pseudo-OD zoning under the main Model C setting with $\lambda_b = 300$. Increasing the number of zones improves OD-level fit because the decomposition becomes more spatially detailed, but it also reduces support per active pseudo-OD pair and weakens aggregate fit and behavioral explanatory power. Coarser zoning therefore gives an easier inverse problem, whereas finer zoning yields higher OD resolution at the cost of weaker support and greater dependence on residual adjustment. On this basis, 15 pseudo-zones provides a reasonable compromise.

Table 4: Final converged PURC coefficients from the soft joint estimation. Total observations = 18,097.

	Model A		Model B		Model C	
	est.	z-val	est.	z-val	est.	z-val
Constant	0.0052	18.46	0.0072	19.70	0.0072	19.85
Structural Parameters						
log length	-0.0034	-13.00	-0.0026	-10.66	-0.0026	-10.97
underground	-0.0064	-12.92	-0.0070	-12.87	-0.0070	-13.56
stair	0.0039	5.08	0.0044	6.00	0.0044	6.37
escalator	-0.0016	-2.49	-0.0015	-2.18	-0.0015	-2.24
indoor	0.0016	3.62	0.0005	1.00	0.0005	1.08
sidewalk	0.0011	8.01	0.0007	4.02	0.0008	4.28
flow	0.0003	2.72	0.0000	0.10	0.0000	0.11
closeness	–	–	0.0015	10.17	0.0015	10.67
betweenness	–	–	0.0012	5.77	0.0012	6.01
Environmental Parameters						
shop	–	–	0.0011	4.97	0.0011	5.24
crossing-signal	–	–	-0.0014	-4.24	-0.0014	-4.22
disruption	–	–	-0.0023	-3.27	-0.0023	-3.43
transport	–	–	-0.0029	-11.12	-0.0029	-11.14
LLM Latent Factors						
factor 1	–	–	–	–	-0.0004	-1.91
factor 2	–	–	–	–	-0.0004	-2.03
factor 3	–	–	–	–	-0.0021	-8.66
factor 4	–	–	–	–	-0.0017	-9.40
factor 5	–	–	–	–	-0.0001	-0.49
factor 6	–	–	–	–	-0.0009	-2.22
factor 7	–	–	–	–	-0.0000	-0.22
factor 8	–	–	–	–	-0.0009	-6.53
factor 9	–	–	–	–	-0.0002	-1.17
factor 10	–	–	–	–	-0.0030	-9.29
factor 11	–	–	–	–	-0.0003	-2.36
factor 12	–	–	–	–	-0.0005	-1.98
factor 13	–	–	–	–	-0.0006	-1.81
factor 14	–	–	–	–	-0.0003	-1.74
factor 15	–	–	–	–	-0.0005	-3.89
factor 16	–	–	–	–	0.0002	2.69
factor 17	–	–	–	–	0.0008	6.14
factor 18	–	–	–	–	0.0001	0.78
factor 19	–	–	–	–	-0.0000	-0.36
factor 20	–	–	–	–	0.0003	2.79
Adj. R^2	0.1057		0.1398		0.1947	

Semantic Factor Interpretation

The semantic factors in Model C are residualized latent variables derived from the LLM-based link embeddings and are intended to capture contextual information beyond the manually defined structural and environmental attributes. Their contribution appears mainly on the behavioral side: relative to Model B, Model C leaves fitted person-time RMSE essentially unchanged while increasing the adjusted PURC R^2 from 0.1398 to 0.1947.

Table 7 shows that several latent factors are statistically significant and comparable in magnitude to the manually defined continuous covariates. In particular, Factors 10, 3, and 8 have coefficient magnitudes similar to or larger than variables such as sidewalk, shop, closeness, and betweenness. Because the latent factors are residualized with respect to the observed structural and environmental variables, they should be interpreted in two complementary ways. First, they capture additional semantic context from linked geographic descriptions, OSM-based surroundings, and nearby place information that is not adequately summarized by the manually constructed variables alone. Sec-

Table 5: Sensitivity to the behavioral regularization weight under Model C. Plain values use the baseline slack-enabled formulation; values in parentheses use the no-slack formulation.

λ_b	Corr. fit	RMSE total fit	RMSE OD fit	RMSE total movement	RMSE OD movement	PURC R^2	Slack ratio
0	**0.9948**	**288.705**	**11.38631**	288.605	**11.38633**	0.1797	**0.02%**
	(**0.9948**)	(288.705)	(**11.38631**)	(288.705)	(**11.38631**)	(0.1796)	(-)
100	0.9947	288.720	11.38632	**285.831**	11.43695	0.1902	0.77%
	(0.9948)	(288.652)	(11.38637)	(288.652)	(11.38637)	(0.1806)	(-)
300	0.9947	288.795	11.38636	287.005	11.62009	0.1958	1.78%
	(0.9948)	(288.580)	(11.38677)	(288.580)	(11.38677)	(0.1823)	(-)
1000	0.9947	289.126	11.38670	305.175	12.48074	0.2040	4.15%
	(0.9947)	(**288.521**)	(11.38960)	(**288.521**)	(11.38960)	(0.1862)	(-)
3000	0.9947	290.165	11.38888	367.827	15.03673	0.2126	8.17%
	(0.9947)	(289.003)	(11.39933)	(289.003)	(11.39933)	(0.1906)	(-)
10000	0.9946	293.875	11.40536	511.035	21.17615	**0.2221**	15.12%
	(0.9945)	(292.360)	(11.45359)	(292.360)	(11.45359)	(**0.1997**)	(-)

Table 6: Sensitivity to pseudo-OD zoning under Model C with $\lambda_b = 300$. Plain values use the baseline slack-enabled formulation; values in parentheses use the no-slack formulation.

Zones	Active OD pairs	Corr. fit	RMSE total fit	RMSE OD fit	RMSE total movement	RMSE OD movement	PURC R^2	Slack ratio
8	28	**0.9968**	**211.106**	17.67668	**211.132**	17.89976	**0.2447**	**0.91%**
		(**0.9969**)	(**210.977**)	(17.67693)	(**210.977**)	(17.67693)	(**0.2262**)	(-)
10	45	0.9957	247.135	12.59474	245.352	12.84947	0.2075	1.34%
		(0.9957)	(246.970)	(12.59509)	(246.970)	(12.59509)	(0.1954)	(-)
12	57	0.9950	277.517	12.35538	276.029	12.59725	0.1972	1.67%
		(0.9950)	(277.301)	(12.35582)	(277.301)	(12.35582)	(0.1837)	(-)
15	66	0.9947	288.795	11.38636	287.005	11.62009	0.1958	1.78%
		(0.9948)	(288.580)	(11.38677)	(288.580)	(11.38677)	(0.1823)	(-)
20	89	0.9942	310.451	9.39760	311.283	9.75310	0.1680	2.55%
		(0.9942)	(310.137)	(9.39808)	(310.137)	(9.39808)	(0.1506)	(-)
25	119	0.9935	329.714	**7.26091**	335.450	**7.76937**	0.1444	3.56%
		(0.9935)	(329.271)	(**7.26160**)	(329.271)	(**7.26160**)	(0.1267)	(-)

ond, they absorb complex interaction patterns among those variables that are behaviorally relevant but difficult to specify directly in a parsimonious utility model. More broadly, this result highlights the methodological importance of LLM-derived features as a general mechanism for enriching behavioral models when conventional hand-crafted covariates are incomplete.

4 Conclusions

This paper proposes a macroscopic alternative to path-based indoor pedestrian behavior analysis. Rather than requiring accurate individual map matching, it converts Wi-Fi-based link observation probabilities into directed network flow distributions and then estimates behavioral preferences from those flows. This shift is motivated by a practical limitation observed in previous work: path inference from noisy indoor signals does not scale well enough for large-sample route choice analysis.

The empirical results from Shibuya Station show that the soft joint person-time-based model closely matches the observed link person-second field while maintaining a behaviorally interpretable PURC component. The resulting preference parameters are coherent and interpretable, revealing penalties for underground, long, and vertically complex links, as well as semantic effects captured by LLM-derived geolocation features. The sensitivity analysis further shows that behavioral patterns remain stable across alternative behavioral weights and pseudo-OD zoning schemes, while clarifying the roles of slack regularization and zoning granularity in balancing aggregate fit, movement explanation, and OD-level resolution.

The paper therefore suggests a broader methodological point. In noisy 3D indoor environments, aggregate flow distribution may be a more reliable behavioral measurement target than individual

Table 7: Representative semantic factors in Model C.

Factor	Coef.	z-val.	Strongest correlations	Indicative interpretation
10	-0.003	-9.29	indoor +0.260 escalator +0.196 stair −0.186 disruption +0.134 log length −0.073 flow −0.069	• Positive-score links: indoor transfer links and escalator-connected station passages. • Negative-score links: more open street-level links and some stair-heavy outdoor connectors. • Interpretation: enclosed transfer / interiorized vertical circulation.
3	-0.002	-8.66	indoor +0.318 transport −0.303 crossing-signal −0.253 flow +0.241 disruption +0.195 escalator +0.179	• Positive-score links: internal building movement and multi-floor indoor circulation. • Negative-score links: open street and plaza-facing links. • Interpretation: deep indoor circulation versus surface-street context.
8	-0.001	-6.53	underground −0.279 disruption +0.187 closeness −0.184 shop −0.146 escalator +0.136 indoor +0.058	• Positive-score links: upper-level indoor concourse segments. • Negative-score links: basement and street-level passages. • Interpretation: upper-level indoor circulation.
17	0.001	6.14	log length −0.103 escalator +0.102 disruption −0.093 transport −0.080 shop +0.076 betweenness −0.074	• Positive-score links: more open or locally transitional links near plaza-edge contexts. • Negative-score links: more interiorized transfer spaces. • Interpretation: transition-space context with a positive route-choice effect.

path recovery. Several extensions remain for future work. First, the probabilistic person-time measure should be formalized more explicitly within a measure-theoretic framework and linked more tightly to perturbed-utility-based stochastic assignment. Second, the current alternating estimation scheme could be strengthened by integrating flow assignment and parameter estimation into a single optimization framework, such as an MPEC or related dual formulation. Third, the role of LLM-derived semantic features should be clarified further with respect to both utility formation and observed dwell-time generation. Finally, observation uncertainty could be incorporated more explicitly through heteroskedastic or confidence-weighted formulations.

Acknowledgements

This research was performed with funding from the commissioned research grant provided by F-REI (JPFR 26-05-07-01).

References

Alibaba-NLP. (2024). *gte-qwen2-1.5b-instruct.* `https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct`. (Model card accessed for model specification and embedding dimension)

Bell, M. G. H. (1991). The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, *25*(1), 13–22. doi: 10.1016/0191-2615(91)90016-G

de Grange, L., González, F., & Bekhor, S. (2017). Path flow and trip matrix estimation using link flow density. *Networks and Spatial Economics*, *17*(1), 173–193. doi: 10.1007/s11067-016-9322-1

Fukuzaki, Y., Mochizuki, M., Murao, K., & Nishio, N. (2014). A pedestrian flow analysis system using wi-fi packet sensors to a real environment. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 721–730). doi: 10.1145/2638728.2641282

He, J., Nie, T., & Ma, W. (2024). Geolocation representation from large language models are generic enhancers for spatio-temporal learning. *arXiv preprint arXiv:2408.12116*.

Kim, D., & Suh, Y.-J. (2023). Gconvloc: Wifi fingerprinting-based indoor localization using graph convolutional networks. *IEICE Transactions on Information and Systems*, *E106.D*(7), 1247–1250. doi: 10.1587/transinf.2022EDL8081

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*. doi: 10 .48550/arXiv.2308.03281

Lo, H. K., & Chan, C.-W. (2003). Simultaneous estimation of an origin-destination matrix and link choice proportions using traffic counts. *Transportation Research Part A: Policy and Practice*, *37*(9), 771–788. doi: 10.1016/S0965-8564(03)00021-7

Mai, T., Bui, T. V., Nguyen, Q. P., & Le, T. V. (2023). Estimation of recursive route choice models with incomplete trip observations. *Transportation Research Part B: Methodological*, *173*, 313–331. doi: 10.1016/j.trb.2023.05.004

Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project — the case of openstreetmap. *ISPRS International Journal of Geo-Information*, *1*(2), 146–165. doi: 10.3390/ijgi1020146

Olbricht, R. (2011). *Overpass api — an open and slim xapi alternative.* Presentation at State of the Map Europe 2011, Vienna. (Introduces the Overpass API for OpenStreetMap queries)

Qiu, S., Huang, X., Qin, G., Chen, X., & Sun, J. (2025). Generalized route choice modeling via fusion of structural knowledge and llm-inferred context. In *Proceedings of the ieee international conference on intelligent transportation systems.*

Serere, H. N., Resch, B., & Havas, C. R. (2023). Enhanced geocoding precision for location inference of tweet text using spacy, nominatim and google maps: A comparative analysis of the influence of data selection. *PLOS ONE*, *18*(3), e0282942. doi: 10.1371/journal.pone.0282942

Su, J., & Hato, E. (2025). 3d indoor pedestrian route choice modeling using machine learning-based wi-fi fingerprint observations. In *Proceedings of the ieee international conference on intelligent transportation systems.* (Previous study by the authors)

Tsiamitros, D., et al. (2023). Indoor pedestrian flow identification and occupancy prediction from wi-fi probe data. *Sensors*, *23*(18), 7797. doi: 10.3390/s23187797

Xiong, Z., Li, B., & Liu, D. (2021). Map-matching using hidden markov model and path choice preferences under sparse trajectory. *Sustainability*, *13*(22), 12820. doi: 10.3390/su132212820

Yang, Z., Wen, J., & Huang, K. (2022). A method of pedestrian flow monitoring based on received signal strength. *EURASIP Journal on Wireless Communications and Networking*, *2022*(1), 2. doi: 10.1186/s13638-021-02079-y