Beyond Aggregates: A Fine-Grained Analysis of Individual Mobility and Traffic Dependencies

Anne Josiane Kouam, Aline Carneiro Viana, Mariano G. Beiró, Leo Ferres, Luca Pappalardo

Beyond Aggregates: A Fine-Grained Analysis of Individual Mobility and Traffic Dependencies

Anne Josiane Kouam[1,2], Aline Carneiro Viana[2], Mariano G. Beiró[3,4], Leo Ferres[5,6], Luca Pappalardo[7,8]
[1]TU Berlin, [2]Inria, [3]Universidad de San Andrés, [4]CONICET, [5]Universidad del Desarrollo,
[6]ISI Foundation, [7]ISTI-CNR, [8]Scuola Normale Superiore
kouam.djuigne@tu-berlin.de, aline.viana@inria.fr, mbeiro@udesa.edu.ar, lferres@udd.cl, luca.pappalardo@isti.cnr.it

***Abstract*—Understanding mobile-user behavior requires joint modeling of mobility and traffic, as data consumption is shaped by where, when, and how users travel. Despite this clear intuition, most studies still treat the two in isolation, *missing the intricate dependencies between them at the individual level.* This paper propose a novel approach that explicitly captures the interplay between traffic and mobility behaviors using fine-grained mobile datasets. Using week-long eXtended Data Records (XDRs), we identify 13 interpretable features and pinpoint the mobility traits that truly drive traffic variation. These insights support a privacy-preserving user abstraction that represents each timeline as a sequence of discrete mobility–traffic states, capturing temporal dynamics and heterogeneity while generalizing across regions. We then introduce a probabilistic *likelihood model* that scores any mobility–traffic pairing, enabling cross-modality prediction and statistically sound fusion of fragmented logs. Experiments on four provincial datasets covering 1.3 million Chilean users show that the model reliably separates plausible from implausible behavior and generalizes from dense urban cores to mixed rural–urban contexts. The framework is descriptive, generative, and transferable, paving the way for anomaly detection, personalized QoE adaptation, and realistic network simulation.**

***Index Terms*—Traffic-Mobility Dependency, Individual-Level Analysis, eXtended Data Records (XDRs)**

I. Introduction

Mobile networks have become *ubiquitous social sensors*: every data session conveys not only an application request but also the temporal and spatial context of the user who triggers it. A long line of studies has shown that the two dominant behavioral dimensions captured in network traces—**mobility** and **traffic**—are tightly coupled: where and when people move strongly shapes how, when and how much they consume data, and *vice-versa*. Accurate models of this interdependence therefore underpin tasks ranging from demand forecasting and resource allocation to edge caching and anomaly detection [1].

Previous work, however, has explored this coupling mainly at aggregate or city-wide scales. For instance, Wang *et al.* [2] analyzed the geographical distribution of cellular traffic, revealing correlations between distant cell towers due to transportation patterns. Similarly, Wu et al. [3] demonstrated how cellular tower traffic reflects urban land use, classifying areas as residential, transport, office, or entertainment zones. Meanwhile, Chen et al. [4] examined disparities in hourly traffic patterns between urban and rural areas, linking traffic fluctuations to underlying infrastructure. Such macro perspectives are invaluable for city-scale optimisation, yet they blur the heterogeneity that matters at the *user* level. A commuter, a tele-worker and a tourist may share the same cell sector, but their temporal rhythms and application mixes differ markedly—differences that must be captured for realistic simulation and personalized service design.

Attempts to zoom in on individuals do exist, but they remain constrained by both methodology and data. Alipour *et al.* [5] analyzed WLAN traces and labeled devices as "flutes" (mobile) or "cellos" (static), a binary heuristic that overlooks the spectrum of mobility behaviors and captures only WLAN-covered traffic. Paul *et al.* [6] used 3G data-log statistics to correlate kilometers traveled with megabytes consumed, a dataset dominated by voice and SMS events and offering limited visibility into modern app usage. These studies confirm that mobility matters, but their coarse labels and partial traffic views fall short of modeling the joint, time-evolving interplay that behavior-aware systems now require.

A new opportunity arises from the widespread collection of *eXtended Data Records* (XDRs). XDRs log every data session—spanning app usage, web browsing and general IP traffic—with precise timestamps and antenna locations, yielding a continuous, geo-referenced chronicle of user interaction with the network. Their richness makes XDRs an ideal substrate for individual-level traffic–mobility analysis, yet, to our knowledge, no prior work has leveraged nation-scale XDRs to build a principled, probabilistic model of that dependency.

Harnessing these large-scale XDR datasets, our objective is to move beyond coarse correlations and develop a *principled, user-level model* of traffic–mobility dependency that is (i) **descriptive**—faithfully capturing the joint statistical structure of observed behaviors, (ii) **generative**—enabling realistic simulation and data fusion, and (iii) **transferable**—operating across diverse urban contexts while preserving privacy. Realizing this objective centers on three tightly linked research questions, which we articulate and address below:

- *RQ1: How can raw traffic and location logs be distilled into principled insights about traffic–mobility interplay?*
  We address this question in **§IV** by designing a systematic workflow that converts week-long XDR traces into *13 interpretable features* spanning four lenses—traffic usage, spatial mobility, movement dynamics, and social interactions. Leveraging these features, we cluster users by exploration behavior (*regular, routiner, explorer*) and by traffic profile, then quantify how the two taxonomies interact. The analysis

(i) pinpoints the mobility traits—*traveled distance*, *repetitiveness* and *stationarity*—that *truly* drive traffic variation, (ii) shows that these dependencies persist across dense-urban to mixed rural–urban regions, and (iii) yields region-agnostic design guidelines for finer-grained modeling.

- *RQ2: How can we represent traffic–mobility interplay at the user level while preserving heterogeneity, ensuring transferability, and protecting privacy?*
  We tackle RQ2 in **§V** by proposing a novel user abstraction as a *sequence of discrete mobility–traffic states*, derived from the empirical joint distribution of volume categories and mobility classes at every time step. This graph-based representation offers three key advantages: (i) *context neutrality*—by categorizing traffic volumes and mobility metrics *relative* to their locality, it standardizes behavior across datasets and environments; (ii) *fine-grained differentiation*—the temporal sequence captures evolving statistical dependencies, making identical state strings across users increasingly unlikely; (iii) *privacy preservation*—it discards exact coordinates, retaining only anonymized state transitions while still exposing user-specific signatures.
- *RQ3: Given only one behavioral modality, can we assess the plausibility of—and even infer—the other at scale?*
  We investigate this question in **§VI-A** by introducing a *probabilistic framework* that assigns a *behavior-based likelihood score* to any mobility–traffic pairing. The score quantifies statistical coherence and enables two core tasks: (i) *cross-modality prediction*—inferring traffic behavior from mobility traces and vice-versa; (ii) *principled data fusion*—merging disjoint mobility-only and traffic-only datasets by validating the plausibility of sequence matches.
- *Validation and practical impact* (**§VI-B, §VI-C**). Using *four week-long XDR datasets* (over 1.3M users) from Santiago, Elqui, Bio-Bio and Copiapo, we show that our framework (i) reliably separates plausible from implausible behavior pairings, (ii) remains robust to parameter choices such as state duration and likelihood scaling, and (iii) generalises across diverse urban profiles. These results demonstrate its utility for realistic simulation, privacy-preserving trace synthesis, and behavior-aware network analytics.

II. Background and related works

Mobile traffic and human mobility data each provide unique insights into human activity. While often studied separately, their strong interrelation has been demonstrated in various research. This section reviews the state-of-the-art studies on both data types and examines existing research on their correlation.

Human Mobility data. Human mobility is a well-researched area with significant value across several fields, including public health, sociology, transportation, and tourism. Studies have examined it from multiple angles, identifying key laws that govern movements [1]and creating models for simulating [7], predicting [8], and generating [9] mobility patterns at various scales. When analyzed at an individual level, human mobility provides more detailed insights by accounting for personal differences. Research shows that individuals exhibit varied behaviors in their movement patterns [10] with a recent study [11] identifying three mobility profiles (i) *Scouters*, more inclined to explore and discover new areas; (ii) *Routiners*, who maintain a steady routine and rarely break their established patterns; and (iii) *Regulars*, with a moderate behavior balancing between explorations and revisits.

Mobile Traffic data captures daily user activity on cellular networks valuable for tasks like network optimization and resource management. This study focuses on XDR traces, extensively studied in the literature for characterization, prediction [3], or generation [12] purposes. Similar to human mobility data, individual-level analyses reveal distinct user profiles in data generation [13], [14]. [14] distinguishes four profiles: *Light Occasional (LO), Light Frequent (LF), Heavy Occasional (HO), and Heavy Frequent (HF)*. Light users generate up to 20GB per day, while Heavy users exceed this amount. ”Occasional” and ”Frequent” distinctions depend on the number of sessions users generate daily.

Mobility and Traffic data dependency. Several studies have demonstrated the close relationship between mobile traffic and human mobility through statistical correlations [2], [4]–[6], [15]. The movement patterns of mobile network users—shaped by daily routines, points of interest, and behaviors—introduce significant temporal and spatial dynamics into mobile traffic data, which are crucial for understanding, modeling, and predicting cellular traffic at both large and fine scales. Additionally, the prevalence of network events among dispersed urban populations provides rich datasets that reflect underlying mobility patterns and urban dynamics, which are essential for effective city management and planning.

Existing studies on the correlation between mobility and mobile traffic have primarily focused on the city scale. For instance, [16] examined the geographic distribution of cellular traffic and uncovered strong spatiotemporal dependencies, while [4] identified spatial inhomogeneities in traffic patterns between urban and rural areas. Similarly, [15] illustrated how mobile traffic data can reveal city dynamics and infrastructure usage. While these studies offer valuable insights, they do not address the dependencies between traffic and mobility at the individual level. Some works, such as [5], explored mobility-traffic correlations in WLAN settings at the device level, categorizing devices as ”cellos” (stationary) and ”flutes” (mobile), and found that cellos generated traffic more frequently, albeit in smaller volumes. Similarly, [6] investigated the relationship between subscriber mobility and traffic generation in 3G networks, concluding that more mobile users produced higher traffic volumes. While these approaches provide a preliminary understanding of how mobility correlates with traffic, a more detailed examination of the traffic-mobility dependency at the individual behavior level remains largely unexplored, limiting the flexibility and potential for mobile network optimization and personalization applications.

Positioning. This paper introduces a novel structured approach to *identify and quantify direct dependencies between traffic and mobility behaviors at the user level.* Built upon this foun-

TABLE I: Summary statistics of the four provincial datasets.

	#users b/f filtering	#users a/f filtering	#users w/ full sequence	#cells	size
Santiago	1,093,221	787,326	21,849	1,536	$2{,}030km^2$
Elqui	122,602	83,041	2,873	170	$16{,}895km^2$
Bio-Bio	65,049	38,035	1,106	87	$32{,}538km^2$
Copiapo	56,847	3,223	1,115	71	$14{,}987km^2$
Total	1,337,719	950,388	26,943	1,864	$66{,}450km^2$

dation is a unified representation of user behavior that enables flexible analysis, cross-modality inference, and integration of traffic and mobility data in mobile network environments.

III. Data structuring and preprocessing

This section outlines the XDR mobile datasets used in our study, detailing their structure and the preprocessing steps applied to ensure consistency for behavior modeling.

Preliminaries. XDR data captures each mobile user's communication behavior across two dimensions along the *time*: *traffic*, and *mobility*. A mobile user's daily activity, as recorded in an XDR dataset, can be represented as a temporal sequence of events $S^u = [e_1, e_2, \ldots, e_n]$, where each event e_n is a tuple (t_n, v_n, l_n). In this tuple, t_n represents the timestamp, discretized at a chosen granularity (e.g., 15 or 30 minutes) , v_n is the data volume generated during that time granularity, and l_n indicates the corresponding location information.

Raw data description. We use four independent, privately collected XDR datasets from mobile network operators, covering the Chilean provinces of Santiago, Elqui, Biobío, and Copiapó. Each dataset contains anonymized user activity over a regular week (Sunday to Saturday), captured with time granularity, i.e., $t_{i+1} - t_i$, of 30 minutes. The selected period reflects regular weekly behavior, excluding holidays or exceptional events, to focus on typical day-to-day mobile usage patterns. The four provinces were chosen for their contrasting socio-urban characteristics: Santiago represents a dense metropolitan center; Elqui and Bio-Bio are mid-sized regions with distinct mobility and traffic profiles; and Copiapo offers a more peripheral setting with mixed rural–urban dynamics. This diversity supports the evaluation of our method across varied behavioral and network conditions.

Data Preprocessing. To ensure data quality, location reference codes in the raw datasets were first matched to their corresponding geographical coordinates. Users with more than 5% missing location data were excluded, while minor gaps (less than 5%) were imputed using the last known location. The final user counts after filtering are reported in the second column of Table I. The third column indicates the number of users with complete weekly sequences of $7 \times 48 = 336$ events, with no missing data or temporal gaps.

IV. Data-driven behavior characterization

This section develops a data-driven workflow for characterizing user-level traffic–mobility interplay. We first identify key behavioral features in §IV-A, analyze each lens independently in §IV-B, and quantify their cross-dependencies in §IV-C.

A. User behavioral features

Table II summarizes the features computed for each user, organized by XDR dimension (traffic and mobility) and categorized into spatial, structural, and social aspects for mobility.

Traffic Behavior. We use the *average number of events per day*, capturing traffic frequency, and the *average session volume*, capturing data usage per session. Based on these features, we derive a traffic profile for each user in $\{HO, HF, LO, LF\}$ (cf. §II) through hierarchical clustering, as in [14].

Mobility Behavior. We distinguish the following aspects:

- **Spatial features** describe user mobility based on the geographical patterns of their movements. We consider the user's *average traveled distance* and the *radius of gyration* (Rg), a standard metric in mobility analysis. The radius of gyration is computed in two forms: *Rg_unique*, considering each location once, and *Rg_event*, which weights locations based on visit frequency.
- **Structural features** capture the sequence patterns of user movements, focusing on the consistency and variability of their visits. Here, we use literature well-established metrics such as *repetitiveness (rep.), stationarity (sta.), diversity (div.)*, and *predictability (pre.)* [17], [18]. We calculate trajectory predictability using the Kontoyiannis entropy algorithm [19]. Additionally, we track $\#succ\_ret$ (successive returns) and $\#succ\_expl$ (successive explorations) to classify users into mobility profiles in {*routiner, regular, scouter*} using the clustering method from [11].
- **Social features** capture the influence of social interactions on mobility. We propose two metrics: (i) *user popularity influence (pop_infl)*, calculated as the average popularity of the locations visited by the user, where the popularity $p(l_i, t_j)$ is the number of unique users visiting location l_i at time t_j. Additionally, (ii) $flow(l_i, l_{i+1}, t_j)$ is the number of users following the same path from l_i to l_{i+1} during time slot t_j. Each user's *flow measurement (flow_meas.)* averages these flows across their entire trajectory.

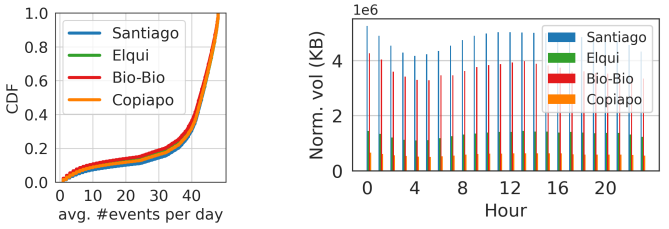
B. Mobility and traffic behavior in isolation

Using the definitions introduced in §IV-A, we now characterize the distribution of behavioral features across the four studied provinces, separately for traffic and mobility.

- Traffic: Fig. 1a depicts the average number of events per user per day, revealing a shared behavior among provinces. Roughly *half of the population engages in 45 to 48 events daily,* a pattern characteristic of XDR datasets collecting even background data traffic. While the average traffic session volume per user is relatively consistent across provinces, Fig. 1b highlights *notable differences in total hourly traffic volume* normalized by the number of users in the province. Users in Santiago generate the most traffic, followed by those in Bio-Bio, while Elqui and Copiapo exhibit lower, yet similar, traffic volumes. These distributions result in approximately 40% of users classified as *LO*, 30% as *LF*, 5% as *HO*, and 25% as *HF*, with negligible variability across datasets.

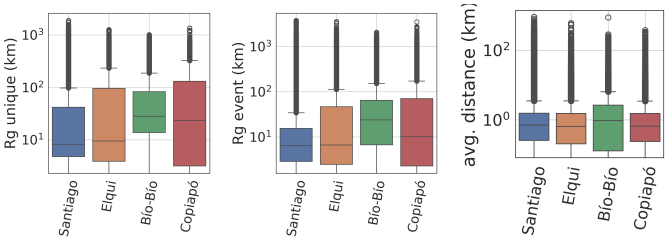
TABLE II: Features description.

Features			Description	Equation (for each user)	Step-level
Traffic	*avg. #events per day* (Fig. 1a)			$(\sum_{\{i\|v_i\neq 0\}} 1)/N_{day}$	/
	avg. session volume (Fig. 1b)			$\frac{1}{n}\sum_{i=1}^{n} v_i$	$trc_i \in \{l, m, h\}$
Mobility	**spatial**	*avg. traveled distance per slot* (Fig. 2c)		$\sum_{i=1}^{n-1} d_i$ with $d_i = \|l_{i+1} - l_i\|$	$disc_i \in \{c, m, f\}$
		Rg_unique (Fig. 2a)	quantifies the spatial extent of users' movements: higher values of r_g indicate that the user covers a large area, while lower values suggest more localized movements.	$\sqrt{\frac{1}{n}\sum_{k=1}^{n} \|l_k - l_{\text{unique}}\|^2}, l_{\text{unique}} = \frac{1}{n}\sum_{k=1}^{n} l_k$	/
		Rg_event (Fig. 2b)		$\sqrt{\frac{\sum_{k=1}^{n} m_k \cdot (\|l_k - l_{\text{event}}\|)^2}{\sum_{i=1}^{n} m_i}}, \quad l_{\text{event}} = \frac{\sum_{k=1}^{n} m_k l_k}{\sum_{i=1}^{n} m_i}$	/
	structural	*Rep.* (Fig. 2d)	Frequency of returns to previously visited locations	$(1 - n_{\text{unique}})/n$	$rep_i = 0/1$
		Sta. (Fig. 2d)	Ratio of user stays at the same location in consecutive intervals	$\sum_{\{i\|l_{i+1}=l_i\}} 1/(n-1)$	$sta_i = 0/1$
		Div. (Fig. 2e)	Number of distinct sub-trajectories in a user trajectory	/	δdiv_i
		Pre. (Fig. 2e)	Entropy rate of the trajectory	/	/
		#succ_ret (Fig. 2h)	Number of successive returns	/	/
		#succ_expl (Fig. 2h)	Number of successive explorations	/	/
	social	*Pop_infl* (Fig. 2f)	avg. visitation rate of a user's visited locations	$\frac{1}{n}\sum_{i=1}^{n} p(l_i, t_i)$	$p(l_i, t_i)$
		Flow_meas. (Fig. 2g)	avg. number of user with the same spatial movements	$\frac{1}{n}\sum_{i=1}^{n} flow(l_i, l_{i+1}, t_i)$	$flow(l_i, l_{i+1}, t_i)$

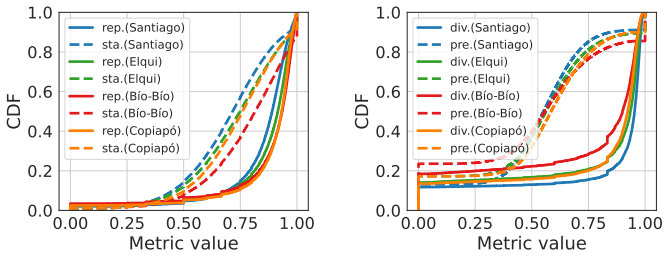


(a) Number of events (b) Hourly normalized volume

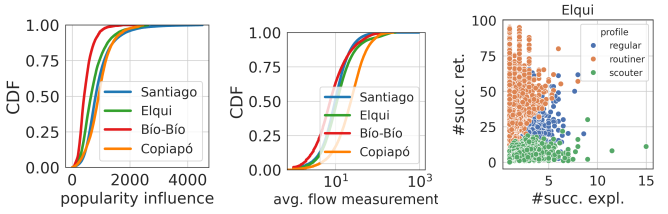
Fig. 1: Mobile users' traffic behavior distributions



(a) Rg_unique (b) Rg_event (c) Avg. distance



(d) Rep. and Sta. (e) Div. and Pre.



(f) Pop_infl (g) Flow_meas. (h) Mobility profiles

Fig. 2: Users' mobility behavior features distributions

- Spatial mobility: Figs. 2a and 2b show that the radius of gyration ranges between 1 km and 200 km. Notably, urbanized provinces like Santiago and Elqui exhibit lower radius of gyration values compared to Bio-Bio and Copiapo, *suggesting that residents travel shorter distances due to nearby facilities in urban environments*. In contrast, the higher variability observed in Elqui and Copiapo highlights the mixed nature of these provinces, which feature both urban and remote areas. Similar trends are captured in Fig. 2c, where the average distance traveled within 30-minute time slots remains relatively low across provinces, between 100 m and 8 km.

- Structural mobility: Figs. 2d and 2e reveal *a high repetitiveness in user trajectories, increasing from the most urbanized province, Santiago, to the least, Copiapo.* Around 80% of the population revisits 75% of the same locations. *Stationarity levels are less pronounced* but still show at least 25% across all provinces, with *higher values observed in less urbanized areas* like Bio-Bio and Copiapo, despite their larger geographical sizes (cf. Table I). This helps explain why nearly 20% of users exhibit almost no diversity in their movements, while the majority show a diversity index of at least 0.75.

- Social mobility: Figs. 2f and 2g reflect user popularity and flow, corresponding to the overall population density in each province. However, Copiapo stands out with significantly higher values, indicating a *strong clustering of the population in certain areas*. Finally, Fig. 2h classifies users into mobility profiles, with approximately 60% identified as regulars, 25% as routiners, and 15% as explorers, highlighting the consistent presence of diverse mobility behaviors across all four datasets.

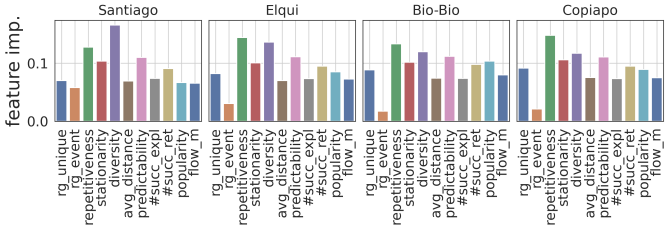


Fig. 3: Mobility feature weight in traffic profile prediction.

C. Mobility and traffic behavior interplay

To explore the relationship between traffic and mobility, we examine the distribution of our two traffic features across mobility profiles and our 11 mobility features across traffic profiles. Since there are only two traffic features, no filtering is needed for mobility profile analysis. However, for traffic profile analysis, we use an Ensemble Trees classifier to predict traffic profiles (*LO, LF, HO, HF*) based on mobility features, selecting the most relevant ones. Feature importance is assessed using Gini score, which quantifies the predictive power of each mobility feature in distinguishing traffic profiles. The results consistently reveal *Rg_unique*, *diversity*, and *popularity influence* as the most influential spatial, structural, and social mobility features, respectively, in all provinces (cf. Fig. 3).

The analyses in Elqui, shown in Fig. 4, enhance our understanding of user behavior yielding the following insights:

Finding 1 (Figs 4a, 4b): *Scouters, or extreme explorers, generate traffic significantly less frequently and at lower volumes compared to regulars and routiners. While the average traffic volume for regulars and routiners is quite similar, it is noteworthy that regulars engage in traffic generation less often than routiners. This suggests that more exploratory users tend to participate in activities that yield less traffic and occur less frequently. In contrast, routiners, being more stationary, have greater opportunities for high-traffic activities.*

Finding 2 (Figs 4c, 4d, 4e): *Users with an occasional traffic profile (LO and FO) demonstrate a lower radius of gyration (Rg) compared to those who engage more frequently in traffic. This aligns with preliminary insights from the literature [5], [6], indicating that spatial extent correlates first with traffic frequency and then with volume. Additionally, users with occasional traffic show lower trajectory diversity, while those with higher frequency traffic profiles (LF and HF) possess greater movement diversity. This trend reflects that increased traffic engagement relates to more varied paths. Finally, the popularity influence is higher for occasional users, suggesting that they are drawn to popular locations that may not require frequent or heavy traffic generation but likely engage them in activities of a different nature, such as shopping or socializing.*

V. Fine-grained behavior representation

Accurately modeling traffic–mobility interplay at the user level demands more than a static feature vector: fixed categories such as *Heavy* or *Light* miss when and how behaviors evolve, obscuring the heterogeneity exposed by XDR data.

This section meets that need with a *dependency-driven user representation* that converts each timeline into a *sequence of discrete mobility–traffic states*. The resulting state graph captures temporal dynamics, standardizes behavior relative to local baselines, and anonymizes raw coordinates while preserving user-specific signatures. §V-A details state construction, §V-B explores intra-sequence dependencies, and §V-C distills the main refinements and practical takeaways.

A. Formal construction

We represent each user as a sequence of discrete, statistically significant mobility–traffic states over time. Each state captures joint mobility and traffic behavior at a given time step t_i, while transitions reflect their temporal evolution. This expressive and privacy-preserving abstraction forms the foundation for the rest of the paper.

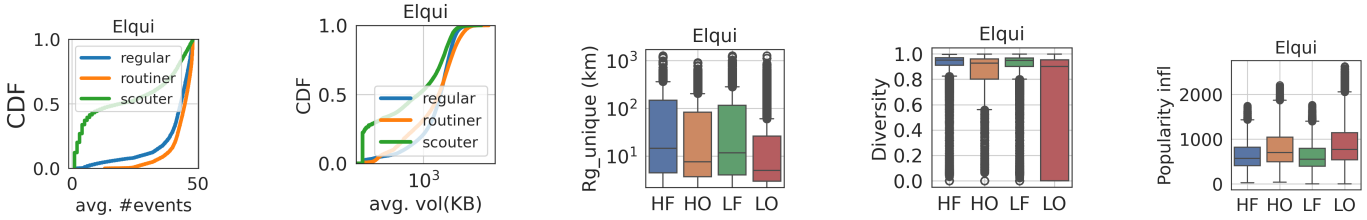
Formally, a user u is represented as a sequence $u = (b_1^u, b_2^u, \ldots, b_n^u)$, where each behavior tuple $b_i^u = (t_i b_{\text{tra-i}}^u b_{\text{mob-i}}^u)$ characterizes the user at time step t_i. Here, $b_{\text{tra-i}}^u$ denotes the traffic-related features, and $b_{\text{mob-i}}^u$ represents the mobility-related features for that time step. These step-level features are derived from the global users' behavioral features in Table II and detailed in the following.

Traffic behavior. We categorize each step-level volume v_i^u as *light* (l), *medium* (m), or *heavy* (h), based on the quantiles of the global session volume distribution. This allows us to track the user's traffic evolution as a sequence $(trc_1, trc_2, \ldots, trc_n)$, where $trc_i \in \{l, m, h\}$. As traffic frequency is captured by the sequence length, the traffic behavior reflects only the volume:

$$b_{\text{tra-i}}^u = (trc_i^u)$$

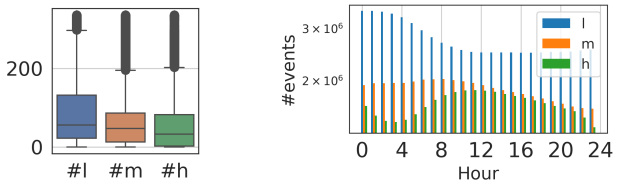
Mobility behavior. For modeling mobility at each step, we use selected spatial, structural, and social metrics. The **spatial features** include the *traveled distance* d_i between consecutive locations. Similar to the traffic classification, we categorize d_i into three groups: *close*, *medium*, or *far*, based on the quantiles of the global step-level distance distribution. The *radius of gyration* is excluded from step-based modeling, as it requires the full trajectory to compute the center of mass. For **structural features**, we track the *repetitiveness* of each step as a binary state, indicating whether the user returns to a previous location (i.e., 1) or explores a new one (i.e., 0). Additionally, we assess *stationarity*, which is another binary state indicating whether the user remains at the same location between two consecutive steps ($l_i = l_{i+1}$). We also evaluate the change in trajectory diversity at each step using *delta-diversity* (δdiv_i), which measures the difference in trajectory diversity from the previous step, reflecting whether the current movement increases, stabilizes, or decreases the diversity of locations visited. We deliberately avoid using a separate *predictability* measure, as it is heavily influenced by the repetititveness and stationarity of visited locations [18], potentially leading to redundancy. For **social features**, we capture the *zone popularity* of the locations visited by the user at each step, denoted as $p(l_i, t_i)$, which represents the number of unique users visiting the user location l_i at the time step t_i. We similarly measure the *flow of users*, denoted as $flow(l_i, l_{i+1}, t_i)$, which tracks how many users follow the same movement transition at the same time step. Hence, the mobility behavior at each step is represented as:

$$b_{\text{mob-i}}^u = (disc_i, rep_i, sta_i, \delta\text{div}_i, p(l_i, t_i), flow(l_i, l_{i+1}, t_i))$$



(a) Avg. #events (b) Avg. volume (c) Rg_unique (d) Div. (e) Pop_infl

Fig. 4: (a)-(b) Traffic features by mobility profile; (c)-(e) Mobility features distribution by traffic profile.



(a) In-sequence distribution (b) Hourly distribution

Fig. 5: Distribution of traffic category count in Santiago.

B. Step-level dependency analysis

We analyze the dependency between traffic (b^u_{tra-i}) and mobility (b^u_{mob-i}) behaviors at an individual basis and per time step (t_i), identifying mobility indicators most affected by traffic patterns. Across all datasets, users tend to generate *light* traffic events most frequently, followed by *medium* and *heavy* traffic (cf. Fig. 5a). Interestingly, only *heavy* traffic volumes exhibit a clear daily pattern, peaking during the day and dropping at night (cf. Fig. 5b). In contrast, *light* traffic is more prominent at night, reflecting background activity when users are generally inactive, and declines during the day as heavier traffic takes over.

To establish dependency, we examine how different mobility features distribute across traffic categories (*light*, *medium*, *heavy*). Significant differences in the distributions suggest strong dependencies between certain mobility features and traffic patterns. Our results in Fig. 6 highlight three mobility features strongly influenced by traffic category: *traveled distance* (cf. Fig. 6a), *repetitiveness* (cf. Fig. 6b), and *stationarity* (cf. Fig. 6c). In contrast, mobility features like *popularity* (cf. Fig. 6e) and *flow* (cf. Fig. 6f) show only moderate dependency with traffic, and *delta diversity* (cf. Fig. 6d) appears to be unaffected. Key insights emerge from these findings:

Finding 3: *Users tend to make shorter trips during light traffic periods, suggesting that light traffic often results from background activities, especially at night and when users are less active with reduced mobility. Additionally, repetitive and stationary movement patterns are more common during light traffic and decrease as traffic intensity rises. Hence, during heavy traffic, users likely engage in more varied or spontaneous activities that disrupt routine behaviors. Moreover, users involved in medium or heavy traffic tend to travel longer distances.* ***This supports the idea that predictable, stationary actions are associated with lighter traffic, while more dynamic behaviors correspond to higher traffic volumes.***

C. User representation takeaways

Building on the step-level dependency analysis, we refine the user behavior representation by retaining only the mobility features that exhibit strong correlations with traffic behavior—namely, *traveled distance*, *repetitiveness*, and *stationarity*. This streamlined representation reduces complexity, improves interpretability and computational efficiency, and retains the key behavioral traits most relevant to traffic dynamics.

Our novel user representation, thus, offers three key benefits: **(i)** First, by focusing on the relative behavior of users within a given location (e.g., relative categorization of traffic volumes as heavy, medium, or light states), *the model standardizes user behavior and provides a generic representation that is both transferrable and adaptable across various contexts*, including diverse datasets and urban or rural environments.
(ii) Second, the sequence structure captures refined statistical dependencies between mobility and traffic dimensions at each time step, enabling evolution capture. As the sequence grows, it encodes increasingly complex and individualized behavior patterns. Unlike fixed categorical profiles (e.g., *Heavy Frequent*), this enables fine-grained user differentiation making identical sequences across users increasingly unlikely.
(iii) Last, the representation avoids inclusion of user identifiers or raw data. Instead, it captures patterns in anonymized state sequences, preserving privacy while enabling analysis.

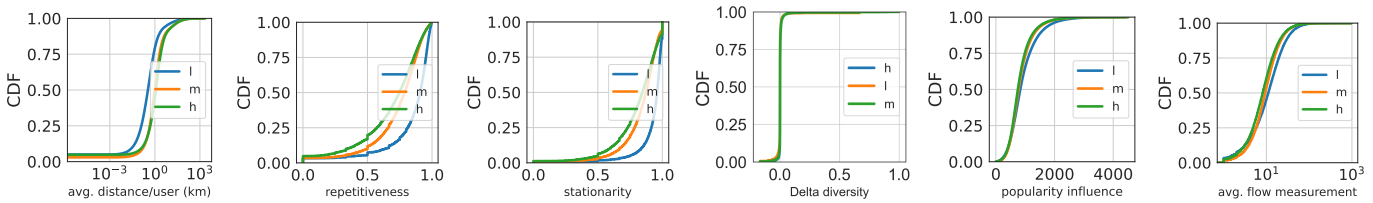
VI. User Behavior Inference

Building on the previous user representation, we introduce a *probabilistic inference model* that learns typical joint mobility–traffic patterns and generalizes user-level reasoning beyond the combinations observed in data. At its core is a *likelihood score* that quantifies the statistical compatibility of any mobility–traffic pair and unlocks two capabilities: (i) *cross-modality inference*, predicting one modality from the other, and (ii) *data integration*, matching and merging fragmented mobility-only and traffic-only traces into complete behavior sequences. §VI-A details the model and likelihood computation, §VI-B evaluates its realism, and §VI-C demonstrates its effectiveness for cross-modality prediction and data merging.

A. Likelihood-based inference model

Given a population $\mathcal{U} = \{u_k\}_{k=1}^{P}$, where each user u_k is modeled as a sequence of temporal behavior states (cf. §V-A),

$$u_k = (t_0 b^{u_k}_{\text{tra-0}} b^{u_k}_{\text{mob-0}}, \ldots, t_n b^{u_k}_{\text{tra-n}} b^{u_k}_{\text{mob-n}})$$



(a) Traveled distance (b) Repetitiveness (c) Stationarity (d) Delta diversity (e) Pop. influence (f) Flow measurement

Fig. 6: Dependency between in-step traffic and mobility features of users in Santiago.

the inference model is a Markov model $M = (S, P)$ where:

- Each state $s \in S$ corresponds to a unique combination of behavior defined as $s_i = b_i = (t_i b_{\text{tra-i}} b_{\text{mob-i}})$. States are global across the population, with no user-specific indexing.
- The transition matrix P defines the probabilities of transitioning between states. Each element $p_t = P(s_j \mid s_i)$ represents the probability of transitioning from state s_i to state s_j across of the population. Hence, $p_t = P(s_j \mid s_i) > 0$, if there is at least one user u_k in the population and state l such that $(s_i, s_j) = (s_l^k, s_{l+1}^k)$.

Likelihood score: From the proposed inference model, we introduce a *likelihood score* to evaluate the realism of matching a sequence of traffic behaviors to a corresponding sequence of mobility behaviors while ensuring it satisfies essential criteria:

- *Distinguish between real and random sequences*: The function should assign higher scores to sequences that align with real-world behaviors while penalizing unrealistic matches.
- *Ensure high likelihood correlates with realism*: If a sequence has a high likelihood, it must be reflective of naturally occurring patterns. Thus, both extremely rare and completely non-existent transitions should be penalized proportionally.
- *Robustness to missing transitions*: The metric should not immediately assign a likelihood of zero if a transition is unseen but should gradually adjust the score based on transition plausibility.
- *Score normalization*: The score should be bounded within a fixed range (e.g., $[0, 1]$) for easier interpretability.
- *Efficient inference application*: The likelihood function should be computable on a per-user-sequence basis without requiring batch processing or excessive computation.

To satisfy the outlined criteria, we propose a *scaled max-normalized likelihood function,* $MaxL_\alpha$ described below:

1) Given two sequences of traffic behaviors $(b_{\text{tra-0}}, \ldots, b_{\text{tra-n}})$ and mobility behaviors $(b_{\text{mob-0}}, \ldots, b_{\text{mob-n}})$—originating from the same or different users, depending on the inference context—we extract the corresponding state sequence $s_0, \ldots, s_n$, where each state is defined as $s_i = (t_i, b_{\text{tra-i}}, b_{\text{mob-i}})$.
2) The likelihood score of the resulting paired sequence is:

$$MaxL_\alpha = \frac{1}{n} \sum_{\substack{t=1 \\ p_t > 0}}^{n} \left[\alpha + (1 - \alpha) \times \frac{P(s_{t+1} \mid s_t)}{\max P(s \mid s_t)} \right], \text{with}$$

- $P(s_{t+1} \mid s_t)$ is the empirical transition probability derived from the trained inference model.
- $\max P(s \mid s_t)$ is the maximum transition probability from state s_t, ensuring normalization.
- α is a scaling parameter that prevents rare transitions from being overly penalized.
- n is the number of transitions in the sequence, and the condition $p_t > 0$ ensures we sum only valid transitions, while missing transitions do not contribute to the score.

This approach assigns likelihoods close to 1 to highly probable sequences, down-weights rare transitions without excluding them entirely, and smoothly decreases the score in case of missing transitions instead of collapsing it to zero.

Model setup: Before performing inference, the model is trained on a large, representative dataset by building a graph structure of observed states and transitions, from which empirical probabilities are estimated. This resulting graph structure captures behavioral dynamics and enables evaluating the realism of new sequences. Its structural formulation and aggregated transition probabilities allow for transfer across scenarios while preserving privacy.

For the following validations and use cases, we split users per province into 70% training and 30% test sets. To enable traffic–mobility sequence shuffling and matching, we retain only users with complete (non-missing) sequences. Table I, col. 3, reports the number of users and sequences per province used in the inference tasks.

B. Model validation and robustness

We validate the inference model by assessing the behavior likelihood score's reliability, sensitivity, and generalizability. We test its ability to reflect plausible traffic–mobility pairings, its robustness to parameters (e.g., state duration), and its generalizability across diverse urban contexts. Results confirm the score's relevance and the model's robustness.

*1) **Validating the likelihood score**.:* We first evaluate whether the likelihood score meaningfully reflects the plausibility of matching traffic and mobility behaviors. To this end, we generate a shuffled dataset where each user's traffic sequence is randomly assigned to a mobility sequence sampled from the dataset (with replacement). This process introduces both plausible and implausible matchings, including the ground truth pairings. For each shuffled pair, we compute the likelihood score and compare it against the behavioral difference between the assigned mobility sequence and the ground truth. As a distance metric, we use the Hamming distance, counting the number of time steps in which mobility states differ between the assigned and true sequences.

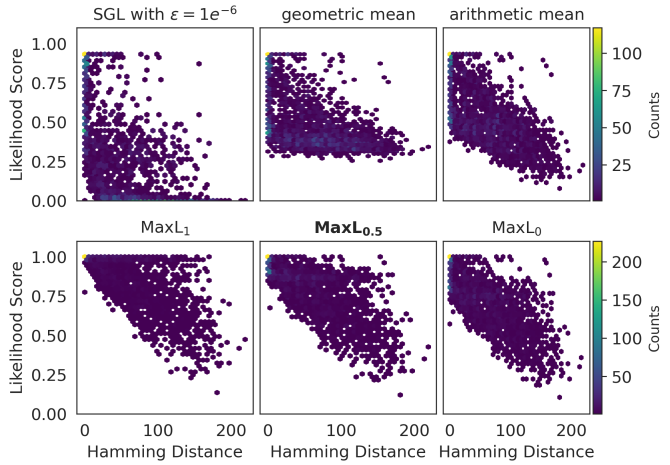


Fig. 7: Comparison of our $\mathbf{MaxL}_\alpha$ and baseline likelihood scores vs. distances to ground truth in Elqui.

A robust likelihood score should exhibit low values for dissimilar pairings (high distance) and high values—close to 1—for near-perfect or ground truth alignments. We visualize this relationship in a hexbin plot (Fig. 7), which shows the density and distribution of matchings as a function of their likelihood and distance to the ground truth. We compare our *max-normalized likelihood score* ($MaxL_\alpha$) —evaluated with three values of the scaling parameter α (1, 0.5, and 0)—against several classical baselines:

- *Smoothed Geometric Likelihood*, with ϵ being a small constant for Laplace smoothing:

$$SGL = \exp\left(\frac{1}{n}\sum_{t=1}^{n}\log\left(\frac{P(s_{t+1} \mid s_t) + \epsilon}{1+\epsilon}\right)\right),$$

- *Geometric mean* of transition probabilities (non-zero only),
- *Arithmetic mean* of sequence transition probabilities.

Results, on Fig. 7, show that log-likelihood and geometric mean offer poor discriminative behavior, with likelihood values clustered near zero for most matchings and no clear correlation with ground truth distance. The arithmetic mean performs better but fails to consistently assign high likelihood to ground-truth matches. In contrast, our max-normalized score spans the full $[0, 1]$ range and shows stronger correlation. The variant with $\alpha = 1$ (transition ratio) produces overly dispersed scores, including many false positives with perfect likelihood. The version with $\alpha = 0$ is too conservative. The intermediate $\alpha = 0.5$ strikes the best balance, yielding a well-shaped oblique distribution that best reflects matchings quality.

Finding 4: *The max-normalized likelihood best reflects the plausibility of traffic–mobility matchings, outperforming baselines in separating realistic from implausible pairings.*

*2) **Analyzing parameter sensitivity**:* We next examine how internal parameters affect the model's ability to discriminate between realistic and unrealistic traffic–mobility matchings. For this, we compare the likelihood distributions of real-world and artificially shuffled datasets, using the Hellinger distance as a measure of separation. The shuffled dataset is constructed to ensure unrealistic matchings by imposing a minimum Hamming distance of 20% (i.e., at least 67 mismatched states) from the ground truth. The Hellinger distance, ranging from 0 (identical) to 1 (completely disjoint), quantifies how well the model separates these two distributions.

(i) Effect of state size and time encoding. We begin by analyzing the influence of the memory length n_{state} in the Markov model, which controls the duration of behavioral context encoded in each state. We vary this from 1 hour to 6 hours and test both the *default* time encoding (hour and minute information) and the *time-hour* variant (hour only). All experiments are performed with $\alpha = 0.5$.

As shown in Fig. 8a, Hellinger distances are consistently high across all provinces, indicating strong discriminative ability. The peak distance is generally observed for state durations of 3 to 4 hours, suggesting that this range captures the most informative behavioral context. Increasing the memory beyond this point slightly degrades performance, possibly due to overfitting of behavioral patterns. The differences between time encoding variants are negligible, indicating that minute-level time granularity adds limited benefit in this context.

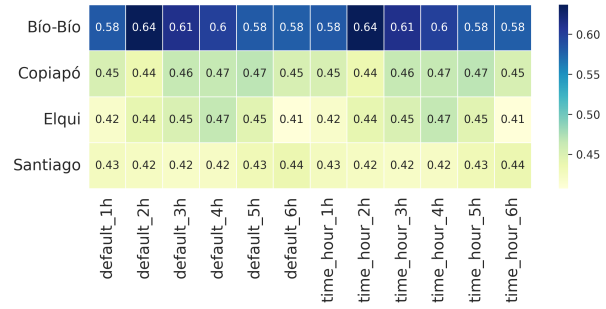
Finding 5: *The model is most effective when using 3–4 hours of behavioral memory, and is robust to moderate variations in temporal encoding granularity.*

(ii) Effect of scaling parameter α. Fixing the state size to 3 hours, we now evaluate the effect of the scaling parameter α in the max-normalized likelihood score, using the *default* time encoding. Fig. 8b shows that Hellinger distances remain relatively stable across values of α, with optimal performance for α of 0.2 to 0.6. Increasing α beyond this point leads to a slight but consistent degradation, particularly at $\alpha = 1$, which corresponds to the transition ratio. This confirms previous observations that overly permissive scores (e.g., transition count-based) fail to distinguish implausible matchings.

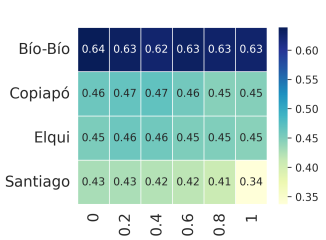
Finding 6: *The likelihood score performs best for α values between 0.2 and 0.6, and remains robust to moderate tuning.*

*3) **Testing model generalizability**:* Finally, we assess whether the model generalizes across heterogeneous urban environments. Using a fixed configuration ($n_{\text{state}} = 3$h, $\alpha = 0.5$), we train the model on one province and evaluate its discriminative performance on others. Fig. 8c reports the resulting Hellinger distances. The results show that the model generalizes well across regions: in most cases, the discriminative performance is comparable to (or slightly better than) that observed with intra-province training. Provinces such as Bio-Bio even exhibit stronger separability when trained on data from other areas, suggesting that core behavioral patterns are shared across regional contexts.

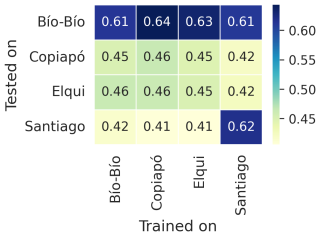
Finding 7: *The model exhibits strong generalization across urban regions, enabling behavior matching and integration even when trained on external data sources.*



(a) Intra-province with $\alpha = 0.5$ w.r.t. n_{state} and time encoding



(b) Intra-province with *default*, $n_{state} = 3h$ w.r.t. α



(c) Inter-province with *default*, $n_{state} = 3h$, and $\alpha = 0.5$

Fig. 8: Hellinger distance between regular and shuffled likelihood distributions

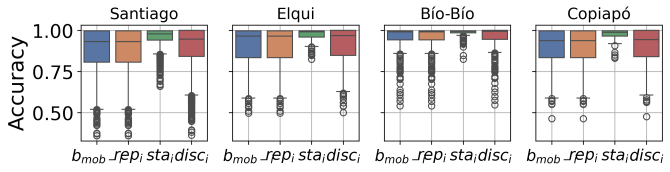


Fig. 9: Mobility from traffic behavior inference accuracy.

C. Inference model applications

We now evaluate the practical use of the inference model in two key applications. First, we assess its ability to infer one dimension of user behavior (mobility or traffic) from the other. Second, we demonstrate how the model supports realistic integration of independently collected mobility and traffic datasets by identifying plausible pairings across them. This capability is essential for enhancing data availability and enabling downstream behavior reconstruction through data augmentation. In both cases, the model leverages the learned dependencies and the associated likelihood score to guide cross-dimensional alignment.

1) Cross-modality inference: We evaluate the model's ability to predict a user's mobility behavior from traffic one, and vice versa. To this end, we repeatedly sample plausible sequences from the trained transition graph, conditioned on the known behavior dimension. These samples reflect the most likely behavioral states, as informed by the learned transition probabilities. This setup enables us to evaluate the model's predictive capacity and gain insight into how well it captures the coupling between traffic and mobility dynamics.

Fig. 9 shows the accuracy of predicting users' mobility behavior from traffic patterns in different provinces. The overall avg. mobility inference accuracy ranges from 88.7% in Santiago, to 94.6% in Bio-Bio, with minimal deviation across users. At the feature level, the model performs best in predicting stationarity, followed by distance and repetitiveness. Conversely, inferring traffic categories from mobility behavior yields slightly lower but still strong performance, with accuracies of $88.7\% \pm 12.3$, $90.4\% \pm 11.7$, $94.7\% \pm 8.8$, and $89.7\% \pm 12.2$ in Santiago, Elqui, Bio-Bio, and Copiapo.

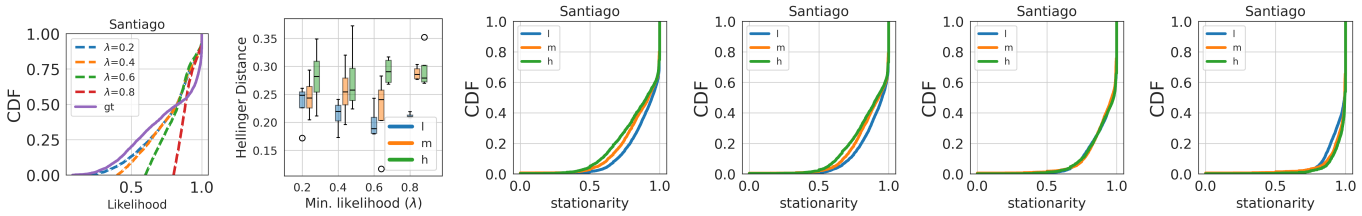
Finding 8: *The model reliably captures the mutual dependencies between mobility and traffic behaviors, enabling effective inference in both directions offering a promising basis for personalized mobile network services.*

2) Data integration through matching: We use the inference model to reconstruct realistic user-level behavior by merging independently collected traffic and mobility sequences. Starting from a fully integrated dataset (used as ground truth), we simulate a disjoint scenario by splitting it into two parts—traffic-only and mobility-only representations. We then randomly shuffle the mobility and traffic behavior sequences across users and use the model's likelihood score to guide the merging process. Only matchings with a likelihood exceeding a given threshold λ are retained. *The parameter λ thus provides a tunable control over matching strictness. As λ increases, fewer matchings are accepted, leading to smaller but behaviorally more consistent datasets.* For example, increasing λ from 0.2 to 0.8, in Santiago, reduces the dataset size from 3,457 to 2,097 users (original size: 3,569).

To assess the impact of λ, we evaluate the reconstructed dataset in terms of (i) its internal behavioral consistency and (ii) its representativeness with respect to the original dataset. First, Fig. 10a shows the likelihood distributions in the ground truth dataset and the merged datasets for different values of λ. As λ increases, the resulting datasets concentrate around higher likelihoods, indicating more realistic user profiles—but at the cost of reduced diversity.

Next, we evaluate how well the reconstructed datasets preserve key traffic-to-mobility behavioral dependencies found in the original population. Given space constraints, we report only *stationarity* per traffic category. Fig. 10b reports the Hellinger distance between the reconstructed and ground truth distributions. As we increase the min. likelihood threshold λ from 0.2 to 0.6, the distance gradually decreases, indicating closer alignment with the original dataset. However, beyond $\lambda = 0.6$, the distance increases again, suggesting that the resulting dataset—though behaviorally consistent—is no longer representative of the population as a whole. This is consistent with the raw distribution plots shown in Figs. 10c–f, where higher thresholds yield narrower and more homogeneous distributions that diverge from the population-wide behavior.

Finding 9: *The likelihood-guided merging framework reliably rebuilds realistic user timelines. Raising the threshold λ improves per-user fidelity but increasingly discards atypical users, reducing population representativeness. High λ is therefore suited to personalized analytics, while broader reconstructions call for more inclusive settings.*



(a) Likelihood scores (b) Hellinger distance (c) $\lambda = 0.2$ (d) $\lambda = 0.4$ (e) $\lambda = 0.6$ (f) $\lambda = 0.8$

Fig. 10: Effect of the min. likelihood λ on dataset realism and representativeness. (a) Likelihood distribution in reconstructed data. (b) Hellinger distance to ground truth stationarity. (c–f) Raw stationarity per traffic state for selected λ values.

VII. Discussion

Our study focuses on building an interpretable, transferable model of traffic–mobility dependency. We show that our model captures the mutual dependency between mobility and traffic, and that it can rebuild realistic user timelines, also generalizing across urban regions in Chile. Here we reflect briefly on (i) other modeling avenues we explored and (ii) practical uses of the framework beyond the scope of this paper.

Alternative modeling. Our user *joint-behavior representation*—which preserves heterogeneity, protects privacy, and transfers across datasets—could feed deep sequential or graph-based inference models. We tested RNN, LSTM, LLM embeddings, and graph-neural variants but found that the sparse, coarse information per time step offered too little signal for them to generalize; they also lacked a transparent, closed-form likelihood for cross-modality validation. The Markov model therefore provides the balance we need: simplicity, interpretability, and efficient likelihood computation. Future work could revisit hybrid schemes (e.g., shallow attention layers atop the Markov backbone) to seek further performance.

Practical applications. Although their implementation lies outside the scope of this paper, the learned likelihood scores enable several individual-level applications in mobile networks. Sudden drops in likelihood can signal mobility–traffic mismatches for *anomaly or fraud detection*. Forecasting traffic from upcoming mobility supports *personalized QoE adaptation* at the network edge. Sampling state sequences produces *privacy-preserving synthetic traces* for simulation or data sharing, and likelihood-based matching can *fuse* separate mobility-only and traffic-only logs. Exploring these directions at operational scale is a promising avenue for future work.

VIII. Conclusion

This work deepens the understanding of mobility–traffic dependencies at the user level, where previous analyses remained coarse or aggregated. We revealed rich behavioral correlations and used them to build a novel representation of users as joint mobility–traffic signatures. A probabilistic inference model with an associated likelihood score was then designed and shown to be robust, generalizable across urban contexts, and useful for key mobile networking applications, including cross-modality inference and the integration of fragmented behavioral data across key mobile network applications. These contributions lay the foundation for integrated behavior-based reasoning in mobile service design and personalization.

Acknowledgments

This work is part of the Mob Sci-Dat Factory project (ANR-23-PEMO-0004) under the France 2030 program and the CAPES-STIC-AMSUD 22-STIC-07 LINT project.

References

[1] M. C. González, C. A. Hidalgo, and A. Barabási, "Understanding individual human mobility patterns," *Nature*, 2008.

[2] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," *IEEE Transactions on Mobile Computing*, 2019.

[3] J. Wu, M. Zeng, X. Chen, Y. Li, and D. Jin, "Characterizing and predicting individual traffic usage of mobile application in cellular network," in *ACM UbiComp*, 2018.

[4] X. Chen, Y. Jin, S. Qiang, W. Hu, and K. Jiang, "Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale," in *IEEE ICC*, 2015.

[5] B. Alipour, L. Tonetto, A. Y. Ding, R. Ketabi, J. Ott, and A. Helmy, "Flutes vs. cellos: Analyzing mobility-traffic correlations in large wlan traces," in *IEEE INFOCOM*, 2018.

[6] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *IEEE INFOCOM*, 2011.

[7] K. Keramat Jahromi, M. Zignani, S. Gaito, and G. P. Rossi, "Simulating human mobility patterns in urban areas," *Simulation Modelling Practice and Theory*, 2016.

[8] J. Wang, X. Kong, F. Xia, and L. Sun, "Urban human mobility: Data-driven modeling and prediction," *SIGKDD Explor. Newsl.*, 2019.

[9] A. Kobayashi, N. Takeda, Y. Yamazaki, and D. Kamisaka, "Modeling and generating human mobility trajectories using transformer with day encoding," in *ACM HuMob-Challenge*, 2023.

[10] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A. Barabási, "Returners and explorers dichotomy in human mobility," *Nature Communications*, 2015.

[11] L. Amichi, A. C. Viana, M. Crovella, and A. A. Loureiro, "Understanding individuals' proclivity for novelty seeking," in *ACM SIGSPATIAL*, 2020.

[12] A. J. Kouam, A. C. Viana, and A. Tchana, "Lstm-based generation of cellular network traffic," in *IEEE WCNC*, 2023.

[13] D. Naboulsi, R. Stanica, and M. Fiore, "Classifying call profiles in large-scale mobile traffic datasets," in *IEEE INFOCOM*, 2014.

[14] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," in *IEEE PerCom*, 2015.

[15] M. G. Demissie, G. H. de Almeida Correia, and C. Bento, "Exploring cellular network handover information for urban mobility analysis," *Journal of Transport Geography*, vol. 31, 2013.

[16] K. Xu, R. Singh, M. Fiore, M. K. Marina, H. Bilen, M. Usama, H. Benn, and C. Ziemlicki, "Spectragan: spectrum based generation of city scale spatiotemporal mobile network traffic data," in *ACM CoNEXT*, 2021.

[17] E. M. R. Oliveira, A. C. Viana, C. Sarraute, J. Brea, and I. Alvarez-Hamelin, "On the regularity of human mobility," *Pervasive and Mobile Computing*, 2016.

[18] D. Teixeira, J. Almeida, and A. C. Viana, "On estimating the predictability of human mobility: the role of routine," *EPJ Data Science*, 2021.

[19] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," *IEEE Transactions on Information Theory*, no. 3, 1998.