Complementing Location-Based Social Network Data With Mobility Data: A Pattern-Based Approach

Elena Daraio, *Student Member, IEEE*, Luca Cagliero, *Member, IEEE*, Silvia Chiusano, *Member, IEEE*, and Paolo Garza, *Member, IEEE*

***Abstract*— Location-Based Social Networks can be profitably exploited to characterize citizens' activities in urban environments. However, collecting LBSN is potentially challenging due to privacy concerns, connectivity issues, and potential imbalances in LBSN service usage. We propose to complement LBSN data with mobility data in the analysis of citizens' activities in urban areas. Unlike the explicit insights provided by LBSN users, mobility data give implicit feedback on citizens' habits. This paper explores the spatial and temporal conditions under which user habits are coherent according to both sources and reports the most reliable common sequences of visited categories of Points-Of-Interests. To this aim, it relies on a multidimensional model in which recurrent citizens' activities are described by a new pattern type, namely the generalized activity pattern. It also detects the eventual presence of bias between LBSN and mobility user activities by customizing the established Statistical Parity metric. The motivations behind the detected bias are explained in terms of combinations of POI categories that are most likely to be the main causes. We evaluate the proposed approach on real-world data achieved from Foursquare check-ins, taxi service, and free-floating car sharing. The results highlight not only the complementarity of the data sources regarding specific POI categories, but also their interchangeability in many spatio-temporal conditions.**

***Index Terms*— Location-based social networks, activity patterns, mobility data, sequential pattern mining.**

I. Introduction

LOCATION-BASED Social Networks (LBSNs) are a particular kind of online social network where geographical locations are the core of the network structure [1]. Unlike traditional social networks, which are mainly based on social relationships among users, LBSNs take advantage of geo-referenced data and location-tagged content to foster user interactions. For example, Foursquare[1] is among the most popular LBSNs. Foursquare users voluntarily 'check in' to places they visit using a mobile application. Hence, they disclose the temporal sequence of their main activities and the related locations (e.g., she has lunch at the restaurant and then she does the shopping).

The availability of LBSN data has prompted their use to study life-style user patterns in urban environments to assess different urban dynamics, such as urban land use and activities [2]–[4]. They allows a better understanding and sense of how citizens interact with the city, such as spatio-temporal citizen activities and services of interest to citizens. Their analysis may reveal differences in users habits, preferred places, and action and location patterns [5]. However, collecting large real-world LBSN datasets can be challenging due to privacy concerns, connectivity issues, and service usage imbalances. For instance, users can either revoke their consent for public use or be reluctant to activate geo-tagging options [6], [7]. Furthermore, check-ins are often unevenly distributed across the urban environment, e.g., they are concentrated in the city center, thus making life-style patterns potentially biased [8].

In situations in which there is a need to overcome the lack of LBSN data, we propose to complement their analysis using mobility data. The key idea is to model user activities in urban environments by profiling their use of various means of transport, such as taxi or free-floating car sharing trips, in terms of origin/destination places and leaving and arrival times. Unlike LBSN data, mobility data provide implicit feedback on the most common user habits in urban activities. By construction, they indirectly express the actual user purposes and preferences even when LBSN data are missing or incomplete. Since mobility service providers routinely gather operational data about service provision, mobility data acquisition is less exposed to privacy concerns and connectivity issues, and the corresponding service usage profiles are potentially complementary to LBSN ones. For these reasons, we deem mobility data as a valid candidate for providing additional content that enriches and supplements citizens' activity monitoring.

In this paper we seek the conditions under which mobility data can effectively complement or even substitute LBSN data in describing citizens' activities in urban environments. To this purpose, we define a multidimensional model that characterizes urban activities based on different dimensions such as *location* (e.g., the city district), *time* (e.g., season, daily time slot), and *data source type* (e.g., taxis, shared bicycles,

Manuscript received 3 September 2021; revised 14 March 2022; accepted 21 May 2022. Date of publication 23 June 2022; date of current version 7 November 2022. This work was supported by the Italian Ministry of Research (MUR) under the Smart Cities and Communities Grant SCN_00325 (Project s[m2]art). The Associate Editor for this article was Z. He. *(Corresponding author: Paolo Garza.)*

Elena Daraio, Luca Cagliero, and Paolo Garza are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, 10129 Turin, Italy (e-mail: paolo.garza@polito.it).

Silvia Chiusano is with the Dipartimento Interateneo di Scienze, Progetto e Politiche del Territorio, Politecnico di Torino, 10129 Turin, Italy.

Digital Object Identifier 10.1109/TITS.2022.3182569

[1]https://Foursquare.com/ (latest access: March 2022)

rented cars, Foursquare). User activities are described by a set of *generalized activity patterns* mined separately from LBSN and mobility data. They model recurrent user activities as temporal sequences of the categories associated with the "visited" Points-of-Interests (POIs). An example of such activity patterns is *Restaurant* $\rightarrow$ *Theater*, which indicates that users frequently go to the restaurant and then to the theater. The aforesaid pattern exemplifies the temporal relationship in user behavior between a specific entertainment place and a typical food service.

Given a specific combination of spatial and temporal dimension values (e.g., *location* = New York City, *time* = May 2020), we study the feasibility of integrating LBSN with mobility data. Specifically, our goal is threefold:

(1) **Bias detection**. Compare the activity patterns discovered from LBSN and mobility data to detect potential bias, which would hinder an effective integration of the aforesaid data sources.
(2) **Bias explanation**. Identify the combinations of categories of Points-of-Interests that are most likely to be the cause of bias (if any).
(3) **Pattern analysis**. Shortlist the top-K most reliable activity patterns in common between LBSN and mobility data.

We accomplish goal (1) by tailoring the Statistical Parity (SP) metric [9], which was originally designed for measuring the fairness of a classification model, to our purposes. Specifically, we verify, via statistical test, the initial hypothesis that the same activity patterns hold for both LBSN and mobility data. The combinations of POI categories for which the initial hypothesis is rejected are returned as potential causes of bias, i.e., goal (2). To deepen the analysis of LBSN and mobility data consistency, i.e., goal (3), we shortlist the top-K most reliable activity patterns and repeat the SP test for different values of K. For the pattern shortlist that is likely to be not affected by any bias, we study the common user activities with the aim at highlighting the complementarity or interchangeability of the two data sources.

We report the results achieved in a real case study, i.e., the comparison between Foursquare check-ins acquired in New York City and Portland and mobility data collected by the taxi and free-floating car sharing services. The top ranked patterns such as *Food* $\rightarrow$ *Entertainment* describe common activities shared by LBSN and mobility users. However, specific patterns such *Health* $\rightarrow$ *Bank* reveal the presence of bias due to specific POI categories, whose corresponding user activities are commonly not disclosed by LBSN users. By comparing the patterns extracted in different contexts, we also provide interesting insights into the temporal changes in citizens' habits.

The main paper contributions are summarized below.

- **Data modelling.** A joint multidimensional data model valid for LBSN and mobility data. The model leverages both explicit and implicit activity-related citizen information (see Section III for a description of the data model).
- **Theory.** (1) A novel type of patterns that describe the citizens' activities in terms of sequences of visited POIs (see Section III-C for a description of the newly proposed patterns). (2) A new pattern-based approach to detect bias between the activity patterns mined from LBSN and mobility data (see Section IV).
- **Methodology.** An in-depth comparison between LBSN and mobility data in terms of activity patterns. It entails exploring contextualized data to detect bias first and then perform detailed comparisons between unbiased pattern shortlists (see Sections III and V).
- **Empirical evidence.** The outcomes achieved in a real case study confirm the applicability of the methodology (see Section VII for a summary of the main relevant results).

II. Related Works

The increasing availability of LBSN data has prompted the study and development of advanced data mining and machine learning solutions to

1) infer individual life-style patterns from activity-location choices revealed in social media [5],
2) recommend locations relevant to specific users [10],
3) predict the next place a user is likely to visit [11],
4) analyze tourist movements by extracting activity patterns [12], and
5) compare user habits across different social platforms [13].

The scope of this work is mainly related to (1). In this regard, a particular attention has been paid to location-based check-in services, where users notify their geographical position to share the activity-related choices. Geo-tagged user preferences are relevant, for instance, to model Point-of-Interest (POI) demand (e.g., [14]–[17]) or to extract sequential user activity patterns (e.g., [12], [18]–[20]). The present study is related to the latter type of applications. However, as pointed out in [6], considering LBSN data is often not sufficient to effectively characterize life-style patterns. Thus, the research community has started to integration of complementary data sources. For example, in [21], [22] the authors compare the movement, mobility, and activity patterns observed in LBSNs with those occurring in cell phone location data. They observe that short-ranged travels are spatially and temporally periodic and weakly effected by the social network structure, whereas long-distance travels are more influenced by social network ties. The work presented in [23] aims at connecting social sensors and road traffic conditions. The key idea is to correlate the habits and routines of Foursquare and Instagram users to the traffic maps acquired from Bing Map. In [24] the authors estimate urban traffic flows using LBSN data. They highlight a promising potential of using LBSN data for urban travel demand analysis and monitoring. However, mobility data are collected through a dedicated survey, which is influenced by the user engagement, availability, and perception. Our approach differs from [23], [24] in (1) the type of complementary data (i.e., taxi and car sharing data), (2) the customization and application of bias detection methods to the problem under analysis, (3) the use of sequence mining techniques to model activity patterns and explain the bias between LBSN and mobility data.

TABLE I
SUMMARY OF NOTATIONS AND THEIR MEANINGS

$\mathbf{D}$	set of data sources
$\mathbf{U}$	set of LBSN users
P	analyzed time period $[t_{start}, t_{end}]$
tw	time window in P
$\mathbf{L}$	set of geo-referenced locations l: $\langle latitude, longitude \rangle$
$c_{\langle u,t,l \rangle}$	check-in of user $u \in \mathbf{U}$ at time t in P at location $l \in \mathbf{L}$
$\mathbf{POI}$	set of analyzed Points-Of-Interest
$\text{POI}_{\langle u,t,l \rangle}$	subset of POIs in $\mathbf{POI}$ associated with check-in $c_{\langle u,t,l \rangle}$
R	spatial region including a set of geo-referenced locations
$\mathtt{C}$	Context corresponding to time span tw, spatial region R, and data source d
$\mathbf{COI}$	set of POI categories
$COI_{\langle u,t,l \rangle}$	POI categories associated with check-in $c_{\langle u,t,l \rangle}$
$\mathbf{Tr}$	set of trips
tr	trips from $l_{\text{source}} \in \mathbf{L}$ to $l_{\text{dest}} \in \mathbf{L}$
$\mathtt{DB}_{\text{LBSN}}$	check-in sequential dataset
$\mathtt{DB}_{\text{tr}}$	trip sequential dataset
$\mathbf{S}_u^{\text{POI}}$	sequence of POIs visited by user u during the check-ins
$\mathbf{S}_{\text{tr}}^{\text{POI}}$	sequence of neighbor POIs for trip $tr \in \mathbf{Tr}$
$\mathbf{S}_u^{\text{COI}}$	sequence of categories of the POIs visited by user u during the check-ins

More recently, in [6], [25] the authors propose ad hoc data simulators to overcome the lack of LBSN data. Since they generate new data by mimicking the original data distribution, they are unable to capture complementary information such as the user interest in particular POI categories that are usually not present in the LBSN check-ins (e.g., bank, hospital, rest home).

To complement Foursquare check-in data, the authors [26] propose to analyze the published tips trips venues and temporal patterns. Since they consist of User-Generated Content, the provided information is substantially different from those provided by mobility data.

III. MULTIDIMENSIONAL DATA MODEL

We present a multidimensional data model, in which the activities of the users of LBSN and mobility services carried out within the same urban context can be jointly analyzed. A summary of the notation used throughout the paper is reported in Table I.

A. Data Sources

Location-based Social Network (LBSN) data collect the history of the social user check-ins. Each check-in $c_{\langle u,t,l \rangle}$ is a triple of user-timestamp-location $\langle u, t, l \rangle$, $u \in \mathbf{U}$, $t \in P$, $l \in \mathbf{L}$. Every check-in is associated with a set of Points-of-Interest $\text{POI}_{\langle u,t,l \rangle}$ that are manually annotated by user u.

Example: A Foursquare user created a check-in through her mobility phone when she visited the *Colosseo* in Rome and tagged the check-in with the corresponding POI name (*Colosseo*) and category (*Tourist attraction*).

Mobility data consist of a set of trips from a location to a destination. Each trip $tr \in \mathbf{T}$ can be modelled as a sequence of location-timestamp pairs respectively denoting the trip start and end: $\langle l_{\text{source}}, t_{\text{source}} \rangle \rightarrow \langle l_{\text{dest}}, t_{\text{dest}} \rangle$, where $l_{\text{source}} \in \mathbf{L}$, $l_{\text{dest}} \in \mathbf{L}$ and $t_{\text{source}} \in \mathbf{T}$, $t_{\text{dest}} \in \mathbf{T}$. Notice that, unlike LBSN check-ins, trip locations are neither necessarily associated with a specific user nor annotated with any POI information. Hence, we implicitly derive the most likely user activities based on the presence of a set of POIs in the source and destination neighborhoods.

Example: A taxi ride from *Piazza del Colosseo* to *Piazza del Popolo* in Rome can be annotated with the corresponding POI names and categories (e.g., *Tourist attraction*).

B. Contextual Model

We study the users activities in specific spatio-temporal conditions and verify the consistency of the patterns extracted from LBSN and mobility data. A sketch of the proposed strategy is depicted in Figure 1.

We rely on a *contextual model*, where LBSN and mobility user activities are described by the following dimensions:

- *Data source*: it reports the data source d used to retrieve the raw data. It encompasses social data sources (e.g., Foursquare) and mobility data (e.g., taxi services, free-floating car sharing services).
- *Time*: it indicates the time span tw (within the analyzed time period P) in which the event (either a LBSN check-in or a urban trip) occurred. It is a calendar data descriptor extracted from the recorded timestamps, which characterizes the periodicity and seasonality of the underlying urban activities. A temporal hierarchy can be used to aggregate the acquired timestamps at multiple abstraction levels (e.g., daily, monthly, or different daily time slot).
- *Space*: it indicates the spatial region R corresponding either to the LBSN check-ins or the trip endpoints (source and destination). It is useful for differentiating the activities carried out in different city areas (e.g., in the city center, in suburbs, or in external hubs).

We define the context $\mathtt{C}$ as a triple of contextual dimensions consisting of time span tw (within P), spatial region R, and data source $d \in \mathbf{D}$.

Example: To analyze the activities of the citizens of San Francisco in the weekends considering only the Bay Area and the Foursquare check-ins, the data are tailored to the context (tw=weekend, R=Bay Area, d=Foursquare).

C. The Generalized Activity Patterns

Each context is described by a set of patterns, namely the *generalized activity patterns*. They represent the underlying user activities in terms of sequences of Points-Of-Interest Categories. The patterns selected from different sources within the same spatio-temporal context can be compared with each other to verify the complementarity and interchangeability of the analyzed sources.

Example: The comparative analysis in Figure 1 studies the similarity between the contexts (tw=May, R=New York City (NYC), d=Taxi) and (tw=May, R=New York City (NYC), d=Foursquare). It quantifies the corresponding coherence level

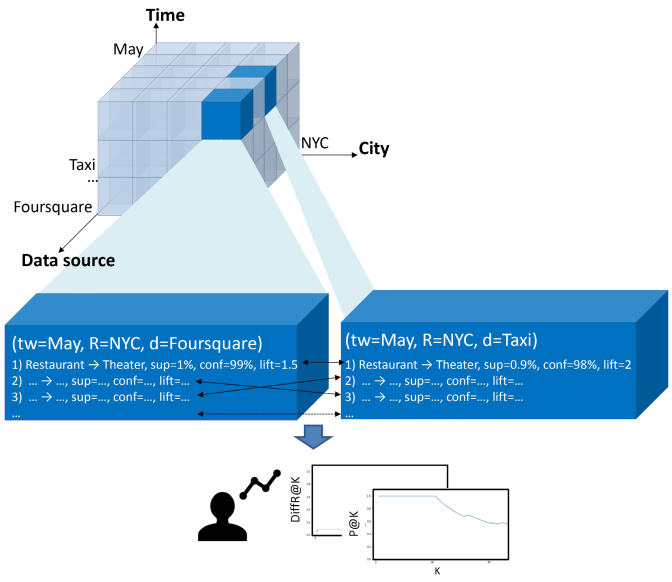


Fig. 1. Multidimensional data model.

based on a set of common generalized activity patterns such as *Restaurant* → *Theater* (i.e., eat out and then go to the theater) and their corresponding quality metrics.

a) Preliminaries: To extract the activity patterns, we need to first build a sequential dataset collecting all the POI sequences within a particular context. To this aim, we formalize the adherence of a user check-in/trip to a specified context.

Definition 3.1 (Check-in Context Adherence): Let $c_{\langle u,t,l\rangle}$ be a LBSN check-in and let C be a context where the source $d \in \mathbf{D}$ is related to the LBSN domain. A check-in $c_{\langle u,t,l\rangle}$ *adheres* to $\mathtt{C}$ if and only if t belongs to tw and l belongs to the spatial region R associated with the context $\mathtt{C}$.

Similarly, in the transportation domain we define the adherence of the trip source and destination to the specified context.

Definition 3.2 (Trip Context Adherence): Let $tr \in \mathbf{Tr}$ be an arbitrary trip and let $\mathtt{C}$ be a context where the source $d \in \mathbf{D}$ is related to the transportation domain. A trip *adheres* to $\mathtt{C}$ if and only if (i) t_{source} and t_{dest} belong to tw and (ii) l_{source} and l_{dest} belong to the spatial region R associated with the context $\mathtt{C}$.

In LBSNs POIs are manually annotated by the social users. The set of POIs associated with a check-in $c_{\langle u,t,l\rangle}$ is denoted by $\text{POI}_{\langle u,t,l\rangle}$, where the corresponding set of POI categories, hereafter denoted as *Category Of Interest* (COI), is denoted by $\text{COI}_{\langle u,t,l\rangle}$. Since we are interested in modeling the general habits of the users, our analyses will be mainly focused on COIs rather than POIs.

In the mobility scenario we first map the trip endpoints (origin and destination) to the set of nearest POIs and then we associate the corresponding COIs.

Definition 3.3 (COI Mapping to Trip Endpoints.): Let $\text{NN}(\cdot)$: $\mathbf{L} \rightarrow \mathcal{P}(\mathbf{POI})$ be a function that maps an arbitrary location to the subset of the POIs with a fixed radius r.[2] The POIs associated with l_{source} and $l_{destination}$ correspond to $\text{NN}(l_{\text{source}})$ and $\text{NN}(l_{\text{dest}})$, respectively. Each set of POIs in $\text{NN}(\cdot)$ is then mapped to the corresponding COIs.

We store the temporal sequence of user check-ins in a *check-in sequential dataset* as follows.

Definition 3.4 (Check-in sequential dataset): Let C be the context under analysis. Let $\mathbf{S}_u$: $c_{\langle u,t_1,l_1\rangle} \rightarrow c_{\langle u,t_2,l_2\rangle} \rightarrow \ldots \rightarrow c_{\langle u,t_n,l_n\rangle}$ be the temporal sequence of all the check-ins made by user u such that the following conditions hold: (1) $t_1 < t_2 < \ldots < t_n$,
(2) $t_1, t_2, \ldots t_n$ belong to P,
(3) t_n - $t_1 \leq \texttt{maxtimegap}$, and
(4) all user check-ins in the sequence adhere to $\mathtt{C}$. Let $\mathbf{S}_u^{\text{COI}}$: $\text{COI}(c_{\langle u,t_1,l_1\rangle}) \rightarrow \text{COI}(c_{\langle u,t_2,l_2\rangle}) \rightarrow \ldots \rightarrow \text{COI}(c_{\langle u,t_n,l_n\rangle})$ be the temporal sequence of COIs built on top of $\mathbf{S}_u$ (hereafter denoted as *input COI sequence*). The check-in sequential dataset DB_{LBSN} consists of the set of all the input COI sequences for every user $u \in \mathbf{U}$.

We characterize user activities according to their short-term movements across different POIs located in the urban area. Hence, we enforce both `maxtimegap` and `mintimegap` constraints. The `maxtimegap` constraint ensures that the user check-ins that are temporally distant are not included in the same sequence. Conversely, according to the `mintimegap` constraint different check-ins that are temporally close are considered as a simultaneous events in a sequence.

Analogously to LBSN data, we build a *trip sequential dataset* on top of the generated temporal sequences.

Definition 3.5 (Trip sequential dataset): Let C be the context under analysis. Let $\mathbf{Tr}^* \subseteq \mathbf{Tr}$ be a subset of trips such that
(1) $t_{\text{source}} < t_{\text{dest}}$,
(2) t_{dest} - $t_{\text{source}} \leq \texttt{maxtimegap}$, and
(3) $\mathbf{Tr}^*$ satisfies C.
The trip sequential dataset DB_{tr} consists of the set of all the temporal sequences of COIs for every trip $tr \in \mathbf{Tr}^*$.

b) Pattern definition: The life-style of a LBSN user can be modelled as an *activity pattern*. It describes either a recurrent sequence of venues associated with the check-ins of a single user [12] or a sequence of venue categories described by the corresponding POIs [27]. In our context, an activity is described by a subset of POIs and their categories, and the pattern captures the temporal sequence of the performed activities (in terms of POI categories).

Our aim is to unify the LBSN and mobility data by proposing a generalized version of the activity pattern including both user check-ins and mobility data. More specifically, a *generalized activity pattern (GAP)* is a temporal sequence of POI categories which represents either explicit social user annotations or implicit neighbor relationships for trip locations/destinations.

Definition 3.6 (Generalized Activity Pattern (GAP)): Let $\mathbf{COI}_i$ be an arbitrary set of COIs recorded at time t_i. A generalized activity pattern is a temporal sequence g: $\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_n$, such that
(1) $t_1 < t_2 < \ldots < t_n$,
(2) $t_1, t_2, \ldots t_n$ belong to P,
(3) $t_{i+1} - t_i \geq \texttt{mintimegap}$ $i = 1, 2, \ldots, n-1$ and
(4) $t_n - t_1 \leq \texttt{maxtimegap}$.

[2]Throughout the paper we will use the Euclidean distance to estimate geographical distances as the crow flies.

We quantitatively estimate the GAP relevance to the generalized check-in and trip sequential datasets, respectively, using the established support, confidence, and lift metrics [28].

Definition 3.7 (Support of a Generalized Activity Pattern.): Let g:$\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_n$ be a generalized activity pattern. The support $\sup(g,\mathtt{DB}_*)$ of g in an arbitrary sequential dataset $\mathtt{DB}_*$ (either check-ins $\mathtt{DB}_{\mathrm{LBSN}}$ or trips $\mathtt{DB}_{\mathrm{tr}}$) is the fraction of input COI sequences in the dataset that either exactly match or contain the corresponding POI category sequence.

Definition 3.8 (Confidence of a Generalized Activity Pattern): Let g: $\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_n$ be a generalized activity pattern. The confidence $\mathrm{conf}(g,\mathtt{DB}_*)$ of g in an arbitrary generalized sequential dataset $\mathtt{DB}_*$ (either $\mathtt{DB}_{\mathrm{LBSN}}$ or $\mathtt{DB}_{\mathrm{tr}}$) is defined as the conditional probability of occurrence of the last set of POI categories $\mathbf{COI}_n$ given the POI category sub-sequence $\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_{n-1}$. It is defined as

$$\frac{\sup(g, \mathtt{DB}_*)}{\sup(\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_{n-1}, \mathtt{DB}_*)}.$$

Definition 3.9 (Lift of a Generalized Activity Pattern): Let g: $\mathbf{COI}_1 \rightarrow \mathbf{COI}_2 \rightarrow \ldots \rightarrow \mathbf{COI}_n$ be a generalized activity pattern. The $\mathrm{lift}(g,\mathtt{DB}_*)$ of g in an arbitrary generalized sequential dataset $\mathtt{DB}_*$ (either $\mathtt{DB}_{\mathrm{LBSN}}$ or $\mathtt{DB}_{\mathrm{tr}}$) is defined as

$$\frac{\sup(g, \mathtt{DB}_*)}{\sup(\mathbf{COI}_1 \rightarrow \ldots \rightarrow \mathbf{COI}_{n-1}, \mathtt{DB}_*) \times \sup(\mathbf{COI}_n, \mathtt{DB}_*)}$$

Example: Let us consider the GAP g: *Restaurant* $\rightarrow$ *Theater*, $\sup(g,\mathtt{DB}_{\mathrm{LBSN}})$=1%, $\mathrm{lift}(g,\mathtt{DB}_{\mathrm{LBSN}})$=1.5, $\mathrm{conf}(g,\mathtt{DB}_{\mathrm{LBSN}})$=60% mined from LBSN data. It means that 1% of the users visit (i.e., check-in) a restaurant and then a theater, and when users go to the restaurant then in 60% of the cases the next destination is a theater. Since lift is greater than one, this generalized activity pattern represents a positive correlation, i.e., the likelihood to go to the theater after going to the restaurant is higher than expected.

c) Pattern extraction: The unified pattern-based model consists of a selection of GAPs extracted from either LBSN or mobility data. The aim of the model is twofold: (1) Provide a social view of urban activities by means of the GAPs extracted from the explicit user check-ins. (2) Provide a mobility-level implicit view of the urban activities by means of the GAPs extracted from the traffic traces.

To obtain (1) we focus on the GAPs g whose

- check-in-related support is above a given threshold, i.e., $\sup(g,\mathtt{DB}_{\mathrm{LBSN}})>\mathrm{minsup}$
- check-in-related confidence is above a given threshold, i.e., $\mathrm{conf}(g,\mathtt{DB}_{\mathrm{LBSN}})>\mathrm{minconf}$
- check-in-related lift indicates a positive correlation, i.e., $\mathrm{lift}(g,\mathtt{DB}_{\mathrm{LBSN}})>1$

Similarly, to achieve (2) we consider the GAPs g whose

- trip-related support is above a given threshold, i.e., $\sup(g,\mathtt{DB}_{\mathrm{tr}})>\mathrm{minsup}$
- trip-related confidence is above a given threshold, i.e., $\mathrm{conf}(g,\mathtt{DB}_{\mathrm{tr}})>\mathrm{minconf}$
- trip-related lift indicates a positive correlation, i.e., $\mathrm{lift}(g,\mathtt{DB}_{\mathrm{tr}})>1$

To extract the GAPs we apply an established sequence mining algorithm, namely cSPADE [28], to both LBSN and mobility data.[3]

IV. Bias Detection and Explanation

AI-based models are known to be susceptible to the presence of bias in real-world data. Hence, an increasing research effort has been devoted to proposing new bias detection and mitigation techniques (e.g., [9], [29], [30]). They address bias detection in classification models by comparing the model predictions for a group with those in the ground truth. For example, according to the Statistical Parity (SP) metric [9], a binary classification model is *unbiased* if the members of two groups are equally likely to be assigned to the positive set, independently of their group membership.

Our purpose is to detect the presence of bias between the GAPs mined from LBSN data (R_{LBSN}) and those extracted from mobility data (R_{tr}) within the same spatio-temporal context, i.e., Goal (1). Firstly, we formulate the initial hypothesis of *unbiased model* as follows:

$$P(\hat{Y}=1|R_{tr})=P(\hat{Y}=1|R_{LBSN})$$

where $P(\hat{Y}=1|R_{tr})$ is the probability of the GAPs in R_{tr} to be assigned to the positive set while $P(\hat{Y}=1|R_{LBSN})$ is the probability of the GAPs in R_{LBSN} to be assigned to the positive set.

Next, by assuming that a GAP is assigned to the positive set if it is present in both sets, we compute the two probabilities and we test the initial hypothesis.

$$P(\hat{Y}=1|R_{tr}) = \frac{|R_{tr} \cap R_{LBSN}|}{|R_{tr}|}$$

$$P(\hat{Y}=1|R_{LBSN}) = \frac{|R_{tr} \cap R_{LBSN}|}{|R_{LBSN}|}$$

The idea behind it to verify whether the most relevant user activities are coherent to a large extent.[4]

If the initial hypothesis (i.e., unbiased model) is rejected, then we leverage differences between GAPs to explain the causes of bias. Specifically, we focus on the GAPs present only in R_{tr} or R_{LBSN} (but not in both sets), as potentially represent user habits. GAPs are sorted by decreasing support to highlight the most recurrent anomalies in citizens' activities.

Since the aforesaid procedure can be influenced by the number of considered GAPs, we run multiple tests by varying the k value to identify the largest, unbiased top-k set of GAPs.

V. Pattern Analysis

We explore the unbiased sets of GAPs identified at the previous step to highlight the main similarities and differences between LBSN and mobility data. The purpose is to provide domain experts with a deep characterization of the interchangeability and complementarity of the two data sources.

To allow a direct comparison, GAPs are first sorted as follows.

[3]URL: https://github.com/zakimjz/cSPADE Latest access: March 2022

[4]Whenever not otherwise specified, we will set the minimum significance level to 99%.

- **Social rank**: GAPs are sorted by decreasing confidence and lift values computed on LBSN data ($\mathtt{DB}_{\mathrm{LBSN}}$) and are stored in the ranked list R_{LBSN}. We will refer to the top-k GAPs in $\mathtt{DB}_{\mathrm{LBSN}}$ as $R_{\mathrm{LBSN}}(k)$.
- **Mobility rank**: GAPs are sorted by decreasing confidence and lift values computed on mobility data ($\mathtt{DB}_{\mathrm{tr}}$) and stored in the ranked list R_{tr}. We will refer to the top-k GAPs in $\mathtt{DB}_{\mathrm{tr}}$ as $R_{\mathrm{tr}}(k)$.

Then, we use standard information retrieval metrics to compare the produced rankings R_{LBSN} and R_{tr} for different k values. Specifically, P@k and R@k are commonly used to compare a ranked list with a reference one (namely, the ground truth) [31]. In our context, we apply them to verify the accuracy of trip sequences to model activities observed in LBSN data.

- **Precision at k** P@k: it is computed as the percentage of GAPs in $R_{\mathrm{tr}}(k)$ that occur in R_{LBSN} as well.
- **Recall at k** R@k: it is computed as the percentage of GAPs in R_{LBSN} that occur in $R_{\mathrm{tr}}(k)$.

Since the number k of selected GAPs is significantly lower than the total number of mined patterns the R@k values are typically rather small.

To get a more intuitive quality score version of R@k, we compute the difference between the actual R@k values and the corresponding best value achievable by setting the same k value. We call this metric Differential R@k (DiffR@k). DiffR@k ranges from 0 to 1. The smaller the value of DiffR@k, more consistent the trip and LBSN data are.

VI. Case Studies

We validate the effectiveness and usability of the proposed methodology in two representative urban scenarios: *New York City*, which is the most populous city in the United States, and *Portland*, the Oregon's largest city. They are examples of complex urban environments characterized by a large availability of open data about mobility service usage and Foursquare check-ins. The city maps are annotated with a large set of Points-of-Interests and the related categories. We study two different types of on-demand mobility services, i.e., the *Yellow* taxi service in New York City and the Free Floating Car-Sharing (FFCS) service in Portland.

Table II summarizes the main characteristics of the analyzed data. In both cases, the cardinality of mobility data is orders of magnitude larger than those of LBSN data. Hence, the integration of mobility data is particularly helpful to address the lack of LBSN data.

A. Mobility Data

Historical data about taxi rides in New York City can be retrieved from the open *NYC - Taxi & Limousine Commission* source [32]. It reports for each taxi ride the geo-coordinates of the trip origin and destination, and the starting and ending timestamps. In New York City taxi services are very popular due to the lack of available parkings. In accordance with local privacy laws and standards, taxi identifiers are hidden. Therefore, all user activities and interests can be derived only in an indirect way. In the performed experiments, we focus on the Manhattan district because of the extensive use of taxi services, the widespread usage of LBSNs, and the high population density. The taxi trips dataset for New York City contains around 170 million trips over 14 months (from January 2012 to February 2013), 8.5 million trips in the Manhattan area.

TABLE II
LBSN Data Characteristics

City	Num. Check-ins	Num. Users	Avg. per-user check-ins
New York City	100,879	1,083	93
Portland	1,700	480	3.54

The FFCS service usage dataset for Portland was obtained from the *car2go* provider.[5] In Free-Floating Car Sharing (FFCS) services the rented vehicles can be taken and left anywhere in the operative area. Preventive car reservation is optional and users can verify car availability in real-time through a GPS-based mobile application. The dataset consists of 485000 trips of 316 cars spread over a time period of 19 months (i.e., from June 2012 to December 2013). The raw FFCS data include the history of all car bookings. For each booking the timestamp and the location of each reserved vehicle are known. For our purposes, raw data are transformed by applying the procedure described in [33]. Specifically, it first identifies and early discards the cancelled car reservations by analyzing the travelled vehicle distance, the booking time duration, and the fuel consumption associated with the booking. Then, it generates the history of past trips.

B. LBSN Data

We collect LBSN data from the Foursquare social network by tracing the user visits to specific geo-referenced locations. Beyond the geographical position of the location, we collect the descriptions of a set of POIs annotated by the user about the geo-referenced venue.

For both New York City and Portland, check-in data are retrieved from [34] by considering the same time period used to collect the mobility data. For New York City we selected the check-in data related to the Manhattan area, resulting in around 100,000 check-ins from approximately 1,000 users, with an average number of check-ins per user equal to 93. For the Portland dataset, we collected 1,700 check-ins by 480 users. The social users in Portland are averagely less active than to those in Manhattan, probably due to the lower number of tourist attractions. The average number of check-ins per user is approximately 3.54.

C. Points of Interest

We spatially stratify each urban area into a square grid and annotate each cell with the set of POIs present in it. POIs are then clustered into well-known categories (e.g., restaurant, museum, square), i.e., the Categories of Interest

[5] www.car2go.com latest access: March 2022

(COIs), to provide end-users with high-level views of the analyzed urban activities.

Both the New York City and Portland city maps are annotated with POIs using the OpenStreetMap tool,[6] which leverages the Overpass API.[7]

COIs are extracted using the Google APIs .[8] To map each location to the most relevant POIs we first identify the POIs present either in the same cell of the location or in its surrounding cells. Then, we pick the nearest POIs among the previously selected ones (see Definition 3.3).

The definition of the POI neighborhood depends on the analyzed mobility service. Specifically, for taxi rides we assume that the trip origin and destination are relatively close to the POI as passengers can be dropped in any place. For this reason, we set the maximum neighborhood distance r to 100 meters. Conversely, in FFCS the rented cars need to be parked. Hence, the trip origin and destination are often farther. Thus, we set r to 500 meters.

In the performed experiments we focus on a subset of most popular COIs while neglecting those categories that are either very rare or not relevant to the analyzed case study (e.g., *room*, *route*). The resulting datasets consists of about 18,000 POIs and 17 distinct COIs for the Manhattan area, and about 1,100 POIs and 18 distinct COIs for Portland. COIs are evenly distributed over the analyzed datasets. For example, COIs such as *Food* and *Store/Shop* frequently occur in both mobility and LBSN data. Conversely, *Health* is the most frequent COI in mobility data whereas rarely occurs in LBSN check-ins, probably due to privacy concerns.

D. Spatial Contexts

We conduct a preliminary analysis of the spatial distribution of the taxi rides in Manhattan and identify three geographical areas characterized by *high*, *moderate*, and *low* frequency of taxi rides, respectively. Specifically, they respectively cover the 86%, 12% and 2% of the overall number of taxi rides. In Figure 2 (a) the selected areas are depicted in orange, yellow, and green, respectively. In the performed experiments, we mainly focus on orange area, which approximately corresponds to the Manhattan centre.

FFCS usage data in Portland shows a relatively homogeneous spatial distribution. Thus, we stratify the Portland urban area in five regions (i.e., the city center and four other regions corresponding to the cardinal point directions) independently of the number of observed trips (see Figure 2 (b)). The city center area is defined as a square with a side length of 3km. The remaining four areas are defined by connecting the edges of the city center area to the edges of the operating area.

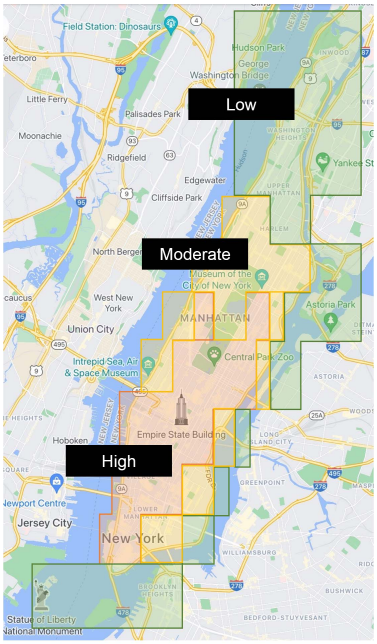
E. Temporal Contexts

We extract GAPs at different temporal granularities to capture a variety of different urban lifestyles.

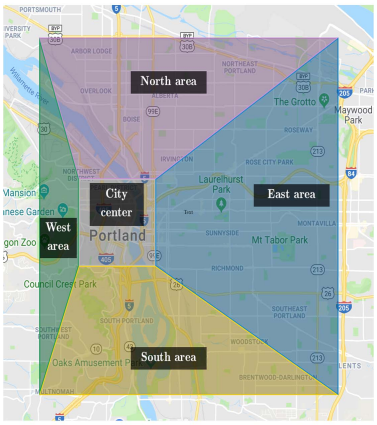
[6]https://www.openstreetmap.org latest access: March 2022

[7]http://www.overpass-api.de latest access: March 2022

[8]https://developers.google.com/maps/documentation/places/web-service/supported_types#table2 latest access: June 2021



(a)



(b)

Fig. 2. Reference spatial contexts in new york city (a) and Portland (b).

Due to the inherent characteristics of the original data distribution, we choose different time granularities for the mobility and LBSN services. Specifically, we separately analyze different monthly periods for the taxi service in New York City and FFCS data in Portland whereas we aggregate LBSN data acquired in consecutive months due to their sparsity (lack of user data).

VII. Experimental Results

We run the experiments were run on a multi-core 2.67 GHz Intel(R) Xeon(R) workstation with 32 GB of RAM with Ubuntu Linux 18.04 LTS.

A. Configuration Settings

To limit the computational complexity in our experiments we enforce (1) a maximum sequence length n equal to 3, i.e.,

we generate sequences no longer than 3 COI sets, and (2) a maximum COI set size $|\mathbf{COI}_i|$ equal to 3, i.e., each COI set in the sequence contains no more than 3 COI categories.

To deepen the analysis of short-term urban movements we extract the sequential patterns (GAPs) by enforcing the `mintimegap` and `maxtimegap` to 15 and 60 minutes, respectively. This implies the covered time span never exceeds one hour and the temporal distance between consecutive events is above 15 minutes.

The value of the support threshold (minsup) is adapted to the cardinality and sparsity input data distribution. Specifically, for New York City we set the support threshold to 50% for the mobility dataset and to 1% for the LBSN dataset, whereas for Portland we set it to 4% for the mobility dataset and to 0.06% for the LBSN dataset. The values of the confidence and lift thresholds are set to 50% and 10, respectively.

B. Bias Detection and Explanation: Results Overview

We report here some examples of bias detection outcomes relative to one representative month (May) in Manhattan (New York City). The selected context is characterized by a relatively high variability in the activity patterns. The results are in line with those achieved in the other months.

Table III reports the distribution, in percentage, of the COIs occurring in the sets of uncommon GAPs.[9] COIs such as *Health*, *Bank*, and *Diplomacy* are present only in the GAPs mined from the mobility data. These exceptions are likely due to privacy concerns, which prevent LBSN users from disclosing such sensitive information. Conversely, GAPs mined from mobility data include the aforesaid COIs, partially compensating the lack of LBSN data. COI *Theater* is instead present only in the LBSN GAPs. Theaters are popular venues for entertainment activitie. They are frequently annotated by LBSN users, whereas appear to be less common in mobility data.

In Table III, we can notice two COIs occurring in both sets: *Food* and *Entertainment*. This means that Food and Entertainment frequently occur in the mined GAPs, independently of the data source. However, they co-occur with different COIs depending on the data source from which they have been mined. For instance, GAP *Diplomacy* $\rightarrow$ *Food* is mined only from mobility data, whereas *Theater* $\rightarrow$ *Food* is mined only from LBSN data. By analyzing only the frequency of single COIs in the data sources we cannot identify this type of discrepancy and bias. GAPs, which identify correlations between sets of COIs, is a valuable support to identify also differences and bias due to combinations of events, which are representative of different user habits.

The bias detection test applied to the whole set frequent and reliable GAPs reject the initial hypothesis of unbiased model. Conversely, by setting k to 65 or less, the top-k GAPs appear to be not affected by bias. Therefore, hereafter we will consider such a pattern shortlist to explore the similarities between the GAPS mined from the two data sources.

[9]Note that the sum of the percentages do not sum to 100% because each GAP contains many COIs.

TABLE III
PERCENTAGE OF GAPS MINED FROM THE MOBILITY/ LBSN DATA INCLUDING A GIVEN COI

Mobility data	
COI	**Perc. of mobility GAPs containing COI**
Health	48.0%
Food	42.4%
Bank	39.2%
Store/Shop	38.1%
Entertainment	28.0%
Public Transport	22.8%
Diplomacy	20.9%
Bar/Cafe	16.3%
Market	14.6%
Post Office	14.4%
intersection	14.1%
Restaurant	10.1%
School/College	6.9%
LBSN data	
COI	**Perc. of LBSN GAPs containing COI**
Theater	71.4%
Food	50.0%
Entertainment	35.7%
Car related places	28.6%

C. Pattern Analysis

We separately analyze the top-k GAPs ($k \leq 65$) in terms of P@k and DiffR@k to explore similarities and differences in the user habits.

1) Precision@k: Figure 3 (a) plots the P@k value for increasing values of k. The experimental results show that the precision has its maximum value (P@k=1) when $k \leq 20$. Then, the precision decreases gradually up to k=50, but with values of P@k greater than 0.55, and more significantly for $k > 50$. Therefore, the top 20 GAPs and approximately 55% of the top 50 GAPs in the mobility dataset are consistent with LBSN data. The pattern-based methodology indicates that, in practice, the user activities discovered in different domains (Foursquare and taxi services) are strongly correlated with each other while focusing on the top ranked patterns. Thanks to their inherent interpretability, end-users can explore these patterns to investigate the underlying motivations behind the reported activities. A qualitative analysis of real-world GAPs is reported in Section VII-D.

2) Differential Recall@k: Figure 3 (b) plots the DiffR@k values achieved by setting various values of k between 1 and 50. When $k \leq 20$ DiffR@k is equal to zero because all the top ranked GAPs extracted from mobility data have an exact correspondence with a GAP in the LBSN data. This supports the hypothesis that the two data sources are fairly consistent with each other. When $k \geq 20$ the DiffR@k value increases, but is always 0.18 until $k = 50$, meaning that the gap between the best R@k value and the actual value is at most 18%.

D. Qualitative Analysis of the Discovered Patterns

a) User activities in common between mobility and LBSN data: We analyze here a selection of representative GAPs that are relevant to both LBSN and mobility data.

Example: Diplomacy $\rightarrow$ *Restaurant* is an example of GAP in common between the analyzed data sources. The aforesaid GAP has a confidence value equal to 100% in both data sources (taxi and Foursquare). It represents a recurrent connection between the visit to a diplomacy POI followed by

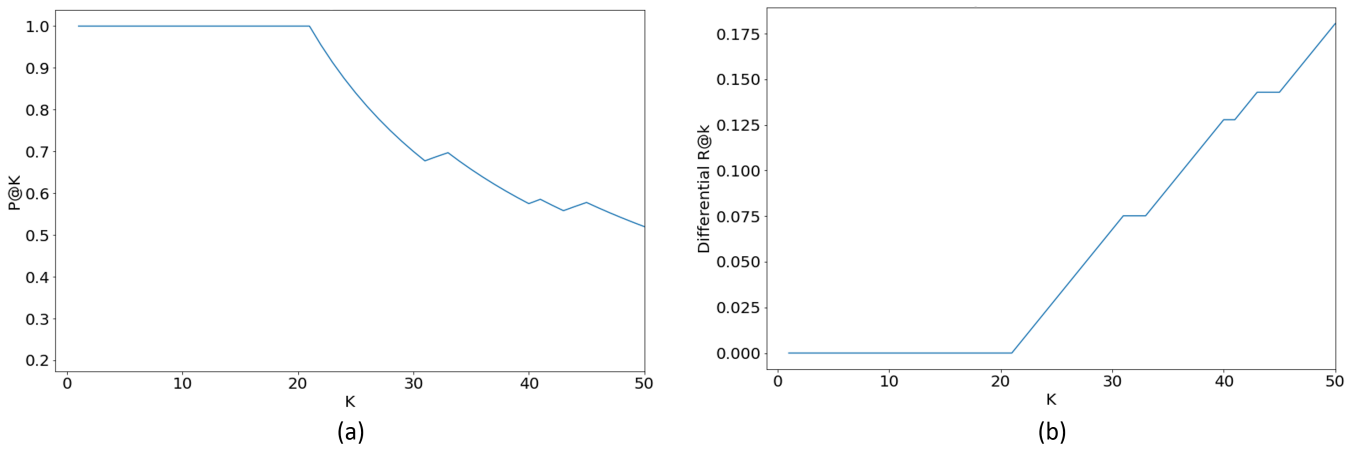


Fig. 3. Comparative analysis. new york city. May 2012.

a lunch/dinner in a restaurant. By exploring the trip sequences covered by the selected GAP we find the following trips: *Consulate general of the Philippines → Cuban restaurant on the 8th avenue*, *Permanent Mission of the Plurinational State of Bolivia to the United Nations → Mexican Restaurant Taco Dumbo*, and *Consulate General of Japan in New York → Thai restaurant on the 8th avenue*. The extracted GAP, and the corresponding trips, describe an activity pattern, which consists in having lunch/dinner after a diplomacy meeting to conclude an event in an official venue. The extracted GAP highlights a user habit that cannot be detected considering POIs instead of COIs. Another example is given by the trips *Courant Institute of Mathematical Science → Times Square* and *NYU Stern School of Business → The Bitter End* that support the GAP *School/College → Entertainment*, which characterizes peculiar students' activities. The municipality of Manhattan could leverage these patterns to plan the most appropriate location for new services to (i) increase the citizens' experience, (ii) reduce the average trip time, and (iii) reduce traffic and pollution.

The heatmap in Figure 4 summarizes the coverage level of the GAPs extracted from the mobility service over LBSN data separately for each of the analyzed months. Specifically, for each 60 GAP mined from the LBSN data it indicates whether it is covered by any top ranked GAP extracted from the mobility data (covered patterns are represented by blue squares). On the x-axis, the GAPs in the ranked list R_{LBSN} are sorted by decreasing confidence value and lift value. The plot shows that (1) the top ranked LBSN patterns are covered by mobility data for most of the analyzed months. (2) Some activities in common between LBSN and mobility services are peculiar to specific months whereas other independent of the analyzed monthly period. (3) As expected, August is an outlier month as citizen activities significantly change.

The importance of the extracted activity patterns, measured in terms of GAP confidence and lift values, varies over the analyzed months. Focusing on the subset of GAPs relative to a specific POI category, we can analyze the evolution of the corresponding patterns' rankings, and thus user activities, over time.

Example: GAPs relative to *Health* and *Theater* are very important in March, whereas the GAPs in the top-20 include

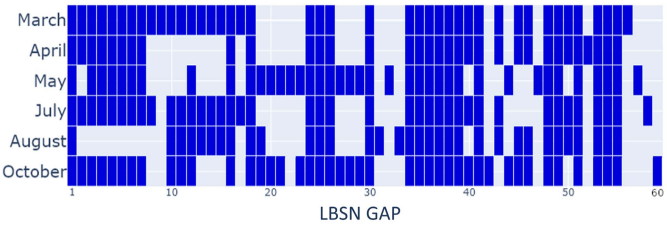


Fig. 4. Heatmap of the 60 check-in-related GAPs mined from new york city matched by a trip-related GAP.

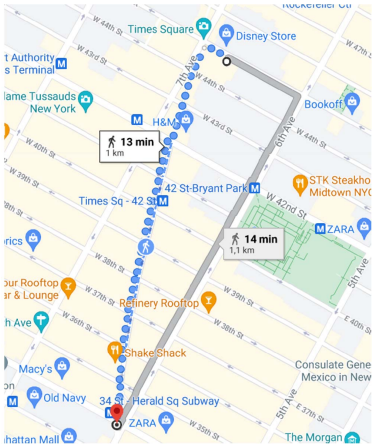


Fig. 5. Examples of check-in sequences associated with GAPs mined from LBSN data.

the category *Diplomacy services* only in July and August. This is probably due to the common need to retrieve the necessary documents to travel in summertime. Categories *School/College* and *Entertainment* have a greater significance in April, May and October. This could be linked to the academic calendar and the greater recreational activity during the spring/autumn months favoured by more pleasant temperatures. Finally, GAPs related to the category *Bars/Cafes* are fairly relevant to most of the analyzed months.

b) User activities peculiar to the mobility service: We seek the top ranked GAPs extracted from the mobility data that are not equally relevant to the LBSN context. They represent user habits that are related to either "sensitive" or less "social" aspects.

Example: Categories *Bank* and *Health* are typical of the subset of GAPs peculiar to the mobility service. They are related

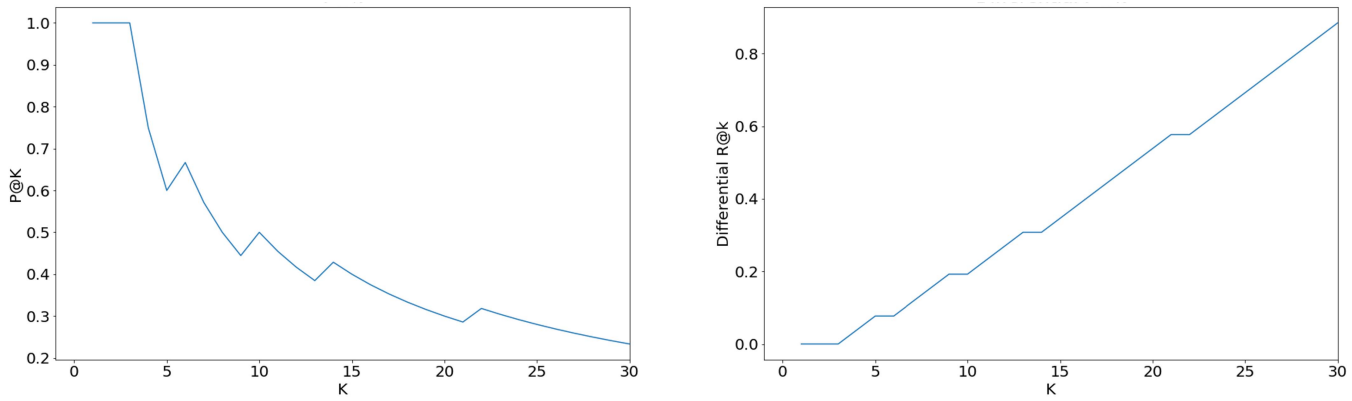


Fig. 6. Comparative analysis. Portland. May 2012.

to healthy conditions, work activities, or routine management of personal business. The associated venues do not correspond to the usual POIs for which users wish to show their presence through check-ins. User movements between these kinds of POIs can thus be captured through the analysis of the mobility data but not through the analysis of the LBSN data.

c) User activities peculiar to the LBSN context: These patterns represent user movements between rather close venues (possibly coupled with a scarce availability in the area of the considered mobility means). Users can make these movements without having to use a mobility means like the taxi. Despite users notify these movements through the LBSN service they do not use any mobility means thus the corresponding activities cannot be revealed by analyzing mobility data.

Example: Theater $\rightarrow$ *Entertainment* and *Restaurant* $\rightarrow$ *Entertainment* are examples of GAPs mined from LBSN data only. Here we discuss three examples of POI sequences matching the GAP *Theater* $\rightarrow$ *Entertainment*. In all the cases, the sequence origin and destination are located at a walking distance, i.e., 2 minutes, 9 minutes and 13/14 minutes on foot.[10] Specifically, one of the three POI sequences refers to a movement started from the Magnet Theater and ended in the Greeley Square Park. The other POI sequence started from David H. Koch Theater and ended in the Ballfields Café located in Central Park. The third POI sequence is from the Lyceum Theater to the Herald Square Plaza. As an example, this latter sequence is shown in Figure 5.

E. The Portland Case Study

We summarize here the main results achieved on the Portland dataset (see the key statistics in Table II), which comprises Foursquare check-ins and Free-Floating Car Sharing service usage data. Despite in Portland the Foursquare social service is quite popular users tend to check-in only few venues located in the city center. The relatively limited number of social user check-ins compared to the number of available car trips (485,000 trips vs. 1,700 check-ins) and their imbalanced spatial distribution make Portland as a particularly challenging scenario in which domain experts could leverage the integration of mobility data.

[10]We estimated the walking distances by means of the GoogleMaps service available at https://maps.google.com latest access: March 2022.

1) Comparison Between Mobility- and Check-in-Based GAPs: We compute the P@k and DiffR@k values to evaluate the similarities between the GAPs mined from the mobility data and those extracted from Foursquare in Portland as well. The LBSN-based patterns are again considered as the ground truth. However, since the cardinality of LBSN data is rather limited, the associated GAPs cannot be considered reliable like those mined from New York City. Due to the inherent sparsity of LBSN data, in Portland only 25 GAPs are mined from Foursquare.[11]

We compare the GAPs mined from mobility data with the 25 extracted from LBSN data.[12] The plots in Figure 6 show the P@k and DiffR@k values by varying the number k of selected GAPs. The precision and recall values are optimal when $k \leq 3$, i.e., the top 3 activity patterns are in common between LBSN and FFCS usage data. Then, the performance worsens but is still acceptance until $k \geq 10$ (e.g., precision above 45%). Notice that when there is a lack of LBSN data the most valuable information is provided by the new activity patterns not present in LBSN data, as mobility data are more likely to represent reliable life-style patterns.

2) Effect of Spatio-Temporal Contexts: We also compare the GAPs extracted from mobility data within different city areas and time periods with those extracted from LBSN data.

Example: In Portland the GAP *Restaurant* $\rightarrow$ *Shop* ranked first in the East area, probably due to the movements of tourists beyond to those of local people. Conversely, *Restaurant* $\rightarrow$ *Restaurant* ranked first in the north, as the density of restaurants within that area is relatively high. The central area is characterized by homogeneous activity trends: the mined activity patterns all have similar quality measures and include most of the GAPs that already appear in North and the East areas.

We explore different time granularities beyond the monthly periods. For example, in Portland a drill down on the temporal dimension to the weekly granularity confirms the main results, in terms of P@k and DiffR@k values, achieved with the monthly periods. Furthermore, the majority of the GAPs extracted at the monthly aggregation level are still relevant at the weekly level as well.

[11]We set the minimum relative support to 0.06% for the check-in data.

[12]In the experiments we set the minimum support threshold to 0.06% to extract the COI sequences that occur at least twice.

VIII. Discussion and Conclusions

In this paper, we proposed to complement the analysis of LBSN data using mobility data. The main purpose was to characterize the activities of citizens in urban environments based on both the explicit annotations made through the LBSN check-ins and the implicit feedback provided by the geo-referenced trip data acquired by mobility service providers.

The main stakeholders are city planners and urban designers, e.g., municipality managers and city council delegates, who are in charge of

- *Shape urban areas* to meet the citizens' demand and *foreseen emerging urban areas*. For example, GAPs such as *Food* → *Entertainment* indicate a joint interest for culinary and entertainment activities, whose related services can be offered within the same urban district (at a short distance). Entrepreneurs can also take advantage of these targeted recommendations to plan business activities and shape the existing services.
- *Suggest* to people having access to a service in a given POI category (e.g., "Food") the next possible POI category of interest (E.g., "Entertainment").
- *Target specific customer needs* by analyzing behavioral data of specific customer segments (e.g., city tourists and residents, young/middle age/old people). For example, the proposed solution can be helpful to mobility managers who are involved in the planning of transport services, with particular attention to issues like accessibility (e.g., offer the access to both kinds of facilities to the disabled).

The takeaways of the research can be summarized as follows.

Explicit vs. implicit tagging. So what? LBSN services like Foursquare foster the user annotation of the published check-ins. Hence, the user intentions are explicit. Conversely, in mobility data the trip endpoints could approximately indicate the POIs of interest thus providing implicit feedback on users' habits. The data-driven methodology presented in this paper bridges the gap between explicit and implicit POI tagging, providing domain experts with a quantitative strategy to assess the coherence of the two data types. Implicit tags are particularly relevant to complement explicit content when there is a lack of User-Generated Content.

Are LBSN and mobility data actually complementary? At a first glance, the whole set of patterns derived from the two data sources show substantial differences (bias). However, based on our empirical evidence, we can identify a shortlist of patterns that are highly similar to each other (approximately 60 GAPs). A detailed analysis of these common patterns revealed that the corresponding user habits are coherent to a large extent. Therefore, under the aforesaid conditions, the complementarity and interchangeability of the two data sources are preserved.

How can we complement LBSN data with mobility data? The experiments carried out on Foursquare data highlighted specific time periods and spatial regions when the POI sequences visited by the social users reflect the taxi rides or the trips associated with car sharing services. This is particularly helpful to overcome the lack of LBSN data for specific, privacy-sensitive services (e.g., finance, healthcare, religion) and for particular spatio-temporal contexts (e.g., when LBSN services are temporarily out of order).

As a future work, we plan to extend the study to other mobility services, such as shared scooter and bikes, and social platforms, such as Instagram. We also aim at developing an integrated context-aware platform, which automatically discovers the underlying correlations between the data sources and recommends to domain experts specific Key Performance Indicators (KPIs).

Acknowledgment

The authors would like to thank Alexander Sebastian Abstreiter for implementing the code used to collect and analyze the Portland data.

References

[1] P. Symeonidis, D. Ntempos, and Y. Manolopoulos, "Location-based social networks," in *Recommender Systems for Location-Based Social Networks*. New York, NY, USA: Springer, 2014, pp. 35–48, doi: 10.1007/978-1-4939-0286-6_4.

[2] T. Üsküplü, F. Terzi, and H. Kartal, "Discovering activity patterns in the city by social media network data: A case study of Istanbul," *Appl. Spatial Anal. Policy*, vol. 13, no. 4, pp. 945–958, Dec. 2020.

[3] A. Dunkel, "Visualizing the perceived environment using crowdsourced photo geodata," *Landscape Urban Planning*, vol. 142, pp. 173–186, Oct. 2015.

[4] D. Tasse and J. I. Hong, "Using social media data to understand cities," Carnegie Mellon Univ., Pittsburgh, PA, USA, 2018, doi: 10.1184/R1/6470645.v1.

[5] S. Hasan and S. V. Ukkusuri, "Location contexts of user check-ins to model urban geo life-style patterns," *PLoS ONE*, vol. 10, no. 5, pp. 1–19, May 2015.

[6] J.-S. Kim *et al.*, "Location-based social network data generation based on patterns of life," in *Proc. 21st IEEE Int. Conf. Mobile Data Manage. (MDM)*, Versailles, France, Jun./Jul. 2020, pp. 158–167, doi: 10.1109/MDM48529.2020.00038.

[7] B. Huang and K. M. Carley, "A large-scale empirical study of geotagging behavior on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2019, pp. 365–373.

[8] B. J. Hecht and M. Stephens, "A tale of cities: Urban biases in volunteered geographic information," in *Proc. 8th Int. Conf. Weblogs Social Media (ICWSM)*, Ann Arbor, MI, USA: AAAI Press, Jun. 2014, pp. 197–205. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114

[9] E. Mishraky, A. B. Arie, Y. Horesh, and S. M. Lador, "Bias detection by using name disparity tables across protected groups," *J. Responsible Technol.*, vol. 9, Apr. 2022, Art. no. 100020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666659621000135

[10] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel, "Recommendations in location-based social networks: A survey," *GeoInformatica*, vol. 19, no. 3, pp. 525–565, 2015.

[11] E. Malmi, T. M. T. Do, and D. Gatica-Perez, "From foursquare to my square: Learning check-in behavior from multiple sources," in *Proc. 7th Int. Conf. Weblogs Social Media (ICWSM)*. Cambridge, MA, USA: AAAI Press, Jul. 2013, pp. 701–704.

[12] A. Talpur and Y. Zhang, "A study of tourist sequential activity pattern through location based social network (LBSN)," in *Proc. Int. Conf. Orange Technol. (ICOT)*, Oct. 2018, pp. 1–8.

[13] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro, "A comparison of foursquare and Instagram to the study of city dynamics and urban social behavior," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput. (UrbComp@KDD)*, Chicago, IL, USA, 2013, pp. 4:1–4:8, doi: 10.1145/2505821.2505836.

[14] Y. Liu, C. Liu, X. Lu, M. Teng, H. Zhu, and H. Xiong, "Point-of-Interest demand modeling with human mobility patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, Aug. 2017, pp. 947–955, doi: 10.1145/3097983.3098168.

[15] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, and Z. Yao, "A general geographical probabilistic factor model for point of interest recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1167–1179, May 2015.

[16] H. Katsumi, W. Yamada, and K. Ochiai, "Generic POI recommendation," in *Proc. Adjunct Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., ACM Int. Symp. Wearable Comput. (UbiComp-ISWC)*, Sep. 2020, pp. 46–49, doi: 10.1145/3410530.3414421.
[17] A. Popescu and A. Shabou, "Towards precise POI localization with social media," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, Oct. 2013, pp. 573–576, doi: 10.1145/2502081.2502151.
[18] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2007, pp. 330–339, doi: 10.1145/1281192.1281230.
[19] S. Kisilevich, D. A. Keim, and L. Rokach, "A novel approach to mining travel sequences using collections of geotagged photos," in *Geospatial Thinking* (Lecture Notes in Geoinformation and Cartography). Guimarães, Portugal: Springer, May 2010, pp. 163–182, doi: 10.1007/978-3-642-12326-9_9.
[20] X. Long, L. Jin, and J. Joshi, "Exploring trajectory-driven local geographic topics in foursquare," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, Pittsburgh, PA, USA, Sep. 2012, pp. 927–934, doi: 10.1145/2370216.2370423.
[21] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 1082–1090, doi: 10.1145/2020408.2020579.
[22] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput. (UrbComp@KDD)*, Chicago, IL, USA, Aug. 2013, pp. 6:1–6:8, doi: 10.1145/2505821.2505823.
[23] A. I. J. T. Ribeiro, T. H. Silva, F. Duarte-Figueiredo, and A. A. F. Loureiro, "Studying traffic conditions by analyzing foursquare and Instagram data," in *Proc. 11th ACM Symp. Perform. Eval. Wireless Ad hoc, Sensor, Ubiquitous Netw. (PE-WASUN)*, 2014, pp. 17–24, doi: 10.1145/2653481.2653491.
[24] A. Kheiri, F. Karimipour, and M. Forghani, "Intra-urban movement flow estimation using location based social networking data," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. XL15, pp. 781–785, Dec. 2015.
[25] H. Kavak, J.-S. Kim, A. Crooks, D. Pfoser, C. Wenk, and A. Züfle, "Location-based social simulation," in *Proc. 16th Int. Symp. Spatial Temporal Databases*, Vienna, Austria, Aug. 2019, pp. 218–221, doi: 10.1145/3340964.3340995.
[26] Y. Chen, J. Hu, Y. Xiao, X. Li, and P. Hui, "Understanding the user behavior of foursquare: A data-driven study on a global scale," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 1019–1032, Aug. 2020.
[27] G. Deeva, J. D. Smedt, J. D. Weerdt, and M. Óskarsdóttir, "Mining behavioural patterns in urban mobility sequences using foursquare check-in data from Tokyo," in *Complex Networks and Their Applications VIII* (Studies in Computational Intelligence), vol. 882. Lisbon, Portugal: Springer, Dec. 2019, pp. 931–943, doi: 10.1007/978-3-030-36683-4_74.
[28] M. J. Zaki, "Sequence mining in categorical domains: Incorporating constraints," in *Proc. 9th Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2000, pp. 422–429, doi: 10.1145/354756.354849.
[29] Y. Horesh, N. Haas, E. Mishraky, Y. S. Resheff, and S. M. Lador, "Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning," in *Machine Learning and Knowledge Discovery in Databases* (Communications in Computer and Information Science), vol. 1167, P. Cellier and K. Driessens, Eds. Würzburg, Germany: Springer, Sep. 2019, pp. 590–604, doi: 10.1007/978-3-030-43823-4_47.
[30] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," in *Proc. Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 525–534, doi: 10.1145/3351095.3372878.
[31] P. Dadure, P. Pakray, and S. Bandyopadhyay, "Mathematical information retrieval trends and techniques," in *Deep Natural Language Processing and AI Applications for Industry 5.0*. Hershey, PA, USA: IGI Global, 2021, pp. 74–92.
[32] *TLC Trip Record Data*. Accessed: Mar. 1, 2022. [Online]. Available: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
[33] E. Daraio, L. Cagliero, S. Chiusano, P. Garza, and D. Giordano, "Predicting car availability in free floating car sharing systems: Leveraging machine learning in challenging contexts," *Electronics*, vol. 9, no. 8, p. 1322, Aug. 2020.
[34] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–23, Apr. 2016.

Elena Daraio (Student Member, IEEE) received the M.Sc. degree in computer engineering from the Politecnico di Torino, where she is currently pursuing the doctoral degree with the Department of Control and Compute Engineering. Her current research focus is on supervised and unsupervised machine learning algorithms. Her research interests are in the field of urban intelligence, and in particular related to the analysis of mobility systems, such as car sharing and on demand systems and to the analysis of physiological signals to detect drivers anomalous conditions. She has been part of the program committee of international conferences (e.g., KDD, LOD, and DARLI-AP).

Luca Cagliero (Member, IEEE) received the M.Sc. degree in computer and communication networks and the Ph.D. degree in computer engineering from the Politecnico di Torino. He has been an Associate Professor with the Dipartimento di Automatica e Informatica, Politecnico di Torino, since January 2020. His current research interests are mainly related to deep natural language processing, machine learning, and applied data science and, specifically, on text summarization, pattern mining, and classification. Currently, he is an Associate Editor of the *Expert Systems With Applications* (Elsevier) and *Machine Learning With Applications* (Elsevier) and serves as a TPC Member/reviewer for various international conferences/journals.

Silvia Chiusano (Member, IEEE) is a Full Professor in computer engineering with the Politecnico di Torino. She works in the area of database systems and machine learning. Her research interests focus on the design of innovative solutions for large-scale data management and mining with interest on health care, smart cities, and healthy cities application domains. She was the Co-Chair of the Workshop Track at ADBIS 2018 and the Co-Chair of six editions of the DARLI-AP Workshop on data analytics, co-located with IEEE Smart Data in 2017 and the EDBT/ICDT Joint Conference from 2018 to 2022.

Paolo Garza (Member, IEEE) received the master's and Ph.D. degrees in computer engineering from the Politecnico di Torino. He has been an Associate Professor with the Dipartimento di Automatica e Informatica, Politecnico di Torino, since December 2018. His current research interests are in the fields of data mining, database systems, and big data analytics. He has worked on classification, clustering, itemset mining, and scalable algorithms.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement