Noname manuscript No.
(will be inserted by the editor)

Bias Characterization, Assessment, and Mitigation in Location-based Recommender Systems

Pablo Sánchez · Alejandro Bellogín · Ludovico Boratto

Received: date / Accepted: date

Abstract Location-Based Social Networks stimulated the rise of services such as Location-based Recommender Systems. These systems suggest to users points of interest (or venues) to visit when they arrive in a specific city or region. These recommendations impact various stakeholders in society, like the users who receive the recommendations and venue owners. Hence, if a recommender generates biased or polarized results, this affects in tangible ways both the experience of the users and the providers' activities. In this paper, we focus on four forms of polarization, namely venue popularity, category popularity, venue exposure, and geographical distance. We characterize them on different families of recommendation algorithms when using a realistic (temporal-aware) offline evaluation methodology while assessing their existence. Besides, we propose two automatic approaches to mitigate those biases. Experimental results on real-world data show that these approaches are able to jointly improve the recommendation effectiveness, while alleviating these multiple polarizations.

Keywords POI recommendation · Bias mitigation · Polarization · Temporal Evaluation

Pablo Sánchez
Instituto de Investigación Tecnológica (IIT), Universidad Pontificia Comillas. C. Alberto Aguilera 25. E28015. Madrid. Spain
Universidad Autónoma de Madrid, Madrid, Spain
E-mail: psperez@icai.comillas.edu
E-mail: pablo.sanchezp@uam.es

Alejandro Bellogín
Universidad Autónoma de Madrid, Madrid, Spain
E-mail: alejandro.bellogin@uam.es

Ludovico Boratto
University of Cagliari, Cagliari, Italy
E-mail: ludovico.boratto@acm.org

1 Introduction

Artificial Intelligence (AI)-based systems are known to typically perform worse for minorities and marginalized groups (Buolamwini and Gebru 2018; Koenecke et al. 2020; Obermeyer et al. 2019). This lower effectiveness might have a concrete impact on the users interacting with these systems, such as allocational and representational harms (Jacobs et al. 2020; Blodgett et al. 2020). One of the research areas where AI-based systems are commonly used and where the analysis of these biases might be particularly relevant is the recommendation domain. Recommender Systems (RSs) are software tools that help users finding relevant items. Due to their ability to adapt to users' needs, they have been applied in various disciplines (Ricci et al. 2015). As such, they are one type of AI technique that is being increasingly used nowadays, and hence, may affect society as a whole by amplifying existing biases or guiding people's decisions. In fact, RSs are known to be multi-stakeholder environments (Abdollahpouri et al. 2019a), since they affect multiple actors in a direct way, mainly the users receiving the recommendations (consumers) and those behind the recommended objects (providers). Because of that, research on bias analysis and fairness measurements is needed; in particular, specific definitions, dependency variables, and mitigation approaches beyond those already studied for general Machine Learning (Zehlike et al. 2020).

Tourism is a domain where the needs of consumers and the services offered by providers naturally meet in the real world. In the tourism industry, travel guides/blogs have always been used to organize trips. However, while travel portals and travel guides tend to focus on the most popular places (which can be useful in many cases), recommendation algorithms should also offer users more novel recommendations, to provide them satisfying experiences (Massimo and Ricci 2022). For this reason, *tourism recommendation*, where AI models automatically support decision-making processes, clearly impacts on society. Hence, it is an area that is particularly sensitive to these effects and biases. Several recommendation tasks related to tourism have been addressed, such as tour recommendation to groups (Herzog and Wörndl 2019), trajectory recommendation (Chen et al. 2016), suggestion of travel packages (Benouaret and Lenne 2016), etc. Probably, the most important recommendation task related to tourism is the Point-Of-Interest (POI) or venue recommendation problem, which focuses on suggesting to users new places to visit when they arrive in a city (Zhang and Chow 2015; Liu et al. 2014). The POI recommendation problem is usually defined upon data stored in Location-Based Social Networks (LBSNs) (Doan and Lim 2019). These social networks allow users to check-in in venues; thanks to these check-ins, platforms such as Foursquare can provide services to the users, like the possibility to share information between them, together with venue search and/or recommendation. At the same time, based on reviews, ratings, and venue check-ins available in LBSNs, users decide what to buy or consume and where to go. However, generating recommendations in LBSNs introduces new challenges with respect to traditional recommendation, such as different contextual dimensions (temporal, geographical, social, and so on), and a higher sparsity on the user preferences (Li et al. 2015; Wang et al. 2013; Liu et al. 2017; Kapcak et al. 2018). From now on, we will refer to RSs that operate in LBSNs as *Location-Based Recommender Systems* (LBRSs).

In this context, it is critical to assess the extent to which LBRSs have a concrete impact on the tourism domain as a whole. Besides the users accepting the recommendations (the consumers), whose experience in a city depends on these suggestions, the business of venue owners/managers (the providers) strongly depends on them. Hence, we must think of properties of a RS that go *beyond accuracy*, to provide equitable suggestions. Thus, RSs might be polarized towards certain undesired properties (e.g., by recommending only popular items) and this would concretely impact the involved stakeholders in different ways. In the end, not exposing the full catalog of candidate venues to the users might not be fair from a business perspective (Wasilewski and Hurley 2018) and may also lead to a lack of novelty and diversity in the recommendations. As a consequence, the most widely known type of polarization in recommender systems is towards *item popularity*, which means that only a subset of popular items is recommended to the user. Polarized recommendations towards popular venues would worsen user experience, since they might get too crowded, and it might also strengthen inequalities between venue owners/managers. *Venue category* can also be characterized by a certain popularity, which can impact POI recommendation and society at a broader (and probably more dangerous) level. Indeed, users might not be recommended possibly interesting but unpopular categories of POIs (thus probably ignoring their fine-grained preferences) and the owners of an entire sector/type of business might be affected as a whole by it. Item popularity may also affect the exposure of the venues, since popular venues are always ranked in higher positions. Hence, these venues would increase their chances of being noticed and selected by the users (Singh and Joachims 2018), while other interesting items may go unnoticed by the user (*exposure bias*). Finally, a *geographical* polarization towards far away or close POIs with respect to those the user is currently visiting, might ignore their preferences and previous interactions. This polarization would affect the trust of the users on the recommender system (and, again, their experience) and impact owners of more relevant venues. The problem of under-recommending and under-exposing providers is well known in the recommender systems literature (Mehrotra et al. 2018), but to the best of our knowledge, it has never been studied for LBRSs.

It should be clear that polarization might be related to the concept of algorithmic bias, which has been widely studied in recommender systems (Jannach et al. 2015; Bellogín et al. 2017; Boratto et al. 2019; Abdollahpouri et al. 2017; Adamopoulos et al. 2015; Adomavicius et al. 2014; Ekstrand et al. 2018; Guo and Dunson 2015; Jannach et al. 2016). Algorithmic bias assumes that RSs reinforce a previously existing bias in the data. While a pre-existing bias might be the cause of polarization, our focus is at a societal level, to study *the impact of polarization for the involved stakeholders*. In other words, it does not really matter in the context of this work if a venue is popular in the recommendations because it already was or because the system made it popular. Heavily polarized recommendations have a negative impact on tourism stakeholders, so we study these phenomena, without any assumption of the prior distribution of the data. To summarize, in this work, we use the term *polarization* to quantify to what extent an algorithm deviates from what it is observed in the training data[1]. We use the term *bias* to describe, in a more generic way, the inclination of an algorithm to go towards polarization. As our results will show, in LBRSs, polarization is a phenomenon that appears independently of how the data were generated. This, in particular, includes cases where data is biased towards some algorithms (such as popularity) or sensitive features of users (gender or race) or items (higher advertising budgets).

In this work, we characterize the four previously mentioned forms of polarization (i.e., towards venue and category popularity, venue exposure, and geographical distance) through metrics that have not been used before. Then, we assess if the use of check-ins to capture the interactions of the users with a LBSN to produce recommendations may lead to polarized suggestions from these perspectives. To do this, we consider an evaluation methodology that mimics the real world, by using a temporal split of the user check-ins. We then compare different families of recommender systems to inspect these forms of polarization. In order to show to what extent a recommender might be affected by different forms of polarization, it is useful to characterize these phenomena independently. However, at the same time, mitigating these forms of polarization separately would not be adequate, since the objective is to produce recommendations that are as non-polarized as possible (regardless of the type of polarization). As previously mentioned, each polarization affects stakeholders in different and negative ways; hence, dealing only with a form of polarization would still lead to negative outcomes. For this reason, we propose two forms of mitigation based on the concept of hybrid recommendation (Burke 2002) and re-ranking (Abdollahpouri et al. 2019b). Both approaches will allow us to deal with multiple forms of polarization at the same time by combining the outcomes of different recommenders.

2 Background and related work

2.1 Recommender Systems

The purpose of a Recommender System is to provide recommendations of different types of items to a particular user by analyzing their interests and tastes (Ricci et al. 2015). These items vary considerably depending on where we apply the recommender (e.g., movies, books, online dating, businesses, etc). This wide variety of applications has led to the development of a large number of different recommendation techniques. The most extended ones are the content-based models (de Gemmis et al. 2015), which exploit the features of

[1] Note that polarization cannot be computed in an absolute way as no ground truth is available. As a surrogate, we assume the *observed* interactions in the system (i.e., training data) represent, to some extent, the target distributions against which we want to compare.

users and items to make the recommendations, and the collaborative-filtering approaches, that can be divided into two different families. The first of them, memory-based or k-nn methods (Ning et al. 2015), compute similarities between users and/or items to build recommendations. The second family, known as model-based algorithms (e.g., classic matrix factorization models or more recent proposals based on neural networks) (Koren and Bell 2015), uses the information of the interactions between the users and items in order to create a predictive model. Finally, another popular technique in the area are hybrid approaches. These methods combine different types of algorithms to alleviate the possible drawbacks that each recommender may have independently (Burke 2002).

Regardless of the recommendation algorithm, normally all of them have to deal with a fundamental problem: sparsity, that is, the ratio between the actual number of interactions made by users on items in the system and the potential number of interactions considering those users and items. Generally, this sparsity is severe, being common to work with datasets with a sparsity higher than 97% (i.e., only 3% of the possible information is available to estimate the recommendations). At the same time, in classical recommendation (e.g., movies) researchers usually make use of the ratings that the users gave to the items explicitly (generally a score between 1 and 5). However, in other recommendation domains such as web, music, or Point-Of-Interest recommendation, there might not be ratings available, but rather the number of times a user has visited/consumed an item (as in the Foursquare dataset used in this paper).

2.2 Location-based recommender systems

While POI recommendation has the same goal as traditional RSs, there are aspects that make LBRSs different. First, the sparsity in these domains is considerable; for example, the densities, i.e., the inverse of sparsity, of the MovieLens20M and Netflix datasets are 0.539% and 1.177%, respectively. On the other hand, the Foursquare dataset we use in our experiments shows a density of around 0.0034%. Second, the use of one-sided or one-class information, where LBSNs normally only record positive values (check-ins) indicating that a user has visited a venue. Besides, users may check-in the same venue more than once, something that it is not considered in the traditional recommendation. And third, and more importantly, venue recommendation is highly affected by geographical, temporal (Sánchez and Bellogín 2022), and sometimes even social (user friends) (Gao et al. 2018) influences. The former is possibly the most critical aspect to consider in LBRSs, as it is usually assumed that users prefer to visit venues that are close to each other (Miller 2004). That is the reason why existing algorithms have incorporated geographical influence for generating recommendations (Liu et al. 2014; Ye et al. 2011; Lian et al. 2014).

Each model incorporates these influences differently, and although there are a large number of LBRSs (see Liu et al. (2017) for an experimental survey of the state-of-the-art models), many of them use traditional recommendation techniques. For example, Matrix Factorization (MF) approaches are used in the IRenMF model (Liu et al. 2014), which also takes into account the neighbor POIs of the target one by distance and uses a clustering algorithm to group all the POIs to model the geographical influence. Similarly, the GeoMF method (Lian et al. 2014), which uses two additional matrices, one to model the user activity areas by dividing the geographical space in a set of grids and the other to represent the influence of the POIs, and the LRT algorithm (Gao et al. 2013), which models the temporal component by factorizing the check-in matrix for every hour in a day. User-neighborhood approaches are also used in some LBRSs, like the USG model (Ye et al. 2011), which computes user similarities based on their check-in activities and combines them with the probability of visiting the target venue. LORE (Zhang et al. 2014) and iGLSR (Zhang and Chow 2013) are two other user-neighborhood approaches, which compute the similarities based on the distance of the users' residences, combined with the geographical influence modeled using Kernel Density Estimation (KDE).

2.3 Realistic evaluation in Recommender Systems

When evaluating recommendation quality in an offline setting, the RSs literature usually considers a random split with cross-validation methods to avoid the overfitting problem (Said et al. 2013). However, a RS should be evaluated as realistically as possible, not knowing anything about future interactions, to avoid obtaining unrealistic results and avoid data leakage (Kaufman et al. 2012).

Because of this, the community is slowly shifting the offline evaluation towards using temporal splits, where the recommendation algorithms should predict the present (or, actually, future) user interactions based on their past activity (Campos et al. 2014). However, different strategies may arise for performing such a temporal split. We can split by selecting a percentage of interactions to use in the training/test splits. A common approach would be to select the 80% of the oldest interactions to build the training set and the rest would form the test set. Other strategies would be to choose a timestamp, so as to use all interactions that happened after that timestamp for testing the recommenders. In alternative, one can order the interactions for each user separately and assign the most recent ratings of each user to the test set.

Each of these strategies has advantages and disadvantages in terms of the characteristics of the training/test splits derived and how close they represent real-world scenarios. Based on these descriptions, the most realistic protocols would be those that allow for a training set temporally separated from the test set, which can be achieved by either using a common splitting timestamp for the entire dataset or by selecting a percentage of the data according to the moment of interaction. This conclusion is in line with recent analyses made by the community regarding data leakage (Meng et al. 2020; Ji et al. 2021).

It is worth noting that, even if some of the existing POI recommenders perform a temporal split (Li et al. 2015; Zhang et al. 2014; Zhang and Chow 2015), to the best of our knowledge there is no thorough research about the effects of this type of evaluation split on typical recommendation approaches in this domain.

2.4 Impact of Recommender Systems

As described before, RSs analyze users' preferences in order to make personalized recommendations to users. However, it has been observed that sometimes the recommendations of the algorithms can be discriminatory for different groups (e.g., by ethnicity, age, occupation, or gender) (Edizel et al. 2019; Sánchez and Bellogín 2019; Weydemann et al. 2019). This effect can also cause certain types of users to receive the same type of items, isolating them according to these biases (the so-called *filter bubble* (Pariser 2011)). This was one of the main reasons to propose metrics in the field so that we could measure complementary dimensions beyond accuracy, such as novelty and diversity (Castells et al. 2015).

One of the most recognizable biases in RSs that has received much attention in recent years is the popularity bias, which shows how the recommendations produced are generally biased (or polarized) to the most popular items, affecting negatively the novelty and diversity of the suggestions. Some researchers have proposed different mechanisms to palliate this problem; for example, Abdollahpouri et al. (2017) presented a regularization framework to retrieve long-tail items with a small performance loss in ranking evaluation, whereas Abdollahpouri et al. (2019b) proposed re-ranking techniques to reduce the popularity bias in recommendations. Alternatively, Bellogín et al. (2017) defined two new split protocols to counter the effect of the popularity bias. Additionally, recent work has focused on the theoretical impact of popularity bias on the algorithms (Cañamares and Castells 2017, 2018). In any case, this is an issue that has been studied in different domains (Jannach et al. 2015; Boratto et al. 2019). Our goal is to go beyond the assessment/reinforcement of pre-existing polarized data recorded in a system or biases in algorithms, to study more broadly polarization in POI recommendation.

Another related topic associated with the societal impact of recommendations on the users is algorithmic fairness. A recent work by Weydemann et al. (2019) studied to what extent LBRSs can provide suggestions to groups characterized by sensitive features. More recently, Sánchez and Bellogín (2021) analyzed the recommendations of two different groups of users using LBRSs, i.e., locals and tourists, concluding that the latter suffers from a greater popularity bias. As we introduced in our motivation, polarized recommendations do not impact only consumers, but also providers (venue owners). In this work, we study a broader phenomenon, which complements and does not overlap with the studies on algorithmic fairness (indeed, we are neither considering demographic information of the users/providers, nor notions of similarity between them), by providing insights on the polarization generated by different algorithms. Hence, no direct comparison is possible and the connection between this study and algorithmic fairness is left as future work.

3 Polarization Characterization

Given the peculiarities of the POI recommendation problem with respect to the traditional recommendation, it is important to control which forms of polarization occur in this domain. In this section, we explain how to measure different forms of polarization: towards popular venues (Section 3.1) and categories (Section 3.2), regarding the venue exposure (Section 3.3), and with respect to the geographical distance (Section 3.4) between the user and the recommended venues. At the end of this section (Section 3.5), we also show several toy examples to better understand the proposed polarization metrics.

3.1 Measuring venue popularity polarization

From the multiple definitions that "novelty" has in the RSs and Information Retrieval areas, one of the most commonly used definitions is that something is novel when it is not popular (Gunawardana and Shani 2015). To measure novelty, Vargas and Castells (2011) defined the Expected Popularity Complement (EPC) metric, by computing the number of users who rated that item, divided by the number of users in the system; then, they proposed to subtract that value to 1, so that values closer to 1 indicate that the items are more novel (less known by the users in the systems). A similar metric called Inverse User Frequency (IUF), defined in Castells et al. (2015) measures novelty in a similar way, but considering the logarithm between the user that rated that item and the total users in the system. However, these metrics are too sensitive to the actual number of ratings, or interactions, in general, received by each item. For instance, if an algorithm always returns the same top-n items but the item distribution is too skewed, we may obtain similar novelty values between that algorithm and another one that recommends more items which have been rated by a similar number of users.

Because of this, in this work, we analyze the polarization towards popular venues by analyzing the popularity distribution derived from each recommendation algorithm. In this way, we can compare whether some algorithms are more or less tailored to return more popular items. Moreover, we propose a metric that summarizes such distribution in an empirical value for each algorithm; however, since we cannot assume the inherent distribution of the data, there is no general skewness function to measure it (such as kurtosis, which assumes data is normal); because of that, we resort to empirical metrics aware of the domain we analyze.

Definition 1 (Venue Popularity Polarization) The polarization of a recommendation model *rec* towards popular venues is the probability that a more popular venue is ranked higher than a less popular one, when considering the top-n items recommended to a user.

Our proposed metric to characterize the polarization of a model towards popular venues is computed by measuring the area under the curve generated by the cumulative distribution of the recommended items by rec; this is done by approximating the analytical integral by the trapezoidal rule. More specifically, given the unique set of items $R(rec, n)$ returned by recommender rec up to cutoff n, i.e., the length of the recommendation list, for all users, we propose the following formulation to measure the venue popularity polarization:

$$\text{PopI@}n(rec) = \frac{1}{2|m|} \sum_{k=2}^{|m|} \left(F_{\text{pop}}^{R(rec,n)}(x_{k-1}) + F_{\text{pop}}^{R(rec,n)}(x_k) \right) \qquad (1)$$

where $|m|$ are the items in the training set, ordering them by the number of times they have been recommended by the recommender rec. $F_{\text{pop}}^{R}(x)$ measures the cumulative popularity distribution[2] for an item x, depending on whether it belongs to R, in such a way that it is updated only for those items contributed by the corresponding recommender used to create such list R. Finally, to measure the popularity of a venue, we count the number of users who visited it divided by the total number of users that visited all recommended venues. By definition, the larger the area, the more uniform (less skewed) the distribution is. Hence, this metric produces lower values for recommenders polarized towards popular items, which is bounded in $[0, 1]$ thanks to the trapezoidal rule applied on a square $[0, 1] \times [0, 1]$. It is important to note that obtaining a high value in this metric does not imply that the ranking accuracy of the recommendations is higher, it implies that more items with different popularity values are being recommended. Therefore, in order to obtain the "expected" value of this metric, we should compute it with the data available in the test set as it represents the real visiting patterns of the users in the dataset. We shall do this later in the experiments by contrasting the behavior of recommenders against a method that provides suggestions based on the test set.

3.2 Measuring category popularity polarization

Intuitively, a user who likes rock music would probably prefer recommendations of groups such as Led Zeppelin or the Rolling Stones rather than classical music. In the case of POI recommendation, users may prefer some venues over others depending on the venue type. In this domain, the venue type is unambiguously linked to the venue category, such as restaurant, museum, public park, etc. Because of this, it is important to consider the polarization with respect to well-known groups of items, such as genres in movies or music, venue categories in POIs, or verticals in e-commerce.

[2] More specifically, since the distribution is discrete, we compute a cumulative histogram.

Moreover, the interactions between users and these groups of items are not uniformly distributed in typical recommendation systems, and in particular in LBSNs, as we show later for different cities with respect to venue categories. Hence, it is important to distinguish the popularity of a specific POI from that of the associated categories (e.g., a particular museum may be the most popular venue in a city, but museums may be the least represented category in that city).

Definition 2 (Category Popularity Polarization) The polarization of a recommendation model *rec* towards popular categories is the likelihood of recommending venues belonging to categories associated with the highest number of user interactions.

We analyze the polarization towards popular categories by grouping the top-n recommended POIs by each category, while sorting the different categories by increasing popularity, measured as the number of interactions each category has received in the entire dataset.

Thus, we summarize this analysis in the following metric value:

$$\text{PopC@}n(L) = \frac{1}{\min(n, |L|)} \sum_{i \in L} \frac{\text{bin}\big(\text{cat}(i)\big)}{|\{\text{cat}(\cdot)\}|} \tag{2}$$

$$\text{PopC@}n(rec) = \frac{1}{|U|} \sum_{u \in U} \text{PopC@}n(R(rec, u, n)) \tag{3}$$

where, as before, n denotes the cutoff at which we measure the metric (i.e., the number of items to consider from the recommendation list), $R(rec, u, n)$ denotes the top-n recommended items to user u by recommender rec. Note that $|L|$ and n will be, in general, equal, since $L = R(rec, u, n)$, except when the recommender has a low recommendation coverage (i.e., number of items that the recommender is returning to the target user). In case of low coverage, $|L|$ might be smaller than n, that is why we prefer to make this situation explicit in the formulation. Here, $\text{cat}(i)$ returns the associated category to each item, and $\text{bin}(\cdot)$ returns the category bin, where the least popular category is associated to the first bin (i.e., $\text{bin}\big(cat(i)\big) = 1$) and popular categories are assigned the last bins. In this way, a larger value is obtained for popular categories and we can use this as an indicator of how polarized an algorithm towards popular categories is. We consider the number of categories, $|\{\text{cat}(\cdot)\}|$, to be fixed[3]. In those cases where the category information is not available, an implicit clustering of the venues might be used (for instance, those items whose name contains a special keyword might be classified into a pre-defined group, e.g., 'Museum' or 'Cafe').

The metric is in the [0,1] range, where 0/1 indicates that a model only recommended venues associated with the most unpopular/popular category.

[3] In Foursquare, as an example, the categories are organized using a 3-layer hierarchy tree structure. This number would then depend on the category level used; for instance, for the first level, the one covering the most generic types of POIs, there are 9 different categories.

As in the previous case, to obtain the expected value, we should compute this metric with the data available in the test set.

3.3 Measuring polarization in terms of item exposure

When measuring the quality of a recommender system, in most cases only the users' opinions are taken into consideration, either in terms of relevance or other dimensions such as novelty and diversity. However, the perspective of the items should be equally important because we may be over-representing the most popular items in the recommendations (Ariza et al. 2021). For several years, researchers in the recommender systems area have analyzed the effect of over-representing the most popular items, observing that the most unpopular items actually belong to the long-tail item distribution (Park and Tuzhilin 2008). Although a large number of users consume popular items, according to Anderson (2006), vendors should focus on such long-tail items as unpopular items are often more profitable. In the POI recommendation domain, the items are venues, ranging from major tourist sites to minor ones, e.g., food establishments, bars, or small businesses. By recommending less popular venues in the long-tail, we may introduce users to new places that they had not thought they might be interested in, and also make these less popular sites receive more visits, which means that they end up having more customers. As these venues are sometimes businesses that generate trade activity in the cities, a poor exposure of these venues might negatively affect the city's economy.

Definition 3 (Venue Exposure Polarization) The polarization of a recommendation model *rec* in terms of exposure is the likelihood of the model to suggest a venue proportionally to the number of times the users will consider that venue in the future.

While, in the characterization of item popularity, we assessed the probability of recommending a popular item, in order to measure the exposure of venues, we compare the number of times an item has been recommended (Recommender Exposure, RE) against its actual exposure (i.e., the number of times that venue should be recommended regarding a subjective policy) (Actual Exposure, AE). However, differently from the metrics proposed by Ariza et al. (2021), instead of dividing RE and AE, we will compute the squared difference since it is a more common mechanism to measure errors, as in Ekstrand et al. (2021b):

$$RE@n(i, rec) = \frac{1}{|U|} \sum_{u \in U} \frac{1/log_2(pos(i, R(rec, u, n)) + 1)}{\sum_{j \in R(rec,u,n)} 1/log_2(pos(j, R(rec, u, n)) + 1)} \tag{4}$$

$$AE(i; \pi) = p(i|\pi) \tag{5}$$

$$IE@n(rec; \pi) = \sum_{i \in I} (RE@n(i, rec) - AE(i; \pi))^2 \tag{6}$$

where $R(rec, u, n)$ denotes, as before, the top-n recommended list by rec for user u, whereas $pos(j, L)$ denotes the position of item j in a recommendation list L. Finally, π denotes the exposure of the item under the target policy (ideal exposure). In this paper, we will work with two different policies – see (Ekstrand et al. 2021b) for an overview of reasonable choices over these policies –, namely: **Parity**, where we assume that all items should be recommended equally, i.e., following a uniform distribution, and **Relevance**, in which we assume that each item should be recommended following the same distribution observed in the test set. According to this, IE would denote the final Item Exposure based on a target policy π, for a recommender rec measured at cutoff n. Hence, the lower the IE, the better (low polarization, similar to the expected exposure) the recommender is.

3.4 Measuring polarization towards geographical distance

According to the first law of geography, "Everything is related to everything else, but near things are more related than distant things" (Miller 2004), which is why many LBRSs model geographical influence. Because of this, exposing the polarization (or the lack of it) towards this aspect might be a critical signal of the type of venues provided by a recommendation algorithm. Despite the fact that geographic influence has been used extensively to make recommendations to users, we have not found many works that analyzes the geographical relationship between the actual recommended POIs (e.g., if they are close to each other or to the user midpoint). Hence, we consider this analysis an important contribution of the presented work.

Definition 4 (Geographical Distance Polarization) The polarization of a recommendation model rec towards geographical distance is the likelihood of the model to suggest a venue that is close to / far from the current position of the user.

As a first approximation, we propose two metrics that consider the distance of recommended POIs in their evaluation. The first one, DistT, shown in Equation (7), sums the distance of the recommended POIs as if the user accepted those recommendations and visited those venues in order[4]. With this metric, we account for the polarization towards longer or shorter recommended routes or *trajectories*, even though this metric could be applied to any type of recommender system, not only for those producing routes. The second metric, DistU, shown in Equation (8), computes the total distance between each recommended POI and the user historical midpoint, obtained by averaging the coordinates of every venue visited by the user in the training set. In this

[4] Existing literature shows that users pose higher trust in highly ranked results, and measures of exposure in a ranking introduce a discount for lower-ranked results (Singh and Joachims 2018). The assumption is that the lower an item is ranked, the lower is the likelihood that the user will choose it. So we can assume that the ranking generated by a recommender system is a proxy for the sequence of choices for the users.

way, we aim to capture how sensitive each recommendation algorithm is to the history of previous locations of the user. This concept connects to the recent literature on calibrated recommendations (Steck 2018), by studying how adherent the recommendations are to the previous behavior patterns of the users, which in our case are modeled by their locations. Note that Equation 8 cannot be used if the user has not checked-in in any venue in the training set. However, in a real environment where a tourist arrives at a city, instead of her midpoint we could make use of the coordinates of the venue she is staying at or the actual geographical position of the user.

$$\mathrm{DistT@}n(R_u) = \sum_{i=2}^{\min(n,|R_u|)} \mathrm{Hav}(R_{u,i-1}, R_{u,i}) \tag{7}$$

$$\mathrm{DistU@}n(R_u) = \sum_{i=1}^{\min(n,|R_u|)} \mathrm{Hav}(\mathbf{u}_m, R_{u,i}) \tag{8}$$

where $R_{u,i}$ is i^{th} item recommended to user u, $\mathbf{u}_m$ is user u historical midpoint, and Hav is the Haversine distance of the coordinates of two geographical points. In order to interpret the geographical polarization values found using these metrics, we need to compare those values obtained by the recommenders we are analyzing with respect to those found using the user's ground truth. Thus, obtaining high values in these metrics (which indicate that the recommended venues are either far away from each other or from the user's midpoint) might not be intrinsically bad if the users actually exhibit those mobility patterns in the ground truth. However, obtaining very different values from those exhibited by the users in the test set would be a sign that the recommenders are actually showing a geographical distance bias *far from the expected one.*

3.5 Toy examples

In this section, we show a toy example for every proposed metric to illustrate how they work, in order to help the reader to have a better understanding of all the different analyzed polarizations.

First, in Figure 1, we compare the performance of two different recommenders (rec_1 and rec_2) using our Venue Popularity Polarization metric (Equation 1), and we compare it against other novelty metrics like EPC and IUF (they were both defined at the beginning of Section 3.1). As we can observe in that figure, both recommenders would obtain the same values in terms of Expected Popularity Complement (EPC) or Inverse User Frequency (IUF) because they are recommending items that have been rated by the same number of users (i.e., their popularity is the same). However, the second recommender is able to recommend both the black and white items while the first one is not. Thus, the area under the curve of the second recommender would be higher, as it is recommending a higher number of items, showing less polarized results.

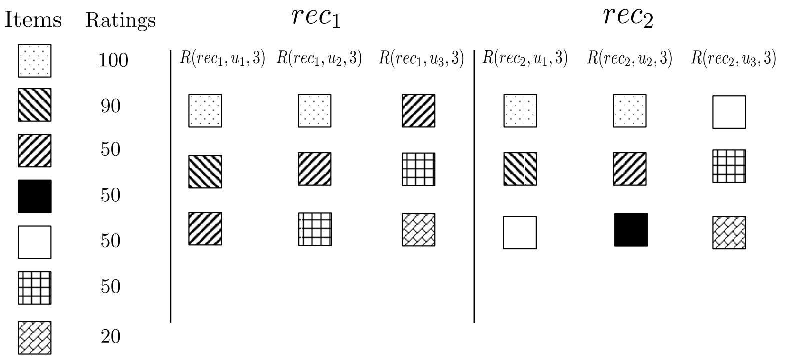

Fig. 1: Visual example of the popularity polarization of two different recommenders, rec_1 and rec_2. The second recommender would obtain higher values in our metric due to the fact that it is recommending more different venues, and hence the area under the curve would be higher than in the first recommender.

Secondly, in Figure 2, we show a comparison between two recommenders in terms of Venue Category Polarization. Both of them recommend three different items, but the first one is only recommending items with the feature denoted as "A", which is the most popular one. On the other hand, rec_2 is recommending venues belonging to all categories, and hence obtaining a lower category polarization.

Thirdly, in Figure 3, we show a comparison between two different recommenders using our formulation for Venue Exposure Polarization by applying a relevance-based target policy or ideal exposure. In that example, we observe that the second recommender obtains a lower result in terms of exposure than the first one due to several reasons. Firstly, rec_2 is not recommending one of the items (the one with the dotted pattern), which actually does not appear in any of the test sets. Secondly, this model is also recommending the black item twice, which is the same number of times that item appears in the test set; however, the first method only recommends this item once. Finally, rec_2 is the only model that recommends the item with vertical lines; moreover, this item appears in as many recommendation lists as in the test set. Hence, rec_2 achieves the expected exposure for this item, and the value of IE is decreased since both RE and AE are closer to each other.

Finally, in Figure 4, we show a comparison between two recommenders in terms of our Geographical Distance Polarization metrics (Equations 7 and 8). In this example, the second recommender would obtain lower values in the metrics as the recommended venues are closer with respect to the user mid-

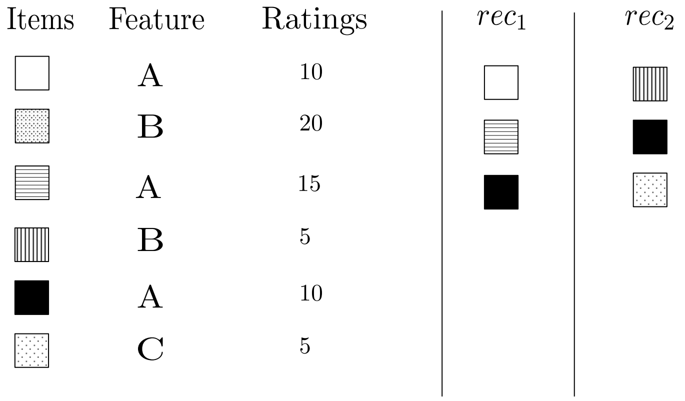

Fig. 2: Visual example of the category popularity polarization of two different recommenders, rec_1 and rec_2. The second recommender would be preferred as it is recommending venues from different categories.

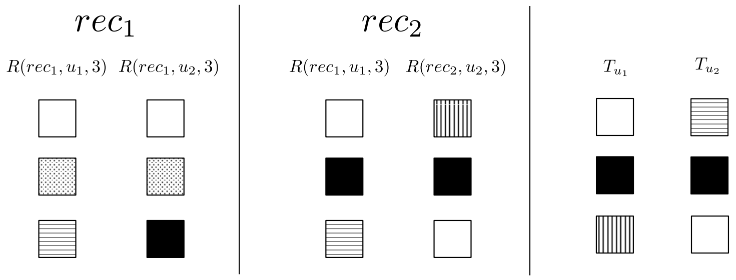

Fig. 3: Visual example of how the value of Item Exposure (IE) changes according to the behavior of recommenders, for a relevance-based policy. Here, rec_1 and rec_2 denote the first and second recommenders, $R(rec_n, u, 3)$ denotes the recommendations from recommender rec_n for user u, while T_{u_1} and T_{u_2} represent the test set of the two users. In this situation, rec_2 would obtain a lower value because it is recommending the black item 2 times, as in the test set, and it is not recommending the dotted item, which does not appear in the test set. Hence, as the recommended items from rec_2 are more similar to the ground truth of the user than the ones recommended by rec_1, the venue exposure polarization of rec_2 would be lower.

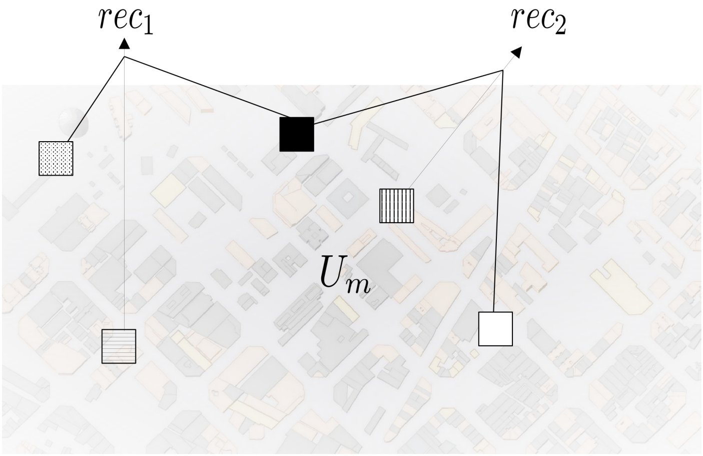

Fig. 4: Visual example of the geographical distance polarization of two different recommenders, rec_1 and rec_2. In this example, the second recommender will be preferred as the recommended venues are more geographically related between them and with respect to the user midpoint (represented by U_m).

point (U_m) and also closer between them than the recommendations produced by the first algorithm.

4 Evaluation Settings

4.1 Evaluation methodology

We performed experiments on the Foursquare global check-in dataset[5] used in (Yang et al. 2016). This dataset is formed by 33M check-ins in different cities around the world. We selected the check-ins from the cities of Tokyo, New York, and London from this dataset and, once we selected the check-ins of all three cities separately, we performed a 5-core, that is, we removed both users and POIs with less than 5 interactions. Next, aiming for a realistic evaluation, we split the check-ins so that the 80% of the oldest interactions were used to train the recommenders and the rest 20% to test them.

[5] https://sites.google.com/site/yangdingqi/home/foursquare-dataset

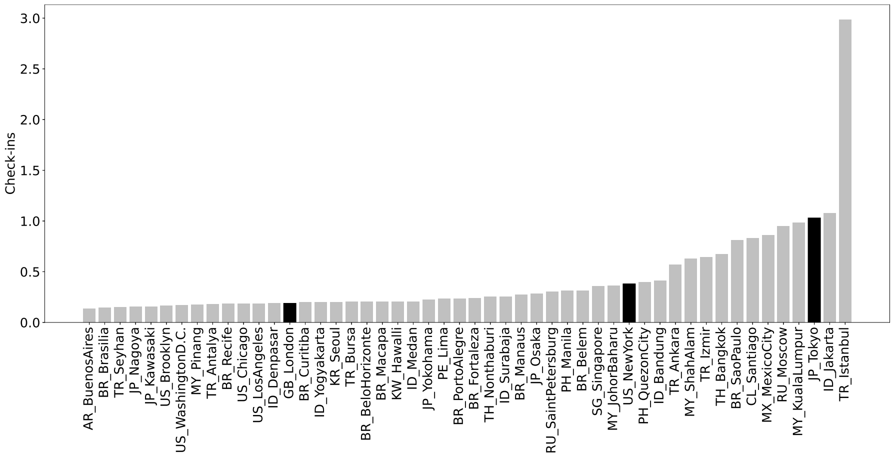

Fig. 5: Plot showing the 50 most popular cities (in terms of number of check-ins) in the Foursquare dataset before preprocessing. In black, the cities of Tokyo, New York, and London are highlighted.

The statistics of the datasets and their splits are shown in Table 1. Finally, we removed from the test set all interactions that appeared in the training set (as the purpose is to recommend new venues to the users) and the repetitions, that is, we consider that the users just visit the same POI once in the test set. These evaluation methodology issues, combined with the sparsity of the dataset and the fact that we do not force test users to have a minimum number of training interactions, means that the results in terms of ranking accuracy will be low. However, we decided to not focus only on those users with enough locations visited in their profile, as this would make our experimental analysis too limited. However, we leave as a future work the analysis of cold-start users (Lika et al. 2014).

Moreover, for the training set, we maintain three different versions due to the intrinsic characteristics of some of the aforementioned models: the one with repetitions (RTr), the one adding all interactions (FTr), and the one binarizing all possible user-POI interactions (BTr). Please note that in this dataset there are no explicit ratings as we typically find in classic recommendation datasets, such as MovieLens. In Foursquare, we only know when a user has visited a certain POI, unlike other LBSNs such as Yelp[6], where we do find ratings and reviews. Hence, the training set with repetitions (RTr) is being used by the recommenders that build sequences for performing the recommendations. The frequency training set (FTr) is being used by the recommenders that can exploit the explicit information, to give more importance to those interactions with a higher score. In this case, by aggregating the check-ins, we can obtain a frequency matrix that can be used in the models as if it was the classic matrix of user ratings. However, these frequencies are not entirely comparable

[6] `https://www.yelp.com/dataset`

Table 1: Statistics of the three cities used in the experiments. We show the number of users ($|\mathbf{U}|$), number of venues ($|\mathbf{V}|$), number of check-ins ($|\mathbf{C_r}|$), number of unique check-ins ($|\mathbf{C_{\overline{r}}}|$, without repeated interactions), and data density computed according to whether repetitions are considered or not. We present these values for the entire city, together with the corresponding training and test splits. In the case of Tokyo, New York, and London, there are 7,301, 4,188, and 1,958 users appearing in both training and test sets, respectively.

City	**Split**	$\lvert\mathbf{U}\rvert$	$\lvert\mathbf{V}\rvert$	$\lvert\mathbf{C_r}\rvert$	$\lvert\mathbf{C_{\overline{r}}}\rvert$	$\frac{\lvert\mathbf{C_r}\rvert}{\lvert\mathbf{U}\rvert\cdot\lvert\mathbf{V}\rvert}\%$	$\frac{\lvert\mathbf{C_{\overline{r}}}\rvert}{\lvert\mathbf{U}\rvert\cdot\lvert\mathbf{V}\rvert}\%$
Tokyo	Complete	10,057	24,892	921,874	381,165	0.3683	0.1523
	Training	9,735	24,614	737,499	317,213	0.3078	0.1324
	Test	7,623	18,901	184,375	97,554	0.1280	0.0677
New York	Complete	7,832	12,975	315,472	154,639	0.3104	0.1522
	Training	7,319	12,713	252,377	126,453	0.2712	0.1359
	Test	4,701	9,275	63,095	37,256	0.1447	0.0855
London	Complete	4,443	7,384	141,402	73,295	0.4310	0.2234
	Training	3,968	7,284	113,121	59,243	0.3914	0.2050
	Test	2,433	5,329	28,281	18,109	0.2181	0.1400

to ratings because they are not bounded at the system level (there may be users with a wide range of frequencies). Finally, the binarized training set is used by both the implicit and explicit recommenders. This final training set will denote with a '1' if a user has visited a particular POI (regardless of the number of times it has been visited) and will present a '0' otherwise. For generating the recommendations, we follow the TrainItems methodology (Said and Bellogín 2014), i.e., we consider as POI candidates for a target user u those venues that appear in the training set but that have not been visited by u.

4.2 Recommenders

In order to analyze and characterize the biases that may exist in the Foursquare dataset, we now describe the state-of-the-art algorithms that have been considered in our experiments, grouped in different families:

- Non-Personalized: we tested a Random (**Rnd**) and a Popularity (**Pop**) recommender. The latter recommends the venues that have been checked-in by the largest number of users.
- Collaborative-filtering: we used a User-Based (**UB**) (non-normalized k-nn algorithm that recommends to the target user venues that other similar users visited before) and an Item-Based (**IB**) (non-normalized k-nn that recommends to the target user venues similar to the ones that she visited previously) collaborative filtering algorithm. We also included a matrix factorization algorithm that uses Alternate Least Squares for optimization (**HKV**) from (Hu et al. 2008), and the Bayesian Personalized Ranking (a

pairwise personalized ranking loss optimization algorithm) using a matrix factorization approach (**BPR**) from (Rendle et al. 2009). For the BPR, we use the MyMediaLite library[7].
- Temporal/Sequential: we include a user-based neighborhood approach with a temporal decay function (**TD**) (that gives more weight to more recent interactions), and several algorithms based on Markov Chains: Factorized Markov Chain (**MC**), Factorized Personalized Markov Chains (**FPMC**) and Factorized Item Similarity Models with high-order Markov Chains (**Fossil**). All three Markov Chains approaches are obtained from (He and McAuley 2016).
- Purely geographical: we used the Kernel Density Estimation (**KDE**) from (Zhang et al. 2014), and a recommender that suggests to the user the closest venues to her centroid (**AvgDis**).
- Point-of-Interest: we used the fusion model proposed by Cheng et al. (2012) that combines the Multi-center Gaussian Model technique (MGM) with Probabilistic Matrix Factorization (PMF) (**FMFMGM**), a POI recommendation approach from (Yuan et al. 2016) that uses BPR to optimize the model (**GeoBPR**), a weighted POI matrix factorization algorithm (**IRenMF**) from (Liu et al. 2014), and a hybrid POI recommendation algorithm that combines the UB, Pop, and AvgDis recommenders (**PGN**).

We also include a perfect recommender that uses the test set as the ground truth, named Skyline. This recommender will return the test set for the user, in order to check the maximum values that we can obtain with ranking-based accuracy metrics (**Skyline**). At the same time, it helps to evidence the biases and polarizations that already exist in the test split.

4.3 Metrics

Since we have already defined in previous sections our proposed metrics to measure different types of polarization, we will now show the formulation of the metrics used for measuring the item accuracy, novelty, and diversity.

- Accuracy: oriented at measuring the number of relevant items recommended to the user (Gunawardana and Shani 2015). We will use Precision (P) and the normalized Discounted Cumulative Gain (nDCG):
  - Precision:
    $$\mathrm{P@n(u)} = \frac{\mathrm{Rel_u@n}}{\mathrm{k}} \tag{9}$$
    where $\mathrm{Rel_u@n}$ denotes the set of relevant items recommended at top n.
  - nDCG:
    $$\mathrm{nDCG@n(u)} = \frac{\mathrm{DCG@}n(u)}{\mathrm{IDCG}(u)\mathrm{@}n} \tag{10}$$

[7] MyMedialite: `http://www.mymedialite.net/`.

$$\mathrm{DCG@}n(u) = \sum_{k=1}^{n} \frac{2^{rel_k} - 1}{\log_2(k+1)} \tag{11}$$

where rel_k denotes the real relevance of item k in the test set. In a rating-based dataset, this real relevance would be the rating that the user gave to that item in the test set. In our case, as we only know whether (and when) a user has performed a check-in, we fix this ideal relevance to 1 as long as the venue appears in the test set of the user (every venue visited by the user in the test set is equally relevant).

Higher values of P and nDCG imply a better recommendation quality.

- Novelty: oriented at measuring the number of popular venues, since they are inversely related to novel venues (Vargas and Castells 2011). We use a simplified version of the Expected Popularity Complement (EPC) metric:
  - EPC:

$$EPC@n(u) = C \sum_{i=1}^{\min(n,|R_u|)} (1 - p(\text{seen} \mid R_{u,i})) \tag{12}$$

where C is a normalizing constant (generally $C = 1/\sum_{i=1}^{\min(n,|R_u|)}$). In our case, $p(seen|i_k) = \frac{|U_i|}{|U_{training}|}$, with U_i being the number of users that checked-in in venue i and $U_{training}$ the set of users in the training set. Higher EPC implies better recommendation novelty.

- Diversity: oriented at measuring how many different venues we are recommending to the user (Vargas and Castells 2011). We use the Gini Coefficient to measure the diversity.
  - Gini:

$$\text{Gini@n} = 1 - \frac{1}{|\mathcal{I}| - 1} \sum_{k=1}^{|\mathcal{I}|} (2k - |\mathcal{I}| - 1) p(i_k \mid s) \tag{13}$$

$$p(i \mid s) = \frac{|\{u \in \mathcal{U} | i \in R^s_{u,n}\}|}{\sum_{j \in \mathcal{I}} |\{u \in \mathcal{U} | j \in R^s_{u,n}\}|} \tag{14}$$

where $p(i_n \mid s)$ is the probability of the n-th least recommended item being drawn from the recommendation list generated by s, that is, when considering all rankings @n ($R^s_{u,n}$) for every user. In this paper, we will use the complementary of the Gini Index proposed in Castells et al. (2015), as defined in Vargas and Castells (2014). Higher Gini implies better recommendation diversity.

- User Coverage: aims to measure whether the recommender system covers all the users or items in the catalog (Gunawardana and Shani 2015). We focus on the User Coverage (UC), that accounts for the number of users to whom at least one recommendation is made. This metric is useful because there might be some models that are not be able to recommend to all users of the test set (e.g., users with very few interactions who are difficult to model properly). Higher user coverage means that our model is able to recommend to more users.

Table 2: Parameters of evaluated recommenders; the values that are not between the symbols {} are considered fixed and not tuned.

Family	**Rec**	**Parameters**
Non-Personalized	Rnd	None
	Pop	None
Collaborative-filtering	UB	$k = \{20, 40, 60, 80, 100, 120\}$, sim = {Jac, Cos }, Tr = {FTr, BTr}
	IB	$k = \{20, 40, 60, 80, 100, 120\}$, sim = {Jac, Cos }, Tr = {FTr, BTr}
	HKV	$k = \{10, 50, 100\}$, $\alpha = \{0.1, 1, 10\}$, $\lambda = \{0.1, 1, 10\}$, Tr = {FTr, BTr}
	BPR	$k = \{10, 50, 100\}$, $\lambda_u = \lambda_i = \{0.001, 0.0025, 0.005, 0.01, 0.1\}$, $\lambda_0 = \{0, 0.5, 1\}$, $\lambda_j = \lambda_u/10$, iter = 50, learnR = 0.05
Temporal/Sequential	TD	$k = \{20, 40, 60, 80, 100, 120\}$, sim = {Jac, Cos }, $\lambda = \{0.1, 0.05\}$, Tr = {FTr, RTr}
	MC	$k = \{2, 5, 10, 20\}$, $\lambda = \{0.1, 0.2\}$, Tr = {FTr, RTr}
	FPMC	$k = \{2, 5, 10, 20\}$, $\lambda = \{0.1, 0.2\}$, Tr = {FTr, RTr}
	Fossil	$k = \{2, 5, 10, 20\}$, $\lambda = \{0.1, 0.2\}$, $L = \{1, 2, 3\}$, Tr = {FTr, RTr}
Geographical	KDE	Tr = {FTr, RTr}
	AvgDis	ScoreFreq = {True, False}
POI	FMFMGM	Factors = {50, 100}, $\alpha = \{0.2, 0.4\}$, $\alpha_2 = \{20, 40\}$, $\theta = \{0.02, 0.1\}$, maxDist = 15,, Tr = {FTr, BTr} iter = 30, $\beta = 0.2$, sigmoid = false, learnR = 0.0001
	GeoBPR	Factors = {10, 50, 100}, $\lambda_u = \lambda_i = \{0.001, 0.0025, 0.005, 0.01, 0.1\}$, $\lambda_0 = \{0, 0.5, 1\}$, iter = 50, learnR = 0.05, MaxDist = 1, 4
	IRenMF	$k = \{50, 100\}$, $\alpha = \{0.4, 0.6\}$, $\lambda_1 = \lambda_2 = 0.015$, $\lambda_3 = \{0.1, 1\}$, #Clust = {5, 50}, GeoNN = 10, Factors = 100, $\alpha = 10$, Tr = {FTr, BTr}
	PGN	$k = \{40, 60, 80, 100, 120\}$, sim = {Jac, Cos }, Tr = {FTr, BTr}
Skyline	Skyline	None

Table 3: Performance results on Tokyo city. All metrics computed at cutoff 5. In bold the best values are shown without considering the Skyline recommender. In italics the best value for each family is shown.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Rnd	0.000	0.000	**0.999**	**0.551**	**0.303**	*0.760*	**0.000**	**0.001**	37.2	34.7	**7,253**
Pop	**0.071**	*0.087*	0.746	0.000	0.000	0.960	0.131	0.121	*24.9*	*26.4*	**7,253**
UB	*0.070*	*0.087*	0.769	0.001	0.002	0.968	0.103	0.093	26.0	25.8	*6,931*
IB	0.063	0.080	0.819	*0.025*	*0.026*	*0.911*	0.064	0.057	23.2	25.0	*6,931*
HKV	0.064	0.078	*0.845*	0.002	0.003	0.921	*0.038*	*0.031*	*22.0*	*21.7*	*6,931*
BPR	0.066	0.081	0.754	0.000	0.003	0.955	0.123	0.112	25.6	27.7	*6,931*
TD	**0.071**	**0.088**	0.776	0.001	0.003	0.965	0.097	0.087	25.9	25.4	*6,931*
MC	0.051	0.062	0.804	0.001	0.003	*0.939*	0.107	0.098	26.5	30.9	6,879
FPMC	0.053	0.064	0.807	0.001	0.001	0.943	0.103	0.096	31.0	30.1	6,884
Fossil	0.058	0.074	*0.851*	*0.003*	*0.006*	0.878	*0.046*	*0.040*	*22.0*	*21.7*	6,879
KDE	*0.004*	*0.005*	**0.999**	*0.318*	*0.212*	0.753	**0.000**	**0.001**	**0.4**	15.5	6,879
AvgDis	0.001	0.001	**0.999**	0.202	0.187	**0.719**	**0.000**	**0.001**	0.6	**4.2**	*6,931*
FMFMGM	0.063	0.079	0.772	0.001	0.002	0.979	0.105	0.095	23.7	22.7	6,931
GeoBPR	0.065	0.081	0.756	0.000	0.001	0.957	0.120	0.110	23.7	24.2	6,931
IRenMF	*0.069*	0.083	*0.799*	0.003	0.008	0.951	*0.072*	*0.063*	23.9	23.8	6,931
PGN	0.068	*0.086*	0.777	*0.014*	*0.023*	*0.932*	0.110	0.100	*23.6*	*20.9*	**7,253**
Skyline	0.784	0.996	0.982	0.231	0.087	0.796	0.000	0.000	17.5	18.8	7,241

5 Polarization Assessment

Tables 3, 4, and 5 show the results of the aforementioned recommenders in terms of accuracy (P and nDCG), novelty (EPC), diversity (Gini), and our metrics to measure popularity polarization (PopI, for item popularity and PopC for category popularity), item exposure (ExpP using Parity and ExpR using Relevance as target policies), polarization towards geographical distance (DistT and DistU), and user coverage (UC). Recall that higher values indicate *better* accuracy, novelty, diversity, and coverage. On the contrary lower values of popularity polarization (except PopI that, in this case, higher values means a higher area under the curve and hence lower popularity bias), exposure, and distance measure the optimal situation with less polarization. The parameters tested of the recommenders can be found in Table 2. We selected the best

Table 4: Performance results on New York city. Same notation as in Table 3.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Rnd	0.000	0.001	**0.999**	**0.579**	**0.347**	*0.702*	**0.000**	**0.001**	35.6	*30.9*	**4,416**
Pop	**0.069**	**0.099**	0.837	0.000	0.000	0.758	0.173	0.154	*26.2*	31.2	**4,416**
UB	0.056	0.087	0.868	0.002	0.005	0.734	0.102	0.087	36.3	34.5	*3,903*
IB	0.025	0.036	*0.967*	*0.175*	*0.118*	0.693	*0.006*	*0.004*	*20.2*	*25.1*	*3,903*
HKV	0.054	0.079	0.907	0.003	0.004	*0.681*	0.044	0.034	31.3	29.9	*3,903*
BPR	*0.060*	*0.092*	0.841	0.000	0.000	0.755	0.167	0.148	29.3	36.8	*3,903*
TD	*0.055*	*0.087*	0.884	*0.005*	0.012	0.736	*0.085*	*0.071*	36.4	35.3	*3,903*
MC	0.048	0.071	0.862	*0.005*	*0.013*	0.773	0.132	0.116	35.7	36.0	3,819
FPMC	0.039	0.057	*0.889*	0.001	0.002	*0.685*	0.116	0.104	*33.1*	33.7	3,819
Fossil	0.053	0.075	0.875	0.001	0.003	0.725	0.099	0.085	35.5	*31.6*	3,819
KDE	*0.005*	*0.006*	0.998	*0.332*	0.219	0.708	**0.000**	**0.001**	**0.5**	13.7	3,821
AvgDis	0.002	0.001	**0.999**	0.247	**0.221**	**0.670**	**0.000**	0.002	0.8	**4.2**	*3,903*
FMFMGM	0.029	0.042	*0.894*	0.001	0.003	*0.716*	*0.127*	0.117	*11.7*	*17.9*	3,903
GeoBPR	0.055	0.068	0.850	0.000	0.001	0.738	0.169	0.154	30.9	30.7	3,903
IRenMF	0.057	0.087	0.854	0.002	0.006	0.743	0.130	*0.113*	32.4	34.1	3,903
PGN	*0.065*	*0.097*	0.859	*0.019*	*0.034*	0.769	0.141	0.124	30.8	26.8	**4,416**
Skyline	0.726	0.991	0.981	0.230	0.114	0.762	0.001	0.000	13.3	14.1	4,390

Table 5: Performance results on London city. Same notation as in Table 3.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Rnd	0.001	0.001	**0.998**	**0.549**	**0.328**	*0.694*	**0.000**	**0.001**	35.8	27.6	**2,301**
Pop	*0.039*	*0.047*	0.898	0.001	0.001	0.878	0.195	0.185	*15.3*	*15.3*	**2,301**
UB	*0.037*	0.048	0.923	0.004	0.007	0.781	0.045	0.036	*20.7*	*22.4*	1,824
IB	0.021	0.028	*0.980*	*0.234*	*0.131*	*0.693*	*0.003*	*0.002*	22.6	26.9	1,824
HKV	0.040	*0.052*	0.933	0.005	0.006	0.781	0.029	0.022	25.5	27.6	*1,826*
BPR	0.034	0.044	0.942	0.012	0.018	0.731	0.045	0.038	24.2	24.5	*1,826*
TD	*0.037*	0.049	0.927	0.007	0.012	0.813	0.044	0.036	24.9	24.6	*1,824*
MC	0.028	0.038	0.920	0.001	0.002	*0.740*	0.136	0.128	31.2	31.2	1,778
FPMC	0.031	0.043	*0.955*	*0.031*	*0.030*	*0.740*	*0.014*	*0.009*	*19.6*	*22.5*	1,785
Fossil	0.039	*0.054*	0.928	0.003	0.006	0.789	0.069	0.006	24.3	25.1	1,778
KDE	*0.007*	*0.009*	**0.998**	*0.372*	0.220	0.691	**0.000**	**0.001**	0.9	13.7	1,779
AvgDis	0.003	0.004	**0.998**	0.319	*0.222*	**0.669**	**0.000**	**0.001**	**0.7**	**0.9**	*1,826*
FMFMGM	0.029	0.039	0.921	0.003	0.006	0.860	0.091	0.083	*11.7*	15.7	1,826
GeoBPR	0.041	**0.054**	0.915	0.002	0.004	*0.780*	0.079	0.070	13.9	17.5	*1,826*
IRenMF	0.036	0.047	*0.932*	*0.021*	*0.024*	0.812	*0.044*	*0.037*	22.2	23.0	*1,826*
PGN	**0.042**	0.052	0.912	0.008	0.015	0.856	0.118	0.109	16.4	*14.4*	**2,301**
Skyline	0.729	0.996	0.986	0.254	0.118	0.742	0.001	0.000	13.7	12.7	2,292

configuration of each recommender according to nDCG@5 obtained in the test set[8].

In order to validate the different forms of polarization we presented, we performed three sets of experiments:

1. **Impact on accuracy metrics.** Before assessing polarization, we evaluate the models shown in Section 4.2, considering the metrics presented in Section 4.3. This will allow us to assess the behavior of these models from accuracy and beyond-accuracy perspectives, to then contextualize it to the polarization these models generate.
2. **Measuring recommendation polarization.** We address to what extent the considered recommendation models are polarized towards the four perspectives considered in this work (i.e., venue and category popularity, venue

[8] It should be noted that other cutoffs do not produce significantly different results. This has been tested in additional experiments not reported here, in agreement with other authors reporting strong correlations between metric values at low and high cutoffs (Valcarce et al. 2018).

exposure, and geographic distance), by measuring the metrics proposed in Section 3.

3. **Polarization mitigation.** We evaluate the capability of hybrid and re-ranking mitigation strategies to counter polarization.

Since no validation set is used in these experiments, the reported performance is an overestimation. Such an experimental setting is not uncommon in recommender systems, especially when dealing with temporal splits as we have here (Sun 2022).

In what follows, we analyze these perspectives in depth.

5.1 Impact on accuracy metrics

The analysis of these results highlighted some interesting behaviors. First, we observe in Tables 3, 4, and 5 that the Skyline does not have full coverage for the users and it is not obtaining a value of 1 in the accuracy metrics. This is because we follow the TrainItems methodology (see Section 4.1) and therefore the items that did not appear previously in the training set cannot be recommended. Besides, there might be some users that have a smaller number of relevant items than the used cutoff. These two reasons could prevent some metrics from obtaining a perfect score.

Regarding the rest of the algorithms, we observe that one of the best performing recommender (in terms of accuracy, if we ignore the Skyline) is the Pop recommender in all cities, even though in Tokyo the TD model and in London the GeoBPR and PGN models obtain a slightly better value than Pop. This could be due to several causes, including (i) the high sparsity found in the datasets, (ii) the test set that only contains new interactions (and hence popular venues are *safe* recommendations), and (iii) the temporal evaluation methodology, as there could be users in the test set that do not appear in the training subset (for whose, again, popular venues can be very useful recommendations). This is an interesting conclusion, because it is a clear sign that this algorithm, despite its simplicity, is able to beat more complex models that incorporate temporal and/or geographical influences. However, this is somewhat surprising, because despite being such a competitive baseline, it is not so common to analyze the performance of this baseline in POI recommendation (Sánchez and Bellogín 2022). Indeed, the authors of IRenMF and the FMFMGM did not test their approaches against the Pop recommender.

With respect to the POI algorithms, we observe that, in terms of accuracy, their performance is very similar to other classical approaches, like the UB or the BPR. This may be due to the high number of both hyper-parameters and parameters that these models have, making it sometimes difficult to find a good configuration of hyper-parameters that obtains a decent performance. In fact, it is interesting to highlight the low values achieved by the FMFMGM algorithm in New York and London, while in Tokyo it is competitive against other models. This demonstrates that although we might find good configurations in terms of accuracy, the parameter settings in some circumstances is critical. In the end, classical proposals such as those based on neighbors, might be easier to explain and optimize due to its simplicity and lower number of parameters (Ning et al. 2015). This also affects the PGN recommender since, despite its simplicity, its performance is rather high. In New York and London it is the best recommender of the POI family and in Tokyo it has a very similar performance to IRenMF. The low number of parameters of this recommender, combined with the fact that it merges different sources of information such as popularity and geographical influence, may be the reason for this behavior.

5.2 Measuring recommendation polarization

When measuring the distance (DistT and DistU), we observe that both Rnd and Pop algorithms obtain high values, showing us that the recommended venues of these models are far from each other. Analyzing this geographical information is also important because, as we observe in the Skyline, users tend to visit POIs that are relatively close to each other, meaning that the distance between the relevant items, and also between the recommended items and the user's center, should be low. Nevertheless, the geographical influence alone is not enough to obtain high values in terms of relevance, as evidenced by the poor performance of the pure geographical algorithms (AvgDis, KDE).

At the same time, if we analyze the rest of the recommenders, we observe that, although all of them seem to perform personalized recommendations, regarding PopI, PopC, ExpP, and ExpR metrics we observe a pronounced popularity bias. Let us focus, for example, in the PopI and exposure metrics. The only recommenders with decent values of accuracy that seem to obtain high values on these metrics are PGN and IB, while the rest only obtain results slightly higher than Pop. In fact, when analyzing the exposure metrics (ExpP, ExpR), the random recommender obtains lower values in terms of ExpP than all algorithms (except the Skyline) due to the fact that it recommends items in an arbitrary manner, without overrepresenting any subset of items. Similarly, this recommender obtains good results in the ExpR metric because it is recommending almost all the venues in the system, so it is very likely that within those recommendations there are relevant venues. However, what the Rnd recommender fails is in recommending the relevant venues to the correct users, as discussed before regarding the accuracy metrics.

Hence, we conclude that most of the recommenders suffer from a great popularity bias, evidencing the difficulty of finding good representatives for all metrics. Therefore, among all the experimented recommenders, we consider IB and PGN to be of particular interest, since even though they do not perform as well in terms of accuracy as Pop, they obtain competitive results in terms of other metrics like novelty, diversity, and item exposure; this is a direct consequence of suffering less from the popularity bias. Let us now analyze the effect of the popularity and the categorical polarization more in detail.

Figure 6 shows the cumulative plot of the cities of Tokyo (top), New York (bottom-left), and London (bottom-right) of the most representative recom-

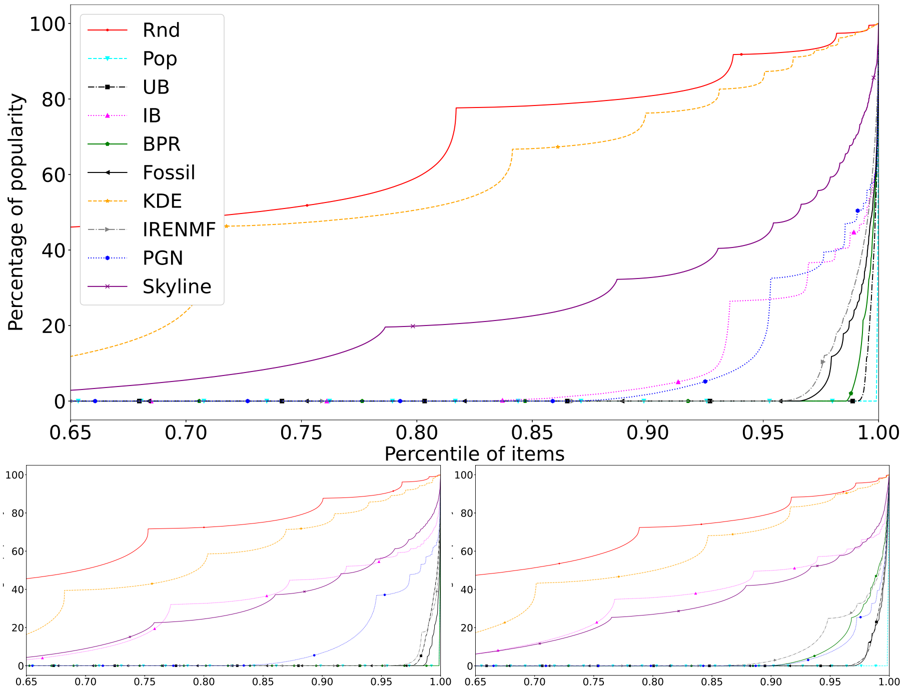

Fig. 6: Popularity cumulative plots from the cities of Tokyo (top), New York (bottom-left), and London (bottom-right) of the recommenders, considering the top-5 items returned by each of them. Showing the 30% most popular POIs.

menders shown in Tables 3, 4, and 5, showing the 30% of the most popular venues. For this selection, we considered those models with better values in any evaluation dimension that belong to different families. By considering those results, we observe that some of the most competitive recommenders like UB, TD, and BPR are just basically returning the most popular POIs (something that we observed in the previous tables thanks to our proposed metrics PopI and PopC). At the same time, those recommenders that are able to obtain a higher area under the curve than the one obtained by the Skyline are the worst in terms of performance (i.e., Rnd and KDE). This is a worrying result that departs from the results previously reported for some recommenders in terms of classical accuracy metrics, which slightly differed from the Pop algorithm. However, when the recommended items are analyzed, a clear, strong popularity bias is observed. In order to better visualize this effect, in Figure 9, in the left column, we show the distribution of the top 30% most popular venues in the three different cities. As we can observe, despite showing only 30% of the most popular venues, most of the check-ins are concentrated in the most popular ones, leaving a large number of other venues in the long-tail unexplored. If, for example, we analyze the same distributions at the user level (distribu-

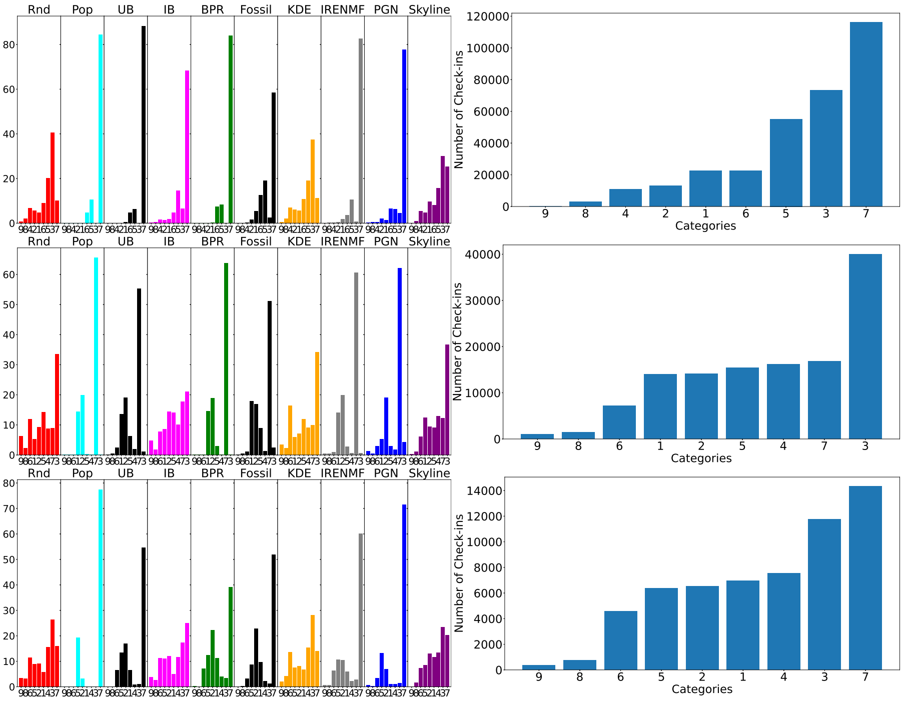

Fig. 7: In the left column, we represent distribution of the categories that appear in the top-5 recommended items for each algorithm in the training set for Tokyo (first row), New York (second row), and London (last row). In the right column, we show the distribution of the categories of the venues in the cities following the same order. The category bins in the latter case are ordered by increasing category popularity.

tion of the check-ins performed by the users, shown in the right column in Figure 9), we can observe how the distribution is not so unbalanced, although we can find that there are a considerable number of users who have made very few check-ins. Nevertheless, we believe there is potential in combining different types of algorithms (those more biased towards popularity and those less so) to see if it is possible to maintain an adequate level of accuracy while increasing at the same time the performance of other metrics such as novelty, diversity, or item exposure.

Now, let us move to the analysis of category popularity polarization. In the Foursquare dataset, we have 9 categories of level 1[9]: Arts & Entertainment (1), Outdoors & Recreation (2), Food (3), Nightlife Spots (4), Shops & Services (5), Professional & Other Places (6), Travel & Transport (7), Colleges & Univer-

[9] There are at least two other levels in Foursquare, each level is more specific than the previous one: level 2 includes 48 categories, whereas level 3 contains 337. An example of the relation between the three levels is a *soccer stadium* (level 3), which would be categorized as *stadium* (level 2) or as *arts & entertainment* (level 1).

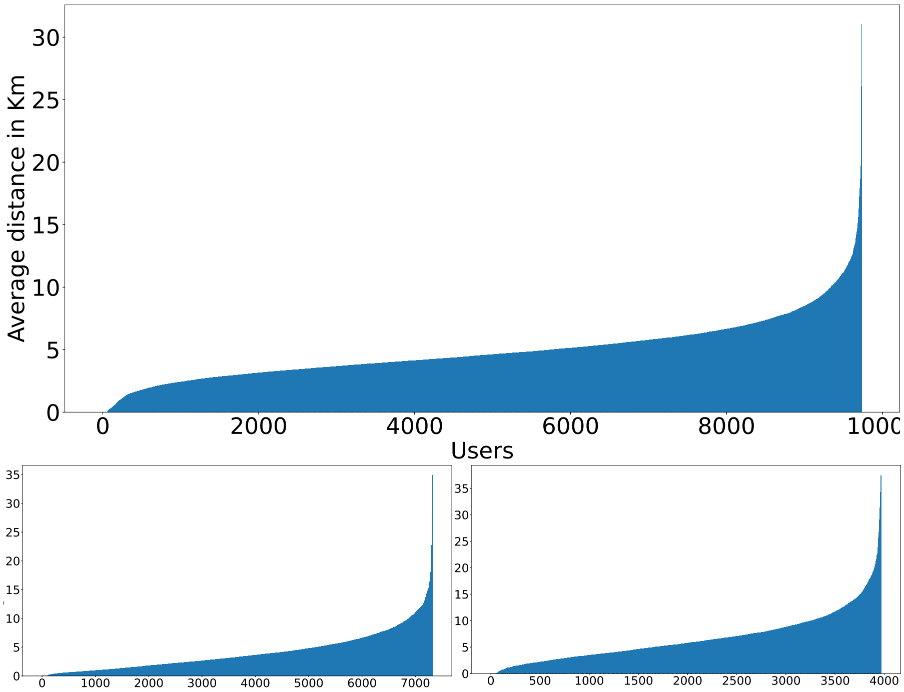

Fig. 8: Distance distribution of the users in the training sets of the cities of Tokyo (top), New York (bottom-left), and London (bottom-right).

sities (8), and Residences (9); due to space restrictions, they will be presented using their numerical IDs. We first show in the right column in Figure 7 the distribution of the venue categories in the training set of the three cities. With this image, we want to show that the categories are not distributed uniformly and that venues related to both transport (airports, train stations, subways, etc.) and food (restaurants) are the most numerous in these cities, while the number of check-ins in residences is negligible. Taking this into account, we show in the right column of Figure 7, the distribution of the categories of the recommended venues by our models using a cutoff of 5, that is, only the top-5 items recommended by each of those models are considered when measuring PopC. In these figures, we observe that the popularity of a category is not always associated with the number of POIs that share that category; more specifically, category 7 (Travel & Transport) concentrates the largest number of check-ins in the city of Tokyo, while category 3 (Food) is the second most popular category; however, since this category covers a large number of different venues, those recommenders with a strong item popularity bias (such as Pop) recommend almost no POIs from this category, since its corresponding items are not globally popular. A similar behavior is observed in New York, where category 3 is the most popular one in the number of check-ins but most personalized recommenders do not suggest as many items belonging to that category as those from categories 7 or 1.

Interestingly, the analysis of the category bias allows discriminating between those recommendation methods that seem to have the same popularity bias, according to Figure 6. For instance, it is now more clear that Pop and BPR are recommending practically the same items. At the same time, IRenMF and UB also include some of the least popular categories, evidencing different patterns on the recommendations that, as we will discuss later, prompts different effects on the accuracy of these algorithms. Finally, those techniques with a less pronounced category bias exploit very different sources of information: Skyline uses the test directly, KDE exploits the geographical coordinates, IB computes collaborative similarities between items (probably favoring the less interacted items, as discussed previously), and Rnd. This is an indication that the mitigation of these types of biases requires additional information sources. These additional sources should, in any case, be balanced with relevant recommendations, since the risk of providing not interesting items is higher for less popular categories; for instance, Skyline and Rnd show similar plots but have very different accuracy levels.

5.3 Polarization mitigation

As we observed in the previously reported results, it is impossible for one algorithm to obtain the best performance in all reported metrics. In fact, the Skyline, which would represent the best recommender in terms of accuracy, performs worse than the Rnd recommender in terms of novelty and diversity. For that reason, and considering accuracy as one of the most critical dimensions to optimize, we aim to combine several algorithms to create models that obtain decent levels of accuracy while overcoming the analyzed polarization measurements: popularity, exposure, and geographical distance. In order to do so, we propose two different but complementary approaches to mitigate the aforementioned biases.

As a first approach, we create hybrid recommenders by combining several models (Burke 2002); we apply simple models based on weighting differently each of the combined recommendation algorithms. By means of these weights, we will be able to enhance the quality of the recommendations by balancing the contribution of the different models depending on the evaluation dimension that we are interested in maximizing in that particular moment, either ranking accuracy, novelty, or diversity. In our second approach, we make use of reranking techniques popular in the Information Retrieval and Recommender Systems fields to address the tradeoff between accuracy and diversity (Santos et al. 2010). In our context, we use these techniques in order to rearrange the top-n recommended items by an algorithm according to another recommendation technique. The objective of both proposals is to generate new recommendation lists that are capable of maintaining acceptable levels of accuracy, while improving performance in other dimensions, such as novelty (since it is the opposite to popularity), diversity, or geographical variability, thus mitigating some of the desired biases.

It should be noted, however, that all these measurements (and whether an improvement was found) depend on having a test set as reference. Such a set may contain biases itself, hence limiting the generalization and impact of the proposed techniques. Collecting and using unbiased datasets is out of the scope of this paper, but it is a direction worth exploring in the future.

To define our hybrid approaches, we assume we have collected the top-n lists of a set of recommenders, denoted as $\mathcal{R}$, and a weight vector W, so that $R^j \in \mathcal{R}$ denotes the recommendations for all the users of the j-th recommender, and $w^j \in W$ denotes the weight for that recommender. As every recommender may have a different range (for the scores generated for every recommended item), we first combine all the recommendation lists using the min-max normalization. The final score user u has for item i is computed as:

$$s(i,u;\mathcal{R},W) = \sum_{j=1}^{|\mathcal{R}|} w^j \frac{s(i,R_u^j) - \min{(R_u^j)}}{\max{(R_u^j)} - \min{(R_u^j)}} \tag{15}$$

where $s(i,L)$ provides the score of item i within the recommendation list L, whereas $\min{(\cdot)}$ and $\max{(\cdot)}$ denote the minimum and maximum score of the list for user u by recommender R^j. Moreover, instead of using all the recommended items from each method (which might be computationally expensive) in our hybrid formulation, we decided to use the top-100 items of each recommender being considered. This top-100 selection is only used for generating recommendations, i.e., it is independent of the cutoff used to measure the quality of the recommendations.

On the other hand, we base our re-ranker approach in the xQuAD framework (Santos et al. 2010). Considering this, our proposed model can be formulated as follows:

$$f_{obj}(u,i;\lambda,R^j,R^k) = \lambda \cdot f_{R^j}(u,i) + (1-\lambda) \cdot f_{R^k}(u,i) \tag{16}$$

where R^j and R^k are the two RSs to be combined (the second one is used to re-rank the results from the first one), f_{obj} is the objective function to be maximized. Consequently, the final score of item i is a combination of the ranking position in the original recommender R^j and the second recommender R^k used to re-rank using the combination parameter λ. In both cases, we use a score derived from the one presented before for the hybrid approach, that is, $f_R(u,i) = \text{rank}(s(i,R_u),R_u)$. Then, a new ranking is created by sorting the combined scores obtained through the objective function. As in this case we re-rank a recommendation using another algorithm, we need to restrict the number of items even more. Otherwise, the second method may push items that are not very relevant since they were originally very low in the ranking. Thus, we consider the top-20 items from R^j.

Even though both approaches may seem similar, there is a substantial difference between them. While in the hybrid approach we combine two independent recommendation lists, in the re-ranking approach the candidate items

Table 6: Performance results on Tokyo city for hybrid (H) and re-ranker (RR) methods to mitigate polarization. Rest of notation as in Table 3.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Pop	0.071	0.087	0.746	0.000	0.000	0.960	0.131	0.121	24.9	26.4	**7,253**
H(0.2 Pop + 0.8 IB)	0.071	0.088	*0.801*	*0.019*	*0.022*	*0.921*	*0.078*	*0.069*	24.0	24.4	**7,253**
H(0.8 Pop + 0.2 IB)	0.072	0.089	0.746	0.000	0.000	0.962	0.131	0.120	24.9	26.3	**7,253**
H(0.5 Pop + 0.5 IB)	0.073	0.089	0.765	0.005	0.007	0.946	0.110	0.100	25.2	25.2	**7,253**
RR(Pop, IB)	*0.074*	*0.093*	0.758	0.000	0.000	0.968	0.111	0.101	*23.7*	*24.2*	**7,253**
UB	0.070	0.087	0.769	0.001	0.002	0.968	0.103	0.093	26.0	25.8	*6,931*
H(0.2 UB + 0.8 IB)	0.065	0.081	*0.811*	*0.020*	*0.022*	**0.918**	*0.069*	*0.061*	24.1	25.2	*6,931*
H(0.8 UB + 0.2 IB)	*0.070*	*0.087*	0.768	0.001	0.002	0.966	0.104	0.094	26.0	25.6	*6,931*
H(0.5 UB + 0.5 IB)	0.068	0.085	0.786	0.008	0.010	0.943	0.089	0.080	25.6	25.2	*6,931*
RR(UB, IB)	0.068	0.086	0.778	0.001	0.006	0.954	0.092	0.083	*23.5*	*24.6*	*6,931*
TD	0.071	0.088	0.776	0.001	0.003	0.965	0.097	0.087	25.9	25.4	*6,931*
H(0.2 TD + 0.8 IB)	0.065	0.081	**0.811**	*0.020*	*0.022*	**0.918**	**0.069**	**0.061**	24.1	25.1	*6,931*
H(0.8 TD + 0.2 IB)	*0.071*	*0.088*	0.773	0.001	0.003	0.963	0.099	0.090	25.8	25.3	*6,931*
H(0.5 TD + 0.5 IB)	0.068	0.085	0.789	0.008	0.010	0.941	0.087	0.078	25.4	25.1	*6,931*
RR(TD, IB)	0.069	0.086	0.780	0.002	0.007	0.952	0.091	0.081	*23.6*	*24.5*	*6,931*
IRenMF	0.069	0.083	0.799	0.003	0.008	0.951	0.072	0.063	23.9	*23.8*	*6,931*
H(0.2 IRenMF + 0.8 IB)	0.066	0.082	*0.811*	**0.020**	*0.020*	**0.918**	*0.069*	*0.061*	23.8	25.0	*6,931*
H(0.8 IRenMF + 0.2 IB)	0.071	0.087	0.788	0.003	0.008	0.954	0.081	0.072	24.4	24.1	*6,931*
H(0.5 IRenMF + 0.5 IB)	*0.071*	*0.088*	0.789	0.008	0.011	0.942	0.083	0.074	24.5	24.4	*6,931*
RR(IRenMF, IB)	0.070	0.087	0.784	0.003	0.007	0.951	0.087	0.078	**23.5**	24.2	*6,931*
PGN	0.068	0.086	0.777	0.014	**0.023**	0.932	0.110	0.100	*23.6*	**20.9**	**7,253**
H(0.2 PGN + 0.8 IB)	0.072	0.089	*0.803*	*0.019*	0.021	*0.922*	*0.076*	*0.067*	24.2	24.2	**7,253**
H(0.8 PGN + 0.2 IB)	0.073	0.091	0.760	0.003	0.006	0.956	0.117	0.107	25.3	23.9	**7,253**
H(0.5 PGN + 0.5 IB)	0.074	0.092	0.772	0.006	0.009	0.947	0.102	0.093	25.5	24.6	**7,253**
RR(PGN, IB)	**0.075**	**0.093**	0.766	0.001	0.003	0.961	0.103	0.093	23.7	23.5	**7,253**

Table 7: Performance results on New York city for polarization mitigation. Same notation as Table 6.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Pop	**0.069**	**0.099**	0.837	0.000	0.000	0.758	0.173	0.154	26.2	31.2	**4,416**
H(0.2 Pop + 0.8 IB)	0.043	0.058	*0.941*	*0.140*	**0.112**	*0.715*	*0.021*	*0.015*	*22.6*	*23.2*	**4,416**
H(0.8 Pop + 0.2 IB)	0.068	0.099	0.837	0.000	0.001	0.761	0.172	0.153	25.4	30.7	**4,416**
H(0.5 Pop + 0.5 IB)	0.062	0.086	0.885	0.060	0.063	0.755	0.086	0.073	34.9	30.9	**4,416**
RR(Pop, IB)	0.063	0.085	0.876	0.001	0.001	0.692	0.082	0.068	29.2	27.2	**4,416**
UB	0.056	0.087	0.868	0.002	0.005	0.734	0.102	0.087	36.3	34.5	*3,903*
H(0.2 UB + 0.8 IB)	0.029	0.042	*0.958*	*0.158*	*0.109*	*0.698*	*0.010*	*0.006*	*21.9*	25.9	*3,903*
H(0.8 UB + 0.2 IB)	*0.057*	*0.088*	0.872	0.004	0.009	0.736	0.098	0.084	35.7	33.7	*3,903*
H(0.5 UB + 0.5 IB)	0.047	0.070	0.915	0.078	0.063	0.720	0.043	0.033	33.0	31.3	*3,903*
RR(UB, IB)	0.045	0.064	0.916	0.011	0.021	0.697	0.035	0.027	23.4	*25.4*	*3,903*
TD	0.055	0.087	0.884	0.005	0.012	0.736	0.085	0.071	36.4	35.3	*3,903*
H(0.2 TD + 0.8 IB)	0.030	0.043	**0.959**	**0.161**	*0.111*	*0.698*	**0.009**	**0.006**	*21.9*	26.1	*3,903*
H(0.8 TD + 0.2 IB)	*0.056*	*0.088*	0.885	0.010	0.017	0.735	0.083	0.070	35.2	34.2	*3,903*
H(0.5 TD + 0.5 IB)	0.047	0.070	0.921	0.088	0.068	0.721	0.039	0.030	32.5	31.6	*3,903*
RR(TD, IB)	0.045	0.064	0.922	0.021	0.031	0.705	0.030	0.022	24.8	*26.0*	*3,903*
IRenMF	0.057	0.087	0.854	0.002	0.006	0.743	0.130	0.113	32.4	34.1	*3,903*
H(0.2 IRenMF + 0.8 IB)	0.030	0.043	*0.956*	*0.155*	*0.108*	0.698	*0.011*	*0.007*	**21.3**	*25.8*	*3,903*
H(0.8 IRenMF + 0.2 IB)	*0.057*	*0.088*	0.854	0.002	0.006	0.743	0.129	0.112	32.4	33.9	*3,903*
H(0.5 IRenMF + 0.5 IB)	0.052	0.076	0.892	0.050	0.047	0.735	0.064	0.053	34.5	32.7	*3,903*
RR(IRenMF, IB)	0.053	0.075	0.896	0.006	0.011	**0.676**	0.053	0.042	27.2	27.9	*3,903*
PGN	0.065	0.097	0.859	0.019	0.034	0.769	0.141	0.124	30.8	26.8	**4,416**
H(0.2 PGN + 0.8 IB)	0.044	0.058	*0.941*	*0.140*	*0.111*	0.714	*0.020*	*0.014*	*22.6*	**23.0**	**4,416**
H(0.8 PGN + 0.2 IB)	*0.066*	*0.098*	0.858	0.015	0.027	0.767	0.140	0.123	31.4	27.4	**4,416**
H(0.5 PGN + 0.5 IB)	0.062	0.086	0.893	0.064	0.062	0.745	0.072	0.060	34.5	29.4	**4,416**
RR(PGN, IB)	0.060	0.080	0.887	0.008	0.014	*0.707*	0.064	0.052	26.7	23.7	**4,416**

come only from the first recommender, i.e., the re-ranked items by the second recommender belong to the first model. Additionally, we can only apply the second approach to a pair of RSs ; hence, for the sake of comparability, we restrict the size of the set $\mathcal{R}$ to hybrid recommenders of size two, although in the future, we would like to investigate how to combine larger pools of recommenders.

Hence, based on the proposed approaches, we present in Tables 6, 7, and 8 the results for the cities of Tokyo, New York, and London of the following recommenders: Pop, UB, TD, IRenMF, and PGN. We decided to select these recommenders because they are the ones that achieve the best values according

Table 8: Performance results on London city for polarization mitigation. Same notation as Table 6.

	Accuracy		**Novelty**	**Diversity**	**Popularity**		**Exposure**		**Distance**		**Coverage**
Recommender	**P**	**nDCG**	**EPC**	**Gini**	**PopI**	**PopC**	**ExpP**	**ExpR**	**DistT**	**DistU**	**UC**
Pop	0.039	0.047	0.898	0.001	0.001	0.878	0.195	0.185	15.3	15.3	**2,301**
H(0.2 Pop + 0.8 IB)	0.031	0.035	*0.958*	*0.168*	**0.125**	*0.740*	*0.020*	*0.017*	20.8	20.8	**2,301**
H(0.8 Pop + 0.2 IB)	0.040	0.048	0.898	0.001	0.001	0.875	0.190	0.180	**14.8**	*15.2*	**2,301**
H(0.5 Pop + 0.5 IB)	0.040	0.047	0.920	0.035	0.041	0.825	0.093	0.086	18.7	17.5	**2,301**
RR(Pop, IB)	**0.044**	*0.051*	0.912	0.002	0.002	0.786	0.069	0.059	18.0	17.5	**2,301**
UB	0.037	0.048	0.923	0.004	0.007	0.781	0.045	0.036	20.7	22.4	1,824
H(0.2 UB + 0.8 IB)	0.023	0.031	**0.976**	*0.213*	*0.122*	**0.698**	**0.004**	**0.003**	22.5	26.3	1,824
H(0.8 UB + 0.2 IB)	*0.038*	*0.049*	0.923	0.005	0.008	0.782	0.044	0.036	20.6	22.2	1,824
H(0.5 UB + 0.5 IB)	0.032	0.043	0.948	0.078	0.056	0.735	0.020	0.015	22.6	23.2	1,824
RR(UB, IB)	*0.038*	0.047	0.937	0.013	0.018	0.753	0.026	0.020	*19.9*	*21.9*	1,824
TD	0.037	0.049	0.927	0.007	0.012	0.813	0.044	0.036	24.9	24.6	1,824
H(0.2 TD + 0.8 IB)	0.022	0.030	**0.976**	*0.216*	*0.121*	*0.699*	**0.004**	**0.003**	22.7	26.4	1,824
H(0.8 TD + 0.2 IB)	*0.038*	*0.050*	0.927	0.007	0.012	0.812	0.044	0.036	24.2	24.2	1,824
H(0.5 TD + 0.5 IB)	0.033	0.044	0.950	0.091	0.064	0.751	0.019	0.014	24.1	24.5	1,824
RR(TD, IB)	0.037	0.046	0.942	0.025	0.029	0.764	0.022	0.016	*20.8*	*22.0*	1,824
IRenMF	0.036	0.047	0.932	0.021	0.024	0.812	0.044	0.037	22.2	23.0	1,826
H(0.2 IRenMF + 0.8 IB)	0.023	0.031	**0.976**	**0.221**	*0.124*	*0.702*	**0.004**	**0.003**	22.1	26.2	1,826
H(0.8 IRenMF + 0.2 IB)	*0.038*	0.050	0.933	0.024	0.025	0.810	0.042	0.035	21.8	22.5	1,826
H(0.5 IRenMF + 0.5 IB)	0.035	0.047	0.952	0.104	0.067	0.758	0.018	0.014	21.8	23.1	1,826
RR(IRenMF, IB)	0.040	*0.051*	0.942	0.040	0.035	0.753	0.022	0.016	*19.1*	*21.4*	1,826
PGN	0.042	0.052	0.912	0.008	0.015	0.856	0.118	0.109	*16.4*	**14.4**	**2,301**
H(0.2 PGN + 0.8 IB)	0.031	0.036	*0.959*	*0.171*	*0.124*	*0.739*	*0.020*	0.016	20.7	20.6	**2,301**
H(0.8 PGN + 0.2 IB)	*0.043*	**0.054**	0.911	0.007	0.013	0.856	0.115	0.105	*16.4*	14.5	**2,301**
H(0.5 PGN + 0.5 IB)	0.041	0.049	0.931	0.060	0.057	0.800	0.060	0.053	19.6	17.3	**2,301**
RR(PGN, IB)	*0.043*	0.050	0.921	0.007	0.012	0.769	0.050	0.042	19.0	16.7	**2,301**

to the accuracy metrics. For each recommender, we show three configurations regarding the hybrid approaches denoted as H(R_1, R_2), where each model is combined with the IB recommender with different weights. These weights are designed to balance the contribution of each model in the final recommendations. As there might be a large number of possible configurations, we decided to focus on three weights: 0.2, 0.8, and 0.5. These weights allow us to explore the effect in the recommendations when giving less importance to the first recommender (0.2), the same weight to both models (0.5), and more importance to the first algorithm (0.8). Thus, for example for H(0.2 Pop + 0.8 IB), the final score of every item is created from Pop recommender and IB recommender contributing 20% and 80% to the final score, respectively. We also include one re-ranker configuration, denoted as RR(R_1, R_2), where, as explained before, the IB recommender is used to re-rank the top 20 recommended items from each method. The reason why we selected the IB approach is straightforward: it is the personalized recommender (discarding the pure geographical ones) that achieves the best values in novelty, diversity, and exposure while not being the worst in terms of accuracy.

When analyzing these results, we notice some interesting outcomes. In New York, we observe that the best recommender in terms of accuracy is still the pure Pop model, however, when using the hybrid IB with a weight of 0.5 we reduce the popularity bias (as we can see in the PopI metric) while improving almost in half the exposure values. Better mitigation results are obtained when the weight on IB is higher, but in that scenario, accuracy metrics decrease by more than a 37% (from 0.069 to 0.043 in terms of P). For the rest of the models (UB, TD, IRenMF, PGN) in this city we do observe that using a hybrid with a weight of 0.2 in the IB component allows us to alleviate most of the biases while also obtaining slightly higher values in terms of accuracy. This is particularly interesting because we are able to maintain similar levels of accuracy while improving significantly the results obtained in terms of novelty, diversity, and polarization mitigation using such a simple technique. With respect to comparing the performance of the re-rankers with the hybrids, we can observe that, in general, re-rankers obtain comparable results to those of using a weight of 0.5 for the hybrids, which might be reasonable since the IB re-ranker can only modify the ranking of the top-20 items returned by the recommender, so the (biased) original recommendations still maintain a strong effect in the final ranking. It is important to note that, regarding the geographical polarization, we observe that in the case of New York we are able to reduce this bias when using a weight of 0.8 with the IB approach in the hybrid model (in the case of DistT, more than a 26%, from 30.8 to 22.6) or when using the re-ranker (here, for DistT, more than a 13% improvement, from 30.8 to 26.7). However, the reduction of the bias in these metrics is still far from the values reported in the Skyline of Table 4. In fact, it should be noted that any reduction of this bias would be surprising considering that the IB recommender does not include any geographical component. Regarding this, we performed experiments considering the KDE as a candidate algorithm to build the hybrids and the re-rankers. However, we observed that when we reduced the distance of the recommended venues to the user, the accuracy of the recommendations decreased significantly. For example, in New York, we observed that when using our reranking approach, the performance in terms of ranking accuracy decreases, for all recommenders, more than a 50%, evidencing that the KDE is not a good method to be used with these mitigation proposals.

The results for the Tokyo dataset, shown in Table 6, confirm a very interesting case where the best algorithm in terms of accuracy outperforms the best recommender reported in Table 3 (which was, in fact, the Pop recommender, also reported in this table). Here, the best performing configuration is the PGN with the IB re-ranker. Although this is a promising result, we observe that in this case, the re-ranker is obtaining lower values in terms of novelty and diversity while suffering from a larger popularity bias (but lower category bias). Nevertheless, there is one example that shows a very good tradeoff among all the metrics: H(0.2 PGN + 0.8 IB). In this case, it also obtains a higher performance than the pure PGN; more specifically, we are able to improve the accuracy a 5.88% in terms of P while reducing the ExpP and ExpR by a 30.9% and a 33% respectively when compared against the result obtained by the PGN. In the case of the city of London, we observe in Table 8 that the best performing configuration in terms of P is the Pop algorithm with the item reranker, and in terms of nDCG is the PGN combined with the IB with a weight of 0.2. However, the most important conclusion about these cases is that we again managed to improve performance in terms of ranking accuracy (up to 14% in the case of the Pop), maintaining similar values in novelty and diversity while reducing exposure polarization. This indicates that, as long as we have a test set available, we are able to increase the performance of the different models in other dimensions without degrading the accuracy ranking dramatically. The geographical polarization, on the other hand, is more difficult to improve, as discussed for the New York city. However, all these examples confirm that it is possible to find configurations where better results than the original recommenders are obtained, either in terms of accuracy while keeping similar polarization values, or reduced polarization measurements while keeping comparable accuracies.

6 Conclusions and future work

Research on the characterization of biases in Artificial Systems in general, and Recommender Systems in particular, is an area of growing interest. In this work, we have focused on polarization, that is, how far an algorithm deviates from what was observed in the training data. We have characterized four types of polarizations in Location-Based Recommender Systems, a specific type of algorithms that suggest points-of-interest (or venues) to users, by exploiting their preferences and other inherent characteristics from the touristic domain, such as location and item categories. This type of suggestion is one of the main means for users to explore a city and the business of venue owners is directly affected by them, hence providing equitable recommendations is a key aspect that may have a concrete impact on society. In detail, we have analyzed the popularity polarization (both from venues and categories), the exposure of venues, and the polarization related to geographical distance.

After the characterization, in the experiments, we have assessed these different sources of polarization by comparing several state-of-the-art recommenders. Our results show that popularity polarization is prevalent in many of these recommendation algorithms, both in generic or tailored approaches for location-based recommendation. In terms of exposure and distance, there is a difficult tradeoff to satisfy with respect to accuracy. This is, as discussed in the paper, tied to the test set available which may itself contain bias.

Finally, we propose two techniques based on combining recommendation algorithms (either by building a hybrid or a re-ranker method) that have demonstrated promising results to mitigate the analyzed polarizations. In particular, for some cases, these approaches are able to improve accuracy while reducing the observed polarization. However, this effect depends on how the recommenders to be combined are selected and also on the test set used to analyze the quality of the recommendations.

That is why, in the future, a deeper analysis is necessary to be performed so that other families of algorithms are also included. In particular, more dynamic approaches based on sequences or other contexts available in the tourism domain might have different levels of sensitivity to these biases. Similarly, we believe the polarization assessment performed herein should be extended to analyze how it affects groups of different users, for example, according to sensitive attributes such as user gender, age, or ethnicity. In the same way, a more automatic approach to detect which recommendation algorithm should be used to be combined with when using the proposed techniques, needs to be analyzed to scale these approaches to larger datasets or other recommendation tasks.

At the same time, we would like to explore other strategies for reducing the polarization of recommendations without the need for the users' ground truth, so that the polarization reduction is not so dependent on the test set. Indeed, besides the algorithmic bias discussed in the introduction, another popular source of bias is the fact that the data could be collected in a biased way, or that users interact with the system in such a way that biased interactions are recorded (Chen et al. 2020). In this paper, we aimed at understanding how biased or polarized the recommendations depending on the algorithm are, since, even starting from the same data, some recommenders may output more polarized results than others. However, this only relates to training data, but this could also affect the test data, since the original data from where the training and test splits are generated are the same. To the best of our knowledge, there are not many feasible and realistic solutions to this aspect, and the community is still working on it. One possibility would be to collect complete and unbiased datasets. This has been done for specific domains (Cañamares and Castells 2018), evidencing the very high cost it is required for such constructions. We may also focus on specific subsets of users or items (those items in the long-tail or users with enough interactions in the system), however this is not guaranteed to reduce the bias in the data, and may have generalization problems. A potential solution that would require further analysis and proper formalization is the use of simulations to generate synthetic data without biased ground truth. However, this alternative would depend on the possibility of generating realistic user interactions, which is something quite challenging, even more for location-based information (Ekstrand et al. 2021a; Hazrati and Ricci 2022).

Finally, other ways to mitigate these and other biases should be explored in the field of Point-of-Interest recommendation, beyond exposure polarizations, such as selection biases – where the observed interactions are not a representative sample of all the interactions – and feedback loop effects – where the exposed items by the recommender are used as training data for the same recommender, intensifying the biases over time – (Chen et al. 2020).

Acknowledgements This work has been funded by the Ministerio de Ciencia e Innovación (reference PID2019-108965GB-I00) and by the European Social Fund (ESF), within the 2017 call for predoctoral contracts.

Declarations

Funding

This work has been funded by the Ministerio de Ciencia e Innovación (reference PID2019-108965GB-I00) and by the European Social Fund (ESF), within the 2017 call for predoctoral contracts.

Conflict of interest

The authors declare that they have no conflict of interest.

Availability of data and material

The dataset used in this paper is publicly available. Likewise, the cleaned data will be provided once the paper is accepted.

Code availability

The code will be available once the paper is accepted in the following Bitbucket repository PabloSanchezP/BiasMitigationLBRs/

Appendix

As already discussed in Section 5.2, in Figure 9 we show the distribution of the top 30% most popular venues in the three different cities (left column). On the right column of this figure, the check-in distribution performed by users is depicted. It is remarkable how strong the long-tail effect is in both situations, meaning that there are items and users that concentrate most of the check-ins.

References

Abdollahpouri H, Burke R, Mobasher B (2017) Controlling popularity bias in learning-to-rank recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, ACM, pp 42–46

Abdollahpouri H, Adomavicius G, Burke R, Guy I, Jannach D, Kamishima T, Krasnodebski J, Pizzato LA (2019a) Beyond personalization: Research directions in multistakeholder recommendation. CoRR abs/1905.01986

Abdollahpouri H, Burke R, Mobasher B (2019b) Managing popularity bias in recommender systems with personalized re-ranking. In: Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, AAAI Press, pp 413–418

Adamopoulos P, Tuzhilin A, Mountanos P (2015) Measuring the concentration reinforcement bias of recommender systems. In: Poster Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, CEUR-WS.org, vol 1441

Adomavicius G, Bockstedt J, Curley S, Zhang J (2014) De-biasing user preference ratings in recommender systems. In: Joint Workshop on Interfaces and Human Decision Making in Recommender Systems, p 2

Anderson C (2006) The long tail: Why the future of business is selling less of more. Hachette UK

Ariza A, Fabbri F, Boratto L, Salamó M (2021) From the beatles to billie eilish: Connecting provider representativeness and exposure in session-based recommender systems. In: Hiemstra D, Moens M, Mothe J, Perego R, Potthast M, Sebastiani F (eds) Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol 12657, pp 201–208, DOI 10.1007/978-3-030-72240-1_16, URL https://doi.org/10.1007/978-3-030-72240-1_16

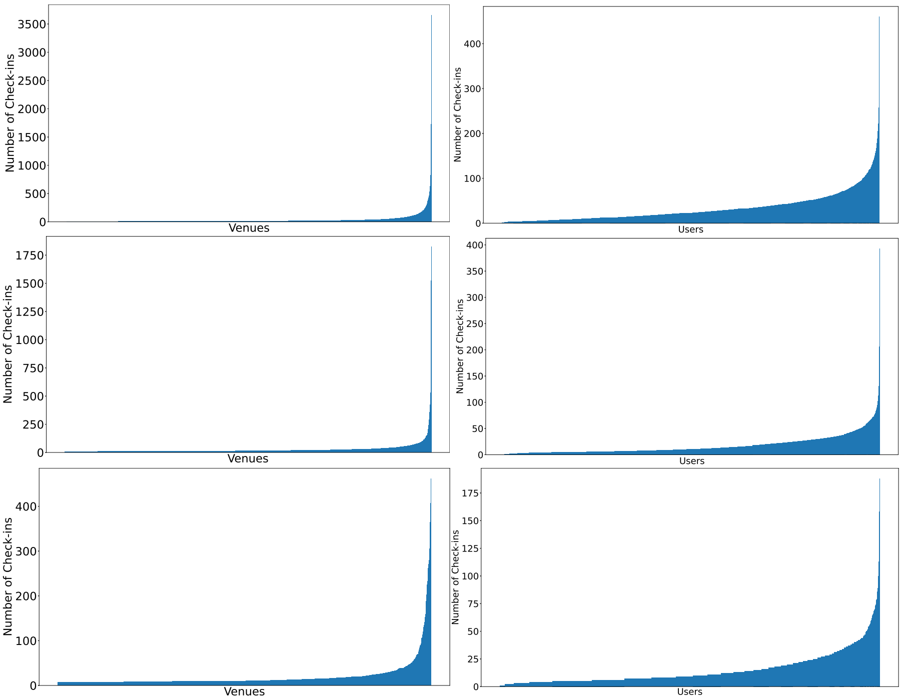

Fig. 9: In the first column, we show the distribution of the 30% of the most popular venues in the cities of Tokyo (first row), New York (second row) and London (last row). In the second column, we show the distribution of the check-ins performed by the users in the same cities

.

Bellogín A, Castells P, Cantador I (2017) Statistical biases in information retrieval metrics for recommender systems. Inf Retr Journal 20(6):606–634

Benouaret I, Lenne D (2016) A package recommendation framework for trip planning activities. In: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, pp 203–206

Blodgett SL, Barocas S, III HD, Wallach HM (2020) Language (technology) is power: A critical survey of "bias" in NLP. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, pp 5454–5476, DOI 10.18653/v1/2020.acl-main.485, URL https://doi.org/10.18653/v1/2020.acl-main.485

Boratto L, Fenu G, Marras M (2019) The effect of algorithmic bias on recommender systems for massive open online courses. In: Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Proceedings, Part I, Springer, vol 11437, pp 457–472

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, PMLR, Proceedings of Machine Learning Research, vol 81, pp 77–91, URL http://proceedings.mlr.press/v81/buolamwini18a.html

Burke RD (2002) Hybrid recommender systems: Survey and experiments. User Model User-Adapt Interact 12(4):331–370

Campos PG, Díez F, Cantador I (2014) Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. User Model User-Adapt Interact 24(1-2):67–119

Cañamares R, Castells P (2017) A probabilistic reformulation of memory-based collaborative filtering: Implications on popularity biases. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 215–224

Cañamares R, Castells P (2018) Should I follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, ACM, pp 415–424

Castells P, Hurley NJ, Vargas S (2015) Novelty and diversity in recommender systems. In: Recommender Systems Handbook, Springer, pp 881–918

Chen D, Ong CS, Xie L (2016) Learning points and routes to recommend trajectories. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, ACM, pp 2227–2232

Chen J, Dong H, Wang X, Feng F, Wang M, He X (2020) Bias and debias in recommender system: A survey and future directions. CoRR abs/2010.03240, URL https://arxiv.org/abs/2010.03240, 2010.03240

Cheng C, Yang H, King I, Lyu MR (2012) Fused matrix factorization with geographical and social influence in location-based social networks. In: Hoffmann J, Selman B (eds) Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada, AAAI Press, URL http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4748

Doan T, Lim E (2019) Modeling location-based social network data with area attraction and neighborhood competition. Data Min Knowl Discov 33(1):58–95, DOI 10.1007/s10618-018-0588-4, URL https://doi.org/10.1007/s10618-018-0588-4

Edizel B, Bonchi F, Hajian S, Panisson A, Tassa T (2019) Fairecsys: mitigating algorithmic bias in recommender systems. International Journal of Data Science and Analytics

Ekstrand MD, Tian M, Azpiazu IM, Ekstrand JD, Anuyah O, McNeill D, Pera MS (2018) All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Conference on Fairness, Accountability and Transparency, FAT 2018, PMLR, vol 81, pp 172–186

Ekstrand MD, Chaney A, Castells P, Burke R, Rohde D, Slokom M (2021a) Simurec: Workshop on synthetic data and simulation methods for recommender systems research. In: Pampín HJC, Larson MA, Willemsen MC, Konstan JA, McAuley JJ, Garcia-Gathright J, Huurnink B, Oldridge E (eds) RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021, ACM, pp 803–805, DOI 10.1145/3460231.3470938, URL https://doi.org/10.1145/3460231.3470938

Ekstrand MD, Das A, Burke R, Diaz F (2021b) Fairness and discrimination in information access systems. CoRR abs/2105.05779, URL https://arxiv.org/abs/2105.05779, 2105.05779

Gao H, Tang J, Hu X, Liu H (2013) Exploring temporal effects for location recommendation on location-based social networks. In: Seventh ACM Conference on Recommender Systems, RecSys '13, ACM, pp 93–100

Gao R, Li J, Li X, Song C, Zhou Y (2018) A personalized point-of-interest recommendation model via fusion of geo-social information. Neurocomputing 273:159–170, DOI 10.1016/j.neucom.2017.08.020, URL https://doi.org/10.1016/j.neucom.2017.08.020

de Gemmis M, Lops P, Musto C, Narducci F, Semeraro G (2015) Semantics-aware content-based recommender systems. In: Recommender Systems Handbook, Springer, pp 119–159

Gunawardana A, Shani G (2015) Evaluating recommender systems. In: Recommender Systems Handbook, Springer, pp 265–308

Guo F, Dunson DB (2015) Uncovering systematic bias in ratings across categories: A bayesian approach. In: Proceedings of the 9th ACM Conference on Recommender Systems, ACM, pp 317–320

Hazrati N, Ricci F (2022) Simulating users' interactions with recommender systems. In: UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022, Adjunct Proceedings, ACM, pp 95–98, DOI 10.1145/3511047.3536402, URL https://doi.org/10.1145/3511047.3536402

He R, McAuley J (2016) Fusing similarity models with markov chains for sparse sequential recommendation. In: IEEE 16th International Conference on Data Mining, ICDM 2016, IEEE, pp 191–200

Herzog D, Wörndl W (2019) User-centered evaluation of strategies for recommending sequences of points of interest to groups. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, ACM, pp 96–100

Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), IEEE Computer Society, pp 263–272

Jacobs AZ, Blodgett SL, Barocas S, III HD, Wallach HM (2020) The meaning and measurement of bias: lessons from natural language processing. In: Hildebrandt M, Castillo C, Celis LE, Ruggieri S, Taylor L, Zanfir-Fortuna G (eds) FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, p 706, DOI 10.1145/3351095.3375671, URL https://doi.org/10.1145/3351095.3375671

Jannach D, Lerche L, Kamehkhosh I, Jugovac M (2015) What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Model User-Adapt Interact 25(5):427–491

Jannach D, Kamehkhosh I, Bonnin G (2016) Biases in automated music playlist generation: A comparison of next-track recommending techniques. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, ACM, pp 281–285

Ji Y, Sun A, Zhang J, Li C (2021) A critical study on data leakage in recommender system offline evaluation. CoRR abs/2010.11060, URL https://arxiv.org/abs/2010.11060, 2010.11060

Kapcak Ö, Spagnoli S, Robbemond V, Vadali S, Najafian S, Tintarev N (2018) Tourexplain: A crowdsourcing pipeline for generating explanations for groups of tourists. In: Proceedings of the Workshop on Recommenders in Tourism, RecTour 2018, CEUR-WS.org, vol 2222, pp 33–36

Kaufman S, Rosset S, Perlich C, Stitelman O (2012) Leakage in data mining: Formulation, detection, and avoidance. ACM Trans Knowl Discov Data 6(4):15:1–15:21, DOI 10.1145/2382577.2382579, URL https://doi.org/10.1145/2382577.2382579

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S (2020) Racial disparities in automated speech recognition. Proc Natl Acad Sci USA 117(14):7684–7689, DOI 10.1073/pnas.1915768117, URL https://doi.org/10.1073/pnas.1915768117

Koren Y, Bell RM (2015) Advances in collaborative filtering. In: Ricci F, Rokach L, Shapira B (eds) Recommender Systems Handbook, Springer, pp 77–118, DOI 10.1007/978-1-4899-7637-6_3, URL https://doi.org/10.1007/978-1-4899-7637-6_3

Li X, Cong G, Li X, Pham TN, Krishnaswamy S (2015) Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 433–442

Lian D, Zhao C, Xie X, Sun G, Chen E, Rui Y (2014) Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, ACM, pp 831–840

Lika B, Kolomvatsos K, Hadjiefthymiades S (2014) Facing the cold start problem in recommender systems. Expert Syst Appl 41(4):2065–2073

Liu Y, Wei W, Sun A, Miao C (2014) Exploiting geographical neighborhood characteristics for location recommendation. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, ACM, pp 739–748

Liu Y, Pham T, Cong G, Yuan Q (2017) An experimental evaluation of point-of-interest recommendation in location-based social networks. PVLDB 10(10):1010–1021

Massimo D, Ricci F (2022) Building effective recommender systems for tourists. AI Magazine 43(2):209–224, DOI https://doi.org/10.1002/aaai.12057, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12057, https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12057

Mehrotra R, McInerney J, Bouchard H, Lalmas M, Diaz F (2018) Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In: Cuzzocrea A, Allan J, Paton NW, Srivastava D, Agrawal R, Broder AZ, Zaki MJ, Candan KS, Labrinidis A, Schuster A, Wang H (eds) Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, pp 2243–2251, DOI 10.1145/3269206.3272027, URL https://doi.org/10.1145/3269206.3272027

Meng Z, McCreadie R, Macdonald C, Ounis I (2020) Exploring data splitting strategies for the evaluation of recommendation models. In: Santos RLT, Marinho LB, Daly EM, Chen L, Falk K, Koenigstein N, de Moura ES (eds) RecSys 2020: Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, September 22-26, 2020, ACM, pp 681–686, DOI 10.1145/3383313.3418479, URL https://doi.org/10.1145/3383313.3418479

Miller HJ (2004) Tobler's first law and spatial analysis. Annals of the Association of American Geographers 94(2):284–289

Ning X, Desrosiers C, Karypis G (2015) A comprehensive survey of neighborhood-based recommendation methods. In: Recommender Systems Handbook, Springer, pp 37–76

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453, DOI 10.1126/science.aax2342, URL https://www.science.org/doi/abs/10.1126/science.aax2342, https://www.science.org/doi/pdf/10.1126/science.aax2342

Pariser E (2011) The Filter Bubble: What the Internet Is Hiding from You. The Penguin Group

Park Y, Tuzhilin A (2008) The long tail of recommender systems and how to leverage it. In: Pu P, Bridge DG, Mobasher B, Ricci F (eds) Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, 2008, ACM, pp 11–18, DOI 10.1145/1454008.1454012, URL https://doi.org/10.1145/1454008.1454012

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: bayesian personalized ranking from implicit feedback. In: UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp 452–461

Ricci F, Rokach L, Shapira B (2015) Recommender systems: Introduction and challenges. In: Recommender Systems Handbook, Springer, pp 1–34

Said A, Bellogín A (2014) Comparative recommender system evaluation: benchmarking recommendation frameworks. In: Eighth ACM Conference on Recommender Systems, RecSys '14, ACM, pp 129–136

Said A, Bellogín A, de Vries A (2013) A top-n recommender system evaluation protocol inspired by deployed systems. In: Proceedings of the 2013 ACM RecSys Workshop on Large-Scale Recommender Systems

Sánchez P, Bellogín A (2019) Attribute-based evaluation for recommender systems: incorporating user and item attributes in evaluation metrics. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2017, ACM, pp 378–382

Sánchez P, Bellogín A (2021) On the effects of aggregation strategies for different groups of users in venue recommendation. Inf Process Manag 58(5):102609, DOI 10.1016/j.ipm.2021.102609, URL https://doi.org/10.1016/j.ipm.2021.102609

Sánchez P, Bellogín A (2022) Point-of-interest recommender systems based on location-based social networks: A survey from an experimental perspective. ACM Computing Surveys DOI 10.1145/3510409

Santos RLT, Macdonald C, Ounis I (2010) Exploiting query reformulations for web search result diversification. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, ACM, pp 881–890

Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, ACM, pp 2219–2228

Steck H (2018) Calibrated recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, ACM, pp 154–162

Sun A (2022) From counter-intuitive observations to a fresh look at recommender system. CoRR abs/2210.04149, DOI 10.48550/arXiv.2210.04149, URL https://doi.org/10.48550/arXiv.2210.04149, 2210.04149

Valcarce D, Bellogín A, Parapar J, Castells P (2018) On the robustness and discriminative power of information retrieval metrics for top-n recommendation. In: Pera S, Ekstrand MD, Amatriain X, O'Donovan J (eds) Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018, ACM, pp 260–268, DOI 10.1145/3240323.3240347, URL https://doi.org/10.1145/3240323.3240347

Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, ACM, pp 109–116

Vargas S, Castells P (2014) Improving sales diversity by recommending users to items. In: Kobsa A, Zhou MX, Ester M, Koren Y (eds) Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014, ACM, pp 145–152, DOI 10.1145/2645710.2645744, URL https://doi.org/10.1145/2645710.2645744

Wang H, Terrovitis M, Mamoulis N (2013) Location recommendation in location-based social networks using user check-in data. In: 21st SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2013, ACM, pp 364–373

Wasilewski J, Hurley N (2018) Are you reaching your audience?: Exploring item exposure over consumer segments in recommender systems. In: Mitrovic T, Zhang J, Chen L, Chin D (eds) Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018, ACM, pp 213–217, DOI 10.1145/3209219.3209246, URL https://doi.org/10.1145/3209219.3209246

Weydemann L, Sacharidis D, Werthner H (2019) Defining and measuring fairness in location recommendations. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, LocalRec@SIGSPATIAL 2019, ACM, pp 6:1–6:8

Yang D, Zhang D, Qu B (2016) Participatory cultural mapping based on collective behavior data in location-based social networks. ACM TIST 7(3):30:1–30:23

Ye M, Yin P, Lee W, Lee DL (2011) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, ACM, pp 325–334

Yuan F, Jose JM, Guo G, Chen L, Yu H, Alkhawaldeh RS (2016) Joint geo-spatial preference and pairwise ranking for point-of-interest recommendation. In: 28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2016, San Jose, CA, USA, November 6-8, 2016, IEEE Computer Society, pp 46–53, DOI 10.1109/ICTAI.2016.0018, URL https://doi.org/10.1109/ICTAI.2016.0018

Zehlike M, Hacker P, Wiedemann E (2020) Matching code and law: achieving algorithmic fairness with optimal transport. Data Min Knowl Discov 34(1):163–200, DOI 10.1007/s10618-019-00658-8, URL https://doi.org/10.1007/s10618-019-00658-8

Zhang J, Chow C (2013) igslr: personalized geo-social location recommendation: a kernel density estimation approach. In: 21st SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2013, ACM, pp 324–333

Zhang J, Chow C (2015) Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 443–452

Zhang J, Chow C, Li Y (2014) LORE: exploiting sequential influence for location recommendations. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, pp 103–112