Information Technology & Tourism (2025) 27:75–124
https://doi.org/10.1007/s40558-024-00304-0

ORIGINAL RESEARCH

Understanding the influence of data characteristics on the performance of point-of-interest recommendation algorithms

Linus W. Dietz[1] · **Pablo Sánchez**[2,3] · **Alejandro Bellogín**[3]

Received: 19 October 2023 / Revised: 22 November 2024 / Accepted: 6 December 2024 /
Published online: 3 January 2025

Abstract
Point-of-interest (POI) recommendations are essential for travelers and the e-tourism business. They assist in decision-making regarding what venues to visit and where to dine and stay. While it is known that traditional recommendation algorithms' performance depends on data characteristics like sparsity, popularity bias, and preference distributions, the impact of these data characteristics has not been systematically studied in the POI recommendation domain. To fill this gap, we extend a previously proposed explanatory framework by introducing new explanatory variables specifically relevant to POI recommendation. At its core, the framework relies on having subsamples with different data characteristics to compute a regression model, which reveals the dependencies between data characteristics and performance metrics of recommendation models. To obtain these subsamples, we subdivide a POI recommendation data set on New York City and measure the effect of these characteristics on different classical POI recommendation algorithms in terms of accuracy, novelty, and item exposure. Our findings confirm the crucial role of key data features like density, popularity bias, and the distribution of check-ins in POI recommendation. Additionally, we identify the significance of novel factors, such as user mobility and the duration of user activity. In summary, our work presents a generic method to quantify the influence of data characteristics on recommendation performance. The results not only show why certain POI recommendation algorithms excel in specific recommendation problems derived from a LBSN check-in data set in New York City, but also offer practical insights into which data characteristics need to be addressed to achieve better recommendation performance.

Keywords Point-of-interest recommendation · Offline evaluation · Regression analysis · Data characteristics

Extended author information available on the last page of the article

1 Introduction

Understanding which recommendation algorithm is most effective for a specific data set is crucial, as it has been widely acknowledged that no single recommender can achieve optimal performance in all scenarios (Im and Hars 2007; Anelli et al. 2022). Apart from the performance variation of the algorithms across different data sets, it should be taken into account that the quality of any recommender can be evaluated through a wide array of different dimensions (Gunawardana et al. 2022). While accuracy may be the primary concern when recommending items that a user might actually consume, it is equally important to prioritize recommending different items for the users (diversity), items that may be unfamiliar to users and, hence, surprise them (novelty), or ensure that our recommendations are not biased towards individual users or items (fairness) (Castells et al. 2022; Ekstrand et al. 2022). However, designing models that perform well in all these dimensions is challenging, as they need to deal with, for example, the accuracy-diversity trade-off (Isufi et al. 2021). Whereas such analysis of accuracy versus novelty, diversity, and other dimensions has been conducted in traditional recommendation scenarios like movies or books, within the point-of-interest recommendation domain, the problem has not been analyzed in such detail, although some researchers have started to examine these dimensions in this context (Massimo and Ricci 2021; Sánchez and Dietz 2022).

Moreover, while most existing studies have primarily focused on evaluating the quality of recommendations based on the accuracy of recommended venues through offline experimentation metrics (i.e., using ranking accuracy metrics like Precision or Recall), there remains a lack of consensus on the other crucial aspects of evaluation methodology, such as data sets, data filtering, data partitioning, and other evaluation metrics (Sánchez and Bellogín 2022). In addition, it is important to consider that users can be grouped based on simple touristically relevant information, such as their origin or the categories of visited POIs, all of which can correlate with their preferences. It has been shown that recommendation performance may fluctuate substantially depending on the user group a user belongs to, especially between local users and visiting tourists (Sánchez and Dietz 2022).

Deldjoo et al. (2021) proposed a method to analyze the impact of different data characteristics on the accuracy and fairness of matrix factorization algorithms in the movie and book recommendation domains. As opposed to such classical recommendation problems, point of interest (POI) recommendations are influenced in a way larger degree by further factors, such as seasonality, geographical influences of the venues to be recommended, and the type of users of such system (Sánchez and Bellogín 2022). Hence, in this paper, we investigate the success factors of classical and POI recommendation algorithms through the lens of data characteristics present in the data set used as input by the recommendation models. To do so, we incorporate the most relevant influences in a framework derived from (Deldjoo et al. 2021) to analyze the effect they have on the performance of a set of recommenders.

1.1 Using data characteristics to explain recommender performance

When proposing a new recommendation model, researchers often begin with intuitions or anecdotal evidence, until they ultimately obtain empirical evidence through experimentation to validate the model effectiveness. However, relying solely on intuitions or anecdotal evidence to justify the quality of recommendations is insufficient. This approach may explain why some models obtain excellent results on some data sets while, at the same time, they perform poorly in others, as recent efforts on reproducibility have demonstrated (Dacrema et al. 2019; Said and Bellogín 2014).

It should be noted that some studies have analyzed the effect of different aspects in the recommendations, like the data partitioning (Meng et al. 2020) and the hyperparameters of the models (Anelli et al. 2019). The experiments presented in this paper are based on the approach by the works of Deldjoo et al. (2021, 2020), where the authors defined a set of explanatory variables that model the characteristics of the data set (e.g., ratings per user, per item, population bias, etc.). On the basis of those works, herein, we use a similar methodology as proposed by Deldjoo et al. (2021), but adapt the framework towards POI recommendation by incorporating additional variables that capture the unique dynamics of the POI recommendation domain. The core idea is that recommendation performance is influenced by quantifiable patterns in the data, which result in easier or more difficult recommendation problems. For example, a high density of the user-rating matrix, i.e., where many users have already rated a large portion of items, generally provides recommendation models with ample signal to compute suitable recommendations. Hence, many such data characteristics can potentially influence performance. In this paper, we study not only the impact of these data characteristics on ranking quality, but we also analyze the effect on the recommendations in terms of both novelty and the amount of exposure each item receives across all users. These aspects are important as they help identifying recommenders that may be amplifying unfairness in the exposure of items. For example, if a model recommends popular POIs significantly more frequently than they appear in the test set, it can lead to low novelty values and greater disparities in item exposure. In e-tourism, this is a key consideration as it can decide which businesses thrive.

Despite the commonalities with the approach presented in Deldjoo et al. (2021), we further develop several aspects of the general method, which requires generating a sufficiently large number of recommendation data sets with varying data characteristics to compute a regression model. To achieve this, we start from a widely-used check-in data set based on the location-based social network Foursquare and generate domain-driven subsamples; that is, considering characteristics of special importance in both traditional and POI recommendation. The subsamples are created using filter-like rules targeting the interaction density, popularity of venues, seasonality and origin of users, e.g., locals visitors of a travel destination. This is in contrast with the original approach which used a constraint-based *random* sampling method to derive subsamples. By subdividing the set of all POI visits in the data set based on the aforementioned *domain-specific* rules, we can better steer the subsampling process and simultaneously obtain interpretable subsamples, which can be used to understand inherent attributes and characteristics of

the POI recommendation domain. As an outcome, we obtain individual subsamples of the data set with varying data characteristics.

This variability in the data characteristics in the individual subsamples is important as, in the next step, we independently perform recommendation experiments with each subsample and record the outcome variables in terms of accuracy, novelty, and item exposure. Herein, it is important to note that these subsamples are synthetic simulations of recommendation data sets. We compute regression models using the data characteristics of the individual subsamples as independent variables and the performance metrics as dependent variables. In other terms, we quantify data characteristics using *explanatory* variables to explain the performance changes of the recommenders in terms of ranking accuracy, novelty, and item exposure using the regression model. Through the quantification of the statistical significance of the explanatory variables within the regression model, we ensure that the determined influences are robust and not spurious effects. To capture all potential influences on the recommendation outcome in the POI recommendation domain, we further extend the variables proposed for classical recommendation domains (Deldjoo et al. 2021, 2020; Adomavicius and Zhang 2012) with spatio-temporal features. An analysis of the statistical significance of the coefficients in the regression model reveals which data characteristics are needed to explain the recommenders' performance.

1.2 Overview of contributions

While the core concepts of this paper are based on previous approaches of Deldjoo et al. (2021) and Adomavicius and Zhang (2012), we go beyond of simply adapting an established method to the domain of POI recommendations. We make the following conceptual and methodological contributions:

1. We extend the corpus of explanatory variables for analyzing the effect of different data characteristics, including geographic aspects, in a varied set of state-of-the-art POI recommendation algorithms (Sect. 3.2). This analysis considers three complementary evaluation dimensions: ranking accuracy, novelty, and item exposure.
2. We introduce a domain-specific subsampling algorithm for POI recommendation (Sect. 4). This algorithm ensures that the simulated data sets are grounded in the domain instead of random subsampling, as done in previous works.
3. We perform a comprehensive analysis of a set of recommendation algorithms by considering 144 different simulations (Sect. 5). Each simulation corresponds to a subsampled recommendation data set of a specific city within the Foursquare check-in data set. In this way, we conduct an analysis of different samples with disparate characteristics to detect which explanatory variables help us to better explain the performance of the recommenders.

1.3 Impact on e-tourism

This research can have significant implications for the tourism industry. POI recommendations play a crucial role in shaping tourists' experiences and guiding their choices of which places to visit; thus, they are a key factor in decision-making. We offer valuable insights that can enhance the tourism industry's ability to provide more tailored and satisfying experiences to travelers by determining which recommendation algorithms should be used in which kind of specific recommendation scenarios. The analysis of the beyond-accuracy metrics, i.e., novelty and item exposure, offers valuable perspectives that cater to the user needs and local businesses. Novelty is especially relevant to local users, as it fosters the exploration of their city, whereas item exposure is a precondition to ensure that the flow of visitors is dissipated on many venues, contributing to provider fairness (Deldjoo et al. 2023). By harnessing data characteristics such as density, popularity bias, and user activity duration, the proposed methods can be used to select algorithms that better align with the business goals of destination management stakeholders. Thus, understanding how user behavior varies in different parts of a destination can enable businesses to tailor their offerings and marketing efforts more effectively. Our research underscores the value of data-driven decision-making in the tourism sector. By leveraging the insights gained from this study, both providers of POI recommendation platforms and the tourism industry can enhance their ability to provide personalized and engaging experiences to users.

2 Background

2.1 Point-of-interest recommendation

The POI recommendation problem is typically defined as suggesting venues in a city or particular region that the target user has not previously visited (Massimo and Ricci 2022). These venues have a specific location, typically expressed as latitude and longitude, and might be varied in nature, including museums, parks, restaurants, or bars, among others. As in the traditional recommendation scenario, the objective is to maximize the number of relevant items (in this case, venues) that are being recommended to the user. Formally, as pointed in a recent survey by Sánchez and Bellogín (2022), the POI recommendation problem can be formulated as follows:

$$p^*(u) = \arg\max_{p \in \mathcal{P}} g(u, p, \boldsymbol{\Phi}) \tag{1}$$

where $\mathcal{P}$ denotes all POIs available in the region, p^* is the optimal venue that maximizes the utility for user u among all POIs[1] in $\mathcal{P}$, as defined by the utility function g, and $\boldsymbol{\Phi}$ represents the set of influences, also referred to as contextual information in some works (Adomavicius et al. 2022). This contextual information should

[1] Even though we use the symbol $\mathcal{P}$ to refer to the POIs, in line with the standard notation from the traditional recommendation problem, we shall use the letter I to refer to the items of the system, i.e., the POIs.

be considered to perform meaningful POI recommendations (Manotumruksa et al. 2018).

Temporal, sequential, social, categorical, and, most importantly, geographical information are normally exploited in most POI recommendation approaches (Li et al. 2015; Griesner et al. 2015; Zhang and Chow 2015; Liu et al. 2014). In order to perform POI recommendations, researchers often use the information available in location-based social networks (LBSNs). Foursquare, Gowalla, or Yelp are examples of this type of application where users are allowed to register the check-ins they perform at the different venues they visit. Data sets extracted from such LBSNs are invaluable for understanding the visiting behavior of users. Information about friendship links, along with the geographical coordinates of the venues, their categories, and the timestamps of the check-ins, can be used to model the aforementioned influences and generate potentially interesting recommendations to the users who are new in a specific geographical region, sometimes even requiring completely different approaches, such as reinforcement learning (Massimo and Ricci 2023).

It is important to note that LBSNs typically contain check-ins from different cities around the world. However, as this type of recommendation is affected by all the aforementioned influences (especially the geographical information), many researchers perform recommendations considering each city/region as independent data sets (Liu et al. 2014; Li et al. 2015). This strategy is not only practical from an experimental point of view, but it is also quite reasonable since when a user is in a particular city, she will be interested in visiting venues belonging to that region and not from distant cities. Furthermore, each city may exhibit a unique distribution of POIs to visit, along with distinct cultural characteristics and specific urban planning. In fact, this is one of the reasons why POI recommendation is closely related to the tourism industry (Wang et al. 2020a; Santos et al. 2019), since tourists, when they arrive to a new city, are usually interested in visiting the most relevant venues of that specific city and immerse themselves in the local culture. Besides, we need to consider that tourism is the base of many economies, such as some countries in southern Europe (Cortés-Jiménez 2008), due to the large number of stakeholders involved, including tourists, venue owners, and local residents.

2.2 Specific considerations of the POI recommendation domain

When addressing the POI recommendation problem, it is necessary to regard several domain-specific considerations and problems. Some of them include:

Sparsity: The ratio between the stored preferences in the data set and all the possible interactions between the users and the venues is extremely low. While the densities of LBSNs data sets from Foursquare and Gowalla are approximately 0.0034%, the density of the Netflix and Movielens20M data sets, typically used in classical recommendation, are around 1.77%, hence showing much higher values of sparsity.

Additional influences: As discussed in Sect. 2.1, POI recommendation is influenced by geographical aspects, social connections, and temporal information. Due to the high data sparsity, it is crucial to leverage additional information to enhance

the performance of the algorithms. Among all the information sources, geographical influence plays the most important role since users often prefer to visit nearby POIs in accordance to Tobler's law: *"[...]Everything is related to everything else, but near things are more related than distant things"* (Tobler 1970). However, temporal influence can also provide valuable insights, such as the duration of users' visits and their movement patterns between POIs. Therefore, exploiting information suitable to the respective use cases is essential for the success of the recommendations.

Implicit information: Traditionally, classical recommender systems model user-item interactions using ratings. However, in POI recommendation data sets, we typically lack of explicit ratings and we only have timestamps of user visits. Moreover, users may check in multiple times at the same POI, which classical recommendation systems typically do not account for Nikolakopoulos et al. (2022). In POI recommendation, repeated check-ins to the same venue can serve as implicit information, refining the model of a user's preference similar to explicit ratings. As normally we do not have explicit information to create a user-item matrix, these repeated check-ins provide valuable implicit feedback. POI recommender systems capture latent user preferences using frequency matrices, providing better recommendations.

Popularity bias: Popularity bias is a well-studied problem in the recommender systems domain that occurs when popular items are recommended more frequently than less popular ones, regardless of whether they actually match the interests of the target user (Abdollahpouri et al. 2019). The effect of popularity bias is evident in multiple layers within the context of POI recommendation. Firstly, at the city level, an analysis of the original Foursquare data set reveals that out of the 415 cities worldwide, the top 1% of the most popular cities (based on the highest number of check-ins) represent the 20% of the total check-ins in the data set. However, when considering the top 2% of the most popular cities, this percentage increases to 28% of the total check-ins. Additionally, within each specific city, we can observe the impact of popularity bias on individual POIs. Taking New York City and Tokyo as examples, two extensively studied cities in the Foursquare data set (Sánchez and Bellogín 2022), we find that in New York City, the top 1% of the most popular venues are responsible for 27% of all check-ins, while the top 2% of venues account for 36% of the total check-ins. Similarly, in Tokyo, the top 1% of venues comprise 48% of all check-ins, and the top 2% represent 57% of the check-ins in the city.

2.3 Offline evaluation in point-of-interest recommendation

In offline evaluation of POI recommendation methods, most works follow the same protocols used in the traditional recommendation scenario. The data set is split into a training and test set with, occasionally, an additional validation set being used for model parameter tuning. All models are trained on this data, and for each user in the test set, a top-N list of recommendations is generated based on predicted user satisfaction (Cremonesi et al. 2010).

As in classical machine learning, subsets are often generated through random splits or cross-validation from the original data set (Cheng et al. 2016; Wang

et al. 2020b). However, recently, temporal dimension has been considered in these splits (Zhao et al. 2019; Huang et al. 2020). Currently, two main types of temporal splits are common: per user, where the n oldest interactions for each user are used for training, and the rest for validation and testing, and a global split, where interactions before a specific timestamp are used for training and the rest, for validation and test (Sánchez and Bellogín 2022). The latter is more natural, mimicking production-scale recommender system evaluation and avoiding data leakage to the test set (Ji et al. 2022).

Different metrics are used to evaluate recommendation algorithms, and most of them focus on measuring recommendation accuracy. This is determined by the overlap between recommended and actually visited venues in the test set. Greater overlap implies better recommendation quality. However, there is a recognition in the community that solely focusing on items in the ground truth may overlook other user-centric evaluation dimensions such as novelty (recommending non-popular items), diversity (recommending items that are different), and serendipity (recommending items that are novel and not easy to discover) (Castells et al. 2022). In the POI recommendation domain, which is influenced by categorical, geographical, and social factors, additional metrics can be used. For instance, category-level accuracy metric (Zhao et al. 2015) and the error in geographical distance between recommended and visited POIs (Yin et al. 2015) are used for measuring additional dimensions.

Another aspect that has received considerable attention in recent years in the recommender systems community, as in other areas of Artificial Intelligence and Machine Learning, is trying to understand the inherent biases learned by these systems and how they get reinforced by the recommendations. Thus, many researchers have focused on analyzing potential biases that may be found in either the data sets or the recommendations produced by the models. These biases vary widely, ranging from gender (Ekstrand and Kluver 2021; Melchiorre et al. 2021) to popularity bias (Abdollahpouri et al. 2019; Cañamares and Castells 2017). Moreover, whereas in the classical recommendation scenario it is currently established that analyzing different types of biases is important, the POI recommendation domain appears to lack comprehensive exploration in this regard. To the best of our knowledge, only a limited number of studies have analyzed this aspect. For example, Sánchez and Dietz (2022) observed biases in the recommendations provided to different groups of tourists and locals; and Weydemann et al. (2019) studied three types of fairness in this domain, i.e., fairness regarding the popularity of the venues, fairness with respect to the nationalities of the users, and an assessment if the recommendations are aligned with the category distribution observed in previous visits.

In light of the complexities and influences in the POI recommendation domain, there is little known about the impact of data characteristics on the recommendation performance. We address this gap using an explanatory framework, which was adapted to this domain from Deldjoo et al. (2021).

3 An explanatory framework for POI recommendation

The overall goal of this work is to understand which factors influence the performance of different recommendation models in the POI recommendation domain in terms of different evaluation dimensions such as ranking accuracy, novelty, and item exposure. Similar to previous approaches addressing this challenge (Adomavicius and Zhang 2012; Deldjoo et al. 2021), we use data characteristics to describe subsamples of a recommendation data set and use a regression model to capture the impact of each feature on the recommendation outcome. The idea behind analyzing different subsamples of the same data set is that the regression analysis would reveal the influence of variations in data characteristics on dependent variables. The main difference to previous studies is that our analysis is regarding POI recommendation, which enables us to define further explanatory variables capturing the geographical influences of users visiting venues in a city. Moreover, we are able to use a domain-driven subsampling approach, which yields additional insights into the performance of recommenders.

In this section, we describe the explanatory framework, which is a regression model applied to a series of data characteristics for capturing the interactions of users with the respective venues in a city. These data characteristics are computed for each subsample independently using a domain-driven approach outlined in the subsequent Sect. 4.

3.1 Regression model

Given all subsamples, we aim to model the relationship between the data characteristics and the recommendation performance of each individual recommendation algorithm. This allows us to test hypotheses regarding which explanatory variables are able to describe the variations in the dependent variables in a statistically significant way. Equation (2) shows the regression model, which is the core of the explanatory analysis.

$$y^r = \epsilon^r + \theta_0^r + \sum_{ev \in EV} \theta_{ev}^r x_{ev} \tag{2}$$

where ϵ is the error term (residuals), θ_0 is the intercept, i.e., the mean value of the dependant variable when the rest of the independent variables are zero, θ_{ev} is the regression coefficient of the respective explanatory variable ev (among the set of variables EV), x_{ev} represents the value of the explanatory variable in the current training example, and y is the value of the dependent variable according to the recommendation models. Since some of these values will depend on a specific recommendation model r, the notation shows this with a superscript. In particular, this means that, for a specific recommender system r, the value of the dependent variables will be potentially different for each r. Then, the performance would be modeled upon the set of explanatory variables x_{ev}, that will depend on the characteristics of the data set. Based on this, the regression model will produce coefficients θ_{ev}^r specifically tailored for this particular recommender, as it will consider the explanatory

variables and the dependent variables at the same time. When using the EVs within the regression model, we apply min-max normalization to obtain coefficients that are directly comparable.

3.2 Explanatory variables

We define 32 explanatory variables (EVs) which serve as independent variables in the regression analysis. Unlike the approaches of Adomavicius and Zhang (2012) and Deldjoo et al. (2021), we do not have a user-rating matrix, but a user-check-in matrix, since the interaction between the users and POIs is a visit and not a rating. While the user-check-in matrix (UCM) is conceptually very similar to a user-rating matrix (URM), the UCM is established based on unique visits, i.e., the cell values of the UCM are 1 if a user has visited a venue; otherwise, it is 0. Multiple visits of one user to the same venue also result in a value of 1. Despite this small conceptual difference, most EVs proposed and used by Adomavicius and Zhang (2012) and Deldjoo et al. (2021) can also be computed for a UCM. Since there is a significant geographic influence in POI recommendation (Li et al. 2015), we also propose some further EVs that capture such geographic information about the visited venues. Thus, the EVs we use can be categorized in the following four categories:

1. EVs that describe the structure of the UCM.
2. EVs that describe the check-in distribution of the UCM.
3. EVs that are based on item and user properties in the UCM.
4. EVs that capture the underlying user activity and mobility.

3.2.1 EVs based on the structure of the UCM

These EVs capture the general structure of the UCM and are well established to describe properties of recommendation data sets. Thus, we keep the discussion around them succinct.

Definition 1 (SpaceSize) Given a UCM, SpaceSize is defined as:

$$ev_1 = SpaceSize(UCM) = |U| \cdot |I| \tag{3}$$

We use the SpaceSize instead of its components, the number of users, $|U|$, and the number of items $|I|$ since it reduces the number of variables. As pointed out in Sect. 2.1, although we are dealing with POIs, denoted in that section as $\mathcal{P}$, we use the letter I to refer to items in general.

Definition 2 (Shape) Given a UCM, shape is defined as:

$$ev_2 = Shape(UCM) = \frac{|U|}{|I|} \tag{4}$$

The ratio between the number of users and the number of items can be an initial indicator of whether user-based collaborative filtering or item-based collaborative filtering approaches might be more successful (Nikolakopoulos et al. 2022).

Definition 3 (Density) Given a UCM, density is defined as:

$$ev_3 = Density(UCM) = \frac{|C|}{|U| \times |I|} \tag{5}$$

Density, or its inverse, sparsity, is a commonly reported metric to give an estimation of the recommendation difficulty for collaborative recommendation algorithms. Here we use C to refer to all the check-ins performed by the users and registered in a data set. Generally, the higher the density, the more signal is available for the algorithm to compute fitting recommendations. Density typically varies a lot depending on the data set and the domain, as mentioned in Sect. 2.2.

Definition 4 (Cp_u, Cp_i) Given a UCM, check-ins per user (Cp_u) and check-ins per item (Cp_i) are defined as:

$$ev_4 = Cp_u(UCM) = \frac{|C|}{|U|} \tag{6}$$

$$ev_5 = Cp_i(UCM) = \frac{|C|}{|I|} \tag{7}$$

The number of interactions per user/item is also a simple but effective measure to put the recommendation quality into perspective. If the number of interactions is remarkably small for a user or an item, it can be regarded as *"cold"*, indicating that there is not sufficient information to compute meaningful recommendations. Given the high sparsity of the data sets in the POI recommendation domain, it is very common to impose a minimum number of interactions for both items and users, i.e., enforcing a k-core, cf. Sect. 4.1. This is done to avoid evaluating cold-start recommendations.

3.2.2 EVs based on the check-in distribution of the UCM

Naturally, some users are more active than others, and not all items get the same attention. In the POI recommendation domain, this is a very natural phenomenon since major highlights inherently attract more visits. This is a major challenge with respect to several dimensions of the recommendation algorithms, including

accuracy, novelty, and fairness, as there is a delicate trade-off between recommending popular POIs and items in the long-tail (Rahmani et al. 2022).

Definition 5 (Gini_I, Gini_U) Given a UCM, $|C_i|$ and $|C_u|$ be the number of check-ins associated with item i and user u and the users/items are sorted according to C_i and C_u, respectively; then Gini_I and Gini_U are defined as follows (Deldjoo et al. 2021):

$$ev_6 = Gini_I(UCM) = 1 - 2\sum_{i=1}^{|I|} \frac{|I| + 1 - i}{|I| + 1} \times \frac{|C_i|}{|C|} \tag{8}$$

$$ev_7 = Gini_U(UCM) = 1 - 2\sum_{u=1}^{|U|} \frac{|U| + 1 - u}{|U| + 1} \times \frac{|C_u|}{|C|} \tag{9}$$

The Gini coefficient captures the frequency distribution of the check-ins for users or items. It is scaled between [0, 1], where a score of 0 would correspond to a uniform popularity distribution, and 1 to the extreme case of all check-ins being concentrated on one user/item.

3.2.3 EVs based on the item and user properties

The following two EVs, popularity bias and long tail items, have been motivated by Deldjoo et al. (2021) to be included in their explanatory framework. Arguably, the POI recommendation domain is even more severely impacted by popularity bias (Massimo and Ricci 2021; Sánchez and Dietz 2022), thus, it is imperative to include them in our framework. Given the distribution of the metrics with outliers, we augment the aggregation methods used in Deldjoo et al. (2021) (mean, standard deviation, skewness, and kurtosis) with the median value, as the median is more robust against outliers compared to the mean value. Note that a higher kurtosis is related to heavier tails, and hence, more outliers, while the skewness is related to the symmetry of the distribution. If the skewness is positive, it means that the right-hand tail is longer than the left-hand tail. If the skewness is negative, then the right-hand tail is shorter than the left-hand tail.

Definition 6 (Popularity bias) We follow the commonly accepted definition of popularity bias proposed by Abdollahpouri et al. (2019), which should not be confused with the popularity bias produced by the recommendation algorithm, as this applies to the bias that exists in the original data:

$$ev_{8:12} = f\left(\left\{\frac{\sum_{i \in C_u} \phi(i)}{|C_u|}\right\}_u\right) \tag{10}$$

where $\phi(i)$ is the popularity scoring function for an item i. The notation $\{\cdot\}_u$ aims to indicate that we iterate over the users and compute the value inside the brackets,

which is then processed by the outer function f. An item's popularity score is thus defined as the number of users who visited i over the entire number of users, and $|C_u|$ is the number of check-ins of user u. The term f is an aggregation operator over users to capture inter-user differences in the popularity profiles of users. They include average popularity bias (ev_8, APB), median popularity bias (ev_9, MedPB), standard deviation of popularity bias scores (ev_{10}, StPB), skewness popularity bias (ev_{11}, SkPB), and kurtosis popularity bias (ev_{12}, KuPB).

Definition 7 (Long tail items) Analyzing the popularity of items, they can be separated into a short-head and a long-tail.

$$ev_{13:17} = f\left(\left\{\frac{|i, i \in (C_u \cap \Gamma)|}{|C_u|}\right\}_u\right) \tag{11}$$

where C_u are, again, the check-ins of user u, Γ, on the other hand, represents long-tail items, which is determined by splitting the items into short-head and long-tail items. We define the split between short-head and long-tail in terms of the number of different users that have visited the item. In the literature, typical cutoffs for separating the short-head from the long-tail are at 20–80%, cf. Yin et al. (2012), Abdollahpouri et al. (2017), Deldjoo et al. (2021), which we also use in our experiments.

As before, f is an aggregation operator over users to capture inter-user differences in long-tail profiles of users. They include average long tail items (ev_{13}, ALT), median long tail items (ev_{14}, MedLT), standard deviation of long-tail items scores (ev_{14}, StLT), skewness long tail items (ev_{15}, SkLT), and kurtosis long tail items (ev_{17}, KuLT).

3.2.4 EVs based on the user activity and mobility

In this section, we introduce a family of data characteristics that are specific to the POI recommendation domain. The radius of gyration captures the size of a user's activity area. The distance from the city center is defined similarly, but the information it captures is more about how central the check-ins of a user are. Finally, the user's activity duration is interesting to include as well, as one can assume that the longer a user is within a city, the more familiar she becomes with the city, and it might result increasingly difficult for recommendation algorithms to propose interesting items.

Definition 8 (Radius of gyration) This is a common metric to capture the geographic extent of user mobility.

$$ev_{18:22} = f\left(\left\{\sqrt{\frac{\sum_{i \in C_u} \text{distance}((\text{lat}_i, \text{lon}_i), (c_u^x, c_u^y))}{|C_u|}}\right\}_u\right) \tag{12}$$

$$c_u = (c_u^x, c_u^y) = \left(\frac{1}{|C_u|}\sum_{i \in C_u} \text{lat}_i, \frac{1}{|C_u|}\sum_{i \in C_u} \text{lon}_i\right) \tag{13}$$

where c_u is the centroid of all the user's visited venues (González et al. 2008) and lat_i and lon_i represent the latitude and the longitude of item i respectively.

Again, f is an aggregation operator over users to capture inter-user differences in the radius of gyration of users. They include average radius of gyration (ev_{18}, ARG), median radius of gyration (ev_{19}, MedRG), standard deviation of radius of gyration scores (ev_{20}, StRG), skewness of radius of gyration (ev_{21}, SkRG), and kurtosis of the radius of gyration (ev_{22}, KuRG).

Definition 9 (Distance to city center) This EV is very similar to the radius of gyration, however, the center is not set to the centroid of the venues visited by the user, but to the center of the city. It is useful to differentiate between users who perform activities near the center of a city—typically of historic significance—and users who are more active in the outskirts.

$$ev_{23:27} = f\left(\left\{\sqrt{\frac{\sum_{i \in C_u} \text{distance}((\text{lat}_i, \text{lon}_i), (cc^x, cc^y))}{|C_u|}}\right\}_u\right) \tag{14}$$

where $cc = (cc^x, cc^y)$ is the geographic location of the city center. Again, we use different aggregation functions: the average distance to the city center (ev_{23}, ADCC), median distance to the city center (ev_{24}, MedDCC), standard deviation of the distances to the city center (ev_{25}, StDCC), skewness of distance to the city center (ev_{26}, SkDCC), and kurtosis of the distance to the city center (ev_{27}, KuDCC).

Definition 10 (Duration active) This EV is useful to provide insights into the effects of how long the duration of user activity was. In this context, users who have been active for a shorter duration may correspond to tourists, whereas those who have been performing check-ins for a longer period can be considered local residents of the city (Sánchez and Bellogín 2021).

$$ev_{28:32} = f\left(\left\{t(i)_l - t(i)_0 | i \in C_u\right\}_u\right) \tag{15}$$

where $t(i)_0, t(i)_l$ is the time of the first and last check-in of the items i the user u has visited, respectively. Note that in this case, we took the duration from all check-ins of the user, including repeated check-ins. Again, we use different aggregation functions: the average duration active (ev_{28}, ADA), median duration active (ev_{29},

MedDA), standard deviation of duration active (ev_{30}, StDA), skewness of duration active (ev_{31}, SkDA), and kurtosis of the duration active (ev_{32}, KuDA).

3.3 Dependent variables

For the dependent variables, we decided to analyze three different dimensions of the recommendations:

Ranking accuracy: We will focus on measuring how many recommended items actually match the ground truth of the user. For this purpose, we will use the nDCG metric (Järvelin and Kekäläinen 2002) that is defined in Eqs. (16) and (17).

$$\text{nDCG@k} = \frac{1}{U} \sum_{u \in U} \frac{\text{DCG}\ (u)@k}{\text{IDCG}\ (u)@k} \tag{16}$$

$$\text{DCG}\ (u)@k = \sum_{i_n \in RL_u@k} \frac{2^{rel_n} - 1}{\log_2(n+1)} \tag{17}$$

where RL_u is the recommendation list for user u, rel_n denotes the real relevance of item i_n, and k denotes the first k items of RL_u. In explicit rating data sets, this relevance value is normally bounded in the [0, 5] interval, with 0 representing a non-relevant value; in our experiments, as we do not have explicit ratings but check-ins, the relevance of the items appearing in the test set will always be 1. IDCG represents the ideal DCG, and it is computed in the same way as DCG but using the ground truth as the ranking. Higher values in nDCG mean that more relevant recommendations are being provided to the users; that is, more recommended venues are actually visited by the user according to the test set.

Novelty: The novelty of a recommendation can be assessed by measuring the proportion of popular venues being recommended. If a high percentage of the recommended venues are already well-known or frequently consumed/visited, it indicates that the recommendations lack novelty. To measure novelty, popularity is often used as a proxy, especially in offline evaluation where direct feedback from users is not possible. This is because it is generally assumed that whatever is popular within a community is likely to be known by most users and, thus, not novel. To measure this dimension, we use the Expected Popularity Complement (EPC)[2] metric (Vargas and Castells 2011), defined in the following equation:

$$EPC@k = \frac{1}{U} \sum_{u \in U} Z(u) \sum_{i_n \in RL_u@k} (1 - p(\text{ seen } \mid i_n)) \tag{18}$$

[2] Please, note that the original definition of the metric provided by Vargas and Castells (2011) also incorporates a discount model (as the one used in the nDCG metric) and a relevance model, in order to measure both the relevance and the novelty of the recommendations together. However, in our work, as we are evaluating ranking accuracy with nDCG, we will use the pure definition of EPC.

where RL_u is again the recommendation list for user u, $Z(u)$ is a normalizing constant (generally $Z(u) = 1/\sum_{i \in RL_u@k} 1$), and $p(\text{ seen } \mid i_n)$ represents the probability of item i_n to be consumed. This probability is estimated as $|U_i|/|U|$, that is, the number of users who have visited POI i in the training set, divided by the total number of users in the training set. Higher values in EPC imply that more novel recommendations are provided to the users.

Item exposure: In traditional recommendation domains, many algorithms tend to emphasize only a few items from the entire catalog (Liu and Zheng 2020; Abdollahpouri et al. 2019), leading to a popularity bias which we discussed in Sect. 2.3. This bias results in models favoring the most popular items, regardless of their relevance. In our study, to account for this effect, we measure item exposure by means of the so-called *expected exposure loss*, which slightly differs from plain popularity (Shih et al. 2016; Ekstrand et al. 2022). While popularity bias means recommending the most popular items without considering the distribution in the ground truth, poor performance with respect to expected exposure loss occurs when items are over- or under-represented in recommendations compared to the test set. Thus, we compare the number of times an item is recommended against the number of actual interactions in the test set (Ekstrand et al. 2022):

$$IE@k = \sum_{i \in I} \left| \frac{U_{test}(i)}{U_{test}} - \frac{Rec@k(i)}{U_{test}} \right| \tag{19}$$

where U_{test} denotes the number of users in the test set, $U_{test}(i)$ refers to the number of users that visited item i in the test set, and $Rec@k(i)$ is the number of times item i has been recommended considering all recommendations (i.e., rankings) until position k, i.e., at cutoff $@k$. As we are comparing the recommended exposure and the exposure of the items in the test set, the lower the values obtained in this metric, the better the performance of the recommenders in terms of item exposure.

4 Constructing subsamples with different data characteristics

To apply the explanatory framework described in the previous section, it is necessary to obtain several subsamples from a larger recommendation data set. We now discuss our approach to constructing various data sets with different characteristics that will serve as inputs for the explanatory framework. This aspect poses a challenge since the recommendation data sets, specifically the user-item interaction matrices, exhibit interdependencies that are complex to disentangle. The approach proposed by Deldjoo et al. (2020) constructs subsamples by selecting a random number of users and items while enforcing certain constraints, such as predefined data set densities.

As opposed to prior studies (Adomavicius and Zhang 2012; Deldjoo et al. 2021), we propose to use subsamples created by exploiting different data characteristics that are grounded in the domain instead of randomly sampling users with constraints. Generating the subsamples in such a way has two advantages: first, the

subsamples retain meaningful semantics, which provides additional analytic insights into the domain; second, it allows providers of recommender systems to understand in which real-world situations different recommendation models are advantageous or unfavorable. In the following section, we describe our proposed domain-driven subsampling procedure, which is based on subsetting a recommendation data set by factors that might be relevant to recommending points of interest within a city.

4.1 Data characteristics for creating domain-driven subsamples

Our design goals when developing the methods to construct subsamples are that (i) the subsamples must have a common basis to enable fair and meaningful comparisons, (ii) while at the same time, they show variability in terms of their resulting data characteristics. (iii) Further, we need to generate a sufficiently large number of subsamples to obtain robust results from the regression model, but (iv) each subsample should be a tractable recommendation problem, i.e., there is sufficient signal for the individual recommendation models to produce sensible recommendations.

As mentioned before, we take a different approach to construct subsamples compared to the approaches in the literature (Deldjoo et al. 2021, 2020). We formulate hypotheses regarding relevant factors that might have an influence on the outcome of POI recommendations. The core idea is to introduce a number of data characteristics relevant to the POI recommendation domain and use them as filters to include an interaction in the UCM or not. Thus, each data characteristic represents the explicit hypothesis that changing its value has an influence on the recommendation outcome.

The set of all subsamples is the cross-product of the data characteristics values applied to the original data set. This means that the generation of the subsamples is not only controlled by the outcome of the random processes that define the number of items and users in the interaction matrix but by meaningful subsetting of groups of users, items, or interactions. In the following subsections, we propose different data characteristics for subsampling, as the POI recommendation domain offers possibility for more complex hypotheses, since—unlike most classic recommendation data sets—there is the temporal (when a venue is visited) and the geographical (where the venue is and where the user is from) aspects to analyze. We leverage these aspects to formulate hypotheses along with common strategies employed in the evaluation of POI recommendation data sets to shape the recommendation outcome.

4.1.1 Enforce a minimum *k*-core

To mitigate the extreme sparsity of typical POI recommendation data sets, it is common to remove interactions from the UCM until all users and venues have at least k interactions. This is done to achieve a certain—higher—level of density in the UCM and, thus, fewer 'cold' items/users which usually results in higher accuracy metrics for interaction-based algorithms. Typical values for k are 2 (cf. Gao et al. 2015),

5 (cf. Yuan et al. 2013, 2014; Yao et al. 2015), or 10 (cf. to Nunes and Marinho 2014; Feng et al. 2015; Li et al. 2016).

4.1.2 Drop top *n*% popular venues

As previously discussed, popularity bias plays a large role in POI recommendation with a substantial interplay between the popularity of items and recommendation accuracy (Abdollahpouri et al. 2019; Massimo and Ricci 2021; Sánchez et al. 2023). As check-in-based data sets usually do not come with rating information, we limit the scope of popularity to the number of people that have visited a venue (Jannach et al. 2015).

The method to analyze the impact of popularity bias is to generate different subsamples by removing the most popular venues in the data set. We propose to drop the most $n\%$ popular items from the data set to obtain different distributions of the item popularity (Abdollahpouri et al. 2017). The concrete values depend on the data set at hand, but as a general guideline, we propose values for n between 0 and 5% for the specific point-of-interest recommendation domain after considering the popularity bias discussed in Sect. 2.2.

4.1.3 Filter by season

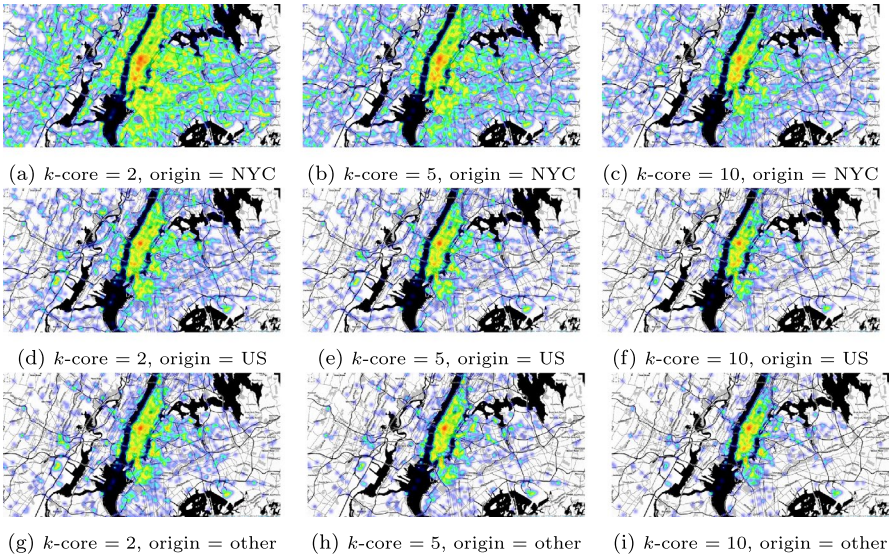
Seasonality is also a relevant factor that undoubtedly has an influence on the visited venues both by locals and tourists (Liu et al. 2011). The exact split between seasons can be tricky to make as a high granularity (e.g., weeks or months) can result in very small subsamples. Further, seasons are not the same in different regions, potentially requiring different segmentations for destinations in different climate zones (Trattner et al. 2018). In the context of an explainability study, we recommend using broad season categories, such as a two-season (warmer and colder months) or a four-season model.

4.1.4 Filter by user residence

In the context of POI recommendation, different groups of users exhibit different behavior (Sánchez and Dietz 2022). Due to the differences in behavior between locals and visitors, we argue that it is very promising to use such a division as a subsampling variable. If the information about the user's home is available in the recommendation data set, the typical groups to analyze would be the locals of the city, domestic visitors, or international travelers.

4.2 Discussion

In this section, we have formulated various hypotheses of what influences POI recommendations. These hypotheses are manifested in different data



(a) k-core = 2, origin = NYC (b) k-core = 5, origin = NYC (c) k-core = 10, origin = NYC

(d) k-core = 2, origin = US (e) k-core = 5, origin = US (f) k-core = 10, origin = US

(g) k-core = 2, origin = other (h) k-core = 5, origin = other (i) k-core = 10, origin = other

Fig. 1 Heat map of the visited venues in New York with different parameters for the k-core and the origin of the users (better viewed in color)

characteristics to enable a rigorous computational analysis. In the choice of data characteristics, we discussed the ones that we deem to be most relevant based on the literature on the analysis POI recommendation algorithms.

When setting up the experiments, it is still necessary to analyze the statistics of the resulting subsamples to understand which value ranges of the data characteristics to test. Here, it is important to retain tractable recommendation problems, i.e., not result in too sparse or small subsamples. Also, depending on the data set, it is not always possible to test all data characteristics discussed in Sect. 4.1. For example, removing resident users in a city that is not very active as a tourist destination could be counterproductive in making interesting recommendations that may attract more tourists.

4.3 Visualizing the subsampling variations

To exemplify the effect of subsampling variables, Fig. 1 visualizes the interplay of two data characteristics: the k-core and the origin of the users. In this heat map showing the density of check-ins in New York City, USA, the difference between the behavior of the locals and travelers becomes apparent: travelers tend to visit venues in Manhattan (with the exception of the airports), while the locals naturally have check-ins all over the map. A higher value for k-core leads to a higher density in the UCM but eliminates many venues, which can be observed in the map visualization.

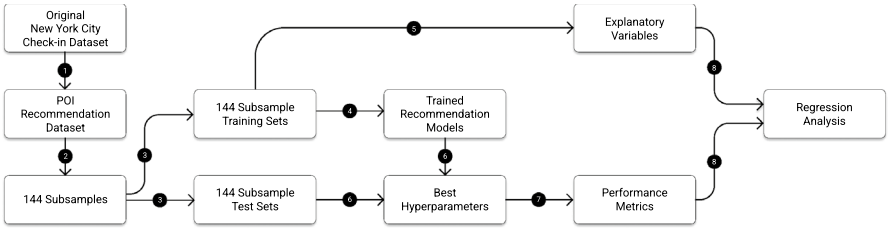


Fig. 2 Diagram representing the methodology followed in the paper. Each number corresponds to a step in the process: Initially, we clean the original check-in data set (1) to obtain a recommendation data set, which we subsequently subdivide into the 144 subsamples (2). Each subsample is further split into training and test sets (3), where a recommendation model is trained on each subsample individually (4) and the explanatory variables are computed based on the training sets (5). We determine the best hyperparameters of each recommender in each subsample using the test set (6), and record the metrics of the best performing recommendation configuration (7). Finally, we perform the regression analysis towards the performance metrics of each recommendation algorithm using the explanatory variables (8)

5 Experimental setup

In this section, we describe the used data set and the process of selecting the subsampling variables to obtain the subsampled recommendation data sets. We provide details about the data preprocessing, how we conducted the recommendation experiments, and give an overview of the outcomes. Finally, we outline the variable selection process for the regression model, which is the core of the explanatory framework. The full process explained in this section is shown in Fig. 2.

5.1 Selecting a suitable data set

To evaluate our proposed approach, it is essential to have a sufficiently large data set for POI recommendation that enables us to perform meaningful subsampling. Revisiting the literature (Bao et al. 2015; Sánchez and Bellogín 2022), we decided to use the Foursquare data set published by Yang et al., which has about 33 million check-ins in 415 different cities of the world (Yang et al. 2015) and has been frequently used to benchmark POI recommendation performance.[3] Although the data set contains check-ins from many cities, we conclude that it is infeasible to include multiple cities in the scope of the analysis. The number of check-ins per city is influenced by the number of people and the popularity of Foursquare in the city, which results in a few cities having a large number of check-ins but many not having sufficient interactions to further subsample them. Furthermore, the behavior of users is influenced by the topological realities of cities, such as centralized vs. decentralized cities. Therefore, each city needs to be analyzed separately to account for the different geographic influences.

In this study, we focus on New York City (NYC), NY, USA, as it is one of the most active cities on Foursquare and, hence, has been a common subject of analysis in the POI recommendation domain, e.g., in Albanna et al. (2016), Maroulis et al.

[3] Data set is available from https://sites.google.com/site/yangdingqi/home/foursquare-dataset.

(2016), Jiao et al. (2019). Concretely, the scope of our analysis is the New York City Metropolitan Area, which comprises the five boroughs of New York City and Newark, NJ. The complete data set from New York City Metropolitan Area consists of 17,467 users, 71,310 venues, and 608,131 check-ins. The geographical scope of the analysis is visualized in Fig. 1. To adapt the data set for the POI recommendation domain, we eliminated venues of the "Residences" category, as we do not consider them as interesting POIs to recommend. Finally, we also removed duplicated check-ins, i.e., check-ins at the same venue and identical timestamps.

5.2 Generation of subsampled recommendation data sets

In Sect. 4, we discussed subsampling data characteristics to be used in an explanatory framework for POI recommendation. In the context of this study we instantiated them as follows: We imposed a minimum k-core of the recommendation data sets, excluded varying levels of the most popular venues, and subdivided by season of the year and the origin of the user.

5.2.1 Subsampling data characteristics

UCM density Traditionally, density (i.e., the inverse of sparsity) has been a key metric to quantify the difficulty of a recommendation problem. The sparsity is normally referred to as the situation where most of the user-item interactions are not observed in the training data (Idrissi and Zellou 2020).

However, density is a dependent variable, which is typically adjusted by enforcing a k-core, i.e., requiring at least k interactions for each user and venue and discarding users and venues that do not fulfill this threshold.

Following the practice in literature, we create subsamples using the following values for k:

- **2:** Enforce a $k = 2$-core.
- **5:** Enforce a $k = 5$-core.
- **10:** Enforce a $k = 10$-core.

Item popularity Due to the large popularity bias in POI recommendation, we argue that it is important to analyze the effect of disregarding the most popular venues. We used the following values to analyze this effect:

- **0.5:** Drop the most popular 0.5% venues from the current data set.
- **1:** Drop the most popular 1% venues from the current data set.
- **2:** Drop the most popular 2% venues from the current data set.
- **5:** Drop the most popular 5% venues from the current data set.

Season The effect of seasonality on the recommendation outcome has not been analyzed in depth so far, providing us with the opportunity to analyze this within the explanatory study. The oceanic climate of New York City comes with relatively similar precipitation throughout the year, thus, we used the temperature aspect of the climate diagram to subdivide the year into the following groups:

- **all:** All check-ins irrespective of the season.
- **summer:** check-ins during the warmer months in New York City from May to October.
- **winter:** check-ins during the colder months in New York City from November to April.

By using only two groups, we hope to achieve a clear separation and keep the size of the resulting subsamples larger.

User origin The Foursquare data set contains check-ins from users from all around the world. Although we are only running the recommendation experiments with check-ins in New York City, we can use the complete data set to determine the home city of the people in the recommendation data set. To achieve this, we use the open-source tripmining library[4] to obtain the residence of the different users (Dietz et al. 2020). This library converts the check-in stream of users into periods of being at home and on travel. It uses the plurality strategy, i.e., selecting the city with the most check-ins as home city, to determine which is the user's home city, which has been shown to be accurate in a ground-truth study (Kariryaa et al. 2018).

We use this user home label as a subsampling data characteristic with the following values:

- **all:** considering all users.
- **US:** only domestic visitors from the United States, but not citizens of New York City.
- **NYC:** citizens of New York City.
- **other:** travelers from outside of the US.

The intuition behind using the users' home as a subsampling data characteristic is that the behavior of locals is different than the one of visitors. This also has a significant influence on the recommendation outcome, as shown by Sánchez and Dietz (2022).

[4] https://github.com/LinusDietz/tripmining.

5.2.2 Summary

Using these aspects, we generate 144 recommendation subsamples in the form of usercheck-in matrices. The result of the cross-product of applying the aforementioned subsampling data characteristics as filters on the original data set is 144:

$$\{\text{origin} = [\text{all, NYC, US, international}]\} \times \{\text{season} = [\text{all, summer,winter}]\} \times \{\text{k-core} = [2, 5, 10]\} \times \{\text{drop top venues} = [0.5, 1, 2, 5]\}$$

5.2.3 Training and test set generation

For each of the subsamples, we perform a temporal split per user, where 80% of the oldest interactions of each user are sent to the training set and the rest to the test set. Foursquare users might have performed check-ins at the same venue more than once, but the algorithms we use are meant to recommend new items, which means that we discard all duplicate check-ins of a user in the same venue both in the training and the test set. We decided to proceed like this in the test set because the goal of recommender systems is to recommend new venues to users to explore, not venues that the user already knows, which is common practice in the POI recommendation domain.

Table 6 (in the Appendix) tabulates statistics regarding the subsamples with the average values of different explanatory variables. For space reasons, we only tabulate the 12 subsampling data characteristics independently to get an impression of their individual impact. The experiments for the explanatory analysis used the cross-product of the subsampling data characteristics, resulting in 144 subsamples.

5.3 Algorithms for POI recommendation

In this section, we explain in detail the algorithms that we have used in our experiments. Due to the nature of the application of the different models, we will divide them into two main groups: classical recommendation algorithms and those specifically designed for point-of-interest recommendation.

5.3.1 Classical recommendation algorithms

Random: performs recommendations of venues randomly.
Pop: recommends to the target user the venues ordered by decreasing popularity. The popularity is measured by the number of different users that have visited that venue.
UB: user-based neighborhood. Non-normalized *k*-nn algorithm that recommends to the target user venues that other similar users visited before (Nikolakopoulos

et al. 2022; Aiolli 2013). We used the cosine similarity and Jaccard Index as similarity metrics.
IB: item-based neighborhood. Non-normalized *k*-nn that recommends to the target user venues similar to the ones that she visited previously (Nikolakopoulos et al. 2022; Aiolli 2013). We use the item variations of the same similarity metrics used for UB.
HKV: matrix factorization (MF) algorithm that uses alternate least squares for optimization proposed by Hu et al. (2008).
BPRMF: matrix factorization (MF) algorithm that uses the pairwise Bayesian personalized ranking loss proposed by Rendle et al. (2009) as optimization algorithm. For our experiments. we used the version from MyMedialite's[5] library.

5.3.2 Point-of-interest recommendation algorithms

IRENMF: weighted matrix factorization method proposed by Liu et al. (2014). This algorithm incorporates geographical information by assuming that users tend to visit neighboring venues (instance-level influence) and also by considering that the users check-ins are shared in the same geographical region (region-level influence).
GeoBPRMF: geographical Bayesian personalized ranking matrix factorization. Algorithm proposed by Yuan et al. (2016) that assumes that the target user will prefer to visit new venues that are close to the ones she visited previously.
RankGeoFM: a ranking-based matrix factorization model proposed by Li et al. (2015). They also incorporate the geographical influence in the recommendations by exploiting the neighboring venues (by geographical distance) with respect to the candidate POIs to recommend.
PopGeoNN: hybrid algorithm combining popularity (Pop), a user-based neighborhood method (UB), and a simple geographical component that recommends to the target user the venues closer to the average geographical position of all the venues visited by the user in the training set. This recommender has been used in previous works such as Sánchez and Dietz (2022). The final score is an aggregation of every item score provided by each recommender after normalizing its values by the maximum score of each method.

To achieve optimal parameters for each recommendation subsample, we systematically chose the optimal hyperparameters for each recommendation model by nDCG@5. We do this as it is a standard procedure in the area, despite the models could be optimized independently for each evaluation dimension, however, this would not be practical, as accuracy is typically understood as a first order-objective of any recommender system. The tested hyperparameter ranges are listed in the Appendix, Table 7.

[5] MyMedialite library: http://www.mymedialite.net/.

5.4 Recommendation results on subsamples

Using the recommendation models presented in Sect. 5.3 with the optimal parameters, we achieve the following recommendation outcomes regarding nDCG (Fig. 3a), EPC (Fig. 3b), and Item Exposure (Fig. 3c) in the 144 subsamples. We also report in Table 1 the average (denoted as avg) and the standard deviation (denoted as std) by each recommender in nDCG, EPC, and Item Exposure, all of them measured at a ranking cutoff of 5.

Table 1 reveals that the performance of the recommenders, in terms of nDCG, is relatively low. This is a common phenomenon in the POI recommendation domain, given the vast number of candidate POIs and the scarcity of user visits to train the models. However, some recommenders, like HKV or RankGeoFM, consistently show a lower performance overall, with limited standard deviation. With respect to novelty, we generally obtain high results, even in the Pop recommender. This can be attributed to the data sparsity, where although certain popular POIs have a significant number of visits, they have been explored by only a small percentage of users relative to the total number of potential users. Notably, the Pop algorithm exhibits the highest deviation, indicating that each subsample may feature different popular POIs. This also shows that the popularity bias is different in each subsample, being more exaggerated in those subsamples where we remove a smaller percentage of popular items (dtv = 0.5). This popularity bias impacts the performance of other recommenders with a bias towards popularity, such as BPRMF or PopGeoNN, while others like IB or RankGeoFM are less affected. In terms of Item Exposure, we can observe notable variations among the algorithms. While the Pop recommender achieves higher exposure results by focusing solely on recommending popular venues, other models such as RankGeoFM, UB, or IB can offer recommendations that encompass a more diverse range of POIs. The behavior of the IB recommender is particularly interesting: as derived from the results, it ranks fourth in terms of Item

Table 1 Mean value of the performance results of the recommenders in all subsamples

Family	Recommender	nDCG		EPC		Item Exposure	
		Mean	Std	Mean	Std	Mean	Std
Classic	Rnd	0.000	0.001	**0.999**	0.001	5.602	0.737
	Pop	0.004	0.004	0.987	0.009	7.820	1.430
	UB	**0.011**	0.006	0.997	0.003	6.128	1.230
	IB	0.010	0.005	**0.999**	0.001	6.137	1.026
	HKV	0.003	0.005	0.998	0.002	7.689	1.482
	BPRMF	0.010	0.008	0.994	0.005	6.618	1.223
POI	IRENMF	**0.011**	0.009	0.996	0.004	6.818	1.089
	GeoBPRMF	0.009	0.010	0.995	0.004	7.223	1.348
	RankGeoFM	0.003	0.004	0.998	0.002	**5.336**	0.708
	PopGeoNN	0.009	0.006	0.992	0.006	6.784	1.123

All values are reported at a cutoff of 5. We represent in bold the best value obtained when computing the mean in each metric

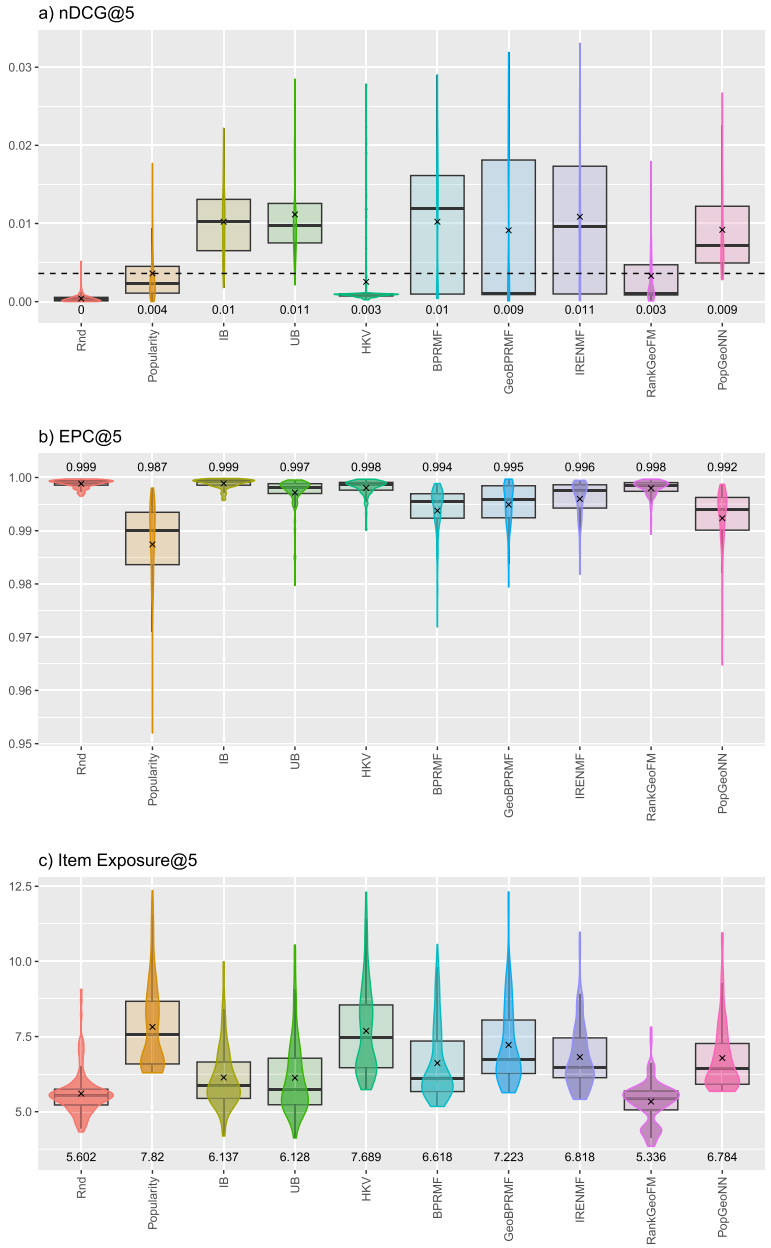


Fig. 3 The recommendation outcomes using the following metrics: **a** nDCG@5, **b** EPC@5, and **c** Item Exposure@5. The boxplot indicates the 25%, the median, and the 75% quantiles. Overlayed is a violin plot emphasizing the density of values and the overall range of the outcomes and the mean value with an x. The dashed line in the nDCG plot of Fig. 3a indicates the mean value of the Popularity algorithm

Exposure, highlighting its ability to provide fair recommendations for items appearing in the test sets; simultaneously, its performance in novelty stands out while it also obtains a competitive level of relevance, falling slightly behind the UB and IRENMF recommenders. However, both UB and IRENMF do not obtain results as competitive as IB in both Item Exposure and Novelty.

Visualizing the distribution of recommendation results in Fig. 3, we see that the recommendation accuracy varies substantially between the 144 subsamples, which is an expected and desired outcome (Fig. 3a). The random recommender is the one with the least variance and performance, however, the HKV matrix factorization model is also consistently low in performance, indicating that it can not deal well with many of the smaller subsamples. Furthermore, the Pop, IB, UB, and RankGeoFM models seem to be more robust regarding their outcomes compared to the purely matrix factorization-based approaches that do not consider the geographical component, as their interquartile ranges are smaller.

There are no surprises regarding the novelty of the recommendations, which we measure using the EPC metric, cf. Sect. 3.3. The models that involve some aspect of popularity (Popularity, PopGeoNN) and the BPR models produce relatively fewer novel recommendations, which is expected due to how the recommendations are computed (Fig. 3b). On the contrary, as mentioned before, the behavior of IB is interesting in terms of novelty since it is the second-best model (after Rnd) in all subsamples, obtaining relatively competitive accuracy results. This makes the IB a model that should be considered to try to achieve a balance between accuracy and novelty.

Finally, the Item Exposure (Fig. 3c) shows a low variation between most models, with the quantiles all being between 5 and 7.5. Again, the recommenders that generated fewer novel recommendations or exhibited a higher popularity bias, such as Pop, BPRMF, GeoBPRMF, or PopGeoNN, also achieved higher scores in terms of item exposure. This indicates a notable disparity in the distribution of recommended POIs compared to the POIs that the user has visited during the test set. This is consistent with previous work (Sánchez et al. 2023), where different biases are analyzed in the POI recommendation domain, and the effect on the performance of a set of recommenders in different cities around the world is compared.

The plots in Fig. 3 also tell much about the outcome of the subsampling process. With respect to the nDCG and the EPC metrics, the mass of the density plots is relatively compact, with some outliers to the top or bottom, respectively. In the Item Exposure plot, the shapes of the density plots are strung out, generally matching the interquartile ranges better. An interesting observation is that in the recommendation outcomes of the worse-performing models in terms of accuracy, i.e., Random, Pop, HKV, and RankGeoFM two groups become visible, i.e., the density plot looks tapered just below the mean value, meaning that the subsamples could be divided into 2 groups according to their performance. Even though a deep analysis of this aspect is out of the scope of this paper, it might be interesting to understand in the future which samples belong to each group and the impact the number (and frequency) of these groups may have on the explanatory power of the methodology followed in this work.

5.5 Excluding low-performing recommendation models from the explanatory study

The pure Popularity-based recommendation algorithm is a very simple, parameter-free model, but nevertheless a useful baseline in POI recommendation due to the inherent popularity bias of the domain (Bellogín et al. 2017). Even though computing the recommendations solely on the popularity of the items contradicts the principle of personalization, visitors tend to visit the popular highlights of a destination. Thus, we argue that any model in POI recommendation should at least outperform the Popularity model in terms of accuracy.

When it comes to the explanatory analysis, we remove the Random, HKV, and RankGeoFM models from the pool of algorithms for the explanatory study since the mean recommendation accuracy over all subsamples is lower than the simple Popularity baseline. The reason for this is that the purpose of the explanatory study is that we want to learn what the success factors of recommendation models are in terms of their data characteristics measured using EVs. By including models that are not "successful" (by outperforming the popularity baseline in terms of nDCG), we would analyze the factors contributing to poor recommendations, which is a meaningless endeavor.

This sets the final pool of 7 recommendation models to Pop, IB, UB, BPRMF, GeoBPRMF, IRENMF, and PopGeoNN.

5.6 Selecting relevant explanatory variables

In Sect. 3.2, we defined 32 potential explanatory variables that can be used to explain the dependent variables (Sect. 3.3) using the regression model. Naturally, not all independent variables possess equal levels of informative signal, i.e., they can be noisy. The regression model is useful in identifying noisy or unrelated independent variables as these will not be statistically significant coefficients with respect to the target variable. However, when using multiple variables in a regression model, multicollinearity can arise, i.e., two or more explanatory variables being correlated with each other. While this does not impede the outcome of the regression model, multicollinearity in the explanatory variables decreases the predictive contribution of the individual coefficients. To obtain meaningful results in the significance analysis of the coefficients, we mitigate collinearity by eliminating such redundant variables.

To do this in a reproducible way, we propose a procedure to remove highly correlated variables until the collinearity is mitigated to an acceptable level. The multicollinearity of a regression model is measured by the variance inflation factor (VIF), but the scientific literature is divided on what maximum VIF value is acceptable (Robinson and Schumacker 2009; O'brien 2007; Stine 1995). Reflecting on this, we systematically analyzed the outcome of applying Algorithm 1 with VIF thresholds between 5 and 25, ultimately choosing 12, which retains 8 EVs and explains on average $R^2 = 0.79$ of the variance in our regression models towards the nDCG@5. The choice of the VIF threshold is a trade-off between the number of variables, the

resulting variance the regression model can explain, and the level of multicollinearity, which any analyst or researcher must consider carefully on a case-by-case basis. Our proposed procedure is useful with an increasing amount of variables, as it removes human judgement from the process of choosing which variables to eliminate. This is an improvement on the previous work (Adomavicius and Zhang 2012; Deldjoo et al. 2020, 2021), where this step was not precisely specified.

Algorithm 1 Elimination of EVs based on correlation analysis

```
features ← {EVs};
while max(VIF(features)) > TR_VIF do
    correlation_matrix ← PCC(features);
    c1, c2 ← argmax( | correlation_matrix | );
    if max(| PCC(c1, { features \ c2 })|) > max(| PCC(c2, { features \ c1 })|) then
        features ← features \ c1 ;
    else
        features ← features \ c2 ;
    end
end
return features;
```

The goal is to determine a set of input variables for the regression analysis that have low collinearity, which is measured by the variance inflation factor (VIF). To determine variables that cause unwanted collinearity, a correlation analysis is required to discard correlated variables. Algorithm 1 describes our proposed procedure: while there is still an EV with a VIF over the threshold (TR_VIF), we compute all pairwise Pearson correlation coefficients (PCC) of the features, obtaining the correlation matrix, and determining the two different features with the highest positive or negative correlation as our candidates for elimination. From these two candidates (c_1, c_2), we eliminate the candidate that has the highest correlation to any other feature in the remaining features. This elimination of EVs is repeated until the VIF values of all remaining EVs satisfy the threshold.

Table 2 shows the outcome of applying Algorithm 1 to the data from the experiments comprising the recommendation outcome of the well-performing recommenders as established previously in Sect. 5.5 on the 144 subsamples. The target metric of the linear model to compute the VIF was nDCG@5. We retain 8 EVs, namely shape, density, Gini_U, StPB, KuPB, StRG, MedDA, and KuDA. This outcome is interesting as it puts emphasis on the EVs that capture the structure of the user-check-in matrix, such as shape, and density. This is not surprising, as these are very common metrics to quantify the difficulty of a recommendation problem. Furthermore, some aspect of most families of EVs was included, with the exception of the distance to city center and long-tail items. While we will come to the explanatory power of the EVs in the following section, this result alone underlines that the newly introduced EVs regarding mobility and user activity broaden the perspective of POI recommendation problems. We plot the pairwise correlations of the EVs in the Appendix, Fig. 7, where we find that the maximum (in absolute terms) pairwise PCC is −0.69 between Gini_U and density, after removing highly correlated variables.

Table 2 Final EVs after controlling for multicollinearity

EV	VIF before	VIF after
SpaceSize	25.90	–
Shape	71.79	1.51
Density	30.95	5.10
Cp_u	69.71	–
Cp_i	304.45	–
$Gini_I$	103.34	–
$\mathbf{Gini}_U$	130.78	6.13
APB	9138.49	–
MedPB	6802.78	–
StPB	316.46	3.27
SkPB	31.19	–
KuPB	15.47	1.36
ALT	519.42	–
StLT	661.18	–
SkLT	567.80	–
KuLT	449.43	–
ADCC	5711.40	–
MedDCC	1229.74	–
StDCC	1731.52	–
SkDCC	354.85	–
KuDCC	225.33	–
ARG	2749.51	–
MedRG	561.53	–
StRG	129.15	11.32
SkRG	1002.16	–
KuRG	204.44	–
ADA	561.36	–
MedDA	66.90	2.77
StDA	52.75	–
SkDA	429.65	–
KuDA	146.96	4.95

6 Results

Recall from Sect. 3 that in its core, the explanatory framework is a linear regression (cf. Eq. 2) with the data characteristics of the 144 subsamples quantified in the explanatory variables as input variables and nDCG, EPC, and Item Exposure as dependent variables. What we are interested in are the coefficients of the model (θ_{ev} in Eq. 2), as these coefficients quantify the impact of the individual EV on the outcome variable. We run the explanatory framework for each recommendation model that produced competitive results (cf. Sect. 5.5) independently using the EVs that did not suffer from multicollinearity (cf. Sect. 5.6).

In the analysis of the results, we assessed three key aspects: Accuracy, Novelty, and Item Exposure. Our findings revealed notably high R^2 values for each of these aspects. This suggests that, even after discarding numerous explanatory variables, our regression models still have substantial explanatory power. In this section, we make a detailed discussion of our results, using a cutoff value of 5 for each metric (i.e., we evaluate the top 5 recommendations), as in the POI recommendation domain, it is common to report small cutoffs (Liu et al. 2014; Li et al. 2015; Yuan et al. 2016). For additional results at cutoff values of 10 and 20, please refer to Appendix C.

Our structure for presenting the experimental results follows this format: firstly, we tabulate the regression coefficients (θ_{ev}) and then visualize these coefficients through coefficient plots. The result tables, i.e., Tables 3, 4, and 5, provide an overview of the goodness-of-fit with the R^2 and adjusted R^2 values. Subsequently, we present the coefficients θ_{ev} for each model. These coefficients are annotated with stars, which indicate the significance level of the relationship between the explanatory variable and the outcome variable. The significance levels and the corresponding p-values used are *** to indicate that $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

Furthermore, we visualize the values of the tables in coefficient plots (Figs. 4, 5, 6). The dots show the coefficient θ_{ev}; the whiskers span the 95% confidence interval.

6.1 Accuracy

First, we analyze the recommendation accuracy in terms of nDCG in Table 3. The R^2 coefficients of determination of the regression models are all between 0.68 and 0.88, indicating that the 8 EVs could explain 68%–88% of the nDCG@5 variation. This result is consistent with the two previous studies of Adomavicius and Zhang (2012) and Deldjoo et al. (2021). The least predictable algorithm was the IB recommendation model, while the GeoBPRMF algorithm had the highest R^2.

Table 3 Coefficients of the regression model for nDCG@5

Name	Popularity	IB	UB	BPRMF	GeoBPRMF	IRENMF	PopGeoNN
R^2	0.768	0.698	0.72	0.845	0.886	0.798	0.799
R^2 (adj)	0.754	0.68	0.704	0.836	0.879	0.786	0.787
Shape	− 0.003***	0.002	− 0.002	0.001	0	0.003*	− 0.003*
Density	0.006***	0.009***	0.009***	0.018***	0.019***	0.009***	0.011***
$Gini_U$	− 0.001	0.015***	0.014***	0.015***	0.012***	0.006	0.003
StPB	0.005***	0.003	0.005*	0.006***	0.011***	0.007**	0.009***
KuPB	− 0.002	0.002	0.002	0.005***	0.006***	− 0.004*	− 0.002
StRG	− 0.006**	− 0.009**	− 0.014***	− 0.009**	− 0.009***	− 0.018***	− 0.01**
MedDA	− 0.004***	0.01***	0.003*	0.011***	0.01***	0.01***	− 0.001
KuDA	0.001	0.011***	0.011***	0.004	0.002	0.009**	0.012***

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

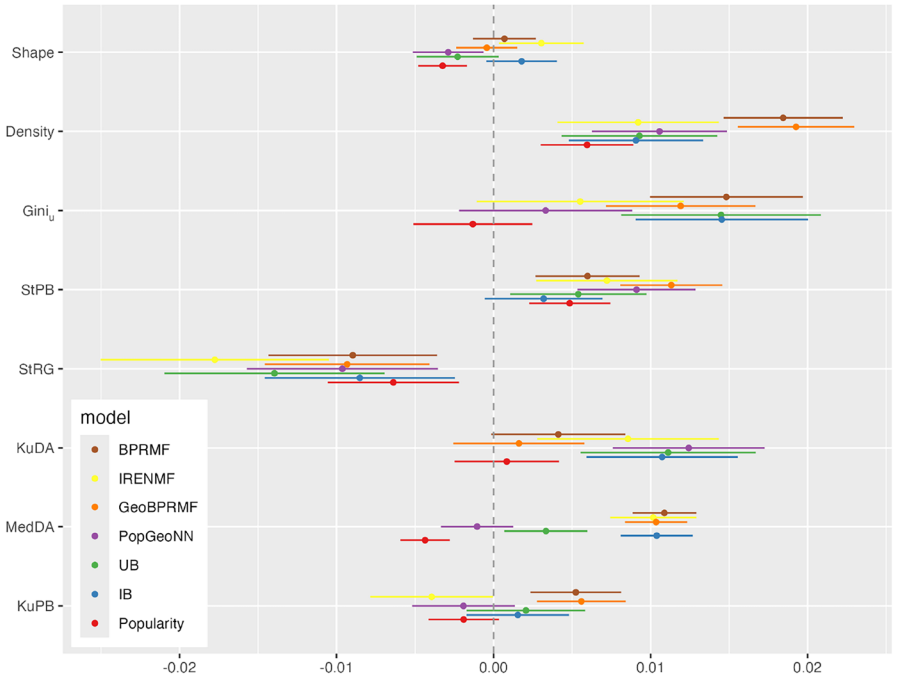


Fig. 4 Coefficient plot for nDCG@5. StRG has a negative influence on nDCG@5, shape, KuPB, and MedDA are mostly neutral, whereas the other EVs have a positive impact

When it comes to the general patterns in the coefficients (cf. Fig. 4), we observe that an increase in the standard deviation of the radius of gyration has clearly the most negative influence on the recommendation accuracy of all variables. On the contrary, a higher density, Gini_U, standard deviation of the popularity bias, and kurtosis of the duration active generally tend to improve the accuracy. The EVs shape, median of duration active, and the kurtosis of the popularity bias have mixed influences, with shape overall having the smallest (but still statistically significant in three cases) influence.

Turning our attention to the significance levels of the coefficients regarding the outcome variable, we observe that density and StRG are always highly significant with $p < 0.01$ for all recommendation models. These consistently low p-values across the board of all recommendation models underline their importance for the success of POI recommendation algorithms. All EVs were a significant predictor towards the nDCG@5 in some of the recommendation models, although the shape was only significant towards the accuracy of the popularity and the PopGeoNN models.

Table 4 Coefficients (θ_{ev}) of the regression model for EPC@5

Name	Popularity	IB	UB	BPRMF	GeoBPRMF	IRENMF	PopGeoNN
R^2	0.88	0.953	0.803	0.769	0.749	0.861	0.869
R^2 (adj)	0.873	0.951	0.791	0.755	0.734	0.852	0.861
Shape	0.01***	0	0.003***	0.004***	0.006***	0.004***	0.006***
Density	− 0.024***	− 0.004***	− 0.008***	− 0.013***	− 0.017***	− 0.01***	− 0.017***
$Gini_U$	0.009**	0	0.001	0.009***	0.007*	0.006***	0.006**
StPB	− 0.021***	− 0.001***	− 0.004***	− 0.01***	− 0.013***	− 0.011***	− 0.014***
KuPB	− 0.004*	0	0.002	0.003	0.002	0.001	0
StRG	0.005	0	0.003	− 0.001	− 0.001	− 0.001	0.002
MedDA	0.004**	0	0.003***	0.004***	0.004**	0.001*	0.003***
KuDA	− 0.002	− 0.001***	0.001	0	− 0.005	− 0.003*	− 0.002

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

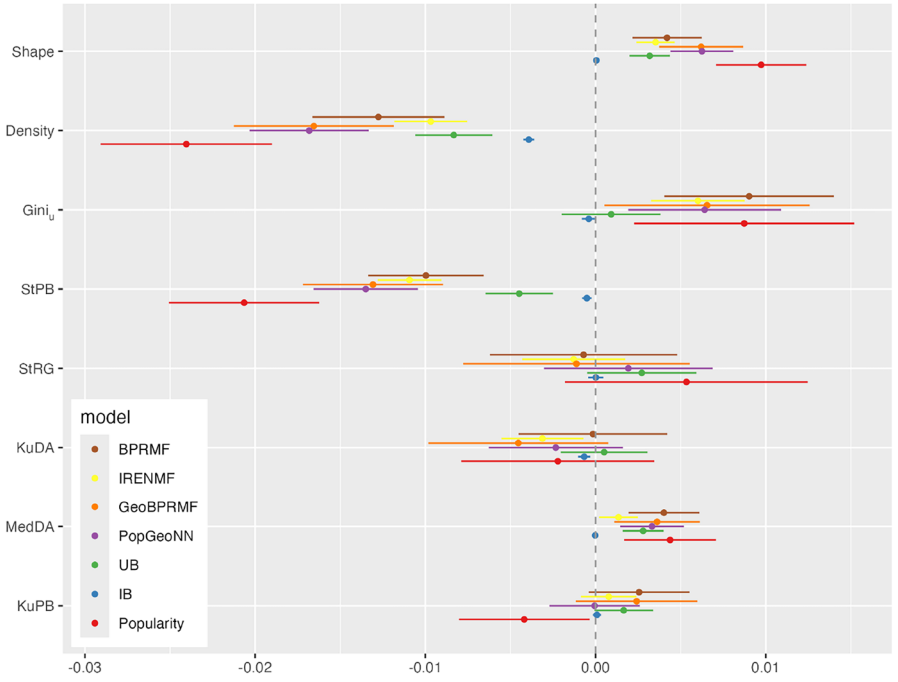


Fig. 5 Coefficient plot for novelty measured using EPC@5. We observe an inverse relationship compared to nDCG @5 with density, StPB, and KuDA having a negative impact and shape, $Gini_U$, and MedDA having a positive impact on the EPC metric

6.2 Novelty

The second aspect of our analysis is novelty, which we measure using the EPC metric. Again, we tabulate the coefficients in Table 4. Comparing the R^2 values to

Table 5 Coefficients (θ_{ev}) of the regression model for average item Exposure@5

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.932	0.907	0.902	0.873	0.874	0.666	0.917
R^2 (adj)	0.928	0.902	0.896	0.866	0.866	0.646	0.912
Shape	− 1.609***	− 2.631***	− 2.421***	− 1.924***	− 2.165***	− 1.515***	− 2.347***
Density	0.547	− 0.069	0.273	0.36	0.106	0.561	− 0.361
$Gini_U$	1.813***	1.302***	1.052*	0.839	0.351	− 1.486*	2.126***
StPB	1.573***	0.985***	1.222***	1.232***	1.743***	2.26***	0.814**
KuPB	− 0.514*	− 0.264	− 0.326	− 0.889***	− 0.525*	− 1.418***	− 0.015
StRG	0.688	0.193	0.477	0.934	1.371**	2.249**	0.849*
MedDA	3.388***	1.719***	2.595***	2.553***	1.988***	1.503***	2.23***
KuDA	0.496	0.019	0.422	1.348***	1.583***	2.517***	0.563

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

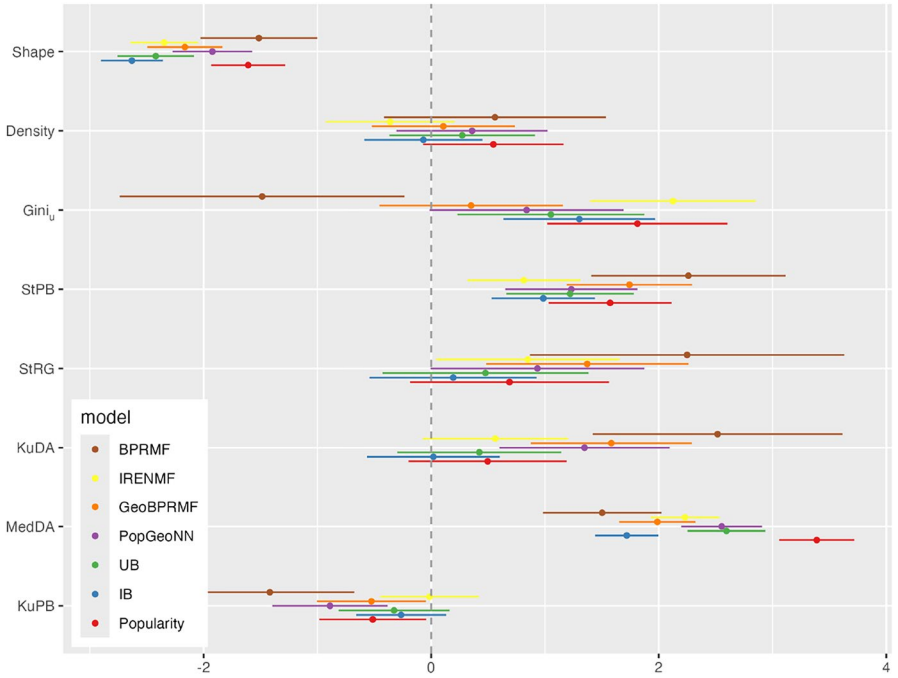


Fig. 6 Coefficient plot for average item Exposure@5. Shape and KuPB exert a negative influence on average item exposure, meaning that the difference to the expected item exposure is smaller, which is better. StPB and MedDA have a clearly positive influence on the item exposure metrics, whereas the remaining EVs are mostly slightly positive or neutral (density)

the ones in Table 3 (accuracy), we see a slightly better regression fit with values ranging from 0.73 (GeoBPRMF model) to 0.95 (Item-based model). This indicates that despite the fact that the explanatory variables were selected based on their

collinearity with respect to the nDCG@5 (cf. Sect. 5.6), the regression model for the EPC metric is similarly accurate and even slightly more expressive compared to the one for nDCG@5.

Analyzing the patterns within the coefficients presented in Fig. 5, we can discern a noteworthy observation: our experiments reveal an inverse relationship between novelty and accuracy, which is a well-known trade-off in recommender systems. For all considered EVs, except for the statistically non-significant StRG and KuPB (except in the Pop model), and across various recommender models a distinct change in the sign of coefficients is evident. Specifically, many coefficients switch from a positive association to a negative one and vice versa. The coefficients for the Item-based model converge near zero, indicating that this model produces recommendations with a stable Novelty regardless of the data characteristics. Among the EVs, the shape, Gini_U, and MedDA display positive coefficients for EPC@5. Conversely, the other EVs exhibit a negative impact (density, StPB) or a neutral influence (StRG, KuDA) on this particular outcome variable. The observation of a neutral influence of the standard deviation of the radius of gyration is bolstered by the finding that its coefficients do not achieve statistical significance across any of the recommendation models. In contrast, we note that certain EVs exhibit a high level of significance ($p < 0.001$) across all recommendation models. Specifically, these influential EVs are density, shape, StPB, and MedDA.

6.3 Item exposure

Lastly, we shift our focus to the assessment of item exposure, as measured by the metric defined in Sect. 3.3. Notably, the R^2 values for this analysis are remarkably high, exceeding 0.86 in all cases, except for BPRMF, where the EVs can still account for 0.65 of the variance, as detailed in Table 5. We would like to emphasize that higher values in the item exposure metric indicate a larger disparity between the number of times items are recommended and the number of times they should ideally be recommended (as defined in Eq. 19).

The coefficients that exhibit high levels of significance across all recommendation models include shape, StPB, and MedDA. In contrast, density fails to achieve significance in any of the models. Lastly, the significance of the coefficient for Gini_U in the IB and UB models is noteworthy as this EV quantifies the inequality in the frequency distribution of item check-ins.

Upon analyzing the grouping of coefficients and models in Fig. 6, again certain patterns emerge: shape and KuPB consistently exert a negative influence on Item Exposure within all recommendation models. As their values increase, the item exposure metric decreases, which means that the item exposure is closer to the expected item exposure from the test set. This observation is in line with the intuition that a relatively higher number of check-ins within the user-check-in matrix results in a more dispersed distribution, which in turn helps mitigate the influence of popularity bias. Conversely, the median of the duration active (MedDA) consistently exhibits a positive impact on the item exposure metric across all the recommendation models. One plausible explanation for this trend is that when users spend longer

periods in a city, they are more likely to explore and visit a large number of popular POIs, further accentuating the effect of popularity bias. The other EVs typically have an overall positive influence on the Item Exposure metric, with many (but not all) being significant predictors.

7 Discussion

The obtained results shed light on the strengths and weaknesses of POI recommendation models in terms of the data characteristics of the recommendation problem. Using the domain-driven subsampling approach to create specific subsamples from a large POI recommendation data set, we noticed that several established POI recommendation models are unsuited for smaller problems (cf. Fig. 3a), which warranted their exclusion from the explanatory analysis.

In terms of the quality of the linear model, it is striking that it was possible to explain 68–88% variation of the nDCG with the remaining 8 out of 32 EVs after the elimination of the collinear EVs. Besides, we also obtained high results in terms of explaining item exposure (65–86%) and novelty (73–95%) variations. This result confirms that the collinearity analysis is necessary and helps to focus on the interesting variables without losing explainability. In terms of accuracy, we find a clearly positive influence of density on the nDCG@5, which confirms the findings from other domains that a higher density creates an easier recommendation problem (Deldjoo et al. 2021). Furthermore, we provide evidence that a higher standard deviation of the radius of gyration leads to diminished accuracy. This shows that geographic information is determinant in predicting the results in terms of ranking accuracy in POI recommendation. In this scenario, an increase in the standard deviation of the radius of gyration suggests a greater diversity in user movement patterns, making it more difficult (for algorithms) to identify consistent global movement trends among users.

In terms of novelty of recommendations, we see a similar trend as for the accuracy; however, our results showed once more that these two concepts are inverse due to the popularity bias and related trade-offs discussed in the literature (Kaminskas and Bridge 2016; Zhao et al. 2019). Unsurprisingly, we could show that a higher variance in the popularity bias in the interaction data helps to promote novel recommendations. Thus, these two targets still need to be balanced in any POI recommender system in accordance with business needs. Emphasizing these dimensions is also important from the user's point of view, as it would allow the system to surprise them with recommendations that are different from what they may be familiar with beforehand.

Regarding Item Exposure, we observe negative coefficients of the shape (which is the ratio between the number of users and items) on this metric, signifying that relatively more users compared to items tend to yield lower values of the item exposure metric. In this context, lower values are preferable, as it means that they are closer to the exposed values in the unbiased test set. However, we should also consider that an increased duration active and standard deviation of the popularity bias leads to a higher item exposure than desired, as this variable always obtains positive

coefficients in the regression model. Platforms need to monitor such effects closely, as the exposure of small local travel enterprises in the recommendations can determine whether they have a means of existence in the market.

7.1 Practical and theoretical implications

In this paper, we go beyond evaluating recommendation models by employing a framework that reveals the associations between various data characteristics and different dimensions of performance. While this framework is not without its imperfections, as discussed below, its application carries significant implications for the training and evaluation of POI recommender systems. For instance, our observations highlight the impact of explanatory variables such as Density, Gini_U, and StPB which have a positive influence on nDCG@5, as opposed to EVs like StRG, which displays a negative influence. Based on our findings, developers, analysts, and researchers of POI recommender systems should be aware of the importance of these data characteristics. In practice this means that countermeasures against adverse data characteristics can be undertaken: The density can be addressed by defining a minimum number of interactions before a user is served by the main recommendation model (for *"warm"* users) instead of computing cold-start recommendations to a user with too little interactions. By directly asking *"cold"* users for venues they have visited, relevant information can be collected. At the same time, the geographic scope of candidate items can be adjusted to deal with the adversary effects of the radius of gyration.

Furthermore, it is worth noting that the effect of these explanatory variables is dependent on the specific recommendation algorithm employed; however, we see systematic common effects on related recommendation models. Nevertheless, researchers should be mindful of this when comparing baseline methods against proposed models. Unintentionally, they might report comparisons that are not fair because some methods are either positively or negatively impacted by the data characteristics. Notably, this observation extends to beyond-accuracy metrics, as we have observed similar trends in novelty and item exposure.

From a more theoretical perspective, this work introduces the prospect of learning these data dependencies directly from the data themselves and incorporating them into recommendation models and user profiles. Specifically, our work gives guidance to the choice of models for recommendations in automated ways, e.g., through Auto-RecSys (Anand and Beel 2020; Vente et al. 2023), analogous to AutoML (Karmaker et al. 2021). As our approach is entirely data-driven, the only ad-hoc decisions we made were related to the selection of the original explanatory variables, which could vary depending on the specific target domain. However, once the explanatory and dependent variables are established, it becomes conceivable to integrate the methodology presented here into a reinforcement learning framework. In such a setup, data could be fed to the recommenders in accordance with the predicted effects they are anticipated to have, thus optimizing recommendation outcomes, as addressed in some works based on specific user characteristics to improve overall performance (Said and Bellogín 2018; Penha and Santos 2020).

Even in the absence of a reinforcement learning framework, the current methodology can provide valuable insights into designing and enhancing existing recommendation algorithms, at least in the POI recommendation domain. For instance, we observed that IB is less sensitive to density than GeoBPRMF across the three metrics analyzed. Given that their accuracy metrics are similar, one way to leverage this insight is to explore how to make GeoBPRMF more resilient in scenarios where data sets exhibit varying levels of density. To achieve this, researchers and practitioners could consider incorporating strategies from IB or similar robust methods into an improved version of the GeoBPRMF algorithm. This integration might involve adapting the underlying model or introducing additional mechanisms that allow GeoBPRMF to handle data sets with different density levels more effectively. By doing so, GeoBPRMF could become more versatile and capable of delivering consistent performance across a wider range of data set characteristics, enhancing its practical applicability in diverse POI recommendation scenarios.

Most importantly, whether it is hotels, restaurants, or other POIs, this work gives guidance to recommendation platforms of the e-tourism sector to characterize their recommendation data and understand the benefits and drawbacks of their recommendation model from the perspective of different users groups and businesses. It can also inform methods to self-audit biases in the recommendations of platforms, e.g., with regards to the expected and achieved item exposure of items in certain recommendation models (Srba et al. 2023).

7.2 Limitations and future work

While being a widespread drawback of offline analyses of POI recommendation algorithms using location-based social network data, we still acknowledge that there is a gap between actual user behavior and what is recorded in LBSNs. Although this is universal for all recommendation models in all studies, it is worth mentioning that such studies analyze a proxy concept of check-ins that have been actively submitted instead of the actual ground truth of the temporal visitation of all POIs, venues, and other places. Thus, the generalization of the analyses in this study is—just as in all other studies— constrained by the limited availability of high-quality POI recommendation data sets. This limitation can only be overcome by collecting data sets that reflect user interactions from actual POI recommender systems, which would be a crucial next step in advancing the field.

Although we utilized the widely employed global Foursquare data set containing 33 million check-ins (Yang et al. 2015), its sparsity posed a significant challenge since the actual number of check-ins per destination dwindled, limiting our ability to create meaningful subsamples for most of the cities within the data set. To prevent biases stemming from geographic influences of the topological features of different cities, we focused our study on a single destination, the New York City metropolitan area, due to its prominence and the number of check-ins in the data set. This decision to focus on one city was essential because many less popular destinations on Foursquare lacked the volume of check-in data to support the explanatory analysis with subsamples of sufficient size, which would lead to a deterioration in

recommendation quality and lower statistical significance. In fact, while New York City provided ample interaction data for addressing the recommendation problem, it is a city less susceptible to seasonal variations in travel behavior, as evidenced by the minor differences in data characteristics between the summer and winter subsamples shown in Table 6, hence the obtained results cannot be considered *universal*, since this analysis in only based on one data set from one city. As future work, analyzing the generalizability of the explanatory framework *between* cities and to further data sources would be an obvious extension to our work. If one would repeat the experiments in many cities, how robust would the set of *significant* explanatory variables (i.e., the coefficients in the respective linear regression models for the individual algorithms) be? Such an analysis would be interesting, however, it would involve computing explanatory variables whose values are comparable, i.e., normalized between different cities. Generalizing the method towards incorporating different data sources would be a further step to understand the impact of different data collection methods on the data characteristics and the performance of POI recommender systems. Herein, the challenge lies in establishing comparable check-ins data sets, both in terms of geographic and temporal coverage.

In this context, another distinct line of future work would be to compute a *universal* regression model, with the aim to *predict* the performance of recommendation algorithms in different cities. The relevant research questions are two-fold: (a) under which circumstances is it permissible to mix data characteristics of samples from different cities, as they have different sizes, which might result in incomparable values of the same explanatory variables?, and (b) are other influences, such aspects that *cannot* be directly quantified as explanatory variables, like cultural or climatic aspects, small enough so that they do not have a practical influence on the predictions?

In our study, we incorporated EVs capturing both spatial and temporal aspects, additional variables capturing more of the users' context were omitted for the subsampling for practical reasons. With the currently used data sets, a finer temporal subdivision of check-ins by different times of the day or weekdays and weekends would be possible in theory but are practically prevented by the size of the resulting subsamples. While further contextual information is not available in the data set, the performance of a POI recommendation is known to be influenced by many contextual factors, such as the weather on a given day (Trattner et al. 2018). User context has been the focus of various approaches in literature (Wörndl et al. 2011; Cai et al. 2017; Yang et al. 2017; Zhao et al. 2019), this gives opportunity for future works to analyze the importance of different contextual factors, including more information about the users and their needs, which has been analyzed in literature on traveler types (Gibson and Yiannakis 2002; Neidhardt et al. 2014; Dietz et al. 2020), which offers various roles that one could use as subsampling data characteristics. Specifically, since it has been shown that locals and tourists showcase different behavior in the same city (Sánchez and Dietz 2022), it would be worthwhile identifying which explanatory variables are more relevant for various user groups.

8 Conclusions

In both the classical recommendation domain and specialized fields, such as point of interest recommendation, a multitude of algorithms have been proposed. However, the effectiveness of these algorithms often varies significantly depending on the data set under evaluation. While addressing this challenge has been explored previously in classical recommendation scenarios, there remains a research gap within the POI recommendation domain. This gap is particularly significant, given the relevance of POI recommendations in the tourism sector, affecting a multitude of businesses and consumers. Unlike the traditional recommendation domain, like purchasing a book or watching a movie, recommending POIs involves algorithms that utilize various signals, such as popularity and geographic locations of venues.

In this paper, we expanded upon the framework introduced by Deldjoo et al. (2021), incorporating additional explanatory variables specific to the POI recommendation domain. Our objective was to investigate which data characteristics affect the performance of both classical and state-of-the-art POI recommendation algorithms. We assessed these algorithms across three dimensions: accuracy, novelty, and item exposure. The results we obtained shed light on the robustness of recommendation models concerning the data characteristics inherent in recommendation data sets, offering valuable insights for the field. Among the various explanatory variables we analyzed, it became evident that certain factors pertaining to the data structure (shape, density), as well as those associated with the distribution of check-ins (Gini_U, StPB, and KuPB), along with geographical and temporal variables (StRG, MedDA, and KuDA) are critical for explaining the performance of POI recommender systems. In terms of ranking accuracy, density, Gini_U, StPB, and KuDA generally tend to be conducive to higher accuracy, whereas an increase in the standard deviation of the radius of gyration is detrimental. The significance of newly introduced spatio-temporal explanatory variables (radius of gyration and duration active) in the coefficient analysis of the regression models underlines our conclusion that the well-known data characteristics from the analysis of classical recommendation domains are insufficient to explain algorithmic performance in the POI recommendation domain.

Appendix

A Statistics of subsamples

Table 6 presents the values obtained by independent subsamples in the explanatory variables defined in Sect. 3.2.

Table 6 Summary statistics of the impact of the subsampling on the explanatory variables

Subsample	SpaceSize	Shape	Density	Cp_u	Cp_i	$Gini_U$	$Gini_I$	APB	ALT	ARG	ADCC	ADA
k = 2	5.57E+08	0.36	4.45E−04	17.44	6.32	0.62	0.66	254.24	0.02	5.12	6.39	119.12
k = 5	1.97E+08	0.47	1.04E−03	21.28	10.07	0.51	0.62	176.72	0.02	5.74	5.93	158.40
k = 10	7.04E+07	0.55	2.24E−03	25.27	13.96	0.40	0.59	120.08	0.02	5.96	5.96	200.77
dtv = 0.5%	5.29E+08	0.35	3.85E−04	15.02	5.22	0.65	0.59	22.50	0.02	3.93	6.03	110.57
dtv = 1%	5.15E+08	0.34	3.68E−04	14.27	4.88	0.65	0.57	17.15	0.03	3.89	6.18	109.86
dtv = 2%	4.98E+08	0.34	3.43E−04	13.18	4.45	0.65	0.54	13.00	0.03	3.83	6.35	108.96
dtv = 5%	4.56E+08	0.33	3.02E−04	11.25	3.70	0.64	0.48	8.41	0.03	3.77	6.77	108.86
o = NYC	1.99E+08	0.14	9.66E−04	36.41	5.09	0.44	0.62	71.39	0.04	6.38	7.41	215.42
o = US	5.09E+07	0.46	6.01E−04	6.31	2.91	0.46	0.56	104.55	0.06	4.68	6.55	81.60
o = other	3.27E+07	0.51	7.81E−04	6.23	3.20	0.48	0.61	88.59	0.05	4.03	4.89	39.66
s = summer	3.67E+08	0.33	3.80E−04	12.62	4.22	0.59	0.61	154.02	0.03	4.75	6.96	108.58
s = winter	3.29E+08	0.32	4.02E−04	12.84	4.15	0.57	0.61	140.04	0.04	4.70	6.71	105.17

Note that for space reasons the table only contains the independent subsamples and not the combination of the subsampling data characteristics. Cp_u and Cp_i correspond to check-ins per user and check-ins per item (definition 4), Gini_U and Gini_I refers to Gini for users and for items (definition 5), and APB, ALT, ARG, ADCC, and ADA terms represent the averages of the popularity bias, long tail items, distance to the city center and duration active (Definitions 6, 7, 9, and 10, respectively). Key for the subsample column: k: value for the k-core, dtv: drop top venues (in terms of popularity), o: origin, s: season (cf. Sect. 5.2.2) The definition of the EV acronyms in the header row can be found in Sect. 3.2

Table 7 Hyperparameters tested in the recommenders

Rec	Hyperparameters
Random	None
Pop	None
UB	Sim = {Vector Cosine, Set Jaccard}, $k = \{20, 40, 60, 80, 100, 120\}$
IB	Sim = {Vector Cosine, Set Jaccard}, $k = \{20, 40, 60, 80, 100, 120\}$
HKV	Iter = 20, Factors = $\{10, 50, 100\}$, $\lambda = \{0.1, 1\}$, $\alpha = \{0.1, 1\}$
BPRMF	Factors = $\{10, 50, 100\}$, BiasReg = $\{0, 0.5, 1\}$, LearnRate = 0.05, Iter = 50, RegU = RegI = $\{0.0025, 0.001, 0.005, 0.01, 0.1\}$, RegJ = RegU/10
GeoBPRMF	Factors = $\{10, 50, 100\}$, BiasReg = $\{0, 0.5, 1\}$, LearnRate = 0.05, Iter = 50, RegU = RegI = $\{0.0025, 0.001, 0.005, 0.01, 0.1\}$, maxDist=$\{1, 4\}$
IRENMF	Factors = $\{50, 100\}$, geo-$\alpha = \{0.4, 0.6\}$, $\lambda_3 = \{0.1, 1\}$, clusters = $\{5, 50\}$
RankGeoFM	Factors = $\{50, 100\}$, $\alpha = \{0.1, 0.2\}$, c = 1, $\epsilon = 0.3$, neighs = $\{10, 50, 100, 200\}$ iter = 120, decay = 1, boldDriver = True, learnRate = 0.001
PopGeoNN	Sim = {Vector Cosine, Set Jaccard}, $k = \{20, 40, 60, 80, 100, 120\}$

The best configurations are selected by maximizing nDCG@5

B Hyperparameters

Table 7 shows the search space of the hyperparameters for each recommendation model. Refer to Sect. 5.3 for context.

C Supplementary result tables

The following tables correspond to the results presented in Sect. 6 at cutoffs of 10 and 20, respectively. The results generally follow the same patterns as presented for a cutoff of 5, however, we can observe a tendency that with a higher cutoff, the absolute impact of coefficients becomes smaller in the item exposure metric (Tables 8, 9, 10, 11, 12, 13).

Table 8 Coefficients of the regression model for nDCG@10

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.801	0.717	0.724	0.827	0.861	0.825	0.817
R^2 (adj)	0.789	0.701	0.708	0.817	0.852	0.815	0.806
Shape	− 0.004***	0.003**	− 0.003	− 0.004**	0	0.002	0.005***
Density	0.008***	0.012***	0.012***	0.012***	0.022***	0.023***	0.011***
$Gini_U$	− 0.002	0.018***	0.018***	0.003	0.014***	0.019***	0.005
StPB	0.007***	0.003	0.008**	0.011***	0.013***	0.007**	0.008**
KuPB	− 0.001	0.003	0.002	− 0.001	0.007***	0.007***	− 0.004
StRG	− 0.009**	− 0.011**	− 0.019***	− 0.013***	− 0.013***	− 0.012***	− 0.021***
MedDA	− 0.006***	0.012***	0.004*	− 0.003*	0.012***	0.012***	0.011***
KuDA	0.002	0.009**	0.011**	0.013***	− 0.001	0.002	0.008*

Table 9 Coefficients of the regression model for EPC@10

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.892	0.955	0.82	0.896	0.739	0.764	0.873
R^2 (adj)	0.886	0.953	0.809	0.89	0.723	0.75	0.865
Shape	0.008***	0	0.003***	0.005***	0.005***	0.004***	0.003***
Density	− 0.021***	− 0.004***	− 0.008***	− 0.017***	− 0.014***	− 0.011***	− 0.009***
$Gini_U$	0.009**	0	0.001	0.007***	0.006*	0.009***	0.006***
StPB	− 0.019***	0**	− 0.004***	− 0.014***	− 0.012***	− 0.01***	− 0.01***
KuPB	− 0.002	0	0.002*	0	0.002	0.002	0.001
StRG	0.004	0	0.002	0.001	− 0.002	− 0.001	− 0.001
MedDA	0.003**	0	0.002***	0.002*	0.003**	0.003***	0.001*
KuDA	− 0.002	− 0.001**	0.001	− 0.002	− 0.004	0	− 0.003**

Table 10 Coefficients of the regression model for ItemExposure@10

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.933	0.851	0.835	0.818	0.811	0.476	0.897
R^2 (adj)	0.929	0.842	0.825	0.807	0.799	0.445	0.891
Shape	− 1.593***	− 3.224***	− 3.208***	− 1.712***	− 2.339***	− 1.509***	− 2.662***
Density	0.518	− 0.277	− 0.026	0.537	− 0.179	0.394	− 0.626*
$Gini_U$	1.843***	2.059***	1.125	1.083*	− 0.188	− 2.266**	2.123***
StPB	1.521***	0.894**	1.45***	0.758*	1.8***	2.282***	0.687**
KuPB	− 0.515*	− 0.015	− 0.258	− 1.241***	− 0.526*	− 1.613***	− 0.074
StRG	0.697	− 0.968	− 0.574	1.009	1.53**	2.4**	0.817
MedDA	3.403***	1.048***	1.844***	2.569***	1.308***	0.77**	1.638***
KuDA	0.483	− 1.908***	− 1.037*	1.598***	1.926***	2.902***	0.839*

Table 11 Coefficients of the regression model for nDCG@20

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.834	0.746	0.747	0.831	0.858	0.844	0.821
R^2 (adj)	0.824	0.731	0.732	0.821	0.85	0.835	0.811
Shape	− 0.005***	0.005***	− 0.003	− 0.005***	0.001	0.003	0.008***
Density	0.011***	0.014***	0.017***	0.014***	0.026***	0.029***	0.014***
$Gini_U$	− 0.005	0.021***	0.024***	0.001	0.018***	0.021***	0.005
StPB	0.011***	0.003	0.01**	0.015***	0.017***	0.012***	0.011***
KuPB	− 0.002	0.004	0.004	− 0.001	0.009***	0.008***	− 0.003
StRG	− 0.01**	− 0.015***	− 0.025***	− 0.016***	− 0.017***	− 0.016***	− 0.025***
MedDA	− 0.009***	0.014***	0.006**	− 0.006***	0.015***	0.015***	0.014***
KuDA	0.003	0.007*	0.009*	0.013***	− 0.005	0	0.007

Table 12 Coefficients of the regression model for EPC@20

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.903	0.96	0.841	0.912	0.731	0.758	0.889
R^2 (adj)	0.897	0.958	0.831	0.907	0.715	0.744	0.882
Shape	0.007***	0	0.002***	0.004***	0.004***	0.003***	0.003***
Density	− 0.018***	− 0.004***	− 0.007***	− 0.015***	− 0.012***	− 0.01***	− 0.007***
$Gini_U$	0.009***	0	0.002	0.007***	0.006*	0.008***	0.005***
StPB	− 0.017***	0*	− 0.004***	− 0.014***	− 0.01***	− 0.009***	− 0.009***
KuPB	0	0	0.001*	0.002	0.002	0.002	0.001
StRG	0.003	0	0.002	0	− 0.002	− 0.001	− 0.001
MedDA	0.002*	0	0.002***	0.001	0.002*	0.003**	0.001
KuDA	− 0.002	0*	0	− 0.003	− 0.003	0	− 0.002**

Table 13 Coefficients of the regression model for ItemExposure@20

Name	Popularity	IB	UB	PopGeoNN	GeoBPRMF	BPRMF	IRENMF
R^2	0.933	0.717	0.714	0.715	0.69	0.29	0.868
R^2 (adj)	0.929	0.7	0.697	0.698	0.672	0.248	0.861
Shape	− 1.58***	− 4.227***	− 4.517***	− 1.474***	− 2.5***	− 1.515***	− 2.915***
Density	0.44	− 0.135	− 0.118	0.635	− 0.473	0.112	− 0.899**
$Gini_U$	1.885***	5.417***	3.738**	1.362*	− 0.73	− 2.924***	2.023***
StPB	1.443***	1.389	2.253*	0.127	1.726***	2.178***	0.463
KuPB	− 0.519*	0.604	0.1	− 2.026***	− 0.496	− 1.795***	− 0.144
StRG	0.682	− 4.454**	− 3.889**	1.088	1.648**	2.376**	0.754
MedDA	3.415***	1.056	1.368*	2.639***	0.405*	− 0.061	0.91***
KuDA	0.446	− 7.898***	− 6.694***	2.106***	2.294***	3.135***	1.103**

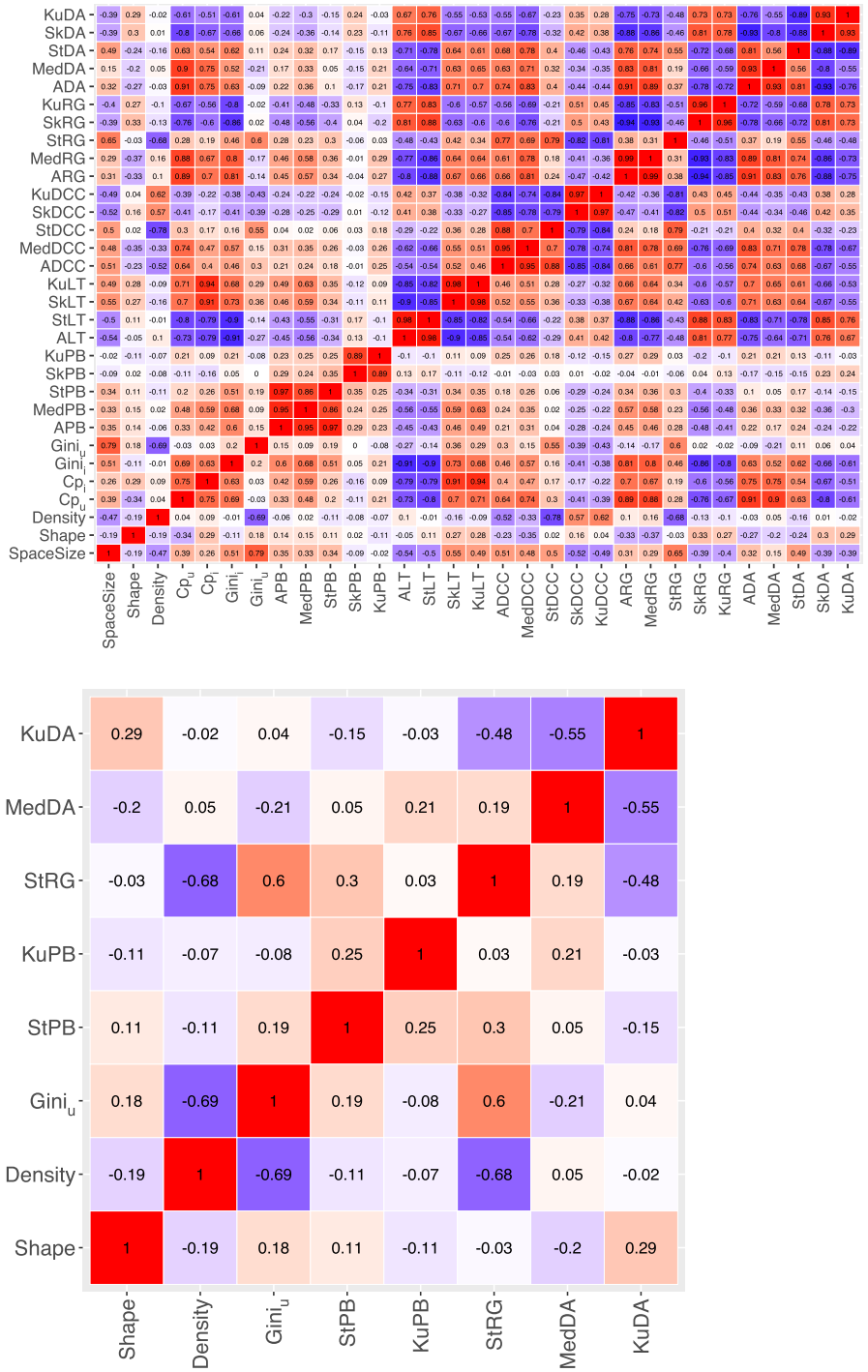


Fig. 7 Above: Pairwise Pearson correlations between all explanatory variables. Below: Pearson correlation between explanatory variables after removing highly correlated variables using Algorithm 1

D Pairwise correlations of explanatory variables

The following Fig. 7 showcases the effect of the reduction of EVs using pairwise correlation plots.

Funding LD acknowledges support from the European Union's Horizon 2020 research and innovation program under grant agreement No. 869764 (GoGreenRoutes). AB acknowledges support from grant PID2022-139131NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe".

Data availability The Foursquare data set used in this study is available from Dingqi Yang's web page https://sites.google.com/site/yangdingqi/home/foursquare-dataset. Further intermediate results are available in the supplementary material shared with this paper.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

References

Abdollahpouri H, Burke R, Mobasher B (2017) Controlling popularity bias in learning-to-rank recommendation. In: RecSys. ACM, New York, pp 42–46. https://doi.org/10.1145/3109859.3109912

Abdollahpouri H, Burke R, Mobasher B (2019) Managing popularity bias in recommender systems with personalized re-ranking. In: FLAIRS conference. AAAI Press, pp 413–418

Adomavicius G, Zhang J (2012) Impact of data characteristics on recommender systems performance. ACM Trans Manag Inf Syst 3(1):1–17. https://doi.org/10.1145/2151163.2151166

Adomavicius G, Bauman K, Tuzhilin A, Unger M (2022) Context-aware recommender systems: from foundations to recent developments. In: Recommender systems handbook. Springer US, pp 211–250

Aiolli F (2013) Efficient top-n recommendation for very large scale binary rated datasets. In: RecSys. ACM, pp 273–280. https://doi.org/10.1145/2507157.2507189

Albanna BH, Sakr MA, Moussa SM, Moawad IF (2016) Interest aware location-based recommender system using geo-tagged social media. ISPRS Int J Geo-Inf 5(12):245. https://doi.org/10.3390/ijgi5120245

Anand R, Beel J (2020) Auto-surprise: an automated recommender-system (autoRecSys) library with tree of Parzens estimator (TPE) optimization. In: RecSys, ACM, RecSys '20. https://doi.org/10.1145/3383313.3411467

Anelli VW, Noia TD, Sciascio ED, Pomo C, Ragone A (2019) On the discriminative power of hyper-parameters in cross-validation and how to choose them. In: RecSys. ACM, pp 447–451. https://doi.org/10.1145/3298689.3347010

Anelli VW, Bellogín A, Noia TD, Jannach D, Pomo C (2022) Top-n recommendation algorithms: a quest for the state-of-the-art. In: 30th ACM conference on user modeling, adaptation and personalization. ACM, New York. https://doi.org/10.1145/3503252.3531292
Bao J, Zheng Y, Wilkie D, Mokbel M (2015) Recommendations in location-based social networks: a survey. GeoInformatica 19(3):525–565. https://doi.org/10.1007/s10707-014-0220-8
Bellogín A, Castells P, Cantador I (2017) Statistical biases in information retrieval metrics for recommender systems. Inf Retr J 20(6):606–634. https://doi.org/10.1007/s10791-017-9312-z
Cai L, Xu J, Liu J, Pei T (2017) Integrating spatial and temporal contexts into a factorization model for POI recommendation. Int J Geogr Inf Sci 32(3):524–546. https://doi.org/10.1080/13658816.2017.1400550
Cañamares R, Castells P (2017) A probabilistic reformulation of memory-based collaborative filtering: implications on popularity biases. In: SIGIR. ACM, pp 215–224. https://doi.org/10.1145/3077136.3080836
Castells P, Hurley N, Vargas S (2022) Novelty and diversity in recommender systems. In: Recommender systems handbook. Springer US, pp 603–646
Cheng C, Yang H, King I, Lyu MR (2016) A unified point-of-interest recommendation framework in location-based social networks. ACM TIST 8(1):10:1-10:21. https://doi.org/10.1145/2901299
Cortés-Jiménez I (2008) Which type of tourism matters to the regional economic growth? The cases of Spain and Italy. Int J Tour Res 10(2):127–139. https://doi.org/10.1002/jtr.646
Cremonesi P, Koren Y, Turrin R (2010) Performance of recommender algorithms on top-n recommendation tasks. In: RecSys. ACM, pp 39–46. https://doi.org/10.1145/1864708.1864721
Dacrema MF, Cremonesi P, Jannach D (2019) Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In: RecSys. ACM, pp 101–109. https://doi.org/10.1145/3298689.3347058
Deldjoo Y, Noia TD, Sciascio ED, Merra FA (2020) How dataset characteristics affect the robustness of collaborative recommendation models. In: SIGIR. ACM, pp 951–960. https://doi.org/10.1145/3397271.3401046
Deldjoo Y, Bellogín A, Noia TD (2021) Explaining recommender systems fairness and accuracy through the lens of data characteristics. Inf Process Manag 58(5):102662. https://doi.org/10.1016/j.ipm.2021.102662
Deldjoo Y, Jannach D, Bellogin A, Difonzo A, Zanzonelli D (2023) Fairness in recommender systems: research landscape and future directions. User Model User-Adap Inter 34(1):59–108. https://doi.org/10.1007/s11257-023-09364-z
Dietz LW, Sen A, Roy R, Wörndl W (2020) Mining trips from location-based social networks for clustering travelers and destinations. Inf Technol Tour 22(1):131–166. https://doi.org/10.1007/s40558-020-00170-6
Ekstrand MD, Kluver D (2021) Exploring author gender in book rating and recommendation. User Model User Adapt Interact 31(3):377–420. https://doi.org/10.1007/s11257-020-09284-2
Ekstrand MD, Das A, Burke R, Diaz F (2022) Fairness in recommender systems. In: Recommender systems handbook. Springer, pp 679–707. https://doi.org/10.1007/978-1-0716-2197-4_18
Feng S, Li X, Zeng Y, Cong G, Chee YM, Yuan Q (2015) Personalized ranking metric embedding for next new POI recommendation. In: IJCAI. AAAI Press, pp 2069–2075
Gao H, Tang J, Hu X, Liu H (2015) Content-aware point of interest recommendation on location-based social networks. In: AAAI, AAAI Press, pp 1721–1727. https://doi.org/10.1609/aaai.v31i1.9462
Gibson H, Yiannakis A (2002) Tourist roles: needs and the lifecourse. Ann Tour Res 29(2):358–383
González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782. https://doi.org/10.1038/nature06958
Griesner J, Abdessalem T, Naacke H (2015) POI recommendation: towards fused matrix factorization with geographical and temporal influences. In: RecSys. ACM, pp 301–304. https://doi.org/10.1145/2792838.2799679
Gunawardana A, Shani G, Yogev S (2022) Evaluating recommender systems. In: Recommender systems handbook. Springer US, pp 547–601
Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: ICDM, IEEE, pp 263–272. https://doi.org/10.1109/ICDM.2008.22
Huang L, Ma Y, Liu Y, Sangaiah AK (2020) Multi-modal Bayesian embedding for point-of-interest recommendation on location-based cyber-physical-social networks. Future Gener Comput Syst 108:1119–1128. https://doi.org/10.1016/j.future.2017.12.020

Idrissi N, Zellou A (2020) A systematic literature review of sparsity issues in recommender systems. Soc Netw Anal Min 10(1):15. https://doi.org/10.1007/s13278-020-0626-2

Im I, Hars A (2007) Does a one-size recommendation system fit all? The effectiveness of collaborative filtering based recommendation systems across different domains and search modes. ACM Trans Inf Syst 26(1):4. https://doi.org/10.1145/1292591.1292595

Isufi E, Pocchiari M, Hanjalic A (2021) Accuracy-diversity trade-off in recommender systems via graph convolutions. Inf Process Manag 58(2):102459. https://doi.org/10.1016/j.ipm.2020.102459

Jannach D, Lerche L, Kamehkhosh I, Jugovac M (2015) What recommenders recommend: an analysis of recommendation biases and possible countermeasures. User Model User-Adapt Inter 25(5):427–491. https://doi.org/10.1007/s11257-015-9165-3

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446. https://doi.org/10.1145/582415.582418

Ji Y, Sun A, Zhang J, Li C (2022) A critical study on data leakage in recommender system offline evaluation. ACM Trans Inf Syst. https://doi.org/10.1145/3569930

Jiao X, Xiao Y, Zheng W, Wang H, Hsu C (2019) A novel next new point-of-interest recommendation system based on simulated user travel decision-making process. Future Gener Comput Syst 100:982–993. https://doi.org/10.1016/j.future.2019.05.065

Kaminskas M, Bridge D (2016) Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans Interact Intell Syst. https://doi.org/10.1145/2926720

Kariryaa A, Johnson I, Schöning J, Hecht B (2018) Defining and predicting the localness of volunteered geographic information using ground truth data. In: Conference on human factors in computing system. ACM. https://doi.org/10.1145/3173574.3173839

Karmaker SK, Hassan MM, Smith MJ, Xu L, Zhai C, Veeramachaneni K (2021) Automl to date and beyond: challenges and opportunities. ACM Comput Surv 54(8):1–36. https://doi.org/10.1145/3470918

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: SIGIR. ACM, pp 433–442. https://doi.org/10.1145/2766462.2767722

Li H, Ge Y, Hong R, Zhu H (2016) Point-of-interest recommendations: learning potential check-ins from friends. In: KDD. ACM, pp 975–984. https://doi.org/10.1145/2939672.2939767

Liu S, Zheng Y (2020) Long-tail session-based recommendation. In: RecSys. ACM, pp 509–514. https://doi.org/10.1145/3383313.3412222

Liu Q, Ge Y, Li Z, Chen E, Xiong H (2011) Personalized travel package recommendation. In: IEEE 11th international conference on data mining. IEEE, Vancouver, pp 407–416. https://doi.org/10.1109/icdm.2011.118

Liu Y, Wei W, Sun A, Miao C (2014) Exploiting geographical neighborhood characteristics for location recommendation. In: CIKM. ACM, pp 739–748. https://doi.org/10.1145/2661829.2662002

Manotumruksa J, Macdonald C, Ounis I (2018) A contextual attention recurrent architecture for context-aware venue recommendation. In: SIGIR. ACM, pp 555–564. https://doi.org/10.1145/3209978.3210042

Maroulis S, Boutsis I, Kalogeraki V (2016) Context-aware point of interest recommendation using tensor factorization. In: BigData. IEEE, pp 963–968. https://doi.org/10.1109/BigData.2016.7840694

Massimo D, Ricci F (2021) Popularity, novelty and relevance in point of interest recommendation: an experimental analysis. Inf Technol Tour 23(4):473–508. https://doi.org/10.1007/s40558-021-00214-5

Massimo D, Ricci F (2022) Building effective recommender systems for tourists. AI Mag 43(2):209–224. https://doi.org/10.1002/AAAI.12057

Massimo D, Ricci F (2023) Combining reinforcement learning and spatial proximity exploration for new user and new POI recommendations. In: UMAP. ACM, pp 164–174. https://doi.org/10.1145/3565472.3592966

Melchiorre AB, Rekabsaz N, Parada-Cabaleiro E, Brandl S, Lesota O, Schedl M (2021) Investigating gender fairness of recommendation algorithms in the music domain. Inform Process Manag 58(5):102666. https://doi.org/10.1016/j.ipm.2021.102666

Meng Z, McCreadie R, Macdonald C, Ounis I (2020) Exploring data splitting strategies for the evaluation of recommendation models. In: RecSys. ACM, pp 681–686. https://doi.org/10.1145/3383313.3418479

Neidhardt J, Schuster R, Seyfang L, Werthner H (2014) Eliciting the users' unknown preferences. In: RecSys. ACM, New York, pp 309–312. https://doi.org/10.1145/2645710.2645767

Nikolakopoulos AN, Ning X, Desrosiers C, Karypis G (2022) Trust your neighbors: a comprehensive survey of neighborhood-based methods for recommender systems. In: Recommender systems handbook. Springer US, pp 39–89

Nunes I, Marinho LB (2014) A personalized geographic-based diffusion model for location recommendations in LBSN. In: LA-WEB. IEEE, pp 59–67

O'brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41(5):673–690. https://doi.org/10.1007/s11135-006-9018-6

Penha G, Santos RLT (2020) Exploiting performance estimates for augmenting recommendation ensembles. In: RecSys. ACM, pp 111–119. https://doi.org/10.1145/3383313.3412264

Rahmani HA, Deldjoo Y, Tourani A, Naghiaei M (2022) The unfairness of active users and popularity bias in point-of-interest recommendation. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Advances in bias and fairness in information retrieval. Springer, Cham, pp 56–68. https://doi.org/10.1007/978-3-031-09316-6_6

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: Bayesian personalized ranking from implicit feedback. In: UAI. AUAI Press, pp 452–461

Robinson C, Schumacker RE (2009) Interaction effects: centering, variance inflation factor, and interpretation issues. Mult Linear Regress Viewp 35(1):6–11

Said A, Bellogín A (2014) Comparative recommender system evaluation: benchmarking recommendation frameworks. In: RecSys. ACM, pp 129–136. https://doi.org/10.1145/2645710.2645746

Said A, Bellogín A (2018) Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. User Model User-Adapt Interact 28(2):97–125

Santos F, de Almeida A, Martins C, Gonçalves R, Martins J (2019) Using POI functionality and accessibility levels for delivering personalized tourism recommendations. Comput Environ Urban Syst. https://doi.org/10.1016/j.compenvurbsys.2017.08.007

Shih T, Hou T, Jiang J, Lien Y, Lin C, Cheng P (2016) Dynamically integrating item exposure with rating prediction in collaborative filtering. In: SIGIR. ACM, pp 813–816. https://doi.org/10.1145/2911451.2914769

Srba I, Moro R, Tomlein M, Pecher B, Simko J, Stefancova E, Kompan M, Hrckova A, Podrouzek J, Gavornik A, Bielikova M (2023) Auditing Youtube's recommendation algorithm for misinformation filter bubbles. ACM Trans Recommend Syst. https://doi.org/10.1145/3568392

Stine RA (1995) Graphical interpretation of variance inflation factors. Am Stat 49(1):53–56. https://doi.org/10.1080/00031305.1995.10476113

Sánchez P, Bellogín A (2021) On the effects of aggregation strategies for different groups of users in venue recommendation. Inform Process Manag 58(5):102609. 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 Comput Surv. https://doi.org/10.1145/3510409

Sánchez P, Dietz LW (2022) Travelers vs. locals: the effect of cluster analysis in point-of-interest recommendation. In: 30th ACM conference on user modeling, adaptation and personalization. ACM, New York, pp 132–142. https://doi.org/10.1145/3503252.3531320

Sánchez P, Bellogín A, Boratto L (2023) Bias characterization, assessment, and mitigation in location-based recommender systems. Data Min Knowl Discov 37(5):1885–1929

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(sup1):234–240

Trattner C, Oberegger A, Marinho LB, Parra D (2018) Investigating the utility of the weather context for point of interest recommendations. Inf Technol Tour 19(1–4):117–150. https://doi.org/10.1007/s40558-017-0100-9

Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. In: RecSys. ACM, pp 109–116. https://doi.org/10.1145/2043932.2043955

Vente T, Ekstrand M, Beel J (2023) Introducing Lenskit-auto, an experimental automated recommender system (autoRecSys) toolkit. In: RecSys, ACM, RecSys'23. https://doi.org/10.1145/3604915.3610656

Wang Q, Yin H, Chen T, Huang Z, Wang H, Zhao Y, Hung NQV (2020a) Next point-of-interest recommendation on resource-constrained mobile devices. In: WWW, ACM/IW3C2, pp 906–916. https://doi.org/10.1145/3366423.3380170

Wang W, Chen J, Wang J, Chen J, Liu J, Gong Z (2020b) Trust-enhanced collaborative filtering for personalized point of interests recommendation. IEEE Trans Ind Inform 16(9):6124–6132. https://doi.org/10.1109/TII.2019.2958696

Weydemann L, Sacharidis D, Werthner H (2019) Defining and measuring fairness in location recommendations. In: LocalRec@SIGSPATIAL. ACM, pp 6:1–6:8. https://doi.org/10.1145/3356994.3365497

Wöörndl W, Hübner J, Bader R, Gallego-Vico D (2011) A model for proactivity in mobile, context-aware recommender systems. In: RecSys. ACM. https://doi.org/10.1145/2043932.2043981

Yang D, Zhang D, Chen L, Qu B (2015) NationTelescope: monitoring and visualizing large-scale collective behavior in LBSNs. J Netw Comput Appl 55:170–180. https://doi.org/10.1016/j.jnca.2015.05.010

Yang C, Bai L, Zhang C, Yuan Q, Han J (2017) Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. In: 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 1245–1254. https://doi.org/10.1145/3097983.3098094

Yao L, Sheng QZ, Qin Y, Wang X, Shemshadi A, He Q (2015) Context-aware point-of-interest recommendation using tensor factorization with social regularization. In: SIGIR. ACM, pp 1007–1010. https://doi.org/10.1145/2766462.2767794

Yin H, Cui B, Li J, Yao J, Chen C (2012) Challenging the long tail recommendation. VLDB Endow 5(9):896–907. https://doi.org/10.14778/2311906.2311916

Yin H, Cui B, Chen L, Hu Z, Zhang C (2015) Modeling location-based user rating profiles for personalized recommendation. TKDD 9(3):19:1-19:41. https://doi.org/10.1145/2663356

Yuan Q, Cong G, Ma Z, Sun A, Magnenat-Thalmann N (2013) Time-aware point-of-interest recommendation. In: SIGIR. ACM, pp 363–372. https://doi.org/10.1145/2484028.2484030

Yuan Q, Cong G, Sun A (2014) Graph-based point-of-interest recommendation with geographical and temporal influences. In: CIKM. ACM, pp 659–668. https://doi.org/10.1145/2661829.2661983

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: ICTAI. IEEE, pp 46–53. https://doi.org/10.1109/ICTAI.2016.0018

Zhang J, Chow C (2015) GeoSoCa: exploiting geographical, social and categorical correlations for point-of-interest recommendations. In: SIGIR. ACM, pp 443–452

Zhao X, Li X, Liao L, Song D, Cheung WK (2015) Crafting a time-aware point-of-interest recommendation via pairwise interaction tensor factorization. In: KSEM, vol 9403. Springer, pp 458–470. https://doi.org/10.1007/978-3-319-25159-2_41

Zhao P, Zhu H, Liu Y, Xu J, Li Z, Zhuang F, Sheng VS, Zhou X (2019) Where to go next: a spatio-temporal gated network for next POI recommendation. In: AAAI. AAAI Press, pp 5877–5884. https://doi.org/10.1609/aaai.v33i01.33015877

Authors and Affiliations

Linus W. Dietz[1] · Pablo Sánchez[2,3] · Alejandro Bellogín[3]

✉ Linus W. Dietz
linus.dietz@kcl.ac.uk

Pablo Sánchez
psperez@icai.comillas.edu

Alejandro Bellogín
alejandro.bellogin@uam.es

[1] King's College London, London, UK

[2] Instituto de Investigación Tecnológica (IIT), Universidad Pontificia Comillas, Madrid, Spain

[3] Universidad Autónoma de Madrid, Madrid, Spain