Exploring the Effect of Context-Awareness and Popularity Calibration on Popularity Bias in POI Recommendations

Andrea Forster
Graz University of Technology
Graz, Austria
andrea.forster@student.tugraz.at

Simone Kopeinik
Know Center Research GmbH
Graz, Austria
skopeinik@know-center.at

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

Stefan Thalmann
BANDAS Center
University of Graz
Graz, Austria
stefan.thalmann@uni-graz.at

Dominik Kowald
Know Center Research GmbH
Graz, Austria
Graz University of Technology
Graz, Austria
dkowald@know-center.at

Abstract

Point-of-interest (POI) recommender systems help users discover relevant locations, but their effectiveness is often compromised by popularity bias, which disadvantages less popular, yet potentially meaningful places. This paper addresses this challenge by evaluating the effectiveness of context-aware models and calibrated popularity techniques as strategies for mitigating popularity bias. Using four real-world POI datasets (Brightkite, Foursquare, Gowalla, and Yelp), we analyze the individual and combined effects of these approaches on recommendation accuracy and popularity bias. Our results reveal that context-aware models cannot be considered a uniform solution, as the models studied exhibit divergent impacts on accuracy and bias. In contrast, calibration techniques can effectively align recommendation popularity with user preferences, provided there is a careful balance between accuracy and bias mitigation. Notably, the combination of calibration and context-awareness yields recommendations that balance accuracy and close alignment with the users' popularity profiles, i.e., popularity calibration.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

POI recommendations,
popularity bias,
popularity calibration,
algorithmic fairness,
user groups,
context-aware recommender systems

1 Introduction

Tourism provides a rich ground for recommender systems (RS), supporting tasks such as destination planning, hotel, transport, or points-of-interest (POI) selection [29]. POI recommendations are particularly challenging, characterized by datasets with high sparsity and the reliance on observational data rather than explicit rating data to infer user preferences [7, 28]. Additional complexity is added by the use of contextual features such as geographic proximity, temporal dynamics, or social network connections to improve personalization [19, 20, 28, 29]. More recently, social media data, such as geotagged images, have also been explored to enhance POI recommendations [29]. A growing body of work highlights fairness and sustainability concerns in tourism and location-based RS. In this regard, popularity bias [2, 16, 18] poses a significant challenge, where popular POI dominate recommendations, limiting exposure to diverse and lesser-known places, attributed to the sensitivity of many well-known RS (especially collaborative filtering) to popularity indications [11, 26]. Popularity bias can lead to unfair treatment of both item providers, where niche locations receive less exposure [1, 5, 26], and consumers (i.e., RS users), not catering to those who prefer less mainstream locations [2, 10, 17, 26]. In tourism, this can elevate overcrowding, environmental degradation, or reduced user satisfaction [1, 5, 26].

Although prior research highlights the potential of context-awareness [5, 26] and calibrated popularity (CP) [4, 15, 21, 33] as effective strategies to mitigate popularity bias, these two approaches have so far been studied in isolation. To our knowledge, CP has not yet been applied in the context of POI recommendation, nor combined with context-aware models. This raises questions about whether these strategies complement each other and how they jointly affect the trade-off between accuracy and popularity bias.

In our work, we address this research gap by systematically evaluating CP and two well-known context-aware models, a *LO*cation *RE*commendation approach (LORE) [36] and a collaborative filtering model that harnesses *U*ser preference, *S*ocial influence, and *G*eographical influence (USG) [35], independently and in combination, and comparing them to the non-contextualized baseline Bayesian Personalized Ranking (BPR) [27], thus contributing empirical insight into how context-aware RS and calibration interact, and how these methods affect different user groups. Using four real-world POI datasets (Brightkite, Foursquare, Gowalla, and Yelp), we group users based on their preference for popular locations (*LowPop*, *MedPop*, and *HighPop*), and analyze how the RS

approaches impact accuracy and popularity within these groups. We formulate two research questions to guide this work and to structure our methodology as well as results presentation:

- **RQ1.** To what extent can context-aware recommendations (LORE, USG) and calibration-based debiasing (CP) individually mitigate popularity bias in POI recommendations, and how does this impact accuracy, compared to a non-contextual baseline (BPR)?
- **RQ2.** Does the combination of context-aware POI recommendations and calibration-based debiasing (CP) improve the trade-off between recommendation accuracy and popularity bias, compared to their respective purely context-aware versions (LORE, USG)?

Our results show that non-contextualized models like BPR disadvantage *LowPop* users, while the performance of context-aware models depends predominantly on the model and dataset. CP helps align recommendations with user preferences, but may affect accuracy (RQ1). Combining CP with context-aware models yields interesting results: USG combined with CP achieves higher accuracy, but retains more popularity bias, while LORE combined with CP increases the popularity of the LORE base version, yet offers the closest match to user popularity profiles out of all methods and combinations studied in our work (RQ2).

Our findings provide valuable insights for researchers and practitioners working on mitigating popularity bias in POI recommendation systems and can inform the design of user studies aimed at evaluating user experience.

2 Background and Method

2.1 Overview of Related Work

Several studies explore how the effects of popularity bias and overcrowding can be mitigated in RS: Ghanem et al. [12] model trade-offs between consumer utility and provider profit; Merinov et al. [23] optimize travel itineraries to reduce POI congestion; Banerjee et al. [6] propose a fairness score incorporating environmental and seasonal constraints and location popularity into recommendations; Massimo and Ricci [22] explore the trade-offs between accuracy and novelty in POI recommendation, concluding that users prioritize precision over novelty and struggle to assess unknown suggestions. Banerjee et al. [5], Rahmani et al. [25, 26] find context-aware POI recommendation models that incorporate geographical, temporal, social, or categorical information to produce more diverse recommendation lists than traditional RS (e.g., collaborative filtering). Abdollahpouri et al. [4] introduce CP, a user-centered debiasing method that aligns the distribution of head, mid, and tail (H, M, T) items in recommendation lists with users' historical interaction patterns. Klimashevskaia et al. [15] evaluate CP in a movie streaming setting, finding that it improves fairness without degrading performance; Ungruh et al. [33] conducted a user study in the music domain showing that a mix of familiar and unfamiliar items leads to more satisfactory recommendations; and Lesota et al. [21] introduce group-specific trade-off parameters to improve fairness across user subgroups, calling for more targeted criteria to select mitigation parameters.

2.2 Datasets

In our work, we leverage four datasets commonly used in tourism and POI recommendation research [5, 26, 28, 32, 38]: Foursquare[1], Yelp[2], Brightkite[3], and Gowalla[4]. Due to limited space, we illustrate results only for Foursquare and Yelp, but provide them for all four datasets on GitHub[5]. Each dataset includes *user ID*, *timestamp*, *check-in* location, and POI coordinates. To reduce the computational costs of our study, we create data samples. The Foursquare sample includes 1,500 users, 2,804 items, and 69,401 unique check-ins (sparsity = 98.4%). Our Yelp sample includes 1,500 users, 4,515 items, and 35,288 unique check-ins (sparsity = 99.5%), as illustrated in Table 1. We group users by their average profile popularity, based on normalized location check-in frequencies [3, 17]: the bottom 20% (*LowPop*), middle 60% (*MedPop*), top 20% (*HighPop*). We similarly classify items as tail (T, bottom 20%), mid (M, middle 60%), and head (H, top 20%) [4, 33]. Finally, we apply a user-based temporal split [28] to create training (65%), validation (15%), and test (20% most recent check-ins) sets.

Table 1: Descriptive statistics of the datasets used in this work. For the sake of space, we only report results for Foursquare and Yelp in the remainder of this paper.

Dataset	Users	Items	Unique check-ins	Sparsity
Brightkite	600	794	15,341	0.967798
Foursquare	1,500	2,804	69,401	0.983500
Gowalla	1,500	7,579	53,679	0.995278
Yelp	1,500	4,515	35,288	0.994790

2.3 Popularity Bias Mitigation

Context-Aware POI Recommendation. We use two well-known POI recommendation models, which are implemented in the CAPRI framework[6]. The first model, LORE, integrates sequential, social, and geographical influences by combining additive Markov chains with a location-location transition graph to model user movement patterns and POI transitions [36]. The second model, USG, is a hybrid model that combines user- and friend-based collaborative filtering with geographical influence using a naive Bayes approach [35]. We do not use data on social relations in our models and experiments, since not all of our datasets contain such information.

Calibrated Popularity (CP). CP is a re-ranking method that adjusts a base recommendation list to better reflect a user's historical preferences for item popularity levels. Proposed by Abdollahpouri et al. [4] and based on the idea of calibrated recommendations by Steck [31], CP selects a refined recommendation list L_u^* of size n from a larger base list L_u of size m, using a weighted optimization that balances item relevance and distributional similarity.

[1] https://www.kaggle.com/datasets/chetanism/foursquare-nyc-and-tokyo-checkin-dataset
[2] https://www.yelp.com/dataset
[3] https://snap.stanford.edu/data/loc-brightkite.html
[4] https://snap.stanford.edu/data/loc-gowalla.html
[5] https://github.com/andreafooo/POI_RS_PopBias_Mitigation
[6] https://github.com/CapriRecSys/CAPRI

Table 2: RQ1. Performance of context-aware recommendation models and CP applied on BPR in relation to the BPR baseline. Metrics include nDCG, ARP, and PopLift. Symbols indicate the preferred direction for each metric: ↓ (lower is better), ↑ (higher is better), and → 0 (closer to zero is better), and best values are shown in bold. For BPR, absolute values are shown; Δ% values for LORE, USG CP_H and $CP_{\mathfrak{J}}$. Significant relations are indicated by ** via t-test ($p < 0.05$), Bonferroni-corrected for each metric.

Group	nDCG ↑ | Δ% nDCG					ARP ↓ | Δ% ARP					PopLift → 0 | Δ% PopLift				
	BPR	LORE	USG	CP_H	$CP_{\mathfrak{J}}$	BPR	LORE	USG	CP_H	$CP_{\mathfrak{J}}$	BPR	LORE	USG	CP_H	$CP_{\mathfrak{J}}$
Foursquare															
LowPop	0.0395	-56.62%**	-43.28%**	**+0.54%**	-21.28%	0.0795	**-91.30%****	+94.71%**	-0.08%	-30.97%**	4.3299	**-110.89%****	+149.05%**	-0.11%	-39.98%**
MedPop	**0.1084**	-88.77%**	-11.81%**	-0.04%	-14.92%**	0.1009	**-93.69%****	+32.79%**	-0.41%**	-17.69%**	2.2977	**-134.47%****	+48.68%**	-0.72%**	-26.28%**
HighPop	**0.1655**	-96.63%**	-2.11%	-0.14%	-6.43%**	0.1089	**-94.31%****	+18.09%**	-0.37%**	-6.85%**	1.2302	**-171.04%****	+33.50%**	-0.82%**	-13.57%**
All	**0.1060**	-88.83%**	-11.13%**	-0.03%	-12.74%**	0.0982	**-93.44%****	+39.55%**	-0.35%**	-17.44%**	2.4906	**-129.88%****	+82.08%**	-0.52%**	-29.79%**
Yelp															
LowPop	0.0192	**+126.17%****	+12.65%	+6.94%	+6.94%	0.0040	**-63.60%****	-15.61%**	-38.46%**	-38.46%**	1.7702	**-101.15%****	-25.27%**	-65.20%**	-65.20%**
MedPop	0.0304	-33.66%**	**+2.22%**	+1.15%	-27.81%**	0.0079	**-75.92%****	-5.00%**	-0.14%**	-35.00%**	2.3983	**-107.50%****	-7.77%**	-0.22%**	-50.67%**
HighPop	**0.0650**	-75.53%**	-0.94%	**+0.00%**	-17.65%**	0.0093	**-74.34%****	+0.54%	-0.01%	-18.58%**	1.0542	**-145.19%****	+1.13%	-0.04%	-37.78%**
All	0.0350	-31.70%**	**+2.19%**	+1.36%	-20.24%**	0.0074	**-74.20%****	-4.74%**	-4.21%**	-31.24%**	2.0039	**-110.34%****	-9.93%**	-11.68%**	-51.88%**

Optimization. The optimization is guided by a trade-off parameter $\lambda \in [0, 1]$, controlling the balance between relevance $\text{Rel}(L_u)$ based on item scores from the recommendation model, and calibration measured by the Jensen-Shannon divergence $\mathfrak{J}(P, Q(L_u))$ between the user's historical popularity distribution P and the recommendation list's distribution Q. Higher divergence indicates a stronger mismatch between what the user prefers and what is recommended, resulting in a greater penalty. This is formally given by:

$$L_u^* = \arg \max_{L_u, |L_u|=n} ((1-\lambda) \cdot \text{Rel}(L_u) - \lambda \cdot \mathfrak{J}(P, Q(L_u))) \quad (1)$$

The final list L_u^* is constructed iteratively via greedy optimization, where at each step, the item that maximizes the weighted trade-off between relevance and calibration is selected. To further personalize the debiasing procedure, we optimize λ separately for each user group (*LowPop*, *MedPop*, and *HighPop*) via grid search, resulting in group-specific parameters that are inserted into Equation (1). The optimal λ values are found by:

CP_H: Maximizing the harmonic mean H of accuracy (nDCG) and calibration (1 - $\mathfrak{J}$), following Lesota et al. [21]:

$$\lambda = \frac{\text{nDCG}_g \cdot (1 - \mathfrak{J}_g(P, Q))}{\text{nDCG}_g + (1 - \mathfrak{J}_g(P, Q))} \quad (2)$$

$CP_{\mathfrak{J}}$: minimizing Jensen-Shannon divergence directly by setting $\lambda = 1$, prioritizing calibration over accuracy, denoted as $CP_{\mathfrak{J}}$.

2.4 Training and Evaluation

We generate baseline recommendations (BPR) using RecBole[7] [37] and the context-aware recommendations (LORE and USG) using CAPRI[8] [32]. To foster reproducibility [30], our code and data samples are publicly available via GitHub[9]. We train the models for 200 epochs, optimizing the learning rate, embedding size, and batch size on the validation sets. The top-150 recommendations per user are stored for CP_H and $CP_{\mathfrak{J}}$ as a basis for re-ranking to evaluate the final top-10 recommendations. In this study, we evaluate accuracy via nDCG[13, 14, 34]. The evolution of the recommendation popularity, and consequently, the level of popularity bias, is measured by average recommendation popularity (ARP) [4, 24] and the popularity lift (PopLift) [24]. While ARP illustrates the popularity of the recommendations of each user, PopLift helps to reflect whether the popularity of the recommendations is higher, lower, or close to the popularity of items in the user profile for each user group.

3 Results

In this section, we present our results according to our two research questions. As mentioned beforehand, for the sake of space, we only discuss our findings for the Foursquare and Yelp datasets.

RQ1. Independent Impact of Context-Awareness and Popularity Calibration. We depict our results for RQ1 in Table 2 and evaluate how the context-aware recommendation models LORE and USG, as well as popularity calibration, (1) optimized for the weighted mean between nDCG and $\mathfrak{J}$ (CP_H), and (2) optimized to minimize $\mathfrak{J}$ ($CP_{\mathfrak{J}}$), perform independently compared to the non-contextualized BPR baseline.

The impact of context-awareness depends on the chosen model and dataset[10]. LORE significantly lowers ARP and PopLift for all user groups in both datasets, effectively diminishing popularity bias and delivering niche recommendations, but often at the cost of lower accuracy, except for niche user groups like *LowPop*, where accuracy is significantly improved in the Yelp dataset. Figure 1 shows that T-items increase in all groups in LORE *Base* compared to the user profile. USG has a contrary effect in Foursquare, producing recommendations that almost exclusively contain H-items, leading to a significant increase in ARP and PopLift for all user groups, thus increasing the level of popularity bias, compared to the general baseline. Simultaneously, accuracy decreases significantly for all user groups, except *HighPop*, where the decrease is non-significant. The context-aware algorithm leads to better results on Yelp with a significant decline in ARP and PopLift and a non-significant increase in nDCG for all user groups except *HighPop*.

CP_H leads to a significant decrease in ARP and PopLift for all user groups except *LowPop* in Foursquare, whilst neither significantly impacting accuracy, nor approximating the user profile distribution as illustrated in Figure 1. In the Yelp dataset, no significant changes are observed using the accuracy-oriented calibration technique,

[7] https://github.com/RUCAIBox/RecBole/tree/master
[8] https://github.com/CapriRecSys/CAPRI
[9] https://github.com/andreafooo/POI_RS_PopBias_Mitigation
[10] For full results on all datasets, please see: https://github.com/andreafooo/POI_RS_PopBias_Mitigation

Table 3: RQ2. Evaluation results for LORE and USG algorithms combined with CP_H and $CP_{\mathfrak{J}}$ methods. Results for LORE *Base* and USG *Base* are equivalent to LORE and USG from Table 2 (best values are shown in bold); Δ% values per metric between *Base* and CP_H/$CP_{\mathfrak{J}}$ in brackets, showing any significant t-test differences ($p < 0.05$ as **), Bonferroni-corrected for each metric.

Model	Group	nDCG ↑ (Δ% nDCG)			ARP ↓ (Δ% ARP)			PopLift → 0 (Δ% PopLift)		
		Base	CP_H	$CP_{\mathfrak{J}}$	*Base*	CP_H	$CP_{\mathfrak{J}}$	*Base*	CP_H	$CP_{\mathfrak{J}}$
Foursquare										
LORE	*LowPop*	0.0172	**0.0255 (+48.51%)**	**0.0255 (+48.51%)**	**0.0069**	0.0119 (+72.26%**)	0.0119 (+72.26%**)	-0.4715	**-0.1371 (+70.92%**)**	**-0.1371 (+70.92%**)**
LORE	*MedPop*	0.0122	**0.0212 (+74.07%**)**	**0.0212 (+74.07%**)**	**0.0064**	0.0132 (+107.75%**)	0.0132 (+107.75%**)	-0.7920	**-0.5712 (+27.88%**)**	**-0.5712 (+27.88%**)**
LORE	*HighPop*	0.0056	**0.0200 (+258.16%**)**	**0.0200 (+258.16%**)**	**0.0062**	0.0149 (+139.93%**)	0.0149 (+139.93%**)	-0.8740	**-0.6986 (+20.06%**)**	**-0.6986 (+20.06%**)**
LORE	*All*	0.0118	**0.0218 (+83.99%**)**	**0.0218 (+83.99%**)**	**0.0064**	0.0133 (+106.32%**)	0.0133 (+106.32%**)	-0.7443	**-0.5099 (+31.50%**)**	**-0.5099 (+31.50%**)**
USG	*LowPop*	0.0224	**0.0228 (+1.72%)**	0.0222 (-1.03%)	0.1548	0.1353 (-12.62%**)	**0.1247 (-19.47%**)**	10.7837	9.0601 (-15.98%**)	**7.9484 (-26.29%**)**
USG	*MedPop*	**0.0956**	0.0937 (-2.03%)	0.0912 (-4.58%)	0.1340	0.1287 (-3.90%**)	**0.1189 (-11.26%**)**	3.4161	3.2405 (-5.14%)	**2.9022 (-15.04%**)**
USG	*HighPop*	**0.1620**	0.1602 (-1.12%)	0.1552 (-4.15%)	0.1286	0.1253 (-2.56%)	**0.1168 (-9.19%**)**	1.6424	1.5728 (-4.24%)	**1.3966 (-14.97%**)**
USG	*All*	**0.0942**	0.0928 (-1.54%)	0.0902 (-4.27%)	0.1371	0.1294 (-5.62%**)	**0.1196 (-12.72%**)**	4.5349	4.0709 (-10.23%**)	**3.6103 (-20.39%**)**
Yelp										
LORE	*LowPop*	**0.0434**	0.0426 (-1.69%)	0.0390 (-10.03%)	**0.0014**	**0.0014 (+0.00%)**	0.0015 (+5.37%)	-0.0203	-0.0210 (-3.66%)	**0.0052 (+125.79%)**
LORE	*MedPop*	0.0201	**0.0254 (+26.22%)**	**0.0254 (+26.22%)**	**0.0019**	0.0022 (+16.65%**)	0.0022 (+16.65%**)	-0.1798	**-0.0473 (+73.71%**)**	**-0.0473 (+73.71%**)**
LORE	*HighPop*	0.0159	**0.0269 (+68.99%)**	**0.0269 (+68.99%)**	**0.0024**	0.0032 (+34.00%**)	0.0032 (+34.00%**)	-0.4764	-0.3043 (-36.12%**)	-0.3043 (-36.12%**)
LORE	*All*	0.0239	**0.0292 (+21.79%)**	0.0284 (+18.77%)	**0.0019**	0.0023 (+18.46%**)	0.0023 (+19.29%**)	-0.2072	-0.0934 (+54.91%**)	**-0.0882 (+57.44%**)**
USG	*LowPop*	0.0216	**0.0246 (+14.09%)**	**0.0246 (+14.09%)**	0.0033	**0.0022 (-35.61%**)**	**0.0022 (-35.61%**)**	1.3230	**0.4271 (-67.72%**)**	**0.4271 (-67.72%**)**
USG	*MedPop*	0.0310	**0.0315 (+1.58%)**	0.0274 (-11.89%)	0.0075	0.0075 (+0.00%)	**0.0048 (-35.75%**)**	2.2118	2.2066 (-0.23%)	**1.0415 (-52.91%**)**
USG	*HighPop*	**0.0644**	**0.0644 (+0.00%)**	0.0496 (-22.87%)	0.0094	0.0093 (-0.26%)	**0.0075 (-19.45%**)**	1.0661	1.0605 (-0.53%)	**0.6456 (-39.44%**)**
USG	*All*	0.0358	**0.0367 (+2.52%)**	0.0313 (-12.70%)	0.0070	0.0068 (-3.55%)	**0.0048 (-31.41%**)**	1.8049	1.6215 (-10.16%**)	**0.8395 (-53.49%**)**

except for *HighPop*. This can also be seen in Figure 1, where BPR *Base* is heavily biased towards the most popular H-items, while CP_H approximates the distribution for *LowPop* users and increases accuracy, yet fails to do so for the other groups. $CP_{\mathfrak{J}}$, optimized to align the popularity of the recommendations with the user profile, has more profound effects on both accuracy and popularity bias, leading to significant decreases in accuracy and popularity bias for all users in the Foursquare dataset, except for the *LowPop* users' accuracy. In Yelp, a significant decline in ARP and PopLift, hence a reduced level of popularity bias, is evident for all user groups. However, this also impacts accuracy for all groups, except *LowPop*, where a non-significant increase in nDCG is achieved after calibration. $CP_{\mathfrak{J}}$ approximates the distribution of H- and M-items to the user profiles, yet T-items remain underrepresented (see Figure 1).

In the PopLift metric, the baseline and calibrated BPR recommendations and USG produce recommendations that exceed the user profile's popularity level. In contrast, LORE delivers recommendations below the user profile's popularity level, yielding the highest accuracy for *LowPop* users, also suggesting potential for better personalization among underrepresented users.

RQ2. Combined Impact of Context-Awareness and Popularity Calibration. To address RQ2, we evaluate the effects of combining context-aware models USG and LORE with CP_H and $CP_{\mathfrak{J}}$ against their respective *Base* (top-10 recommendations without calibration) results, and depict them in Table 3.

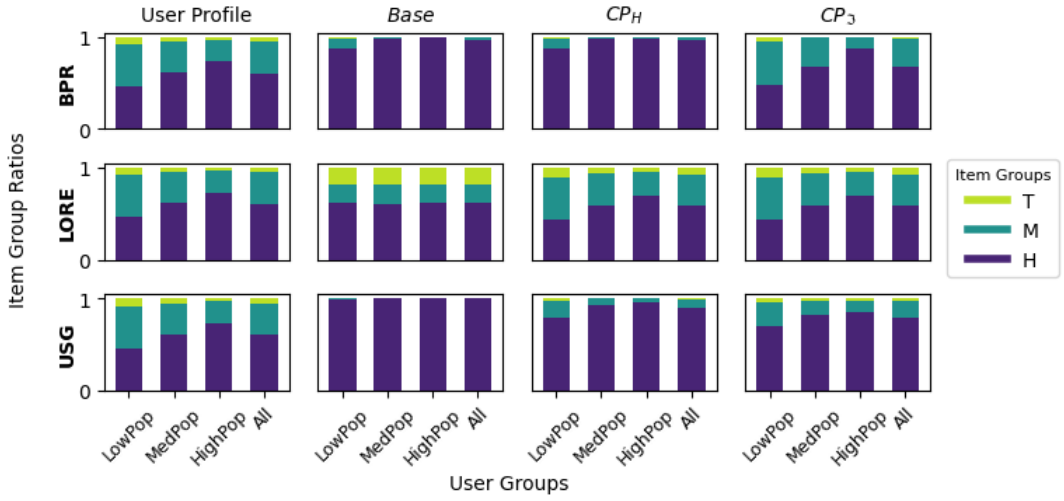
LORE (*Base*) recommends less popular items than the user profile, as highlighted by the negative PopLift values in Foursquare and Yelp, or the high proportion of T-items in Figure 1, yet lacks distributional similarity to the user profiles. Applying CP-based re-ranking to LORE improves accuracy in both datasets for all groups except the *LowPop* user group in Yelp, but also increases the ARP of the recommended items, thereby reducing diversity. Nonetheless, our analysis of the PopLift metric suggests that, despite this increase in popularity, the combination of LORE with CP aligns the popularity distribution of the recommendations more closely with the user profile than other methods and combinations, also managing to include T-items in the recommendation lists of all user groups, as opposed to applying CP to BPR.

USG (*Base*), in combination with the accuracy-oriented CP_H, shows no significant impact on nDCG in either dataset. We generally observe a positive, yet non-significant, effect on the *LowPop* accuracy in both datasets. Concerning popularity bias, in Foursquare, the combination significantly reduces ARP for all groups except *HighPop*, and improves PopLift significantly for *LowPop* and All. However, as illustrated in Table 2, USG *Base* exhibits a particularly high level of popularity bias in Foursquare, creating recommendations more than 10 times as high as the user profile for *LowPop* users, vs. more than 4 times as high for BPR. In Yelp, where BPR and USG yield similar results, combining USG and CP_H significantly reduces both popularity bias metrics for *LowPop*, while leading to equal levels or non-significant increases in accuracy for the user groups. The combination of USG with $CP_{\mathfrak{J}}$ consistently reduces ARP and PopLift across all user groups in both datasets. These reductions are statistically significant, indicating popularity bias mitigation while maintaining accuracy and reducing bias.

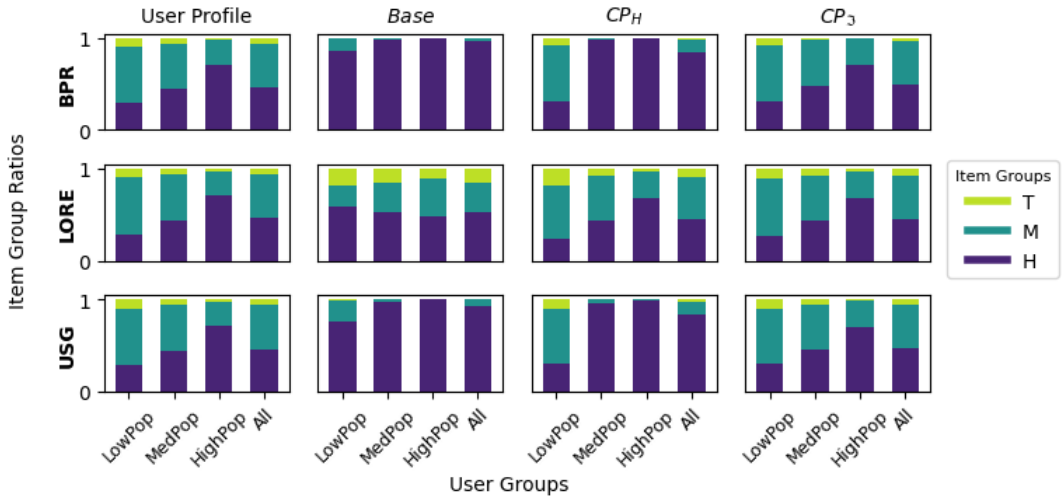
4 Discussion and Conclusion

In this paper, we evaluated the effectiveness of context-awareness and popularity calibration to mitigate popularity bias in POI recommendations. In general, we find that the effectiveness of context-awareness varies substantially between models, while the effectiveness of CP can vary for different user groups and is dictated by the chosen λ that balances accuracy and calibration. The combination of context-awareness and CP helps counteract the shortcomings of both methods when applied individually.

The two context-aware models yield contrasting results, with LORE producing the most niche recommendations and USG producing the most biased recommendations in all datasets except Yelp. In all four datasets, the combination of BPR and CP_H hardly improves the biased item distribution and popularity bias, especially for *MedPop* and *HighPop* users, whereas $CP_{\mathfrak{J}}$ approximates the distribution in the user profile, except for T-items that remain



(a) Foursquare



(b) Yelp

Figure 1: RQ1 & RQ2. Comparing the item group ratios (T, M, H) in the groups' user profiles to BPR, LORE, and USG *Base*, CP_H, and $CP_\mathfrak{J}$ in (a) Foursquare and (b) Yelp.

underrepresented, suggesting a lack of T-items in the top-150 recommendations produced by BPR that form the basis for the re-ranking (RQ1). Combining context-aware models and CP, especially LORE, increases the accuracy and popularity of recommendations, while recommending slightly more T-items than the user profiles. This can closely align recommendations with the user profiles' item distributions and PopLift values, potentially counteracting popularity bias over time. In our study, the combination of LORE and CP is the only method that accurately reflects the distribution of T-items, suggesting that the combination can counteract the algorithms' low user fairness that was discussed in previous research [26] (RQ2). Therefore, if mitigating popularity bias is the primary objective, our findings suggest combining LORE with CP. Conversely, if preserving accuracy while slightly reducing bias is preferred, general models such as BPR or context-aware models like USG combined with CP and tuned towards accuracy (e.g., CP_H) are more suitable.

To validate these findings and assess their practical relevance, in our future work, we plan to investigate the effects of context-aware and calibration-based combinations in user studies, particularly focusing on their impact on user satisfaction and perceived recommendation quality across different user groups. In addition, we plan to study how these methods impact other RS stakeholders, in particular POI providers, as well, by following principles of multi-stakeholder RS evaluation [8, 9].

Acknowledgments

This research was supported by the Austrian FFG COMET program and the FFG Femtech project Radreisen4All.

References

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158* (2019).

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided exposure bias in recommendation. *arXiv preprint arXiv:2006.15772* (2020).

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).

[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*. 119–129.

[5] Ashmi Banerjee, Paromita Banik, and Wolfgang Wörndl. 2023. A review on individual and multistakeholder fairness in tourism recommender systems. *Frontiers in big Data* 6 (2023), 1168692.

[6] Ashmi Banerjee, Tunar Mahmudov, Emil Adler, Fitri Nur Aisyah, and Wolfgang Wörndl. 2025. Modeling sustainable city trips: integrating CO 2 e emissions, popularity, and seasonality into tourism recommender systems. *Information Technology & Tourism* (2025), 1–38.

[7] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic, and Markus Strohmaier. 2015. Photowalking the City: Comparing Hypotheses About Urban Photo Trails on Flickr. In *Social Informatics*, Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu (Eds.). Springer International Publishing, Cham, 227–244.

[8] Robin Burke, Gediminas Adomavicius, Toine Bogers, Tommaso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Soledad Pera, Nava Tintarev, and Jürgen Ziegler. 2025. De-centering the (Traditional) User: Multistakeholder Evaluation of Recommender Systems. *arXiv preprint arXiv:2501.05170* (2025).

[9] Robin Burke, Gediminas Adomavicius, Toine Bogers, Tommaso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Soledad Pera, and Jürgen Ziegler. 2024. Dagstuhl Seminar on Evaluation Perspectives of Recommender Systems: Multistakeholder and Multimethod Evaluation. *Dagstuhl Report on Evaluation Perspectives of Recommender Systems: Driving Research and Education* (2024).

[10] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[11] Yashar Deldjoo, Alejandro Bellogin, and Tommaso Di Noia. 2021. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Information processing & management* 58, 5 (2021), 102662.

[12] Nada Ghanem, Stephan Leitner, and Dietmar Jannach. 2022. Balancing consumer and business value of recommender systems: A simulation-based analysis. *Electronic Commerce Research and Applications* 55 (2022), 101195.

[13] Aryan Jadon and Avinash Patil. 2024. A comprehensive survey of evaluation techniques for recommendation systems. In *International Conference on Computation of Artificial Intelligence & Machine Learning*. Springer, 281–304.

[14] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[15] Anastasiia Klimashevskaia, Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Astrid Tessem, and Christoph Trattner. 2023. Evaluating the effects of calibrated popularity bias mitigation: a field study. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1084–1089.

[16] Dominik Kowald and Emanuel Lacic. 2022. Popularity bias in collaborative filtering-based multimedia recommender systems. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.

[17] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 35–42.

[18] Emanuel Lacic, Leon Fadljevic, Franz Weissenboeck, Stefanie Lindstaedt, and Dominik Kowald. 2022. What drives readership? an online study on user interface types and popularity bias mitigation in news article recommendations. In *European Conference on Information Retrieval*. Springer, 172–179.

[19] Emanuel Lacic, Dominik Kowald, Denis Parra, Martin Kahr, and Christoph Trattner. 2014. Towards a scalable social recommender engine for online marketplaces: The case of apache solr. In *Proceedings of the 23rd International Conference on World Wide Web*. 817–822.

[20] Emanuel Lacic, Dominik Kowald, Paul Christian Seitlinger, Christoph Trattner, and Denis Parra. 2014. Recommending Items in Social Tagging Systems Using Tag and Time Information. In *In Proceedings of the 1st Social Personalization Workshop co-located with the 25th ACM Conference on Hypertext and Social Media*. ACM.

[21] Oleg Lesota, Stefan Brandl, Matthias Wenzel, Alessandro B Melchiorre, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization.. In *MORS@ RecSys*.

[22] David Massimo and Francesco Ricci. 2021. Popularity, novelty and relevance in point of interest recommendation: an experimental analysis. *Information Technology & Tourism* 23, 4 (2021), 473–508.

[23] Pavel Merinov, David Massimo, and Francesco Ricci. 2022. Sustainability Driven Recommender Systems. *CEUR Workshop Proceedings, Vol.3177, 22* (June 2022).

[24] Peter Müllner, Elisabeth Lex, Markus Schedl, and Dominik Kowald. 2024. The impact of differential privacy on recommendation accuracy and popularity bias. In *European Conference on Information Retrieval*. Springer, 466–482.

[25] Hossein A Rahmani, Yashar Deldjoo, and Tommaso Di Noia. 2022. The role of context fusion on accuracy, beyond-accuracy, and fairness of point-of-interest recommendation systems. *Expert Systems with Applications* 205 (2022), 117700.

[26] Hossein A Rahmani, Yashar Deldjoo, Ali Tourani, and Mohammadmehdi Naghiaei. 2022. The unfairness of active users and popularity bias in point-of-interest recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 56–68.

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[28] Pablo Sánchez and Alejandro Bellogín. 2022. Point-of-interest recommender systems based on location-based social networks: a survey from an experimental perspective. *ACM Computing Surveys (CSUR)* 54, 11s (2022), 1–37.

[29] Joy Lal Sarkar, Abhishek Majumder, Chhabi Rani Panigrahi, Sudipta Roy, and Bibudhendu Pati. 2023. Tourism recommendation system: A survey and future research directions. *Multimedia tools and applications* 82, 6 (2023), 8983–9027.

[30] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. 2025. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine* 46, 2 (2025), e70002.

[31] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*. 154–162.

[32] Ali Tourani, Hossein A Rahmani, Mohammadmehdi Naghiaei, and Yashar Deldjoo. 2024. CAPRI: Context-aware point-of-interest recommendation framework. *Software Impacts* 19 (2024), 100606.

[33] Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. 2024. Putting Popularity Bias Mitigation to the Test: A User-Centric Evaluation in Music Recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 169–178.

[34] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on learning theory*. PMLR, 25–54.

[35] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 325–334.

[36] Jia-Dong Zhang, Chi-Yin Chow, and Yanhua Li. 2014. Lore: Exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems*. 103–112.

[37] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *proceedings of the 30th acm international conference on information & knowledge management*. 4653–4664.

[38] Xu Zhou, Zhuoran Wang, Xuejie Liu, Yanheng Liu, and Geng Sun. 2024. An improved context-aware weighted matrix factorization algorithm for point of interest recommendation in LBSN. *Information Systems* 122 (2024), 102366.