Behaviour-aware Tourist Profiles Data Generation

Pavel Merinov, David Massimo and Francesco Ricci

Free University of Bozen-Bolzano, Italy

Abstract

We propose a computational model to synthesise individual-level user profiles from scarce population-level data in tourism domain. Namely, our model exploits, as input, summary information about the items (and item attributes) selected by users, and, as output, builds individual-level user profiles that respect the provided input. As a key contribution, we utilise discrete choice behavioural model to conjoin (via chi-square divergence minimisation) the choices made by the synthesised population of users with the choices observed in the real data. To showcase our idea, we release a code and a dataset that includes synthesised profiles of 10,000 users that interacted with circa 200 items.

Keywords

population synthesis, aggregated data, dataset, user modelling, tourist profiles, operations research

1. Introduction

Research on recommender systems (RSs) in tourism domain relies significantly [1, 2] on the quality of available data of tourist movements, including details about points of interest (POIs, items) and the preferences of tourists (users). Sadly, this data is often hard to access: either it is not available or it is sensitive due to privacy regulations. To overcome these limitations, our approach aims to synthesise tourist profiles – preferences towards POIs attributes – so that the tourists choices based on these profiles are in statistical agreement with the actual POIs visit popularity distribution. We make two key contributions. First, we propose a computational model for generating individual-level tourist profiles by exploiting assumptions about rational tourist behaviour. Second, we release a source code and a dataset with generated tourist profiles as the study artefacts. The implementation is publicly accessible to be adapted for similar use cases in the tourism domain.

2. Related work

Our research on user profiles synthesis relates to the class of Synthetic Reconstruction techniques [3] and the class of Copula-based Population Generation techniques [4, 5]. According to the literature, a typical population reconstruction model goes through a two-step process. At the first step, a joint distribution of attributes is fitted to match known marginal sums, generally requiring a small real sample of data. At the second step, simulated users and items are sampled from an estimated joint distribution. Overview of reconstruction methods is presented in a

IIR2023: 13th Italian Information Retrieval Workshop, 8th - 9th June 2023, Pisa, Italy

pmerinov@unibz.it (P. Merinov); davmassimo@unibz.it (D. Massimo); fricci@unibz.it (F. Ricci)

survey [6]. Our research also relates to the area of user behaviour (choice) models. Knowledge about how a user chooses can be exploited to enrich the population synthesis process; user behaviour, items visit popularity distribution, and user profile are jointly dependent. Choice models are typically based on the assumption that the behaviour of the decision maker maximises the own utility [7]. In pursue to answer a question on how to generate high-fidelity tourist profiles, our approach links together (1) reconstruction methods with (2) user behaviour models. We apply utility maximisation framework and solve the matching problem between predicted and marginal sums. Technical details are provided in the next section.

3. Mathematical model

Our model requires structured input that can be tweaked depending on target constraints of the tourism environment. In general, the computational model relies on (1) prior knowledge about the environment presented in a form of summary tables, (2) POI attributes model, and (3) choice model that determines which POIs a tourist with a given profile will visit. As output, the computational model generates tourist profiles that set tourist preferences towards POIs. We discuss the algorithm that links input and output together in a unified optimisation framework.

3.1. Summary tables of tourist visits

The computational model by design relies on two marginal distributions that provide a summary of tourist behaviour. The first marginal sum is a POIs visit popularity distribution q^*. We model a distribution q^* as a long-tail with probabilities $q^*_j \propto 1/r_j$, where r_j is an assigned rank for j-th POI and the most popular one has rank 1. In a population of N users and J POIs on average $q^*_j N$ users visit the j-th POI. Importantly, probabilities do not sum to 1, as users can visit more than one POI. Second marginal sum is a tourist level of activity a^*. This is the percentage of tourists who visit a given number of POIs during a trip. We model a discrete distribution a^* as a non-negative probability mass vector over possible number of visited POIs. Similarly, we assume that on average $a^*_t N$ users visit exactly t POIs during a trip (see Figure 1 and Table 1).

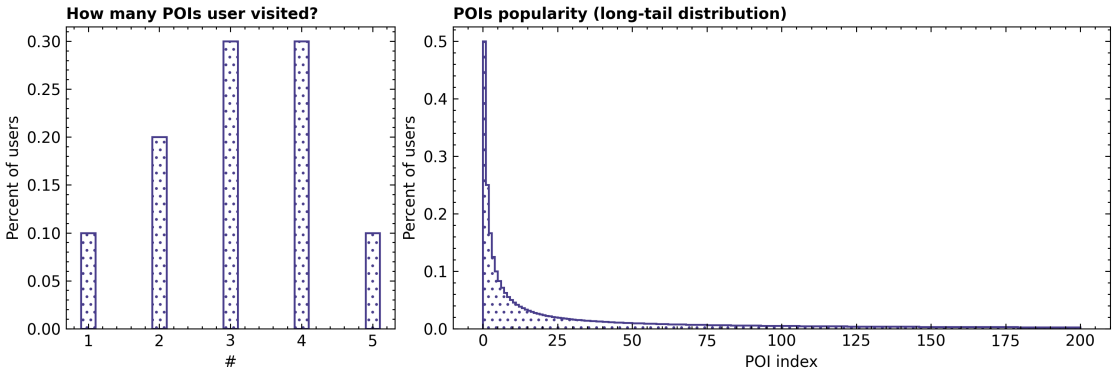

Figure 1: Marginal sums. Left figure shows tourist activity distribution (a^*). Here, tourists visit from 1 to 5 POIs during their trip. Right figure shows long-tail POIs popularity distribution (q^*). The first POI gets the attention of half the visitors, the last is about 200 less popular than the first

3.2. POI profile

Each POI j has a unique representation f_j that we model as a vector with d components $(f_{j1}, f_{j2}, \ldots, f_{jd}) \in [0,1]^d$ that describes a given POI. Each component determines (on a scale of 0 to 1) how much each attribute (POI category or feature) is present in this POI. By design, if POI profile f_j information is available from domain experts, it can be used as-is in our computational model. Otherwise, we run a procedure to create POI profile from scratch. While there is freedom on which attributes to use, the joint distribution of attributes is not arbitrary. First, we assume that each POI should belong to one or more categories that inherit similarities: all POIs within one category share common attributes. Second, a POI profile tend to be sparse, reflecting the presence of only a few attributes. In our case study we generated POI attributes that respect established assumptions, corresponding correlation structure is shown in Figure 2.

Table 1
Summary of the simulation parameters

Parameter	Value	Parameter explanation
d	64	dimension of hidden attributes
J	200	number of items (POIs) in the universe
N	10000	number of users (tourists) in a population
beta	6	inverse temperature
lambda wg and lambda bg	0.001	within and between groups regularisation coefficient
G	15	number of groups for users (typologies)
g^* vec	see source code and [8] for details	sizes of user groups
α^* vec	(0.1, 0.2, 0.3, 0.3, 0.1)	users exploration activity marginal sum
q^* vec	0.5*(1/r)	items popularity marginal sum

3.3. Tourist profile

Each tourist n has a unique representation z_n that we model as a vector with d components $(z_{n1}, z_{n2}, \ldots, z_{nd})$ in the same vector space as POIs. Components of this vector represent preferences towards corresponding POI attributes. From tourism literature, tourists can be grouped on the basis of their preferences [8]. In our case study we assume that tourists can be divided in G groups, such that tourists within a group are more look alike rather than between groups: preferences within a group are distributed (no correlation between preferences) Gaussian around group centroid. Looking ahead, optimised tourist profiles – the output of computational model and the outcome of our case study – are shown in Figure 2.

3.4. Choice model

Each user, choosing an items, tries to maximise their own utility. Choice protocol is as follows, user n estimates a linear utility $u_{nj} = z_n^T f_j$ for each item j in the collection of POIs and then selects top t (in terms of utility) items from this collection. Parameter t reflects how many items will be selected by the user during his/her trip (in our example Figure 1 it can be from 1 to 5). This parameter usually depends on available time budget, time required to visit a POI, and distances between POIs. Even this simple choice model makes synthesis of tourist profiles – our main goal in this study – very difficult, since user decisions are based on a non-differentiable

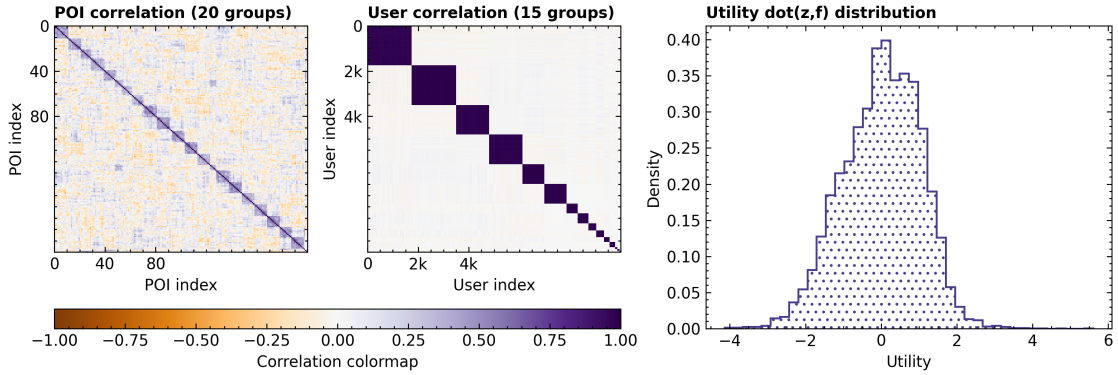

Figure 2: POI attributes and learned tourist profiles. Left figure shows 200 POIs grouped into 20 clusters. Middle figure shows a population of 10000 users grouped into 15 typologies: within each group users behave alike (and share common preferences). Right figure shows utility distribution in this population

operator arg max. In the next section we discuss an approach to relax this operator and, as a consequence, simplify the synthesis problem.

3.5. Link model

The aim of this research is to synthesise a population of N users $\{z_n\}$ that interact with a collection of J items $\{f_j\}$ in accordance with utility $u_{nj} = z_n^T f_j$ maximisation framework. It can be considered as an optimisation problem over dN parameters. Namely, we are searching for user profiles such that user choices are compatible with POIs visit popularity distribution q^*: popular items appear more often in the top choices because the utility of popular items should be higher than the utility of long-tail items. To fit the population we optimise the quadratic loss:

$$L_z = \sum_{j=1}^{J} (p_j - q_j^*)^2 / q_j^* + R_z \longrightarrow \min_{z_1, z_2, \ldots, z_N}$$

$$R_z = \sum_{g=1}^{G} \lambda_{wg} \sum_{n=1}^{N} \mathbb{1}_{\{n \in g\}} \cdot \|z_n - \bar{z}_g\|^2 + \lambda_{bg} \sum_{h>g}^{G} \operatorname{cossim}(\bar{z}_g, \bar{z}_h).$$

p_j is the probability that the j-th item is chosen. We approximate p_j as the fraction of users who prefer the j-th item based on utility maximising choice model. Regularisation R_z contains two parts: the first part imposes Gaussian prior on user profiles within each cluster g with centroid $\bar{z}_g$, and the second part penalises for similarities between each two cluster centroids $\bar{z}_g$ and $\bar{z}_h$. To optimise this quadratic loss with standard gradient-based techniques we re-parameterise (to relax arg max structure of choice model) probability vector for each component p_k as:

$$p_k = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} a_t^* \cdot \left(1 + \sum_{j=t+1}^{J} \exp(\beta u_{n(j)} - \beta u_{nk})\right)^{-1}$$

$u_{n(j)}$ means that the sum is taken over an ordered in decreasing order subsequence of $u_{n(1)}, u_{n(2)}, \ldots, u_{n(J)}$ starting from index $t+1$. For high β values, this estimates the share

of users for whom item k belongs to top-t choices, i.e., $u_{nk} > u_{n(t+1)}$. Expectation over distribution a^* (over a possible number of visits t) reflects the fact that users, even with the same preferences z_n, may visit different number of POIs during a trip depending on the available time budget.

4. Experimental setup

To showcase our approach, we synthesised tourist profiles. Table 1 summarises experimental setup. Figure 2 shows correlation structure of the synthesised tourist profiles. Figure 3 shows reconstructed tourist choices: each black dot is a POI (choice) favoured by a particular tourist according to his/her utility model, while the right side portrays marginalised over all tourists choices. Marginalised fit – our measure of goodness of fit – is consistent: proposed computational model learned tourist profiles in such a way that corresponding tourist choices match with the ground truth visit popularity distribution, small deviations are due to substitution of combinatorial hard-argmax optimisation problem (that is intractable) with a differentiable problem with soft-argmax indicators. The source code is available on Github[1].

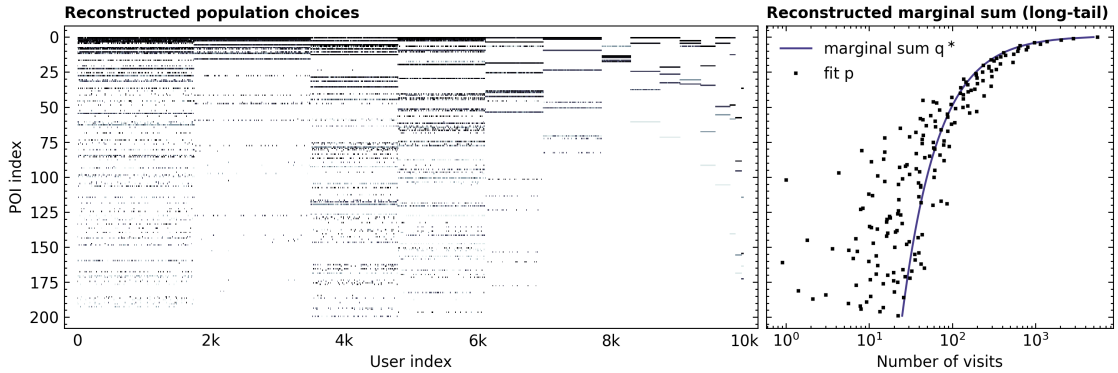

Figure 3: Learned tourist profiles. Left figure shows tourist choices pursuing utility-maximising behaviour: choices within a group tend to be similar, choices between groups are more diverse. Right figure shows that aggregated over tourists (marginalised) choice profile resembles original distribution

5. Conclusions

We have proposed a computational model to synthesise tourist profiles from aggregated data and released a synthetic dataset with a source code to generate it. Table 1 shows our configuration, which can be tweaked based on the target tourism environment. Our model exploits gradient-based optimisation, making it scalable to very large environments (millions of tourists and thousands of POIs). Further extensions require the close collaboration with experts in the domain to expand knowledge about the tourist population: improve user generation process (enforce correlation structure), and support context-aware user behaviour.

[1]https://github.com/pashaPASHaa/tourist-profiles-data-generation

References

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

[2] P. Merinov, D. Massimo, F. Ricci, Sustainability driven recommender systems, in: G. Pasi, P. Cremonesi, S. Orlando, M. Zanker, D. Massimo, G. Turati (Eds.), Proceedings of the 12th Italian Information Retrieval Workshop 2022, Milan, Italy, June 29-30, 2022, volume 3177 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3177/paper22.pdf.

[3] J. Barthélemy, P. L. Toint, Synthetic population generation without a sample, Transp. Sci. 47 (2013) 266–279. URL: https://doi.org/10.1287/trsc.1120.0408. doi:10.1287/trsc.1120.0408.

[4] Z. Li, Y. Zhao, J. Fu, Sync: A copula based framework for generating synthetic data from aggregated sources, in: G. D. Fatta, V. S. Sheng, A. Cuzzocrea, C. Zaniolo, X. Wu (Eds.), 20th International Conference on Data Mining Workshops, ICDM Workshops 2020, Sorrento, Italy, November 17-20, 2020, IEEE, 2020, pp. 571–578. URL: https://doi.org/10.1109/ICDMW51313.2020.00082. doi:10.1109/ICDMW51313.2020.00082.

[5] F. Benali, D. Bodenes, N. Labroche, C. de Runz, Mtcopula: Synthetic complex data generation using copula, in: K. Stefanidis, P. Marcel (Eds.), Proceedings of the 23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP) co-located with the 24th International Conference on Extending Database Technology and the 24th International Conference on Database Theory (EDBT/ICDT 2021), Nicosia, Cyprus, March 23, 2021, volume 2840 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 51–60. URL: https://ceur-ws.org/Vol-2840/paper8.pdf.

[6] K. Müller, K. W. Axhausen, Population synthesis for microsimulation: State of the art, Arbeitsberichte Verkehrs-und Raumplanung 638 (2010).

[7] K. E. Train, Discrete choice methods with simulation, Cambridge university press, 2009.

[8] H. Gibson, A. Yiannakis, Tourist roles: Needs and the lifecourse, Annals of tourism research 29 (2002) 358–383.