Identifying and Addressing Disparities in Public Libraries with Bayesian Latent Variable Modeling

Zhi Liu[1, 2], **Sarah Rankin**[2], **Nikhil Garg**[1]
[1]Cornell Tech
[2]The New York Public Library
zl724@cornell.edu, sarahrankin@nypl.org, ngarg@cornell.edu

Abstract

Public libraries are an essential public good. We ask: are urban library systems providing equitable service to all residents, in terms of the books they have access to and check out? If not, what causes disparities: heterogeneous book collections, resident behavior and access, and/or operational policies? Existing methods leverage only system-level outcome data (such as overall checkouts per branch), and so cannot distinguish between these factors. As a result, it is difficult to use their results to guide *interventions* to increase equitable access. We propose a Bayesian framework to characterize book checkout behavior across multiple branches of a library system, learning heterogeneous book popularity, overall branch demand, and usage of the online hold system, while controlling for book availability.

In collaboration with the New York Public Library, we apply our framework to granular data consisting of over 400,000 checkouts during 2022. We first show that our model significantly out-performs baseline methods in predicting checkouts at the book-branch level. Next, we study spatial and socioeconomic disparities. We show that disparities are largely driven by disparate use of the online *holds* system, which allows library patrons to receive books from any other branch through an online portal. This system thus leads to a large outflow of popular books from branches in lower income neighborhoods to those in high income ones. Finally, we illustrate the use of our model and insights to quantify the impact of potential interventions, such as changing how books are internally routed between branches to fulfill hold requests.

1 Introduction

Libraries play a key role in how humans preserve, share, and access knowledge. Public libraries in particular are an important social space and resource. Urban library systems are not monoliths, but rather spatially distributed throughout the city, with each *branch* primarily serving its neighborhood population. As a prominent example, the New York Public Library (NYPL), has 92 branches in the boroughs of Manhattan, the Bronx, and Staten Island,[1] serving roughly 867 thousand active *patrons* (users)(NYC Mayor's Office of Operations 2023); similarly, the Chicago Public Library has 81 branches, and the Seattle Public Library has 27 branches. These branches are both locally and centrally operated – e.g., branch librarians may curate their collections, but resources and books are also shared across branches.

An important component of library usage is the *holds* system: an online portal allowing patrons to request a book available at *any branch* in the system, which is then transferred to their local branch to be picked up. Such hold systems increase effective branch collection quality and thus *efficiency*: books desired in one branch are not sitting unused in another; however, as we show, such systems can also be a source for *inequity*, if usage of the holds system is disparate.

We ask: are urban library systems providing equitable service to all residents, in terms of the books they have access to and check out? If not, what causes disparities: heterogeneous branch collection quality, availability, resident behavior and access, and/or operational policies? Existing research on library equity largely focuses on high-level questions, such as large-scale disparities in funding (Sin 2011) or accessibility (Cheng et al. 2021; Guo, Chan, and Yip 2017) across an entire country or region, or comparison between rural and urban libraries (Real, Bertot, and Jaeger 2014).

While essential, such analyses are insufficient to inform operational decisions, such as how much to invest to improve collection quality in each branch, and how to fill hold requests across branches. In particular, coarse analyses *confound* many potential reasons that there may be fewer checkouts by patrons of one branch than another:

(a) Collection desirability. Does a branch collection contain books that potential patrons desire to read?

(b) Latent demand. What is the overall latent demand for the branch, due to factors such as population size, geographic distance to the branch, and outside options for books (such as personal purchases)?

(c) Book availability. Are desired books actually available for checkout by branch patrons? If not, where is the book?

(d) Differential mode usage. Books can be checked-out via different means. Patrons using the **hold** system have access to books from *any* branch in the system and can request books online. Other patrons physically visit their local branch and **browse** books on the shelves. Differential mode usage (due to, e.g., mode familiarity or access to home internet) can induce disparities, due to de facto access to different collection sizes and qualities, as well as how such holds lead to transfers between branches.

[1]The boroughs of Queens and Brooklyn have separate systems.

In collaboration with NYPL, we develop a Bayesian model, informed by two crucial desiderata: (a) Disambiguating the above factors, which is necessary to design effective *interventions* to increase equity; (b) Efficiently leveraging granular operational data available to many library systems, without sacrificing individual patron privacy. Our model contains latent parameters for each *book* (representing overall popularity) and *branch* (characterizing overall demand and mode usage). We leverage daily availability and checkout information for each book at each branch and fit the model using Markov Chain Monte Carlo (MCMC).

After fitting our model, we apply it to characterize library usage patterns and potential sources of disparities, including branch-specific book collection quality and mode usage. In particular, we find that *differential mode usage* leads to a large outflow of popular books from branches in lower income neighborhoods to those from high income ones; thus, even though there is little disparity in *branch collection quality*, there are large disparities in de facto *book availability and checkouts* for patrons at different branches.

Finally, we use our model and calibrated simulations to inform policy decisions. For example, practitioners at NYPL are interested in optimizing the *paging* list – when a hold is placed, from which branch should the book be pulled? We show that – by protecting pulls from branches with relatively high *browser* usage – the system can mitigate much of the above disparities, without sacrificing the efficiency benefits of the holds system. We are discussing the implementation of this intervention with our NYPL collaborators.

Supplementary analyses and code can be found at https://github.com/ZhiLiu724/library_disparities.

Related work This work relates and contributes to the literature on disparities in the allocation of resources to and the usage of libraries, concerns of digital inclusion and exclusion (Thompson et al. 2014), and the role of digital divide in knowledge acquisition (Van Dijk 2020), by developing a model that can disambiguate various potential sources of inequity, thus guiding specific interventions.

We analyze how books flow between branches, as a result of patron behavior. Such dynamics also occur in the sharing economy, such as ride-hailing and bike sharing. For example, one line of literature in AI explores spatial characteristics of usage and develops optimization approaches to rebalance bikes across stations; see, e.g., Singhvi et al. (2015); Singla et al. (2015); Pan et al. (2019)). For example, O'Mahony and Shmoys (2015) similarly need to account for censored demand estimates due to station capacity. To this literature, our work contributes a focus on how usage dynamics can create disparities across branches, and how disparities can be mitigated by system design.

More broadly, our work connects to works that quantify and address spatial and demographic equity in urban systems, such as in crowd-sourcing hotlines (Liu, Bhandaram, and Garg 2023; Agostini, Pierson, and Garg 2024; Laufer, Pierson, and Garg 2022), policing (Goel, Rao, and Shroff 2016), ride-sharing and other emerging mobility (Yan and Howe 2020). Methodologically, this work also connects to literature in machine learning and AI that models complex data-generating processes in a Bayesian manner, producing rich yet empirically tractable models. These include many contexts, most notably in disease outbreak identification (Wilder, Mina, and Tambe 2021; Campbell et al. 2019) and criminal justice (Pierson, Corbett-Davies, and Goel 2018).

2 Model

Whether a patron checks out a given book on a given day is a complex process that depends on their demand for the book, their propensity to check out books at their local branch versus placing a hold through the online system, and the specific books available to them via either mode. We develop a model to distinguish between these factors, informed by our available granular data.

Consider a library with N *branches* and M unique *book titles*. Each branch has its own collection, available to be checked out in person. A branch may have multiple copies of the same book title, and multiple branches may have copies of the same title. Our data indicate two type of checkouts: **browser checkouts** and **hold checkouts**. Browser checkouts are the traditional way of using the library: patrons browse the currently available collection of a branch in person and check out books. Hold checkouts originate from a 'hold' placed by a patron using an online system, for a book available at *any* branch in the system. After a hold is placed, the library transfers the book to a branch of the patron's choice, provided that it is available in at least one branch.[2] The patron is notified when it is available for pick up and checks out the book.

We develop a model for how these two types of checkouts are generated. For each branch $b \in [N]$, we associate it with two parameters. Parameter $p_b \in [0,1]$ characterizes the **patron demand size** of a branch (representing overall latent demand for books); parameter $h_b \in [0,1]$ characterizes the **hold usage fraction**, the fraction of that branch's demand which is through the holds system. Each title $t \in [M]$ has a parameter $d_t \in [0,1]$, representing title **desirability**. Our generative model at each branch b and title t is as follows.

1. If title t is available at branch b, *at least one* browser checkouts happen with probability
$$\mathbf{p}_{bt}^{\text{browser}} = p_b(1-h_b)d_t.$$
If not available, a browser checkout does not happen.
2. If title t is available in the system (that is, available in *at least* one branch, including branch b), *at least one* hold checkout at branch b happens with probability
$$\mathbf{p}_{bt}^{\text{hold}} = p_b h_b d_t.$$
Otherwise a hold checkout does not happen.

We assume that checkouts are independent across days, books, branches, and checkout types, *conditional on capacity*: intuitively, this means that if one book is checked out at a branch today, it should not affect other checkouts (on any day, at any branch, for any title), except through its effect on the checked out title's availability.[3]

[2] If not available, the patron enters a queue, that is served as first-come-first-served. We do not observe the queue, only checkouts.

[3] This assumption does not hold if, for example, patrons *substitute* their demand for one book with another.

The data includes the following: for each branch and title pair $\{b, t\}$, we observe the number $K_{b,t}$ of days that the book is available at branch b; of these $K_{b,t}$ days, $x_{b,t}^{\text{browser}}$ denotes the number of days with browser checkouts. We further observe $L_t \geq K_{b,t}$ days that the book is available anywhere in the system; of these days, hold checkouts occur on $x_{b,t}^{\text{hold}}$ days. Our model is thus:

$$x_{b,t}^{\text{browser}} \sim \text{Binomial}(K_{b,t}, \mathbf{p}_{bt}^{\text{browser}}),$$
$$x_{b,t}^{\text{hold}} \sim \text{Binomial}(L_t, \mathbf{p}_{bt}^{\text{hold}}).$$

We also consider a **zero-inflated** variant (Hall 2000), where the number of days with hold checkouts instead follows

$$x_{b,t}^{\text{hold}} \sim \text{Bernoulli}(1-\gamma) \times \text{Binomial}(L_t, \mathbf{p}_{bt}^{hold}).$$

This corresponds to a data-generating process that only a fraction γ of days, we would see hold checkouts. This attempts to capture the complex aspects of back-end operations associated with the hold system not incorporated in the model, for example, when a book is being processed or in transit, and thus is actually not available.

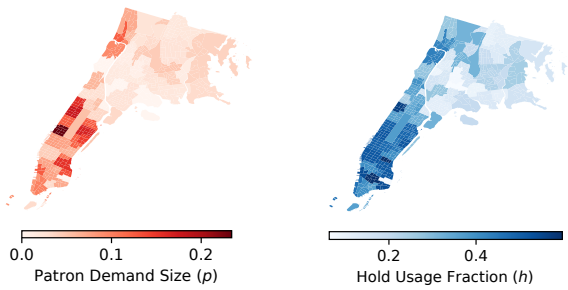
Research questions We jointly estimate the three sets of parameters – *patron demand size* $\{p\}$, *hold usage fraction* $\{h\}$, and *book desirability* $\{d\}$ – given data $\{x^{\text{browser}}\}, \{x^{\text{hold}}\}, \{L\}$ and $\{K\}$.[4]

Estimating these parameters enables analyzing usage and the *causes* of disparities. For example, in our empirical application, we find that there are only small differences in the quality of book collections in each branch; however, heterogeneous hold usage across branches (in ways that correlate with socioeconomic characteristics) leads to a large outflow of books from libraries in lower-income areas to those in high-income areas. As highlighted by our NYPL collaborators, such a model also provides counterfactual estimates: suppose a book was additionally available at a given branch for a period of time, how many more browser checkouts would we expect? Such analyses inform collection planning and, as we analyze, interventions to decrease disparities.

Modeling choices and limitations Our model is purposely simplified, to focus our analysis on our collaboration-informed research questions.

First, we learn single-dimensional branch demand parameters and title desirability scores – instead of higher dimensional embeddings as in a matrix factorization model. This structure provides parameter interpretability, and best describes the interplay of hold usage and overall demand size in determining the number of checkouts. Our empirical results suggest that this structure already leads to a good fit of the observed data, though extensions, where the set of parameters p, h, and d are higher-dimensional, could be of interest.[5] Similarly, we do not model an individual's checkout dynamics via a *discrete choice modeling* approach, in which books may substitute or complement each other. While such modeling is important for questions such as *which specific* books each branch should acquire or display, we believe that it is less relevant for our research questions.

Second, we fit the probability of observing *at least one* checkout for a given book and branch, as opposed to fitting the *number* of checkouts per day that the book is available. In our dataset, we rarely observe more than a single checkout for a given branch-title pair in a day; of all the days that we observe checkouts of a given title-branch pair, we observe only one checkout in 96.96% of the days; and two checkouts in 2.67% of the days; with more than two checkouts in only the remaining 0.37%. Thus, characterizing *whether* checkouts appear approximates the *number* of checkouts, without needing to model multiple checkouts as a function of the number of copies available for that title in that branch.



(a) (b)

Figure 1: Patron demand size and hold usage fraction parameter estimates in each branch, plotted by branch service area, in Manhattan (lower left) and the Bronx (upper right). The map with Staten Island is in Appendix Figure 8. Higher patron demand size (p) indicates a higher level of branch usage, and a higher hold usage fraction (h) indicates that a higher *fraction* of checkouts are from holds rather than browsing. We observe that these parameters are correlated: the same branches that have higher overall usage also have higher hold fractions. At a high level, branches in Manhattan (more socioeconomically privileged than the Bronx) have higher demand, *and* more of this demand is from holds.

3 Empirical Methods and Data

3.1 Data Pre-processing

The raw data consists of two parts. The NYPL actively maintains a transaction log that keeps track of all transactions (check-in, checkout, renewal, etc.) including, for each transaction: the item copy involved (including a unique ID, and the copy's home branch), the branch where it was checked out, and the time. Additionally, every month the NYPL takes a 'snapshot' of their entire circulating collection and documents the availability status of each item copy in each branch. We focus on data from calendar year 2022, during which most branches have (at least partially) recovered from the impact of the COVID-19 pandemic on browsing.[6] No

[4] In the **zero-inflated** model, we additionally estimate γ.

[5] Expanding on this idea, we additionally train a hierarchical Bayesian model, the results of which are presented in Appendix A.5, and do not differ much from the results we present in the main text

[6] There are time-varying effects that we do not model. For example, there was a clear Omicron effect in Winter 2021-22 where

Variable	Coef.	p-value	r^2
log(Population)	0.011	0.269	0.029
Female fraction	0.449	0.041	0.053
White fraction	0.116	<0.001	0.428
log(Median income)	0.061	<0.001	0.455
English speaker fraction	0.153	<0.001	0.317
Spanish speaker fraction	-0.125	<0.001	0.319
Poverty fraction	-0.242	<0.001	0.279
No Internet access fraction	-0.535	<0.001	0.402
Average household size	-0.066	<0.001	0.345

(a) Coefficients w/ patron demand size (p) as response.

Variable	Coef.	p-value	r^2
log(Population)	0.000	0.457	0.007
Female fraction	-0.087	0.899	0.000
White fraction	0.296	<0.001	0.291
log(Median income)	0.175	<0.001	0.397
English speaker fraction	0.438	<0.001	0.270
Spanish speaker fraction	-0.396	<0.001	0.335
Poverty fraction	-0.649	<0.001	0.210
No Internet access fraction	-1.384	<0.001	0.283
Average household size	-0.269	<0.001	0.596

(b) Coefficients w/ hold usage fraction (h) as response.

Table 1: Regression coefficients of socio-economic variables, when included in a linear regression one-by-one (with an intercept) with the patron demand size (p) or hold usage fraction (h) as the response variable. These results show largely consistent differences across demographic groups: i.e., higher median income areas both have larger overall demand, *and* a higher *fraction* of their demand is via hold usage. One outlier is the fraction of female residents in the service area: though significantly (p-value $< .05$) associated with the patron demand size, it is not correlated with hold usage.

patron-identifying data is included.

We filter the data to focus on our research questions. We only include transactions involving books, rather than other material types such as DVDs or periodicals. We filter out all books intended for children or young adults, for which the hold usage pattern differs from books intended for adults. Finally, we filter out branches that were closed during the majority of 2022, since any checkouts during closure are uncharacteristic of normal patron behavior.

After the above filtering, we are left with 300,389 book titles across 84 branches, for which 1,399,351 checkouts happened during 2022. However, the vast majority of these books have very few checkouts, and are only available in a very limited number of branches, as shown in Appendix Figures 6 and 7. Crucially, more than 130,000 book titles were never checked out during 2022, and more than 170,000 books titles were only available at one of the 84 branches. Including these books in our estimation would affect identification (between these book parameters and the branch parameters) and would lead to a large number of d parameters having very low values, of little relevance in practice.

As practitioners would like to primarily focus on books that are of higher interest when measuring collection equity, in our primary analyses we further only include books that were available in *20 or more branches*. This leads us to 3809 book titles (1.27% of total) for which a total of 433,998 checkouts (31.01%) occurred. We note that this filter retains almost all the popular books (Appendix Figure 6).

In the Appendix, we reproduce the results with a set of 4000 book titles that were sampled out of the 84,485 books available at 3 or more branches, with sampling weights proportional to their number of copies. This randomly sampled set of books accounts for only 187,286 checkouts (13.38% of all checkouts) but is more representative of all books. Results are qualitatively similar.

For each title-branch combination, we calculate the number of days the book was available as follows. The library-initiated collection snapshots give availability for each title at each branch monthly. *Within* a month, we iterate through the transaction logs, and for each relevant transaction (checkout and check-in), we update the availability of the title at the branch. We verify that our calculated availabilities approximately match the library's monthly snapshots, with a few discrepancies due to factors like transit time.

This process produces an aggregated dataset where each data point represents a title-branch combination with columns: the IDs of the title and the branch, the number of days the title is available at the branch, and the number of days with hold/browser checkouts. This amounts to 319,956 entries, of which 75% were randomly split into the training set with the rest 25% as the test set.

As the NYPL does not store any individual patron-level demographic information, we conduct socioeconomic analyses in terms of the demographics of the NYPL-provided service areas for each branch, using Census data.

3.2 Model Training and Analysis

We fit the models using Stan (Carpenter et al. 2017), a Bayesian probabilistic programming language that uses Hamiltonian Monte Carlo for posterior inference. Stan code for the Zero-inflated model is in the Appendix, along with all prior information. Models were trained using 500 warm-up iterations and 100 sampling iterations on 4 chains. We observe a maximum $\hat{R}$ of 1.03 in both models for the parameters p, h, and d, indicating convergence. We use draws of the parameters from the posterior distribution to predict the number of browser/hold checkouts at test-time, to evaluate model performance.[7]

We compare our models against two baseline models. The first baseline model is a 'pooled regression': using the number of total checkouts at each branch for each book title as the response, and regressing against branch and book title fixed effects, alongside an intercept term; the second naive method uses the same branch and book title fixed effects as predictors, while training two separate regressions using browser and hold checkouts as responses. We present model results in Appendix Tables 2 (test set) and 3 (train

the share of checkouts from holds went up and then back down again across branches.

[7]The procedure is detailed in the Code Appendix. For each branch and book title, the predictions form a sample, and we use the sample mean in calculating the metrics for comparison.

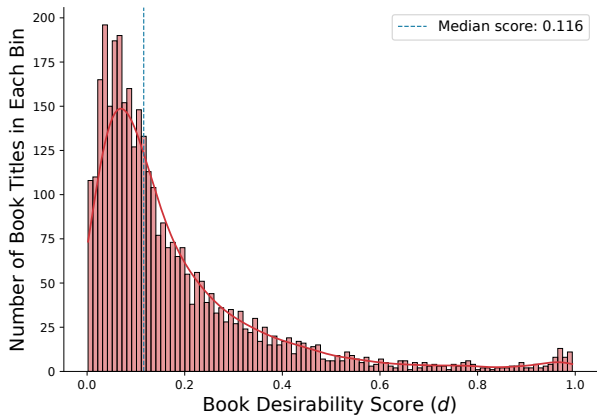


Figure 2: Histogram of desirability scores (d) of all book titles included in the model. Commercially successful or critically acclaimed books published in recent years are more inclined to have higher desirability scores, as opposed to older classics.

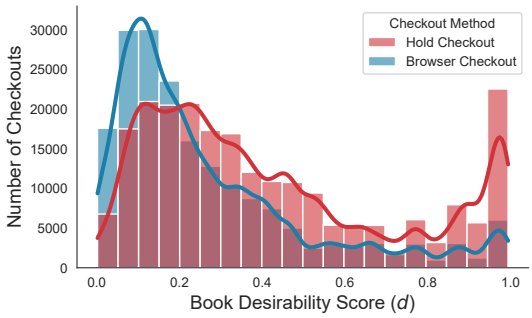


Figure 3: Histogram of checkouts according to the desirability of the book being checked out, and split by checkout method. Patrons are using holds to check out more desirable books, with books in the most desirable bin seeing number of hold checkouts 2 times more than browser checkouts.

set), respectively. Notably, our models perform substantially better than these naive approaches (in predicting overall, browser, and hold checkouts), in terms of the mean squared error, r^2, and Pearson correlation between predictions and observed values. These performance results reflect our insight that naive approaches cannot disentangle differences between browser checkouts and hold checkouts, and do not take into account the effect of book availability on usage.

Comparing our model and the **zero-inflated** variant, they perform similarly in predicting overall checkouts, while the addition of zero-inflation achieves a better balance of performance between hold and browser checkouts. Thus, for the remainder of the analysis, we only include results from the zero-inflated model.[8]

[8]The estimates themselves are very similar and highly correlated between two models, with correlation coefficients of 0.98 for d and 0.99 for p and h. All results are thus near identical.

4 Results

We first illustrate our parameter estimates and then use these estimates to document the driving factors of disparities in usage. Finally, we design an operational intervention that mitigates inequity resulting from disparate hold usage, without sacrificing efficiency.

4.1 Model Parameter Estimates

Spatial and socioeconomic differences in usage The posterior means of p and h parameters are shown spatially in Figure 1, overlayed on the defined service area of each branch. Results from linear regressions using these estimates as the response variable against these socioeconomic variables are shown in Table 1. We find that both these parameters vary spatially, and are strongly correlated with socioeconomic characteristics – branches in socioeconomically advantaged areas have both higher *overall* usage, and a higher *fraction* of their usage from the online *holds* system.

One of the strongest correlations is with (log) median income, with wealthier neighborhoods more likely to make use of the library, and more likely to use holds for checkout. This observation relates to and is qualitatively similar to a vast literature on disparities in *active usage* of public resources (e.g. public service hotlines (Liu, Bhandaram, and Garg 2023) and recreational facilities (Dahmann et al. 2010)), and to the concept of digital divide (Van Dijk 2020) that advantages patrons who have access to the use of online holds.

Title desirability The histogram of title desirability is shown in Figure 2. There is a heavy tail characteristic in the distribution of desirability scores, with most books having relatively low scores. We find that newer, more commercially successful, critically acclaimed books have higher desirability, supporting the external validity of our results.

Holds are used to check out more desirable books One long-standing concern among library practitioners is that more popular books are largely only available to hold users, as these books are often scarce and result in hold checkouts as soon as they become available. Our modeling and estimates provide a method to measure such concerns. In Figure 3, we plot the histogram of all checkouts according to the checkout mode, and the desirability of the book checked out. The results confirm the concern that patrons are using holds to check out more desirable books, and further illustrate the problem's scale: for books in the most desirable bin, with desirability scores between 0.95 and 1.00, around 78.25% of all checkouts are a result of holds.

4.2 Quantifying Potential Sources of Disparities

We now use our model estimates to measure inequity in the library system, and in particular analyze the effects of the hold usage disparities highlighted in Table 1. Our library collaborators are concerned that since hold users can potentially utilize a larger set of collections, disparities in hold usage may result in patrons from different areas receiving different levels of service quality. Our method provides a principled approach to measure such differences.

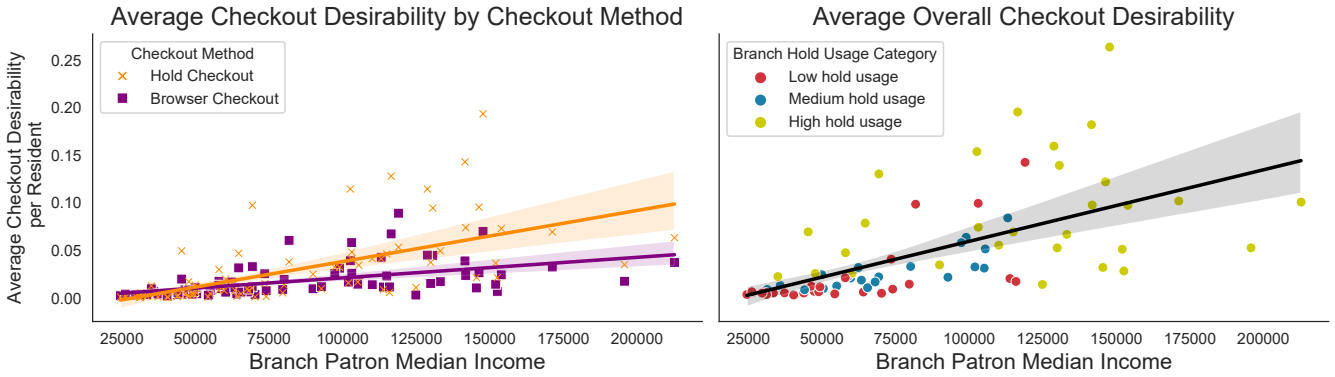


Figure 4: Left: average total desirability of books checked out per resident in each branch, split by checkout method. Right: average total desirability of books checked out per resident in each branch. High hold usage fraction in higher-income branches, combined with the finding that holds are used to checkout more desirable books, exacerbates the disparity. All three regression lines shown have p-value $< .001$. Hold usage categories are defined by dividing branches into three equal groups based on h. Regression coefficients when using median income measured in \$100,000 for browser, hold, and overall checkouts are 0.022, 0.042 and 0.075, respectively.

Small disparities in collection desirability As a first step, we measure the desirability of the collection of each branch – available for browser checkouts – to serve as the baseline of our analysis. Recall that the number of days a book t is available at branch b is denoted $K_{b,t}$; then, the average collection desirability per resident (ACDR) is:

$$\text{ACDR}_b = \frac{\sum_{t \in [M]} d_t K_{b,t}}{\text{Population}_b},$$

where the population served is calculated using the service areas of that branch. ACDR measures the average collection shared by each resident in their respective branch, irrespective of usage – all residents deserve access to a good collection, as part of their share of public goods. ACDR is compared against the median branch income in Appendix Figure 10.

We find that branch collection quality ACDR is only weakly associated with the median income of the neighborhood (p-value $= .055$).[9] However, as we show next, *shelf-availability* does not fully capture *effective-availability* – as discussed above, more socioeconomically privileged areas also make more use of the holds system.

Differences in overall *checkouts* Though the branch collections are similarly desirable on average, the actual books that were utilized by patrons exhibit substantial differences across neighborhoods of different income levels, and are exacerbated by differential hold usage.

To quantify this effect, we calculate the average total desirability of books checked out via browsing, holds, and overall, per resident in the service area. On the left of Figure 4, we find that on average, residents in wealthier neighborhoods check out more desirable books, both by holds and by browsing. Combining with the finding that wealthier neighborhoods also use the hold system more, the average overall checkout desirability exhibits more substantial disparities, as shown on the right of Figure 4 (p-value $< .001$).

Appendix Figure 9 shows the same pattern when we average over the number of checkouts instead of per resident.

Disparate hold usage *lowers effective access* due to book flows A major concern with the holds system is that some branches may *deplete* the books of other branches, leading to *effective-availability* collection inequities. In other words, high hold usage in one branch *lowers* the books available in other branches, to checkout either via holds or browsing. To quantify such flow of books due to holds, for each branch, we calculate the sum of the desirability of the books *pulled from other branches* to fill holds in that branch as the total *desirability inflow*, and the sum of the desirability of the books from the branch used to fill holds at other branches as the total *desirability outflow*. Formally, let set T index all transactions, t_i be the book title involved in transaction i, b_i^{in} be the branch where the checkout takes place, and b_i^{out} be which branch the book comes from. Total desirability inflow and total desirability outflow are defined as

$$\text{Total desirability inflow}_b = \sum_{i \in T} d_{t_i} \mathbf{1}\{b_i^{\text{in}} = b\},$$

$$\text{Total desirability outflow}_b = \sum_{i \in T} d_{t_i} \mathbf{1}\{b_i^{\text{out}} = b\}.$$

The *net desirability inflow* is the difference between total inflow and outflow. Intuitively, negative net inflow indicates that a branch's collection is being depleted in order to help other branches fill holds, and vice versa.

Figure 5 shows that net inflows vary substantially and that the scale of their effect is large. For example, the highest inflow branch has a total collection desirability score of 60,593;[10] further, this branch had a net inflow of 4,512 – this branch pulled an equivalent of 7.5% its collection from other

[9] Many factors are not included in this metric, such as average distance to the library or other dimensions of branch quality.

[10] Defined as ACDR multiplied by the branch population.

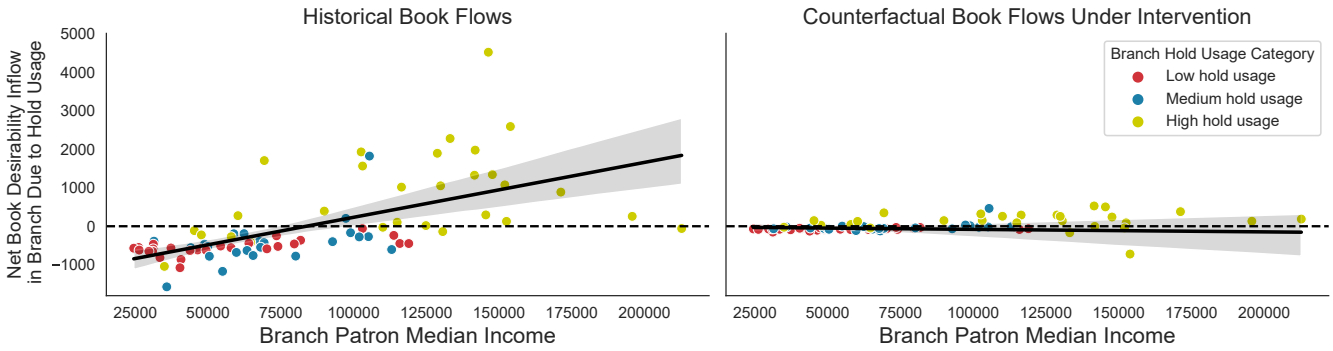


Figure 5: Left: historically observed net book desirability inflow in each branch due to hold usage. Lower-income branches with low hold usage fraction are more likely to be depleting their collection in order to help other branches fill holds (p-value $< .001$). Right: Counterfactual net book desirability inflow due to hold usage under our proposed policy. Net flow in each branch is now close to zero, and there is no observable relationship with patron median income (p-value $= .705$).

places in terms of book desirability.[11] On the other hand, many branches had substantial outflows.

Putting things together, we measure potential *sources* of disparities: (a) book collection quality available to browsers is relatively balanced across branches, only weakly correlated with median income; however, (b) more popular books are predominantly checked out via the holds system, whose usage differs substantially across socioeconomic lines; (c) as a result, the holds system depletes many (predominantly socioeconomically disadvantaged) branches of popular books, so that they are only available to hold users (predominantly from relatively privileged places).[12] How do we intervene to reduce such disparities, without losing the substantial efficiency benefits of the hold system? We explore an intervention below. We note that an equitable system would potentially have better browser-available collections in branches with a relatively higher fraction of browser users.

4.3 Intervention to Mitigate Book Flow Disparities

We now propose and evaluate a novel approach for alleviating the disparities in net book inflows caused by hold usage – re-prioritize where books are pulled *from* to fill holds – an intervention that can be easily integrated into the current library operational workflow. We simulate this proposal, using the historical transaction logs.

[11] 7.5% is an underestimate that assumes that a hold inflow is equivalent to a *single day* of being on the shelves.

[12] Note that the holds system does not, anymore, improve *ACDR* – collection quality as observed by browsers – for high-hold branches. Prior to 2019, a system called 'floating' was in place, where a book returned to any branch would stay at that branch, to be available for browser checkouts at that branch. An internal audit revealed drastic disparities in the effect of floating in conjunction with hold usage: almost all branches in low-income neighborhoods were being depleted as a result. Since then, a new system has been put into place, where each returned book copy would be transferred to its home branch, regardless of where it was returned. However, this intervention does not mitigate the fact that many books are only available via holds, disproportionately affecting non-hold users.

Historically, when a hold request is placed at a certain branch, that request is served by another branch largely at random – only considering where the book is available, irrespective of the overall flow of books happening at the branch. Our proposed policy maintains a *paging list* of branches according to their net inflow of book desirability up to then. For each day, whenever a hold request is received, we find the branch highest on the list where the book is available and use that branch copy to fill the request (i.e., pulling books from branches with the highest net inflow). At the end of each day, we update the priority list. We simulate the effect of such a policy using historical transactions and availability of books, with results in the right of Figure 5. Almost all branches have net inflows much closer to zero, suggesting a better balance of book flows, and moreover, no significant correlation could be observed between net inflows and the level of hold usage.

5 Discussion

Motivated by practical needs to quantify library usage patterns, we develop a novel approach based on Bayesian latent variable modeling that, when applied to granular operational data gathered by NYPL, identifies potential sources of usage disparities. We are discussing operational interventions informed by our analysis with practitioners at NYPL. Our method has wide applicability: hold systems are implemented in many other multi-branch public libraries (e.g., both Chicago and Seattle), and though not publicly available, granular transaction data are often kept internally. We do not have NYPL approval to share data or data processing code publicly, to maintain data confidentiality; however, all of our processing steps are described in detail for reproducibility, and we share our Stan model code.

Acknowledgements

The authors would like to thank Sidhika Balachandar, Matt Franchi, and colleagues at NYPL for constructive feedback. This work was conducted while Liu was an intern at NYPL, supported by the Siegel PiTech Impact Fellowship.

References

Agostini, G.; Pierson, E.; and Garg, N. 2024. A Bayesian Spatial Model to Correct Under-Reporting in Urban Crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Campbell, F.; Cori, A.; Ferguson, N.; and Jombart, T. 2019. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology*, 15(3): e1006930.

Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M. A.; Guo, J.; Li, P.; and Riddell, A. 2017. Stan: A probabilistic programming language. *Journal of statistical software*, 76.

Cheng, W.; Wu, J.; Moen, W.; and Hong, L. 2021. Assessing the spatial accessibility and spatial equity of public libraries' physical locations. *Library & Information Science Research*, 43(2): 101089.

Dahmann, N.; Wolch, J.; Joassart-Marcelli, P.; Reynolds, K.; and Jerrett, M. 2010. The active city? Disparities in provision of urban public recreation resources. *Health & place*, 16(3): 431–445.

Goel, S.; Rao, J. M.; and Shroff, R. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy.

Guo, Y.; Chan, C. H.; and Yip, P. S. 2017. Spatial variation in accessibility of libraries in Hong Kong. *Library & Information Science Research*, 39(4): 319–329.

Hall, D. B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4): 1030–1039.

Laufer, B.; Pierson, E.; and Garg, N. 2022. End-to-end Auditing of Decision Pipelines. In *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments*.

Liu, Z.; Bhandaram, U.; and Garg, N. 2023. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 1–9.

NYC Mayor's Office of Operations. 2023. Mayor's Management Report.

O'Mahony, E.; and Shmoys, D. 2015. Data analysis and optimization for (citi) bike sharing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Pan, L.; Cai, Q.; Fang, Z.; Tang, P.; and Huang, L. 2019. A deep reinforcement learning framework for rebalancing dockless bike sharing systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 1393–1400.

Pierson, E.; Corbett-Davies, S.; and Goel, S. 2018. Fast threshold tests for detecting discrimination. In *International conference on artificial intelligence and statistics*, 96–105. PMLR.

Real, B.; Bertot, J. C.; and Jaeger, P. T. 2014. Rural public libraries and digital inclusion: Issues and challenges. *Information Technology and Libraries*, 33(1): 6–24.

Sin, S.-C. J. 2011. Neighborhood disparities in access to information resources: Measuring and mapping US public libraries' funding and service landscapes. *Library & Information Science Research*, 33(1): 41–53.

Singhvi, D.; Singhvi, S.; Frazier, P. I.; Henderson, S. G.; O'Mahony, E.; Shmoys, D. B.; and Woodard, D. B. 2015. Predicting Bike Usage for New York City's Bike Sharing System. In *AAAI Workshop: Computational Sustainability*.

Singla, A.; Santoni, M.; Bartók, G.; Mukerji, P.; Meenen, M.; and Krause, A. 2015. Incentivizing users for balancing bike sharing systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Thompson, K. M.; Jaeger, P. T.; Taylor, N. G.; Subramaniam, M.; and Bertot, J. C. 2014. *Digital literacy and digital inclusion: Information policy and the public library*. Rowman & Littlefield.

Van Dijk, J. 2020. *The digital divide*. John Wiley & Sons.

Wilder, B.; Mina, M.; and Tambe, M. 2021. Tracking disease outbreaks from sparse data with Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4883–4891.

Yan, A.; and Howe, B. 2020. Fairness-aware demand prediction for new mobility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1079–1087.