Ordinal Classification with Probability Constraints *

Harvey Barnhard

Government programs often aim to benefit the least advantaged. In the presence of measurement error, however, even the most ardent redistribution program can backfire. When noisy, place-based measures of advantage are used to guide policy, the least advantaged might be made worse off if uncertainty is not properly taken into account. In this paper, I focus on the problem of using noisy measures to define 'best' and 'worst' sets of observations. For example, certain housing assistance programs encourage families in 'bad' neighborhoods to move to 'good' neighborhoods, where the same noisy measure is used to determine good and bad. The concern is that some good neighborhoods might be no better—or worse than—the origin neighborhoods. I show how existing rank inference methods can help solve the dual problem of maximizing social welfare subject to a constraint on the probability of a 'best' observation actually being worse than a 'worst' observation. However, existing methods are under-powered when inferring which observations are best and worst. I show that this lack of power arises because the methods do not directly aim to control the probability of pairwise directional errors between the best and worst sets. In response, I develop frequentist and Bayesian methods that increase power while continuing to control the probability of pairwise directional errors. I argue that Bayesian posterior inference is better-suited for this problem because the approach produces a tighter bound on the probability of a directional error, substantially increasing power and resulting in attractive frequentist error rates. I compare these methods both in simulations and to infer geographic patterns of intergenerational mobility across counties in the United States and across Census tracts in the Chicago metropolitan area.

Keywords: Rank inference, multiple comparisons, Bayesian inference, statistical decision theory

*__Current version__: November 30, 2021

Table of Contents

Introduction

Measurement error can compromise redistributive policy in two ways. First, measurement error in estimates used to determine the least-advantaged can lead to poorly targeted policy that fails to benefit the least-advantaged, a form of Type II error.[1] For example, household income measured by tax returns determined who received stimulus checks during the COVID-19 pandemic, but low-income non-filers were not eligible to receive these checks. Second, measurement error in the potential outcomes of a policy can lead to social outcomes that are the same or worse than the status quo, a form of Type I or Type III error, the latter sometimes referred to as a directional or Type S (for sign) error (Gelman and Tuerlinckx 2000). In this paper, I focus on settings where the same measure of advantage is used to determine the least-advantaged and to determine the policy treatments aimed to benefit them.

Place-based redistribution policies can be effective methods to improve social equity (Gaubert, Kline, and Yagan 2021). The characteristics of place can disadvantage incumbent residents through tangible forces like pollution and more intangible forces like cultural persistence. Place based policies can directly target these characteristics, improving outcomes both for current and future generations. However, place-based measures of advantage, such as poverty share or economic mobility, can be noisy due to sampling variation and within-place heterogeneity.

Some housing assistance programs use place-based measures to set restrictions on where families can relocate to use their housing vouchers. The Massachusetts Rental Voucher Program restricted eligible destination neighborhoods to those where at least 40% of the residents had incomes at or below the current Federal Poverty Line. The Creating Moves to Opportunity (CMTO) experiment provides financial assistance for moves to high-opportunity neighborhoods, defined as Census tracts with historical rates of upward income mobility in the top third of tracts in the Seattle and King County Area (Bergman et al. 2019). In such settings, the planner must be confident that families who were incentivized to move by the policy have better outcomes than if they were to have stayed in their original neighborhoods. In this paper, I focus on how the planner can control the probability of such a directional error.

I introduce a model of residential choice in which households choose where to live and a scalar, noisy measure of neighborhood quality is used to determine program eligibility. This same measure of neighborhood quality is used to restrict potential destination neighborhoods. The social planner's objective is to maximize social welfare subject to a constraint on the probability that a family will be worse off if they are induced to move by the policy. There are two important features of this model. First, there exists considerable uncertainty over the true levels and rankings of neighborhood quality. Second, households are heterogeneous in their preferences for neighborhoods. The key result of the model is that the planner's problem is solved by maximizing the number of eligible origin and destination neighborhoods subject to a constraint on the probability of pairwise directional errors across the likely-best and likely-worst sets.

I introduce computationally feasible methods for solving the stated planner's problem. The rank-inference methods discussed in Mogstad et al. (2020) help us approach the planner's problem by formalizing the inherent multiple testing problem in ranking places by noisy measures of advantage. I extend the methods of Mogstad et al. (2020) to better align the statistical machinery with the problem at hand, the practical benefit of this alignment being a tighter bound on the probability of pernicious outcomes, leading to a substantial increase in power. Existing methods also assume fixed thresholds for what constitutes the best and worst. That is, observers categorize observations into the best and worst based on pre-determined rank thresholds, such as "bottom half" or "bottom third." I show how the methods I introduce can be extended to instead let the data decide these thresholds at a given confidence level. The methods discussed in this paper complement the selective-inference procedures of Andrews, Kitagawa, and McCloskey (2019),

[1] Type I errors occur when a researcher incorrectly rejects a true null hypothesis. Type II errors occur when a researcher fails to reject a false null hypothesis. Type III errors occur when a researcher correctly rejects a false null hypothesis, but assigns the wrong direction

procedures that produce median unbiased estimates for the best observations. Conceptually, the methods in this paper are different because they focus on the classification of two sets such that all pairwise differences between the two sets (but not within the two sets) are significant in the proper direction. In this regard, these methods are closer to the literature on testing for superior performance relative to a benchmark (Romano and Wolf 2005).

The paper is organized as follows. The first section establishes a model of residential choice where a planner tries to improve social welfare subject to a constraint. The second section discusses how likely-best and likely-worst sets can be constructed such that the probability of a directional error is constrained. The third section evaluates the performance of these methods in simulations. The fourth section applies these methods to infer geographic differences in intergenerational mobility using data from the The Opportunity Atlas (Chetty, Friedman, et al. 2018).

A Model of Residential Choice

This section builds a simple model of residential choice in which policymakers are trying to maximize social welfare by subsidizing moves. This model is similar to the place-based transfer model constructed in Gaubert, Kline, and Yagan (2021) where wealth is transferred from the richest places to the poorest places. Whereas Gaubert, Kline, and Yagan (2021) focus on policies that redistribute wealth across place, the model I introduce focuses on the redistribution of *place* by subsidizing moving costs.

Household Preferences

A family of type $\theta \in \Theta$ that starts in neighborhood i and moves to neighborhood j will have household utility

$$u_j(y, \bar{C}, \theta) = U(y_j, \bar{C}_{ij}) + \varepsilon_j(\theta)$$

where $y \in \mathbb{R}^N$ is a vector of true neighborhood characteristics of the N neighborhoods. $\bar{C} \in \mathbb{R}^{N \times N}$ is the matrix of neighborhood living costs. $\bar{C}_{ii}$ represents the cost of living in neighborhood i while $\bar{C}_{ij}$ represents the living costs of neighborhood j plus the cost of moving from i to j. A family's additive idiosyncratic utility derived from living in neighborhood j is represented by $\varepsilon_j(\theta)$. We further assume that the idiosyncratic utilities are mean zero across family types θ and across locations j.

$$\mathbb{E}_\theta[\varepsilon_j(\theta)] = \int_\Theta \varepsilon_j(\theta) dF(\theta) = 0$$

Before any policy intervention, families have already solved their optimization problem. Unfortunately, families only observe noisy estimates of neighborhood quality for a given neighborhood, $\hat{y}_j = y_j + \delta_j$, so they must instead solve their optimization problem with uncertainty:

$$v(\hat{y}, \bar{C}, \theta) = \max_j \mathbb{E}[U(\hat{y}_j, \bar{C}_{ij})] + \varepsilon_j(\theta)$$

Planner's Preferences

The planner maximizes the following social welfare function:

$$S(y, \bar{C}) = \int_\Theta w(\theta) v(\hat{y}, \bar{C}, \theta) d\theta$$

where $w(\theta)$ represents the Pareto weight for families of type θ. The planner's sole policy lever is adjusting the original living cost matrix $\bar{C}$ to a living cost matrix with additional taxes and subsides, C. Before including normative constraints, the planner's problem is to maximize expected social welfare by subsidizing moving costs subject to a budget constraint.

$$\max_{C} \mathbb{E}\left[S(\hat{y}, C)\right] \quad \text{s.t.} \quad \sum_{i,j} n_{ij}(C_{ij} - \bar{C}_{ij}) \leq M$$

The budget constraint states that the grand sum of the difference in living cost matrices, weighted by the number of individuals who move from i to j, must be less than M, an exogenous budget. When M is zero, any subsidies to households living in distressed neighborhoods must be paid for by increasing taxes in advantaged neighborhoods.

Simplifying the Planner's Problem

As stated, the planner's problem is intractable with a large number of neighborhoods because the living costs of N^2 neighborhood pairs can be modified. However, the planner can justifiably constrain the problem by only subsidizing the cost of moving from the estimated set of distressed neighborhoods $\hat{W}$ to the estimated set of advantaged neighborhoods $\hat{B}$. These subsidies can be offset by increasing taxes for incumbent residents in advantaged neighborhoods. Mathematically,

$$C_{ij} < \bar{C}_{ij} \quad \text{if} \quad i \in \hat{W}, j \in \hat{B}, \quad C_{ii} > \bar{C}_{ii} \quad \text{if} \quad i \in \hat{B}, \quad \text{and} \quad C_{ij} = \bar{C}_{ij} \quad \text{otherwise}$$

The planner's problem is further simplified by only considering subsidies that equate the costs of moving to and living in $\hat{B}$ with living costs of staying in $\hat{W}$, $C_{ij} = C_{ii} = \bar{C}_{ii}$ if $i \in \hat{W}$ and $j \in \hat{B}$. Taxes can then be uniformly increased across advantaged areas to offset these subsidies. While not necessarily optimal, this policy is a simple representation of a tax schedule that aligns with general principles of redistributive justice. In this paper, I focus on the welfare of the least-advantaged, so it is assumed that the most advantaged experience only light tax consequences with respect to concave utility functions.

Because the budget constraint is satisfied by construction of C, the planner's problem is reduced to maximizing expected social welfare by selecting the estimated best and worst sets.

$$\max_{\hat{W},\, \hat{B}} \mathbb{E}\left[S(\hat{y}, C)\right] \quad \text{s.t.} \quad C = C\left(\hat{W}, \hat{B}, M\right)$$

The naive approach to maximizing social welfare is making $\hat{W}$ as big as possible, expanding housing assistance eligibility to as many families as possible, and only incentivizing families to move to the most advantaged neighborhood, $j^* = \arg\max_{j \in \hat{B}} \hat{y}$. However, multiple features of this model make the naive solution untenable. In this model, heterogeneous household preferences suggest that estimating a larger set of advantaged neighborhoods is preferable for maximizing welfare of the least advantaged. Households will only move if their overall expected utilities, including idiosyncratic preferences, are sufficiently high to induce a move. Expanding the choice set of eligible destination neighborhoods will induce more moves, allowing for a greater increase in social welfare.

Household Response

In the absence of externalities, subsidies will only benefit families that originally optimized to live in distressed neighborhoods. Such eligible families will only choose to move if there is at least one neighborhood $j \in \hat{B}$ that yields higher expected utility than living in neighborhood $i \in \hat{W}$. That is, a family will only move if $i \in \hat{W}$ and there exists a $j \in \hat{B}$ such that

$$\mathbb{E}[U(\hat{y}_i, C_{ii})] + \varepsilon_i(\theta) < \mathbb{E}[U(\hat{y}_j, C_{ij})] + \varepsilon_j(\theta)$$

Note that the decision to move will be driven not only by the cost-quality trade-off, but by idiosyncratic preferences for certain neighborhoods. The move is a "mistake" for a family if the realized utility of moving to neighborhood j is less than the realized utility if the family were to have stayed in neighborhood i.

Ordinal Classification

A family living in a distressed neighborhood might be incentivized to move to a neighborhood that is worse than the neighborhood from which they moved. The planner wants to ensure that families living in $\hat{W}$ are actually more disadvantaged than families living in $\hat{B}$. Mathematically, the goal is to say that $\hat{B} \succ \hat{W}$ according to a specified partial order relation.

Stochastic dominance is a natural partial order for thinking about inequality (Atkinson 1970). Treat neighborhood characteristics y_i as random with potentially differing distributions within non-random $\hat{W}$ and $\hat{B}$. Letting $F_{\hat{W}}$ and $F_{\hat{B}}$ be the cumulative distribution functions of y_i in $\hat{W}$ and $\hat{B}$, first order stochastic dominance states that $\hat{B} \succ \hat{W}$ if $F_{\hat{B}}(y) \leq F_{\hat{W}}(y)$ for all y with strict inequality at some y. However, the stochastic dominance partial order requires estimation of both $F_{\hat{W}}$ and $F_{\hat{B}}$ using noisy estimates $\hat{y}_i$. An empirical Bayes shrinkage approach may be used to produce estimates of these CDFs in the presence of measurement error. However, such an approach assumes that $\hat{W}$ and $\hat{B}$ are non-random, when in fact they are random sets selected by the data.

I propose an alternative partial order over all possible sets of $\hat{W}$ and $\hat{B}$ that treats $\hat{W}$ and $\hat{B}$ as random. Let τ_{ji} represent a test for whether or not neighborhood j is more advantaged than neighborhood i,

$$\tau_{ji} = \mathbb{1}\{y_j \text{ significantly greater than } y_i\}$$

Following sections discuss how to construct $\hat{B}$ and $\hat{W}$ using τ_{ij} and what test statistic to choose. I define $\succ_\alpha$ as a partial order relation over sets at confidence level $\alpha \in [0, 1]$ as follows:

$$\hat{B} \succ_\alpha \hat{W} \qquad \text{if} \qquad \tau_{ji} = 1 \text{ for all } i \in \hat{W},\ j \in \hat{B} \quad \text{and} \quad P(V' \geq 1) < \alpha$$

where $V' = \sum_{i \in \hat{W}, j \in \hat{B}} V_{ij}$ and $V_{ij} = \mathbb{1}\{\tau_{ji} = 1 \text{ and } y_i \geq y_j\}$ represents a Type I or Type III error. Intuitively, this definition states that $\hat{B}$ is greater than $\hat{W}$ if the smallest estimate of $\hat{B}$ is greater than the largest estimate of $\hat{W}$, and the probability is less than α that the smallest true value of $\hat{B}$ is greater than the largest true value of $\hat{W}$.

Ordinal Classification in the Planner's Problem

The constrained planner's problem can be mathematically stated as

$$\max_{\hat{W}, \hat{B}} \mathbb{E}\left[S(\hat{y}, C)\right] \quad \text{s.t.} \quad \hat{B} \succ_\alpha \hat{W}$$

With a large heterogeneous population, some families are bound to err and be worse off due to a move. The planner can therefore focus on the unintended consequences on the "average" family eligible for a housing subsidy. The construction of the partial order $\succ_\alpha$ allows the planner to be confident, at level α, that the average family incentivized to move by the policy will not yield lower realized utility for that family. Because the distribution of idiosyncratic utility is mean zero in every neighborhood, the average realized utility across family types for $i \in \hat{W}$ and $j \in \hat{B}$ is given by

$$\mathbb{E}_\theta[U(y_j, C_{ij}) + \varepsilon_j(\theta)] = U(y_j, C_{ij})$$

The event V_{ij} has been defined as saying that y_j is greater than y_i when in reality y_j is less than y_i. In this, context, V_{ij} is equivalent to the event where a move from neighborhood i to neighborhood j would decrease realized utility on average across families.

The planner uses noisy estimates to decide which neighborhoods are included in $\hat{W}$ and $\hat{B}$. The planner's problem boils down to how estimates of the joint distribution of neighborhood estimates, $\hat{F}(y)$, are used to select $\hat{W}$ and $\hat{B}$. Mathematically, the planner is constructing a set-valued map φ_α such that

$$\varphi_\alpha\left(\hat{F}(y)\right) = \left\{\hat{W}, \hat{B}\right\} \qquad \text{while} \qquad \hat{B} \succ_\alpha \hat{W}$$

In the following sections, I discuss direct methods of choosing φ_α using both a frequentist multiple testing approach and a Bayesian approach. The frequentist approach takes the empirical distribution of $\hat{F}(y)$ as given and focuses on the construction of φ_α. The Bayesian approach differs from the frequentist approach by first estimating the posterior distribution $\hat{F}(y)$, and then choosing φ_α to constrain the posterior probability of a pairwise directional error between $\hat{W}$ and $\hat{B}$.

Ordinal Classification

The previous section discussed the nature of the planner's optimization problem when the planner wants to be confident that intended beneficiaries are not made worse off as a result of the policy change. This section outlines methods for how the planner can solve the problem.

First, I establish notation that clarifies the underlying statistical machinery of the decision-making problem. Second, I discuss how existing multiple-testing methods provide solutions to the planner's problem but are overly conservative in satisfying the probability constraint. Third, I develop a multiple-testing method using the generalized familywise error rate that permits tighter bounds on the probability of an ordinal error. Fourth, I introduce the false discovery rate approach which increases power but fails to properly control the probability of an ordinal error. Fifth and finally, I introduce the Bayesian approach which substantially increases the power to infer best from worst while constraining the posterior probability of an ordinal error. The results of this section easily generalize to any problem in which a decision-maker wants to maximize the size of the likely-best and likely worst sets while constraining the probability of a pairwise directional error.

Setup and Notation

Let $r_x : \mathbb{N} \to \mathbb{N}$ be the unobserved bijective map[2] from true ranks of y_i to arbitrary indices so that

$$y_{r_N} < y_{r_{N-1}} < \cdots < y_{r_2} < y_{r_1}$$

We can estimate the true mapping r using the observed rank of y_i, mathematically represented as

$$\hat{r}_i = 1 + \sum_{j \neq i} \mathbb{1}\{\hat{y}_i < \hat{y}_j\}$$

However, we only observe noisy estimates $\hat{y}_i$, so we also want to create a simultaneous confidence set around $\hat{r}$ that consists of one confidence set for each observation being ranked:

$$\hat{r}_1 \in [\hat{l}_1, \hat{u}_1], \ldots, \hat{r}_n \in [\hat{l}_N, \hat{u}_N]$$

A general approach to constructing these confidence sets is to start by controlling both the Type I and Type III error rates on each pairwise comparison using two one-sided tests for the difference between the

[2]Here we are assuming that there are no "ties" in true ranks. This assumption is reasonable when the outcome of interest is continuous.

ith and jth observation.[3] Letting $\hat{t}_{ij} = (\hat{y}_i - \hat{y}_j)/\hat{\sigma}_{ij}$ be the studentized test statistic, the paired one-sided tests are

$$\tau_{ij} = \mathbb{1}\left\{\hat{t}_{ij} > c_{ij}\right\} \qquad \text{and} \qquad \tau_{ji} = \mathbb{1}\left\{\hat{t}_{ji} > c_{ji}\right\},$$

where c_{ij} and c_{ji} are constants chosen to control the Type I and Type III error rates and $\hat{\sigma}^2_{ij}$ is an estimator of the variance of the difference between $\hat{y}_i$ and $\hat{y}_j$. The naive method of simultaneously controlling the two error rates is to control the piecewise error rate (PWER) by setting c_{ij} and c_{ji} as the $(1-\alpha)$th percentile of the null distribution of the test statistics. In practice, we are accustomed to perform a multiplicity correction when using the two-sided test by instead setting c_{ij} and c_{ji} as the $(1-\alpha/2)$th percentile of the null distribution of the test statistics.

The lower and upper bounds of the confidence set around $\hat{r}_i$ are constructed by counting the number of estimates that are significantly greater and less than y_i, respectively:

$$\hat{l}_i = \sum_{j \neq i} \tau_{ij} + 1 \qquad \hat{u}_i = n - \sum_{j \neq i} \tau_{ji}$$

The general procedure to construct these confidence sets can be condensed into the following three steps: (1) Calculate all standardized pairwise differences. (2) Select c_{ij} in order to control a selected error rate. (3) Construct $\left[\hat{l}_i, \hat{u}_i\right]$ by counting the number of significant pairwise differences corresponding to the ith observation. The researcher holds a considerable amount of discretion in how the second step is approached; there are many error rates to target and multiple methods of constructing c_{ij} for each error rate.

The values c_{ij} are constructed to control a pre-specified error rate. Following the notation in the planner's problem, let V be the number of false discoveries and S the number of true discoveries

$$V = \sum_{i,j} V_{ij} \qquad S = \sum_{i,j} S_{ij}$$

where $V_{ij} = \mathbb{1}\{\tau_{ji} = 1 \text{ and } y_i \geq y_j\}$ and $S_{ij} = \mathbb{1}\{\tau_{ji} = 1 \text{ and } y_i < y_j\}$. The remainder of this section focuses on the selection of error rates and how we should interpret control of these error rates in relation to the planner's problem.

Naive Approach

Returning to the planner's problem, the naive approach to defining the mapping φ is to disregard variance components and take the point estimates as given. Then determine the likely-worst and likely-best sets by picking the bottom m and top n observations, respectively:

$$\varphi\left(\hat{F}(\hat{y})\right) = \varphi(\hat{y}) = \left\{\hat{W} = \left\{i : \hat{r}_i \geq N - m + 1\right\}, \hat{B} = \left\{j : \hat{r}_j \leq n\right\}\right\}$$

However, assigning group categories based on the raw point estimates and disregarding the uncertainty of those point estimates cannot guarantee a bound on the probability of a directional error between $\hat{B}$ and $\hat{W}$. The performance of this approach will vary greatly depending on the noisiness of the data.

Piecewise Error Rate

The simplest improvement upon the naive approach that takes into account uncertainty of the point estimates is to assuming that all test statistics $(\hat{y}_i - \hat{y}_j)/\hat{\sigma}_{ij}$ are independent and normally distributed. Then

[3] Some situations might call for only one one-sided test to be performed for each pairwise comparison, but two tests are required when controlling directional errors.

set both c_{ij} and c_{ji} to the $\left(1-\frac{\alpha}{2}\right)$th critical value of the normal distribution, and calculate the confidence sets around each estimated rank. Construct φ so that the ith neighborhood is in the "worst" set $\hat{W}$ if the upper bound on its confidence set is below the average rank, $\hat{u}_i < N/2$. The jth neighborhood is in the "best" set if the lower bound on its confidence set is above the average rank, $\hat{l}_j > N/2$

$$\varphi\left(\hat{F}(\hat{y})\right) = \varphi(\hat{y}, \hat{\sigma}) = \left\{\hat{W} = \left\{i : \hat{u}_i < N/2\right\}, \hat{B} = \left\{j : \hat{l}_j > N/2\right\}\right\}$$

Although this approach improves upon the naive approach by taking uncertainty into account, it only controls the piecewise error rate (PWER).

$$\text{PWER} = P\left(V_{ij}\right) < \alpha \quad \forall (i,j) \in \mathcal{I} \times \mathcal{I}$$

Where $\mathcal{I}$ is the set of neighborhood indices. Controlling the marginal probability of a false positive or directional error fails to guarantee the proper ordering of sets, $\hat{B} \succ_\alpha \hat{W}$. To see that this is the case, assume that the PWER has successfully been controlled at level α, $P(V_{ij} = 1) < \alpha$, for all i, j. Then

$$P\left(\sum_{i \in \hat{W},\, j \in \hat{B}} V_{ij} \geq 1\right) = 1 - P\left(\sum_{i \in \hat{W},\, j \in \hat{B}} V_{ij} = 0\right) \geq 1 - (1 - P(V_{ij} = 1))^{|\hat{W}|+|\hat{B}|} > \alpha$$

Beyond failing to control the ordinal error rate, the piecewise approach fails to capture the dependence structure between tests. By taking advantage of the dependence structure between tests, the following methods decrease the Type II error, increasing a researcher's ability to reject false null hypotheses (Romano and Wolf 2005).

Familywise Error Rate

The family wise error rate (FWER) is the probability that at least one null hypothesis out of all the pairwise comparisons is falsely rejected. In this paper, we only consider the mixed-directional familywise error rate, which includes errors of falsely rejecting true null hypotheses and errors of rejecting false null hypotheses, but assigning the wrong sign to the difference.

$$\text{FWER} = P\left(V \geq 1\right)$$

The simplest approach to controlling the FWER is the Bonferroni correction. This correction, however, is extraordinarily conservative for large N. Instead, we can use resampling methods to asymptotically bound the FWER. When controlling the FWER in this paper, I use the resampling methods discussed by Mogstad et al. (2020). Controlling the FWER of all pairwise comparisons ensures that $\hat{B} \succ_\alpha \hat{W}$, but the bound on the error is not tight:

$$\text{FWER} = P\left(V \geq 1\right) > P\left(V' \geq 1\right) \quad \text{where} \quad P\left(V' \geq 1\right) < \alpha \implies \hat{B} \succ_\alpha \hat{W}$$

The realization of the event on the left hand side of the inequality results in at least one of the N confidence sets being at least one rank too short. As the number of observations being ranked increases, this error becomes increasingly conservative. For example, suppose the planner is ranking 100 neighborhoods using a noisy measure of neighborhood quality. If observation i has a true rank of $r_i = 51$ and an estimated confidence set of $[1, 2, \ldots, 50]$, then the confidence set does not cover the true rank. Such a confidence set occurs when there exists no j such that $\hat{y}_j$ is significantly greater than $\hat{y}_i$, but 50 other observations such that $\hat{y}_j$ is significantly less than $\hat{y}_i$. Assuming a perfect ranking of true values exists, then the confidence set fails to cover the true value because there exists at least one j such that $\hat{y}_i$ was found to be significantly

greater than $\hat{y}_j$ despite the fact that $y_i < y_j$. Because directional errors are being controlled using two one-sided tests, no errors occur in isolation—the existence of one error implies that the mirrored test also resulted in a directional error (i.e. $\hat{y}_j$ was found to be significantly greater than $\hat{y}_i$).

Depending on the policy context, a one-rank error might be qualitatively significant when comparing, say, 100 observations. However, the severity of a one-rank error qualitatively decreases with the number of observations being compared. A one-rank error when comparing 1,000 observations is likely not a concern. In other words, the FWER does not qualitatively adapt to the number of observations being ranked. In order to help the FWER adapt, the only available option is to increase α along with the number of observations being compared. The *ad hoc* nature of this "solution" highlights the conceptual gap between the FWER and the probability of a pairwise directional error between $\hat{W}$ and $\hat{B}$.

There are two primary reasons why we should be careful when using the FWER to control the probability of a pairwise directional error. First, these FWER controlling methods do not adapt to the number of observations being compared. The probability of at least one false positive is proportional to the *square* of the number of observations. For every N observations that are ranked, there are $\frac{N^2}{2} - N$ unique pairwise comparisons. Twice as many tests must be performed to control directional Type III errors. Therefore, as the number of observations being ranked increases, controlling the FWER with nominal coverage α is the same as controlling the probability of a false positive among all $N^2 - 2N$ tests. Second, FWER controlling methods furnish coarse upper bounds on the probability of a pairwise directional error between $\hat{W}$ and $\hat{B}$. If the thresholds for defining $\hat{W}$ and $\hat{B}$ are separated by one rank[4], then the FWER seemingly corresponds to the relevant pairwise directional error, because it might only take one directional error for an observation in $\hat{B}$ to actually be worse than an observation in $\hat{W}$. However, for observations well within either threshold of classification into $\hat{W}$ or $\hat{B}$, many directional errors will need to occur for an observation in $\hat{B}$ to actually be worse than an observation in $\hat{W}$. The next section extends the FWER approach to improve power in settings where the thresholds defining $\hat{B}$ and $\hat{W}$ are separated by more than one rank.

Generalized Familywise Error Rate

The generalized familywise error rate (FWER$_k$) is the probability that at least k null hypotheses out of all pairwise comparisons are falsely rejected. We continue to consider directional errors among the errors being controlled.

$$\text{FWER}_k = P\left(V \geq k\right)$$

Note that when $k = 1$, the generalized familywise error rate reduces to the standard familywise error rate. Controlling the FWER$_k$ places a tighter bound on the probability of a pairwise directional error. In this paper, I control the FWER$_k$ using the resampling procedures proposed in Romano and Wolf (2007). In particular, I lower the computational burden by using the streamlined versions of their algorithms.

There are two reasons why the FWER$_k$ approach might be preferable to the FWER approach. First, when the occurrence of one error is not qualitatively significant, as when ranking thousands of observations, then it often makes more sense to control the FWER$_k$ with k set as a number that is qualitatively important. Second, the FWER$_k$ approach can better align with the policy-relevant notion of error, just as in the context of the planner's problem in this paper.

The choice of k need not be arbitrary. In fact, the k can be chosen to be as large as possible while ensuring that $\hat{B} \succ_\alpha \hat{W}$. For example, suppose that the planner is ranking ten neighborhoods and wants to determine which sets are within the top and bottom three while constraining the probability that any neighborhood in the bottom three is actually better than any neighborhood in the top three. An error occurs if a neighborhood i has a confidence set $\hat{r}_i \in \left[\hat{l}_i, \hat{u}_i\right]$ where $\hat{u}_i \leq 3$ but actually the true ranking is at least as bad as eight, $r_i \geq 8$. Such an egregious mislabeling would only occur if there were at least

[4]For example, if $\hat{B}$ and $\hat{W}$ are defined as the likely best-half and the likely bottom-half.

5 false positives or directional errors. In particular, at least five errors must have accumulated within the pairwise tests for neighborhood i. It is therefore reasonable to set $k = 5$ when controlling the FWER_k in this setting, because this would ensure that any neighborhood in the best set is significantly better than any neighborhood in the worst set at confidence level α. However, because each pairwise comparison corresponds to two mirrored tests to control for directional errors (80 tests total when 10 observations are being compared), errors will always occur in pairs. Therefore, one can increase power by instead selecting $k = 10$ when controlling the FWER_k. Note that this does not place a tight bound on the probability of such a severe misranking. With normally distributed estimates, one-rank errors might be common depending on the chosen level of α, but multiple errors for a single observation is more unlikely because such an event occurs when estimates contain sufficient noise to lead to opposite-signed differences, but the estimated standard errors are sufficiently small so that multiple tests erroneously produced significant pairwise differences corresponding to a single observation.

The current approach to controlling the FWER_k assumes pre-specified thresholds, and I propose a method for choosing an optimal k. Intuitively, expanding the threshold for what can be considered the best or worst will increase the number of observations with estimates significantly within the best and worst. Changing a top-third threshold to a top-half threshold should increase the number of observations within $\hat{B}$. But because the gap in ranks between $\hat{W}$ and $\hat{B}$ decreases as these thresholds are expanded, there is an increased probability of an ordinal error between $\hat{W}$ and $\hat{B}$. In other words, when the gap between $\hat{W}$ and $\hat{B}$ is small, only a few directional errors among all pairwise tests will lead to an ordinal error to occur between $\hat{W}$ and $\hat{B}$.

The gap in ranks between $\hat{W}$ and $\hat{B}$ can be optimized while continuing to control the generalized familywise error rate. Algorithm 1 presents a method for producing larger $\hat{W}$ and $\hat{B}$ by varying the gap between $\hat{W}$ and $\hat{B}$, all while nominally controlling the probability of an ordinal error, $\hat{B} \succ_\alpha \hat{W}$. However, this algorithm introduces a selective inference problem where the optimal k is treated as fixed, when it is really a random variable that varies depending on the noise in the data. Future work is required to determine how k can be chosen so that it does not depend on the same noise present when constructing $\hat{W}$ and $\hat{B}$.

False Discovery Rate and Empirical Bayes

The false discovery rate (FDR) is the expected proportion of falsely rejected null hypotheses out of all rejected null hypotheses.

$$\text{FDR} = \mathbb{E}\left[\frac{V}{V+S}\right]$$

Controlling the FDR is more liberal than controlling the FWER in that the FDR represents a lower bound on the FWER.

$$\underbrace{\mathbb{E}[V]/N^2}_{\text{PCER}} \leq \underbrace{\mathbb{E}\left[V/(V+S)\right]}_{\text{FDR}} \leq \underbrace{P(V \geq 1)}_{\text{FWER}}$$

where PCER represents the per-comparison error rate, an even more liberal form of the false discovery rate (Tang and Zhang 2007). Because the tests τ_{ij} are not independent, control of the FDR should be performed using methods that allow for dependency such as the methods proposed by Benjamini and Yekutieli (2001), rather than the more standard FDR controlling methods proposed by Benjamini and Hochberg (1995).

There exists a deeper connection between the FDR and the empirical Bayes method as introduced by Efron, Tibshirani, et al. (2001). Efron and Tibshirani (2002) discuss this connection in the context of genetic microarray experiments. Efron (2008) shows that the Benjamini-Hochberg procedure to controlling the FDR is equivalent to empirical Bayes shrinkage on the test statistics. The false discovery rate does not cleanly correspond to the probability of a pairwise directional error between $\hat{W}$ and $\hat{B}$, but the connection to empirical Bayes suggests that the flexibility of a fully Bayesian treatment might yield improvements.

Algorithm 1: maxFWER

Input: $\hat{y}$ and $\hat{\sigma}_{\hat{y}}$, vectors of length n, and $\alpha \in (0, 1)$.

Output: Estimated best and worst sets: $\hat{B}, \hat{W}$

begin

  $k^* \leftarrow 0$

  for $k \in \{1, \ldots, n/2\} \cap \{2\mathbb{N} + 1\}$ **do**

    ▷ Initialize estimated best and worst sets.

    $\hat{B}_k \leftarrow \emptyset, \quad \hat{W}_k \leftarrow \emptyset$

    ▷ Calculate thresholds for defining estimated best and worst sets.

    $u_k \leftarrow (n - k + 1)/2 + 1, \quad l_k \leftarrow n - u_k + 1$

    ▷ Perform FWER$_k$ procedure on pairwise comparisons.

    $\left[\hat{l}, \hat{u}\right] \leftarrow$ rankFWER$(\hat{y}, \hat{\sigma}_{\hat{y}}, \alpha, k)$

    ▷ Assign observations to estimated best and worst sets.

    for $i \in 1, \ldots, n$ **do**

      if $\hat{u}_i < u_k$ **then**

        $\hat{B}_k \leftarrow \hat{B}_k \cup \{i\}$

      else if $\hat{l}_i > l_k$ **then**

        $\hat{W}_k \leftarrow \hat{W}_k \cup \{i\}$

  ▷ Find k that produced the largest best and worst sets.

  $k^* = \arg\max_k \left|\hat{W}_k \cup \hat{B}_k\right|$

  $\hat{B}, \hat{W} \leftarrow \hat{B}_{k^*}, \hat{W}_{k^*}$

Hierarchical Bayes

Xie, Singh, and Zhang (2009) show that nonparametric bootstrap approaches for rank inference, such as those presented by Goldstein and Spiegelhalter (1996), perform poorly in the presence of near ties, $y_i \approx y_j$. Xie, Singh, and Zhang (2009) impart a greater degree of uncertainty, however, of how well Bayesian approaches perform in these scenarios. My approach is much more in line with the Bayesian rank inference procedures of Lin et al. (2006). Lin et al. (2006) focus on alternative loss functions for ranking, including a loss function for optimizing classifications of populations into bottom and top percentiles at a fixed threshold (e.g. classifying into the top 10% and bottom 10%). The Bayesian approach presented in this paper differs in two ways. First, the approach to top-bottom classification focuses on constraining the probability of an ordinal misclassification instead of minimizing a loss function corresponding to classification error. Second, rather than fixing thresholds for what constitutes the top and bottom populations, the Bayesian method presented here focuses instead on maximizing the number of observations that can be said to be in the top or bottom while constraining the probability of a misclassification error. There are many reasons why we would prefer a Bayesian approach (Gelman, Hill, and Yajima 2012). Most of all, we are not concerned with Type I errors. We are much more concerned with Type III errors. In fact, we do not seriously believe that $y_i = y_j$ is a reasonable null hypothesis—we are confident that the underlying characteristics of neighborhoods are truly different, we are just concerned with the ordering of the neighborhoods according to those characteristics. Moreover, Type III errors are much less frequent than Type I errors (Gelman, Hill, and Yajima 2012). When using the frequentist error rates, we waste a lot of power when attempting to control Type III errors by simultaneously controlling Type I and Type III errors.

We can model mean outcomes in neighborhood i as

$$y_i \sim \mathcal{N}(\theta_i, \sigma_i^2) \qquad \theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \qquad p\left(\mu_\theta, \sigma_\theta^{-1}\right) \propto 1$$

where σ_i^2 is assumed known from the data. Non-informative prior distributions are assigned to the hyperparameters μ_θ and σ_θ as suggested by Gelman (2006). The posterior is sampled from using Markov chain Monte Carlo algorithms. In this paper, I sample from the posterior using the `rstan` interface to the Stan programming language (Carpenter et al. 2017; Stan Development Team 2018a; Stan Development Team 2018b). Using the simulated posterior of the fitted hierarchical model, we are able to choose $\hat{W}$ and $\hat{B}$ in order to *directly* constrain the posterior probability that the maximum true value in $\hat{W}$ is greater than the minimum true value in $\hat{W}$. However, the definition of the partial order between $\hat{W}$ and $\hat{B}$ must be adapted to accommodate the Bayesian approach. I propose the following partial order definition:

$$\hat{B} \succ_\alpha \hat{W} \qquad \text{if} \qquad \tau_{ij} = 1 \text{ for all } i \in \hat{W},\ j \in \hat{B} \quad p\left(M_{\hat{W}} > m_{\hat{B}} \mid \hat{y}, \hat{\sigma}_y^2\right) < \alpha$$

where $M_{\hat{W}} = \max_{i \in \hat{W}} y_i$ and $m_{\hat{B}} = \min_{j \in \hat{B}} y_j$. Going back to the planner's problem, if $\alpha = 0.05$, then $\hat{B} \succ_\alpha \hat{W}$ says that the planner is 95% confident that any neighborhood in $\hat{B}$ is more advantaged than any neighborhood in $\hat{W}$.

Without initial restrictions, this discrete optimization problem remains intractable. However, we can constrain the objective space using a three step algorithm. The algorithm maxBAYES is given in Algorithm 2. The first step is to simulate the posterior distribution of true values. The second step is a rough pass of defining $\hat{W}$ and $\hat{B}$ as observations with Bayesian credible sets with lower bounds greater than $N/2$ and upper bounds less than $N/2$, respectively. The marginal credible sets are constructed similarly to the confidence sets in the frequentist sense, but instead of standard test-statistics, we say that y_i is significantly greater than y_j if $y_i > y_j$ in at least $100(1-\alpha)\%$ of simulations.

$$\tau_{ij} = p(y_i > y_j \mid \hat{y}, \hat{\sigma}_y^2) > 1 - \alpha$$

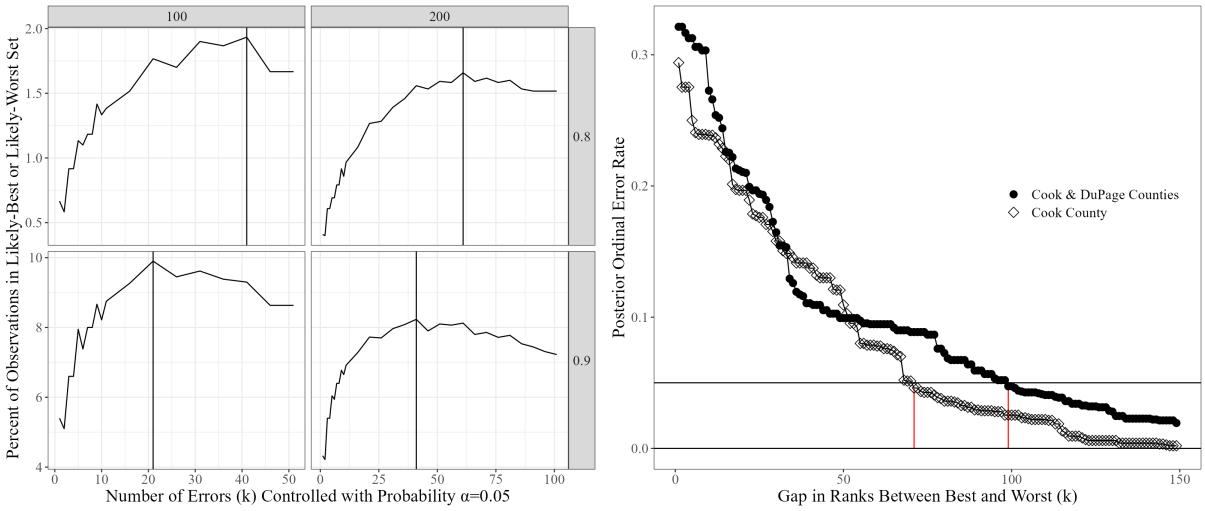
and then the upper and lower bounds of the Bayesian rank credible sets are constructed based on values of τ_{ij} as described in the Setup and Notation section. The third step repeatedly expands the thresholds

for defining $\hat{W}$ and $\hat{B}$ away from the center until $\hat{B} \succ_\alpha \hat{W}$ is achieved. In other words, $\hat{W}$ and $\hat{B}$ are determined by selecting the largest k as in the maxFWER algorithm such that defining $\hat{W}$ and $\hat{B}$ with a gap of k ranks ensures that $\hat{B} \succ_\alpha \hat{W}$, but a gap of $k+1$ ranks does not ensure $\hat{B} \succ_\alpha \hat{W}$. The third step of the algorithm is visualized in Figure 1ii using tract-level data on economic mobility in Cook and DuPage Counties in Illinois. At $k = 1$, the posterior probability of an ordinal error rate is around 30%. But as k increases, the posterior probability falls below $\alpha = 0.05$ at $k \approx 70$ when just focusing on tracts within Cook County, and $k \approx 100$ when also including tracts within DuPage County. The red lines indicate the values of k used to construct $\hat{W}$ and $\hat{B}$ as selected by the maxBAYES algorithm.

In terms of computational efficiency, the maxBAYES algorithm tends to be somewhat faster and more memory-efficient compared to the maxFWER algorithm, because although it may take a substantial amount of time to simulate the posterior distribution, once the posterior has been simulated the optimization steps are quick. However, the maxFWER algorithm requires repeated computation over a grid of k values in order to find the optimal k.

Performance in Simulations

Figure 1i and 2i both show the simulated average percent of observations categorized among the likely best or the likely worst for the provided method. Without measurement error, the maximum attainable value would be 100%, where each observation can be categorized either among the best or the worst. The simulation was conducted as follows. First, simulate $N \in \{10, 20, \ldots, 200\}$ estimates and standard errors using a normal means model where the theoretical ratio of signal to total variance is either 0.8 or 0.9. With estimates and standard errors constructed, use the FDR, FWER_1, maxFWER, or maxBAYES approaches to compute the proportion of groups within the likely best or likely-worst sets at confidence level $\alpha = 0.05$.



(i) maxFWER, Searching for Optimal k

(ii) maxBAYES at $\alpha < 0.05$

Figure 1: maxFWER and maxBayes Algorithms

Figure 2i displays the performance of the three methods using the simulated draws where the reliability ratio is equal to 0.9. For the FDR and FWER_1 methods, the likely-best set is defined using all groups with simultaneous confidence sets contained below $N/2$. The Bayesian approach stands out for two reasons. First, the Bayesian approach displays an impressive increase in power over the next best method, and a nearly ten-fold increase in power over the FWER_1 method of Mogstad et al. (2020). Second, the Bayesian approach best adapts to an increasing number of observations being ranked as the percent of observations

Algorithm 2: maxBAYES

Input: $\hat{y}$ and $\hat{\sigma}_{\hat{y}}$, vectors of length n, and $\alpha \in (0, 1)$.

Output: Estimated best and worst sets: $\hat{B}, \hat{W}$

begin

▷ Step 1: Fit hierarchical model to simulate posterior joint distribution

$p(\theta \mid \hat{y}, \hat{\sigma}_y^2) \leftarrow \frac{p(\hat{y}, \hat{\sigma}_y^2 | \theta) p(\theta)}{p(\hat{y}, \hat{\sigma}_y^2)}$

▷ Step 2: Initialize marginal confidence sets

for $(i, j) \in \mathcal{I},\ i \neq j$ **do**

$\hat{t}_{ij} \leftarrow p(y_i > y_j \mid \hat{y}, \hat{\sigma}_y^2)$

$\tau_{ij} \leftarrow \mathbb{1}\{\hat{t}_{ij} > 1 - \alpha\}$

▷ Construct $\left[\hat{l}, \hat{u}\right]$ based on τ_{ij}.

▷ Step 3: Find largest $\hat{W}$ and $\hat{B}$ such that $\epsilon < \alpha$.

$\epsilon \leftarrow 1,\ k \leftarrow -1$

while $\epsilon \geq \alpha$ **do**

$k \leftarrow k + 2$

▷ Initialize estimated best and worst sets.

$\hat{B}_k \leftarrow \emptyset, \qquad \hat{W}_k \leftarrow \emptyset$

▷ Calculate thresholds for defining estimated best and worst sets.

$u_k \leftarrow (n - k + 1)/2 + 1, \qquad l_k \leftarrow n - u_k + 1$

▷ Assign observations to estimated best and worst sets.

for $i \in 1, \ldots, n$ **do**

if $\hat{u}_i < u_k$ **then**

$\hat{B} \leftarrow \hat{B} \cup \{i\}$

else if $\hat{l}_i > l_k$ **then**

$\hat{W} \leftarrow \hat{W} \cup \{i\}$

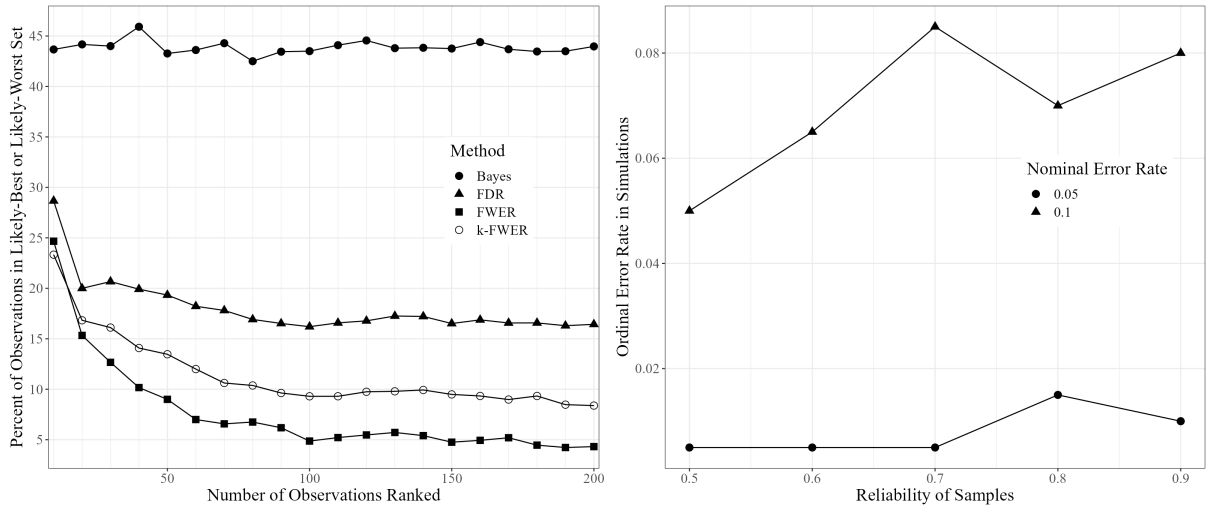
▷ Calculate posterior error, where the maximum of $\hat{W}$ is larger than the minimum of $\hat{B}$.

$M_{\hat{W}} \leftarrow \max_{i \in \hat{W}} y_i, \qquad m_{\hat{B}} \leftarrow \min_{j \in \hat{B}} y_j$

$\epsilon \leftarrow p\left(M_{\hat{W}} > m_{\hat{B}} | \hat{y}, \hat{\sigma}_y^2\right)$

classified as the best or worst remains relatively stable at 44%. The other methods, including the FDR approach, display decreases in power as the number of observations being ranked increases.

Figure 2ii shows that the maxBAYES algorithm displays desirable frequentist coverage properties. In particular, the y-axis displays the true ordinal error rate of the maximum of $\hat{W}$ being greater than the minimum of $\hat{B}$ over 200 samples. The x-axis shows the reliability of the samples, with greater reliability indicating less noisy data. The two trends represent a nominal coverage rate of $\alpha = 0.05$ and $\alpha = 0.10$. The maxBAYES procedure succeeds in covering the errors. In fact, the maxBAYES procedure appears to over-cover, and produce error rates lower than the nominal error rate α. Although the Bayesian approach is somewhat conservative, the other methods over-cover so substantially that no errors were observed in 1000s of simulations.



(i) % of Obs. in Likely-Best or Likely-Worst Sets

(ii) maxBAYES Frequentist Error Rate

Figure 2: Performance of Error Controlling Methods in Simulations

Examples

Upward Mobility Across Counties in the United States

This section takes the error-controlling methods of the previous section to practice using mobility estimates from The Opportunity Atlas (Chetty, Friedman, et al. 2018). Mobility estimates are displayed in Figure 3.[5] Figure 4i replicates the results of Mogstad et al. (2020) which shows the likely-best and likely-worst sets of counties controlling the FWER at $\alpha = 0.05$. Figure 4ii extends the results by instead controlling the generalized familywise error rate with k selected using the maxFWER algorithm. Counties that are are not among the best or worst sets using the FWER_1 approach are shaded darker with black outlines. Although the maxFWER approach was shown to remain conservative in the previous section, there are many notable additions relative to the FWER_1 approach. Additions to the likely-worst include the counties containing Cleveland and Toledo in Ohio, Dallas and Waco in Texas, Atlantic City in New Jersey, and Peoria in Illinois. Two counties contained within the Standing Rock Indian reservation in North and South Dakota are also included in the likely-worst set for economic mobility, notable exceptions in an otherwise highly upwardly mobile Upper Midwest. Additions to the likely-best places for economic mobility include

[5]The estimates used in this paper are the predicted mean household income ranks in adulthood for children raised at the 25th percentile of income (pooled across race).

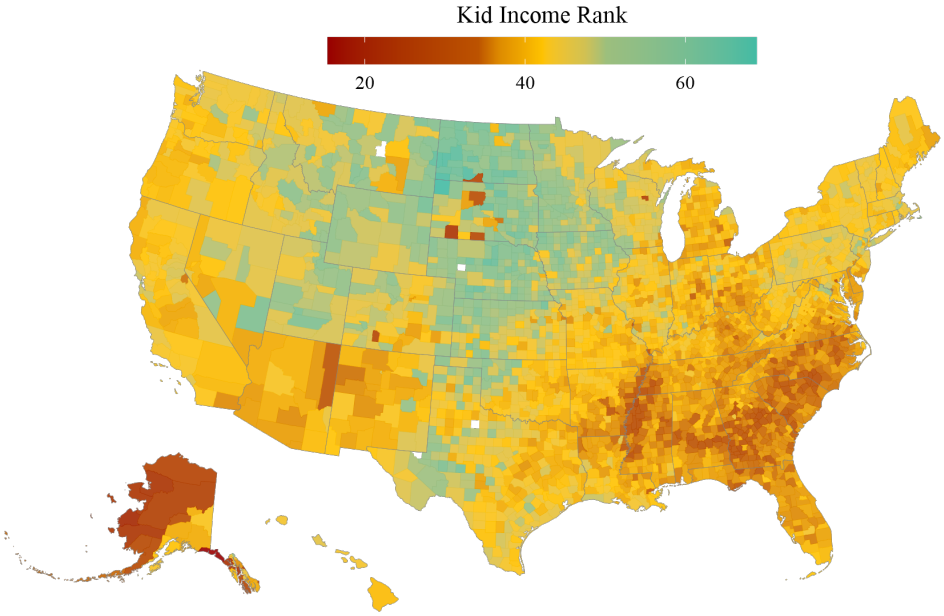


Figure 3: Household Income Rank in Adulthood, All US Counties

places like Arlington, Virginia and Orange County, California. The FDR and Bayesian approach in Figures 4iii and 4iv show substantial increases in the proportion of counties within the likely-best and likely-worst sets.

Upward Mobility Within Cook and DuPage Counties, IL

This section focuses on upward mobility rankings at the level of the census tract within Cook and DuPage Counties, Illinois. Figure 5 displays upward mobility estimates for both counties with the borders of Chicago's community areas overlaid. The City of Chicago clearly has lower upward mobility rates relative to the immediate suburbs. Indeed, Chetty and Hendren (2018) find that DuPage County to the west of Chicago is forecasted to generate the highest incomes for children growing up in low income households out of the 100 most populous counties. Cook County has one of the lowest rates of upward mobility. Chetty and Hendren (2018) estimate that moving from Cook County to DuPage County at birth would increase a child's income by about 30% on average.

In the following figures, the community area of Englewood is outlined in black, a Chicago neighborhood historically known for high rates of violent crime, low high school graduation rates, and substantial depopulation (Morenoff and Sampson 1997). However, taking the FWER_1 approach in Figure 6i, we are unable to say whether Englewood is among the best or among the worst in terms of upward mobility rates when considering rankings within Cook County alone. When considering Cook and DuPage Counties together in Figure 7i, some of the tracts within Englewood are among the worst in terms of upward mobility. The maxFWER approach in Figure 6ii shows that Englewood and surrounding neighborhoods can confidently be stated as having among the lowest upward mobility rates in Cook County. The FDR and Bayesian approaches in Figures 6iii and 6iv display great confidence in our ability to state that the west and south sides of Chicago exhibit low upward mobility relative to the suburbs. In other words, the Bayesian approach allows us to be 95% confident that the best tract in the west or south side of Chicago will still be worse than the worst tract in the suburbs in terms of upward mobility.

Despite the fact that DuPage has much higher upward mobility estimates relative to Cook County, not a single neighborhood in DuPage county can be said to be among the best 95% confidence when using the

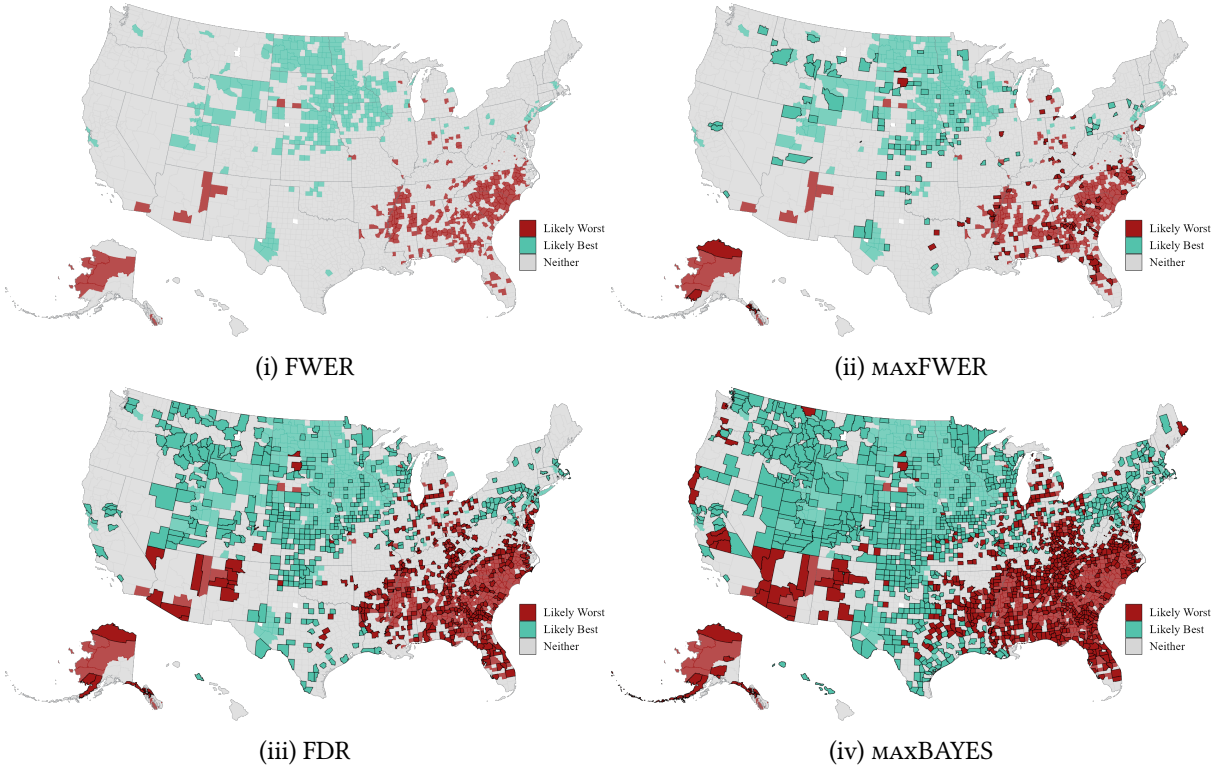


Figure 4: Likely Best and Worst Neighborhoods in terms of Upward Mobility Across US Counties

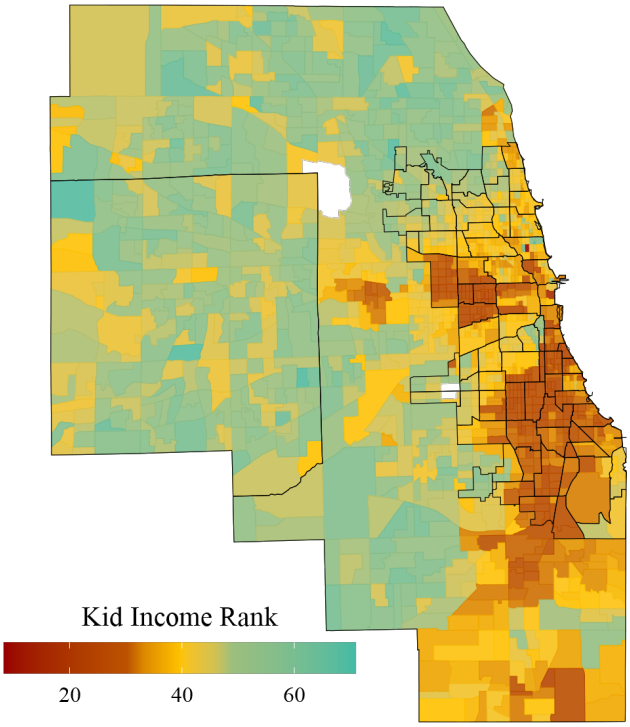


Figure 5: Household Income Rank in Adulthood, Cook and DuPage Counties, IL

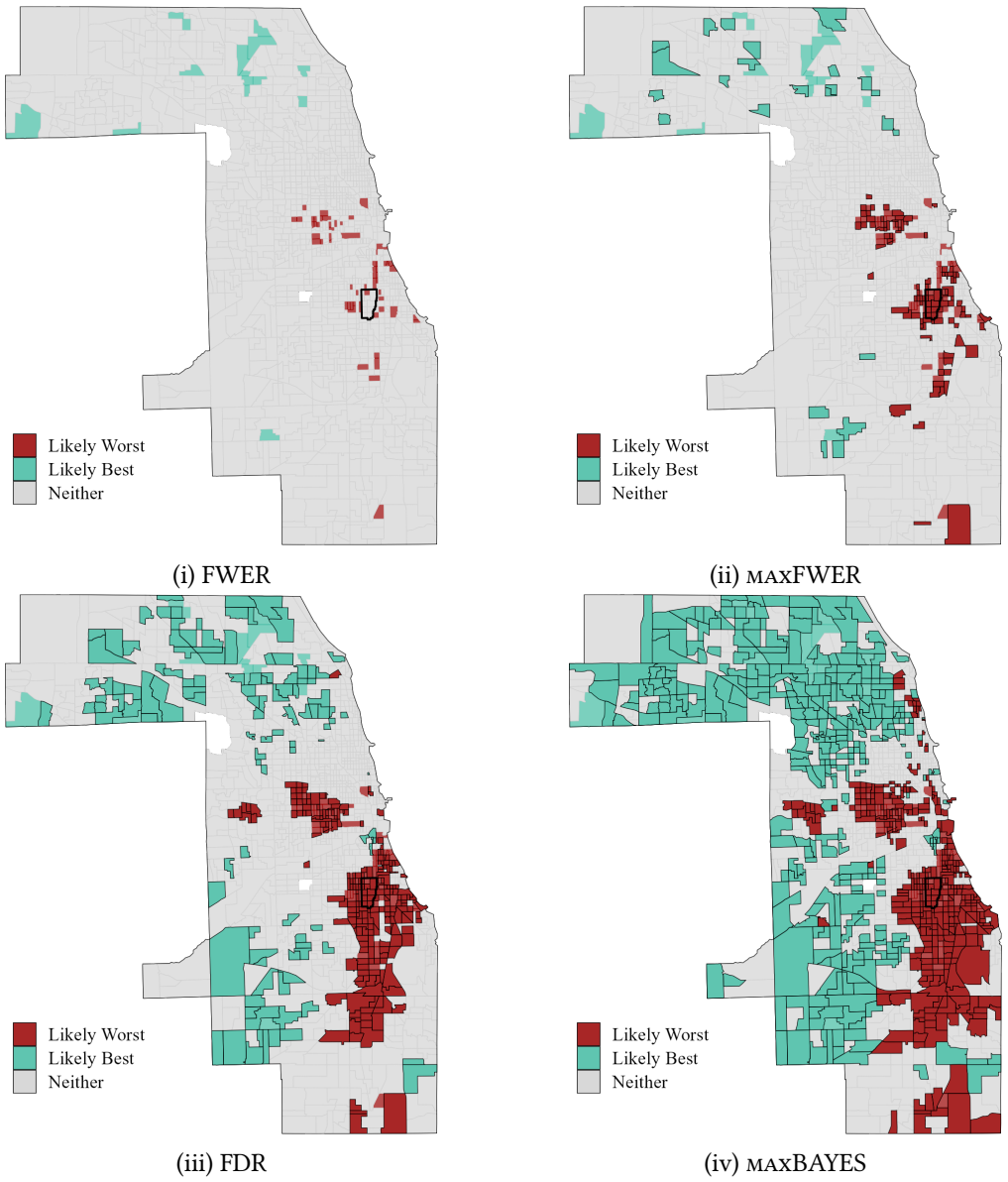


Figure 6: Likely Best and Worst Neighborhoods in terms of Upward Mobility in Cook County, IL

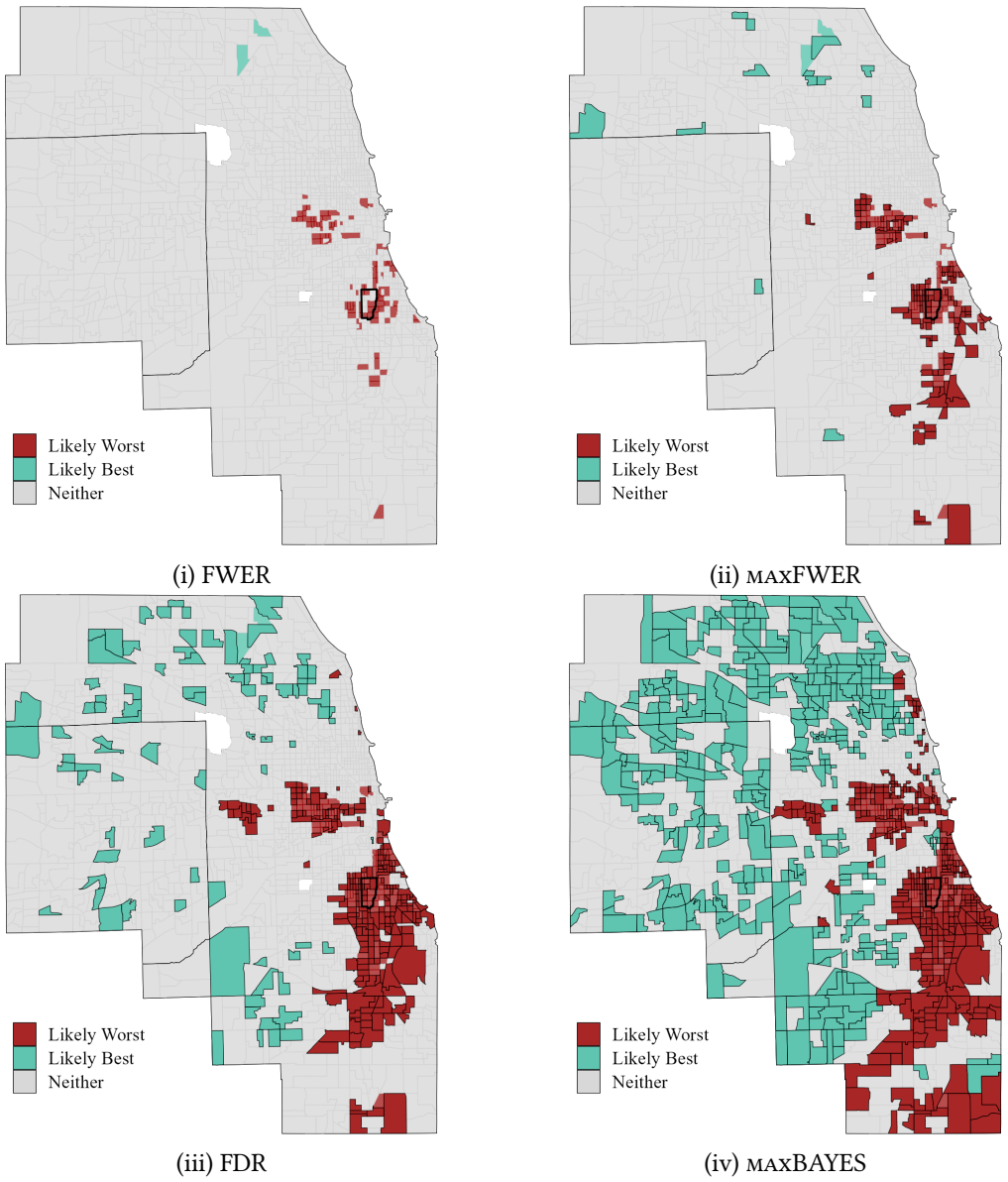


Figure 7: Likely Best and Worst Neighborhoods in terms of Upward Mobility in Cook and DuPage Counties

FWER_1 approach in Figure 7i. Furthermore, when considering rankings of Cook and DuPage counties, *fewer* census tracts in Cook County can be considered among the best in terms of upward mobility. That is, the FWER_1 approach does not adapt to the number of observations being ranked, otherwise we would not expect a decrease in the number of observations estimated to be among the best.

The FWER_k approach improves inference without changing the interpretation of the constraint on a directional error between likely-best and likely-worst sets. Census tracts that were not categorized among the best or worst using the FWER_1 method but are within either category using the FWER_k method are shaded and outlined darker. First of all, many more census tracts are among the best and worst when using the FWER_k approach with $k = 81$ selected using the maxFWER algorithm. Second of all, the reduction in the number of census tracts that are among the best when you include DuPage county is less severe than when using the FWER_1 method. So the FWER_k approach better adapts to an increasing sample size when k is appropriately adjusted, but still suffers from a decrease in relative power as the number of observations grows.

The Bayesian approach considerably improves our ability to infer which neighborhoods are among the best and worst in terms of upward mobility. Heavy spatial autocorrelation shows how swathes of south and west Chicago are the lowest in terms of upward mobility, and outlying suburbs are among the best. Notably, spatial proximity has not even been considered when simulating the posterior distribution. Including DuPage County when ranking tracts does *not* substantially reduce the number of tracts we are able to say are among the top. In this regard, the Bayesian approach adapts to the number of observations being ranked and the reliability of the signal in the estimates, just as the simulation results suggested.

This section conveys two main points. First, the FWER_1 approach allows for only a very limited interpretation of which neighborhoods are best and which neighborhoods are worst. But the FWER_k and Bayesian approaches show that we are able to expand which neighborhoods are significantly among the best and worst while continuing to constrain the probability of a relevant directional error. Second, this example shows how each method adapts to the number of observations being ranked. By contrasting rank inference procedures within Cook County to rank inference procedures within Cook County and DuPage Counties together, I show how the Bayesian approach adapts to an increasing number of observations being ranked.

Conclusion

Motivated by a model of residential choice, this paper introduces the problem of classifying places into best and worst sets while constraining the probability of pairwise directional errors between the best and worst sets. I show that controlling the FWER succeeds in controlling the probability of such an error, but greatly limits our ability to infer best from worst. I introduce two alternative classification methods, the maxFWER and maxBAYES algorithms, that improve inference ability. The key finding is that the maxBAYES algorithm substantially increases our ability to infer best from worst, exhibits stable performance as the number of observations being compared increases, and displays attractive frequentist error rates.

This paper suggests several directions for future work. I list three such directions. First, the frequentist approaches provide very coarse upper bounds on the probability of an ordinal error; there may exist alternative frequentist approaches that more directly constrain the probability of an ordinal error. Second, more work is required to understand the finite-sample and asymptotic properties of the various classifiers. In particular, understanding the conditions under which the maxBAYES algorithm provides desirable frequentist coverage properties will better justify its use. Third, I have so far assumed that the level of aggregation is fixed when drawing comparisons. In the examples of this paper, when inferring geographic patterns of economic mobility, the level of geography is fixed at the county or Census tract level. However, policy can be targeted at the state, block, or even the housing unit level. Moreover, the

level of geography can vary across place in order to adapt to varying levels of local heterogeneity. Developing clustering approaches that abide by policy-relevant constraints will extend the methods of this paper, allowing the data to select the optimal level of geographic aggregation.

References

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey (Jan. 2019). "Inference on Winners". In: 25456. Series: Working Paper Series.

Atkinson, Anthony B (Sept. 1970). "On the measurement of inequality". en. In: *Journal of Economic Theory* 2.3, pp. 244–263.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1. Publisher: [Royal Statistical Society, Wiley], pp. 289–300.

Benjamini, Yoav and Daniel Yekutieli (Aug. 2001). "The control of the false discovery rate in multiple testing under dependency". In: *The Annals of Statistics* 29.4. Publisher: Institute of Mathematical Statistics, pp. 1165–1188.

Bergman, Peter et al. (Aug. 2019). "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice". en. In: w26164, w26164.

Carpenter, Bob et al. (2017). "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1.

Chetty, Raj, John N. Friedman, et al. (Oct. 2018). "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility". In: 25147. Series: Working Paper Series. (Visited on 11/07/2021).

Chetty, Raj and Nathaniel Hendren (Aug. 2018). "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*". en. In: *The Quarterly Journal of Economics* 133.3, pp. 1163–1228.

Efron, Bradley (2008). "Microarrays, Empirical Bayes and the Two-Groups Model". In: *Statistical Science* 23.1. Publisher: Institute of Mathematical Statistics, pp. 1–22.

Efron, Bradley and Robert Tibshirani (June 2002). "Empirical bayes methods and false discovery rates for microarrays". eng. In: *Genetic Epidemiology* 23.1, pp. 70–86.

Efron, Bradley, Robert Tibshirani, et al. (2001). "Empirical Bayes Analysis of a Microarray Experiment". en. In: p. 20.

Gaubert, Cecile, Patrick M. Kline, and Danny Yagan (Jan. 2021). "Place-Based Redistribution". In: 28337. Series: Working Paper Series.

Gelman, Andrew (Sept. 2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". In: *Bayesian Analysis* 1.3. Publisher: International Society for Bayesian Analysis, pp. 515–534.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima (2012). "Why We (Usually) Don't Have to Worry About Multiple Comparisons". In: *Journal of Research on Educational Effectiveness* 5.2, pp. 189–211.

Gelman, Andrew and Francis Tuerlinckx (Sept. 2000). "Type S error rates for classical and Bayesian single and multiple comparison procedures". en. In: *Computational Statistics* 15.3, pp. 373–390.

Goldstein, Harvey and David J. Spiegelhalter (1996). "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159.3. Publisher: [Wiley, Royal Statistical Society], pp. 385–443.

Lin, Rongheng et al. (2006). "Loss function based ranking in two-stage, hierarchical models". English (US). In: *Bayesian Analysis* 1.4. Publisher: Carnegie Mellon University, pp. 915–946.

Mogstad, Magne et al. (Mar. 2020). "Inference for Ranks with Applications to Mobility across Neighborhoods and Academic Achievement across Countries". en. In: ID 3557282.

Morenoff, Jeffrey D. and Robert J. Sampson (1997). "Violent Crime and the Spatial Dynamics of Neighborhood Transition: Chicago, 1970-1990". In: *Social Forces* 76.1. Publisher: Oxford University Press, pp. 31–64.

Romano, Joseph P. and Michael Wolf (2005). "Stepwise Multiple Testing as Formalized Data Snooping". en. In: *Econometrica* 73.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2005.00615.x, pp. 1237–1282.

Romano, Joseph P. and Michael Wolf (Oct. 2007). "Control of generalized error rates in multiple testing". In: *arXiv:0710.2258 [math, stat]*. arXiv: 0710.2258.

Stan Development Team (2018a). *RStan: the R interface to Stan*. R package version 2.17.3. URL: http://mc-stan.org/%205.

— (2018b). *The Stan Core Library*. Version 2.18.0. URL: http://mc-stan.org/%205.

Tang, Weihua and Cun-Hui Zhang (2007). "Empirical Bayes methods for controlling the false discovery rate with dependent data". In: *arXiv:0708.0978 [stat]*. arXiv: 0708.0978, pp. 151–160. (Visited on 11/12/2021).

Xie, Minge, Kesar Singh, and Cun-Hui Zhang (2009). "Confidence Intervals for Population Ranks in the Presence of Ties and Near Ties". In: *Journal of the American Statistical Association* 104.486. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 775–787.