Data Aggregation and Resource Allocation

Zach Siegel
advised by Auyon Siddiq
UCLA Anderson School of Management

1 Introduction

Individual-level features are a natural consideration in centralized resource allocation: for example, public transit stops should optimally be placed near those without cars, and vaccine outreach should be directed towards those who are unvaccinated. While individual addresses are publicly available in much of the US, individual car-ownership and vaccination status are generally private. What is publicly known about these and other private individual data are *aggregate measures*. The US census reports on car ownership in each tract, and public health organizations commonly report on zip-code-wide vaccination levels. This model of uncertainty in individual-level features is ubiquitous, and it is worth devoting special attention to optimization in its setting.

Not only do individual-level features improve resource allocation, but failure to consider them may result in adverse outcomes for groups of individuals that become crystal clear when individual-level data is separately obtained and analyzed. In fact, *inequality in outcome* between members of different groups - even when unavoidable, or when group membership is unobservable to a decisionmaker - is a widely recognized form of *algorithmic bias*. It may be necessary to *include* sensitive features in a model to mitigate unequal outcomes; if features are unknown, it may be necessary to include them as uncertain parameters.

By leveraging known individual-level group membership data, via *ecological inference*, we may be able to gain insight into the exceedingly useful individual-level private features (such as car ownership or vaccination status). In short, if there are many aggregate units (e.g. many zip codes), and zip codes with more individuals from one group tend to have higher rates of a private feature (e.g. car ownership), then ecological inference frameworks estimate the fraction of individuals from each group with that private feature, within each aggregate unit (e.g. the number of white and non-white car owners). These estimates can greatly refine the existing aggregate knowledge.

Privacy norms lead to uncetainty due to aggregation, and while it may be incumbent upon decision-makers to address this uncertainty to improve outcomes or avoid algorithmic bias, problems that affect large populations may be intractable at the individual level of analysis. We propose an approach that naturally decomposes the population into smaller aggregate units, making an

inroad to the tradeoff between data pooling and tractability by utilizing shared information over a population to refine individual-level parameters. Furthermore, we identify a setting in which addressing this type of uncertainty does not substantially increase the computational complexity of a resource allocation problem. We also explain how ecological inference reduces the uncertainty space in a broader family of optimization problems, full treatment of which we leave to further work.

In short, this paper aims to describe the value in resource allocation optimization of refining a statistic like "number of car owners" to "numbers of Black car owners and non-Black car owners", both for the sake of improving outcomes overall and for avoiding unequal outcomes for racial groups.

2 Literature Review

2.1 Ecological Inference

To ensure privacy or simply to summarize, datasets concerning individuals' features are almost always made publicly available only in aggregate form. For example, the US Census reports averages and total counts for individuals in a given census tract, census block group, or census block. An individual's name, address, age, racial group, and voting history (whether they voted, not how they voted) are often available from US states in the form of "voter files". So, for example, public data in the US keeps private individual car ownership and income, but reports in aggregate the number of car owners and the average income in the neighborhood around each individual.

"Ecological inference" attempts to use multiple statistics that are reported in aggregate to infer individual-level correlations to answer questions like "how many Black individuals voted for Democrats?" or "how many Black car-owners are out there?" In this literature, an "ecological correlation" is a correlation (in the usual sense) between the aggregate statistics sampled over many aggregations.

For example, in the classic example, election precincts with high Democratic voting rates may have high numbers of Black individuals: a positive correlation may exist between the aggregate statistics "fraction of Black individuals" and "fraction voted Democratic", both of which are publicly available. That positive correlation, however, does not necessarily imply that Black individuals are more likely to vote Democratic, an "individual-level correlation" between Blackness and Democratic preference. Supopse there are N precincts, and precinct i has known percentage of democratic voters T_i (democratic Turnout), as well as known fraction of Black voters X_i (eXplanatory variable). Ecological attempts to predict the "individual-level correlations": the fraction of Black and non-Black individuals who voted Democratic, represented by β_i^B and β_i^w, respectively. In other words, the inner cells in Table 1 are inferred from many samples of the marginal cells:

		Voted Democratic		
		Yes	No	Tot:
Race	Black	β_i^B	$1-\beta_i^B$	X_i
	Non-Black	β_i^w	$1-\beta_i^w$	$1-X_i$
	Tot:	T_i	$1-T_i$	1

Figure 1: *A two-way (or "four-fold") table describing the relationship between known aggregate-level statistics and unknown individual-level statistics/parameters. X_i represents the (known) fraction of voting age people in district i who are Black, T_i represents the (known) fraction of people in district i who voted Democratic, and β_i^B and β_i^w represent the (unknown) fraction of Black and white people who vote Democratic, respectively.*

If an ecological correlation exists in the absence of an individual-level one, or vice-versa, it is known as "aggregation bias", which implies that the groupings defined by the aggregation process (e.g. grouping by geographic proximity) are more informative of individual-level correlation than the aggregate features are. Ecological inference typically begins with an assumption regarding a limit to (or entirely the absence of) aggregation bias. Ecological inference has had successes in different domains, mostly related to social science and election analysis; there are several existing inference frameworks, each of which makes different assumptions about aggregation bias, and each of which has motivating advantages and disadvantages within different domains.

The concept of ecological inference was introduced in the 1950s [29], extended slightly in the 1970s [31, 18, 9], and then not largely ignored until the work of Gary King beginning in 1997 [21, 23], which combined existing regression techniques with a treatment of the fact that individual-level correlations can be deterministically bounded for each aggregate sample. Applications of these methods appear in recent work in various domains, such as [25].

We leave it to other work (such as [20]) to describe the limitations of ecological inference. Beyond the deterministic bounds the method generates, it is impossible to know how reliably accurate ecological inferences are for data in any domain. The only datasets for which the efficacy of ecological inference can be evaluated are those for which the individual-level correlations are already known (and thus inference is unnecessary). Still, like other popular statistical methods, successesful application in various domains ultimately gauges the usefulness of these methods.

2.2 Differential Privacy and Prevalence of Aggregate Data

It is not a historical coincidence that so much useful social science data is presented in the aggregated form described in the previous section. The field of *differential privacy* studies the mathematical requirements for presenting data regarding individuals in such a way that information about any one individual cannot be precisely inferred. The essential toolkit to differential privacy is *aggregation* along with *introduction of noise* [13, 17, 15]. Without either of these components, a tool such as ecological inference could discover correlations that reveal more about the population in the aggregate data than the individuals involved agreed to release (though even in a setting that is insufficiently differentially private may permit such revelation [14]).

Anonymized datasets that are not differentially private have been particularly influential both as benchmarks in the advancement of classification and machine learning algorithms, and as a

raw ingredient in online advertising. Recently, however, the insufficiency of anonymization in "private" datasets has been well-documented [2, 14, 26]. An example of a shift in attitudes towards anonymization is the phasing out of anonymized third-party cross-site "cookies" in web browsing; a recent effort to replace these with "federated learning of cohorts" essentially gives advertisers access not to individual-level (anonymized) browsing histories, but to the aggregate browsing histories of cohorts of several hundred or thousand "similar" users.

Whether a dataset guarantees a level of differential privacy or not, aggregation is the nearly ubiquitous resolution to the problem of sharing data without conceding privacy. It is possible that as privacy norms shift, so too will the classification algorithms that are the lifeblood of the internet ecosystem, and a proliferation of aggregate datasets from sources other than the US Census will fuel the social sciences for years to come, making ecological inferences increasingly necessary and appealing.

2.3 Fairness

It has been recently well-documented [12, 10] that excluding protected features of individuals from a model ("anti-classification" in [10]) does not preclude discrimination via that model. The consensus alternative in the literature can be thought of as "fairness through awareness," as in the eponymous [12], which is to treat group membership explicitly to achieve an objective fairness goal.

Most efforts towards algorithmic fairness are in the context of classification algorithms, which can be thought of as basic resource allocation problems in which individuals with uncertain parameters are the unit of analysis (and their label assignments are the "resources"). Even in this most basic resource-allocation setting, consideration of uncertain individual-level to avoid algorithmic bias often yields difficult numerical optimization problems [39, 40, 16, 19, 24, 37, 38]. Little work has been done to incorporate fairness into the ubiquitous resource allocation optimization problems in Operations Research (though [8] is a good start, and [34] does so in the limited context of fairness as "equality" without group membership), and indeed the computational expense of incorporating fairness concerns into even a basic classification problem may explain this. There are mature methods, on the other hand, to evaluate whether an outcome is fair after the event, as in [5] [11].

While consideration of uncertainty in individual-level features may broadly enable *fairness-inducing* optimization strategies, simply treating these uncertain features accurately (to the extent possible) may naturally reduce algorithmic bias, as will be demonstrated in Section 3.7, by improving outcomes for members of a group with systematically higher costs, and therefore reducing the between-group disparity in outcomes.

2.4 Optimization Under Uncertainty

Many central resource planning problems are unavoidably NP-hard, suggesting that *the individual* (rather than a coarser aggregation) is an ill-advised unit of analysis. Furthermore, optimization problems are generally substantially complicated by considering uncertainty in model parameters. In summary, it is challenging to include uncertain (private), individual-level features in an opti-

mization model.

On the other hand, Operations Research has seen a proliferation of treatments of uncertainty in optimization in the last two decades. Often robust convex optimization problems have efficient reformulations [6, 35, 4, **?**], stochastic optimization permits inexpensive yet accurate sample average approximations [36], and even distributional robustness has seen tractable and effective treatments [28, 7].

The running example in this work is a robust facility location problem. Numerous treatments of facility location under uncertainty exist [1, 30, 32, 3, 27], and we hope that our future work will connect extant sensitivity analyses of these problems with the type of uncertainty we attempt to understand in this paper.

3 Optimization Model

The "individual" is our unit of analysis, and individuals have uncertain type (an *unknown* discrete feature), and the problem parameters are functions of this type (and are therefore themselves uncertain). While individual types are not known, it is known what number of individuals have each type; furthermore, individuals may be partitioned into subsets (by a *known* discrete feature) in each of which the number of individuals of each type are known. We explore in depth the value of a single partition (known binary feature) of the individuals, and how to utilize this reduction of uncertainty.

We describe a broad optimization setting in which this framework may be desirable, which we then narrow to a specific problem type that we treat in depth and later associate with a real-world problem. We discuss the advantages gained by a partition of the individuals both in this specific problem and the broader setting.

3.1 Model of Uncertainty Set Ω

We are motivated by the idea that individuals with unknown "type" experience costs that are functions of their type, where we know, in aggregate, how many individuals there are of each type, but don't know which individuals are which. We can think of individuals belonging to a "high-cost type" as having a "disturbance" added to some low-cost "nominal" parameter. We return to the model of individuals later in Section 3.4, and for now discuss the more general (and less cumbersome) setting in which nominal cost vectors are added to a known number of "disturbance" cost vectors.

Suppose c is an uncertain parameter from the uncertainty set $\{c_\omega : \omega \in \Omega\}$. To specify a resolution of c_ω, we introduce the binary vector w and the possible disturbances c_i, $i = 1, \ldots, n_c$ to the nominal value c_0. Similar to the framework introduced in [6], we consider an integer "budget"

of uncertainty Γ_c. The uncertainty set consists of all the folowing:

$$c_\omega \in \left\{ c : \begin{array}{ll} c & = c_0 + \sum_{i=1}^{n_c} w_i c_i \\ \sum_{i=1}^{n_c} w_i & \leq \Gamma_c \\ w_i & \in \{0,1\} \end{array} \right\} \tag{1}$$

In other words, c_ω is subjected to at most Γ_c disturbances from its nominal value c_0, and d_ω is subjected to at most Γ_d

We also introduce another uncertain vector, $d \in \{d_\omega : \omega \in \Omega\}$, which appears in the model in the next section, but whose uncertainty, we will demonstrate in Section 3.3, can be omitted from the model without loss of generality:

$$d_\omega \in \left\{ d : \begin{array}{ll} d & = d_0 + \sum_{i=1}^{n_d} z_i d_i \\ \sum_{i=1}^{n_d} z_i & \leq \Gamma_d \\ z_i & \in \{0,1\} \end{array} \right\} \tag{2}$$

3.2 Multi-Stage Problem

Our approach is broadly relevant to two-stage optimization problems in which there is uncertainty in the objective:

$$\begin{aligned} & f^{outer}(y) + \mathcal{R}_\Omega[Q_\omega(y)], \\ & Q_\omega(y) = \min_{x \in X} \left\{ f_\omega^{inner}(x,y) \,|g(x,y) \leq b \right\} \end{aligned} \tag{3}$$

where $\omega \in \Omega$ denotes a resolution of the uncertainty (a "scenario") from the set Ω, and $\mathcal{R}$ is an operator that defines the desired response to uncertainty, such as $\mathcal{R}_\Omega[\cdot] = \max_{\omega \in \Omega}[\cdot]$ (robust optimization) or, when Ω is associated with a probability meassure, $\mathcal{R}_\Omega[\cdot] = E_{\omega \in \Omega}[\cdot]$ (stochastic optimization).

We propose an approach for problems distinguished by the nature of their uncertainty set, rather than a specific optimization methodology. A robust formulation has a straightforward solution that aligns with our motivation to focus on the nature of the uncertainty (rather than on a solution algorithm), and to avoid a "worst case" of algorithmic bias. So, for the remainder of this work, we focus our attention on the robust mixed-integer linear program

$$\begin{aligned} & \min_{y \in Y} d_0^T y + \max_{\omega \in \Omega}[Q_\omega(y)] \\ & Q_\omega(y) = \min_{x \in X} \left\{ c_\omega^T x + d_\omega^T y \,|Ax + By \leq b \right\} \end{aligned} \tag{4}$$

where $Y = \{y \geq \mathbf{0} : y_j \in \mathbb{Z} \; \forall \; j \in J\}$ (J consists of indices of integer-constrained components of y), and $X = \{x : x \geq \mathbf{0}\}$. We also assume that c_ω and d_ω have all positive components, for all $\omega \in \Omega$ for simplicity.

The form of problem (3) is extremely general, whereas the problem (4) is linear and has uncertainty only in costs, and is straightforward to analyze. In our case, it reduces to a mixed integer

linear program.

Note that this optimization model distinguishes between the decision variables x and y by the fact that x represents decisions that have recourse after uncertainty has been resolved, whereas y represents decisions that are made before uncertainty has been resolved. This distinction is in general meaningful in optimization settings subsumed by formulation (3), but we shall see that in the robust linear program (4), recourse does not play a role, and so the variable y will eventually come to have a different meaning.

3.3 Solving and Simplifying the Robust Model

We can simplify the description of uncertainty in (1) for our purposes by reformulating the robust optimization problem (4) in a way that makes it clear that x and y - variables differentiated by the fact that decisions represented by x are recourse after uncertainty has been resolved - end up analytically equivalent, as the order of optimization and realization of uncertainty does not turn out to matter.

Combining the problem (4) and the model of uncertainty (1), we end up with a robust optimization problem that reduces to a linear program. The problem can be written explicitly as follows:

$$\begin{array}{lllll}
\min\limits_{y}\max\limits_{w,z}\min\limits_{x} & & c_0^T x + d_0^T y + \sum\limits_{i=1}^{n_c} w_i c_i^T x + \sum\limits_{i=1}^{n_d} z_i d_i^T y & & \\
\text{s.t.} & Ax + By & \leq b & & \\
& \sum\limits_{i=1}^{n_c} w_i & \leq \Gamma_c & & \\
& \sum\limits_{i=1}^{n_d} z_i & \leq \Gamma_d & & \\
& 0 \leq w_i & \leq 1 & & \forall\ i = 1, \ldots, n_c \\
& 0 \leq z_i & \leq 1 & & \forall\ i = 1, \ldots, n_d \\
& x, y & \geq \mathbf{0} & & \\
& y_j & \in \{0,1\}\ \forall\ j \in J, & &
\end{array} \tag{5}$$

where J denotes the set of indices in y that are binary-constrained. This problem reduces to

$$\begin{array}{lllll}
\min\limits_{x,y,\rho_c,\rho_d} & & c_0^T x + d_0^T y + \Gamma_c \rho_d + \Gamma_d \rho_d + \sum\limits_{i=1}^{n_c} (c_i^T x - \rho_c)^+ + \sum\limits_{i=1}^{n_d} (d_i^T y - \rho_d)^+ & & \\
\text{s.t.} & Ax + By & \leq b & & \\
& \rho_c, \rho_d & \geq 0 & & \\
& x, y & \geq \mathbf{0} & & \\
& y_j & \in \{0,1\}\ \forall\ j \in J. & &
\end{array} \tag{6}$$

Note that the coefficients of the variables x and y in equation (6) have the same exact, symmetrical structure, even though x is optimized *after* the uncertainty is resolved and y is optimized *before* the uncertainty is resolved. In other words, optimal decisionmaking *in anticipation of* the resolution of uncertainty and *in response to* that resolution are in this case equivalent. As a consequence, the

following formulation is equivalent to (6), which demonstrates that the uncertainty budgets Γ_c and Γ_d need not make the distinction in this case between variables with recourse over uncertainty and those without:

$$\begin{array}{lll} \min\limits_{y} \max\limits_{w,z} \min\limits_{x} & \dots & \\ \text{s.t.} & \dots & \end{array} = \begin{array}{lll} \min\limits_{x,y} \max\limits_{w,z} & \dots & \\ \text{s.t.} & \dots & \end{array}$$

in which neither decisions represented by y nor those represented by x have recourse over uncertainty. So, in the context of the robust linear program with recourse (4), we need not make a distinction between a variable x with recourse and a variable y without recourse, i.e. we can combine the "inner" and "outer" variables so that there are only "outer" variables

$$y' = \begin{bmatrix} x \\ y \end{bmatrix}, d'_\omega = \begin{bmatrix} c_\omega \\ d_\omega \end{bmatrix}, B' = \begin{bmatrix} A & B \end{bmatrix} Y' = \left\{ y \begin{bmatrix} x \\ y \end{bmatrix} : x \in X, y \in Y \right\}$$

and can equivalently write the "min max min" problem without recourse.

$$\left\{ \begin{array}{ll} \min_{y \in Y} d_0^T y + \max_{\omega \in \Omega} [Q_\omega(y)] & \\ Q_\omega(y) = \min_{x \in X} \left\{ d_\omega^T y + c_\omega^T x \,|Ax + By \leq d \right\} & \end{array} \right\} = \left\{ \begin{array}{ll} \min_{y' \in Y'} & \max_{\omega \in \Omega} d_\omega'^T y' \\ s.t. & B'y' \leq d \end{array} \right\}.$$

Now we make a different distinction between two subsets of decision variables x and y (rather than "with" and "without" recourse, respectively): those whose cost coefficients *do* and *don't* have uncertainty associated with them. We retain the convention that some variables may have integer constraints, and in particular we restrict interest to the case where only variables without uncertain coefficients can have integer constraints. This is motivated by the source of uncertainty being unknown individual-level features in an assignment problem in which y represents global decisions and, for any fixed value of y, x dictates what amounts to a min-cost flow. We again use the symbols x and y, but this time they correspond to variables *with* and *without* uncertain coefficients, respectively. The resulting model, in which there are n possible disturbances to the nominal cost c_0, is the following mixed-integer nonlinear program:

$$\begin{array}{lll} \min\limits_{x \in X, y \in Y} \max\limits_{w} & c_0^T x + d_0^T y + \sum\limits_{i=1}^{n} w_i c_i^T x & \\ \text{s.t.} & Ax + By \leq b & \\ & \sum\limits_{i=1}^{n} w_i \leq \Gamma & \\ & 0 \leq w_i \leq 1 & \forall\ i = 1, \dots, n \\ & x, y \geq \mathbf{0} & \\ & y_j \in \{0, 1\}\ \forall\ j \in J, & \end{array} \tag{7}$$

whose solution can be found by solving the following mixed-integer linear program:

$$\begin{array}{lll} \min\limits_{x\in X, y\in Y, \rho} & c_0^T x + d_0^T y + \Gamma\rho + \sum\limits_{i=1}^{n}(c_i^T x - \rho)^+ & \\ \text{s.t.} & Ax + By & \leq b \\ & \rho & \geq 0 \\ & x, y & \geq \mathbf{0} \\ & y_j & \in \{0,1\}\ \forall\ j \in J. \end{array} \tag{8}$$

An Optimality Condition

In the single-group robust formulation (8), for a fixed value of x, the optimal value of ρ can be described in terms of the order statistics of the values $\{c_i^T x, i = 1, \ldots, n\}$. First, note that a value of x induces an ordering $c_{[i|x]}$ on $c_i, i = 1, \ldots, n$ such that

$$c_{[1|x]}^T x \geq c_{[2|x]}^T x \geq \ldots \geq c_{[n_1|x]}^T x.$$

Then the optimal values of ρ will take on the value $\rho^* = c_{[\Gamma|x^*]}^T x^*$. This is because the marginal increase in the objective function for a marginal increase in ρ is

$$\frac{\partial}{\partial\rho} c_0^T x + d_0^T y + \Gamma\rho + \sum_{i=1}^{n}(c_i^T x - \rho)^+ = \Gamma - \sum_{i=1}^{n} \mathbf{1}_{c_i^T x - \rho \geq 0}.$$

So, the optimality condition $\frac{\partial}{\partial\rho} = 0$ when $\rho = \rho^*$ implies that

$$\begin{aligned} \Gamma &= \sum_{i=1}^{n} \mathbf{1}_{c_i^T x - \rho^* \geq 0} \\ &= |\{i : c_i^T x \geq \rho^*\}|. \end{aligned}$$

Since any increase in ρ beyond that point will increase the value $\frac{\partial}{\partial\rho}$ to be positive (and thereafter increase the objective function), we know that ρ^* will take on the value of $c_{[\Gamma|x^*]}^T x^*$. This is an optimality condition for problem (8).

This optimality condition reflects the fact that in the primal formulation (7), the variables w_i maximize the sum of the Γ largest costs (single-group).

3.4 Facility Location with Uncertain Individual Cost Types

Each individual has an unknown binary type and a known binary type. The known type denotes "group membership": individuals with known type 1 type are in group 1, etc., and group membership is observable. The unknown type denotes the individual's private "cost type": individuals with the higher cost type experience higher costs, reflected by greater cost coefficients in the objective function of an optimization problem. We know how many individuals have the high cost type, but we do not know which individuals those are. Here, we treat only the cost type of individuals in

a "single-group" problem setting; later, in Section 3.6, we introduce the group membership as a "partition" of the population of individuals.

Let $\mathcal{I}$ now denote the set of individuals. The collision with the previous use of $\mathcal{I}$ to denote a set of disturbances to nominal costs is intentional, as those disturbances will now be associated with the resolving of private features of individuals. We have restricted the uncertainty in the model to cost coefficients of decision variables in the vector labeled x (while decisions represented by the vector y have known coefficients), and we retain this structure by specifying that decisions represented by x have associated costs that depend on individuals' cost types.

Let the indices of individuals with the "high cost type" be represented by the set $\widetilde{\mathcal{I}} \subseteq \mathcal{I}$. For each $i \in \mathcal{I}$, the vector $\mathbf{x}_i \geq 0$ corresponds to decisions that determine the outcome for individual i; the vector $\mathbf{x}$ contains all these decision variables. The vector $\mathbf{y}$ corresponds to decision variables that do not relate specifically to one individual, and cost coefficients of $\mathbf{y}$ do are not uncertain.

To ground the discussion in a tangible example, we use a capacitated facility location problem in which the goal is to locate m new facility locations from the index set $\mathcal{J}$, within which $\mathcal{J}_0 \subset \mathcal{J}$ denotes facilities that already exist. For each facility location index j, there is a cost d_j associated with that location, as well as a capacity b_j such that no more than b_j individuals can be assigned to that location.

Each individual must be assigned to exactly one facility, and each assignment of individual i to facility j has associated cost c^0_{ij} plus an extra cost ("disruption") $\overline{c}_{ij}$ if individual i has the private high-cost type. The robust formulation assigns facilities so that in the worst possible realization of cost types, the cost is minimized.

As a running example, we can think of the known group-membership type as "race" and the private cost type as "car ownership". This is motivated by the fact that in many states, individuals' addresses and race are publicly available through "voter files" compiled by election officials, whereas car ownership (a vital parameter in a spatial optimization problem) is only available via the aggregate statistics provided by the US Census. The method for inferring the number of car owners in each racial group (a quantity leveraged in Sections 3.5 and 3.6) is described in Section 2.1.

$$
\begin{array}{lll}
\min & \displaystyle\sum_{j\in\mathcal{J}} d_j y_j + \sum_{i\in I}\sum_{j\in\mathcal{J}} c^0_{ij} x_{ij} + \max_{\substack{\widetilde{\mathcal{I}}\text{s.t.} \\ |\widetilde{\mathcal{I}}|\leq\Gamma}} \left\{ \sum_{i\in\widetilde{\mathcal{I}}}\sum_{j\in\mathcal{J}} \overline{c}_{ij} x_{ij} \right\} & \\
\text{s.t.} & \displaystyle\sum_{i\in\mathcal{I}} x_{ij} \leq b_j y_j & \forall\ j \in \mathcal{J} \\
& \displaystyle\sum_{j\in\mathcal{J}} y_j \leq m + |\mathcal{J}_0| & \\
& \displaystyle\sum_{j\in\mathcal{J}} x_{ij} = 1 & \forall\ i \in \mathcal{I} \\
& y_j = 1 & \forall\ j \in \mathcal{J}_0 \\
& x_{ij} \geq 0 & \forall\ i \in \mathcal{I}, j \in \mathcal{J} \\
& y_j \in \{0,1\} & \forall\ j \in \mathcal{J}.
\end{array}
\tag{9}
$$

It should be clear that this problem has the structure of formulation (7), and so it has the same

mixed integer linear programming reformulation as in (8):

$$
\begin{array}{lll}
\min\limits_{x,y,\rho} & \sum\limits_{j\in\mathcal{J}} d_j y_j + \Gamma\rho + \sum\limits_{j\in\mathcal{J}} c^0_{ij} x_{ij} + \sum\limits_{i\in I}\left(-\rho + \sum\limits_{j\in\mathcal{J}} \bar{c}_{ij} x_{ij}\right)^+ & \\
\text{s.t.} & \sum\limits_{i\in\mathcal{I}} x_{ij} \leq b_j y_j & \forall\ j\in\mathcal{J} \\
& \sum\limits_{j\in\mathcal{J}} y_j \leq m + |\mathcal{J}_0| & \\
& \sum\limits_{j\in\mathcal{J}} x_{ij} = 1 & \forall\ i\in\mathcal{I} \\
& y_j = 1 & \forall\ j\in\mathcal{J}_0 \\
& x_{ij} \geq 0 & \forall\ i\in\mathcal{I}, j\in\mathcal{J} \\
& \rho \geq 0 & \\
& y_j \in \{0,1\} & \forall\ j\in\mathcal{J}.
\end{array}
\tag{10}
$$

The following table summarizes the connections between the running example and the general optimization framework:

	General Optimization Setting	Facility Location Problem
Indices for Disturbances	$\mathcal{I}$ contains the indices of possible disturbances to nominal costs.	$\mathcal{I}$ indexes all individuals, each of whom may experience the "disturbance" of having the high cost private type.
Total Number of Disturbances	$n = \|\mathcal{I}\|$ denotes the number of possible disturbances to nominal costs.	$n = \|\mathcal{I}\|$ is the total number of individuals.
Disturbances	$\{c_i : i \in \mathcal{I}\}$ are extra costs that may be experienced above the nominal cost c_0.	$\{c_i : i \in \mathcal{I}\}$ are the extra costs for each individual of the "high cost" type defined by a private binary feature (e.g. car ownership).
Number of Disturbances Realized in Each Scenario	Γ is the number of cost disturbances that it is known will be experienced in a ground-truth resolution of uncertainty.	Γ is the (known) total number of individuals of the high cost type.

3.5 Partitioning the Uncertainty Budget

In the setting we are developing, "worse outcomes" for a group correspond to higher costs for their individuals. On the other hand, higher coefficients in an objective function assign a higher weight or "priority" to their variables: in general, all else being equal, raising a cost coefficient reduces the optimal value of a variable, at the cost of increasing costs associated with other variables. If an optimization model's coefficients reflect higher costs for individuals in one group, then the variables

associated with members of that group are in effect given higher "weight" in the model, and the corresponding individuals will, to an extent, be assigned better outcomes than they would otherwise (even if they are still worse outcomes than individuals in the other group experience). Thus, any effort to capture in an optimization model the true higher costs experienced by one group will serve to improve their outcomes in the solution.

Efforts to improve outcomes for groups with systematically worse outcomes can be characterized as "affirmative action" when an exogenously-determined resource allocation is devoted to that group. In contrast, improved accuracy in model parameters (to reflect true higher costs for one group) *endogenously* determine a level of effort (via higher coefficient weights) towards improving outcomes for that group. As mentioned in Section 2.3, we aim to achieve "fairness through awareness". We ultimately seek to avoid algorithmic bias by addressing unobserved cost disparities between groups whose membership is known.

Recall the setting mentioned in Section 3.1 of considering both known individual-level features and private ones that are only observed through aggregation. We refer to individuals sharing the same known individual-level feature as a "group", and define "between-group" bias as a disparity in outcomes for individuals in the two groups. It is possible that in some cases, due to existing differences in parameters corresponding to individuals in two groups, individuals in one group will have measurably worse outcomes in any feasible solution to a problem. We motivated a desire for "fairness" in Section 2.3.

The idea of an uncertainty "budget" Γ in an optimization model has been a successful tool (beginning with [6]) to understand the tradeoff between conservatism towards extreme scenarios in uncertainty space and average or nominal performance of an optimization model. In our model, on the other hand, Γ corresponds to a level of uncertainty derived directly from aggregated data representing private individual-level features (as described in Section 2.1).

We return to the model (7) and its simplified mixed-integer linear form (8). Let $\mathcal{I} = \{1, \ldots, n\}$ denote the set of indices for the uncertain "disturbances" to the nominal cost c_0, of which at most Γ are considered in the optimization ($\mathcal{I}$ will come to index individuals in the next section). Now suppose we partition $\mathcal{I} = \mathcal{I}_1 \sqcup \mathcal{I}_2$, where $n_1 = |\mathcal{I}_1|, n_2 = |\mathcal{I}_2|, n_1 + n_2 = n = |\mathcal{I}|$, as well as the uncertainty budget $\Gamma = \Gamma_1 + \Gamma_2$, where $\Gamma_1 \leq n_1, \Gamma_2 \leq n_2$. As long as $n_1, n_2 > 0$, the number of possible "scenarios" in which uncertainty resolves is considerably reduced by only considering those in which the number of disturbances $c_i, i \in \mathcal{I}_1$ is at most Γ_1 and the number of disturbances $c_i, i \in \mathcal{I}_2$ is at most Γ_2. Namely, scenarios respecting the partitioned structure are always possible in the single-group formulation, but many in the single-group formulation are not included in the partitioned formulation.

The "partitioned" formulation yields the following optimization problem:

$$\begin{array}{lll} \min\limits_{x\in X, y\in Y} \max\limits_{w} & c_0^T x + d_0^T y + \sum\limits_{i\in\mathcal{I}_1} w_i c_i^T x + \sum\limits_{i\in\mathcal{I}_2} w_i c_i^T x & \\ \text{s.t.} & Ax + By \leq b & \\ & \sum\limits_{i\in\mathcal{I}_1}^{n} w_i \leq \Gamma_1 & \\ & \sum\limits_{i\in\mathcal{I}_2}^{n} w_i \leq \Gamma_2 & \\ & 0 \leq w_i \leq 1 & \forall\ i = 1, \ldots, n \\ & x, y \geq \mathbf{0} & \\ & y_j \in \{0,1\}\ \forall\ j \in J, & \end{array} \tag{11}$$

whose solution can be found by solving the following mixed-integer linear program:

$$\begin{array}{lll} \min\limits_{x\in X, y\in Y, \rho_1, \rho_2} & c_0^T x + d_0^T y + \Gamma_1\rho_1 + \Gamma_2\rho_2 + \sum\limits_{i\in\mathcal{I}_1} (c_i^T x - \rho_1)^+ + \sum\limits_{i\in\mathcal{I}_2} (c_i^T x - \rho_2)^+ & \\ \text{s.t.} & Ax + By \leq b & \\ & \rho_1, \rho_2 \geq 0 & \\ & x, y \geq \mathbf{0} & \\ & y_j \in \{0,1\}\ \forall\ j \in J. & \end{array} \tag{12}$$

To summarize the difference between the "single-group" and the "partitioned" models is summarized by the following:

	Single-Group Problem	Partitioned Problem
Indices for Disturbances	$\mathcal{I}$	$\mathcal{I}_1 \sqcup \mathcal{I}_2 = \mathcal{I}$
Total Number of Disturbances	$n = \lvert\mathcal{I}\rvert$	$n_1 = \lvert\mathcal{I}_1\rvert, n_2 = \lvert\mathcal{I}_2\rvert$
Disturbances	$\{c_i : i \in \mathcal{I}\}$	$\{c_i : i \in \mathcal{I}_1\} \sqcup \{c_i : i \in \mathcal{I}_2\}$
Number of Disturbances Realized in Each Scenario	Γ	$\Gamma_1 + \Gamma_2 = \Gamma$.

By the above reasoning regarding scenarios included in the single-group versus the partitioned formulation, it is clear that the optimal value of the partitioned formulations (11) and (12) is less than or equal to the optimal value of the single-group formulations (7) and (8).

Similar to the optimality condition in Section 3.3 for the single-group formulation (8), in the partitioned problem, the optimal solution will satisfy the condition

$$\rho_1^* = c_{1[\Gamma_1|x^*]}^T x^*$$
$$\rho_2^* = c_{2[\Gamma_2|x^*]}^T x^*.$$

3.6 Facility Location with Two Groups and Groupwise Uncertainty Budgets

We grounded the general robust optimization setting (7) in a specific running facility location example (9). In the previous section, we treated the known group membership types via a partitioning of the uncertainty set, resulting in the "partitioned" formulation (11). We introduce the same partitioning (by group membership) to the facility location problem (9) to yield the following

$$
\begin{array}{lll}
\min & z_1 + z_2 & \\
\text{s.t.} \quad z_1 & = \displaystyle\sum_{i\in I_1}\sum_{j\in\mathcal{J}} c^0_{ij}x_{ij} + \max_{\substack{\widetilde{\mathcal{I}}_1 \text{s.t.} \\ |\widetilde{\mathcal{I}}_1|\le\Gamma_1}} \left\{ \sum_{i\in\widetilde{\mathcal{I}}_1}\sum_{j\in\mathcal{J}} (\overline{c}_{ij} - c^0_{ij})x_{ij} \right\} & \\
z_2 & = \displaystyle\sum_{i\in I_2}\sum_{j\in\mathcal{J}} c^0_{ij}x_{ij} + \max_{\substack{\widetilde{\mathcal{I}}_2 \text{s.t.} \\ |\widetilde{\mathcal{I}}_2|\le\Gamma_2}} \left\{ \sum_{i\in\widetilde{\mathcal{I}}_2}\sum_{j\in\mathcal{J}} (\overline{c}_{ij} - c^0_{ij})x_{ij} \right\} & \\
\displaystyle\sum_{i\in\mathcal{I}} x_{ij} & \le b_j y_j & \forall\, j \in \mathcal{J} \\
\displaystyle\sum_{j\in\mathcal{J}} y_j & \le m + |\mathcal{J}_0| & \\
\displaystyle\sum_{j\in\mathcal{J}} x_{ij} & = 1 & \forall\, i \in \mathcal{I} \\
y_j & = 1 & \forall\, j \in \mathcal{J}_0 \\
x_{ij} & \ge 0 & \forall\, i \in \mathcal{I}, j \in \mathcal{J} \\
y_j & \in \{0,1\} & \forall\, j \in \mathcal{J}.
\end{array}
\tag{13}
$$

In the same way that partitioned problem formulation (11) is equivalent to the MILP (12), the partitioned facility location problem (13) is equivalent to the following:

$$
\begin{array}{lll}
\min & z_1 + z_2 & \\
\text{s.t.} \quad z_1 & = \rho_1\Gamma_1 + \displaystyle\sum_{j\in\mathcal{J}} c^0_{ij}x_{ij} + \sum_{i\in I_1}\left(-\rho_1 + \sum_{j\in\mathcal{J}} (\overline{c}_{ij} - c^0_{ij})x_{ij} \right)^+ & \\
z_2 & = \rho_2\Gamma_2 + \displaystyle\sum_{j\in\mathcal{J}} c^0_{ij}x_{ij} + \sum_{i\in I_2}\left(-\rho_2 + \sum_{j\in\mathcal{J}} (\overline{c}_{ij} - c^0_{ij})x_{ij} \right)^+ & \\
\displaystyle\sum_{i\in\mathcal{I}} x_{ij} & \le b_j y_j & \forall\, j \in \mathcal{J} \\
\displaystyle\sum_{j\in\mathcal{J}} y_j & \le m + |\mathcal{J}_0| & \\
\displaystyle\sum_{j\in\mathcal{J}} x_{ij} & = 1 & \forall\, i \in \mathcal{I} \\
y_j & = 1 & \forall\, j \in \mathcal{J}_0 \\
x_{ij} & \ge 0 & \forall\, i \in \mathcal{I}, j \in \mathcal{J} \\
\rho_1, \rho_2 & \ge 0 & \\
y_j & \in \{0,1\} & \forall\, j \in \mathcal{J}.
\end{array}
\tag{14}
$$

The following table summarizes the connections between the running example and the general optimization framework within the context of the partitioned population:

	General Optimization Setting	Facility Location Problem
Indices for Disturbances	$\mathcal{I}_1 \sqcup \mathcal{I}_2 (= \mathcal{I})$ contain the indices of possible disturbances to nominal costs, partitioned into two subsets.	$\mathcal{I}_1 \sqcup \mathcal{I}_2 (= \mathcal{I})$ contain the indices of all individuals in two groups defined by an observable "group membership" type (e.g. race).
Total Number of Disturbances	$n_1 = \lvert\mathcal{I}_1\rvert$ and $n_2 = \lvert\mathcal{I}_2\rvert$ denotes the total number of possible disturbances to nominal costs in each subset.	$n_1 = \lvert\mathcal{I}_1\rvert$ and $n_2 = \lvert\mathcal{I}_2\rvert$ denotes the number of individuals in the two groups defined by observable type (e.g. number of individuals from each racial group).
Disturbances	$\{c_i : i \in \mathcal{I}_1\} \sqcup \{c_i : i \in \mathcal{I}_2\}$ are the extra costs that may be experienced above the nominal cost c_0, partitioned into two subsets.	$\{c_i : i \in \mathcal{I}_1\} \sqcup \{c_i : i \in \mathcal{I}_2\}$ are the extra costs for each individual of the "high cost", partitioned by group membership.
Number of Disturbances Realized in Each Scenario	$\Gamma_1 + \Gamma_2 (= \Gamma)$ are the numbers of cost disturbances from each subset that it is known will be experienced in a ground-truth resolution of uncertainty.	$\Gamma_1 + \Gamma_2 (= \Gamma)$ are the (known) total numbers of individuals of the high cost private type from each of the groups (e.g. the known number of car owners within and without a given racial group).

3.6.1 Probability of Improvement Given Identically Distributed Parameters

Note that the single-group and partitioned models share all parameters except for the partition itself. That is, the constraint matrices A and B, the base costs c_0 and d_0, and the cost disturbances $c_1, \ldots, c_n$ are the same. Now consider the idea that these parameters are generated from a distribution such that the $c_1, \ldots, c_n$ are independently and identically distributed, and such that the partitioning of these disturbances $\mathcal{I}_1 \sqcup \mathcal{I}_2 = \mathcal{I}$, and the partitioning of the uncertainty budget $\Gamma_1 + \Gamma_2 = \Gamma$ is a random partitioning (subject to $\Gamma_1 \leq n_1 = \lvert\mathcal{I}_1\rvert$ and $\Gamma_2 \leq n_2 = \lvert\mathcal{I}_2\rvert$) in which all possible partitionings are equally likely. In this setting, the only systematic difference between the two groups $\mathcal{I}_1$ and $\mathcal{I}_2$ is their uncertainty budgets Γ_1 and Γ_2. Also suppose that the c_i parameters are generated from a continuous distribution so that there is zero probability of any two sharing an equal inner product with any vector x: that is, for any given value of x, $\{c_1^T x, \ldots, c_n^T x\}$ constitutes a set of n distinct real numbers.

It is then possible to answer the question, "What is the probability that the partitioned problem

actually has optimal cost that is exactly as high as the single-group problem?" The objective of the single-group problem contains the term

$$\sum_{i=1}^{\Gamma} c_{[i|x]}^T x, \tag{15}$$

and we recall the definition $c_{[1|x]}^T x \geq \ldots \geq c_{[\Gamma|x]}^T x \geq \ldots \geq c_{[n|x]}^T x$. The objective of the partitioned problem contains the term

$$\sum_{i=1}^{\Gamma_1} c_{1[i|x]}^T x + \sum_{i=1}^{\Gamma_2} c_{2[i|x]}^T x, \tag{16}$$

and we recall that $c_{1[1|x]}^T x \geq \ldots \geq c_{1[\Gamma_1|x]}^T x \geq \ldots \geq c_{1[n_1|x]}^T x$ and $c_{2[1|x]}^T x \geq \ldots \geq c_{2[\Gamma_2|x]}^T x \geq \ldots \geq c_{2[n_2|x]}^T x$. Furthermore, the cost vectors are the same ones, partitioned into two groups:

$$\{c_1, \ldots, c_n\} = \{c_{11}, \ldots, c_{1n_1}\} \sqcup \{c_{21}, \ldots, c_{2n_2}\}.$$

By assumption the set $\{c_1^T x, \ldots, c_n^T x\}$ consists of n distinct real numbers, and so the only possible way for the sum of the Γ_1 cost terms with indices in $\mathcal{I}_1$ and the Γ_2 cost terms with indices in $\mathcal{I}_2$ that yield the order statistics in expression (16) to yield the same sum as in expression (15) is for the exact same cost vectors to be represented in the single and partitioned sums of order statistics.

There are $\binom{\Gamma}{\Gamma_1} = \binom{\Gamma}{\Gamma_2}$ total ways to partition the Γ cost vectors whose inner products with a value of x have the highest order statistics into two subsets of size Γ_1 and Γ_2. On the other hand, in total there are 2^Γ total ways to distribute these cost vectors into two subsets such that any number of them are in each of the two subsets (we assume that $n_1, n_2 > \Gamma$ for simplicity - otherwise the result is not so much different). Thus, the probability of the partitioned problem having an objective value that is *as costly* as the objective value of the single-group problem is

$$\frac{\binom{\Gamma}{\Gamma_1}}{2^\Gamma}.$$

This value is in general low, and it is highly unlikely that the partitioning does not yield an increase in objective value. That is, the less conservative uncertainty set is highly likely to yield a lower objective value (and therefore an objective value that is closer to the objective value of the problem in which the ground-truth costs are all known, of which the robust formulation is always an over-estimate).

3.7 The Uncertainty Space Before and After Partitioning the Uncertainty Budget

We model the variables w_i, which resolve the uncertainty in (1), as continuous rather than discrete because the resulting optimization model is equivalent with this slightly less constrained setting: optimal values of w_i will be binary in the optimization problems (7) and (11). However, our interest is in discrete "disturbances" to nominal costs, and by considering the uncertainty space Ω

in which the uncertainty-resolving variables w_i only do take on discrete values, we can articulate some advantages to the partitioning in the formulation in the previous Section 3.5. The uncertainty space Ω consists of all feasible binary assignments to all of the w_i variables, and so the uncertainty space itself is discrete.

We refer to one element of the uncertainty set as a possible "scenario". This terminology is commonly associated with stochastic optimization, rather than robust optimization, but our model of uncertainty is relevant to both these types of problems, as is articulated in Section 3.2. So, by reducing the uncertainty in the model via partitioning, we reduce the conservatism of the model (in a typical robust optimization sense), and we also reduce the number of possible scenarios (in a typical stochastic optimization sense). We also discuss the extent to which parameters are "accurate" in a randomly selected scenario, a notion that may only be relevant to stochastic optimization.

For one, the partitioning of the uncertainty budget restricts the total number of possible configurations in which uncertainty can resolve. Before partitioning, there are $\binom{n}{\Gamma}$ total possible scenarios in which uncertainty can resolve. After partitioning, there are $\binom{n_1}{\Gamma_1}\binom{n_2}{\Gamma_2}$ possible scenarios in which uncertainty can resolve.

Perhaps more interestingly, the expected number of "correct" parameters in a randomly selected scenario increases as a result of the partitioning, as discussed in the next section.

3.7.1 Correct Parameters in a Randomly Selected Scenario

In the single-group model of uncertainty, there are Γ "true" cost disturbances from among n possibilities. By randomly selecting some Γ cost disturbances from among n, the number of "correctly" selected disturbances follows a hypergeometric distribution, with expected value Γ^2/n.

Now we consider the case then the true uncertainty space consists of all scenarios in which there are exactly Γ_1 "true" disturbances from the first $n_1 = |\mathcal{I}_1|$ possibilities, and exactly Γ_2 from the other $n_2 = |\mathcal{I}_2|$ possibilities. In this case, we can compare the (partitioned) optimization model that reflects this uncertainty space, versus the single-group optimization model that does not distinguish between the two groups.

First, we define $X = \frac{|\mathcal{I}_1|}{n}, (1-X) = \frac{\|\mathcal{I}_2|}{n}$ denote the fractions of disruptions in the two partitions $\mathcal{I}_1$ and $\mathcal{I}_2$, $T = \frac{\Gamma}{n}$ denote the total fraction of "correct" disruptions, and we let $\beta_1 = \frac{\Gamma_1}{n_1}, \beta_2 = \frac{\Gamma_2}{n_2}$ denote the fraction of "correct" disruptions in the two partitions.

Define a randomly selected scenario in the "Single Group Problem" setting as one in which Γ disruptions are selected at random from among all n of the possible disruptions. Define a randomly selected scenario in the "Partitioned Problem" setting as one in which Γ_1 disruptions are selected at random from among all n_1 possible disruptions with indices in $\mathcal{I}_1$, and Γ_2 disruptions are selected at random from among all n_2 possible disruptions with indices in $\mathcal{I}_2$.

The following table gives the expected number of "correctly-identified" disturbances in a randomly selected scenario in both problem settings. That is, the expected number of indices i such that the cost c_i is a true cost in both the real system represented by the model and in a given scenario randomly-selected from the uncertainty space.

	Single-Group Problem	Partitioned Problem
$M_{tot} = E[\#$ correctly identified$]$	$M^1_{tot} = \frac{\Gamma^2}{n}$ $= (\beta_1 X + beta_2(1-X))^2 n$	$M^2_{tot} = \frac{\Gamma_1^2}{n_1} + \frac{\Gamma_2^2}{n_2}$ $= (\beta_1^2 X + \beta_2^2(1-X))n$
$M_1 = E[\#$ correctly identified and in $\mathcal{I}_1]$	$M^1_1 = \Gamma_1 \frac{\Gamma}{n}$ $= \beta_1 X T n$	$M^2_1 = \frac{\Gamma_1^2}{n_1}$ $= \beta_1^2 X n$
$M_2 = E[\#$ correctly identified and in $\mathcal{I}_2]$	$M^1_2 = \Gamma_2 \frac{\Gamma}{n}$ $= \beta_2(1-X) T n$	$M^2_2 = \frac{\Gamma_2^2}{n_1}$ $= \beta_2^2(1-X) n$

This implies the following ratios, which express the advantages of using the "Partitioned Problem" that reflects the reality of the two groups indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$:

- $\frac{M^2_1}{M^1_1} = \frac{\beta_1}{T}$ $\left(= \frac{\text{\# correctly identified disturbances in partitioned problem setting}}{\text{\# correctly identified disturbances in single-group problem setting}}\right)$
- $\frac{M^2_2}{M^2_2} = \frac{\beta_2}{T}$ $\left(= \frac{\text{\# correctly identified disturbances indexed by } \mathcal{I}_1 \text{ in partitioned problem setting}}{\text{\# correctly identified disturbances indexed by } \mathcal{I}_1 \text{ in single-group problem setting}}\right)$
- $\frac{M^2_{tot}}{M^1_{tot}} = \frac{\beta_1^2 X + \beta_2^2(1-X)}{(\beta_1 X + \beta_2(1-X))^2}$ $\left(= \frac{\text{\# correctly identified disturbances indexed by } \mathcal{I}_2 \text{ in partitioned problem setting}}{\text{\# correctly identified disturbances indexed by } \mathcal{I}_2 \text{ in single-group problem setting}}\right)$

Note that when $\beta_1 = \beta_2$, we also necessarily have $\beta_1 = \beta_2 = T$, and this implies that $\frac{M^2_{tot}}{M^1_{tot}} = 1$. That is, when there are no differences between groups $\mathcal{I}_1$ and $\mathcal{I}_2$ in terms of frequency of a "true" disturbance among the possible disturbances indexed by the two sets, then there is no advantage overall to the number of correctly-identified disturbances.

4 Conclusion

In many centralized decisionmaking settings, the only information about the individuals affected is gleaned through aggregate datasets. Still, as computational capacity and the demand for automation increases, it is difficult to resist considering the individual as the unit of analysis in optimization. Aggregation yields a well-defined uncertainty structure, and existing methods from robust and stochastic optimization can effectively treat this uncertainty, as we demonstrate.

All automated decisionmaking is prone to algorithmic bias via inequality in outcomes for members of different groups. If the individual is to be the unit of analysis in centralized decisionmaking, there are more opportunities to ask what information is known about each individual, and what information we can infer. To the extent that an inference procedure can estimate individual-level

features, and to the extent that a treatment of those features can prevent algorithmic bias, it becomes imcumbent upon decisionmakers to do so.

Aggregation is not just a ubiquitous, but a fundamentally important ingredient of privacy. As data availability increases alongside maturing privacy mechanisms, it is possible that decisionmakers will increasingly have access to valuable yet aggregated data about the individuals whom their decisions affect. Ecological inference not a widely used framework outside of the social sciences, and yet it is extremely well-suited to refining uncertain knowledge of individual features in this setting.

This paper takes an agnostic approach to the ecological inference procedure used, and in fact only analyzes the process of analyzing before and after *perfectly accurate* ecological inferences have been obtained. In reality, several different assumptions regarding aggregation bias have been proposed, all which lead to slightly different ecological estimate procedures with different standard errors that depend on the true aggregation bias. To truly test the efficacy of an ecological inference procedure viz a viz an optimization model, it is necessary to already have access to the individual-level data that is being estimated. We look forward to future work that utilizes such data to test the efficacy of existing ecological inference techniques, and to incorporate the the errors in the estimates into an optimization setting.

References

[1] Igor Averbakh and Oded Berman. Minmax regret median location on a network under uncertainty. *INFORMS Journal on Computing*, 12(2):104–110, 2000.

[2] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 181–190, New York, NY, USA, 2007. Association for Computing Machinery.

[3] Opher Baron, Joseph Milner, and Hussein Naseraldin. Facility location: A robust optimization approach. *Production and Operations Management*, 20(5):772–785, 2011.

[4] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization, 2010.

[5] Dimitris Bertsimas, Vivek Farias, and Nikolaos Trichakis. The price of fairness. *Operations Research*, 59:17–31, 02 2011.

[6] Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[7] Timothy C. Y. Chan, Zuo-Jun Max Shen, and Auyon Siddiq. Robust defibrillator deployment under cardiac arrest location uncertainty via row-and-column generation, 2017.

[8] Violet Xinying Chen and J. N. Hooker. Balancing fairness and efficiency in an optimization model, 2020.

[9] W. A. V. Clark and Karen L. Avery. The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4):428–438, 1976.

[10] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

[14] Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta. Statistically valid inferences from privacy protected data, Working Paper.

[15] Georgina Evans, Gary King, Adam D. Smith, and Abhradeep Thakurta. Differentially private survey research, Working Paper.

[16] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact, 2014.

[17] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, page 351–360, New York, NY, USA, 2009. Association for Computing Machinery.

[18] Michael T Hannan. Approaches to the aggregation problem. *Stanford Sociology Technical Reports and Working Papers 1961-1993*, (46), 1972.

[19] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016.

[20] Wenxin Jiang, Gary King, Allen Schmaltz, and Martin A. Tanner. Ecological regression with partial identification. *Political Analysis*, 28(1):65–86, Aug 2019.

[21] Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, 1997.

[22] Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, 1997.

[23] Gary King, Ori Rosen, Martin Tanner, Gary King, Ori Rosen, and Martin A. Tanner. *Ecological Inference: New Methodological Strategies*. Cambridge University Press, New York, 2004.

[24] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. K-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 502–510, New York, NY, USA, 2011. Association for Computing Machinery.

[25] Devan V. Mehrotra, Fang Liu, and Thomas Permutt. Missing data in clinical trials: control-based mean imputation and sensitivity analysis. *Pharmaceutical Statistics*, 16(5):378–392, 2017.

[26] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, 2008.

[27] Susan Hesse Owen and Mark S. Daskin. Strategic facility location: A review. *European Journal of Operational Research*, 111(3):423–447, 1998.

[28] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review, 2019.

[29] W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950.

[30] Zuo-Jun Max Shen, Roger Lezhou Zhan, and Jiawei Zhang. The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *INFORMS Journal on Computing*, 23(3):470–482, 2011.

[31] W. Phillips Shively. "ecological" inference: The use of aggregate data to study individuals. *American Political Science Review*, 63(4):1183–1196, 1969.

[32] Lawrence V. Snyder. Facility location under uncertainty: a review. *IIE Transactions*, 38(7):547–564, 2006.

[33] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. *CoRR*, abs/1807.00787, 2018.

[34] A. Tang, J. Wang, and S.H. Low. Is fair allocation always inefficient. In *IEEE INFOCOM 2004*, volume 1, page 45, 2004.

[35] Tara L. Terry. Robust linear optimization with recourse: Solution methods and other properties. *[Unpublished doctoral dissertation]. The University of Michigan*, 2009.

[36] Bram Verweij, Shabbir Ahmed, Anton Kleywegt, George Nemhauser, and Alexander Shapiro. The sample average approximation method applied to stochastic routing problems: A computational study. *Computational Optimization and Applications*, 24:289–333, 02 2003.

[37] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment and disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017.

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification, 2017.

[40] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.