Mixed Kernel Density Estimation of Procurement Auctions

Nick Fertitta, Nick Hayeck and Skylor Matsuda

University of Virginia Department of Economics.

Contributing authors: naf3wk@virginia.edu; hayeck@virginia.edu; smm7rv@virginia.edu;

Abstract

In this paper, we present a method for the estimation of cost distributions within the context of procurement auctions for infrastructure projects in the state of California. A non-parametric mixed kernel method is selected for this task and the cost distributions are recovered. These distributions are then used to explore an alternative auction structure which the California Department of Transportation (CalTrans) may implement, finding that CalTrans could decrease their expected cost by roughly fifty percent.

1 Introduction

For over a hundred years, Caltrans has been a government agency that has serviced all kinds of transportation systems in the state of California. People often see construction happening on their roads, in their airports, or with the train system, but they seldom stop to think about what happens behind the scenes. How do people decide on which firms are going to do a certain job? And at what price will they do the given job? Caltrans is there to answer these questions and it is their concern that this gets done in the most fair and competitive way possible. After all, Caltrans wants to pay the lowest price that they can and competition helps them do that. Caltrans addresses this problem by holding first price auctions where the lowest bid wins.

One problem that Caltrans faced was that small firms seemed to be at a severe disadvantage to large firms in terms of bidding power. They dealt with this in 1973 by creating a small business preference that discounted the bids of small business to allow them to better compete in these auctions. We will

explain what all this means and the potential implications of it all as we move through the paper.

Our main goal of the paper is to recover the cost distributions of large and small businesses to see if there truly is a difference in their costs. We will first analyze the data and then begin constructing our model of how we will tackle this task. We will then use said model to carry out our estimation. Finally, we will analyze the results and also run a counterfactual to see what would happen if Caltrans did things a different way.

2 Data

This data was obtained from the California Department of Transportation (Caltrans) database. Caltrans is a governmental agency owned by the state of California that is tasked with managing the multi-modal transportation system that runs throughout the entire state. This includes overseeing intercity rail services, airports, hospital heliports, and also managing the highways and freeways throughout the state for which they are responsible for over 50,000 miles. Caltrans breaks itself down into six primary programs to specifically address the needs of its large array of transportation methods. These six primary programs are: Aeronautics, Highway Transportation, Mass Transportation, Transportation Planning, Administration, and Equipment Service Center. Caltrans emphasizes serving all people, no matter the existing the conditions of each community, while making transportation safe and respectful of the environment. Caltrans works in conjunction with the Division of Procurement and Contracts to decide on which firms are assigned to certain jobs and the agreements of those jobs in the fields of IT, service contracts, architectural and engineering contracts, minor public works, etc. As it pertains to this particular dataset, these are bids from auctions for infrastructure projects that are to be carried out by the winning firm.

Auctions for these projects are held every Friday at 10:00 a.m. PST and parties interested may participate via teleconference. The bids are sealed until after all bids have been received. The winner of the auction is the firm that submits the lowest bid, however, this is contingent on the status of said firm as a "Large" or "Small" business. A firm is considered a "Small" business if it: has its principal office in California, its owners are domiciled in California, is independently owned and operated, is not dominant in its field of operations, and either is a manufacturer with less than 100 employees or is a business with less than 100 employees and less than $12 million in gross receipts over the past three years. If all of these things apply and a business is a "small" business, then they get a 5 percent preference in the auction. This means that if a small business is the lowest bidder, it wins the auction, but it can also win the auction if it is within 5 percent of the lowest bid and all other lower bidders are not small businesses. The bids are revealed after the auction concludes and the winning bidder must agree to the job unless there are other issues with licensing or qualifications.

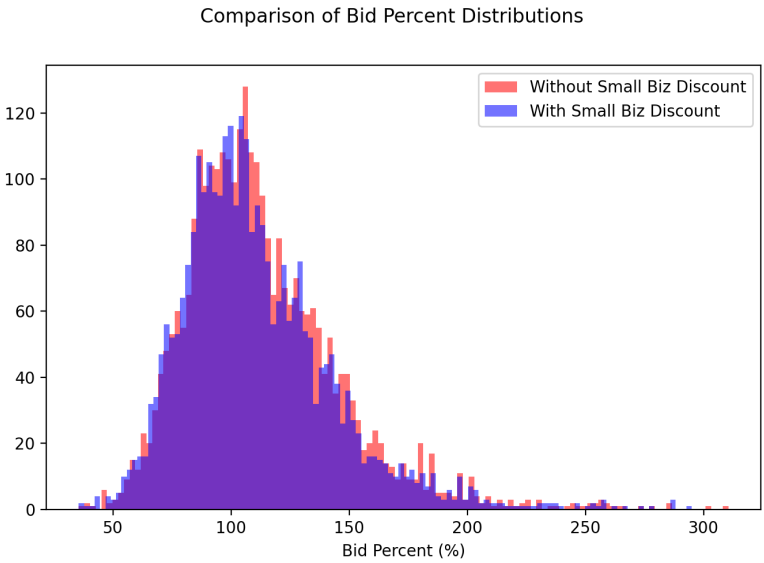

Fig. 1 Two histograms illustrating the left-bound translation of the bid percent distribution when the small business preference rule is taken into account

Summary statistics for the data are shown below in Table 1. There were 3078 bids submitted to 705 auctions. The "estimate" variable as listed in the table below is an engineer's approximation for the total construction costs for a given project. The "# Small Business Bidders" and "# Large Business Bidders" are the number of small and large business bidders for each auction.

Table 1 Compiled Summary Statistics for the Caltrans Dataset, prior to data cleaning

Metric	n_{obs}	mean	std	min	max
Bid	3078	986241.7	3.3m	44655	58.5m
Estimate	705	890364.3	3.2m	74000	60.1m
# Small Business Bidders	705	1.70	1.87	0	13
# Large Business Bidders	705	2.66	1.61	0	9

Since the auction winners are determined by not only who submits the lowest bid, but also by the five-percent small business preference rule (as described above), we introduce a transformation of the bids, labeled "Preferenced Bid" wherein the bids are transformed according to the following rule: if the bid was submitted by a small business, bid $\mapsto 0.95 * \text{bid}$; otherwise, there is no change. The change to the bid percent distribution is, predictably, a shift to the left and is illustrated in Figure 1 below. Although this shift doesn't appear to have the biggest impact on the outcomes of auctions, the small business preference can be the difference in the winner of the auction on occasion.

Additionally, since the bid is assumed to be a function of the estimate to some degree, we control for this across all bids by introducing another transformation of the data labeled "Preferenced Bid Percent" or "PBP" which is the preferenced bid, normalized by it's estimate: PBP $= \frac{\text{Preferenced Bid}}{\text{Estimate}}$. PBP will be useful in graphically displaying distributions of the data set and regressing on various characteristics of the data. Summary statistics for these new data transformations are shown below in Table 2.

Table 2 Compiled Summary Statistics for the Add-in Data Points, prior to data cleaning

	n_{obs}	mean	std	min	max
Preferenced Bid	3078	975944.67	3.3m	44655	58.5m
Pref. Bid Percent	3078	111.78%	35.9%	35.5%	705.8%

It was found that two different filters should be applied to the data to adjust for anomalous bids that violate assumptions about the rationality of the agents or auctions with that do not fit the assumptions of the first-price auction theorem. This was carried out in two parts. First, we removed auctions containing bids with large outliers and auctions with anomalously high engineer's estimates. Second, auctions which contained less than two bids were discarded, due to the assumption of FPA requiring that N $>$ 1. Adjusted summary statistics can be found below in Table 4.

Table 3 Compiled Summary Statistics, after data cleaning

	count	mean	std	min	max
Bid	3038	987292.8	3.3m	44655	58.5m
Preferenced Bid	3038	977079.7	3.3m	44655	58.5m
PBP	3038	111%	32.6%	35.5%	295%
Estimate	669	903523.6	3.3m	91000	60.0m
# Small Business Bidders	669	1.8	1.89	0	13
# Large Business Bidders	669	2.768311	1.57	0	9

Once these filters have been applied, we discover that the number of bids per auction is typically between three and six, with several outliers that indicate the distribution for the number of bidders is likely fairly fat-tailed. This is illustrated with a boxplot in Figure 2 below. This Interquartile Range of three to six bidders is music to the ears of the small business that Caltrans is trying to give slight preference to. The reasoning behind this will be explained more easily with the next graph.

Additionally, we explored the relationship between the number of bidders and size of the winning bid, as a percentage of the estimate. This regression reveals what one would expect given the FPA theorem, i.e. the size of bids decrease as the number of bidders increases. The 99% t-interval estimate states that the true parameter for the slope of this line lies within $-4.040.039$. This

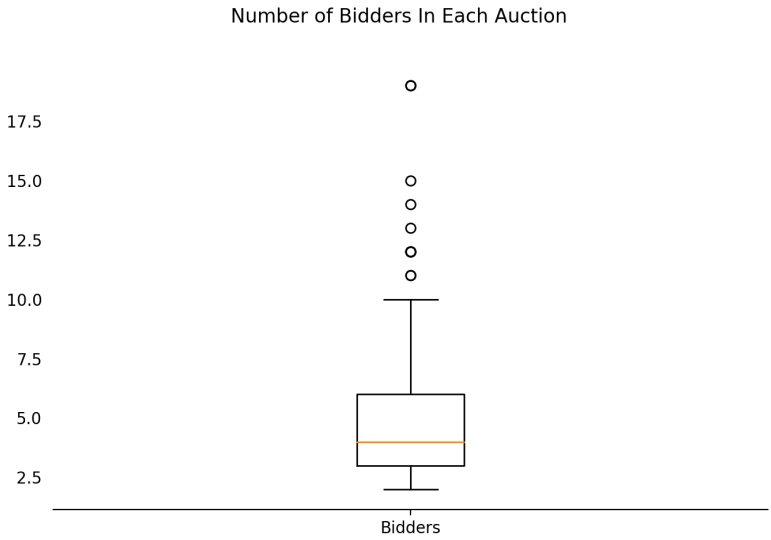

Fig. 2 Boxplot indicating the distribution for N

implies that for each additional bidder that comes into the fray, the winning bid as a percent of the estimate will decrease by around four percent. Small businesses get a preference in the form of a five percent decrease in each bid which allows them to compete a little bit better. Another important takeaway from this graph is that since more bidders is correlated with a lower winning price, Caltrans is incentivized to have as many firms involved in an auction as possible. This is good because it means the barrier to entry of an auction is very low, which checks out, given that any firm can hop on a teleconference and participate, so long as they meet qualifications. The regression chart and its residual plot are shown below in Figure 3. The residual plot seems to indicate the regressors are linear and exogenous, but potentially heteroscedastic. This implies that while the OLS estimator is unbiased, it is not strictly the most efficient, which should be acceptable for data exploration purposes.

3 Model

The structure of auctions that Caltrans uses to sell jobs is what we call asymmetric first-price auctions where the lowest bidder wins. We have already discussed the ins and outs of these auctions used by Caltrans in the previous section, but we still need to address some of the assumptions and designations that are used for our model. In each auction, there are n_s small business and n_b large business bidders, such that, for each of the L auctions, $N_i = n_s + n_b \geq 2$ for $i = 1, 2, ..., L$. Our model aims to recover the distributions of firm costs $c_s \sim F_s(\cdot)$ and $c_b \sim F_b(\cdot)$ (for small and large businesses respectively) over the domain $[c_{\min}, c_{\max}$. As observers, we have no way of knowing exactly what

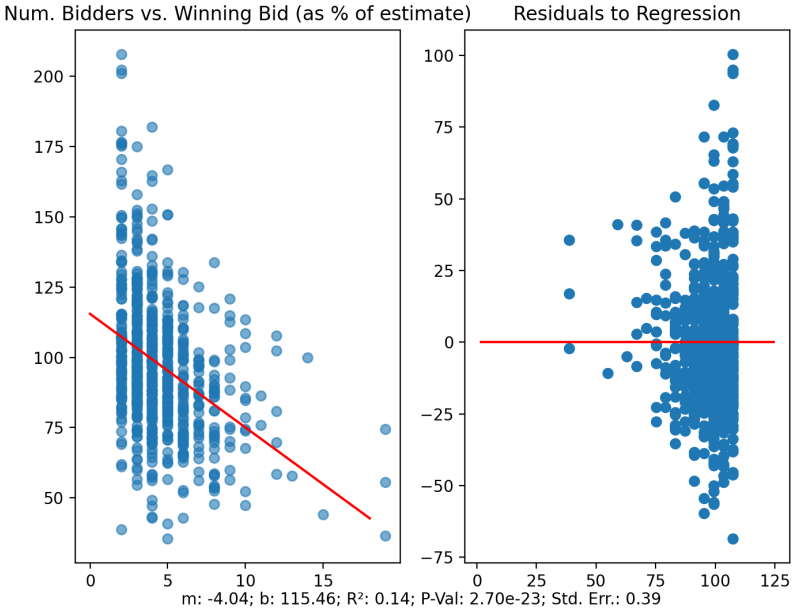

Fig. 3 Regression showing relationship between the number of bidders in an auction and the winning bid as a percentage of that auction's estimate. A negatively correlated, linear trend is observed.

each firm's cost is since we only observe bids, so we make several assumptions about their distributions.

We operate under a few assumptions:

A] Distributions for each type of bidder (small or large business) are identically and independently distributed (thus representing independent and private valuations)

This tells us that each bid by each agent within the two types of bidders is chosen from the same distribution, greatly increasing the mathematical depth of analysis we may perform. Although this assumption may not hold completely true in practice, it is certainly approximately true, especially in an industry as mature as infrastructure construction where any firm having a technological edge over all others is rare.

B] large businesses are stronger bidders than small ones (i.e. $\forall c, F_b(c) \leq F_s(c)$).

This is why these auctions are called asymmetric. Large businesses have the ability to go lower with their bids than small businesses because they almost always have lower cost than a small business. Therefore, we cannot assume that their distributions are symmetric.

C] Additionally, we assume each contract is for a similar project/job, each bidder is acting rationally, and the number of bidders are independent of the specific contract.

Under these assumptions, suppose that for a bidder of type $\tau \in \{s, b\}$ we denote their bidding strategy, a monotonic function that maps costs to bids, as $\beta_\tau(\cdot)$ and can then compute their probability of winning an auction as:

$$\begin{aligned}\Pr.(\text{"bidder wins with bid b"}) &= \Pr.(n_\tau - 1 \ \& \ n_{\bar{\tau}} \text{ bids are all greater than b}) \\ &= \Pr.(n_\tau - 1 \text{ bids} \geq b) \times \Pr.(n_{\bar{\tau}} \text{ bids} \geq b) \\ &= [1 - F_\tau(\beta_\tau^{-1}(b))]^{n_\tau - 1} \times [1 - F_{\bar{\tau}}(\beta_{\bar{\tau}}^{-1}(b))]^{n_{\bar{\tau}}}\end{aligned}$$

Which then yields the following expectation for the the bidder's profit, denoted $\Pi(b, c)$:

$$\begin{aligned}\mathbb{E}[\Pi(b,c)] &= (b-c) \times \Pr.(\text{"bidder wins with bid b"}) &(1)\\ &= (b-c) \times [1 - F_\tau(\beta_\tau^{-1}(b))]^{n_\tau - 1} \times [1 - F_{\bar{\tau}}(\beta_{\bar{\tau}}^{-1}(b))]^{n_{\bar{\tau}}} &(2)\end{aligned}$$

Note that the bid b above is "preferenced bid" (the effective bid when including the small business preference factor), this notation will be used throughout the remainder of the paper. The above then leads to the expected value maximization problem:

$$\max_b [(b-c) \times [1 - F_\tau(\beta_\tau^{-1}(b))]^{n_\tau - 1} \times [1 - F_{\bar{\tau}}(\beta_{\bar{\tau}}^{-1}(b))]^{n_{\bar{\tau}}}] \tag{3}$$

Taking the first order condition $\frac{\partial \mathbb{E}[\Pi(b,c)]}{\partial b} = 0$ yields the (rather onerous expression):

$$\begin{aligned}0 = &\left[1 - F_\tau\left(\beta_\tau^{(-1)}(b)\right)\right]^{n_\tau - 2} \left[1 - F_{\bar{\tau}}\left(\beta_{\bar{\tau}}^{(-1)}(b)\right)\right]^{n_{\bar{\tau}}} \\ &(1 - \frac{n_{\bar{\tau}}(b-c)\left(F_\tau\left(\beta_\tau^{(-1)}(b)\right) - 1\right) F'_{\bar{\tau}}\left(\beta_{\bar{\tau}}^{(-1)}(b)\right)}{\left(F_{\bar{\tau}}\left(\beta_{\bar{\tau}}^{(-1)}(b)\right) - 1\right) \beta'_{\bar{\tau}}\left(\beta_{\bar{\tau}}^{(-1)}(b)\right)} \\ &- F_\tau\left(\beta_\tau^{(-1)}(b)\right) - \frac{(n_\tau - 1)(b-c) F'_\tau\left(\beta_\tau^{(-1)}(b)\right)}{\beta'_\tau\left(\beta_\tau^{(-1)}(b)\right)})\end{aligned}$$

Using the fact that $\beta_\tau^{(-1)}(b)$ is equal to c (since $\beta_\tau(\cdot)$ is monotonic, this is bijective), $F'(\cdot) = f(\cdot)$ (the density) and rearranging terms, we come to:

$$(1 - F_\tau(c))^{n_\tau - 1}(1 - F_{\bar{\tau}}(c))^{n_{\bar{\tau}}} - \frac{(n_\tau - 1)(b-c)(1 - F_\tau(c))^{n_\tau - 2} F'_\tau(c)(1 - F_{\bar{\tau}}(c))^{n_{\bar{\tau}}}}{\beta'_\tau(c)}$$

$$-\frac{n_{\bar{\tau}}(b-c)(1-F_{\tau}(c))^{n_{\tau}-1}(1-F_{\bar{\tau}}(c))^{n_{\bar{\tau}}-1}F'_{\bar{\tau}}(c)}{\beta'_{\bar{\tau}}(c)}$$

Dividing out the $(1-F_{\tau}(c))^{n_{\tau}-1}(1-F_{\bar{\tau}}(c))^{n_{\bar{\tau}}}$ term, we can then replace $\tau \in \{\text{small}, \text{big}\}$ and recover a system of partial differentiation equations:

$$1 = (b-c)\left(\frac{(n_s-1)\times f_s(c)}{(1-F_s(c))\times \beta'_s(\beta_s^{(-1)}(b))} + \frac{n_b \times f_b(c)}{(1-F_b(c))\times \beta'_b(\beta_b^{(-1)}(b))}\right)$$
$$1 = (b-c)\left(\frac{n_s \times f_s(c)}{(1-F_s(c))\times \beta'_s(\beta_s^{(-1)}(b))} + \frac{(n_b-1)\times f_b(c)}{(1-F_b(c))\times \beta'_b(\beta_b^{(-1)}(b))}\right)$$

4 Strategy

4.1 Identification

Under the assumption that $\beta(\cdot)$ is monotonic, we directly find that $\beta^{-1}(\cdot) : B \mapsto V$ is well defined. It then follows that:

$$G_{\tau}(b) = Pr.(B_{\tau} \leq b)$$
$$G_{\tau}(b) = Pr.(\beta_{\tau}^{-1}(V) \leq \beta_{\tau}^{-1}(b))$$
$$G_{\tau}(b) = F_{\tau}(\beta^{-1}(b))$$

Where G is the observed bid distribution. We will later condition this variable on the engineer's estimate, the number of small business bidders, and the number of large business bidders, but this has been omitted from the above derivation to reduce notational clutter. Taking the derivative of the previous equality we find:

$$F'_{\tau}(\beta^{-1}(b)) = G'_{\tau}(b)$$
$$\frac{f_{\tau}(\beta^{-1}(b))}{\beta'_{\tau}(\beta_{\tau}^{-1}(b))} = g_{\tau}(b)$$

We then replace $F(\beta^{-1}(b))$ by $G(b)$ and $\frac{f(\beta^{-1}(b))}{\beta'(\beta^{-1}(b))}$ by $g(b)$ in the equation that we found in the previous section, and condition the bid distribution on the engineer's estimate, number of small bidders, and number of large business bidders. We define the engineer's estimate as a random variable X and specific estimates as x; the number of small business bidders as N_s and specific observations n_s; the number of large business bidders as N_b and specific observations n_b. The conditional distribution is then denoted $G_{\tau}(b|x, n_s, n_b)$ for $\tau \in \{s, b\}$. The following set of equations is then obtained:

$$1 = (b-c)\left(\frac{(n_s-1)\times g_s(b|x,n_s,n_b)}{(1-G_s(b|x,n_s,n_b))} + \frac{n_b \times g_b(b|x,n_s,n_b)}{(1-G_b(b|x,n_s,n_b))}\right)$$

$$1 = (b - c)\left(\frac{n_s \times g_s(b|x, n_s, n_b)}{(1 - G_s(b|x, n_s, n_b))} + \frac{(n_b - 1) \times g_b(b|x, n_s, n_b)}{(1 - G_b(b|x, n_s, n_b))}\right)$$

Solving for c:

$$c = b - \frac{1}{\left(\frac{(n_s-1)\times g_s(b|x,n_s,n_b)}{(1-G_s(b|x,n_s,n_b))} + \frac{n_b \times g_b(b|x,n_s,n_b)}{(1-G_b(b|x,n_s,n_b))}\right)} \tag{4}$$

$$c = b - \frac{1}{\left(\frac{n_s\times g_s(b|x,n_s,n_b)}{(1-G_s(b|x,n_s,n_b))} + \frac{(n_b-1)\times g_b(b|x,n_s,n_b)}{(1-G_b(b|x,n_s,n_b))}\right)} \tag{5}$$

From this we can see that, using only the observed bid distributions, we can estimate the costs associated with each bid, and thereby estimate the cost distribution. We now turn our attention to more practical matters regarding the estimation.

4.2 Estimation Strategy

The above result mandates that, to recover $F_\tau(\cdot)$, we estimate $G_\tau(b|n_s, n_b, x)$ and $G_{\bar{\tau}}(b|n_s, n_b, x)$ for $\tau \in \{s, b\}$. Under the assumption of independence between number of small bidders, number of large bidders, and the engineer's estimate (both pair-wise and jointly), we arrive at the expression:

$$G_\tau(b|n_s, n_b, x) = \frac{G_\tau(b, n_s, n_b, x)}{Pr.(N_s \leq n_s) \times Pr.(N_b \leq n_b) \times Pr.(X \leq x)} \tag{6}$$

We estimate each distribution non-parametrically using kernel density estimation. Since $G(b, n_s, n_b, x)$ is a joint cumulative distribution of both continuous and discrete random variables, we employ a kernel method developed by Li, et al. (2001) and Aitchison, et al. (1976) for estimating joint distributions of mixed discrete and continuous variables, using standard normal kernel for the continuous variables and a categorical kernel for the discrete variables. We estimate other distributions using univariate standard normal kernels. Once this conditional distribution $G_\tau(b|n_s, n_b, x)$ is estimated for both small and large businesses, we utilize equations (6) and (7) to recover pseudo-costs from each bid, and then use kernel density estimation, again with a standard normal kernel, to estimate the cost distributions F for each type of bidder. The exact equations used and details of their implementation are covered in the following section.

5 Results

Our goal is to obtain an estimate for the cost distribution of both small and large bidders. Our strategy is as follows:

We first aim to test the assumptions of our model. Namely, the assumption of pairwise independence among N_s, N_b, and X (this is achieved by comparing the distance correlations between the three variables). From there, we define the kernel method we use for KDE on the joint distribution $G_\tau(b, x, n_s, n_b)$, benchmark its main features, and select appropriate bandwidths for use in this kernel method. Next, we perform gaussian KDE on N_s, N_b, and X, followed by KDE using our custom kernel on the joint distribution. These four kernel density estimates are then combined to obtain the pseudo-costs from each bid, finding that we are able to recover costs for 2208 of the bids. KDE is performed on these pseudo-costs, and the cost distribution is obtained.

The assumption of independence is verified below:

Table 4 Distance Correlations between variable pairs,

Variable Pair	Distance Correlation
(N_s, N_b)	-0.0543
(N_s, X)	-0.2110
(X, N_b)	0.15631

Racine, et al. and Aitchison, et al. define methods for computing a kernel density estimate of both categorical and continuous variables. For two continuous (b and x) and two discrete (n_s and n_b) variables with N observations, this takes the following form:

$$\hat{f}(b, x, n_s, n_b) = \frac{1}{h_1 h_2 N} \sum_{i=1}^{N} K_1\left(\frac{b - b_i}{h_1}\right) K_2\left(\frac{x - x_i}{h_2}\right) L_{1,\lambda_1}(n_s, n_{si})\, L_{2,\lambda_2}(n_b, n_{bi})$$

Where K are standard normal pdfs and L are functions such that $L_{k,\lambda_k}(a, b) \equiv \lambda_k$ when $a = b$ and $L_{k,\lambda_k}(a, b) \equiv \frac{1-\lambda_k}{ncat-1}$ when $a \neq b$, where $ncat$ is the total number of categories that the discrete variable can take on. The λs and hs that you see above are called the bandwidths of the kernel density estimate and are a free parameter chosen by the user (we use a method called cross-validation; more on that later).

The CDF was implemented using Monte Carlo Integration. This method samples N random vectors $x_i = (b, x, n_s, n_b)$ uniformly within some region Ω and then computes:

$$\int_\Omega f(x)\, dx \approx \frac{V}{N} \sum_{i=1}^{N} f(x_i)$$

By the law of large numbers:

$$\lim_{N\to\infty} \frac{V}{N} \sum_{i=1}^{N} f(x_i) = \int_\Omega f(x)\, dx$$

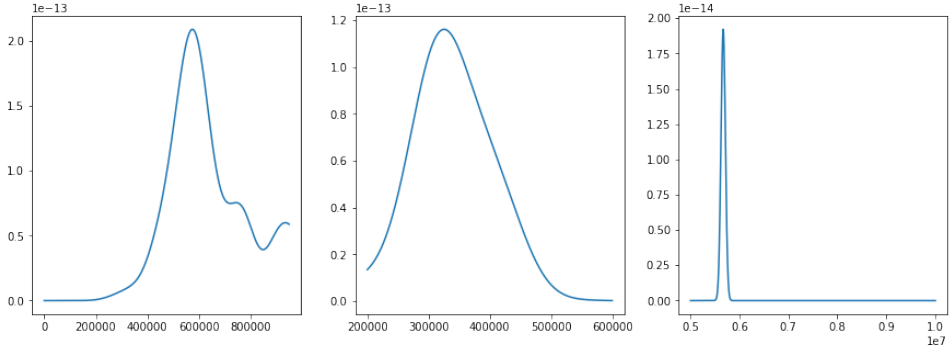

Fig. 4 The joint bid PDF $g_\tau(b, n_s, n_b, x)$ shown around the points (656000,0,4), (293000,8,6), and (6254000,1,4) for small, small, and large businesses, respectively. The x-axis is bid price.

Both this Monte Carlo method, and the kernel density estimate of the pdf are implemented in an extension of standard Python 3.9 called Cython. This extension of the language allows you to build in more assumptions about what sorts of data you're dealing with and allows code to be sped up dramatically. Benchmarking revealed that the main bottleneck of the program (the PDF function) required 141 milliseconds per execution in pure Python, whereas a single execution requires only 46.5 microseconds in the optimized code. This is a speedup of 3065 times. Tests also showed that normalization is roughly achieved with only $N = 100000$, which is quite good for a space of dimension four.

To choose the bandwidths for each variable, we employ a technique from machine learning called cross-validation. This splits the data up into k random partitions (usually called "folds" in machine learning literature), and trains the KDE on all but a single partition, and then computes the probability of the missing partition. This is repeated for each partition and the resulting probabilities are summed. This is similar to jackknife resampling in statistics.

Using our mixed kernel and bandwidth selection methods we obtained $G_\tau(b, n_s, n_b, x)$ for $\tau \in \{s, l\}$. Plots of this distribution around some in-sample points are shown below. We also estimate the PDF for the estimate X, shown in Figure 5.

Using these distribution estimates and the equation we previously derived to convert from bids to costs, we estimate the "pseudo-cost" distributions for both small and large businesses. PDFs and CDFs for both small and large businesses are shown in Figure 6.

One can see a clear difference in the cost distributions of the two types of bidder, with the larger businesses clearly showing an higher willingness to pay, all else equal. One may speculate this is result of economies of scale or of larger corporations being able to more effectively rein in fixed costs as a percentage of contribution margin.

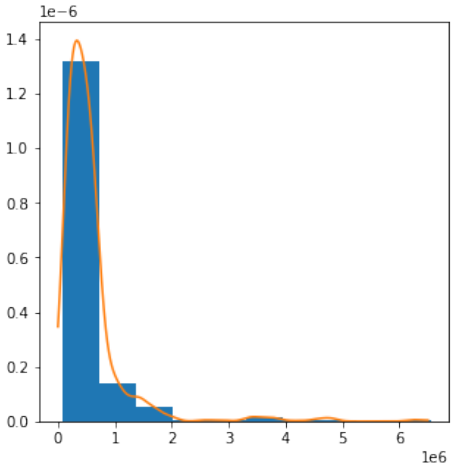

Fig. 5 Kernel Density Estimation for the engineer's estimate X

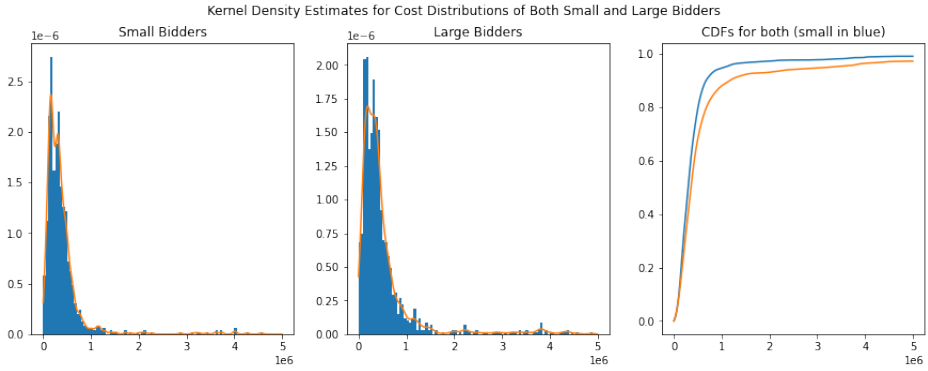

Fig. 6 PDFs and CDFs obtained via the above method. The distribution for large business bidders is shown to be clearly distinct from that of the small business bidders. Note that the x-axis has been truncated to bids between zero and five million so that the region of interest (i.e., where the distributions differ) is more clear.

6 Counterfactual Simulation

In this section, we perform a simulation using the cost distribution obtained in the previous section that aims to answer the question: what if the California Department of Transportation had implemented a different auction structure? Namely, we examine the case of a second-price sealed bid auction with a reserve price equal to the engineer's estimate. We assume that if no bids are made below the reserve price, the cost incurred by CalTrans is that of the engineer's estimate.

With this goal in mind, we first performed kernel density estimation on the joint distribution of costs and estimates. We then performed KDE on only the estimates, and sample from this function. Our samples from this distribution

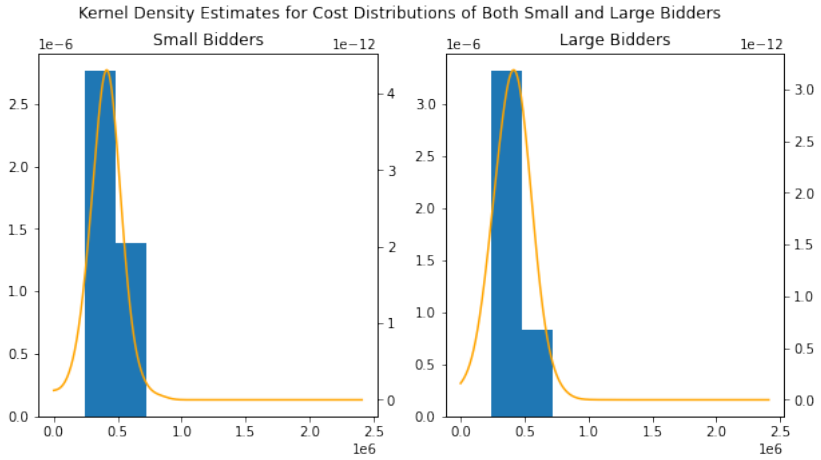

Fig. 7 PDFs and CDFs for the estimate of the joint distribution of pseudo-costs and estimates.

are then used to perform several auctions with a randomly sampled number of bidders from each category (in conformance with the distribution within the data). The second price auction winner is determined according to the small-business preference rule and the winning bid catalogued. These winning bids are used to compute an average cost to CalTrans within this auction structure and this is then compared to the average cost to CalTrans resulting from the first-price, no-reserve auction presented by the data.

Figure 7 shows the joint estimation near the median estimate ($441,000) in-sample points. Observe the distribution is, as one may have expected, heavily skew-right and peaked around the engineer's estimate.

Using this distribution estimate, in addition to the estimated PDF of the engineer's estimate and the PDFs of N_s & N_b, we performed re-sampling of the data set and applied an auction rule consistent with a second-price, sealed-bid auction with reserve price equal to the engineer's estimate.

The results of this simulation show that, had CalTrans implemented a second-price, sealed-bid auction with reserve price equal to the engineer's estimate, the would reap enormous cost savings. The average cost under the proposed method was equal to $252,110.54, whereas the expected cost found in the data (for the relevant auctions not filtered out via methods found in our data section) was equal to $542,341.99. This represents a cost savings of 53.5%, just by tweaking the auction mechanism, without detriment to the small business preference rule or the inclusion of complicated incentive structures.

7 Discussion

Over the course of this paper, we set out to answer the question of whether or not the costs of "large" businesses and "small" businesses, involved in the

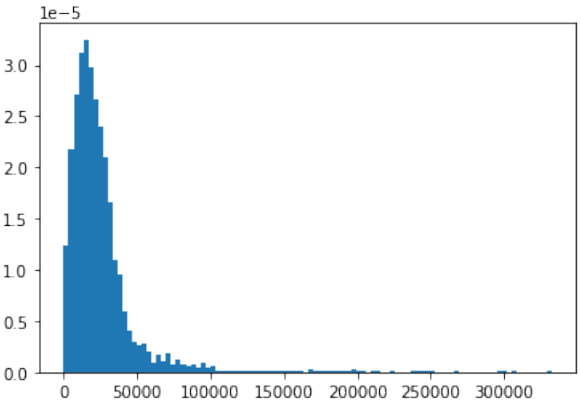

Fig. 8 Winning bids in the simulation under the proposed auction scheme. Note that the x-axis is divided by a factor of ten and represents the winning bid. One may note that the distribution is significantly less leptokurtic than the winning bids distribution under the first-price mechanism, a feature which is certainly to be desired from the perspective of CalTrans.

auctioning off of construction jobs for transportation in the state of California, were actually different. We took a sample of auction data from Caltrans, the agency that coordinated the assignment and sale of these construction jobs, and created a model that used the data and the distributions of bids from large and small bidders to estimate the distributions of costs for the two types of businesses. This model used kernel density estimation methods and monte carlo integration to recover the desired distributions. We were able to successfully recover the two distributions of interest, finding clear differences in the two distributions, namely that the CDFs show that the costs of small businesses consistently exceed those of large businesses. This confirmed our suspicions. We then analyzed an alternative auction structure under the same cost distributions and found that the inclusion of a reserve price and changing the auction to a second-price format would greatly decrease the expected cost to CalTrans for any given project.

Further work is required to confirm the practicality of such changes to the auction structure, and perhaps an investigation into more complex auction structures that seek to level the playing field between small and large businesses is required. Additionally, a more granular analysis of what factors affect business costs within these infrastructure projects could lead to a semi-parametric estimation scheme that would likely have significantly more practical application to CalTrans, or indeed anyone seeking to model procurement auctions with some bidder asymmetry in the future.

References

[1] Li, Q., & Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. journal of multivariate analysis, 86(2),

266-292.

[2] Aitchison, J., & Aitken, C. G. (1976). Multivariate binary discrimination by the kernel method. Biometrika, 63(3), 413-420.

[3] Li, Q., & Racine, J. S. (2007). Nonparametric econometrics: theory and practice. Princeton University Press.

[4] Kroese, D. P., Taimre, T., & Botev, Z. I. (2013). Handbook of Monte Carlo methods (Vol. 706). John Wiley & Sons.