Analyzing the Effects of Judicial Rotation on Criminal Sentencing: An Operations Perspective

Baris Ata
Booth School of Business, University of Chicago, Baris.Ata@chicagobooth.edu

Rhys Hester
College of Behavioral, Social and Health Sciences, Clemson University, rhysh@clemson.edu

Lawrence M. Wein
Graduate School of Business, Stanford University, lwein@stanford.edu

Yuwei Zhou
Booth School of Business, University of Chicago, Yuwei.Zhou@chicagobooth.edu

Objectives. To understand how the impact of judicial rotation and subsequent judge shopping on the defendant's sentence length is mediated by three operational characteristics: the amount of judicial rotation, the allowable shopping time window for defendants, and the capacity utilization of the judicial system.

Methods. Using data from South Carolina in 2000-2001, we formulate and calibrate a mathematical model in which judges rotate across counties, defendants shop for judges, and the sentencing (either by plea or trial) is the result of strategic interactions among the defendant, the judge and the prosecutor. We vary the three operational characteristics via simulation.

Results. The mean and standard deviation of the defendant sentence length decreases (with decreasing returns to scale) in the amount of judicial rotation and the allowable shopping window for defendants, and increases in the capacity utilization, with judicial rotation and the shopping window exhibiting synergistic behavior. The average reduction is modest ($\leq 10\%$), although a small proportion of defendants are impacted in a significant way. In a variant of the model adapted to an urban setting where all defendants have access to all judges, the mean and standard deviation of the sentence length decreases in the number of judges, even in the absence of intertemporal judge shopping.

Conclusions. Judicial rotation in a rural setting can lead to a modest reduction in the mean sentence and to more equitable sentencing. These effects can occur naturally in an urban setting.

Key words: Plea bargaining, inequality, queueing

1. Introduction

This study is motivated by the empirical findings in Hester (2017), which looks at criminal sentencing outcomes in South Carolina during 2000-2001. A distinctive aspect of South Carolina's judicial system, which consisted of 50 judges presiding over 46 counties, was judicial rotation: although judges spent much of their time in their home circuit, they traveled to an average of 12 counties (typically holding court for a week in each county) and counties encountered an average of 13 different judges throughout the year. As is the case in other U.S. jurisdictions, more than 98% of sentenced cases ended in a plea bargain and less than 2% went to trial. Results in Hester (2017) revealed that judicial rotation had two behavioral effects: it led to judge shopping, where defendants would strategically wait for a lenient judge before agreeing to a plea bargain, and to the cross-pollination of ideas and norms by increasing the interactions among judges and prosecutors. Consequently, judicial rotation led to a decrease in both the mean and standard deviation of the sentence length in South Carolina. Indeed, even though South Carolina was a non-guidelines state, the inter-county variability in sentence lengths was smaller than in most guideline states. See Hester (2017) for a fuller discussion of the contextual setting and the results.

In this study, we construct a mathematical model that attempts to capture the effects of judge shopping but does not explicitly incorporate cross-pollination among judges. We model the process in which defendants shop for a judge, the prosecutor proposes a plea deal to the defendant, where both sides are aware of the leniency of the chosen judge, and the defendant either accepts the plea deal or goes to trial. A distinctive aspect of our model is the consideration of queueing and congestion: arriving defendants can only choose judges that have slack capacity in their schedule.

There were not sufficient data to perform an econometric analysis of the interactions among the judge, prosecutor and defendant. Specifically, while we know which judge was chosen by each defendant, we do not know which judges were previously rejected by the defendant during the shopping process. Nonetheless, we are able to estimate the model parameters from the South Carolina data. By simulating the mathematical model, we compute the mean and standard deviation of the sentence length as a function of three operational characteristics: the amount of judicial rotation, the allowable shopping time window for defendants, and the capacity utilization of the judicial system (i.e., the total case workload divided by the total judicial capacity).

Because the full statewide judicial rotation scheme of South Carolina appears unique to that jurisdiction, we also adapt the model to an urban setting where defendants have the opportunity to shop for judges. This context of multiple judges within an urban jurisdiction has been the more typical application of the ideas of individual versus master calendaring, and likely has more far-reaching implications in larger jurisdictions where parties can attempt strategically to identify favorable judges. Applying our model in this setting helps to quantify the effects of shopping availability and capacity utilization in large jurisdictions.

2. Literature Review

By some accounts, the identity of the sentencing judge may matter more to a case outcome than the facts of the case or the background of the defendant. A robust line of sentencing literature focuses on the importance of the identity of the sentencing judge (e.g., Frazier and Bock (1982); Johnson (2006); Myers and Talarico (1987); Spohn (1990); Steffensmeier and Britt (2001)). Consistent with Ulmer (2019) application of Inhabited Institutions Theory to the study of courts and sentencing, we consider the broader impact that court infrastructure characteristics can have on outcomes through the mechanism of the judicial calendaring system. Inhabited Institutions Theory emphasizes how individual actors exercise discretion, reacting to and contributing institutional rules and cultures. In our context, we consider how actors are able to use judge assignment rules to effect more optimal sentencing outcomes—that is, how parties are able to use calendaring systems and local rules and norms to strategically shop for more favorable judges. Early work by Eisenstein et al. (1988) and Ulmer (1997) highlighted differences between individual calendaring systems and master calendaring systems. Under individual calendaring systems judges are assigned a case and retain control of it while in master calendaring systems different judges may handle various tasks, such as arraignment, motions, presiding over the plea or trial, and sentencing. Because the style of calendaring system can be a matter of local rules, systems may differ across counties even within the same jurisdiction (Ulmer (1997)). This early work on court communities found that in master calendaring systems, parties were sometimes able to influence court administrators or otherwise strategically engage in motions or delays in order to “judge shop.” Hester (2017: 218) found that South Carolina’s statewide master calendaring system in conjunction with the practice of judicial rotation led to “an exaggerated form of shopping” for judges. Using a mixed methods approach he found the practice of regular judge rotation (in which judges routinely traveled from county to county holding court) led to the influence of “plea judges”—lenient judges whose sentencing preferences established baseline norms or going rates for sentencing. Since judges rotated, defendants could strategically choose to enter guilty pleas when plea judges were holding court in their jurisdiction. This reality led other pragmatic judges to adopt plea judge norms for the sake of efficiency. Hester (2017)’s work was largely based on qualitative interviews with judges. We extend this prior work by formulating and calibrating a mathematical model of judicial rotation. We also adapt this model to settings involving defendants in a large urban setting, which may offer insights into how our formal model of rotation in South Carolina would generalize to other settings involving judge shopping.

The bulk of our modeling involves the dynamic judge shopping process, which is influenced by the recent models in Yang (2016), Silveira (2017) and Wang (2019). These are part of a much larger literature that analyzes models of the plea versus trial game between a defendant and a prosecutor (e.g., Gross and Syverud (1991) and references therein). Among the three operational characteristics we study, only capacity constraints have been modeled in the judicial context (e.g., Ostrom et al. (1999)).

3. Model

Our simulation model of the judicial process involves three agents for each case – the judge, the prosecutor and the defendant – and results in either a plea bargain or a trial. The output of the simulation model includes three primary performance measures that are defined in §3.4: the mean and the standard deviation of the sentence length of all plea bargains (the mean sentence at trials is assumed to be the same for each judge, and hence is omitted) and the standard deviation of the mean sentence length across counties. The goal of our analysis is to understand how these performance measures are impacted by three key operational characteristics: the amount of judicial rotation, the defendant shopping window, and the aggregate judge utilization.

Our simulation model consists of three modules that are described in §3.1-3.3: the defendant arrivals, the judge schedule and the plea bargain, with each module containing one of the three key operational characteristics. The defendant arrivals module describes the timing and characteristics of defendants that arrive to the system, where the arrival rate depends upon the specified judge utilization. The judge schedule module assigns judges to counties each week based on the specified amount of judicial rotation, and computes the number of pleas that they can process. The plea bargain module includes the interactions among the three agents, and the final sentence imposed on a defendant is determined by the judge he chooses (which in turn depends on the defendant shopping window), the plea offer recommended or presented by the prosecutor and approved (or specified, if a straight plea) by the judge, and whether the defendant accepts the plea offer (i.e., an agreement is reached) or refuses the plea offer and goes to trial. Figure 1 shows a high-level description of our model structure. The estimation of the parameters for each module is described in §4. For ease of reference, all model parameters are described in Table 1.

3.1. Defendant Arrivals

Defendants in our model are indexed by $i = 1, 2, \dots$ and each arriving defendant is assigned a set of covariate values $\boldsymbol{x}_i$, which for ease of presentation is suppressed and embedded in the subscript i. The number of defendants who arrive to county c during a week is modeled by a Poisson random variable N_c with mean λ_c. The value of λ_c is dictated by the key operational characteristic ρ, which is the desired judge utilization (i.e., the overall proportion of available time that judges spend presiding over cases) in our simulation model. In §4.3, we compute λ_c, see Equation (6), and describe the assignment of the covariate values $\boldsymbol{x}_i$.

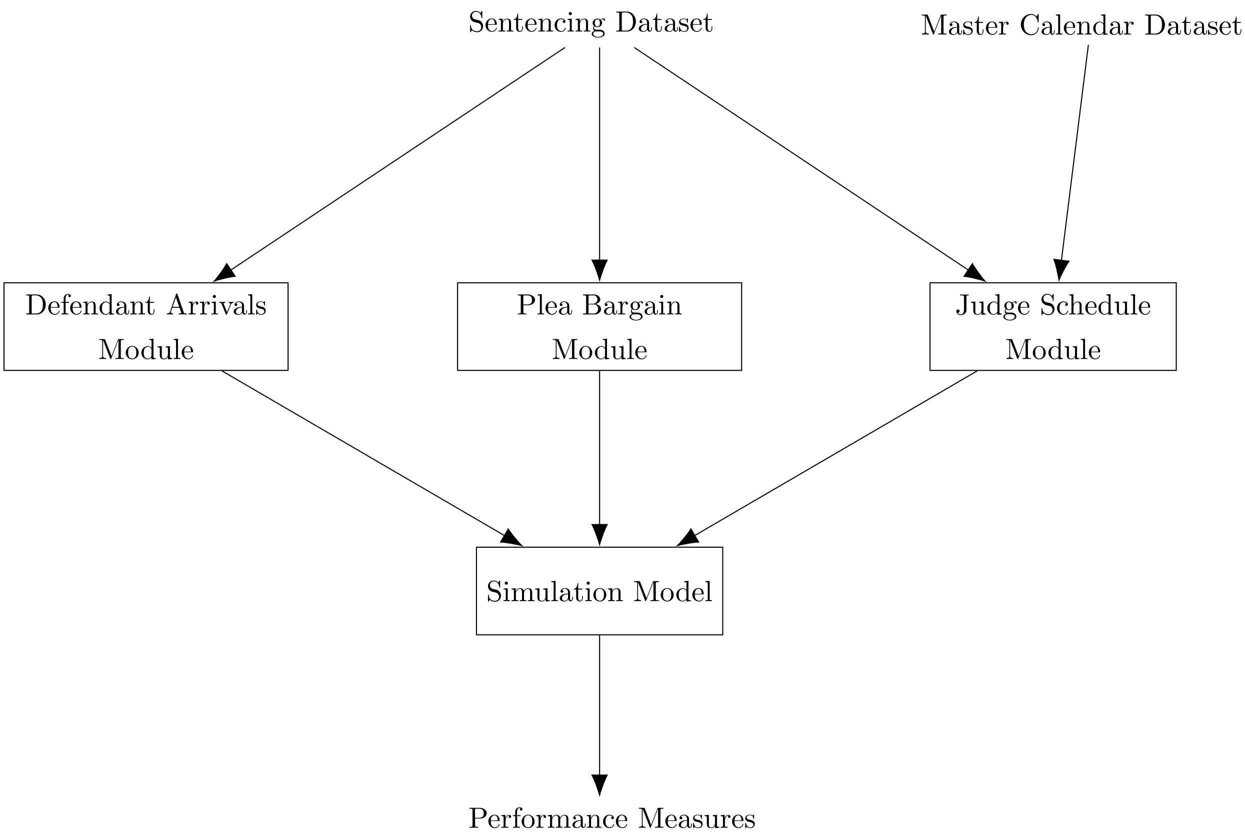


Figure 1 **The simulation model contains three modules: the defendant arrivals (§3.1), the judge schedule (§3.2) and the plea bargain (§3.3). The performance measures are defined in §3.4, and the two datasets are described in §4.1.**

Table 1 **The model parameters**

Parameter	Description	Module	Value
ρ	Judge utilization	Defendant arrivals	Specified
λ_c	Defendant arrival rate		(6)
$\boldsymbol{x}_i$	Defendant covariates		§4.3, Table 2
η	Judge travel probability	Judge schedule	Specified
T_{jct}	Available hours for judge j to work in county c in week t		Master calendar
n^T_{jct}	Number of trials in county c assigned to judge j in week t		Master calendar
γ_P	Mean processing time for a plea		0.018 weeks
γ_T	Mean processing time for a trial		0.839 weeks
r	Defendant shopping window	Plea bargain	Specified
p_i	Probability defendant i receives a zero-length plea sentence		(7), Table 10
θ_i	Probability of conviction at trial for defendant i		(8), Table 11
τ_i	Expected sentence if convicted at trial for defendant i		(9), Table 12
$\ell_j(\cdot)$	Lower bound on the approved sentence length from judge j		§4.4, Fig. 18
$u_j(\cdot)$	Upper bound on the approved sentence length from judge j		§4.4, Fig. 18
$c_d(i)$	Trial cost for defendant i (in months)		§4.4, Fig. 20
d	Defendant waiting cost per week (in months per week)		0.1

3.2. Judge Schedule

We index judges by $j=1,\dots,J$ and counties by $c=1,\dots,C$. The judge schedule module specifies the location (county or counties) of judge j in each week t, along with the number of plea cases that judge j can process in week t. Each judge is assigned one (and occasionally more than one) home county, and the weekly judge locations are driven by the key operational characteristic η, which is the probability that a judge is traveling (i.e., working at a non-home county) in any given week. The weekly judge locations in the module output are denoted by C_{jct}, which is the proportion of judge j's time in week t that is assigned to county c.

For $j=1,\dots,J$ and $t=1,\dots,52$ in the master calendar dataset, let T_{jct} be the proportion of week t that judge j is available to process pleas and trials in county c, and let n_{jct}^T be the number of trials in county c assigned to judge j in week t. Let γ_P and γ_T be the mean processing time (independent of judge) for a plea and a trial, respectively. Then the number of pleas that judge j can process in week t in county c is denoted by N_{jct} and modeled as a Poisson random variable with mean $(T_{jct}-n_{jct}^T\gamma_T)^+/\gamma_P$.

The parameters T_{jct}, n_{jct}^T, γ_P and γ_T are estimated in §4.2, and η is the specified operational characteristic. The outputs of the judge schedule module are C_{jct} and N_{jct}.

3.3. Plea Bargain Process

The key operational characteristic in the plea bargain module is the defendant shopping window r, which is an integer number of weeks. An arriving defendant chooses the available judge within his shopping window that minimizes his total cost. In our model, defendant i receives a zero-length sentence (e.g., probation or supervision) with probability p_i, which is a function of the defendant's covariates but (see §4.4 for a justification) independent of the presiding judge.

We now describe the plea bargain process under the assumption that defendant i does not receive a zero-length sentence. Defendants may reject a plea deal and decide to go to trial. Depending on a defendant's covariates $\boldsymbol{x}_i$, if he goes to trial then he is convicted with probability θ_i, in which case he receives the expected sentence length τ_i. In addition, if the defendant goes to trial, which is a longer judicial process, he incurs the additional cost $c_d(i)$, which is in time units. Hence, the expected total cost of resolving defendant i's case via trial is $\theta_i\tau_i + c_d(i)$, which is observable by all three agents.

Crucially, judges vary in their leniency, which is modeled using two leniency functions $\ell_j(\cdot)$ and $u_j(\cdot)$ for each judge j. These functions define a lower and upper bound, respectively, on the sentence length as a function of $\theta_i\tau_i$, which is the unconditional mean sentence length if defendant i goes to trial. More specifically, judge j only approves plea deals with a sentence length in the range $[\ell_j(\theta_i\tau_i), u_j(\theta_i\tau_i)]$, where $\ell_j(\theta_i\tau_i) \leq u_j(\theta_i\tau_i)$ for all j and for all values of $\theta_i\tau_i \geq 0$.

We assume that prosecutors try to avoid trials by resolving cases through plea bargaining. However, in doing so, prosecutors aim to maximize the sentence length, leading to three possible outcomes:

- If $\theta_i\tau_i + c_d(i) > u_j(\theta_i\tau_i)$, then the prosecutor offers $u_j(\theta_i\tau_i)$, which the defendant accepts, and the judge approves the plea.
- If $\theta_i\tau_i + c_d(i) < \ell_j(\theta_i\tau_i)$, then there are no plea offers that both the defendant would accept and the judge would approve. The case goes to trial and the defendant's expected cost is $\theta_i\tau_i + c_d(i)$.
- If $\ell_j(\theta_i\tau_i) \leq \theta_i\tau_i + c_d(i) \leq u_j(\theta_i\tau_i)$, then the prosecutor offers $\theta_i\tau_i + c_d(i)$, which the defendant accepts and the judge approves.

To summarize, $\min\{\theta_i\tau_i + c_d(i), u_j(\theta_i\tau_i)\}$ is the expected sentence that defendant i receives either through a plea bargain or at trial, provided that he is not offered a zero-length sentence during the plea bargaining process, which has probability p_i. To pick judge j, the defendant needs to delay going to court until judge j works in the defendant's county and has sufficient capacity to hear the case. Let d denote the defendant's cost of waiting per week and $w_i(j)$ denote the number of weeks that he needs to wait for judge j. Recall that defendants are given the window of r weeks to shop for judges. Letting $J_i(r)$ denote the set of available judges, i.e., judges who have assigned sessions with remaining capacity within r weeks upon defendant i's arrival, the defendant chooses judge j^* such that

$$j^* \in \arg\min_{j \in J_i(r)} \left[(1 - p_i)\min\{\theta_i\tau_i + c_d(i), u_j(\theta_i\tau_i)\} + w_i(j)d\right]. \tag{1}$$

The parameters p_i, θ_i, τ_i, $c_d(i)$, $\ell_j(x)$, $u_j(x)$ and d are estimated in §4.4, whereas the shopping window r is a key operational characteristic that we specify. The parameter $w_i(j)$ and the set $J_i(r)$ are dictated by the weekly judge-location assignments C_{jct} and the weekly judge capacities N_{jct}, both of which are outputs of the judge schedule module. That is, we make judge j unavailable in week t after she has been assigned N_{jct} plea cases, and update the $J_i(r)$ sets accordingly. The output of the plea bargain module is the expected (nonzero) sentence for each defendant i, $\min\{\theta_i\tau_i + c_d(i), u_{j^*}(\theta_i\tau_i)\}$.

3.4. Performance Measures

Because the probability of a zero-length sentence is independent of the judge in our model, our three performance measures include only cases with nonzero sentences. Let I_c be the set of simulated defendants from county c whose sentence length is positive, and let $I = \cup_{c=1}^{C} I_c$. Let $s_i = \min\{\theta_i\tau_i + c_d(i), u_{j^*}(\theta_i\tau_i)\}$ be defendant i's sentence length in the model output. The mean and standard deviation of the plea sentence length are given by

$$\mu = \frac{\sum_{i\in I} s_i}{|I|},$$

$$\sigma = \sqrt{\frac{\sum_{i\in I}(s_i - \mu)^2}{|I|}}.$$

Our final performance measure is the standard deviation across counties of the mean plea sentence length:

$$\sigma_{\mathrm{c}} = \sqrt{\frac{\sum_{c=1}^{C}(\mu_c - \mu)^2}{|C|}},$$

where

$$\mu_{\mathrm{c}} = \frac{\sum_{i\in I_c} s_i}{|I_c|}.$$

4. Parameter Estimation

In this section, we estimate the parameters in our model. We describe the data in §4.1, and estimate the parameters of the judge schedule module, the defendant arrivals module and the plea bargain module, respectively, in §4.2–4.4.

4.1. Data

We use two datasets from the South Carolina circuit court (i.e., the court of general jurisdiction) between July 1, 2000 and June 30, 2001: the master calendar and the sentencing dataset. The master calendar describes the weekly schedule (i.e., location among the $C = 46$ counties) of the $J = 50$ judges between July 2000 and June 2001. South Carolina trial judges preside over both criminal and civil matters with terms of court typically lasting one week and consisting of only criminal matters (general session) or civil matters (common pleas) for the given term. We restrict our attention in the master calendar to judge-weeks in which the judge is in general session for at least a portion of the week (judges can also, e.g., be on vacation, out sick, in circuit court or involved with the state grand jury), which comprises 854.4 (33.1%) of the 2582 judge-weeks.

The sentencing dataset contains offenders that are convicted of a felony or serious misdemeanor. There are 17,516 sentencing events in the dataset, and 246 (1.4%) of these events were discarded because we could not impute the correct court dates. Of the remaining 17,270 sentencing events (17,012 pleas and 258 trials), 14,977 (14,748 pleas, 229 trials) involve cases where the judge was in general session. Table 2 shows the covariates that we have for each sentencing event, which describe characteristics of the offense and the defendant.

4.2. Judge Schedule Parameters

In this subsection, we compute the weekly judge schedules C_{jct}, the available times in general session T_{jct} and the trial allocations n^T_{jct}, and the mean processing times γ_P and γ_T.

The starting point for our determination of the weekly judge schedules is a mathematical program that assigns judges to counties so as to cover the plea cases from each county while minimizing the degree of judge traveling. In our assignment problem, we focus on satisfying the sentencing cases that were resolved through plea bargains (i.e., excluding those cases that were resolved through trials) in the sentencing dataset. There are two reasons to exclude trials: (i) the primary goal of this work is to understand the impact of the key operational characteristics on plea bargains, which represent 98.5% of the cases in our sentencing dataset, and (ii) trials have a more complex

Table 2 Covariates for each sentencing event

Variable	Description	Values
Offense seriousness	Five-level ordinal score; From the South Carolina crime classification scheme	1 = Misdemeanor 2 = Felony (Class F) 3 = Felony (Class E) 4 = Felony (Class D) 5 = Felony (Class A, B, C or Unclassified)
Commitment score	12-level ordinal measure; Number of commitment offenses	1 = Least serious 12 = Most serious
Offense type	Four-category indicator of the classification of crime committed	1 = Violent 2 = Drug 3 = Property 4 = Other
Mandatory minimum	Minimum prison sentence	1 = Yes 0 = No
Criminal history	Five-level ordinal score; Derived from the number and severity of prior offenses	1 = None 2 = Minimal 3 = Moderate 4 = Considerable 5 = Extensive
Black	Race	1 = Yes 0 = No
Male	Sex	1 = Female 0 = Male
Black $\times$ offense seriousness	Interaction term	$(1,0)\times(1,2,3,4,5)$
Black $\times$ crime history	Interaction term	$(1,0)\times(1,2,3,4,5)$

and longer timeline than pleas and the datasets do not gives us enough information on the trial schedules.

Let κ_j be the number of weeks in the year (July 2020 - June 2021) that judge j had a general session assignment in the master calendar (see Table 5 in §A.1 for values), and let d_c equal γ_P (which is estimated below) times the number of pleas processed in general session by county c in the year in the master calendar (see Table 6 in §A.1 for values). Our decision variable is the fraction of judge j's capacity that is assigned to county c, x_{jc}, and the quadratic function $x_{jc}(1-x_{jc})$ in (2) forces the optimal x_{jc} values towards 0 or 1, thereby discouraging judge travel. This yields the optimization problem

$$\min \quad \sum_{c=1}^{C}\sum_{j=1}^{J}\kappa_j x_{jc}(1-x_{jc}) \tag{2}$$

$$\text{s.t.} \quad \sum_{j=1}^{J}\kappa_j x_{jc} \geq d_c, \quad \forall c=1,\ldots,C, \tag{3}$$

$$\sum_{c=1}^{C} x_{jc} = 1, \quad \forall j = 1, \ldots, J, \tag{4}$$

$$0 \le x_{jc} \le 1, \quad \forall j = 1, \ldots, J, \quad c = 1, \ldots, C. \tag{5}$$

The solution x^*_{jc} to this optimization problem is given in Table 7 in §A.1. If $x^*_{jc} > 0$, then we say that county c is one of judge j's home counties. Because the objective is to minimize the degree of judge traveling, the solution we obtain assigns only six of 50 judges to more than one home county. In §A.1, we provide an algorithm that maps x^*_{jc} into C_{jct}, where $\sum_{c=1}^{C} C_{jct} = 1$ for all j and t.

To obtain accurate estimates of the mean processing times γ_P and γ_T, we restrict our attention to sentencing events in the general session. Recall T_{jct} be the proportion of week t that judge j is in general session in county c, which can be recovered from the master calendar (this differs from C_{jct}, which is constructed by the algorithm in §A.1) and n^T_{jct} denote the number of trials that judge j handled in county c in week $t = 1, \ldots, 52$ in the master calendar. It follows that κ_j in Table 5 in §A.1 satisfies $\kappa_j = \sum_{c=1}^{C} \sum_{t=1}^{52} T_{jct}$.

We use linear regression to estimate the mean processing times, γ_P and γ_T. In an analagous manner to above, we define n^P_{jct} to be the number of pleas that judge j handled in county c in week $t = 1, \ldots, 52$ in the master calendar. Let $n^T_{jc} = \sum_t n^T_{jct}$ and $n^P_{jc} = \sum_t n^P_{jct}$ be the total number of trials and pleas that judge j handled throughout the year in county c. The regression model, which uses the total number of weeks that judge j worked in county c in the master calendar as the dependent variable, is

$$\sum_{t=1}^{52} T_{jct} = \alpha_c + \gamma_P n^P_{jc} + \gamma_T n^T_{jc} + \epsilon_{jc}, \quad \text{for} \;\; j = 1, \ldots, J, \; c = 1, \ldots, C,$$

where α_c is interpreted as the average idle time (in weeks) in each county. The regression results (Table 9 in the Appendix) imply that it takes on average 0.018 weeks to process a plea (i.e., a judge can process 1/0.018=56.2 pleas per week) and 0.839 weeks to process a trial. With T_{jct}, n^T_{jct}, γ_P and γ_T in hand, we can generate the random variable N_{jct} in §3.2.

As an aside, we note that the total number of judge-weeks in general session during the year is $\sum_{j=1}^{J} \sum_{c=1}^{C} \sum_{t=1}^{52} T_{jct} = 854.4$ weeks. In addition, we know from §4.1 that there are 14,748 pleas and 229 trials in general session in the sentencing dataset. Hence, we estimate the judge utilization in the data to be

$$\frac{0.839(229)+0.018(14,748)}{854.4}=0.536.$$

4.3. Defendant Arrival Parameters

In this subsection, we compute the weekly arrival rate for each county, λ_c, and determine the covariate values for each arrival, $\boldsymbol{x}_i$.

The quantity $\sum_{j=1}^{J}\kappa_j x_{jc}^*$, where x_{jc}^* is the solution to the mathematical program in (2)-(5), is interpreted as the total number of judge-weeks in a year devoted to county c. Recall that γ_P is the mean processing time (in weeks) for a plea, which is estimated in §4.2. Given a desired value for the utilization ρ, we compute the defendant weekly arrival rate to county c by

$$\lambda_c=\frac{\rho\sum_{j=1}^{J}x_{jc}^*\kappa_j}{52\gamma_P}, \tag{6}$$

where κ_j appears in Table 6 in §A.1 and x_{jc}^* is given in Table 7 in §A.1.

Let D_c denote the set of all defendants from county c in the sentencing dataset and recall that N_c is a Poisson random variable with mean λ_c. Then each week we randomly draw (with replacement) N_c defendants from the set D_c and – following Hester and Hartman (2017) – use the values for the covariates $\boldsymbol{x}_i$ in Table 2.

4.4. Plea Bargain Parameters

In this subsection, we estimate the defendant probability of incarceration during a plea bargain (p_i), the expected sentence length at trial ($\theta_i\tau_i$), the judge leniency functions ($l_j(\cdot),u_j(\cdot)$), the defendant cost of going to trial ($c_d(i)$) and the defendant delay cost (d).

Defendant Probability of Incarceration During a Plea Bargain. To estimate the probability p_i, we use the covariates $\boldsymbol{x}_i$ in Table 2 and define the binary variable y_i to equal 1 if defendant i is incarcerated in the sentencing dataset, and to equal 0 otherwise. We fit a logistic regression model using the 17,012 sentencing cases that were resolved by plea bargaining,

$$p_i=\Pr(y_i=0|\boldsymbol{x}_i)=\frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}, \tag{7}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients for the variables in $\boldsymbol{x}_i$. The estimated coefficients are provided in Table 10 in the Appendix. An alternative regression model incorporated the judge identity as an additional covariate, but the results were very similar and we used the sparser model.

Defendant Expected Sentence Length at Trial. Because θ_i and τ_i appear in our plea bargain module only as the product $\theta_i\tau_i$, it suffices to estimate the product. To predict defendant i's expected sentence length at trial $\theta_i\tau_i$, we use the hurdle regression model in Hester and Hartman (2017) and fit the model using the 258 cases that were resolved by trial. The hurdle model uses a logistic regression to predict observations that are zero, and a zero-truncated count model (negative binominal) to predict the remaining nonzero cases:

$$\Pr(\tau_i = 0|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{w})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{w})} = 1-\theta_i, \tag{8}$$

$$\Pr(\tau_i|\tau_i > 0, \boldsymbol{x}_i) = \theta_i \left[\frac{\frac{\Gamma(\tau_i+\alpha^{-1})}{\tau_i!\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{\tau_i}}{1-(1-\alpha\mu)^{-1/\alpha}} \right] \text{ for } \tau_i > 0, \tag{9}$$

where $\boldsymbol{w}$ is the set of regression coefficients for the variables in $\boldsymbol{x}_i$ (Table 11 in the Appendix), $\Gamma(\cdot)$ is the gamma function, α is a dispersion parameter, and μ is the mean of the negative binomial model (i.e., $\mu = \exp(\boldsymbol{x}_i^T\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the set of regression coefficients for the negative binomial regression model; see Table 12 in the Appendix).

For defendant i, we predict his probability of conviction at trial, $\hat{\theta}_i$, using (8) and the coefficients in Table 11 in the Appendix, and predict his sentence length upon conviction at trial, $\hat{\tau}_i$, using (9) and the coefficients in Table 12 in the Appendix. The expected trial sentence length of defendant i is estimated to be $\hat{\theta}_i\hat{\tau}_i$.

Judge Leniency Functions. To estimate the judge leniency thresholds $\ell_j(\cdot)$ and $u_j(\cdot)$ for each judge as functions of the expected sentence length at trial, $\theta_i\tau_i$, we restrict attention in the sentencing dataset to those defendants who accepted a plea bargain and were incarcerated. In the sentencing dataset, let s_i be defendant i's sentence and $\mathcal{I}_j$ be the set of plea bargains handled by judge j for $j = 1, \ldots, J$.

To estimate the leniency functions for a particular judge j, we create a scatter plot of $(\hat{\theta}_i\hat{\tau}_i, s_i)$ for $i \in \mathcal{I}_j$. The key idea is to construct a convex hull for judge j generated by the origin $(0,0)$ and the points $(\hat{\theta}_i\hat{\tau}_i, s_i)$ for $i \in \mathcal{I}_j$, and use the upper and lower boundary of the convex hull (given by the two red dots in Figure 2) to estimate the maximum and the minimum sentence length that judge j would approve for defendant i with a mean sentence length at trial equal to $\hat{\theta}_i\hat{\tau}_i$ (i.e., values of $u_j(\theta_i\tau_i)$ and $\ell_j(\theta_i\tau_i)$).

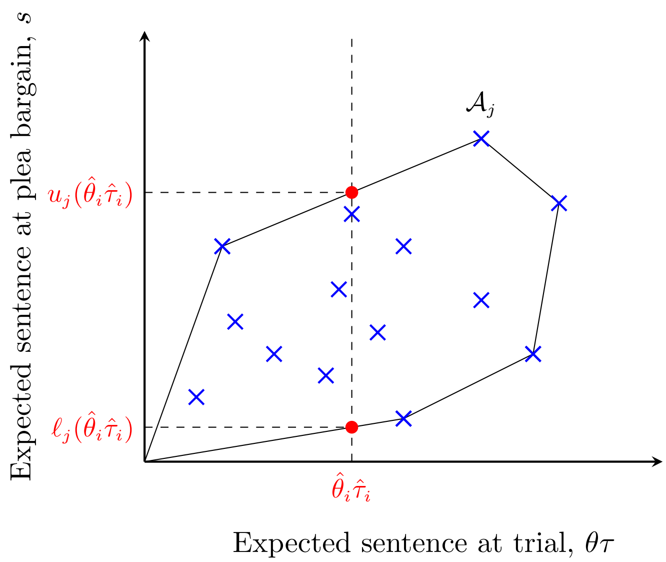


Figure 2 **A scatter plot of the sentencing handled by judge j and the constructed convex hull $\mathcal{A}_j$. This figure is for illustration purposes and is not based on the data.**

We modify this approach in two ways that are described in §A.2. First we detect and remove outliers from the observations $\{(\hat{\theta}_i\hat{\tau}_i, s_i) | s_i > 0, i \in \mathcal{I}_j\}$ using the Mahalanobis Distance, which results in 3.44% of observations being identified as outliers, with a range from 0% to 5.77% for each judge. Second, so that we can estimate the judge leniency thresholds for defendants whose expected sentence length at trial is greater than $\max_{i \in I_j} \hat{\theta}_i\hat{\tau}_i$, we extrapolate the convex hull by including two artificial points that are determined based on the maximum possible expected trial sentence obtained from the sentencing dataset. After these two modifications, the resulting convex hulls for all judges appear in Fig. 18 in the Appendix.

To provide a sense of how these leniency functions impact the sentences in our model, we use each judge's convex hull to compute the mean sentence for all defendants in the sentencing dataset that received a nonzero sentence (Fig. 19 in the Appendix). The mean overall sentence length is 51.6 months, and the mean for individual judges ranges from 44.6 (Judge 36) to 62.0 (Judge 15).

Defendant Trial Cost. To estimate the cost of going to trial, $c_d(i)$, we let $j(i)$ denote the presiding judge for defendant i's case in the sentencing dataset. We have four cases to consider for defendant i: (1) he received sentence $s_i > 0$ in a plea bargain that satisfies $\ell_{j(i)}(\hat{\theta}_i\hat{\tau}_i) < s_i < u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)$, (2) he received a sentence $s_i = u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)$ in a plea bargain, (3) he received his sentence through a plea bargain and was not incarcerated, i.e., $s_i = 0$, and (4) his case went to trial.

We partition all cases according to these four outcomes by defining the sets

$$\mathcal{I}^1 = \cup_{j=1}^J \{i \in \mathcal{I}_j | s_i > 0, \text{ and } \ell_{j(i)}(\hat{\theta}_i\hat{\tau}_i) < s_i < u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)\},$$

$$\mathcal{I}^2 = \cup_{j=1}^J \{i \in \mathcal{I}_j | s_i > 0, \text{ and } s_i = u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)\},$$

$$\mathcal{I}^3 = \cup_{j=1}^J \{i \in \mathcal{I}_j | s_i = 0\},$$

$$\mathcal{I}^4 = \cup_{j=1}^J (\mathcal{I}_j \setminus \{\mathcal{I}_j^1 \cup \mathcal{I}_j^2 \cup \mathcal{I}^3\}).$$

We estimate the trial cost $c_d(i)$ when i belongs to $\mathcal{I}^1, \mathcal{I}^2, \mathcal{I}^3$, and $\mathcal{I}^4$ separately. When $i \in \mathcal{I}^1$, our plea bargain model implies that

$$s_i = \min\left\{\hat{\theta}_i\hat{\tau}_i + c_d(i), u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)\right\}, \tag{10}$$

$$= \hat{\theta}_i\hat{\tau}_i + c_d(i) \quad \text{because} \quad s_i < u_{j(i)}(\hat{\theta}_i\hat{\tau}_i). \tag{11}$$

Thus, we can infer defendant i's cost of trial as $c_d(i) = s_i - \hat{\theta}_i\hat{\tau}_i$ in this case.

For the other three cases $l = 2, 3, 4$, we define a set S^l of similar defendants, and use a $k-$nearest neighbor algorithm with $k = 20$ (see §A.3 for a discussion of how we set k) to estimate the trial cost for defendants in groups $\mathcal{I}^2$, $\mathcal{I}^3$ and $\mathcal{I}^4$. More specifically, for $l = 2, 3, 4$, if $|S^l| < 20$ the estimated trial cost is the average value of $c_d(i')$ where $i' \in S^l$. If $|S^l| \geq 20$, then we select 20 observations that are the nearest neighbors to observation $(\hat{\theta}_i\hat{\tau}_i, s_i)$ using the Mahalanobis distance from $\left\{(\hat{\theta}_i\hat{\tau}_i, s_i)\right\} \cup S^l$.

We let $S^{l'}$ denote the set of those selected observations and estimate the trial cost by the average value of $c_d(i')$ where $i' \in S^{l'}$.

It remains to compute the sets S^2, S^3 and S^4. When $i \in \mathcal{I}^2$, we can only infer that $c_d(i) \geq u_{j(i)}(\hat{\theta}_i\hat{\tau}_i) - \hat{\theta}_i\hat{\tau}_i$. Because set $\mathcal{I}^1$ contains most of the defendants with a positive sentence and we can infer their trial costs, we use those estimates to infer the trial costs of the defendants in set $\mathcal{I}^2$. Hence we let

$$S^2 = \left\{ i' \in \mathcal{I}^1 : c_d(i') \geq u_{j(i)}(\hat{\theta}_i\hat{\tau}_i) - \hat{\theta}_i\hat{\tau}_i \right\}.$$

When $i \in \mathcal{I}^3$, we cannot infer any bounds on $c_d(i)$, and therefore let $S^3 = \mathcal{I}^1$. When $i \in \mathcal{I}^4$, we infer that $c_d(i) < \ell_{j(i)}(\hat{\theta}_i\hat{\tau}_i) - \hat{\theta}_i\hat{\tau}_i$, and let

$$S^4 = \left\{ i' \in \mathcal{I}^1 : c_d(i') \leq \ell_{j(i)}(\hat{\theta}_i\hat{\tau}_i) - \hat{\theta}_i\hat{\tau}_i \right\}.$$

The histograms of $c_d(i)$ for $i \in \mathcal{I}^i$, $i = 1, \ldots, 4$ appear in Fig. 20 in the Appendix, where the cost of going to trial is estimated to be negative for many defendants. We interpret this to mean that many defendants should prefer to go to trial rather than to plea bargain.

Defendant Waiting Cost. To estimate d, we simulate ten 50-year replications of the system, and collect performance measures for the middle 30 years of each replication. Because we use the actual judge schedule in the master calendar, the operational characteristics η and ρ do not apply. We set the shopping window to $r = 4$ weeks, which is the base-case value in §5. We use the 14,977 sentencing events in general session to estimate the weekly arrival rate and allow it to vary across months to more accurately mimic the seasonality in the data, and randomize the timing and characteristics of the defendant arrivals as described earlier. Because τ_i and $c_d(i)$ are measured in months whereas $w_i(j)$ is measured in weeks in (1), the cost d is in units of months per week. Consequently, with an assumption of four weeks per month, $d = 0.25$ means that the cost of delaying a court case for one week is equal to one week of detention. We consider $\{0, 0.1, 0.25\}$ as possible values for d. The three performance measures are quite insensitive to d (Table 3) and we use $d = 0.1$.

Table 3 Statistics from the sentencing dataset and simulation results with weekly waiting cost $d \in \{0, 0.1, 0.25\}$ and a defendant shopping window $r = 4$

	Data	$d = 0$	$d = 0.1$	$d = 0.25$
Mean plea sentence (for nonzero sentences)	54.27	50.21	50.29	50.29
Standard deviation of plea sentence	55.18	52.36	52.34	52.34
Percentage of trial cases	1.51%	1.94%	1.99%	1.99%
Standard deviation across counties (for nonzero sentences)	11.74	12.52	12.50	12.50

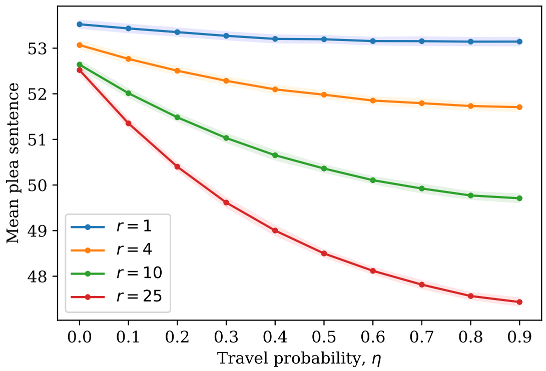
5. Results

We investigate the impact of the key operational characteristics in §5.1, and in §5.2 consider a hypothetical urban model in which all defendants and judges are located in one place (i.e., all defendants have access to all judges). All simulation results consider 10 independent replications of 50 years, with the first and last 10 years of each replication discarded.

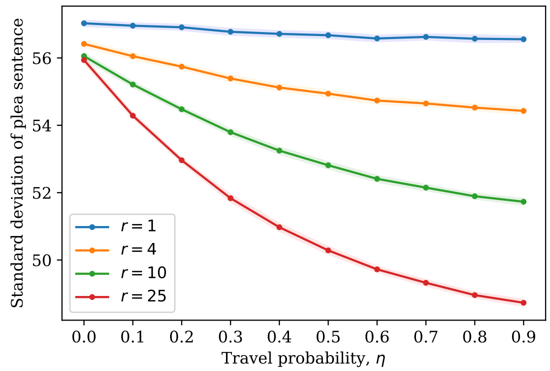
5.1. Impact of the Key Operational Characteristics

In this subsection, we investigate the impact of the judge travel probability η, the defendant shopping window r and the judge utilization ρ on the three performance measures. We consider base values of $\eta = 0.5$, $r = 4$ weeks (although South Carolina published annual schedules for judges, these schedules were updated frequently) and $\rho = 0.5$. Figs. 3-8 show the performance measures (the shaded areas surrounding the curves are 95% confidence intervals) plotted against one of the operational characteristics (with values $\eta \in \{0, 0.1, 0.2, \ldots, 0.9\}$, $r \in \{1, 2, 3, 4, 5, 6, 8, 10, 14, 18, 22, 25\}$ weeks and $\rho \in \{0.5, 0.7, 0.9, 0.95, 0.99\}$), with a second operational characteristic set at its base value, and several different curves for changes in the third operational characteristic. In addition, three-dimensional plots of the performances measures versus two operational characteristics with the third set at its base value appear in Figs. 21-23 in the Appendix.

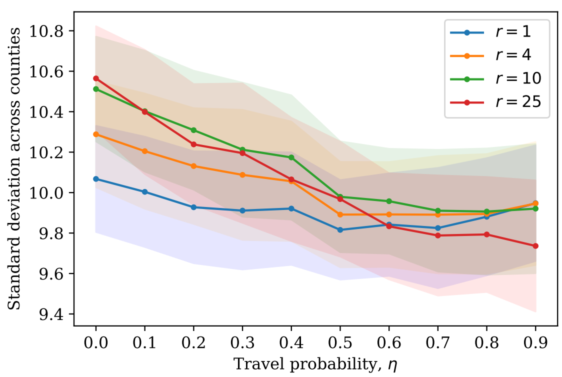
We defer a discussion of the county variation (Figs. 3c-8c) and focus on the impact on the mean and standard deviation of the plea sentences. As the travel probability η increases (Figs. 3-4), the mean and standard deviation of the plea sentences decreases with diminishing impact (i.e., the curves are convex), and the impact is larger when the shopping window r is large or the judge utilization ρ is small. Judges travel more as η increases, increasing defendants' shopping opportunities and therefore their likelihood of securing a more lenient judge. A larger shopping window gives defendants a longer period to choose their preferred judges. However, when judge utilization is high, defendants cannot always choose a more lenient judge because the lenient judges may be overwhelmed with other defendants' pleas.



(a) Mean plea sentence



(b) Standard deviation of plea sentence

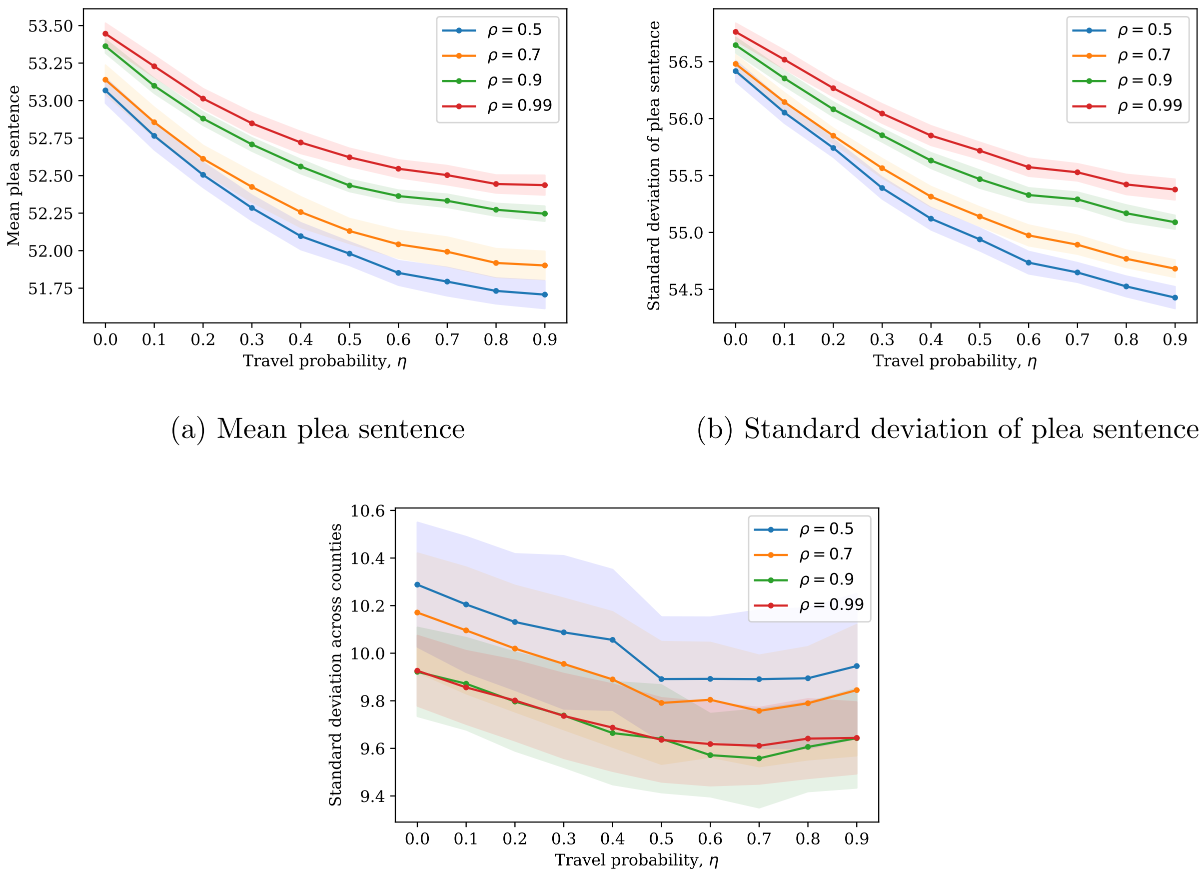


(c) Standard deviation across counties

Figure 3 **Performance measures versus the travel probability** $\eta \in \{0, 0.1, \ldots, 0.9\}$**, where the defendant shopping window is** $r \in \{1, 4, 10, 25\}$ **weeks and judge utilization is fixed at** $\rho = 0.5$**.**

Similarly, as the shopping window r increases (Figs. 5-6), the mean and standard deviation of the plea sentence decrease with decreasing returns, with the impact being larger when the travel probability η is large or the judge utilization ρ is small.

In contrast, when the judge utilization ρ increases (Figs. 7-8), the mean and standard deviation of the plea sentence increase, and are slightly convex (i.e., exhibit increasing returns), with the



(a) Mean plea sentence

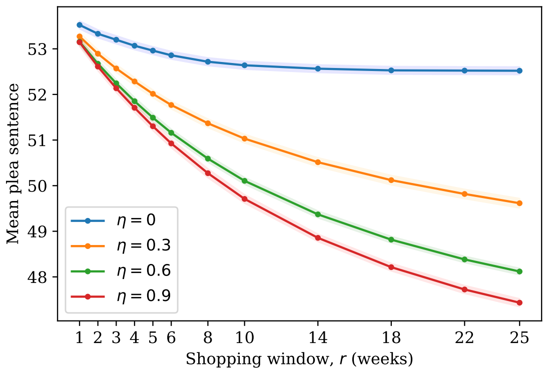
(b) Standard deviation of plea sentence

(c) Standard deviation across counties

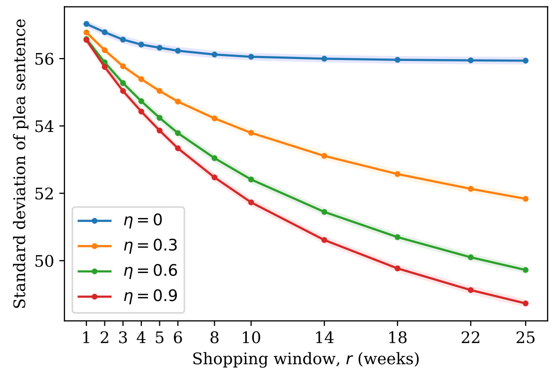
Figure 4 **Performance measures versus the travel probability** $\eta \in \{0, 0.1, \ldots, 0.9\}$**, where the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **and the shopping window is fixed at** $r = 4$ **weeks.**

impact being greater for larger values of the travel probability η or the shopping window r. Overall, at least over the ranges considered here, the impact of judge utilization is smaller than the impacts of the travel probability and the shopping window.

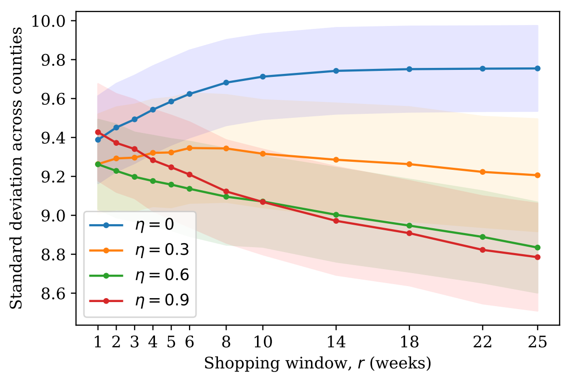
Returning to the impact of the key operational characteristics on the standard deviation of the mean plea sentence across counties, we see that the county variation typically (with a few exceptions) decreases as the travel probability η or the judge utilization ρ increases (Figs. 3c-4c, 7c-8c); in the latter case, a high utilization reduces the chance of choosing a lenient judge for defendants across all counties rather than benefiting some counties more.



(a) Mean plea sentence



(b) Standard deviation of plea sentence

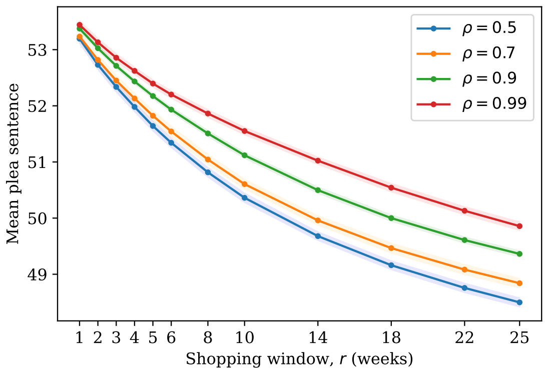


(c) Standard deviation across counties

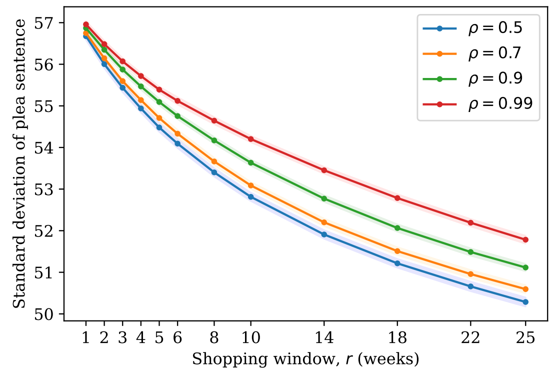
Figure 5 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks, where the travel probability** $\eta \in \{0,0.3,0.6,0.9\}$ **and judge utilization is fixed at** $\rho = 0.5$**.**

However, the behavior of county variation is more complex when the shopping window r increases, both with a fixed ρ or a fixed η. When we fix the judge utilization at $\rho = 0.5$ (Fig. 5c), if the travel probability η is small then increasing the shopping window r primarily impacts the counties with multiple home judges, which initially increases the county variation. In contrast, when η is large, an increase in the shopping window allows defendants in all counties to shop, leading to a decrease in county variation.

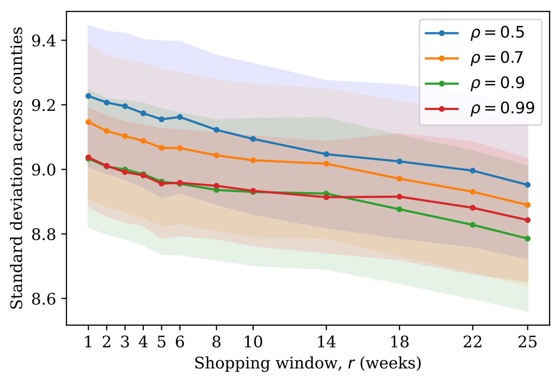
When we fix the travel probability at $\eta = 0.5$ (Fig. 6c), an increasing shopping window r enlarges the defendants' shopping options. A longer shopping window provides a larger benefit to defendants in counties with more visiting judges. When r is large enough, however, the defendants from most



(a) Mean plea sentence



(b) Standard deviation of plea sentence

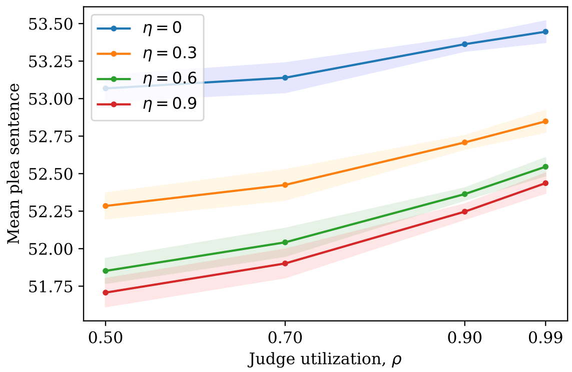


(c) Standard deviation across counties

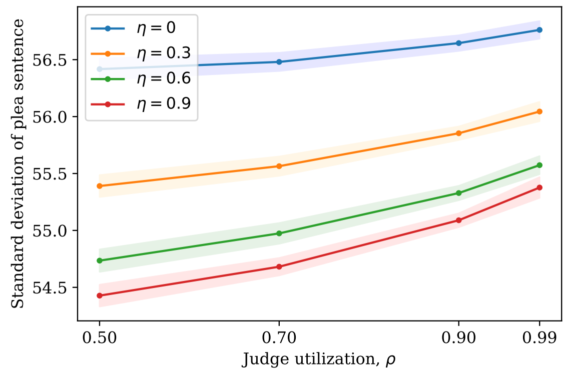
Figure 6 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks, where the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **and the travel probability is fixed at** $\eta = 0.5$**.**

counties are able to see all possible judges from their shopping window, which stabilizes the county variation.

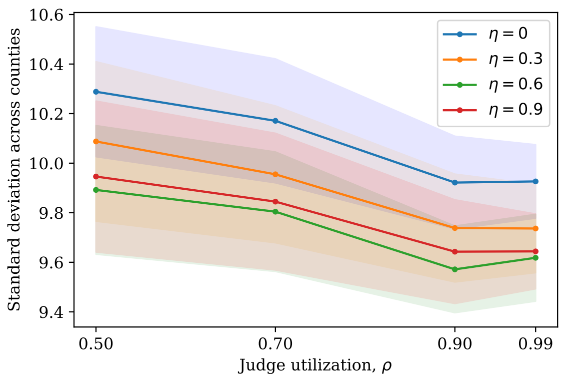
We perform three more calculations in this subsection in an attempt to gain more insight. First, to put the range of performance in Figs. 3-8 in perspective, we simulate an idealized scenario where defendants have access to all judges ($\eta = 1,\ r = \infty$), and judges have infinite capacity to handle pleas ($\rho = 0$). The results of this idealized scenario (Table 4) suggest that large values of the travel probability and the shopping window (e.g., Figs. 3-5) can achieve over half of the impact of the idealized scenario.



(a) Mean plea sentence



(b) Standard deviation of plea sentence



(c) Standard deviation across counties

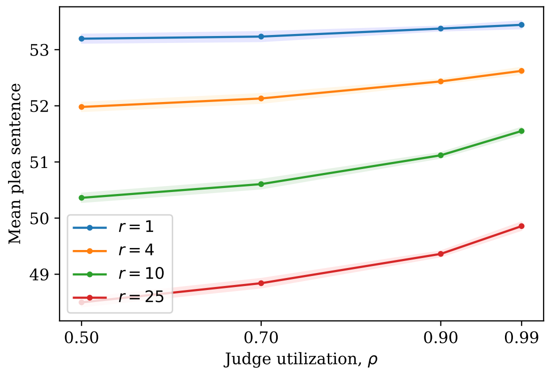
Figure 7 **Performance measures versus the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.95, 0.99\}$**, where the travel probability** $\eta \in \{0, 0.3, 0.6, 0.9\}$ **and the shopping window is fixed at** $r = 4$ **weeks.**

Table 4 **Results from an idealized scenario (**$\eta = 1$, $r = \infty$, $\rho = 0$**)**

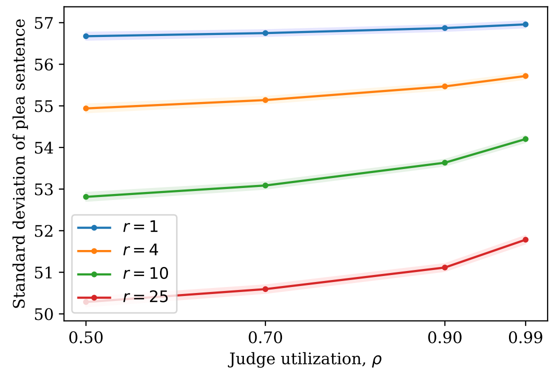
Mean plea sentence	43.75
Standard deviation of plea sentence	49.28
County variation	9.12

Second, referring back to the three possible outcomes in the plea bargain model in §3.3, we simulate the model under the base-case values ($\eta = 0.5$, $r = 4$, $\rho = 0.5$) and find that 6.4% of cases go to trial, the plea sentence equals the judge's upper allowable sentence in 9.2% of the cases, and the plea sentence is equal to the defendant's expected total trial cost in 84.5% of the cases.

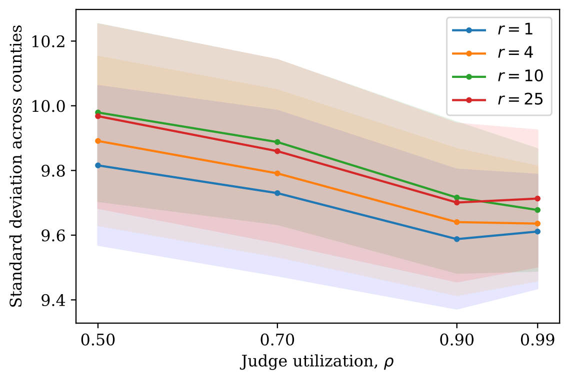
Finally, we elaborate on (and provide a visualization of) this last calculation by roughly estimating an upper bound on the proportion of defendants who are impacted by judge shopping. We



(a) Mean plea sentence



(b) Standard deviation of plea sentence



(c) Standard deviation across counties

Figure 8 **Performance measures versus the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.95, 0.99\}$**, where the shopping window** $r \in \{1, 4, 10, 25\}$ **weeks and the travel probability is fixed at** $\eta = 0.5$**.**

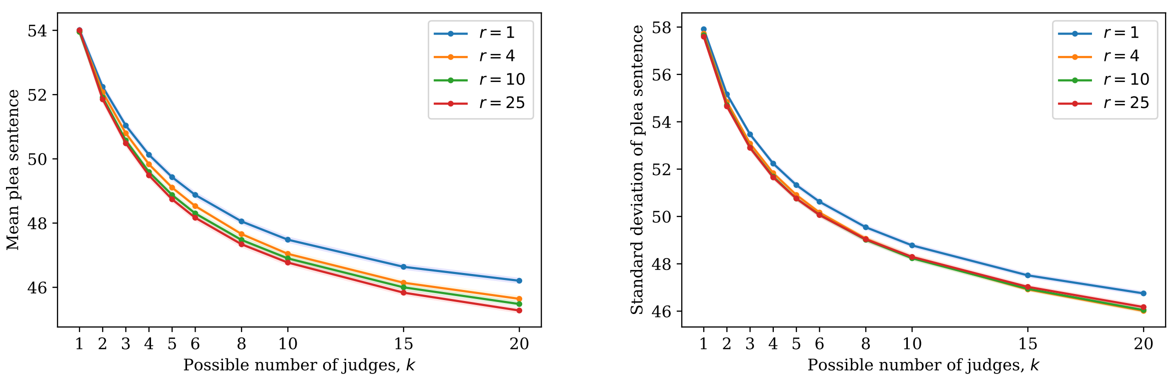
consider a hypothetical two-judge scenario in which each defendant is assigned the most lenient judge (Judge 36 in Fig. 19 in the Appendix) with probability 0.60 and the harshest judge (Judge 15 in Fig. 19 in the Appendix) with probability 0.40, where these probabilities maintain the mean nonzero sentence length to be 51.6 months, which is the overall mean sentence length in Fig. 19 in the Appendix. The defendants whose $(\theta_i \tau_i, s_i)$ fall in the interior of both convex hulls in Fig. 24 in the Appendix, which accounts for 90.5% of the defendants, are indifferent between Judges 15 and 36 because they would receive the same sentence length, $\theta_i \tau_i + c_d(i)$, from either judge. In contrast, the defendants with higher $\theta_i \tau_i + c_d(i)$ are more likely to be sensitive to the choice of judge. These last two calculations suggest that a small proportion of defendants are impacted, but that the impact on some of these individuals is large.

5.2. Urban Model

In this subsection, we adapt our model from the South Carolina setting to an urban setting, where all defendants and judges are located in the same place. We assume that defendants have access to all judges. This assumption is not realistic in that a case would generally be assigned to a judge without input from a defendant. However, local rules and practices may allow parties to engage in some strategic maneuvering, and our analysis is intended to gain insight into the impact of such maneuvering. The traveling probability η now plays no role, and we introduce another parameter whose value is specified: k, which represents the number of judges. Our simulation model is the same as before, except we now allow all $J = 50$ judges to be accessible to each defendant. Recall that an arriving defendant in our simulation model is assigned a set $J_i(r)$ of judges assigned to the defendant's county with remaining capacity within r weeks, and then chooses a judge according to (1). We change the model by allowing $J_i(r)$ to include all $J = 50$ judges that have remaining capacity within r weeks of the defendant's arrival. If the number of judges in this set, $|J_i(r)|$, satisfies $|J_i(r)| \leq k$, then the defendant chooses the best among the $|J_i(r)|$ judges according to (1). If $|J_i(r)| > k$, then the simulation chooses k out of $|J_i(r)|$ judges at random, and the defendant chooses the best among these k judges according to (1). Hence, the system behaves as if there are k judges, but this construction allows us to use the characteristics of all $J = 50$ judges in the dataset.

As in §5.1, we plot our performances measures (county variation no longer plays a role) against two key operational characteristics, keeping the third one fixed (Figs. 9-14, with the three-dimensional plots displayed in Figs. 25-27 in the Appendix), considering $k \in \{1, 2, 3, 4, 5, 6, 7, 10, 15, 20\}$ and a base-case value of $k = 4$ judges.

We highlight the differences in the results between the urban model and the South Carolina model. As expected, the mean and standard deviation of the sentence length decreases as the number of judges k increases because increasing k provides more shopping options. While this impact increases with a larger shopping window r and a smaller judge utilization ρ, these dependencies (Figs. 9-10) are much smaller than in the corresponding results (Fig. 3) in the South Carolina model. Most notably, a shopping window of $r = 1$ week still allows defendants shopping opportunities when k is large. In addition, a comparison of Figs. 6 and 11 reveals that the system performance is very insensitive to the judge utilization ρ in the urban model when the shopping window r is large, because the defendant gets more choices in the urban model than the South Carolina model when ρ and r are large. In Fig. 13, the mean plea sentence is insensitive to the judge utilization ρ when the shopping window r is large because most defendants have access to k judges when r is large, regardless of the utilization.

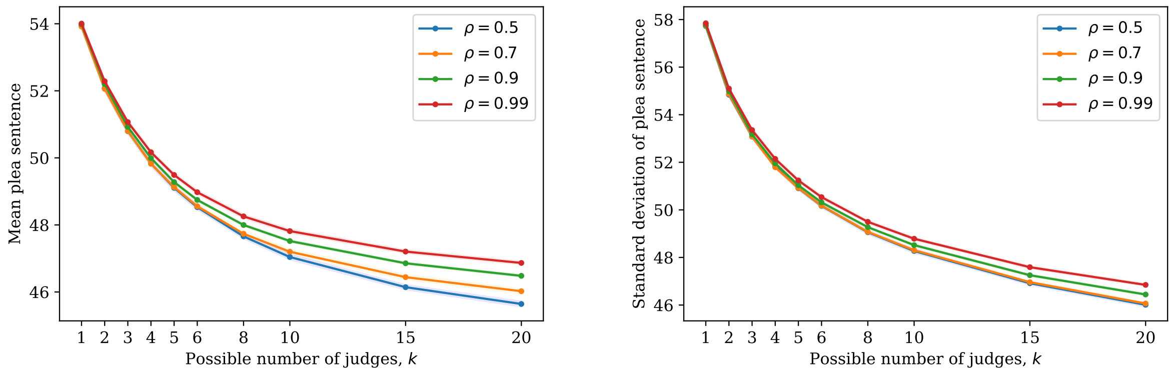


(a) Mean plea sentence (b) Standard deviation of plea sentence

Figure 9 **Performance measures versus the number of judges** $k \in \{1, 2, 3, 4, 5, 6, 8, 10, 15, 20\}$ **in the urban model, where the shopping window** $r \in \{1, 4, 10, 25\}$ **weeks and the judge utilization is fixed at** $\rho = 0.5$.

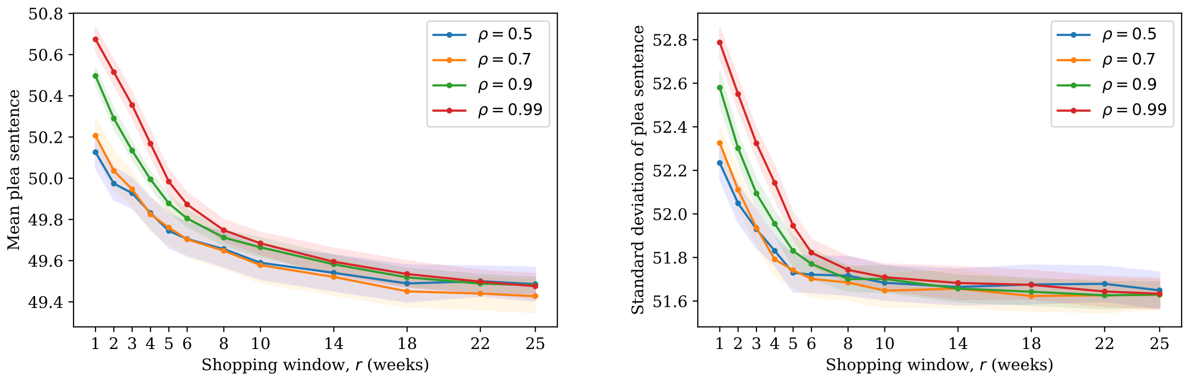
6. Discussion

Motivated by the insights and data in Hester (2017), we formulate and calibrate a mathematical model that allows judges to rotate among counties, allows defendants to shop for lenient judges, and incorporates a sentencing model that captures the strategic interactions among a judge, a defendant and a prosecutor. Our goal is to gain an understanding of how three key operational characteristics – the amount of judicial rotation (as measured by the judge travel probability), the amount of leeway defendants have in shopping (as measured by the shopping time window) and



(a) Mean plea sentence (b) Standard deviation of plea sentence

Figure 10 **Performance measures versus the number of judges** $k \in \{1,2,3,4,5,6,8,10,15,20\}$ **in the urban model, where the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **and the shopping window is fixed at** $r = 4$ **weeks.**

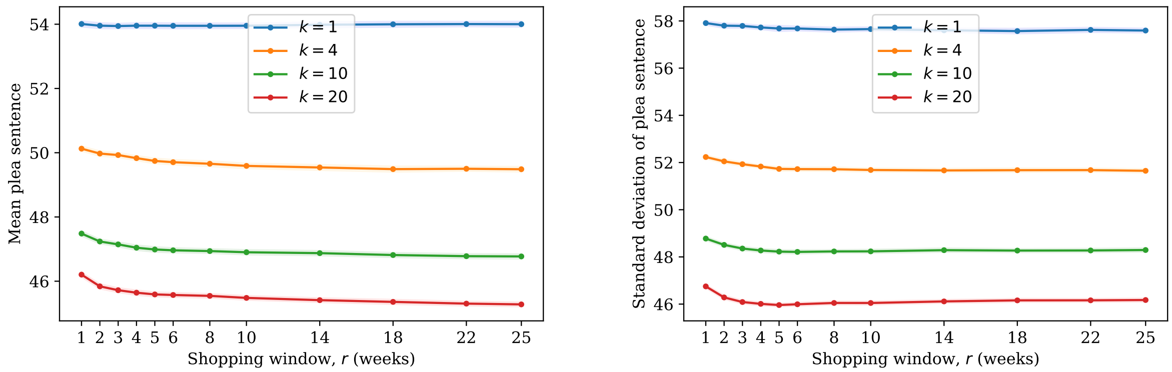


(a) Mean plea sentence (b) Standard deviation of plea sentence

Figure 11 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks in the urban model, where the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **and the number of judges is fixed at** $k = 4$**.**

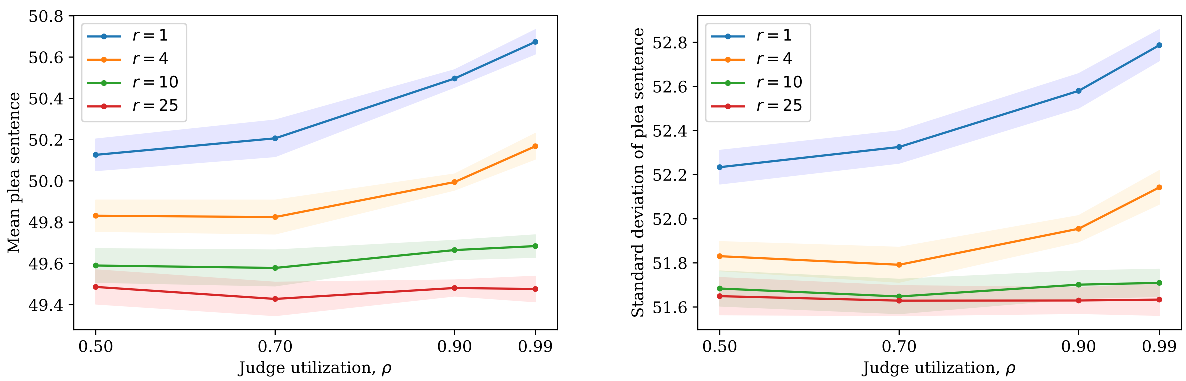
the system congestion (as measured by the judge utilization) – impact the mean and standard deviation of nonzero sentence lengths, and the county variation in mean sentence lengths. The latter two of these three performance measures quantify the amount of sentencing inequity across defendants.

Our first-order findings (Figs. 3-8) are intuitive: the mean and standard deviation of the plea sentence length decrease when judges travel more among counties, defendants have a longer shop-



(a) Mean plea sentence (b) Standard deviation of plea sentence

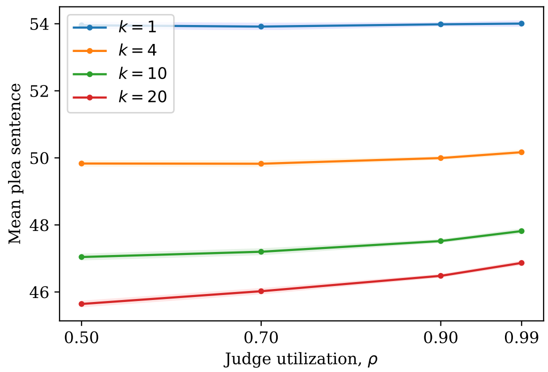
Figure 12 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks in the urban model, where the number of judges** $k \in \{1,2,3,4,5,6,8,10,15,20\}$ **and the judge utilization is fixed at** $\rho = 0.5$**.**



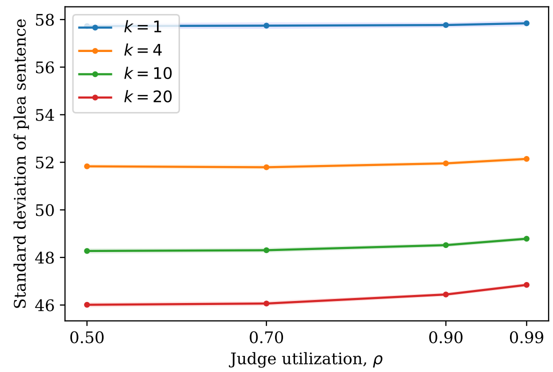
(a) Mean plea sentence (b) Standard deviation of plea sentence

Figure 13 **Performance measures versus the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **in the urban model, where the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks and the number of judges is fixed at** $k = 4$**.**

ping window, and judges have more excess capacity. All three characteristics make it easier for a defendant to plea in front of a lenient judge. The impact of all three characteristics have decreasing returns to scale, in that a moderate amount of judge shopping, a moderate shopping window, and a moderate amount of excess judicial capacity achieve a significant proportion of the potential effects. In terms of interaction effects, we find that the judge travel probability and the shopping window



(a) Mean plea sentence



(b) Standard deviation of plea sentence

Figure 14 Performance measures versus the judge utilization $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ in the urban model, where the number of judges $k \in \{1, 4, 10, 20\}$ and the shopping window is fixed at $r = 4$ weeks.

are complements: each variable generates a bigger impact when the other variable is large. That is, very little impact occurs when judges travel but defendants cannot shop, or when defendants can shop but the judges do not travel. In contrast, high judge utilization reduces – but does not negate – the first-order and synergistic effects of increasing the judge travel probability and the shopping window. In addition, although the overall reduction on the mean and standard deviation is modest (≈10%), we find that the changes in sentencing as a result of these key operational characteristics affect a small proportion of defendants – typically those who have a high cost of going to trial – but the impact on these individual defendants is large.

The Hester (2017) study uses data from South Carolina in 2000-2001, when judicial rotation was used and most counties had a single home judge Hester (2017). Very few states have employed judicial rotation during the last 20 years. In contrast, urban areas throughout the US typically have multiple judges in the same county, which may allow for some judge shopping by defendants. Hence, we adapt the South Carolina model into an urban model, where the number of judges replaces the judge travel probability as the third key operational characteristic. As expected, the mean and standard deviation of the plea sentence length are reduced when the number of judges and the shopping window are large and the judge utilization is low. However, in contrast to the county model, the impact of the number of judges is much larger than the impact of the shopping window, and there is very little synergy: even if intertemporal shopping is not allowed (i.e., $r = 1$ week), the defendant still gets to choose the most lenient among the available judges and can lower his mean sentence length. The impact of judge utilization is smaller than it is in the South Carolina model.

The system we attempt to model is very complex, and our model should be viewed as a mere caricature of the actual process. In particular, we use a highly idealized model for the plea bargain process (e.g., we do not include costs of going to trial for the prosecutor, as in Silveira (2017)), and a simplified model for how judges spend their time and process cases (e.g., judicial decisions do not depend on the amount of congestion, as in Yang (2016)). Moreover, there are no data to directly estimate some of the parameters, particularly the cost of going to trial and the waiting cost. However, some of our methods may be useful, particularly our construction of the convex hulls to estimate judge leniency.

These limitations lead us to conclude that the exact numerical results in Figs. 3-8 should not be interpreted as accurate counterfactual predictions of the system behavior in South Carolina in 2000-2001. However, we believe that our broad qualitative insights are likely to be robust. For one, in relation to previous work in South Carolina, a jurisdiction with heavy utilization of judicial rotation, our findings articulate distinct dimensions of rotation operationalized as the amount of rotation, the shopping window, and capacity utilization. The findings suggest that increased rotation and a longer shopping window can lead to larger decreases in the mean and standard deviation of sentence length, with a small impact of capacity utilization. Thus, expanding choice of judges could assist in aims of incarceration reduction and achieving greater uniformity in sentencing. Second, as applied in the urban models, introducing greater choice of judge could advance similar aims, even with short shopping windows. We do note the paradox, however, that introducing greater shopping choice in an urban jurisdiction may facilitate inter-judge uniformity within the urban jurisdiction but thereby exacerbate county differences within a state. As a result, some defendants in rural settings may suffer in comparison if their county has a single harsh judge.

References

Eisenstein J, Flemming RB, Nardulli PF (1988) *The Contours of Justice: Communities and Their Courts.* (Boston, MA: Little, Brown).

Frazier CE, Bock EW (1982) Effects of court officials on sentence severity: Do judges make a difference? *Criminology* 20(2):257–272, URL `http://dx.doi.org/https://doi.org/10.1111/j.1745-9125.1982.tb00460.x`.

Gross SR, Syverud KD (1991) Getting to no: A study of settlement negotiations and the selection of cases for trial. *Michigan Law Review* 90:319–393.

Hester R (2017) Judicial rotation as centripetal force: Sentencing in the court communities of South Carolina. *Criminology* 55(1):205–235, URL `http://dx.doi.org/https://doi.org/10.1111/1745-9125.12132`.

Hester R, Hartman TK (2017) Conditional race disparities in criminal sentencing: A test of the liberation hypothesis from a non-guidelines state. *Journal of Quantitative Criminology* 33(1):77–100, URL `http://dx.doi.org/10.1007/s10940-016-9283-z`.

Johnson BD (2006) The multilevel context of criminal sentencing: Integrating judge- and county-level influences*. *Criminology* 44(2):259–298, URL `http://dx.doi.org/https://doi.org/10.1111/j.1745-9125.2006.00049.x`.

Myers MA, Talarico SM (1987) *The Social Contexts of Criminal Sentencing* (Springer New York, NY), ISBN 978-0-387-96483-6.

Ostrom B, Hanson R, for State Courts NC, Institute APR (1999) *Efficiency, Timeliness, and Quality: A New Perspective from Nine State Criminal Trial Courts* (National Center for State Courts), ISBN 9780896561991, URL `https://books.google.com/books?id=8KdgvgEACAAJ`.

Silveira BS (2017) Bargaining with asymmetric information: An empirical study of plea negotiations. *Econometrica* 85(2):419–452, URL `http://dx.doi.org/https://doi.org/10.3982/ECTA12974`.

Spohn C (1990) The sentencing decisions of black and white judges: Expected and unexpected similarities. *Law & Society Review* 24(5):1197–1216, ISSN 00239216, 15405893, URL `http://www.jstor.org/stable/3053666`.

Steffensmeier D, Britt CL (2001) Judges' race and judicial decision making: Do black judges sentence differently? *Social Science Quarterly* 82(4):749–764, URL `http://dx.doi.org/https://doi.org/10.1111/0038-4941.00057`.

Ulmer JT (1997) *Social Worlds of Sentencing: Court Communities Under Sentencing Guidelines* (SUNY Press, NY).

Ulmer JT (2019) Criminal courts as inhabited institutions: Making sense of difference and similarity in sentencing. *Crime and Justice* 48:483–522, URL `http://dx.doi.org/10.1086/701504`.

Wang C (2019) *Three Data-driven Model-based Policy Analyses in Criminology*. Ph.D. thesis, Stanford University.

Yang CS (2016) Resource constraints and the criminal justice system: Evidence from judicial vacancies. *American Economic Journal: Economic Policy* 8(4):289–332, URL `http://dx.doi.org/10.1257/pol.20150150`.

APPENDIX

The Appendix contains supplementary material pertaining to the parameter estimation procedure in §A. Tables 9-12 and Figs. 19-27 are referred to in the main text.

Appendix A: Parameter Estimation

We map from x^*_{jc} to C_{jct} in §A.1, modify the convex hulls in §A.2, and choose the value of k in the k-nearest neighbor analysis in §A.3.

A.1. Mapping from x^*_{jc} to C_{jct}

We first provide an algorithm for the judge schedule C_{jct} when the judge travel probability $\eta = 0$, and use that as a basis for generating judge schedules when $\eta > 0$.

C_{jct} when $\eta = 0$. The mathematical program (2)-(5), which has input values in Tables 5-6, has solution x^*_{jc} (Table 7), which gives the fraction of each judge's capacity that is allocated among her home counties. This solution yields a target capacity for county c of $\sum_{j=1}^{J} \kappa_j x^*_{jc}$, and our goal is to generate a schedule that satisfies these target capacities. We use the master calendar to determine the timing of judge assignments and only change the assigned county for each judge-week in order to satisfy the target allocation for each county. The algorithm consists of the following steps.

- **Step 1.** Obtain the judge assignment for each judge-week directly from the master calendar.
- **Step 2.** For each judge-week, if the judge is assigned to one county as in Table 7, assign the county as the working county for this assignment. Else, move to **Step 3**.
- **Step 3.** Let d_{jc} be the residual target allocation from judge j to county c as in Table 7. Initially, we set $d_{jc} = \kappa_j x^*_{jc}$. Let $S^0_j = \{c | x^*_{jc} > 0\}$ be the set of possible home counties for judge j.
- **Step 4.** Randomly select county $c \in S^0_j$ with probability $d_{jc} / \sum_{c \in S^0_j} d_{jc}$.
  - —If d_{jc} is greater than the judge's working time of this judge-week, meaning that this judge-week can be fully used for the selected county, assign the selected county as the working county. Update d_{jc} by decreasing the judge's working time of this judge-week.
  - —Else, we know that only part of the judge-week is needed to satisfy the target allocation from the selected county. Assign the selected county to the judge-week, update the judge's residual capacity by setting $d_{jc} = 0$, and repeat this step with the residual capacity of this judge-week.

C_{jct} **when** $\eta > 0$. To generate a judge schedule when $\eta > 0$, we start with the judge schedule when $\eta = 0$ and then determine whether the judge of each judge-week is a home judge or a traveling judge based on the value of η. We then assign the same working county as in the $\eta = 0$ judge schedule for judge-weeks whose sentencing judge is a home judge, and use the capacity of the judge-weeks whose judge is a traveling judge to satisfy the target allocation that is not covered by the home judges for each county. The algorithm consists of the following steps.

- **Step 1.** Generate a base judge schedule with $\eta = 0$.
- **Step 2.** For each judge-week, decide whether the judge is a home judge or a traveling judge using η randomly.
- **Step 3.** For each judge-week whose judge is a home judge, assign her the same working county as in the base judge schedule from **Step 1**. Let d_c^t be the uncovered target allocation by home judges for county c.
- **Step 4.** For each judge-week whose judge is a traveling judge, let $S_j = C \setminus \{c | x_{jc}^* > 0\}$ be the set of counties such that judge j can satisfy some of their uncovered demand when traveling from Step 3. Randomly select county $c \in S_j$ with probability $d_c^t / \sum_{c \in S_j} d_c^t$.
- If d_c^t is greater than the judge's working time of this judge-week, meaning that this judge-week can be fully used for the selected county, assign the selected county as the working county. Update d_c^t by decreasing the judge's working time of this judge-week.
- Else, we know that only part of the judge-week is needed to satisfy the target allocation from the selected county. Assign the selected county to the judge-week, update the judge's residual capacity by setting $d_c^t = 0$, and repeat this step with the residual capacity of this judge-week.

A.2. Modifying the Convex Hulls

In this subsection, we modify the convex hulls derived in §4.4 by detecting and removing outliers from the judge sentences, and extrapolating the convex hull for each judge.

Detecting Outliers. We detect outliers from the set of observations $\{(\theta_i \tau_i, s_i) | s_i > 0, i \in \mathcal{I}_j\}$ for judge j using the Mahalanobis distance. Using generic notation, suppose we are given a two-dimensional dataset where $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ denotes the i^{th} observation for $i = 1, \ldots, n$. We detect the outliers using the following steps.

Table 5 Values for κ_j

Judge Number	κ_j	Judge Number	κ_j
1	25.0	26	21.2
2	24.4	27	12.6
3	15.0	28	19.6
4	17.5	29	20.4
5	15.4	30	20.2
6	18.2	31	12.0
7	34.0	32	13.6
8	14.8	33	20.0
9	21.3	34	15.2
10	22.2	35	11.0
11	22.0	36	10.0
12	16.8	37	9.5
13	21.9	38	13.0
14	15.0	39	15.8
15	11.0	40	8.0
16	19.0	41	8.0
17	14.6	42	9.0
18	25.4	43	14.6
19	19.0	44	16.0
20	4.8	45	18.2
21	14.2	46	10.33
22	20.7	47	20.2
23	13.2	48	16.0
24	22.7	49	27.8
25	17.8	50	26.3

Step 1. Find the covariance matrix M between the two dimensions for the dataset.

Step 2. Find the center point of the dataset by taking the average value of each variable as

$$\boldsymbol{x}_c = (\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{i1}, \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{i2}).$$

Step 3. Compute the Mahalanobis Distance between each observation $\boldsymbol{x}_i$ and the center point $\boldsymbol{x}_c$ as

$$d_i^2 = (\boldsymbol{x}_i - \boldsymbol{x}_c)^T M^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_c).$$

Step 4. Depending on the percentage of the observations that the user wants to detect as outliers, choose a cutoff value m from the chi-squared distribution. Observation i is deemed an outlier if $d_i^2 > m$.

We fix the threshold from the chi-squared distribution to be 10.6 for each judge, which leads to 3.44% of observations being detected as outliers (and a range of 0%–6% among judges). The results of the outlier detection for each judge appears in Fig. 15, where the header of each figure also shows the percentage of the sentencing events that are detected as outliers for each judge.

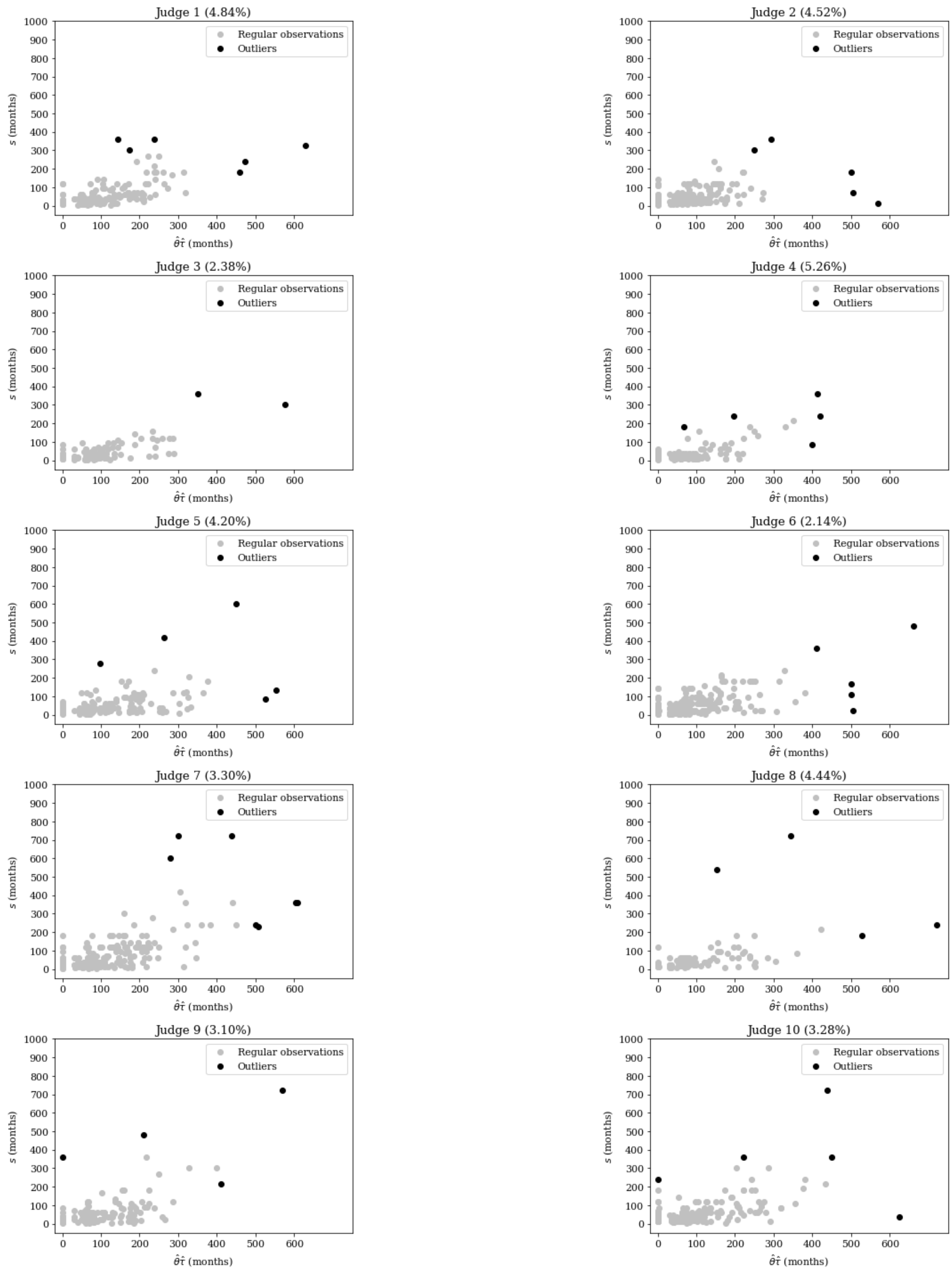


Figure 15 **Outlier distribution among judges, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial. The number in parenthesis above each subfigure is the percentage of the observations that are outliers.**

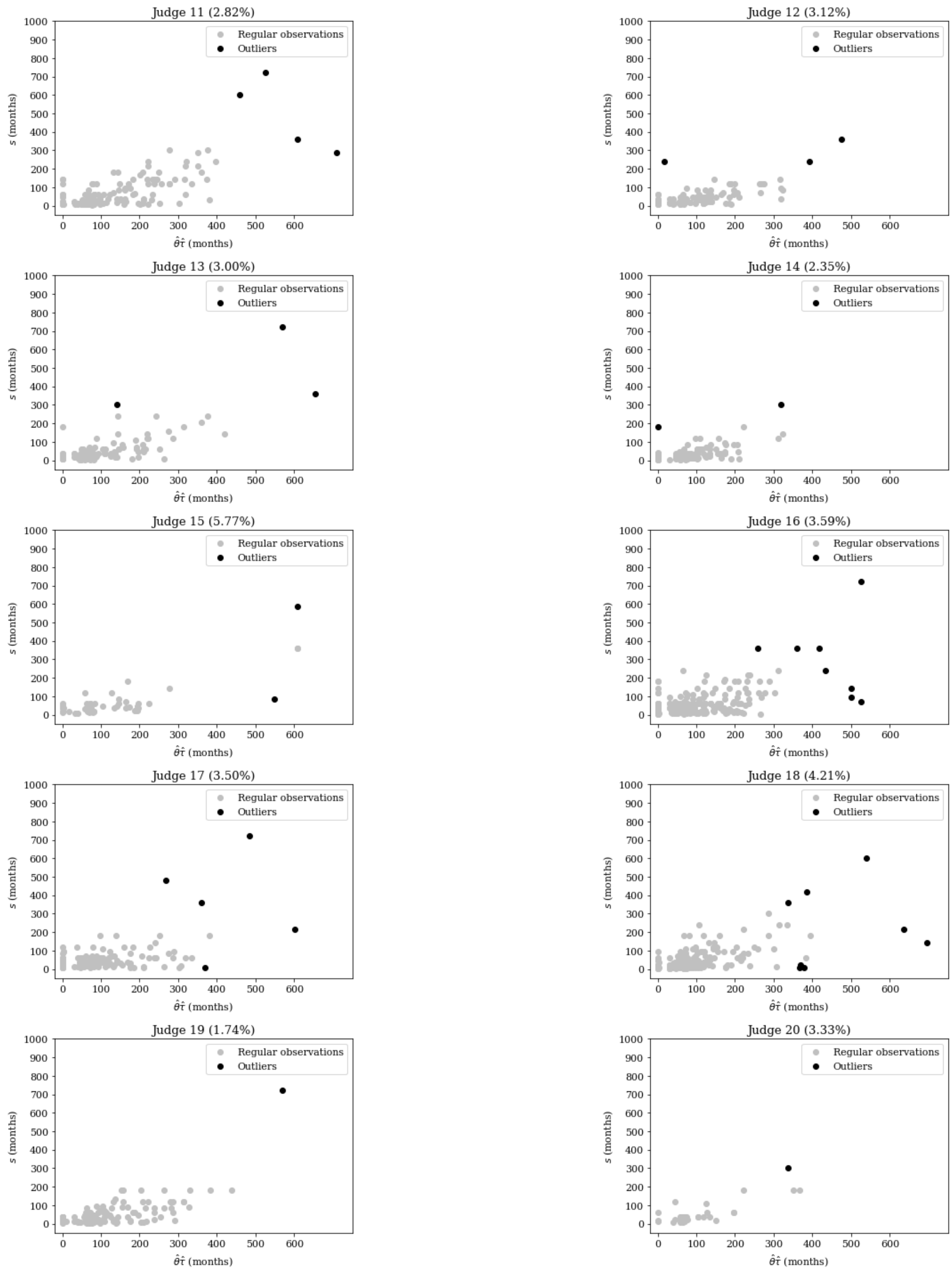


Figure 15 **(Continued) Outlier distribution among judges, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial. The number in parenthesis above each subfigure is the percentage of the observations that are outliers.**

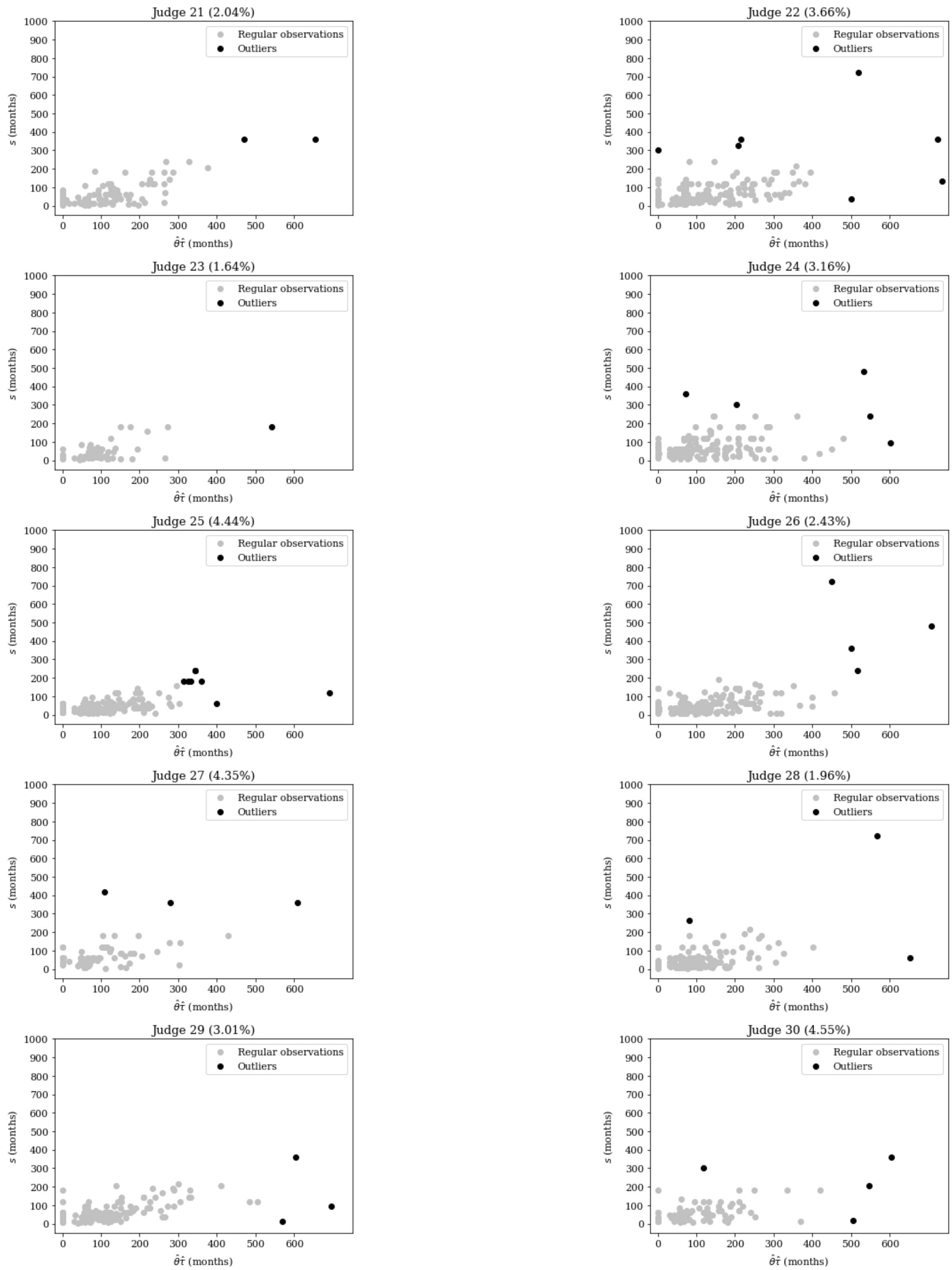


Figure 15 **(Continued) Outlier distribution among judges, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial. The number in parenthesis above each subfigure is the percentage of the observations that are outliers.**

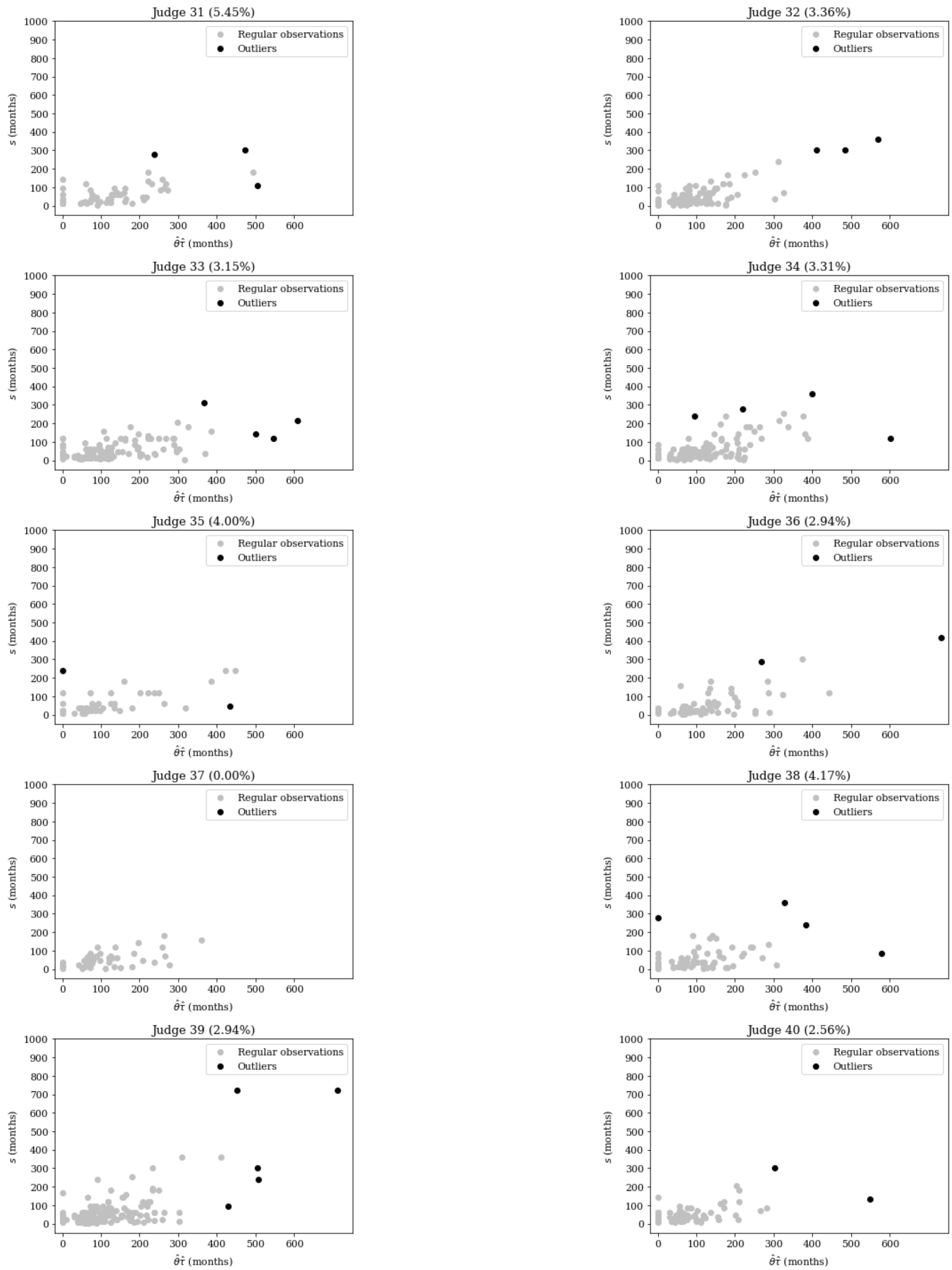


Figure 15 **(Continued) Outlier distribution among judges, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial. The number in parenthesis above each subfigure is the percentage of the observations that are outliers.**

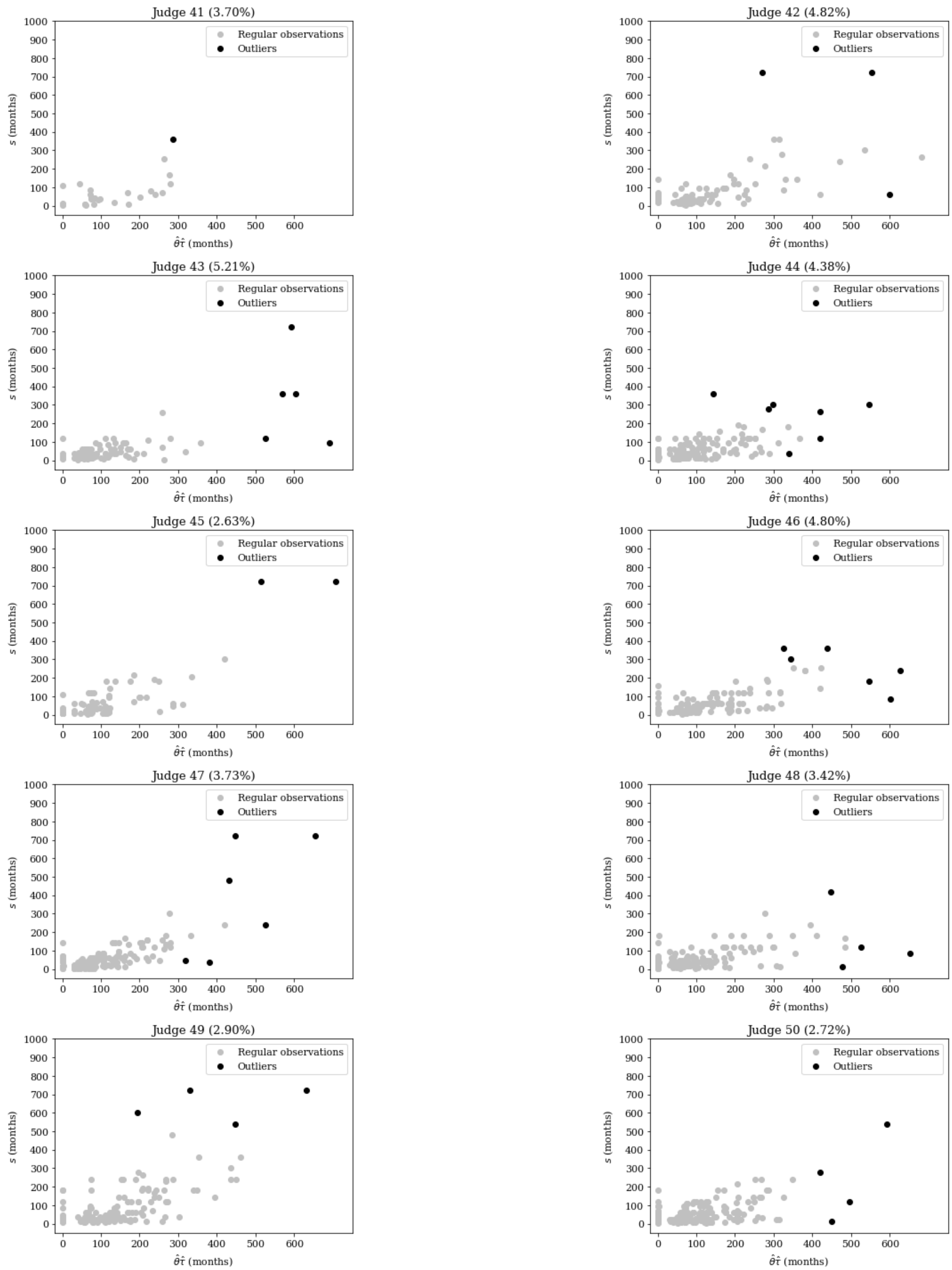


Figure 15 **(Continued) Outlier distribution among judges, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial. The number in parenthesis above each subfigure is the percentage of the observations that are outliers.**

Table 6 Values for d_c

County Number	d_c	County Number	d_c
1	13.581	24	2.741
2	2.812	25	2.617
3	2.100	26	10.644
4	4.788	27	11.463
5	1.958	28	0.481
6	5.910	29	6.408
7	1.691	30	28.569
8	2.617	31	0.854
9	4.948	32	1.2816
10	0.552	33	11.606
11	2.207	34	13.635
12	6.177	35	1.086
13	6.835	36	3.115
14	3.969	37	0.837
15	4.058	38	2.083
16	23.211	39	2.848
17	2.011	40	2.599
18	1.780	41	1.050
19	4.343	42	1.299
20	27.270	43	5.589
21	3.097	44	1.371
22	2.812	45	0.374
23	19.010	46	5.020

Extrapolating the Convex Hulls Let $\mathcal{A}_j$ denote the convex hull generated by the origin $(0,0)$ and the observations from $\{(\hat{\theta}_i\hat{\tau}_i, s_i)|s_i > 0, i \in \mathcal{I}_j\}$ (Fig. 2 in the main text). To add the two artificial points, we let $\mathcal{I}_j^4$ be the set of trial cases handled by judge j for $j = 1, \ldots, J$ and calculate the maximum expected sentence length predicted for all defendants, including both plea bargains and trials, via $\overline{\theta\tau} = \max_{i \in \cup_{j=1}^{J}(\mathcal{I}_j \cup \mathcal{I}_j^4)} \hat{\theta}_i\hat{\tau}_i$. The two artificial points are denoted by $(\overline{\theta\tau}, \hat{s}_u)$ and $(\overline{\theta\tau}, \hat{s}_\ell)$, where $\hat{s}_u$ and $\hat{s}_\ell$ are interpreted as the maximum and minimum, respectively, plea offers that the judge allows for the defendant whose expected trial sentence length is $\overline{\theta\tau}$.

Let A be the set of observations that forms the boundary of the convex hull $\mathcal{A}_j$. To compute $\hat{s}_u$, we let x_1 be the observation in A with the largest sentence s, and x_2 be the observation in A that is found by traversing A in the counterclockwise direction starting from x_1 (Fig. 16). We then set

$$\hat{s}_u = s_1 + \left(\frac{s_1 - s_2}{\hat{\theta}_1\hat{\tau}_1 - \hat{\theta}_2\hat{\tau}_2}\right)\left(\overline{\theta\tau} - \hat{\theta}_1\hat{\tau}_1\right),$$

which is the point where the line containing x_1 and x_2 intersects with the dashed vertical line in Fig. 16.

Table 7 Judge assignment obtained by solving (2)-(5)

Judge Number	Assigned counties (Percentage)
1	County 34 (100%)
2	County 37 (100%)
3	County 46 (100%)
4	County 13 (100%)
5	County 35 (100%)
6	County 45 (4.0%), County 12 (58.0%), County 1 (38.0%)
7	County 2 (100%)
8	County 6 (100%)
9	County 20 (100%)
10	County 44 (100%)
11	County 38 (100%)
12	County 21 (100%)
13	County 31 (100%)
14	County 23 (100%)
15	County 5 (100%)
16	County 35 (100%)
17	County 14 (100%)
18	County 16 (100%)
19	County 13 (100%)
20	County 4 (100%)
21	County 36 (100%)
22	County 43 (100%)
23	County 33 (100%)
24	County 36 (100%)
25	County 38 (100%)
26	County 16 (100%)
27	County 10 (100%)
28	County 30 (100%)
29	County 33 (100%)
30	County 39 (17.0%), County 26 (5.0%), County 24 (29.0%), County 25 (25.0%), County 40 (25.0%)
31	County 35 (100%)
32	County 3 (23.0%), County 20 (14.0%), County 41 (12.0%), County 4 (1.0%), County 9 (50.0%)
33	County 27 (100%)
34	County 30 (100%)
35	County 13 (100%)
36	County 20 (100%)
37	County 27 (100%)
38	County 1 (100%)
39	County 26 (100%)
40	County 39 (40.0%), County 17 (60.0%)
41	County 15 (100%)
42	County 42 (21.0%), County 4 (0.0%), County 26 (1.0%), County 8 (55.0%), County 32 (22.0%)
43	County 11 (100%)
44	County 29 (53.0%), County 26 (13.0%), County 19 (34.0%)
45	County 18 (100%)
46	County 22 (100%)
47	County 7 (100%)
48	County 23 (100%)
49	County 28 (100%)
50	County 11 (100%)

Similarly, to compute $\hat{s}_\ell$, we let x_3 be the observation in A with the smallest sentence s, and x_4 be the observation in A that is found by traversing A in the clockwise direction starting from x_3 (Fig. 17), and set

$$\hat{s}_\ell = s_3 + \left(\frac{s_3 - s_4}{\hat{\theta}_3 \hat{\tau}_3 - \hat{\theta}_4 \hat{\tau}_4} \right) \left(\overline{\theta\tau} - \hat{\theta}_3 \hat{\tau}_3 \right).$$

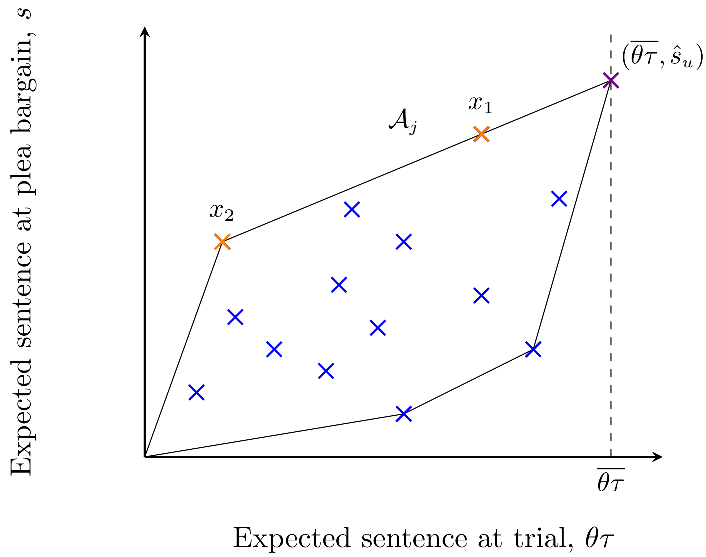


Figure 16 **The modified convex hull for judge j, which preserves the upper leniency threshold $u_j(\cdot)$. This figure is for illustration purposes and is not based on the data.**

The modified convex hull for judge j, denoted by A_j^m, is generated by the origin $(0,0)$, the observations from $\{(\hat{\theta}_i\hat{\tau}_i, s_i)|s_i > 0, i \in \mathcal{I}_j\}$, and the two artificial observations $(\overline{\theta\tau}, \hat{s}_u)$, and $(\overline{\theta\tau}, \hat{s}_\ell)$ (Fig. 17). Taken together, the lower and upper leniency thresholds of judge j is given by the functions

$$\ell_j(\hat{\theta}_i\hat{\tau}_i) = \min\left\{s \middle| (\hat{\theta}_i\hat{\tau}_i, s) \in \mathcal{A}_j^m\right\},$$
$$u_j(\hat{\theta}_i\hat{\tau}_i) = \max\left\{s \middle| (\hat{\theta}_i\hat{\tau}_i, s) \in \mathcal{A}_j^m\right\}.$$

The convex hull for each of the 50 judges appears in Fig. 18.

A.3. Choice of k in the Trial Cost Estimation

In this section, we describe how we choose $k = 20$ for the $k-$nearest neighbor method used to estimate the trial cost. A value of k that is too small may overfit while a value of k that is too large may include distant neighbors that introduce inaccuracies. Also, we hope to choose a k such that most of the defendants from $\mathcal{I}^2$ and $\mathcal{I}^4$ have their trial cost imputed based on k neighbors. The 25$^{\text{th}}$, 50$^{\text{th}}$ and 75$^{\text{th}}$ percentiles are 22, 362, and 5074 for S^2, and 1, 132.5, and 869 for S^4. These percentiles lead us to consider $\{10, 15, 20, 25\}$ as candidate values for k.

To assess these candidate values, we compute the sentence based on our model for each of the 17,270 sentencing events in our dataset. That is, for each defendant i and his judge $j(i)$ from the sentencing dataset,

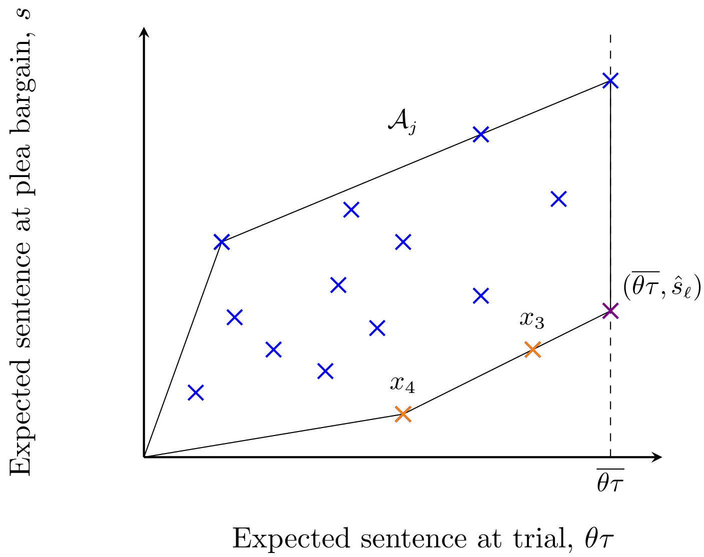


Figure 17 A modified convex hull, A_j^m, for judge j, which preserves the judge's upper and lower leniency thresholds. This figure is for illustration purposes and is not based on the data.

we use §4.4 to estimate his incarceration probability p_i, predict his expected trial sentence $\hat{\theta}_i\hat{\tau}_i$, estimate the upper leniency level $u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)$ of judge $j(i)$ and estimate his trial cost $c_d(i)$ with $k \in \{10, 15, 20, 25\}$, and then determine his sentence length via

$$s_i(j(i)) = \min\left\{\hat{\theta}_i\hat{\tau}_i + c_d(i), u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)\right\}$$

when $\hat{\theta}_i\hat{\tau}_i \le u_{j(i)}(\hat{\theta}_i\hat{\tau}_i)$, and send him to trial if $\hat{\theta}_i\hat{\tau}_i < \ell_{j(i)}(\hat{\theta}_i\hat{\tau}_i)$.

We conduct a simulation study following this process for different values of k, and consider the three performance measures defined in §3.4 along with the fraction of cases that are resolved through trials. Although all four values of k perform well, we choose $k = 20$, which is closest to the data for two of these four performance measures (Table 8).

Table 8 Statistics from the sentencing dataset and simulation results with different values of k in the trial cost estimation

	Data	$k=10$	$k=15$	$k=20$	$k=25$
Mean plea sentence (for nonzero sentences)	54.27	50.79	51.27	51.71	51.65
Standard deviation of plea sentence	40.52	40.21	40.43	40.69	40.45
Percentage of trial cases	1.51%	2.10%	1.40%	1.49%	1.44%
Standard deviation across counties (for nonzero sentences)	11.74	12.55	12.61	12.52	13.36

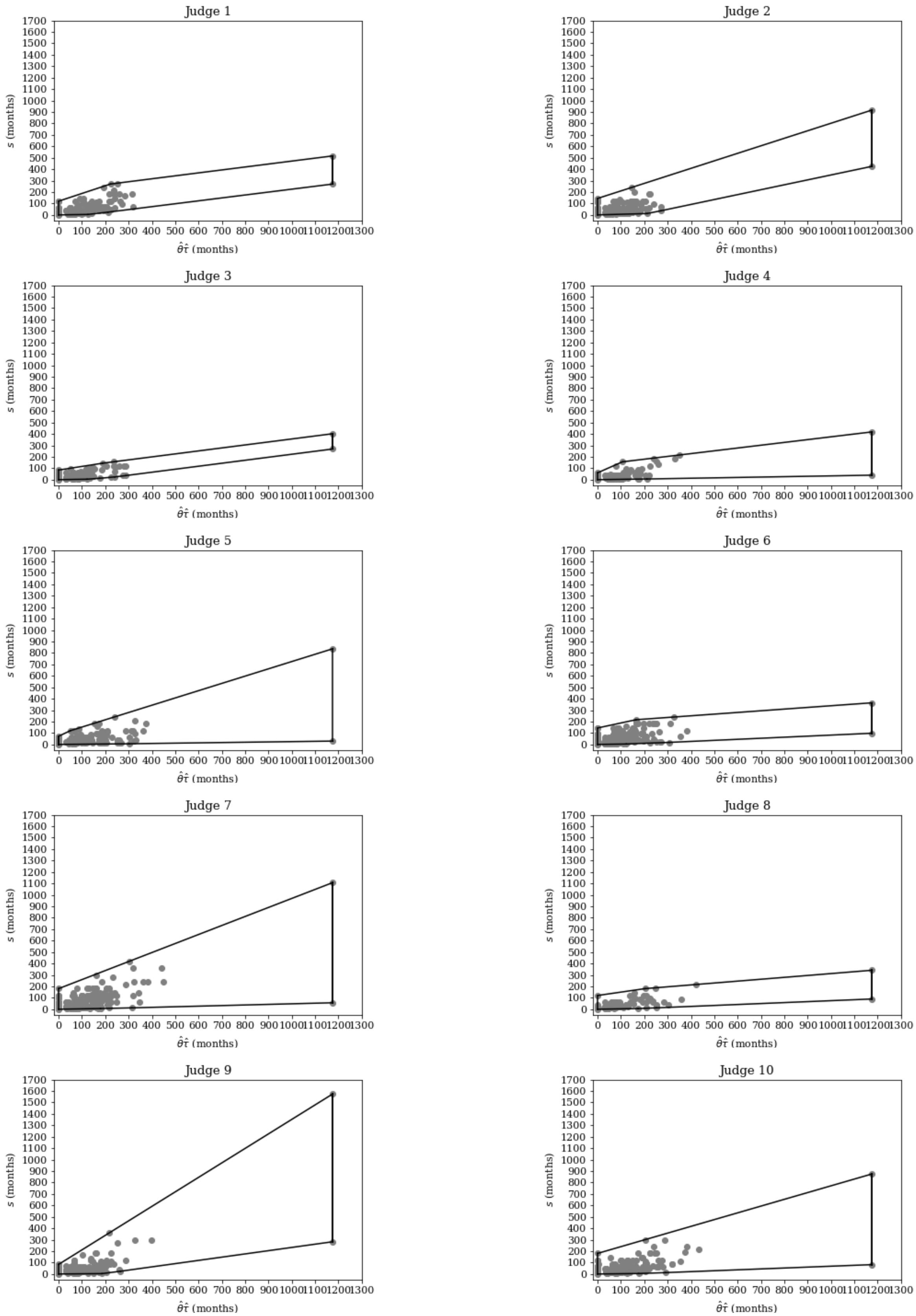


Figure 18 **Judge convex hull with outliers removed, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial.**

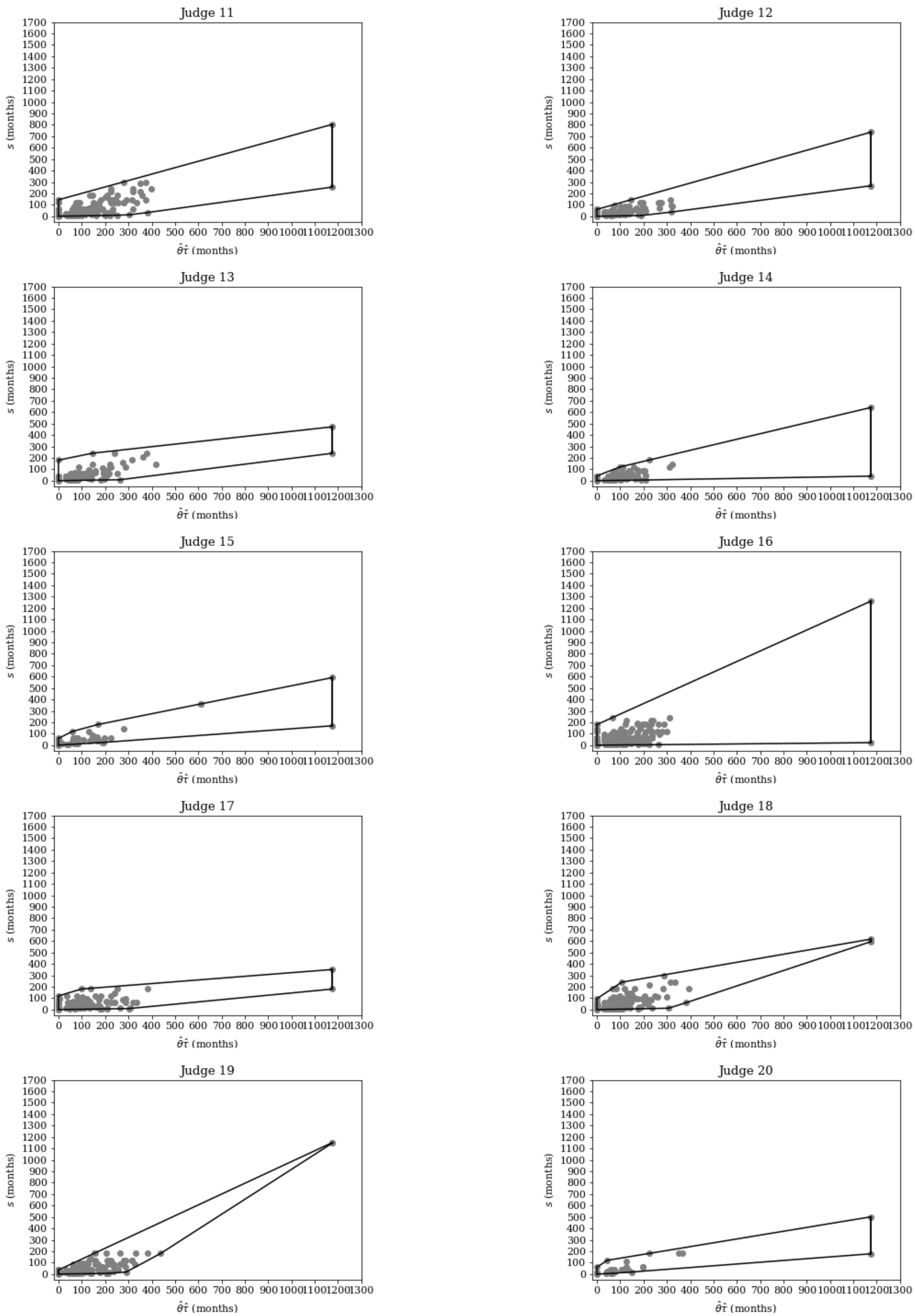


Figure 18 **(Continued) Judge convex hull with outliers removed, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial.**

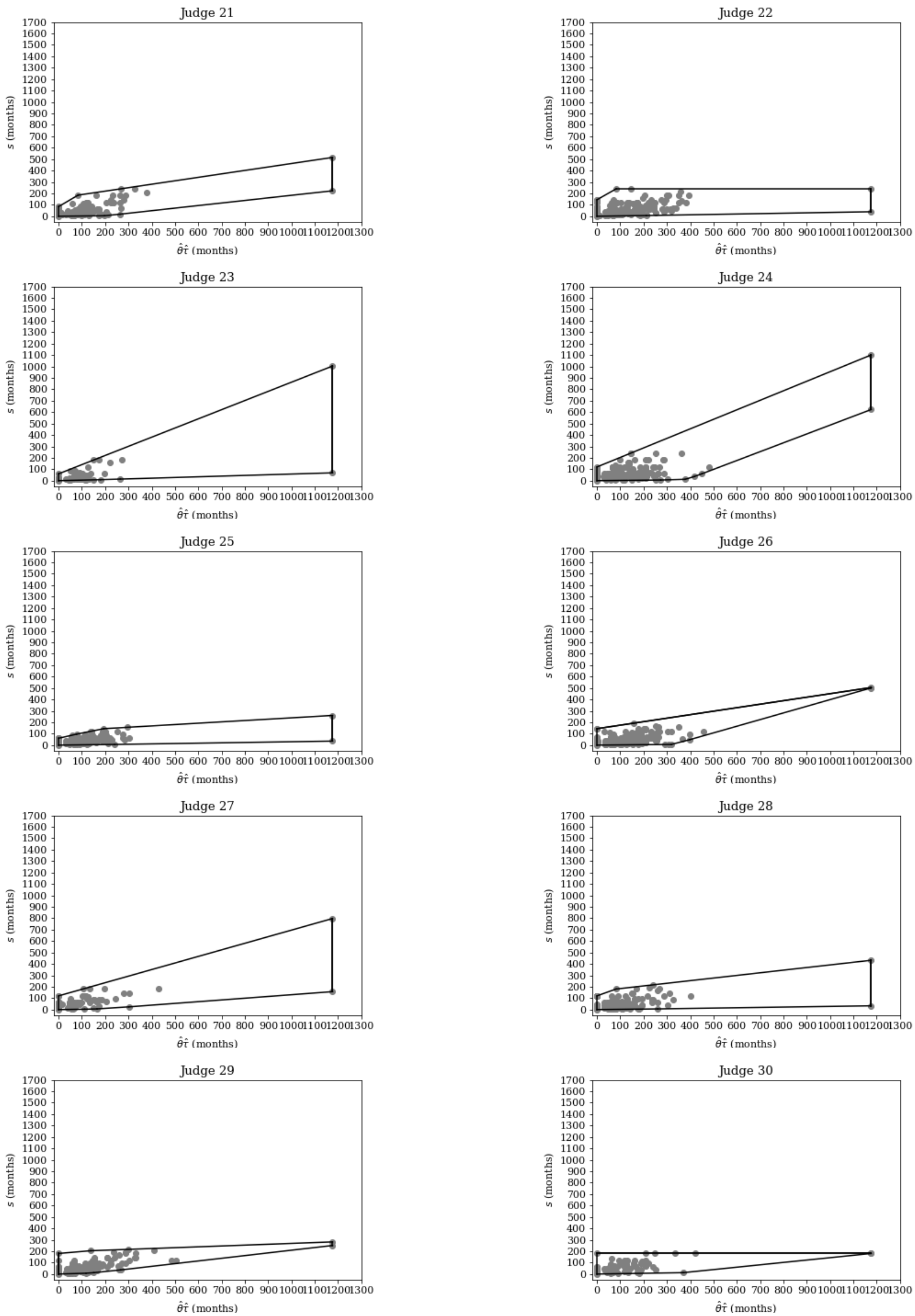


Figure 18 **(Continued) Judge convex hull with outliers removed, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial.**

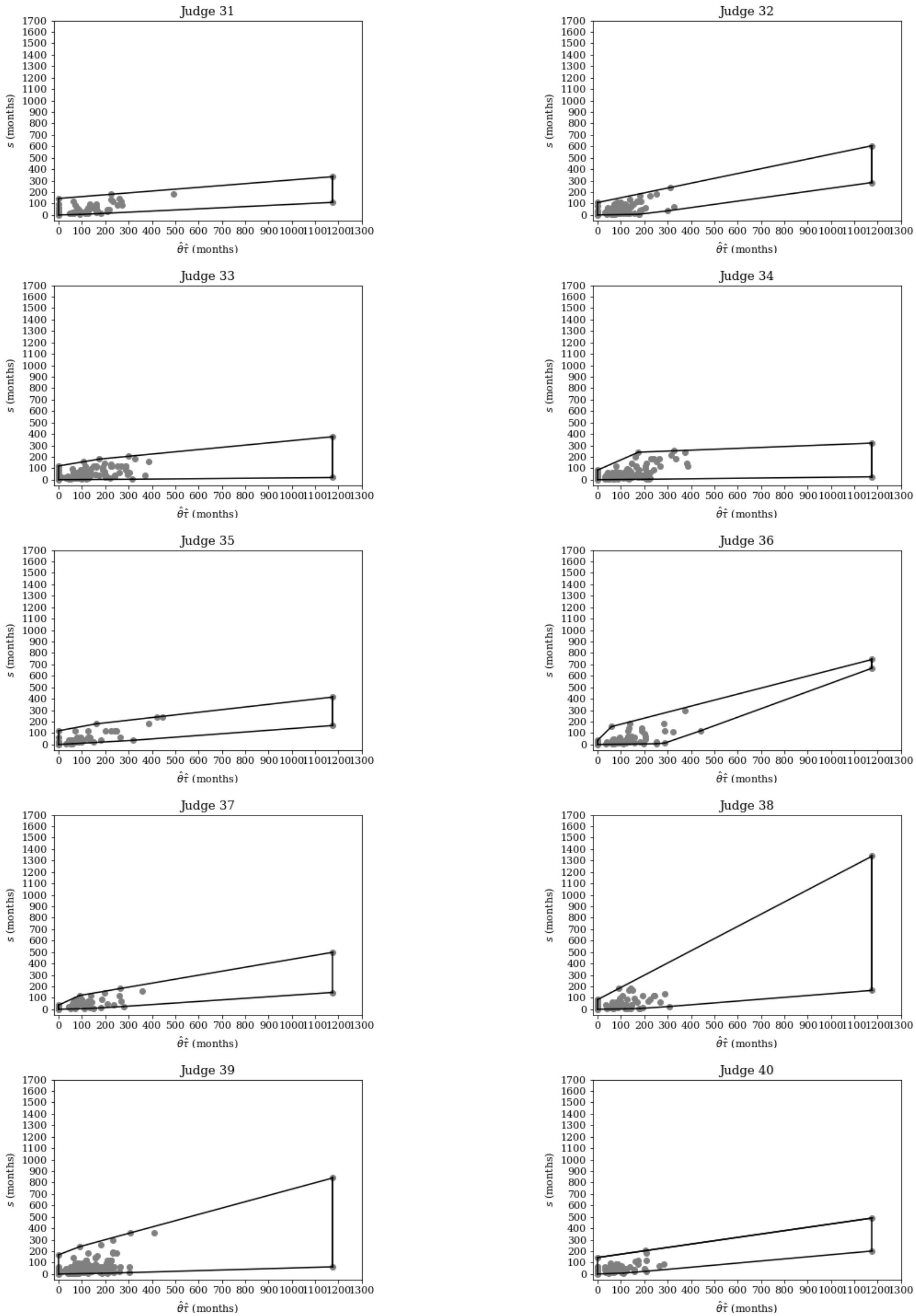


Figure 18 (Continued) Judge convex hull with outliers removed, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial.

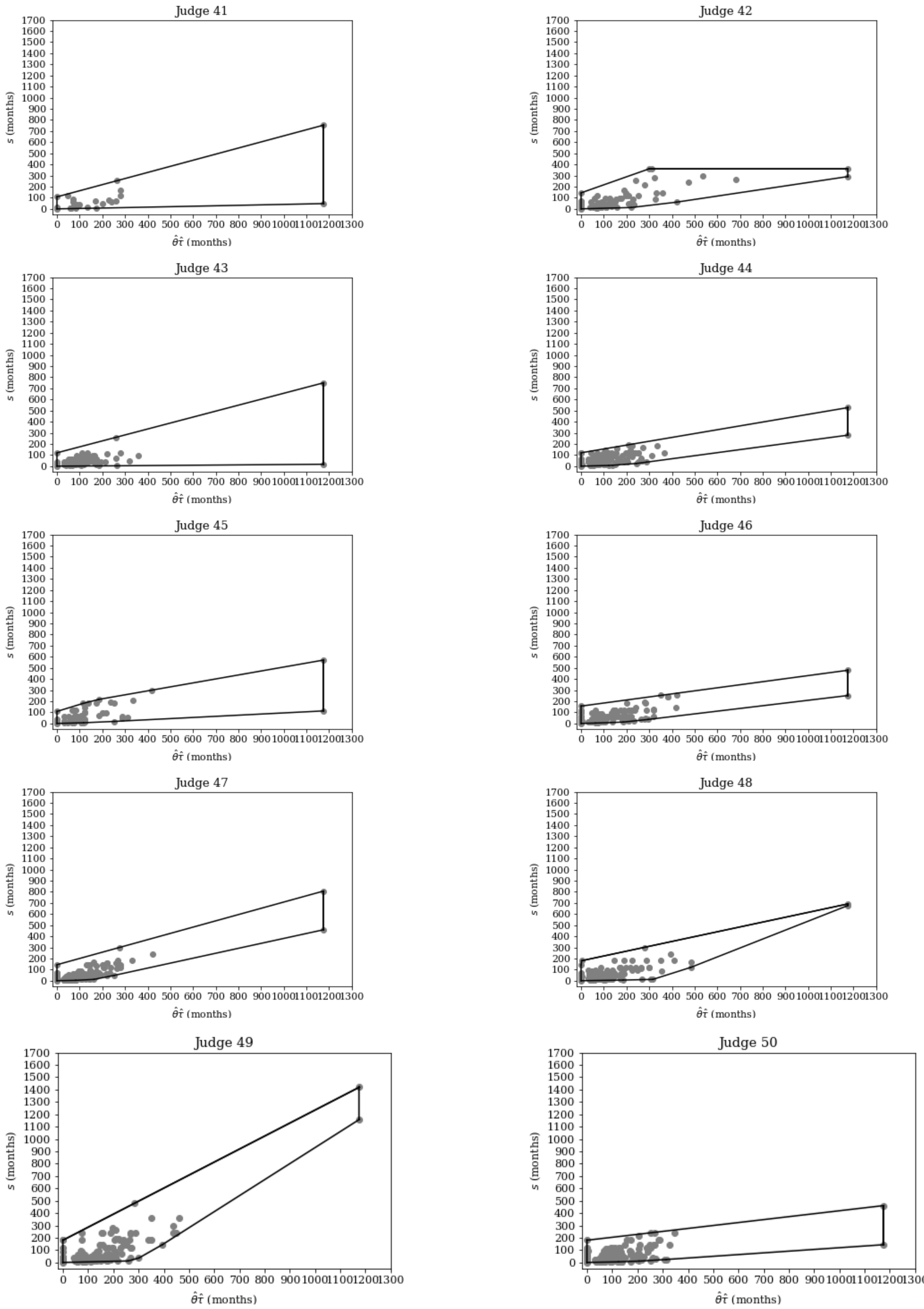


Figure 18 **(Continued) Judge convex hull with outliers removed, where s is the expected sentence at plea bargain and $\hat{\theta}\hat{\tau}$ is the expected sentence at trial.**

Table 9 Regression results to predict mean processing times

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3421	0.755	1.777	0.077	-0.147	2.831
C(County)[T.County 6]	-0.4733	0.977	-0.485	0.628	-2.398	1.451
C(County)[T.County 45]	-0.4664	1.152	-0.405	0.686	-2.738	1.805
C(County)[T.County 33]	-0.0304	0.954	-0.032	0.975	-1.911	1.850
C(County)[T.County 31]	0.2318	1.306	0.177	0.859	-2.343	2.806
C(County)[T.County 7]	-0.5174	1.152	-0.449	0.654	-2.787	1.753
C(County)[T.County 36]	0.3841	0.906	0.424	0.672	-1.402	2.170
C(County)[T.County 13]	-0.5385	0.946	-0.569	0.570	-2.403	1.326
C(County)[T.County 28]	-0.5686	1.152	-0.493	0.622	-2.840	1.703
C(County)[T.County 16]	0.6730	0.870	0.774	0.440	-1.041	2.387
C(County)[T.County 3]	-0.5891	1.069	-0.551	0.582	-2.697	1.519
C(County)[T.County 44]	1.5773	1.152	1.369	0.172	-0.693	3.847
C(County)[T.County 38]	-0.0326	1.012	-0.032	0.974	-2.027	1.962
C(County)[T.County 22]	0.1647	1.012	0.163	0.871	-1.829	2.158
C(County)[T.County 11]	-0.3337	0.945	-0.353	0.724	-2.197	1.530
C(County)[T.County 39]	0.1713	0.973	0.176	0.860	-1.747	2.090
C(County)[T.County 42]	0.1987	1.012	0.196	0.845	-1.796	2.193
C(County)[T.County 46]	-0.7331	1.074	-0.682	0.496	-2.851	1.385
C(County)[T.County 17]	0.3225	1.152	0.280	0.780	-1.949	2.594
C(County)[T.County 35]	-0.1665	1.068	-0.156	0.876	-2.271	1.938
C(County)[T.County 34]	0.6753	0.986	0.685	0.494	-1.269	2.620
C(County)[T.County 15]	-0.1267	1.013	-0.125	0.901	-2.124	1.870
C(County)[T.County 20]	0.3506	0.914	0.384	0.702	-1.451	2.153
C(County)[T.County 29]	-0.0127	0.975	-0.013	0.990	-1.934	1.909
C(County)[T.County 10]	-0.4797	1.067	-0.450	0.653	-2.583	1.623
C(County)[T.County 27]	-0.1582	0.913	-0.173	0.863	-1.958	1.641
C(County)[T.County 41]	-0.6377	1.152	-0.554	0.580	-2.908	1.632
C(County)[T.County 4]	-0.0218	0.946	-0.023	0.982	-1.886	1.843
C(County)[T.County 2]	0.9968	1.067	0.935	0.351	-1.105	3.099
C(County)[T.County 9]	0.2133	0.974	0.219	0.827	-1.706	2.133
C(County)[T.County 18]	-0.8087	1.012	-0.799	0.425	-2.803	1.186
C(County)[T.County 26]	-0.3238	0.894	-0.362	0.717	-2.085	1.437
C(County)[T.County 21]	-0.2242	1.067	-0.210	0.834	-2.327	1.878
C(County)[T.County 24]	1.1038	1.067	1.034	0.302	-0.999	3.207
C(County)[T.County 37]	-0.0203	1.152	-0.018	0.986	-2.291	2.251
C(County)[T.County 8]	-0.0319	1.012	-0.032	0.975	-2.026	1.962
C(County)[T.County 14]	-0.6984	0.974	-0.717	0.474	-2.617	1.220
C(County)[T.County 43]	0.2391	0.946	0.253	0.801	-1.625	2.103
C(County)[T.County 19]	-0.0200	0.974	-0.021	0.984	-1.939	1.899
C(County)[T.County 30]	1.1201	0.878	1.276	0.203	-0.610	2.850
C(County)[T.County 32]	-0.3813	1.152	-0.331	0.741	-2.651	1.889
C(County)[T.County 23]	-0.5872	0.942	-0.623	0.534	-2.443	1.269
C(County)[T.County 12]	-0.0360	0.946	-0.038	0.970	-1.900	1.828
C(County)[T.County 25]	0.0323	1.012	0.032	0.975	-1.961	2.026
C(County)[T.County 40]	0.0330	0.946	0.035	0.972	-1.831	1.897
Trial	0.8393	0.066	12.626	0.000	0.708	0.970
Plea	0.0178	0.002	10.795	0.000	0.015	0.021

Table 10 Regression results to predict defendant incarceration probability

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.4793	0.598	9.156	0.000	4.306	6.652
C(OffenseSeriousness)[T.five]	-1.0464	0.531	-1.969	0.049	-2.088	-0.005
C(OffenseSeriousness)[T.four]	-1.8369	0.485	-3.785	0.000	-2.788	-0.886
C(OffenseSeriousness)[T.one]	-3.8575	0.492	-7.840	0.000	-4.822	-2.893
C(OffenseSeriousness)[T.seven]	-0.1329	0.742	-0.179	0.858	-1.586	1.320
C(OffenseSeriousness)[T.six]	-0.6906	0.655	-1.054	0.292	-1.975	0.593
C(OffenseSeriousness)[T.three]	-2.9696	0.481	-6.179	0.000	-3.912	-2.028
C(OffenseSeriousness)[T.two]	-3.4596	0.483	-7.157	0.000	-4.407	-2.512
C(CommitmentScore)[T.eleven]	-0.1026	0.563	-0.182	0.855	-1.207	1.002
C(CommitmentScore)[T.five]	-0.0849	0.348	-0.244	0.807	-0.766	0.596
C(CommitmentScore)[T.four]	0.0748	0.326	0.229	0.819	-0.565	0.714
C(CommitmentScore)[T.nine]	0.0249	0.505	0.049	0.961	-0.965	1.014
C(CommitmentScore)[T.one]	-1.7656	0.310	-5.697	0.000	-2.373	-1.158
C(CommitmentScore)[T.seven]	-0.1487	0.401	-0.371	0.711	-0.934	0.637
C(CommitmentScore)[T.six]	0.2599	0.378	0.688	0.492	-0.481	1.001
C(CommitmentScore)[T.ten]	-0.7402	0.580	-1.277	0.202	-1.876	0.396
C(CommitmentScore)[T.three]	-0.3092	0.316	-0.977	0.329	-0.929	0.311
C(CommitmentScore)[T.twelve]	-0.3866	0.400	-0.966	0.334	-1.171	0.398
C(CommitmentScore)[T.two]	-0.8926	0.311	-2.868	0.004	-1.503	-0.283
OffenseType[T.Other]	-0.2961	0.072	-4.134	0.000	-0.436	-0.156
OffenseType[T.Property]	-0.6523	0.059	-11.042	0.000	-0.768	-0.536
OffenseType[T.Violent]	-0.1157	0.088	-1.318	0.187	-0.288	0.056
CrimHist[T.minimal]	-3.0625	0.172	-17.769	0.000	-3.400	-2.725
CrimHist[T.moderate]	-0.4687	0.177	-2.645	0.008	-0.816	-0.121
CrimHist[T.none]	-3.4818	0.175	-19.930	0.000	-3.824	-3.139
CrimHist[T.voluminous]	-0.1359	0.202	-0.672	0.502	-0.532	0.261
MandatoryMinimum	3.5936	0.423	8.496	0.000	2.765	4.423
Male	0.5332	0.065	8.221	0.000	0.406	0.660
Black	0.5576	1.168	0.477	0.633	-1.732	2.848
Black:C(OffenseSeriousness)[T.five]	0.0210	1.198	0.017	0.986	-2.328	2.370
Black:C(OffenseSeriousness)[T.four]	-0.8221	1.158	-0.710	0.478	-3.092	1.447
Black:C(OffenseSeriousness)[T.one]	-1.2033	1.161	-1.037	0.300	-3.479	1.072
Black:C(OffenseSeriousness)[T.seven]	-1.8279	1.325	-1.380	0.168	-4.424	0.768
Black:C(OffenseSeriousness)[T.six]	-0.7504	1.384	-0.542	0.588	-3.464	1.963
Black:C(OffenseSeriousness)[T.three]	-0.6751	1.156	-0.584	0.559	-2.941	1.591
Black:C(OffenseSeriousness)[T.two]	-0.4628	1.155	-0.401	0.689	-2.726	1.800
Black:CrimHist[T.minimal]	0.6749	0.204	3.303	0.001	0.274	1.075
Black:CrimHist[T.moderate]	0.1766	0.211	0.835	0.404	-0.238	0.591
Black:CrimHist[T.none]	0.6827	0.207	3.298	0.001	0.277	1.088
Black:CrimHist[T.voluminous]	0.0558	0.243	0.230	0.818	-0.420	0.532

Table 11 Regression results to predict defendant probability of conviction at trials

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.4472	1.99e+06	-7.29e-07	1.000	-3.89e+06	3.89e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.five]	-260.9821	1.4e+06	-0.000	1.000	-2.75e+06	2.75e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.four]	-225.9694	1.83e+06	-0.000	1.000	-3.59e+06	3.59e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.one]	-192.6918	1.01e+06	-0.000	1.000	-1.98e+06	1.98e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.seven]	31.9071	9.67e+05	3.3e-05	1.000	-1.89e+06	1.89e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.six]	-135.3771	7e+05	-0.000	1.000	-1.37e+06	1.37e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.three]	-83.3723	1.1e+06	-7.61e-05	1.000	-2.15e+06	2.15e+06
C(OffenseSeriousness, levels=all_offense_seriousness)[T.two]	-248.9797	1.91e+06	-0.000	1.000	-3.75e+06	3.75e+06
C(CommitmentScore, levels=all_commitment_score)[T.eleven]	-0.4994	5.81e+05	-8.6e-07	1.000	-1.14e+06	1.14e+06
C(CommitmentScore, levels=all_commitment_score)[T.five]	-11.5056	6.15e+05	-1.87e-05	1.000	-1.21e+06	1.21e+06
C(CommitmentScore, levels=all_commitment_score)[T.four]	1.8214	4.08e+05	4.47e-06	1.000	-7.99e+05	7.99e+05
C(CommitmentScore, levels=all_commitment_score)[T.nine]	44.5600	1.03e+06	4.31e-05	1.000	-2.03e+06	2.03e+06
C(CommitmentScore, levels=all_commitment_score)[T.one]	-1.8799	3.22e+05	-5.84e-06	1.000	-6.31e+05	6.31e+05
C(CommitmentScore, levels=all_commitment_score)[T.seven]	-53.9157	9.81e+05	-5.49e-05	1.000	-1.92e+06	1.92e+06
C(CommitmentScore, levels=all_commitment_score)[T.six]	-0.2755	3.46e+05	-7.97e-07	1.000	-6.78e+05	6.78e+05
C(CommitmentScore, levels=all_commitment_score)[T.ten]	2.405e-13	6.36e-09	3.78e-05	1.000	-1.25e-08	1.25e-08
C(CommitmentScore, levels=all_commitment_score)[T.three]	12.2886	9.25e+05	1.33e-05	1.000	-1.81e+06	1.81e+06
C(CommitmentScore, levels=all_commitment_score)[T.twelve]	-0.5905	3.28e+05	-1.8e-06	1.000	-6.43e+05	6.43e+05
C(CommitmentScore, levels=all_commitment_score)[T.two]	19.3078	3.23e+05	5.98e-05	1.000	-6.33e+05	6.33e+05
OffenseType[T.Other]	-42.8970	3.96e+04	-0.001	0.999	-7.77e+04	7.77e+04
OffenseType[T.Property]	89.0265	7.55e+04	0.001	0.999	-1.48e+05	1.48e+05
OffenseType[T.Violent]	23.2533	3.92e+04	0.001	1.000	-7.69e+04	7.69e+04
CrimHist[T.minimal]	97.0734	9.23e+05	0.000	1.000	-1.81e+06	1.81e+06
CrimHist[T.moderate]	100.3873	9.38e+05	0.000	1.000	-1.84e+06	1.84e+06
CrimHist[T.none]	91.8205	8.6e+05	0.000	1.000	-1.68e+06	1.68e+06
CrimHist[T.voluminous]	121.0368	7.33e+05	0.000	1.000	-1.44e+06	1.44e+06
MandatoryMinimum	-86.0137	1.76e+06	-4.9e-05	1.000	-3.44e+06	3.44e+06
Male	176.9426	1.14e+05	0.002	0.999	-2.23e+05	2.24e+05
Black	-32.4115	1.02e+06	-3.17e-05	1.000	-2e+06	2e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.five]	353.5097	7.09e+05	0.000	1.000	-1.39e+06	1.39e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.four]	71.9624	5.33e+05	0.000	1.000	-1.05e+06	1.05e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.one]	-192.6918	1.01e+06	-0.000	1.000	-1.98e+06	1.98e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.seven]	68.8902	5.19e+05	0.000	1.000	-1.02e+06	1.02e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.six]	157.8687	9.11e+05	0.000	1.000	-1.78e+06	1.79e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.three]	9.5656	8.84e+05	1.08e-05	1.000	-1.73e+06	1.73e+06
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.two]	161.8671	3.86e+05	0.000	1.000	-7.56e+05	7.56e+05
Black:CrimHist[T.minimal]	29.5136	1.03e+06	2.87e-05	1.000	-2.01e+06	2.01e+06
Black:CrimHist[T.moderate]	-100.0981	1.3e+06	-7.7e-05	1.000	-2.55e+06	2.55e+06
Black:CrimHist[T.none]	-145.9120	1.01e+06	-0.000	1.000	-1.98e+06	1.98e+06
Black:CrimHist[T.voluminous]	-109.9934	9.05e+05	-0.000	1.000	-1.77e+06	1.77e+06

Table 12 Regression results to predict defendant sentence upon conviction at trials

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.3967	0.285	22.422	0.000	5.838	6.956
C(OffenseSeriousness, levels=all_offense_seriousness)[T.five]	-0.6401	0.197	-3.250	0.001	-1.026	-0.254
C(OffenseSeriousness, levels=all_offense_seriousness)[T.four]	-0.9893	0.192	-5.162	0.000	-1.365	-0.614
C(OffenseSeriousness, levels=all_offense_seriousness)[T.one]	3.092e-15	8.38e-16	3.688	0.000	1.45e-15	4.73e-15
C(OffenseSeriousness, levels=all_offense_seriousness)[T.seven]	-0.4552	0.121	-3.766	0.000	-0.692	-0.218
C(OffenseSeriousness, levels=all_offense_seriousness)[T.six]	-0.3651	0.264	-1.381	0.167	-0.883	0.153
C(OffenseSeriousness, levels=all_offense_seriousness)[T.three]	-1.5581	0.158	-9.843	0.000	-1.868	-1.248
C(OffenseSeriousness, levels=all_offense_seriousness)[T.two]	-1.8983	0.212	-8.955	0.000	-2.314	-1.483
C(CommitmentScore, levels=all_commitment_score)[T.eleven]	0.4391	0.236	1.858	0.063	-0.024	0.902
C(CommitmentScore, levels=all_commitment_score)[T.five]	-0.0753	0.203	-0.371	0.710	-0.473	0.322
C(CommitmentScore, levels=all_commitment_score)[T.four]	0.3451	0.185	1.866	0.062	-0.017	0.708
C(CommitmentScore, levels=all_commitment_score)[T.nine]	0.6666	0.440	1.514	0.130	-0.196	1.529
C(CommitmentScore, levels=all_commitment_score)[T.one]	-0.1567	0.144	-1.092	0.275	-0.438	0.125
C(CommitmentScore, levels=all_commitment_score)[T.seven]	0.2551	0.171	1.491	0.136	-0.080	0.590
C(CommitmentScore, levels=all_commitment_score)[T.six]	0.1648	0.153	1.076	0.282	-0.135	0.465
C(CommitmentScore, levels=all_commitment_score)[T.ten]	-4.851e-16	3.18e-16	-1.527	0.127	-1.11e-15	1.37e-16
C(CommitmentScore, levels=all_commitment_score)[T.three]	-0.1046	0.153	-0.682	0.495	-0.405	0.196
C(CommitmentScore, levels=all_commitment_score)[T.twelve]	0.3219	0.152	2.122	0.034	0.025	0.619
C(CommitmentScore, levels=all_commitment_score)[T.two]	0.0281	0.148	0.190	0.849	-0.262	0.318
OffenseType[T.Other]	-0.5583	0.141	-3.949	0.000	-0.835	-0.281
OffenseType[T.Property]	-0.1477	0.084	-1.763	0.078	-0.312	0.016
OffenseType[T.Violent]	0.3120	0.066	4.748	0.000	0.183	0.441
CrimHist[T.minimal]	-0.6490	0.230	-2.822	0.005	-1.100	-0.198
CrimHist[T.moderate]	-0.6713	0.217	-3.095	0.002	-1.096	-0.246
CrimHist[T.none]	-0.4947	0.213	-2.319	0.020	-0.913	-0.077
CrimHist[T.voluminous]	-0.3147	0.235	-1.342	0.180	-0.774	0.145
MandatoryMinimum	-0.1338	0.080	-1.681	0.093	-0.290	0.022
Male	0.1590	0.095	1.674	0.094	-0.027	0.345
Black	-0.3301	0.219	-1.510	0.131	-0.759	0.098
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.five]	-0.6001	0.211	-2.837	0.005	-1.015	-0.186
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.four]	-0.0223	0.216	-0.103	0.918	-0.446	0.401
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.one]	3.109e-16	1.67e-16	1.866	0.062	-1.57e-17	6.38e-16
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.seven]	-0.2301	0.139	-1.653	0.098	-0.503	0.043
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.six]	-0.5742	0.368	-1.561	0.119	-1.295	0.147
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.three]	-0.0195	0.168	-0.116	0.908	-0.350	0.311
Black:C(OffenseSeriousness, levels=all_offense_seriousness)[T.two]	0.1410	0.215	0.657	0.511	-0.279	0.562
Black:CrimHist[T.minimal]	0.4260	0.245	1.742	0.081	-0.053	0.905
Black:CrimHist[T.moderate]	0.5853	0.231	2.536	0.011	0.133	1.038
Black:CrimHist[T.none]	0.3383	0.229	1.480	0.139	-0.110	0.786
Black:CrimHist[T.voluminous]	0.2712	0.252	1.076	0.282	-0.223	0.765

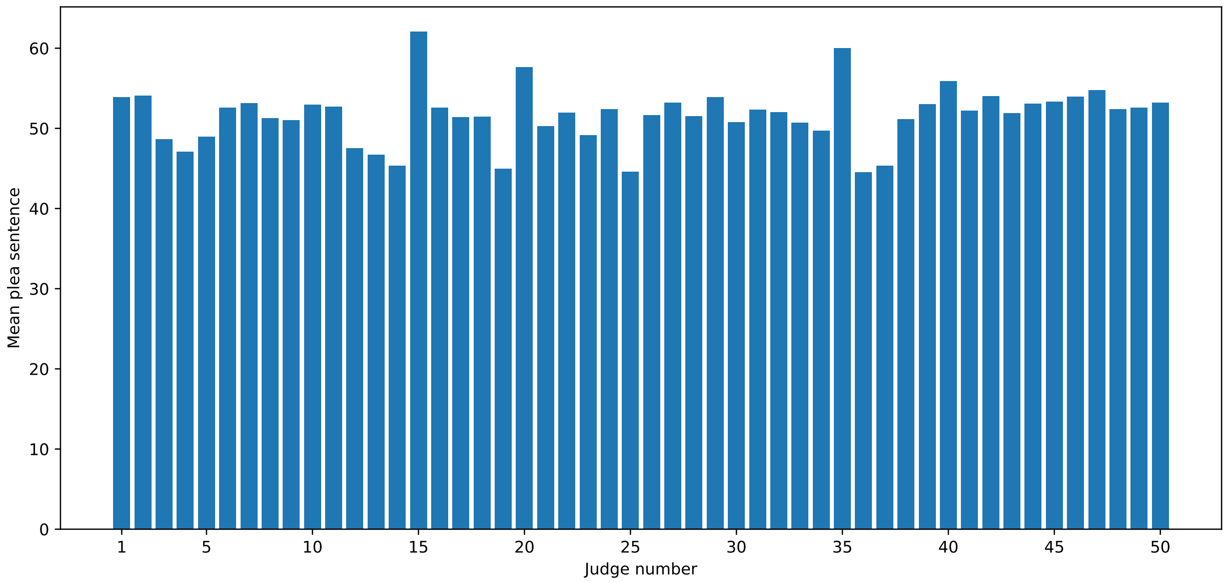
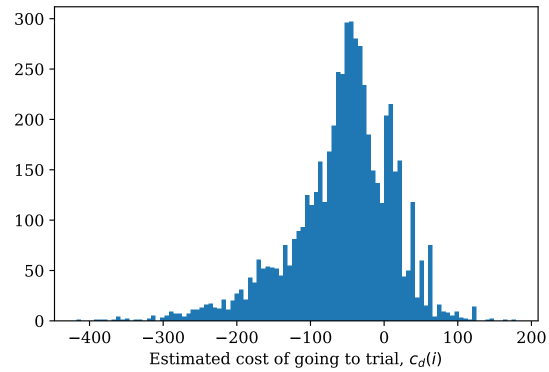
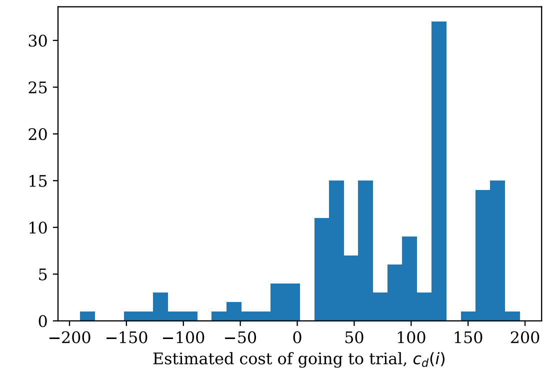


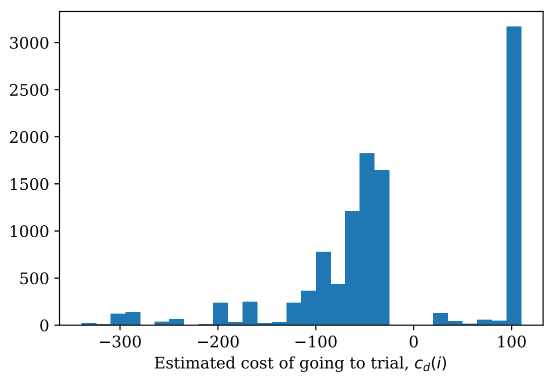
Figure 19 **Mean plea sentence length decided by each judge based on all 17,516 defendants from the sentencing dataset and the corresponding convex hull.**



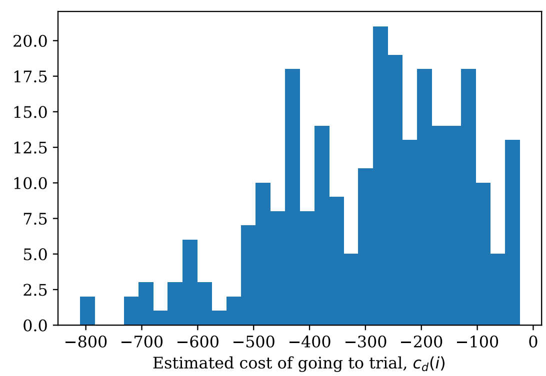
(a)



(b)

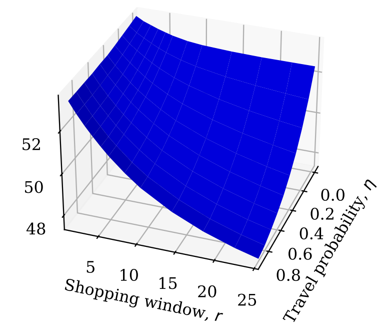


(c)

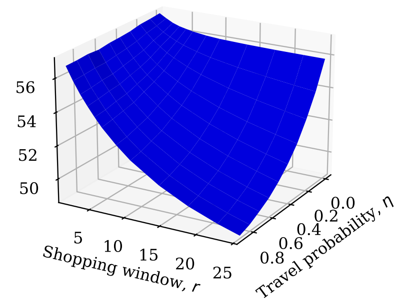


(d)

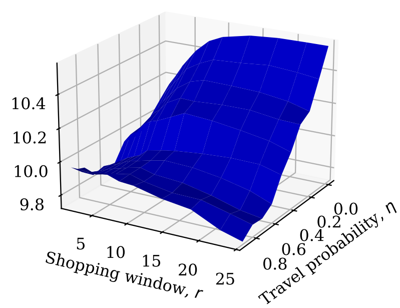
Figure 20 **Histogram of the estimated cost (in months) of going to trial, $c_d(i)$, for (a) $i \in \mathcal{I}^1$ (5731 cases, mean = -56.4, standard deviation = 69.5), (b) $i \in \mathcal{I}^2$ (153 cases, mean = 78.1, standard deviation = 78.3), (c) $i \in \mathcal{I}^3$ (10,920 cases, mean = -22.4, standard deviation = 101.6), and (d) $i \in \mathcal{I}^4$ (258 cases, mean = -292.8, standard deviation = 169.3).**



(a) Mean plea sentence

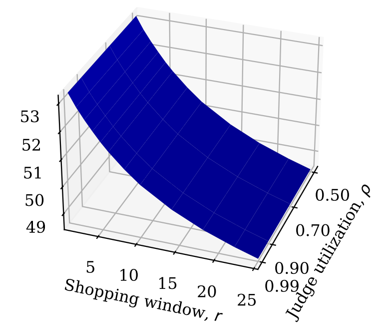


(b) Standard deviation of plea sentence

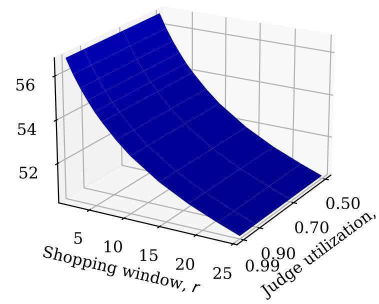


(c) Standard deviation across counties

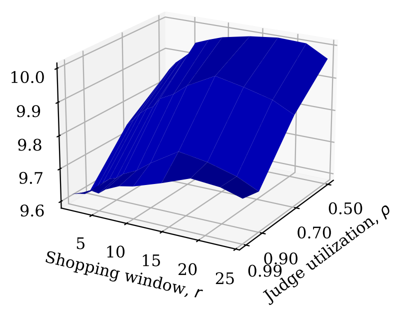
Figure 21 **Performance measures versus the shopping window** $r \in \{1, 2, 3, 4, 5, 6, 8, 10, 14, 18, 22, 25\}$ **weeks and travel probability** $\eta \in \{0, 0.1, \ldots, 0.9\}$**, with a fixed judge utilization** $\rho = 0.5$**.**



(a) Mean plea sentence

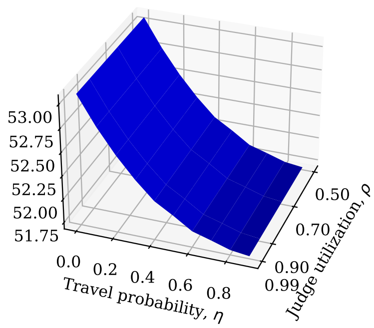


(b) Standard deviation of plea sentence

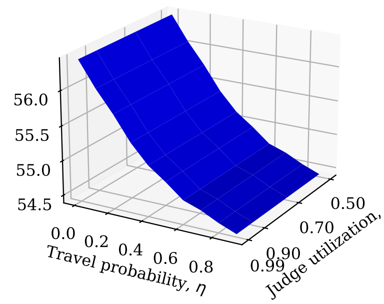


(c) Standard deviation across counties

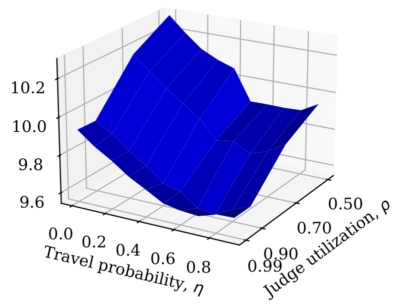
Figure 22 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks and the judge utilization** $\rho \in \{0.5,0.7,0.9,0.99\}$**, with a fixed travel probability** $\eta = 0.5$**.**



(a) Mean plea sentence

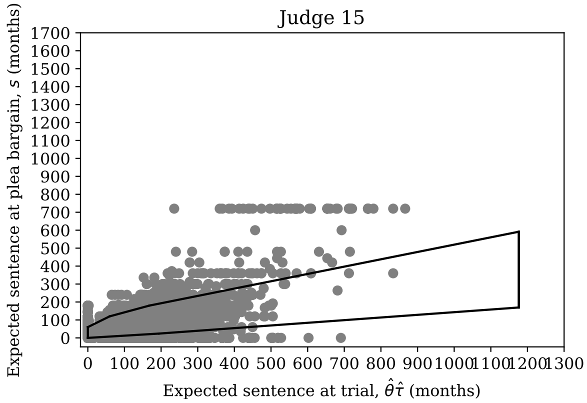


(b) Standard deviation of plea sentence

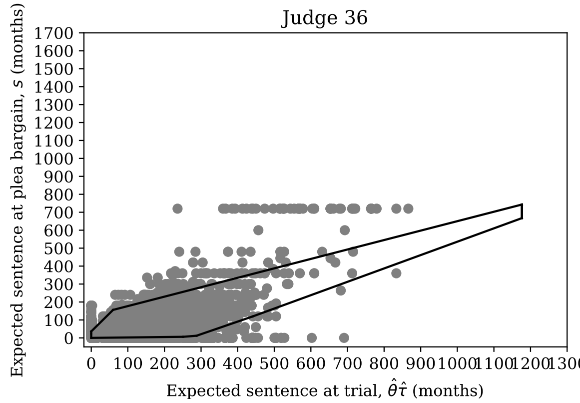


(c) Standard deviation across counties

Figure 23 Performance measures versus the travel probability $\eta \in \{0, 0.1, \ldots, 0.9\}$ **and the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$**, with a fixed shopping window** $r = 4$ **weeks.**

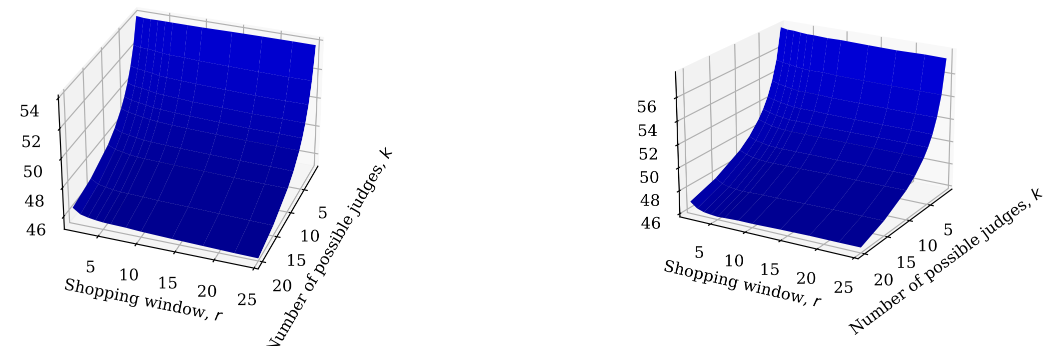


(a) Judge 15



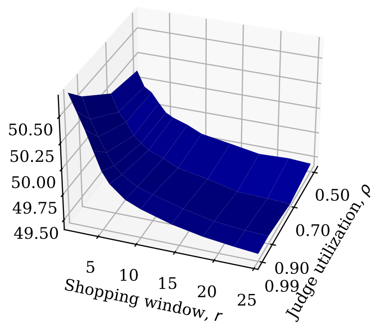
(b) Judge 36

Figure 24 **The defendants characterized by $(\hat{\theta}_i\hat{\tau}_i, s_i)$ and the constructed convex hulls for Judges 15 and 36.**

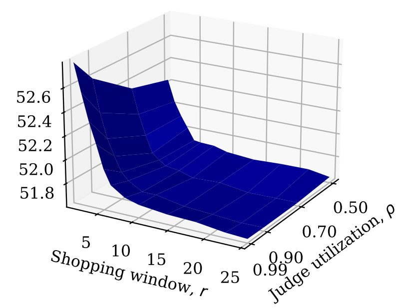


(a) Mean plea sentence (b) Standard deviation of plea sentence

Figure 25 **Performance measures versus the shopping window** $r \in \{1,2,3,4,5,6,8,10,14,18,22,25\}$ **weeks and the number of judges** $k \in \{1,2,3,4,5,6,8,10,15,20\}$ **in the urban model, with the judge utilization fixed at** $\rho = 0.5$**.**

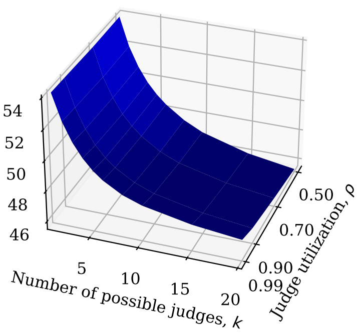


(a) Average plea sentence

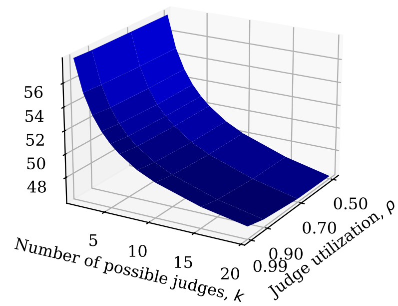


(b) Standard deviation of plea sentence

Figure 26 **Performance measures versus the shopping window** $r \in \{1, 2, 3, 4, 5, 6, 8, 10, 14, 18, 22, 25\}$ **weeks and the judge utilization** $\rho \in \{0.5, 0.7, 0.9, 0.99\}$ **in the urban model, with the number of judges fixed at** $k = 4$**.**



(a) Average plea sentence



(b) Standard deviation of plea sentence

Figure 27 **Performance measures versus the number of judges** $k \in \{1,2,3,4,5,6,8,10,15,20\}$ **and the judge utilization** $\rho \in \{0.5,0.7,0.9,0.99\}$ **in the urban model, with the shopping window fixed at** $r = 4$ **weeks.**