Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN | EISSN | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Reducing Judicial Delay in Resource-Constrained Settings: A Data-Driven Queueing Approach

Shortage of judicial capacity leads to costly delays, stunted economic development, and even failure to deliver justice. This problem is endemic not only in the developing world but also in the congested appeals courts of wealthier nations. Using the Supreme Court of India as an exemplar for such resource-constrained settings, we develop a framework with *data-driven queueing simulations* for estimating performance metrics such as the expected case-disposition time (delay) and expected number of cases awaiting adjudication (pendency). This allows us to not only calibrate the status quo for judicial performance but to also perform counterfactual analysis that evaluates the relative effectiveness of various interventions such as additional judges, process re-engineering, and workload management. We find that the Supreme Court of India operates in a nearly *critically-loaded* regime (nearly 100% utilization of capacity) that is characterized by substantial delays, and small perturbations to capacity or process efficiency have dramatic impact on system performance. In particular, increasing judge capacity by 7% (adding a bench) results in a $75\% - 90\%$ reduction in average delay. Alternatively, capping the number of adjournments allowed in a case to the recommended number three, also results in a comparable delay reduction. Our findings bode well for the ability to tackle persistent delay, but scrutiny of the court's workload-management practices points to the possibility of perpetually languishing in congestion unless more effective winnowing of pent-up demand (e.g., accepting fewer appeals) is ensured.

Key words: judicial delay | case-management queues | data-driven simulation
History: January 2, 2021

1. Introduction

India has a population of about 1.37 billion people (The World Bank 2019). The current backlog of cases in the Indian judicial system stands at nearly 38.5 million. About a quarter of these cases have been pending for more than five years (The Government of India 2020). Such judicial delays have adverse economic and social impact. For instance, in terms of "enforcing contracts" India is ranked 163 out of 190 nations (The World Bank 2020), and about 70% of India's prison population comprises pre-trial detainees (World Prison Brief 2020).

Despite the workload, only about $21,000$ judge positions have been sanctioned to address the litigation that arises (Financial Times 2016). This translates to less than 16 judges per million people in contrast to the more aspirational figure of 50 judges per million (The Supreme Court of India 2012). Moreover, about $25\% - 33\%$ of India's sanctioned judicial strength remains unfilled

(Singhvi 2017). While adding judges is one possible response to the situation, the judiciary itself has proposed multiple ways of tackling the problem of judicial delays, including improving process/procedural efficiency and using techniques for workload management (e.g., alternative dispute resolution mechanisms) (The Supreme Court of India 2018). The multiple options notwithstanding, progress in addressing judicial delays has been painfully slow. In a developing country such as India, there are numerous competing demands on the limited financial resources of the nation. Hence, unless a scientific and compelling argument is made for the proposed remedy, it is likely that pleas for assistance will fall on deaf ears.

In this paper, we develop a *queueing-theoretic* simulation framework which we then calibrate with data to evaluate the relative effectiveness of interventions for managing judicial delay. We also contrast our framework with those currently in use to highlight opportunities for improvement.

While the problem of congestion is much more severe in India's lower judiciary, we apply our framework and demonstrate its utility in the context of the Supreme Court of India (SCI). The reason to do so is twofold. First, the quality of data for courts other than the Supreme Court is not as good (The Supreme Court of India 2012). Second, in addition to hearing constitutional matters, the SCI also serves as a court of appeal. Therefore, it displays characteristics (high congestion levels) similar to other appeals courts even in developed nations (Bray et al. 2016, Green and Yoon 2017). Thus, the SCI provides an ideal testbed for our framework.

1.1. Judicial Delay and Management Science

Justice delayed is justice denied, is an age-old adage that is accompanied by periodic acknowledgement of the severity of the problem, particularly in appeals courts (Carrington 1969, Meador 1974, Adler 2014). The remedies are broadly classified into two categories: boosting infrastructure, and eliminating process inefficiency (e.g., Castro and Guccio 2015). The experience in India has been qualitatively similar albeit distinct in practice and procedure (Law Commission of India 1987, Malimath 2003, Law Commission of India 2014, The Supreme Court of India 2018). It was recognized a while back that the tools and concepts from Management Science, in combination with data, have the potential to make a big dent in addressing this critical societal challenge (Blumstein and Larson 1969, Nagel et al. 1978). However, notable contributions in this regard have been scarce. One recent exception is Bray et al. (2016), which investigates improved scheduling of cases within the Italian labor court of appeals. Specifically, the paper studies when the scheduling logic of *oldest-hearing-first* versus *oldest-case-first* performs better from the perspective of a judge who maximizes their long-term "reward." Although we model oldest-hearing-first, scheduling interventions will be of limited value in the Indian context until the challenge of profligate *adjournments* is addressed, which relates to the unexpected rescheduling of a hearing due to unavailability of participants. Instead, our focus is on using a queueing framework to quantify the impact of interventions in judicial capacity and controlling adjournments on metrics such as delay and pendency.

2. The Supreme Court of India (SCI)

The highest court in India's judicial hierarchy is its Supreme Court. It was established in 1950 with eight justices. Currently it has a sanctioned strength of 34 judges but only 31 sitting judges, including the Chief Justice of India. The SCI has original jurisdiction on matters such as protecting human rights; appellate jurisdiction over decisions of lower courts that are appealed; and advisory jurisdiction on matters referred to it by the President of India (The Supreme Court of India 2020). Appeals dwarf all other forms of workload, and this makes the SCI distinct from other institutions such as the U.S. Supreme Court which restricts itself to constitutional matters and questions of extraordinary legal importance. Moreover, the nine justices of the U.S. court decide all matters together, or *en banc*. In contrast, to deal with its immense workload, the SCI adjudicates matters mainly through two-judge benches. A few exceptional cases are presided over by three-judge benches, while constitutional matters are heard by five-judge benches (Robinson 2013a). Each judge is meant to hear matters on every working weekday for 4.5 hours (The Supreme Court of India 2017). As a result, the backlog of cases (pendency) at the SCI seems to have been relatively stable in the recent past, hovering at around $60,000$ cases (The Supreme Court of India 2019).

3. Model

It would be futile to try to precisely model every detail of how the SCI functions in our paper. Instead, we capture the salient logistical aspects of SCI's work in our model, as described next. (Some additional assumptions are highlighted in Appendix B, Assumptions in Simulation Model.)

A case that comes to the SCI typically has multiple hearings before it is resolved. In any given year, many thousands of cases arrive at the SCI and require scarce judicial time, which results in congestion. Queueing theory is the mathematical discipline that studies congested systems. Specifically, we model a *case-management queue* wherein a case can have multiple hearings across time but always with the same panel of judges — similar to a patient–doctor relationship in a medical setting. We simulate the performance of a case management system and use it to study judicial delays at the SCI.[1] We calibrate our model using "case status" data downloaded from the SCI website (The Supreme Court of India 2020). We downloaded the details for cases that arrived between 2009 and 2016. Table 1 summarizes the arrival information between 2009 and 2016 based on the downloaded data.[2] Our data also tells us when a case was registered, the date for each

[1] Simulation is the appropriate tool for analysis here because even the simplest variant of a case-management queue is known to be analytically intractable (Campello et al. 2017). Our setting is even more complex than the canonical queueing settings for multiple reasons, e.g., the arrival process is batched, the service time is not exponential, and judges take vacations.

[2] We do not use the arrival information documented in the annual reports of the SCI; those numbers are tailored to the idiosyncratic accounting practices of the SCI which can lead to double counting of certain types of cases, and also suffer from other shortcomings that have been documented before (Robinson 2013a).

Table 1 Arrival of new cases to the SCI

Year	New cases registered
2009	39,811
2010	38,625
2011	38,032
2012	39,831
2013	39,834
2014	40,298
2015	39,130
2016	40,135
Total	315,696

hearing that it had, and the court's written orders associated with each hearing. Three elements that form the foundation for our model of the case-management queue are:

1. **The arrival process**. The details of any new case are recorded by the registrar's office first. This requires complete and accurate documentation, hence, hearings practically cannot be scheduled on the same day that the case is registered. This results in *batched* arrival of cases. Using the arrival information for cases from 2009 to 2016 we create an empirical distribution for the daily arrival count. We uniformly sample from this distribution to simulate the number of arrivals on any given day; the arrival sequence of cases is maintained as in the data. For long simulation runs that require more observations than can be supported by the arrival stream from 2009 to 2016, we repeat the sequence of arrivals as many times as needed.
2. **The service process**. Work done to dispose off a case is referred to as service. It can be spread across multiple service encounters or hearings. Cases are heard by an assigned bench or panel of two supreme court justices until the matter is resolved.[3] We also model 14 such homogeneous benches since the strength of the Supreme Court was 28 when we initiated this project in mid-2017. Each bench maintains a dedicated queue of cases that are assigned to it. A new case is assigned to a judicial bench either because it has the shortest queue-size when the case arrives, or because it is connected to another case involving similar legal considerations, which has already been assigned to the panel.[4]

   A critical requirement for simulating the service process is the distribution of time spent on an individual hearing. To estimate this distribution, we collected actual hearing times for cases heard by the SCI as shown in real time by the online display-board on the court's website. We collected this data for seven months, from November 12, 2018 to May 10, 2019; which resulted in 15,725 observations. We identify the distribution which gives us the best fit with

[3] Consistent with the bulk of the workload at the SCI, we model only two-judge benches.

[4] The information about connected matters is available in our data. In practice, certain cases may be assigned to specialist benches (e.g., income tax or environment related) or per other considerations of the Chief Justice. However, such assignments are the exception not the norm. Hence, we do not try to capture them.

data (details in Appendix A, Estimating the Hearing-Time Distribution). Hearing times can vary a lot depending on the nature of the proceedings. Litigants might share a new document which warrants careful offline study and preparation from the counter party – this would make for a short hearing that lasts a few minutes. At other hearings, lawyers may present lengthy arguments running into hours. Our statistical analysis yields an expected hearing time of 461 seconds.[5] Sometimes a scheduled hearing cannot be conducted and has to be adjourned, i.e., rescheduled for another day; we model the time spent on granting an adjournment to be a deterministic 100 seconds (this estimate is based on the authors' actual observation of court proceedings, and corroborated through expert input).

3. **The hearing outcome**. There are three possible outcomes for a hearing. If the judges have arrived at the final decision after a hearing then the case is said to be *disposed* and it leaves the system. If the hearing is inconclusive and the case needs more judicial time before a decision can be reached, then it is referred to as a *regular* hearing. Scheduled hearings may get *adjourned* on occasion if the litigants, lawyers or judges are unavailable. After either a regular hearing or an adjournment, we assume that a case is sent to the back of the queue to await its next hearing. (We refer the reader to Appendix B, Assumptions in the Simulation Model, for additional details.) Figure 1 provides a high-level schematic for the life cycle of a case in the SCI. The outcome of a hearing is not recorded explicitly anywhere in the data. Fortunately, we have access to the written court orders issued at the end of every hearing. We conducted keyword-based text analysis of the court order files to classify the outcomes (details in Appendix D, Classification of Judicial Orders). We have made the simplifying but practical assumption that the outcome of a specific hearing is independent of the history of the case (see §5 for a discussion on robustness). Thus, for any hearing, we estimate the probability of disposal to be 0.29, the probability that the hearing is inconclusive is 0.28, and the probability that the case is adjourned is 0.43. To limit the bias due to pending cases (e.g., underestimating disposal probability), we based these estimates on the hearing outcomes for cases that were registered in 2012.[6] We use these numbers to simulate hearing outcomes in our model.

An important advantage of using data to estimate the hearing time distribution and the hearing outcome probabilities is that it minimizes the impact of different scenarios (e.g., in counterfactual analysis) on the legal considerations in judicial decision making.

[5] Another factor contributing to short hearings is that the SCI grants audience to any case filed before it as an "admissions matter." Only a subset of these matters are accepted for further hearing as "regular matters." The display-board (or even "case status") data does not distinguish between admission and regular matters, thus precluding further analysis along this dimension.

[6] The corresponding probability estimates were similar for cases registered in 2010, 2011, 2013, and 2014.

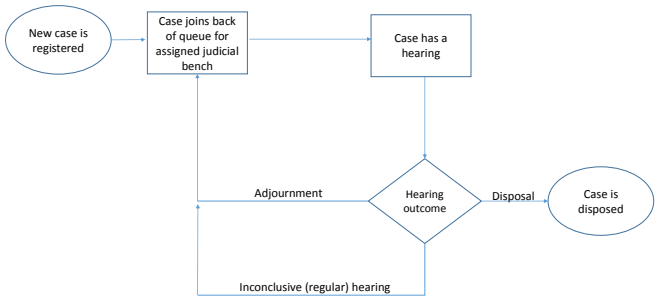

Figure 1 Life cycle of a case at the SCI.

4. Results

4.1. Theoretical foundations

When analyzing queueing systems, the expected delay (disposition time) experienced by the cases is a natural metric to consider. Another metric of interest is the expected backlog of pending cases (pendency). These two metrics of delay and pendency are related through *Little's law*: expected backlogs = expected arrival rate × expected delay. Formally, one is interested in the long-run average measure of these quantities, assuming the system is stable and ergodic (Wolff 1989). Intuitively, if cases arrive at a higher rate than can be processed, the system would be unstable and the backlog of cases would continue increasing over time. This notion is captured via the (theoretical) *utilization* of the system, denoted by ρ, which measures the ratio of system demand to the system capacity. A situation of $\rho \geq 1$ implies instability and would be associated with very high delays and pendency, which theoretically would continue to increase with increase in time. If utilization is less than 100%, i.e. $\rho < 1$, then ρ also equals the fraction of time the system capacity is busy in processing work. Further, in this case, one can analyze the system's steady state characteristics. Unfortunately, queueing analysis can be very complex and analytical characterization of the expected delay is not possible in generality. For the most rudimentary queueing system, the so-called $M/M/1$ queue, in which the time between the arrival of two consecutive jobs, and the service times are both exponentially distributed, the expected delay can be characterized as (Wolff 1989):

$$\text{Expected delay} = \frac{1}{1-\rho} \times \text{Expected service time.}$$

A key insight to take away from this expression is the impact of ρ on the expected delay. Specifically, as ρ approaches 1, the expected delay grows very rapidly. To illustrate, an increase in utilization from 90% to 99%, amplifies the expected delay ten-fold. This qualitative behavior is in fact a fundamental attribute of any queueing system, no matter how complex: as the utilization approaches 100%, the expected delay of the system grows in a highly non-linear fashion.[7]

[7] Some queueing networks are known to become unstable even when they are subcritical (less than 100% utilization) resulting in an unbounded growth in delay (Bramson 2008).

4.2. Analysis

Our first step is to calibrate the simulations with data along the lines of the description in the model section. However, we are missing a critical piece of information: we do not know the exact number of hours each day that the court operates. SCI's procedure handbook does prescribe a daily duration of 4.5 hours for judges to conduct hearings; this excludes time spent offline in reviewing case briefs and writing judgments (The Supreme Court of India 2017). But a number of hearings (e.g., urgent bail applications) may be conducted within the chambers of the judges and not in the courtrooms, and are thus not observable on the display-board (The Supreme Court of India 2015, p. 75). As a result, we do not have complete visibility into the realized duration of a workday. To circumvent this problem, we vary the duration of the workday in our simulation, and compare the simulated expected disposition time for cases with the expected disposition time observed in the data. For the latter calculation, we worked with all disposed cases in our dataset (besides the cases mentioned in Table 1, we also had access to partial data for the years 2008 and 2017), to arrive at an expected disposition time of 275.6 days based on $319,471$ observations.

Our simulation with a workday duration of 4.51 hours results in an expected disposition time equal to 134.6 days, while a workday of 4.5 hours leads to an expected disposition time of 398.6 days. These values sandwich the number 275.6 days observed in the raw data. The full distribution associated with the disposition time of cases is reported in Figure 2. The details of the simulation, especially related to convergence of the estimates, are fairly technical, and hence, are provided in Appendix C, Confidence Intervals for Simulated Estimates. Looking at Figure 2 we notice the contrast between the two distributions. Specifically, for the 4.5 hour workday, nearly 10% of cases take 3 years or more to be disposed, whereas for the 4.51 hour workday this amount of delay is experience by less than 0.5% of cases. The figure illustrates that not only is the mean disposition time much lower with the 4.51 hour workday, but the overall distribution of disposition times is much more condensed (there is a first-order stochastic dominance ordering).

Thus, based on the expected disposition times, our analysis implies that the SCI effectively operates in a range that is equivalent to somewhere between a 4.5 to 4.51 hour workday. This finding has a major implication: SCI operates close to a *critically loaded regime* (100% utilization) wherein even a small change in process characteristics (e.g., a change of 0.01 hours in the workday) can lead to a dramatic change in system performance. Indeed, we find that a 4.51 hour workday corresponds to 99.67% utilization, while a 4.5 hour workday corresponds to 99.97% utilization. This implies that any attempt to reduce the utilization by even a minimal amount should result in a disproportionate decrease in the expected disposition time, and thus lead to a significant improvement *without overly sacrificing throughput* (processed cases). We next investigate and contrast the magnitude of this potential alleviation for different kinds of interventions.

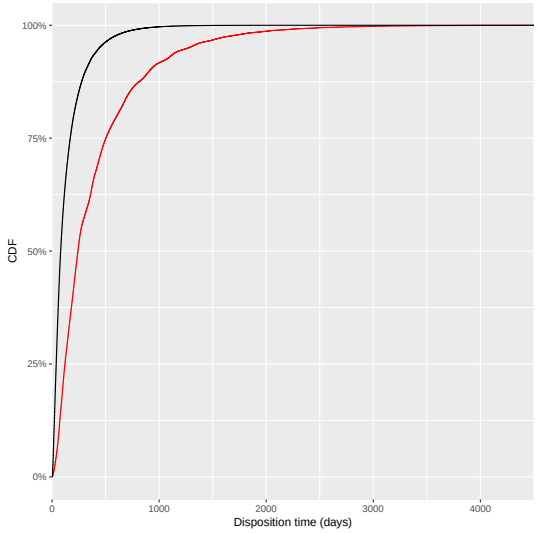

Figure 2 **The CDF (cumulative distribution function) for case disposition time for baseline with** 14 **benches:** 4.51 **hour workday (black and higher line), and** 4.5 **hour workday (red and lower line).**

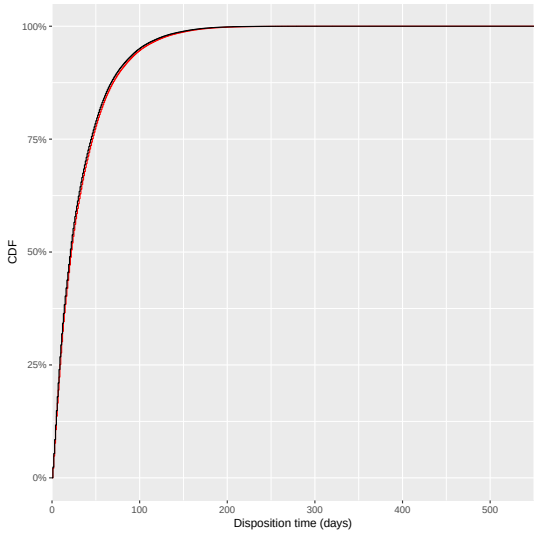

Figure 3 **The CDF (cumulative distribution function) for case disposition time for counterfactual with** 15 **benches:** 4.51 **hour workday (black line), and** 4.5 **hour workday (red line).**

4.3. Counterfactuals

We focus on two interventions guided by policy debates on this topic within the Indian judiciary. The first intervention is the most intuitive manner of tackling congestion, that of adding capacity in the form of judges. The second intervention is more process oriented and entails decreasing the number of adjournments granted in a case. While various Law Commissions in India have argued for additional judges to reduce backlogs (Law Commission of India 1987, 2014), India's Code of Civil Procedure, 1908 also lays down the guidelines and recommended time-frames for the completion of various procedural steps. These guidelines of often flouted thereby contributing to congestion.

Table 2 Impact of Adding Judges

Workday duration	Expected disposition time (14 benches)	Expected disposition time (15 benches)	% reduction
4.5 hrs	398.6 days	33.5 days	91.6%
4.51 hrs	134.6 days	31.6 days	76.5%

A particularly troublesome feature of the current equilibrium is the high number of adjournments granted in some cases. An amendment to the Code of Civil Procedure in 1999 required a cap of three adjournments in any suit. However, subsequent decisions by the SCI have diluted the status of this amendment to a mere recommendation (Ranjan 2016). We study both these interventions to ascertain their impact. Our finding that the SCI operates close to a critically loaded regime implies that either intervention will have a significant positive benefit in our performance metrics.

We have modeled only two-judge benches in our case-management queue, hence, the smallest unit of increase in judicial capacity would be the addition of a 15^{th} bench, i.e., increasing the number of judges from 28 to 30. We summarize our findings in Table 2. We note that this additional bench decreases the expected disposition time by a factor of 10 for the scenario with a 4.5 hour workday. The distribution of the disposition time of cases is reported in Figure 3. Note that the distributions are much narrower than in the baseline case (14 benches); now nearly 99.9% cases are disposed in less than a year. We also note that the plots for workday duration of 4.5 hours and 4.51 hours are indistinguishable. These findings are consistent with the predictions from queueing theory because the increase in capacity by $1/14 = 7\%$ leads to a utilization of about 93% (we confirm this in Appendix C, Confidence Intervals for Simulated Estimates). At this utilization level, which is well short of being critical, the performance with a workday of 4.5 hours is very similar to that with 4.51 hours. As we approach criticality, this small difference of 0.01 hours (less than a minute per day) translates to a large difference in performance, as observed in Figure 2. It is interesting that in this counterfactual, a 93% utilization, which can also be construed as being "high," results in a very manageable expected disposition time of about a month. Thus, our simulations illustrate that it is possible to run the SCI at high utilization with reasonable delay.

Turning to capping the number of adjournments allowed per case, we summarize the simulation outcomes in Table 3. Specifically, we simulate hearing outcomes for any case with the original outcome probabilities (when adjournments are not capped) until the maximum allowed adjournments are reached; thereafter, the hearing outcomes are restricted to disposal or regular hearing with updated probabilities that reflect their relative proportion in the data. Similar to the impact of adding judges, a cap on adjournments reduces the expected disposition time to around a month, both for the case of 4.5 hour and 4.51 hour workdays. We further note that a more restrictive cap on adjournments (below the recommendation of three) has limited additional benefit. (We also

Table 3 Impact of Capping Adjournments

Workday duration	Expected disposition time (no cap)	Adjournment cap	Expected disposition time (with cap)	% Reduction
4.5 hrs	398.6 days	3	39.3 days	90.1%
		2	35.6 days	91.1%
		1	30.0 days	92.5%
4.51 hrs	134.6 days	3	38.3 days	71.5%
		2	35.7 days	73.5%
		1	29.4 days	78.2%

found that the distribution of disposition time was similar to that in Figure 3, and therefore, we omit it.)

To summarize, we find that SCI is operating close to a critically loaded regime such that any improvement in operations, either via increase in capacity or by decreasing work via reducing adjournments, has considerable beneficial impact. We find that the process-level change of capping adjournments at three has an impact comparable to that of adding a two-judge bench.

5. Limitations and Robustness

We have developed a data-driven simulation model based on queueing theory. Data limitations and time constraints make it impractical to exactly replicate the functioning of a complex institution such as the SCI. Instead, our goal is to create a model that is tractable yet realistic enough to generate robust results. We highlighted certain assumptions that we have made in our analysis in Appendix B, Assumptions in Simulation Model. Below, we revisit two key assumptions and discuss why we believe our analysis and insights are robust.

Assumption 1: Ignoring history-dependence

We have assumed that the outcome of a hearing for a case is independent of how long the case has been pending. This may not be entirely accurate because, for instance, the annual reports of the SCI describe ad-hoc initiatives of the court to expedite disposal of excessively delayed matters, e.g., (The Supreme Court of India 2015, p.72-77), (The Supreme Court of India 2019, p.83-84). It is infeasible to model such interventions with reasonable precision, and they are targeted at only a small proportion of cases. Moreover, to the extent that expediting leads to prioritization in scheduling a hearing, independent of the remaining work needed to resolve the case, and does not indicate a change in the expected outcome of a hearing, we can invoke the principle of work-conservation in queueing systems. Correspondingly, we would expect the first moments of our key metrics (e.g., average disposition time) to be robust to any such scheme for prioritizing among cases (Wolff 1970). In the event that such conservation does not hold, the first moments will change, however, the system would remain in the critically-loaded regime and thus the higher-order insight

that a small improvement (in capacity or process efficiency) can result in significant benefits would still hold.

Assumption 2: Ignoring case heterogeneity

The SCI classifies cases by the nature of the matter under consideration. However, we have not tried to capture this scheme to segment the population of cases. Instead, we have treated the cases as homogeneous. This assumption simplifies our exposition and treatment while still being aligned with our high-level objective of tracking performance statistics at the level of the entire population (not by case type). For instance, we have modeled a single hearing-time distribution for all case types. In this regard, the display-board data as well as prior work indicate that more than 80% of the workload of the SCI pertains to dealing with appeals of two kinds: criminal matters and civil matters (Robinson 2013a). Based on our data, we find that the average hearing time for criminal matters and civil matters is not statistically distinguishable, nor are these times statistically distinguishable from the overall average across all case types, thus suggesting that our assumption that cases are homogeneous is not restrictive. Besides, even if variations did exist across case types, intuitively, such variations should "average out" in population-wide metrics.

A violation of both assumptions above (e.g., classifying cases into categories with different hearing time distributions that can be ranked, and then prioritizing "shorter" cases) has the potential to influence even the first moment of simulated metrics. Although we have no reason to believe that prioritization of this kind is an important consideration for the SCI, such procedures would still not influence the level of utilization of the system. Further, as we show next, the utilization of the system is very close to 100% even using simple "first-order" or "back-of-the-envelope" calculations. Therefore, our main insight that the SCI is operating in the nearly critical regime remains valid.

Robustness of the operating regime

As defined previously, the utilization is calculated as the ratio of the rate of incoming workload brought by arriving cases to the available judicial capacity. It is helpful to step back from the details of the simulation and compute a theoretical utilization of the SCI using first-order workload information. We note that we have a non-stationary arrival process with variations in annual arrival of cases (as illustrated in Table 1). A rough workaround is to calculate utilization per year, i.e., looking at the work arriving in a given year and computing a ratio to the judicial capacity (which remains constant in our analysis). For the latter, we use the estimated hearing-time statistics along with the hearing-outcome probabilities (as reported in the Model section) and maintain a 4.5 hour workday. Doing so, we obtain a minimum annual "theoretical utilization" of 98.2%.[8] While our

[8] The highest utilization is 104.1% which exceeds 100% and correspondingly is interpreted as the *offered load* being greater than the judicial capacity.

back-of-the-envelope calculation is not a precise predictor of utilization, and correspondingly, that of stability, it is in the ball-park of what we observed from our simulation study, and is near 100%. Thus, the operating regime is either overloaded or nearly critical. The first-order analysis presented here ignores many aspects of SCI that we have incorporated in the simulation, however, this provides us some assurance that the insights obtained from the simulation are consistent with what one may expect with an even simpler calculation.

6. Discussion

We have determined that the baseline operating regime of the SCI corresponds to very high utilization, and is therefore quite sensitive to changes in attributes. Although the two interventions that we described led to similar reduction in delay, in practice, they can have considerably different financial and institutional costs. This can influence the choice of policymakers regarding the preferred means for controlling congestion.

Adding judges not only has financial implications related to salary and infrastructure, but its feasibility is also linked to the availability of suitably qualified judges in such large numbers. The judicial strength of the SCI has progressively increased over the years: it was eight at inception in 1950; the current sanctioned number is 34 (The Supreme Court of India 2019). Moreover, given their large number, it is impractical for SCI judges to adjudicate *en banc* (together as a full court), instead they typically decide cases as two-judge benches. This has led to concerns about the court being *polyvocal*: not building systematically on legal precedent, a fundamental precept of *common law* (Robinson 2013b). Adding judges further aggravates this concern (The Wire 2019).

By contrast, capping adjournments is a form of process re-engineering that seems to be "free." However, the SCI's own previous ruling in the 2005 case of Salem Advocate Bar Association-II (2005 (6) SCC 344), expresses misgivings about capping adjournments if the litigants have a genuine reason. Thus, the downside of capping adjournments is viewed more in terms of its impact on procedural fairness.

Our results strengthen the ability of policymakers to conduct a cost-benefit analysis to choose between various options that alleviate congestion. The scope for dramatic improvement revealed by our results seems to bode well for the prospects of reducing judicial delay. However, despite the increase in judicial strength at the SCI over the years, this strategy has not been very successful in eliminating backlogs, which have continued to grow (The Wire 2019). This motivates us to take a closer look at the nature of the underlying workload.

In general, judicial strength may be expected to keep pace with population/litigation growth. However, supreme courts are typically shielded from this concern because they focus on mainly constitutional matters. As mentioned previously, the SCI is distinct from its counterparts in other

countries because it also functions as the court of final appeal. Moreover, based on data from $2005-2011$, prior work concludes that appeals constitute a majority of the court's workload, and the growth rate of appeals to the SCI outstrips the growth rate in the output of the state high courts (Robinson 2013a). This evidence is consistent with the vision laid out in the Constitution of India of providing wide access to the common man to justice ("docket inclusion"), and given the inadequacies of India's lower judiciary, access to the SCI. In fact, "docket exclusion" is viewed as a critical national policy challenge that is not revealed by data (The Supreme Court of India 2012). Thus, there is an apparent trade-off at play between acceptable judicial delay and providing citizens with wider access to courts.

To comprehend and balance this trade-off, it is important to first lay out the policy objectives, and then use scientific methods to optimize the relevant decisions such as judicial capacity and accepted workload (appeals). Regarding the former, besides keeping a check on the expected disposition time, the SCI has stated its aspiration to prevent excessive delay to any individual case (The Supreme Court of India 2012). In fact, as per the Malimath Committee's recommendations, any matter that is delayed for more than two years should be treated as *arrears* and disposed with high priority (Malimath 2003). It is clear from the above discussion that policy makers also care about the distribution of disposition time, not just the expected value.

The other piece of the puzzle is a scientific framework for planning, in particular, for determining the required number of judges for a given rate of litigation. The report of the 245^{th} Law Commission of India lists various options in this regard, and favors the "Rate of Disposal Method" (Law Commission of India 2014, p. 24).[9] An important shortcoming of this approach is that, unlike *queueing theory*, it fails to fully account for the variability in the arrival and service processes. Instead, it focuses on ensuring that "... the number of disposals [of cases] equals the number of institutions in any one year." This condition is equivalent to ensuring stability, i.e., $\rho < 1$. However, the expected disposition time blows up as server utilization approaches 100%, which is undesirable for obvious reasons. The SCI is operating in precisely this brittle regime, seemingly due to its focus on only ensuring $\rho < 1$. With such a focus, we conjecture that the relief provided by measures such as adding judges will be temporary as it will soon be offset by admitting more cases — via appeals — to the SCI's docket. Our study provides a framework to resolve these shortcomings. It not only highlights the perils of operating in a stable but high utilization regime, but also offers a rigorous approach to determine the combination of judicial capacity and workload that avoids such an outcome. Although we have demonstrated the utility of our framework in the context of procedures followed at the SCI, it can be adapted and applied

[9] The report considers, but does not find suitable, another capacity planning framework that it refers to as "Time-Based Method," which is also used in the United States.

fruitfully to address the capacity planning needs of India's lower judiciary as well. Equally, we hope that our framework will help in combating congestion in courts in other parts of the world.

Acknowledgement. The authors would like to gratefully acknowledge helpful conversations with the following people: Justice Deepak Gupta and Justice Madan B. Lokur, former judges of the Supreme Court of India; Mr. Gaurav Pachnanda and Mr. Sahil Tagotra, senior advocates practicing in the Supreme Court of India; Justice Sanjay Parihar, former Registrar of the Supreme Court of India; Mr. B.S. Surya Prakash from DAKSH, an organization that promotes judicial access and accountability in India.

References

Adler, Andrew L. 2014. Extended vacancies, crushing caseloads, and emergency panels in the federal courts of appeals. *J. App. Prac. & Process* **15** 163.

Asmussen, Søren. 1992. Queueing simulation in heavy traffic. *Mathematics of Operations Research* **17**(1) 84–111.

Blumstein, Alfred, Richard Larson. 1969. Models of a total criminal justice system. *Operations Research* **17**(2) 199–232.

Bramson, Maury. 2008. *Stability of queueing networks*. Springer.

Bray, Robert L, Decio Coviello, Andrea Ichino, Nicola Persico. 2016. Multitasking, multiarmed bandits, and the Italian judiciary. *Manufacturing & Service Operations Management* **18**(4) 545–558.

Campello, Fernanda, Armann Ingolfsson, Robert A Shumsky. 2017. Queueing models of case managers. *Management Science* **63**(3) 882–900.

Carrington, Paul D. 1969. Crowded dockets and the courts of appeals: the threat to the function of review and the national law. *Harvard Law Review* 542–617.

Castro, Massimo Finocchiaro, Calogero Guccio. 2015. Bottlenecks or inefficiency? an assessment of first instance italian courts' performance. *Review of Law & Economics* **11**(2) 317–354.

Financial Times. 2016. India's top judge Thakur pleads for help with avalanche of cases. https://www.ft.com/content/788fd8f8-0aac-11e6-b0f1-61f222853ff3?mhq5j=e7.

Green, Andrew, Albert H Yoon. 2017. Triaging the law: Developing the common law on the supreme court of india. *Journal of Empirical Legal Studies* **14**(4) 683–715.

Law Commission of India. 1987. Manpower Planning in Judiciary: A Blueprint. http://lawcommissionofindia.nic.in/old_reports/rpt120.pdf.

Law Commission of India. 2014. Arrears and Backlog: Creating Additional Judicial (wo)manpower. http://lawcommissionofindia.nic.in/reports/Report_No.245.pdf.

Malimath, V.S. 2003. Committee on Reforms of Criminal Justice System, Government of India, Ministry of Home Affairs. https://indialawyers.files.wordpress.com/2009/12/criminal_justice_system.pdf.

Meador, Daniel J. 1974. Appellate courts: Staff and process in the crisis of volume. *St. Paul, MN: West Publishing Co* .

Nagel, Stuart, Marian Neef, Nancy Munshaw. 1978. Bringing management science to the courts to reduce delay. *Judicature* **62** 128.

Ranjan, Brajesh. 2016. What causes judicial delay? Judgments diluting timeframes in code of civil procedure worsen the problem of adjournments. https://timesofindia.indiatimes.com/blogs/toi-edit-page/what-causes-judicial-delay-judgments-diluting-timeframes-in-code-of-civil-procedure-worsen-the-problem-of-adjournm

Robinson, Nick. 2013a. A quantitative analysis of the indian supreme court's workload. *Journal of Empirical Legal Studies* **10**(3) 570–601.

Robinson, Nick. 2013b. Structure matters: The impact of court structure on the Indian and US Supreme Courts. *The American Journal of Comparative Law* **61**(1) 173–208.

Singhvi, Abhishek. 2017. Open letter to incoming CJI: The most pressing priorities before India's judiciary, and how to address them.

The Government of India. 2020. National Judicial Data Grid. https://njdg.ecourts.gov.in/njdgnew/index.php.

The Supreme Court of India. 2012. National Court Management System – Policy and Action Plan. https://main.sci.gov.in/pdf/NCMSP/ncmspap.pdf.

The Supreme Court of India. 2015. Annual Report 2014. https://main.sci.gov.in/publication.

The Supreme Court of India. 2017. Handbook On Practice and Procedure and Office Procedure. https://main.sci.gov.in/practice-and-procedure.

The Supreme Court of India. 2018. Conference Proceedings of National Initiative to Reduce Pendency and Delay in Judicial System. https://districts.ecourts.gov.in/sites/default/files/Proceeding%20Book%20Supreme%20Court.pdf.

The Supreme Court of India. 2019. Indian Judiciary: Annual Report 2018-19. https://main.sci.gov.in/publication.

The Supreme Court of India. 2020. Website of the Supreme Court of India. https://main.sci.gov.in/.

The Wire. 2019. Why CJI Gogoi's Proposal to Increase the Supreme Court's Strength Is Misplaced. https://thewire.in/law/cji-ranjan-gogi-proposal-increase-supreme-court-strength.

The World Bank. 2019. Population, total - India. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=IN.

The World Bank. 2020. Ranking economies by ease of doing business. https://www.doingbusiness.org/en/data/exploreeconomies/india.

Whitt, Ward. 1989. Planning queueing simulations. *Management Science* **35**(11) 1341–1366.

Wolff, Ronald W. 1970. Work-conserving priorities. *Journal of Applied Probability* **7**(2) 327–337.

Wolff, Ronald W. 1989. *Stochastic modeling and the theory of queues*. Pearson College Division.

World Prison Brief. 2020. World Prison Brief data. https://www.prisonstudies.org/country/india.

Appendix A: Estimating the Hearing-Time Distribution

The details available on the case-status page of Supreme Court of India (SCI)'s website do not include information about the time spent on individual hearings in a case. This data is required in our model to simulate the service process. Hence, we collected actual hearing times shown in real time by the online display-board on the court's website. The data was collected for a period of seven months, from November 12, 2018 to May 10, 2019. This resulted in 15,725 observations. Those observations that spanned the lunch hour were appropriately corrected by deducting the extra hour from the hearing time. Moreover, as per the court website, the display-board automatically updates every 30 seconds. To overcome the resulting limitations from the perspective of fitting a distribution to the data, we "jitter" the observations by adding a Gaussian noise with 0 mean and a standard deviation equal to 3 seconds.

A peculiarity of the adjudication process is that the duration of hearings can vary considerably: from short hearings that last a few minutes to longer deliberations that run into hours. Therefore, we find that the best fit with data is obtained by using a mixture of nine Gaussian random variables and a random variable with a Log-normal distribution (this heavy-tailed distribution helps capture the long right tail of the distribution for hearing times). See Figure 4. Table 4 provides the details of the individual distributions that are mixed to provide the overall distribution for hearing time. The table also provides the weight associated with each distribution used in the mixture. Our statistical analysis yields an expected hearing time of 461 seconds. We evaluated the quality of the fit using the Kolmogorov-Smirnov test; the p-value is close to 0.4.

Table 4 **Parameters and weight for the mixture of distributions used to model hearing time**

Sr. No.	Distribution Type	Weight	Mean	Std. Dev.
1	Gaussian	0.373	52.13 sec.	3.19 sec.
2	Gaussian	0.039	188.78 sec.	69.50 sec.
3	Gaussian	0.086	156.33 sec.	4.34 sec.
4	Gaussian	0.051	208.87 sec.	4.76 sec.
5	Gaussian	0.030	312.73 sec.	7.67 sec.
6	Gaussian	0.036	261.15 sec.	5.95 sec.
7	Gaussian	0.025	363.87 sec.	13.19 sec.
8	Gaussian	0.002	60.97 sec.	1.38 sec.
9	Gaussian	0.151	104.20 sec.	3.98 sec.
10	Log-normal	0.204	1797.37 sec.	4110.98 sec.

Appendix B: Assumptions in the Simulation Model

We have developed a data-driven simulation model based on queueing theory. The model has three main components: the arrival process; the service process; and the hearing outcomes. Data limitations and time constraints make it impractical to exactly replicate the functioning of a complex institution such as the SCI. Instead, our goal is to create a model that is tractable yet realistic enough to generate robust results. To this end, in addition to the model description provided in the main manuscript, we highlight certain assumptions that we have made.

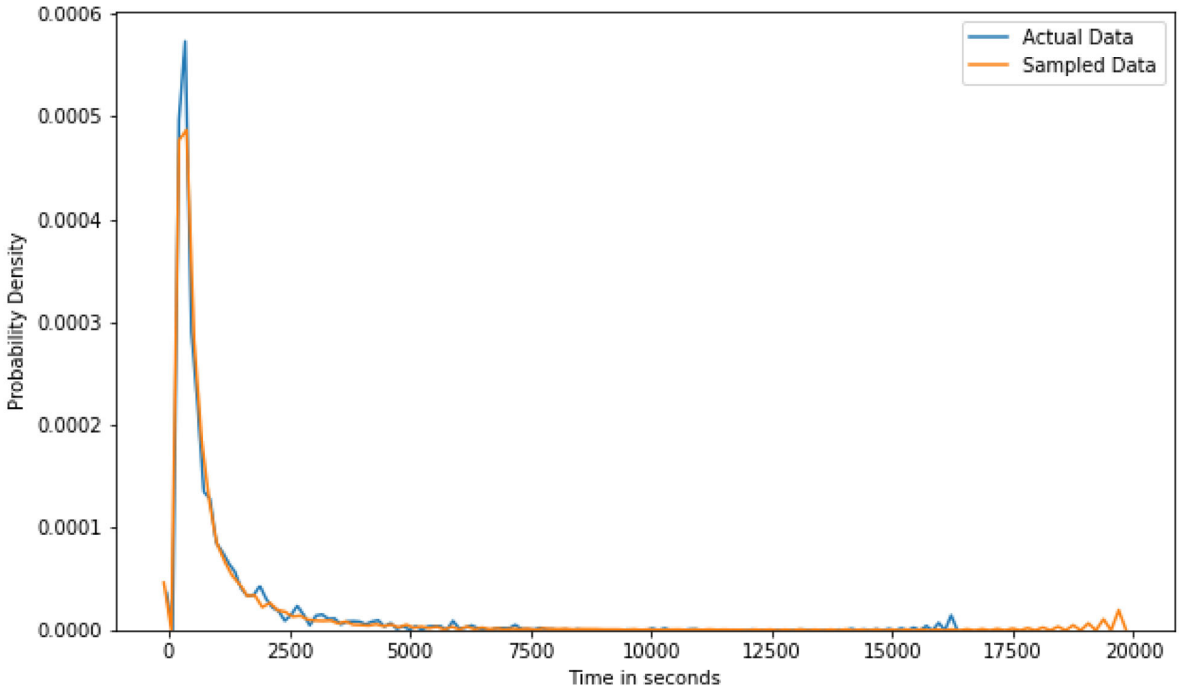

Figure 4 Best Fit for Hearing Time Data.

The Arrival Process

A1. A case cannot have multiple hearings on the same day. This is not only intuitive because a single long hearing can replace multiple short hearings, but also consistent with the data.

A2. A case cannot be registered and called for a hearing on the same day. This allows us to accommodate non-modeled aspects of the registration process such as document verification and validation.

The Service Process

A3. We model fourteen, homogeneous, two-judge benches. Even though some cases are heard by three-judge and five-judge benches, these are very few in number, as explained in the main manuscript. Moreover, we do not model appointment of additional judges or retirement of existing judges.

A4. We model calendar days in the simulation, and assume that no work happens on holidays. Besides weekends, holidays include summer and winter vacations as per the calendar for 2020 displayed on the court website. Festivals are also holidays that are uniformly distributed through the year. In all, a total of 180 working days are modeled in each year.

A5. After the completion of a hearing for a case, it goes to the back of the queue associated with its assigned bench. In practice, there is scope for the next hearing of a case to be prioritized for various reasons, including the discretion of the Chief Justice. Since the data does not provide any visibility into such instances, we do not attempt to model them. Importantly though, this assumption does not affect the total judicial time spent on any case. Hence, we expect the first moments of our key metrics to be robust to this assumption.

A6. Cases are classified by the SCI into case-types based on their subject matter. However, in order to streamline our analysis and exposition, we ignore the heterogeneity in cases. We discuss the robustness of our analysis to this assumption in the "Limitations and Robustness" section in the main paper.

The Hearing Outcome

A7. We have made the simplifying but practical assumption that the outcome of a hearing for a case is independent of the case's history. This may not be entirely accurate because, for instance, the annual reports of the SCI describe ad-hoc initiatives of the court to expedite disposal of excessively delayed matters. It is infeasible to model such interventions with reasonable precision. We provide a more detailed discussion of the robustness of our analysis to this assumption in the "Limitations and Robustness" section in the main paper.

Appendix C: Confidence Intervals for Simulated Estimates

One of the key conclusions of our study is that the SCI operates in a regime that is characterized by very high utilization of its service capacity (i.e., the judges) such that the queue is nearly critically loaded (close to 100% utilization). It is well-known that the convergence of simulation estimates in this regime requires a disproportionately large number of observations for the *central limit theorem* to kick in; fewer observations lead to gross under-estimates (Whitt 1989, Asmussen 1992). In particular, the number of observations required is proportional to $\frac{1}{(1-\rho)^2}$, where ρ is the utilization of capacity which tracks the proportion of time for which the judges are busy. Formally, it is equal to ratio of the expected capacity to the expected demand, i.e., $\rho =$ (expected hearing time)/(expected time between arrival of two cases).

The literature prescribes the following scheme for determining the number of observations required for our simulations (Whitt 1989):

1. Simulate the performance of the queueing system while keeping the level of utilization low, to about 0.3. This is achieved in our setting by increasing the workday duration to a suitably high number compared to the 4.5 hours that the SCI judges are expected to work, and then tracking the utilization achieved in the simulation run. At this lower level of utilization, the number of required observations, t, for the estimates to converge is "typical," and standard techniques for determining the size of the confidence interval apply.

2. Filter the number of observations (disposed cases) obtained from this simulation run, and calculate the corresponding one-sided width of the confidence interval for the metric of interest (e.g., expected disposition time). For each width of the confidence interval, we are able to determine the corresponding constant-of-proportionality for the relationship, $t \propto \frac{1}{(1-\rho)^2} \Rightarrow t(1-\rho)^2 = \text{constant}$.

3. This constant-of-proportionality can now be employed in the high-utilization regime to recommend the number of observations required to achieve the same confidence-interval size (expressed as a percentage) as in the low utilization regime.

Conversely, we can fix the number of observations in the high-utilization regime. This allows us to determine the constant-of-proportionality, which we then use to come up with the corresponding number of observations in the low-utilization regime. Finally, we use this number to determine the associated size of the confidence

Table 5 Summary of Simulation Outcomes: Baseline results

Workday duration	Utilization	Expected disposition time (14 benches)	Confidence Interval One-sided width (%)
4.5 hrs	99.97%	398.6 days	79.2%
4.51 hrs	99.69%	134.6 days	10.4%

Table 6 Summary of Simulation Outcomes: Counterfactuals with additional judges

Workday duration	Utilization	Expected disposition time (15 benches)	Confidence Interval One-sided width (%)
4.5 hrs	93.8%	33.5 days	0.7%
4.51 hrs	92.82%	31.6 days	0.7%

Table 7 Summary of Simulation Outcomes: Counterfactuals with capped adjournments

Workday duration	Max. adjournments	Utilization	Expected disposition time (with cap)	Confidence Interval One-sided width (%)
4.5 hrs	3	96.67%	39.3 days	1.4%
do	2	94.67%	35.6 days	1.0%
do	1	91.11%	30.0 days	0.6%
4.51 hrs	3	96.36%	38.3 days	1.4%
do	2	94.69%	35.7 days	1.0%
do	1	90.70%	29.4 days	0.6%

interval (for the metric of interest) using standard simulation analysis. The relative size of the confidence interval (expressed as a percentage) is the same for both regimes (low and high utilization). For the simulation runs reported in the main manuscript, we provide the average disposition time and associated confidence interval in Tables 5, 6, and 7 below.

For our baseline simulations (Table 5), we work with 14 benches. The workday durations that we report are 4.5 hours and 4.51 hours. In these scenarios, we stopped simulating after collecting 24 million observations (disposed cases). The data that we have downloaded includes arrivals from 2008 to 2016, and we track output from 2009 onward. This data by itself will not allow us to collect anywhere near the 24 million observations that we need. (From Table 1 in the main manuscript, we know that the arrival rate of cases is about $40,000$ per year. Hence, 24 million disposed cases requires about 600 years worth of arrivals.) We achieve this goal by repeating the arrivals from 2008 to 2016 as many times as required to generate the targeted number of observations. Yet, the confidence interval reported for a workday of 4.5 hours is quite wide (the one-sided width is 79.21%). We do not believe it is worth striving for more accuracy. Each simulation run already took multiple days of actual run time on servers in a high-performance computing laboratory, and a tighter confidence interval for the 4.5 hour workday will not affect the robustness of our conclusions. A look at the pendency plot in Figure 5 reveals that a workday duration of 4.51 hours leads to a stable system since the backlogs (pendency) are not growing without bound with time. However, it is less clear whether a workday duration of 4.5 hours leads to stability.

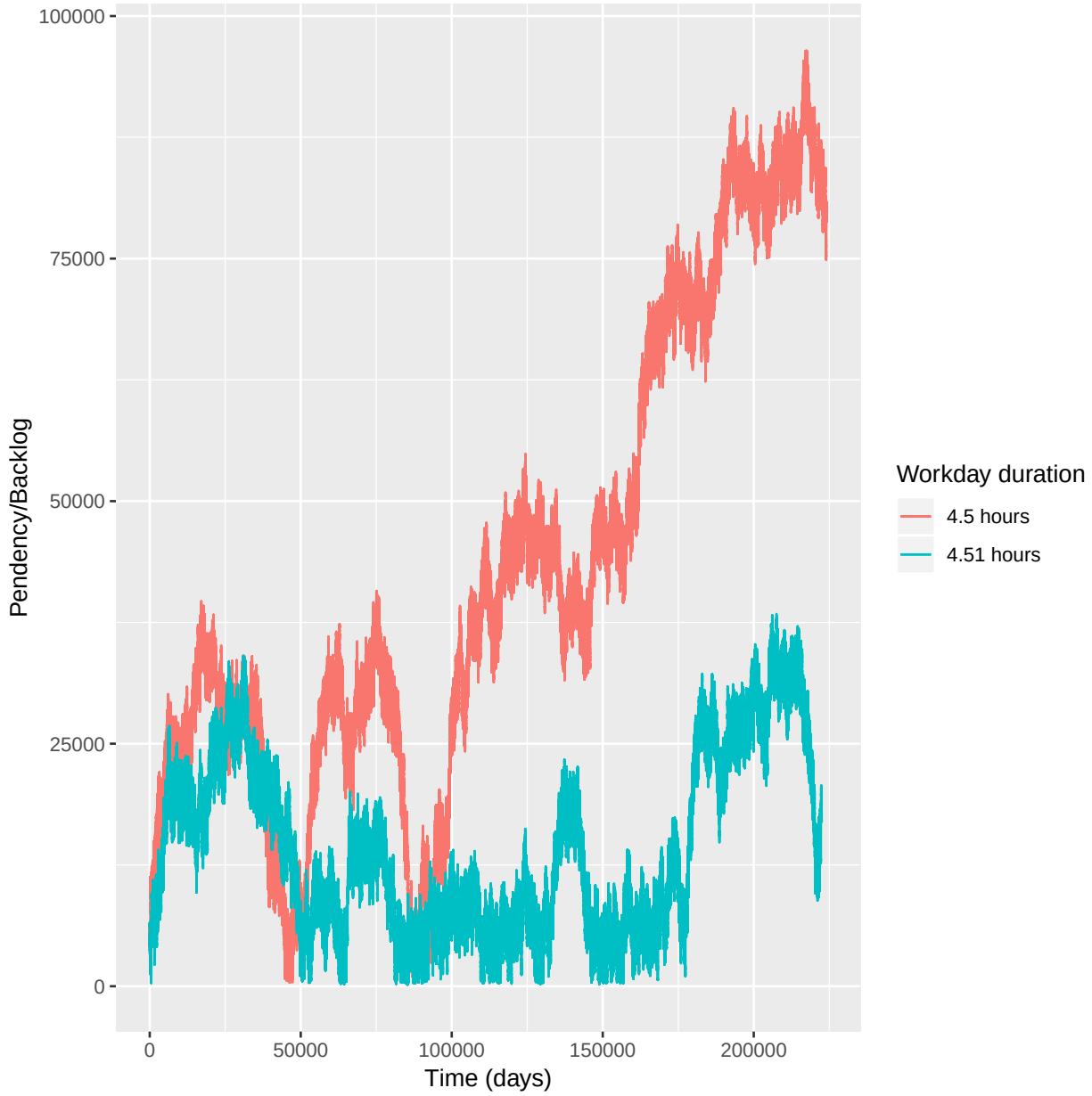

Figure 5 Pendency plots for workday duration 4.51 hours and 4.5 hours (with 14 benches)

Appendix D: Classification of Judicial Orders

Each SCI hearing results in one of the following outcomes: *decision/disposal*, *adjournment*, or *regular hearing*. These categories are not formally recorded anywhere in our data. Hence, one needs to parse the actual text of the order to ascertain the outcome of the hearing. This assessment is quite challenging. We read through a large number of orders and noticed significant variety in how various outcomes were described. While some outcomes were straightforward and used language such as "dismissed" or "granted" to indicate a decision, other times language such as "leave granted" was used to indicate the decision. Using a combination of key words and length of the order as the means for classification, we created a decision tree that is depicted

in Fig. 6. (We also sought expert inputs from practicing supreme court lawyers to refine our classification logic.)

To assess the accuracy of our decision tree, we randomly sampled 572 orders and applied our classifier to it. We also manually read each order to ascertain the outcome directly. We believe that our algorithm correctly classified 554 hearings, which implies an accuracy of 96.9%.

Based on this classification algorithm, we determined the fraction of hearings that resulted in a *disposal*, *adjournment*, or *regular hearing*. We then interpreted these fractions as the probability of the corresponding outcome for a hearing. To ensure that these probabilities are estimated on a coherent set of orders, and to minimize the impact of censoring (due to pending cases), we based our estimates on cases that were registered in 2012. This resulted in the following probability estimates: 0.29 for disposal, 0.43 for adjournment, and 0.28 for regular hearing. We also ran the analysis for the years 2010, 2011, 2013, and 2014. The probability estimates are very close to those for 2012.

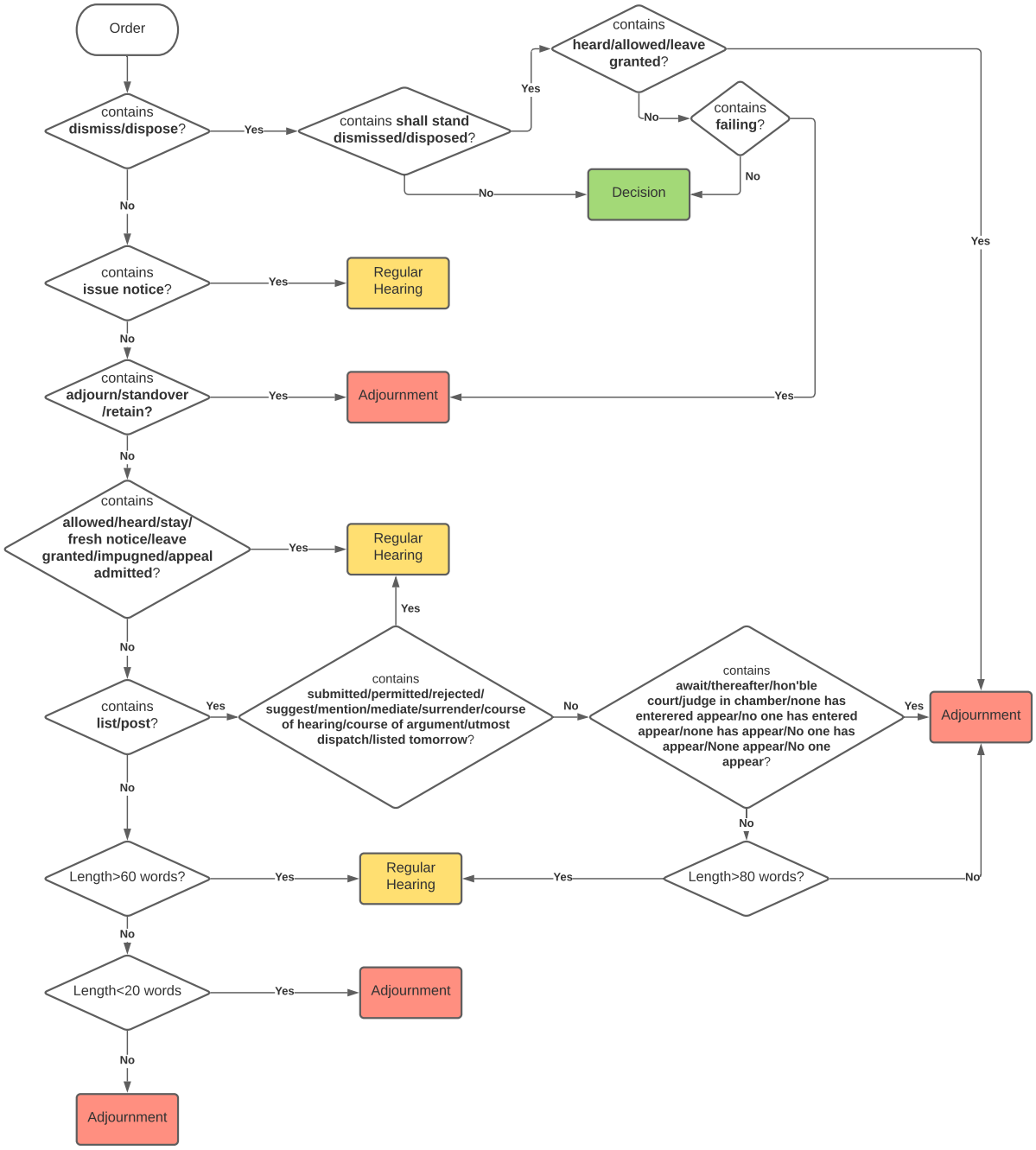

Figure 6 Decision tree to classify hearing as decision, adjournment or regular.