Bachelor of Techology Project Report

Analysis of Indian Legal Dataset

Mayank Sharan

2014CS10234

Advisor

Prof. Mausam

Department of CSE

IIT Delhi

2017-2018

Abstract

India is the world's largest democracy. Judiciary is a critical pillar of this democracy. Since independence the functioning of our judicial system has not been able to keep up with the requirement of the citizens of this country. The judicial system is highly overburdened with over 20 million cases currently pending in district courts of which 10% have been pending for over 10 years. There is a large shortage of judges. India has one judge for every 73000 people which is an average 7 times worse than in the United States. Bulk of the cases that the judiciary handles is done through their lowest tier of execution i.e. the district courts.

Data analysis and Artificial intelligence can help with making this system better. Descriptive analyses such as performance monitoring and anomaly detection can help identify superlative performances for further investigations and perhaps emulate the model being used elsewhere to get better results. Predictive analyses such as predicting expected completion time of the case or the next hearing date can help people set expectations from the judicial system. Prescriptive analyses such as scheduling and allocation methodologies can help improve the performance of the courts and achieve better satisfaction for the public from the system.

Overview

This project entails obtaining large amounts of legal data and drawing data analysis driven inferences and building models on the data for numerous purposes. The work over 2 semesters has been focused on the following :-

1. Data procurement

2. Statistical Analysis of disposal time data

3. Anomaly Detection using disposal time data

4. Case Type Clustering

5. Predictive Modeling

6. Web portal to present data and findings

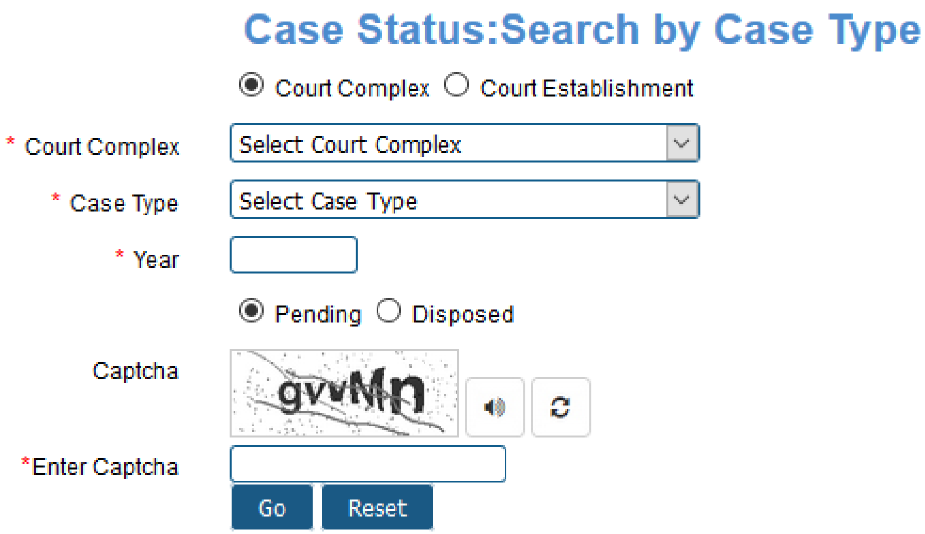
Data Procurement

The data is obtained from http://services.ecourts.gov.in. For each case the website provides numerous details like date of filing, date of disposal, intermediate hearing date, hearing details, case type, acts under which the case has been filed, litigant, defendant, judge presiding and court of filing.

This data from the website is scraped in the form of a text file case by case by an automated script. This script takes the state and the district as an input and then runs through all possible values of court establishments case types, years between 2002 and 2018 and whether the case is pending or disposed. For each of these value tuples if a list is provided by the website the script extracts the data for all cases in the list or as many as it can randomly in 30 minutes. The data obtained for a district is then databased into a mysql database using a second script.

The data is available on the website for 594 districts. The progress of data procurement at this point is

- Data from over 400 districts has been scraped
- Data from over 250 districts has been databased



Case Status:Search by Case Type

● Court Complex ○ Court Establishment

* Court Complex: Select Court Complex

* Case Type: Select Case Type

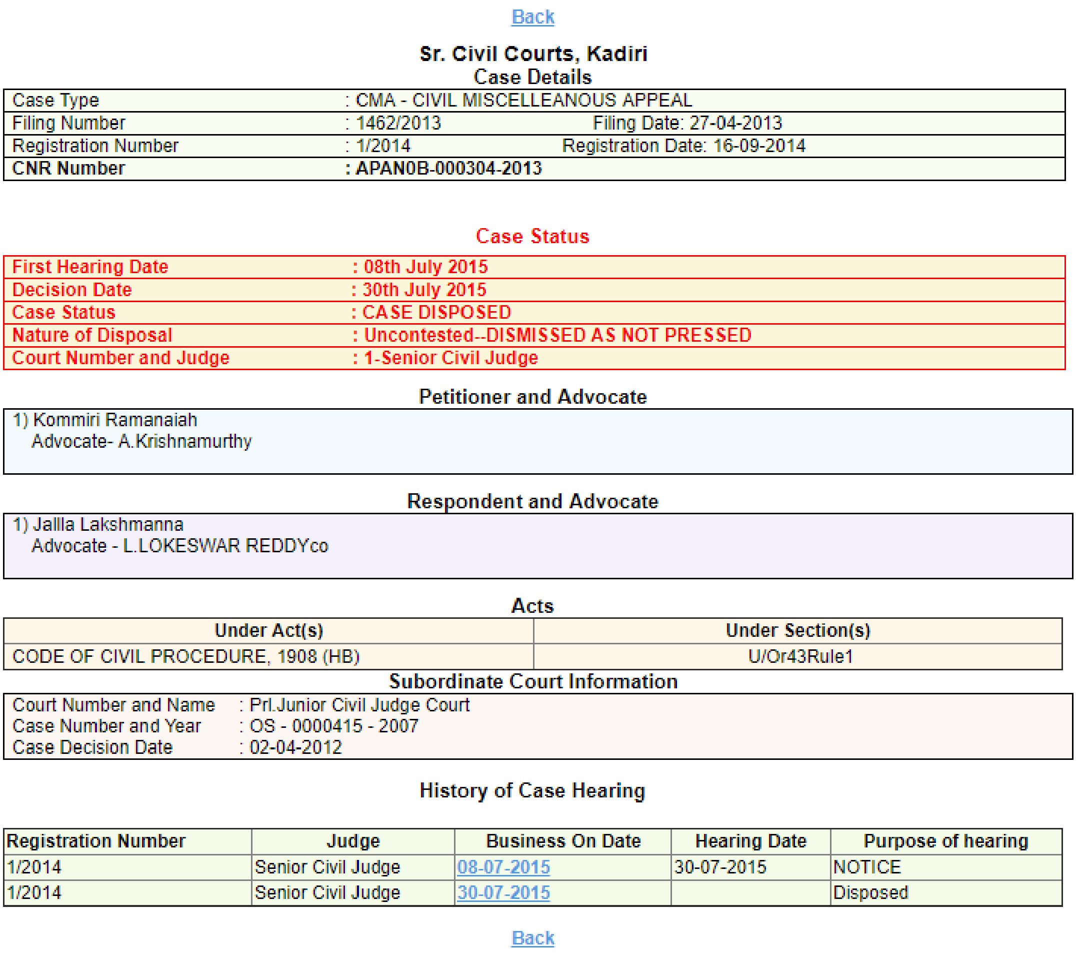
* Year:

● Pending ○ Disposed

Captcha

*Enter Captcha:

Figure 1. Website page



Back

Sr. Civil Courts, Kadiri

Case Details

Case Type	: CMA - CIVIL MISCELLEANOUS APPEAL	
Filing Number	: 1462/2013	Filing Date: 27-04-2013
Registration Number	: 1/2014	Registration Date: 16-09-2014
CNR Number	**: APAN0B-000304-2013**	

Case Status

First Hearing Date	: 08th July 2015
Decision Date	: 30th July 2015
Case Status	: CASE DISPOSED
Nature of Disposal	: Uncontested--DISMISSED AS NOT PRESSED
Court Number and Judge	: 1-Senior Civil Judge

Petitioner and Advocate

1) Kommiri Ramanaiah
Advocate- A.Krishnamurthy

Respondent and Advocate

1) Jallla Lakshmanna
Advocate - L.LOKESWAR REDDYco

Acts

Under Act(s)	Under Section(s)
CODE OF CIVIL PROCEDURE, 1908 (HB)	U/Or43Rule1

Subordinate Court Information

Court Number and Name	: Prl.Junior Civil Judge Court
Case Number and Year	: OS - 0000415 - 2007
Case Decision Date	: 02-04-2012

History of Case Hearing

Registration Number	Judge	Business On Date	Hearing Date	Purpose of hearing
1/2014	Senior Civil Judge	08-07-2015	30-07-2015	NOTICE
1/2014	Senior Civil Judge	30-07-2015		Disposed

Back

Figure 2. Sample data snippet

Statistical Analysis of disposal time data

Disposal time of a case is one of it's most critical attributes. Disposal time is also a good indicator to aggregate and compute feature values representative of the efficiency of the judicial system.

The following features were computed at different aggregate levels, district to get characteristics of the data.

- **Mean Disposal Time**

  This is a weighted mean of the disposal times in an aggregate fashion. Represents what it says the mean time taken in days to dispose off a case.

- **Zero Day Disposal Percentage**

  This is a weighted mean of the percentage of cases that are disposed on the day of filing. It often is a good indicator of the nature of the cases pursuant in the distribution.

- **Tail Bounds**

  This value is the fraction of cases that take more than a certain number of days to get disposed. We compute this for 90 days, 180 days, 365 days and 1095 days. This metric is a very important indicator of the health of a court or an aggregate since it captures the proportion of cases in some of the worst scenarios.

- **Standard Deviation**

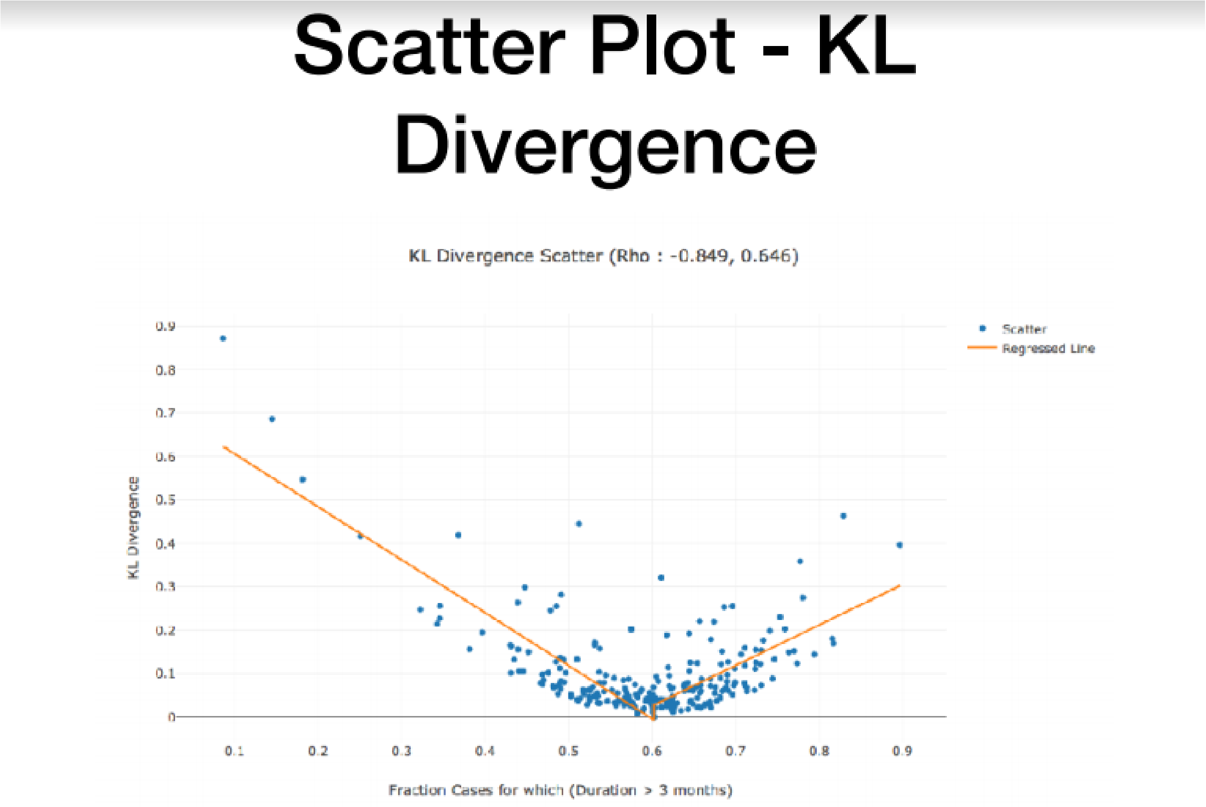
  The standard deviation captures the spread of the distribution and gives an indication along with the mean about the nature of the distribution.

- **KL Divergence**

  The Kullback-Leibler divergence provides a quantitative value representing separation between two probability distributions. For a given aggregate distribution we calculate the KL divergence w.r.t. the national aggregate distribution.

We also use this metric as an overall comparison score by using a negative sign appropriately to convert the symmetric function to one that represents good or bad performance of the metric. The computation is performed with discrete data using the following formula

$$D_{KL}(P||Q) = \sum_{x \epsilon X} P(x) \log_2 P(x)/Q(x)$$



Scatter Plot - KL Divergence

Figure 3. Scatter plot : KL divergence value vs tail bound value

The justification to use the KL divergence score as a metric to rank the districts in aggregate is strengthened by the strong correlation between the KL divergence values and the corresponding tail bound values.

Anomaly Detection using disposal time data

This step is the descriptive part of the analysis. Here we drew up rankings based on the metrics described in the previous section and identified the worst and best performing districts. This information was also captured in GeoJSON files that we shall see a sample of in the section describing the website. These files are a map based representation of these metric values with districts or states being color coded as per their values for a given metric and those values being displayed on hovering over them alongside the rankings.Further attempts at anomaly detection were made using parametric modeling

Parametric Modeling

Parametric modelling captures the information of an aggregate distribution using approximated parameters to get a structured model that can be compared in a standard fashion using a continuous function.Here we model the disposal time distribution using 3 parameters namely, zero day disposal percentage, exponential scaling coefficient α and exponential decay coefficient β. The best fit is found for the distribution without zero day disposals by adjusting the exponential coefficients. The function to fit the curve on is $\alpha \exp(-\beta x)$. These values were used to compare districts and also compute a KL divergence using the following formula -

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log p(x)/q(x) dx$$

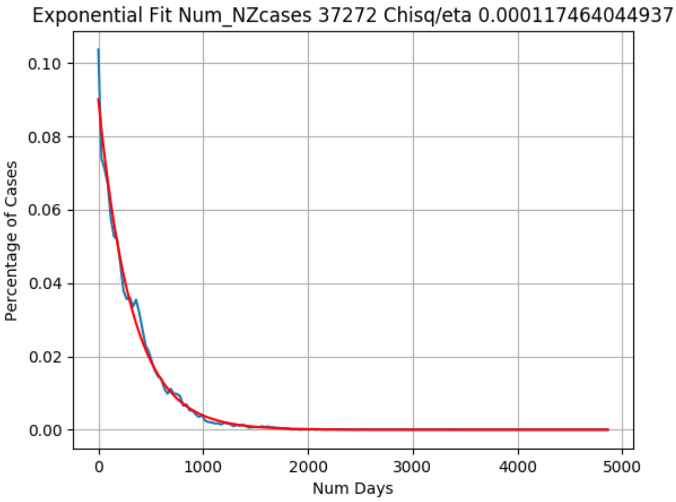


Figure 4. 3 parameter model of a district

Case Type Clustering

Case type is a critical exercise to move further with analysis. This unifies the large number of case types into a small number so that the analysis is simplified and carries more semantics in terms of interpretability and comparability. This standardization also is needed to make the dataset workable which would not b possible in it's primitive state due to problems like

- No fixed case type naming ontology across states. A case with a very specific type in one state can be categorized in a miscellaneous category in a different state due to the absence of that specialized category from the available types for the latter state.

- Token groups like A.. Appl., Appln., Application and spelling mistake variations all mean the same thing i.e. application. This disparity needs to be handled to eliminated redundancy.

- Ambiguous acronyms exist in data such as CriMA which can map to any one of Criminal Miscellaneous Application, Criminal Miscellaneous Appeal, Criminal Municipal Appeal, Criminal Municipal Application etc.

- Lack of standardization in the form of acronyms A.R.B.O.P. maps to Arbitration Original Petition while the acronym suggests that each letter has a separate expansion.

- Some case types like zzEP_CHIT Don't use exist which is hard to infer on what it means or stands for.

- Semantically null case types like R-13 exist which we have no idea what they stood for even after consultation with legal experts.

In this form we had the data for 237 districts which had a total of 16,397 distinct case types. This was brought down to 16 clusters due to our clustering efforts. We implemented a 2-tiered 4-step process to obtain the case type mapping to a cluster. This was done in part programmatically and in part manually due to complex nature of the task.

Server to Website mapping (16,397 types -> 12,821 types)

Server type refers to the case type available in the data that is scraped. This is the data manually entered at the other end and is often error prone. To move towards cleaner and more standardized data we look for mapping to one of the case types in the drop down list available on the website for that district (a website type).

This is done programmatically using a high precision word overlap algorithm that accounts for presence of acronyms and uses Levenshtein distance to accommodate spelling errors. Using this 84.5% of the server types find a match, the rest are carried forward as it is. The performance of the algorithm was measured on a manually labelled sample of 150 types. We obtain a precision of 97.3% and a recall of 93.1% yielding a F measure score of 95.15.

State Level Collation and Cleaning (12,821 types -> 3,973 types)

After the previous step we collate all the case types for a given state and then that list goes through a 2 step cleaning procedure. In the first step we use a manually created acronym phrase map with over 2500 entries to get standardized types and correct for spelling mistakes. After this if remaining, acronym expansion is applied token by token and then the list goes through a duplicate removal to yield a list of clean case types for every state.

Manual Clustering (3,973 types -> 16 clusters)

The next step is completely manual since we have reached a limit on the clustering that can be done on lexical basis and introducing semantics into a clustering mechanism that works programmatically is extremely difficult if possible. The solution here was to obtain large amounts of legal domain knowledge and use that to perform manual clustering. All case type were clustered as one of - Criminal Application, Criminal Appeal, Criminal Arbitration, Criminal Case, Criminal Act, Criminal Petition, Civil Act, Civil Application, Civil Arbitration, Civil Case, Civil Appeal, Civil Petition, Special Case, Small Cause Case, Sessions Case, Other.

This labelling task was performed by 4 people so we used Cohen kappa scores to verify the validity and consistency of the mappings. The score was calculated as follows -

$$\kappa = 1 - (1 - p_o)/(1 - p_e)$$

Here p_e is the percentage of overlap and p_o is computed as $\sum n_{k1} n_{k2}/N^2$. Across the labelling performed on 22 different states we had a mean Cohen kappa score of 0.96 which reflects superlative agreement in the labelling performed by the labeller and the verifier.

Predictive Modeling

For the next stage we move on to prescriptive analysis. The attempt here is to learn from the vast amounts of data that is available and be able to predict the time that a case will take for disposal at the time of filing. We take the following as inputs to the model -

- State
- District
- Case type cluster
- Year of filing
- Workload on court

Here the values we attempt to predict are the mean, median and standard deviation of the disposal times for unique input vectors. We have over 18,000 unique input vectors to use for the modeling. Random forests and Neural networks were used for prediction.

Random Forest

Simple forest with 10 estimators node splits when at least 2 data points are present and minimum number of samples needed to be a leaf is 1.

Neural Network

The architecture of the neural network is a simple network with one hidden layer having 100 perceptrons. The input is in form of embeddings, where the embedding dimensions are state id - 10, district id - 15 and cluster type - 5. ReLU is used as the non-linearity, MSE loss used to optimize using Adam.

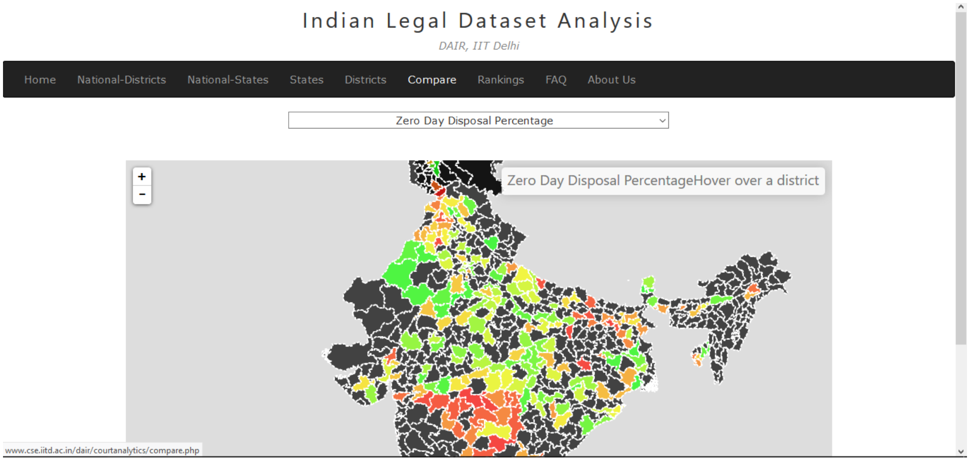
Model	Median absolute error (days)	Median percentage error
Neural Network (Mean)	97.3	27%
Neural Network (Median)	113.2	31%
Neural Network (Std Dev)	46.3	19%
Random Forest (Mean)	69	-

Web Portal

The web portal is available at www.cse.iitd.ac.in/dair/courtanalytics an contains the following panes -

- **Home -** The welcome page of the website

- **National-Districts -** This pane shows the GeoJSON visualization of different metrics on a district level granularity



Indian Legal Dataset Analysis

DAIR, IIT Delhi

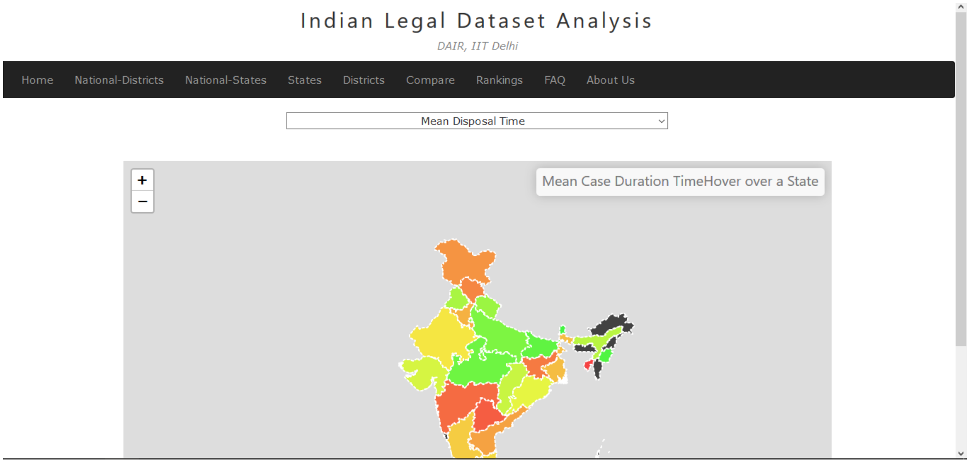
Zero Day Disposal Percentage

Zero Day Disposal PercentageHover over a district

www.cse.iitd.ac.in/dair/courtanalytics/compare.php

Figure 5. National districts pane for zero day disposal percentage

- **National-States -** This pane shows the GeoJSON visualization of different metrics on a state level granularity



Indian Legal Dataset Analysis

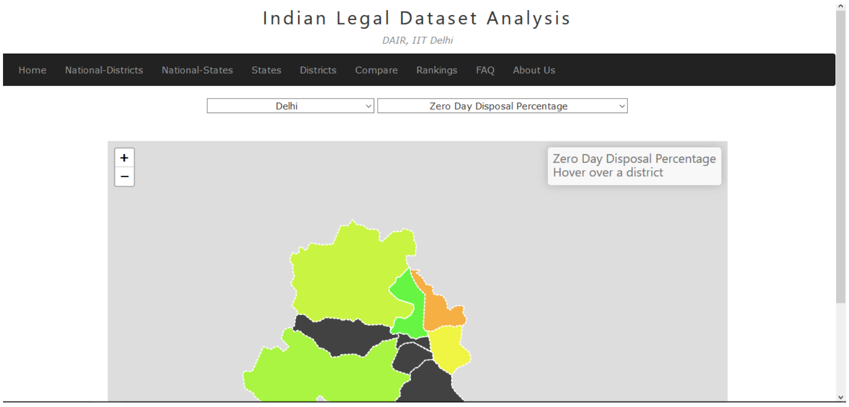
DAIR, IIT Delhi

Mean Disposal Time

Mean Case Duration TimeHover over a State

Figure 6. National states pane for mean disposal time

- **States -** This pane shows the GeoJSON visualization of different metrics for a given state.

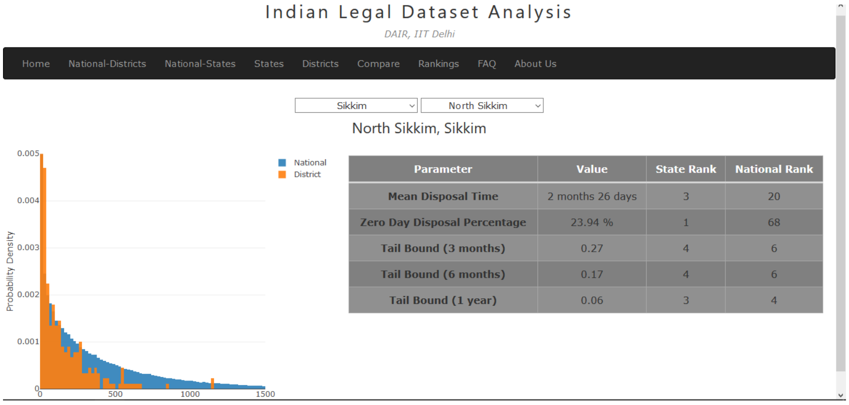


Delhi
Zero Day Disposal Percentage

Zero Day Disposal Percentage
Hover over a district

Figure 7. States pane for Delhi for zero day disposal percentage

- **Districts -** Visualizes the disposal time distribution for a district and provides values and rankings on different metrics.



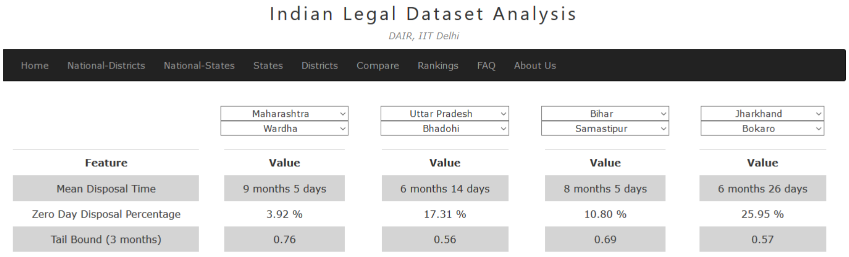
Sikkim
North Sikkim

North Sikkim, Sikkim

Parameter	Value	State Rank	National Rank
Mean Disposal Time	2 months 26 days	3	20
Zero Day Disposal Percentage	23.94 %	1	68
Tail Bound (3 months)	0.27	4	6
Tail Bound (6 months)	0.17	4	6
Tail Bound (1 year)	0.06	3	4

Figure 8. Districts pane for North Sikkim

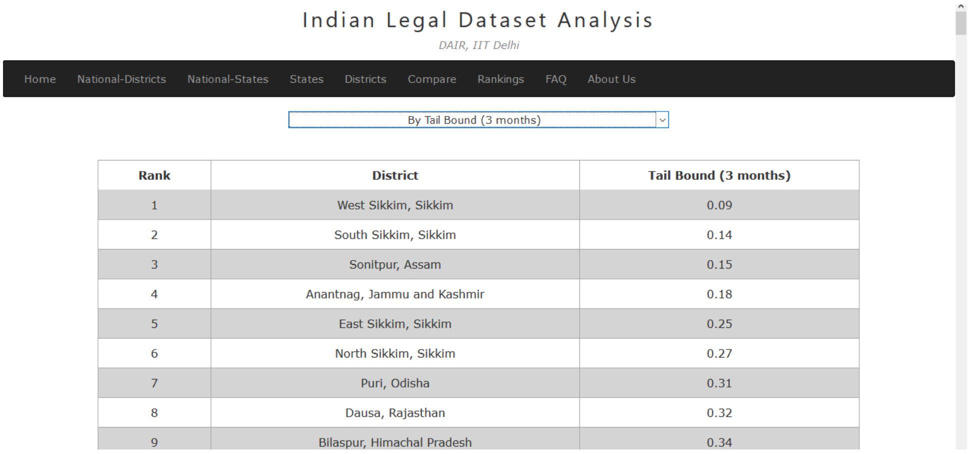
- **Compare -** Compare upto 4 districts on different metrics



	Maharashtra Wardha	Uttar Pradesh Bhadohi	Bihar Samastipur	Jharkhand Bokaro
Feature	Value	Value	Value	Value
Mean Disposal Time	9 months 5 days	6 months 14 days	8 months 5 days	6 months 26 days
Zero Day Disposal Percentage	3.92 %	17.31 %	10.80 %	25.95 %
Tail Bound (3 months)	0.76	0.56	0.69	0.57

Figure 9. Compare pane

- **Rankings -** Gives the compiled list of all districts ranked by a metric



Indian Legal Dataset Analysis

DAIR, IIT Delhi

Rank	District	Tail Bound (3 months)
1	West Sikkim, Sikkim	0.09
2	South Sikkim, Sikkim	0.14
3	Sonitpur, Assam	0.15
4	Anantnag, Jammu and Kashmir	0.18
5	East Sikkim, Sikkim	0.25
6	North Sikkim, Sikkim	0.27
7	Puri, Odisha	0.31
8	Dausa, Rajasthan	0.32
9	Bilaspur, Himachal Pradesh	0.34

Figure 10. Rankings pane

- **FAQ -** Answers basic questions and gives general definitions
- **About Us -** About the team that worked on the project