Date: December 2, 2021
To: IDinsight Leadership Team
From: MPA-ID Advisory Team (Tiara Puteri, Dimas Setyonugroho, Liza Maharjan, Ruoshi Kang, Siew Ng), Harvard Kennedy School of Government

Predicting Out of School Girls for Educate Girls in Rural India

Executive summary

In this study, commissioned by Educate Girls, an NGO working to increase school enrollment and educational attainment of girls in rural India, we developed statistical models to predict the number of out-of-school girls for each village to guide Educate Girls' strategy on specific villages to be targeted.

Using Village Census, District Information System for Education (DISE), and Door-to-Door data with a total of 6,697 observations and 517 variables), we built 12 models using ordinary least square regression (OLS), machine learning techniques of Least Absolute Shrinkage and Selection Operator (LASSO) and Random Forest as well as using a two-tiered parent-child modeling approach. We use the lowest average mean-squared error (MSE) to guide our model selection process.

We considered incorporating Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) data. However, due to missing observations in SHRUG, the increase in the number of variables, on average, did not meaningfully reduce MSE.

Based on our analysis, we find:

1. **OLS models generally perform less robustly** compared to simple LASSO and Random Forest models.
2. **The two-tiered parent-child modeling approach performs better than the simple LASSO and Random Forest models**. However, the parent-child approach requires two stages of prediction which increases the uncertainty of its predictions.
3. Therefore, **we recommend Educate Girls to adopt a pure random forest model with an MSE of 1,092.6 as the best algorithm.**

However, the study is limited by the following factors:

1. **High variabilities in the number of out-of-school girls (OOSG) and number of villages in each district**, affecting the certainty of the predictions. Instead of using a simple average for MSE across all districts, a better approach might be to scale the MSE according to variance of OOSG to evaluate each algorithm.
2. **Limited amount of data** since 14 districts constitute less than 2% of India's 718 districts. Improving data availability will improve model performance.
3. **Key gaps in understanding of the socio-political, administrative, and geographic conditions** that might be additional barriers hindering girls' school attendance. This requires local contextual expertise on the significance of specific predictors.

That said, we believe this initial prediction will enable Educate Girls to build its expansion strategy in the short and medium run.

Introduction and hypothesis

We hypothesized that there could be several binding constraints on the supply and demand side preventing girls from attending schools. Hence, in the OLS models, we selected up to 15 variables which could indicate:

Demand-side Predictors:
1. Costs of schooling
2. Lifetime returns to schooling
3. Parental preferences
4. Children's endowments
5. Liquidity or credit constraints

Supply-side Predictors:
1. Adequate number of schools
2. Quality of schools
3. Student teacher ratio

To avoid relying on human discretion, we also used all available variables in the OLS, LASSO, and Random Forest models.

As an attempt to use a combination of models to improve prediction, we tested a two-tiered parent-child approach by:

1. Categorizing all variables into 9 socio-economic categories and hypothesizing that
    a. **Parent's literacy:** Literate parents are more likely to enroll their daughters in school
    b. **Caste and religion:** Selected religions or castes may hinder girls from attending school
    c. **Occupation:** More occupation opportunities indicate higher opportunity cost of going to school and so, might decrease girls' enrollment
    d. **Assets & Housing condition:** Wealthier families are more likely to have funds to enroll their daughters in school
    e. **Size of Household:** Larger families may only be able to afford enrolling some of the children in school and boys might be prioritized over girls
    f. **School facilities:** Poor school facilities might be regarded by parents as indicators of poor education quality and so, might decrease girls' enrollment
    g. **School type:** Availability of secondary schools or other forms of further education in the village might motivate parents to enroll their daughters in school
    h. **Sex ratio in school:** Higher ratio of enrolled boys to girls at various levels of schooling may reveal gender preferences among the community.
    i. **Others:** Include factors available in Census data that are not accounted above

2. Identifying the model (OLS / LASSO / Random Forest) that resulted in lowest MSE for each category - this creates 9 individual child models

3. Constructing a parent model either by computing a simple average or regressing predictions from all 9 child models to produce final MSE for each district

Data preparation

We used the following data sources in the analysis for the years 2016, 2017, 2019, and few additions from 2020:

1. Door-to-door data from Educate Girls (EG),
2. DISE data, and
3. Village Census data,

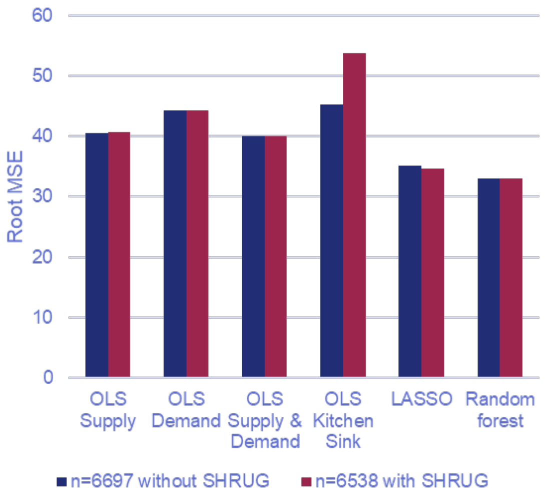


Figure 1: MSE without and with SHRUG data

We initially included SHRUG data to utilize the additional socio-economic variables at the village level. After merging, the number of observations were reduced from 6,697 to 6,538 observations due to missing observations in SHRUG, and the number of variables increased from 517 to 718 variables. However, on average, this did not meaningfully reduce MSE. This could be due to additional variables in the SHRUG dataset being correlated with or proxied by existing variables in DISE and Village Census dataset. Since incorporating SHRUG data did not add value to the overall predictions, the final analysis excludes SHRUG data.

Modeling and Findings

Selected Supply & Demand Variables

Based on the hypothesized demand-side predictors, we shortlisted 12 variables indicating demand-side constraints. Based on the hypothesized supply-side predictors, we shortlisted 10 variables indicating supply-side constraints. An OLS model was developed each for the (1) demand-side predictors, (2) supply-side predictors, and (3) a combination of demand and supply side predictors.

All models were tested using a holdout sample. The model combining demand and supply side predictors produced the lowest average MSE of 1,604.9 for the holdout sample (see Table 2 in Technical Appendix) compared to only using demand-side or supply-side predictors. However, since the variables were shortlisted based on discretion, significant variables may be omitted.

Using All Available Variables

To mitigate selection bias, all available variables were used to develop an OLS model (termed as 'kitchen sink'), a LASSO model and a Random Forest model. For the OLS model - although average MSE on the training dataset significantly decreased, average MSE on the test dataset significantly increased. This could be due to overfitting of the training data

For the LASSO and Random Forest models - both average MSE on the training dataset and average MSE on the test dataset significantly decreased. Random Forest using all variables produced the lowest average MSE of 1,092.6 and we find that machine learning techniques yield better predictions than OLS models.

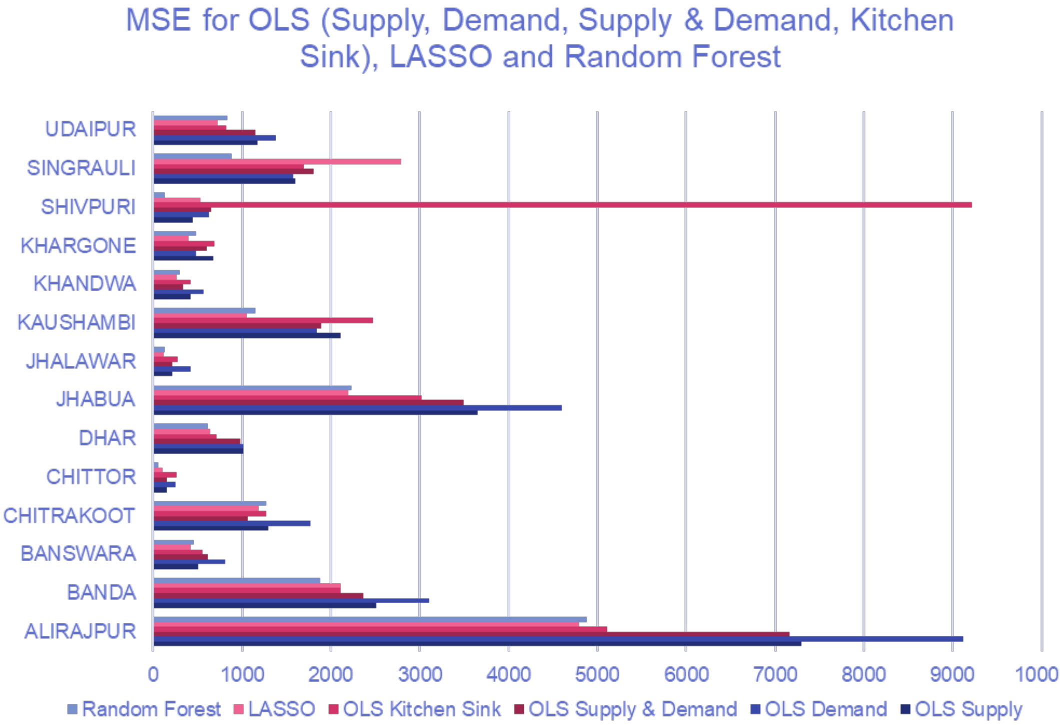


Figure 2: MSE for OLS (Supply, Demand, Supply & Demand, Kitchen Sink), LASSO and Random Forest

Given the OLS model with combination of demand and supply side predictors performed better than the OLS model with all available variables, we explored categorizing all variables into 9 socio-economic categories and fitting models on all the categories to yield better predictions.

Two-tiered Parent-Child Approach

All variables are categorized into one of the 9 socio-economic categories: (1) parent's literacy, (2) caste and religion, (3) occupation, (4) assets and housing condition, (5) household size, (6) school facilities, (7) school type, (8) sex ratio in school, and (9) other factors. This creates 9 individual child models.

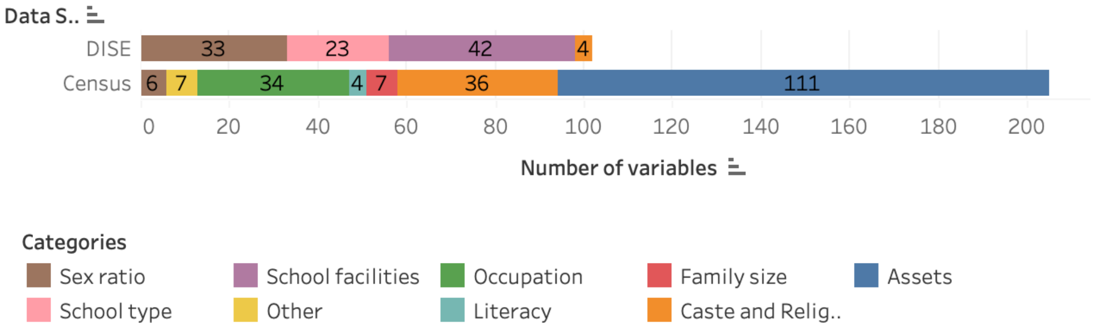


Figure 3: Categories of variables

To aggregate the predictions from all 9 child models, we explored 4 different methods for the parent model:

Method 1: For each category, all three types of models (OLS, LASSO, and Random Forest) were developed. The MSE for each district was computed based on a simple average according to each model type. Of all model types, Random Forest yields the lowest MSE for the test data.

Method 2: For each category, the model type which yields the lowest MSE was selected. For category (1) parent's literacy and (7) school type, LASSO was selected and for all other categories, Random Forest was selected. The MSE for each district was computed based on a simple average, which is lower than only using Random Forest as the model type in Method 1.

Method 3: Combination of LASSO and Random Forest models as per Method 2 were developed, using all available observations. Instead of using a simple average, an OLS model is used to regress the predictions and calculate the weightage for each child model. This method yields the lowest MSE of 913.5 for the test data, compared to all other methods.

Method 4: Combination of LASSO and Random Forest models as per Method 2 were developed, using only observations from 13 districts. Instead of using a simple average, a Random Forest model is used to regress the predictions and calculate the weightage for each child model.

Limitations and Recommendations

We recommend Educate Girls to adopt a pure random forest model with a MSE of 1,092.6 as the best algorithm. This is to avoid any discretion bias, account for interaction between the variables, and ensure that the model can be generalized to future datasets that EG might have.

However, the study is limited by

1. High variabilities in number OOSG and number of villages in each district, affecting the uncertainty of the predictions. A better approach might be to scale the MSE according to variance of OOSG to evaluate each algorithm
2. Limited amount of data since 14 districts constitute less than 2% of India's 718 districts. Improving data availability will improve model performance
3. Key gaps in understanding of the socio-political, administrative and geographic conditions that might be additional barriers hindering girls' school attendance. This requires local contextual expertise on the significance of specific predictors
4. Missing values leading to elimination of some data points. Having a more complete data or using imputation to retain could help improve prediction
5. Two levels of uncertainty in the two-tiered parent-child approach. This requires further research of Bayesian hierarchical modeling to generate a reasonable confidence interval

Technical Appendix

This section of the study explains in more detail the data, models, and other statistics used in this study, including the findings that will help readers better understand the context of the results shown in the previous sections.

Data

There are several datasets considered for this study,

1. Door-to-door (D2D) data from Educate Girls (EG)
   The D2D data from EG has a number of out-of-school girls aged 5 to 14 from 8,980 villages in India. The data covers 4 years (2016, 2017, 2019, and 20214) and 14 districts (Alirajpur, Banda, Banswara, Chitrakoot, Chittor, Dhar, Jhabua, Jhalawar, Kaushambi, Khandwa, Khargone, Shivpuri, Singrauli, and Udaipur).
2. District Information System for Education (DISE)
   DISE data covers education-related indicators across 21,044 villages in India. In general, DISE dataset has different sub data as follows:
   a. Basic information; e.g. average number of computers per school, number of schools in village, etc.
   b. General information; e.g. year school is established, school is managed by the Department of Education, etc.
   c. Enrollment data; e.g. number of enrolled girls from grade 1 to 8, number of enrolled boys from grade 1 to 8, etc.
   d. Facility data; e.g. school electricity, number of girl toilets in school, no school building, etc.
   e. RTE data; e.g. school is approachable by road, school receiver textbook, etc.
   f. Teacher data; e.g. number of female teachers in school, number of teachers with graduate level degree or above, etc.
3. Census data
   Census data covers extensive information on socio-economic characteristics of the village. In general, the census data has:
   a. Assets ownership among households in each village, such as television, mobile phone, car, etc.
   b. Cooking fuel used by households in each village
   c. Housing characteristics (e.g. floor, wall, roof, etc.)
   d. Household size
   e. Sanitation
   f. Source of drinking water
   g. Total employment data
   h. Male employment data
   i. Female employment data
   j. Demographic data (e.g. marriage, caste, etc.)

4. Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG)
   SHRUG data contains extensive socioeconomic data across the whole regions of India. Given the extent of extensiveness that this study already has from DISE and census data, this study only picks several topics in the SHRUG data that would probably add value to the analysis.
   a. Socio-economic caste census data
   b. Electrification data
   c. Non-farm employment data
   d. Road and remote sensing data
   e. Night lights data

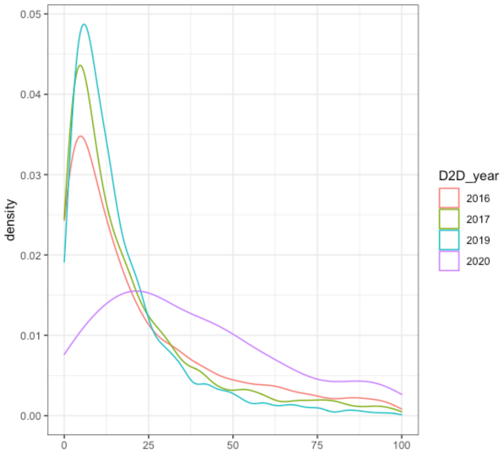
Considering the SHRUG data, there are two possible combinations of number of observations and number of variables in this study. The combinations are:

1. Combination 1: D2D + DISE + Census: 6697 observations across 517 variables
2. Combination 2: D2D + DISE + Census + SHRUG: 6,022 observations across 718 variables

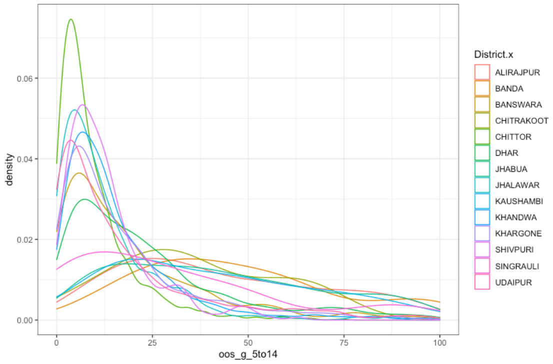
Since this study chooses to employ Combination 1, any prediction results and other related calculations in this study is referring to 6,697 observations that have complete D2D, DISE, and Census data.

Summary statistics

1. Exploratory data analysis
   In analyzing the data, we find the following,
   a. Variance in the number of OOSGs across different years.
      In the 6,697 observations, there is no substantial variation in OOSGs across different years, especially across 2016, 2017, and 2019. We can neglect the differences that are visually visible in 2020 since the data points in this year are very few (i.e. 162) relative to 2016 (i.e. 2,170), 2017 (i.e. 3,075), and 2019 (i.e. 1,290).

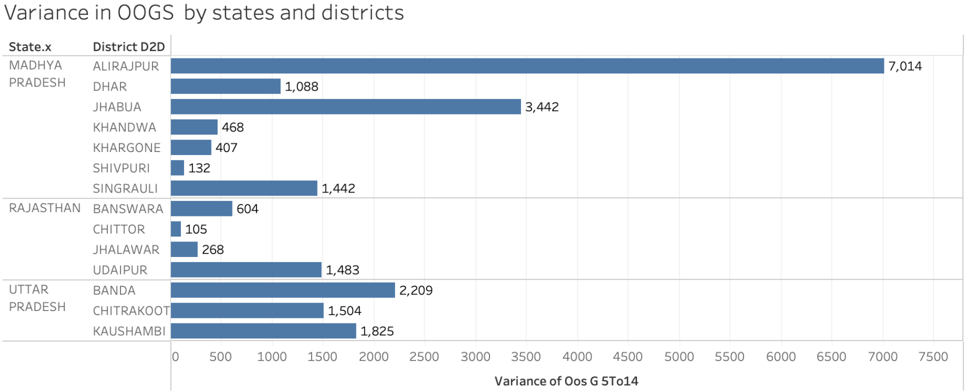


   b. Variation in number of OOSGs across different district
      In the 6,697 observations, there is a variation in OOSGs across different districts. Some districts like Alirajpur, Banda, Chitrakoot, Jhabua, Kaushambi, and Singrauli, on average, seem to have a more intensive number of OOSGs relative to other districts.



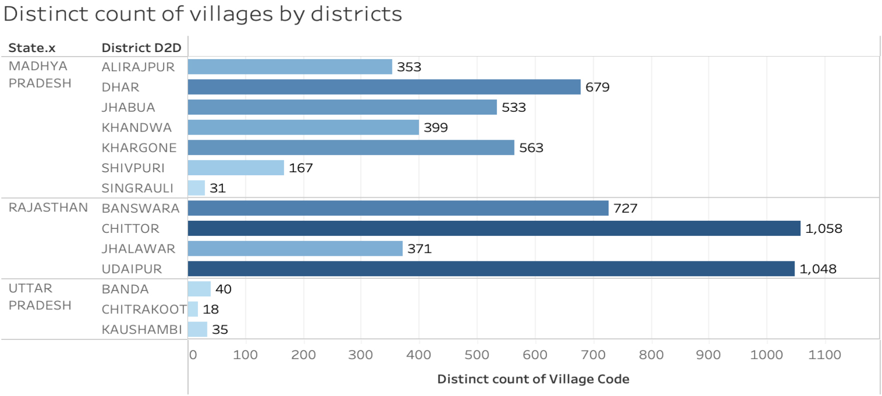
   c. Variance in number of OOSGs across different districts
      The highest variance is found in the Madhya Pradesh state, especially in the Alirajpur

district, with a variance of 7,014; followed by Jhabua, and Bandi.



d. Number of villages per district

The observations mostly come from Madhya Pradesh and Rajasthan states with the most number of villages.



2. Modeling

In this study, we analyze several models in finding the best algorithm, with details in Table 1 below. Except for Model 11, all models were developed using a training dataset of 13 districts and tested on the remaining district. The loop is repeated for every district.

Table 1: Model Description

Model	Technique(s)	Variables	Details
1	Simple OLS	10 Demand variables	Using OLS, variables that may affect demand for education include asset ownership, access to finance, predominantly used cooking fuel, housing characteristics (i.e. main structure, latrine, roof, wall), housing ownership status, source of drinking water, and source of lighting/electricity.
2	Simple OLS	12 Supply variables	Using OLS, variables that may affect supply for education include school approachability by road, school evaluation, electricity in school, English as school's medium of instruction, availability of mid-day meals in school, number of schools in village, student teacher ratio, number of girls toilet in school, textbook availability in school, and availability of water in school.

3	Simple OLS	22 supply and demand variables	Combining variables from the supply and demand model above and then utilizing the OLS regression model to perform statistical analysis.
4	Simple OLS	Kitchen sink (516 variables)	Using all the variables from DISE and census data as the set of independent variables and then utilizing the OLS regression model to perform statistical analysis.
5	Simple LASSO	All 516 variables	This statistical tool performs in a similar way with the OLS but with an effort to shrink, and even nullify, the coefficients of the independent variables by incorporating the coefficients into the minimization algorithm. The goal is to obtain the subset of coefficients that are biased from the standard OLS standpoint but hopefully will reduce the variance and therefore MSEs when this study introduces the LASSO model to out-of-sample data . This model uses all the variables from DISE and census and then imposes a LASSO model on them.
6	Simple Random forest	All 516 variables	Constructing a multitude of decision trees at training using categories of independent variables to make predictions. As this is the regression task, each individual tree returns the mean prediction of the trees. While having many trees improves the prediction job inside the training sample, this will likely overfit on out-of-sample data. In this model, this study is using all the variables from DISE and census and then imposing a random forest model on them.
Parent-child method Grouping the variables into 9 different groups to build the “Child” models: Parent’s literacy, Caste and religion, Occupation, Assets and housing condition, Size of Household, School facilities, School type, Sex ratio in school, and Other factors			
7	Two-tiered Parent Child + OLS	All 516 variables	For each category, an OLS model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
8	Two-tiered Parent Child + LASSO	All 516 variables	For each category, a LASSO model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
9	Two-tiered Parent Child + Random Forest	All 516 variables	For each category, a Random Forest model was developed. The MSE for each district was computed based on a simple average of all 9 child models.
10	Two-tiered Parent Child + Hybrid	All 516 variables	For category (1) parent’s literacy and (7) school type, LASSO was used and for all others, Random Forest was used. The MSE for each district was computed based on a simple average.
11	Two-tiered Parent Child + OLS for	All 516 variables	Combinations of LASSO and Random Forest models as per Model 10 were developed, using all available observations. Instead of using a simple average, an OLS model is used to

	weighted average		regress the predictions and calculate the weightage for each child model. Only at the parent model stage, the OLS model was developed using a training dataset of 13 districts and tested on the remaining district
12	Two-tiered Parent Child + Random Forest for weighted average	All 516 variables	Combination of LASSO and Random Forest models as per Model 10 were developed. Instead of using a simple average, a Random Forest model is used to regress the predictions and calculate the weightage for each child model.

Table 2: MSE of models

Districts	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
ALIRAJPUR	7,295.7	9,113.5	7,167.8	5,107.6	4,790.4	4,875.9	6,767.7	6,747.1	6,332.6	6,286.4	3,931.9	4,734.5
BANDA	2,506.1	3,107.7	2,368.5	2,107.7	2,109.9	1,881.1	3,062.9	3,059.6	2,662.4	2,691.7	1,770.1	2,552.2
BANSWARA	509.8	802.4	608.9	553.7	425.4	457.9	632.7	630.2	667.9	650.8	247.8	421.6
CHITRAKOOT	1,296.3	1,767.4	1,061.1	1,265.2	1,181.2	1,265.3	1,620.1	1,595.5	1342.7	1,342.0	1,051.3	959.6
CHITTOR	157.6	254.9	148.4	263.1	108.6	59.5	229.0	211.3	161.6	165.7	67.3	61.8
DHAR	1,019.7	1017.1	977.2	712.5	635.3	617.5	1042.0	1,016.2	1,018.6	960.6	513.4	643.4
JHABUA	3,645.9	4,593.6	3,498.3	3,021.0	2,189.5	2,230.9	3,745.8	3,411.0	3,027.3	3,041.1	1,874.5	2,033.3
JHALAWAR	210.9	425.4	212.8	276.4	119.9	127.6	263.7	252.9	233.4	239.4	127.5	148.8
KAUSHAMBI	2,109.6	1,846.4	1,892.5	2,478.3	1,050.6	1,149.8	1,879.3	1,772.7	1,743.5	1,742.2	1,252.3	1,296.4
KHANDWA	419.8	564.0	329.2	417.2	260.0	298.1	431.4	407.9	404.2	413.9	233.4	249.7
KHARGONE	671.8	479.5	607.4	686.2	390.1	480.7	543.9	517.1	640.5	588.2	263.2	429.5
SHIVPURI	445.4	626.1	645.3	9,214.4	524.8	128.3	2,979.9	1,073.2	247.8	251.8	115.8	128.7
SINGRAULI	1,595.8	1,577.7	1,806.5	1,692.5	2,782.6	886.2	1,573.2	1,553.8	1,358.7	1,305.3	781.9	879.4
UDAIPUR	1,173.4	1,376.0	1,145.0	826.4	727.6	838.2	1,124.0	1,093.6	1,066.4	1,055.6	558.5	748.4
AVERAGE MSE	**1,647.0**	**1,968.0**	**1,604.9**	**2,044.4**	**1,235.4**	**1,092.6**	**1,849.7**	**1,667.3**	**1493.4**	**1481.1**	**913.5**	**1,092.0**
AVERAGE RMSE	**40.6**	**44.4**	**40.1**	**45.2**	**35.1**	**33.1**	**43.0**	**40.8**	**38.6**	**38.5**	**30.2**	**33.0**