Supplementary Material: Towards Building a District Development Model for India Using Census Data

1 Introduction

This document contains supplementary notes to the original paper. It includes a detailed explanation of several methods, and should be read in conjunction with the relevant sections in the paper.

2 Justification for choice of k

(Refers to Section 3.3 which explains the discretization of variables)

A combination of various tests was carried out to choose the right value of k, ie. the number of levels used to define development in the districts. The value k = 3 for the k-means clustering was carefully chosen after analyzing silhouette plots and elbow plots. A sensitivity analysis was also done by using different values of k to check whether the results remain consistent.

The choice of k = 3 was found to not just be statistically valid, but also makes it simple to interpret the change in levels with 3 classes. The results of individual methods are given in following sections.

2.1 Silhouette plots

The silhouette analysis for k = 2 to 5 shows that the average score is the highest when k = 3 for fuel for cooking, bathroom facility, main source of water, and condition of households. For the type of employment, main source of lighting and asset ownership the average score is higher for k = 2.

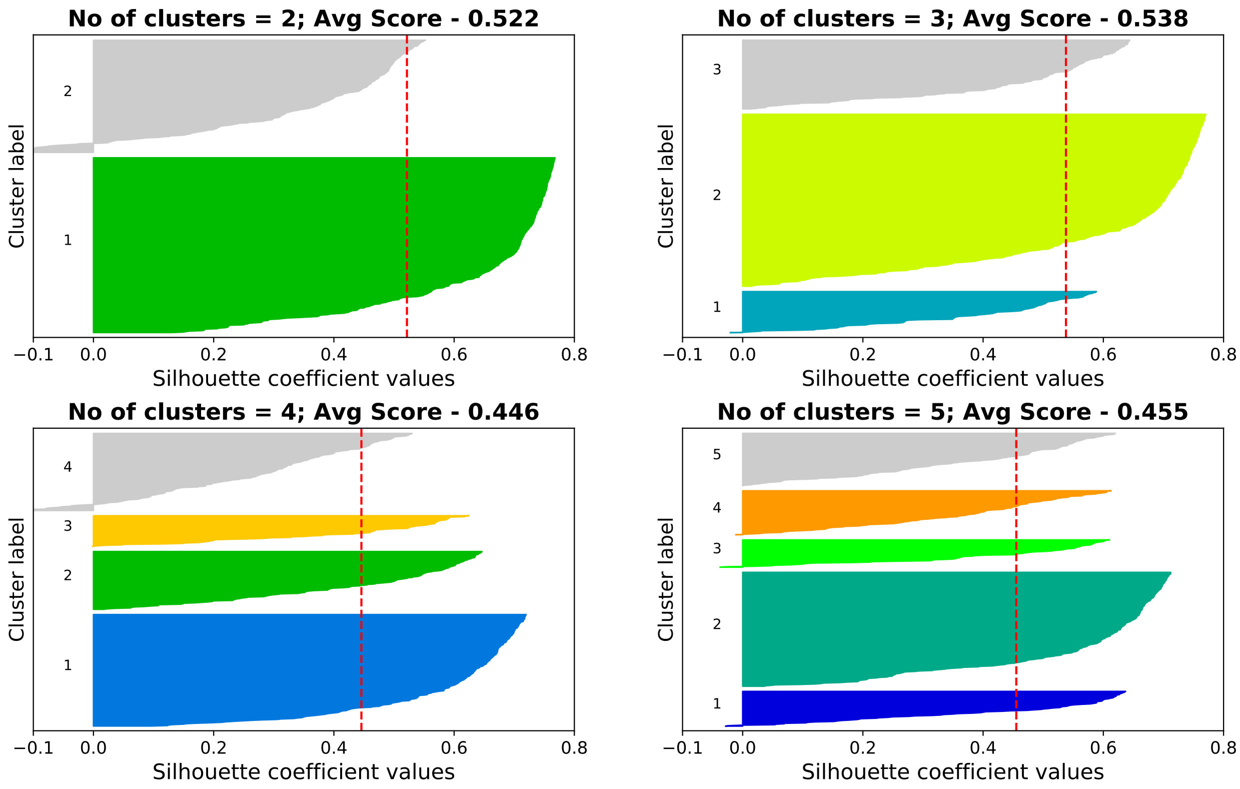


Figure 1: Silhouette plots for clustering with 2-5 clusters : Fuel for Cooking

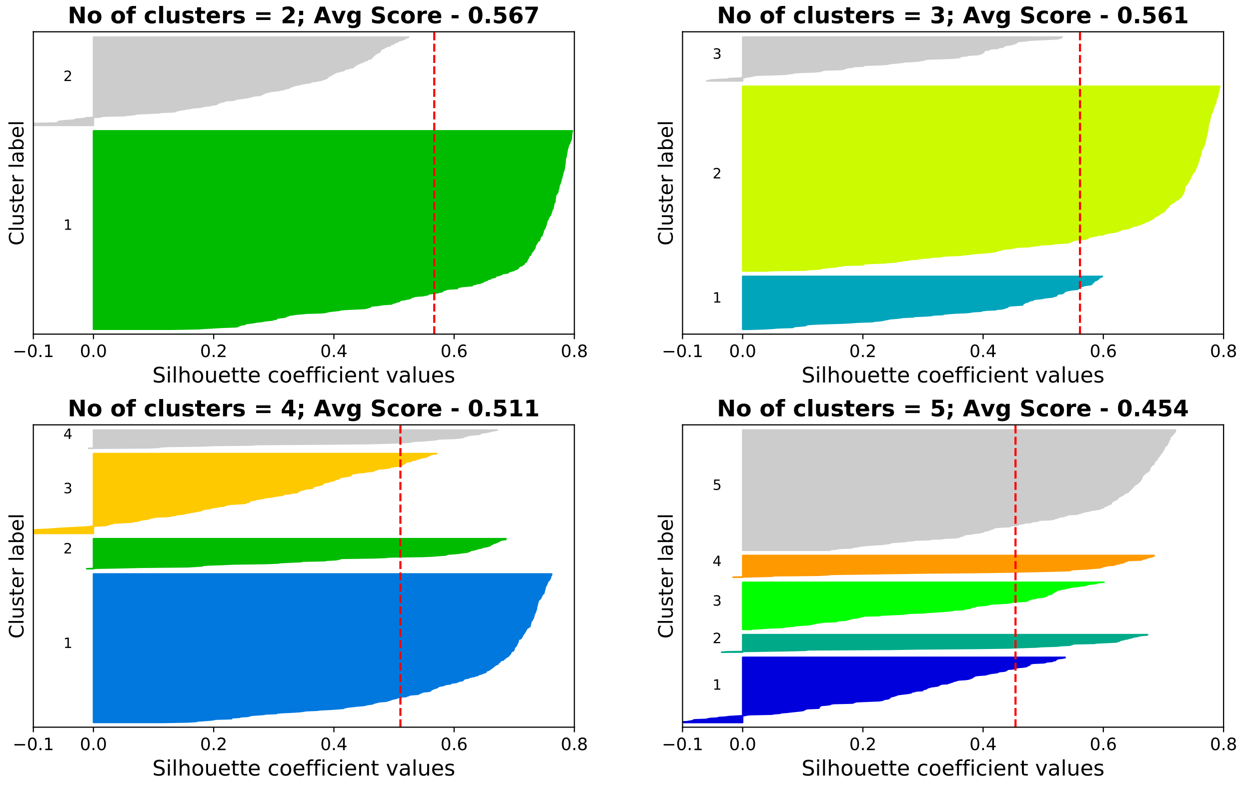


Figure 2: Silhouette plots for clustering with 2-5 clusters : Bathroom facility

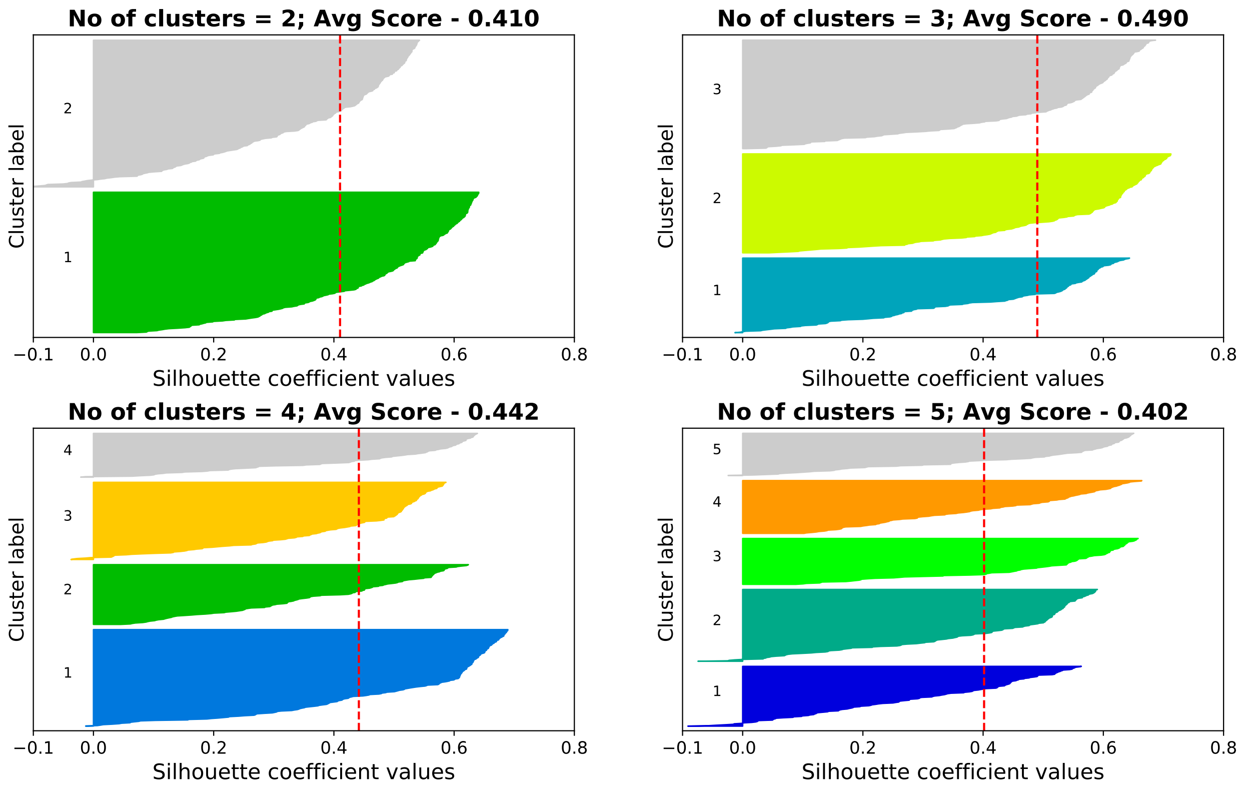


Figure 3: Silhouette plots for clustering with 2-5 clusters : Main source of water

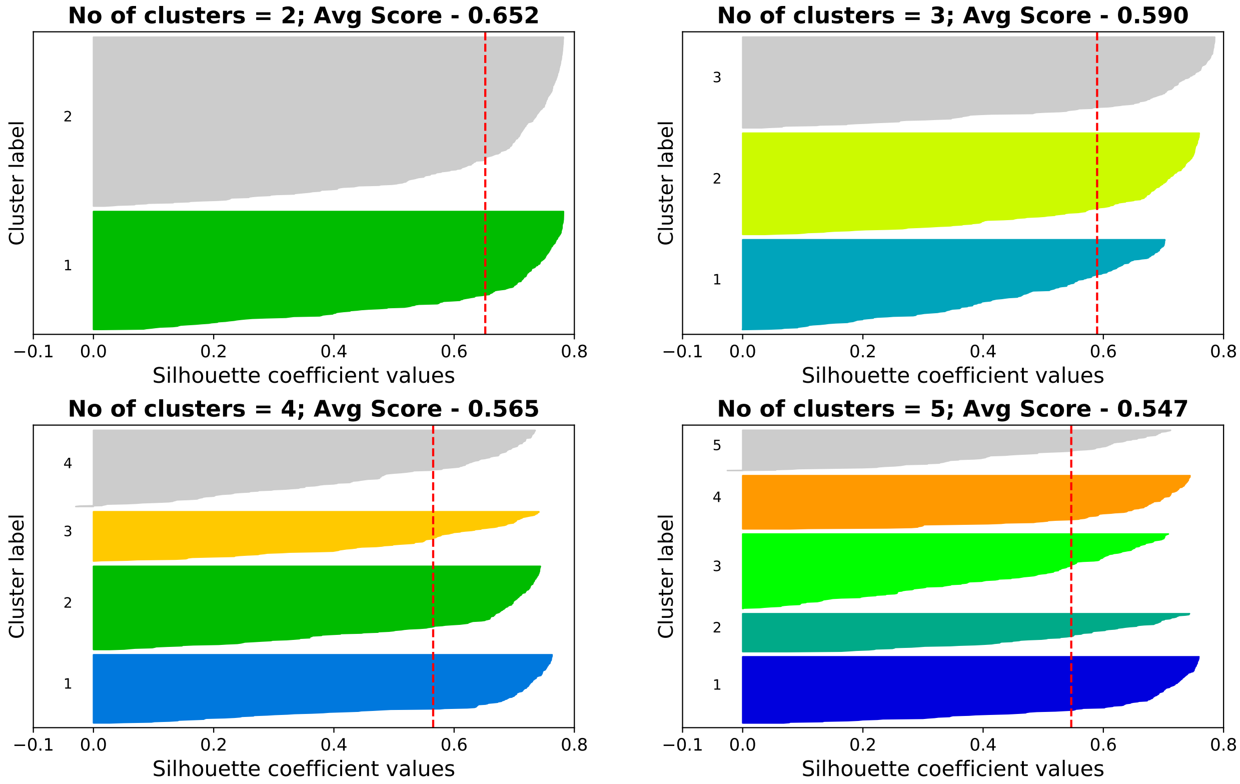
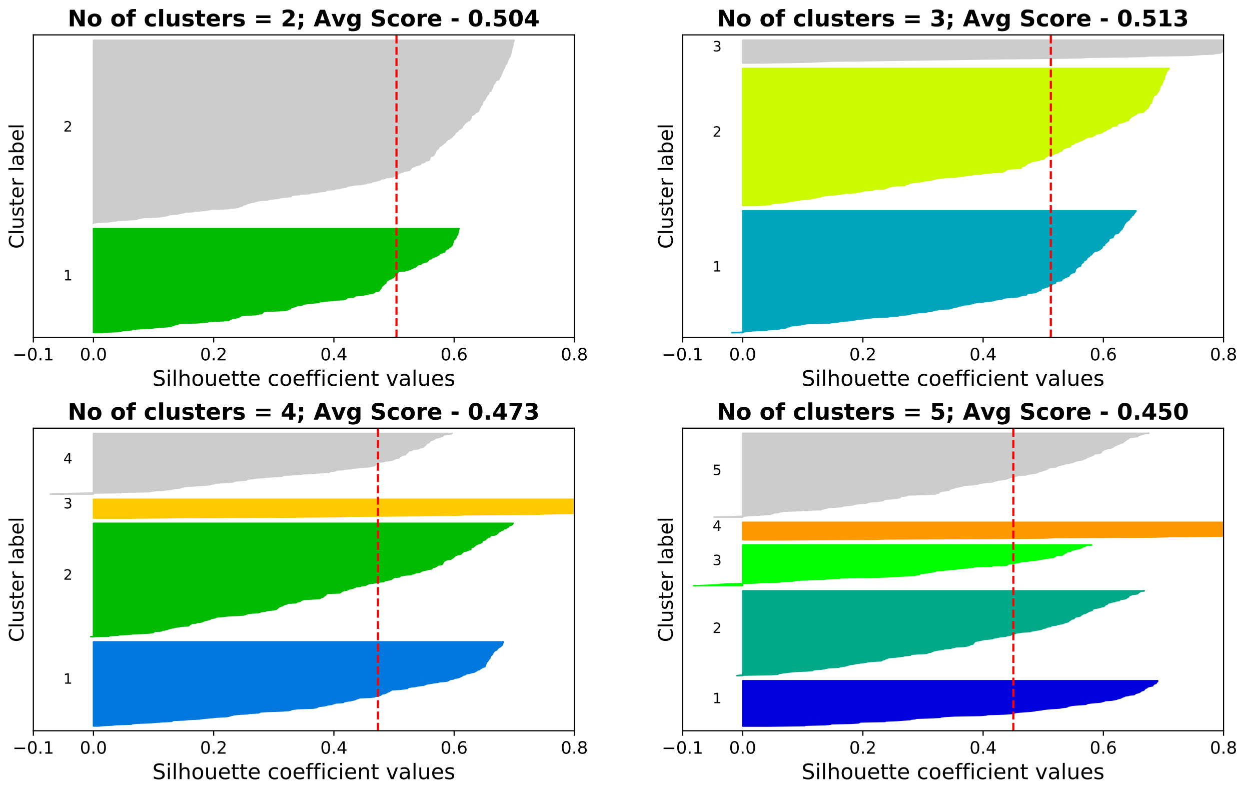


Figure 4: Silhouette plots for clustering with 2-5 clusters : Main source of light

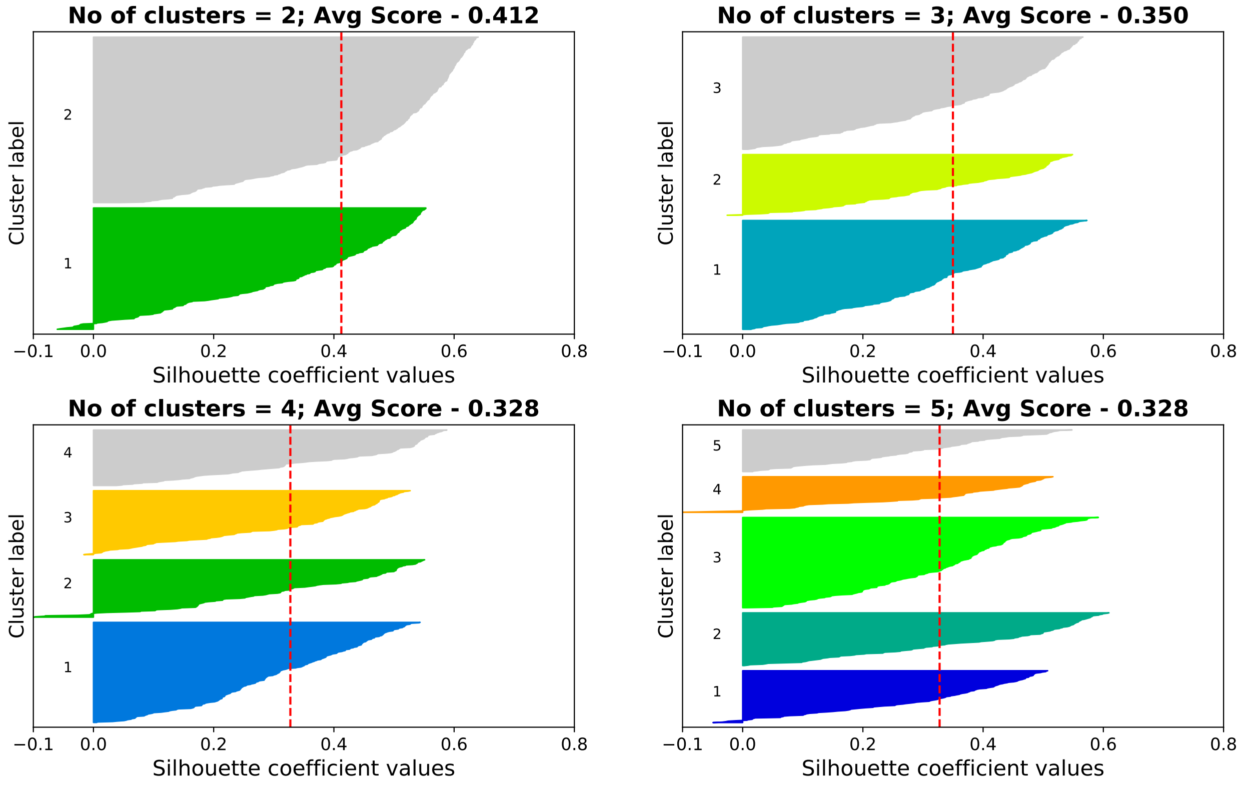

No of clusters = 2; Avg Score - 0.504

No of clusters = 3; Avg Score - 0.513

No of clusters = 4; Avg Score - 0.473

No of clusters = 5; Avg Score - 0.450

Figure 5: Silhouette plots for clustering with 2-5 clusters : Condition of Household


No of clusters = 2; Avg Score - 0.412

No of clusters = 3; Avg Score - 0.350

No of clusters = 4; Avg Score - 0.328

No of clusters = 5; Avg Score - 0.328

Figure 6: Silhouette plots for clustering with 2-5 clusters : Employment

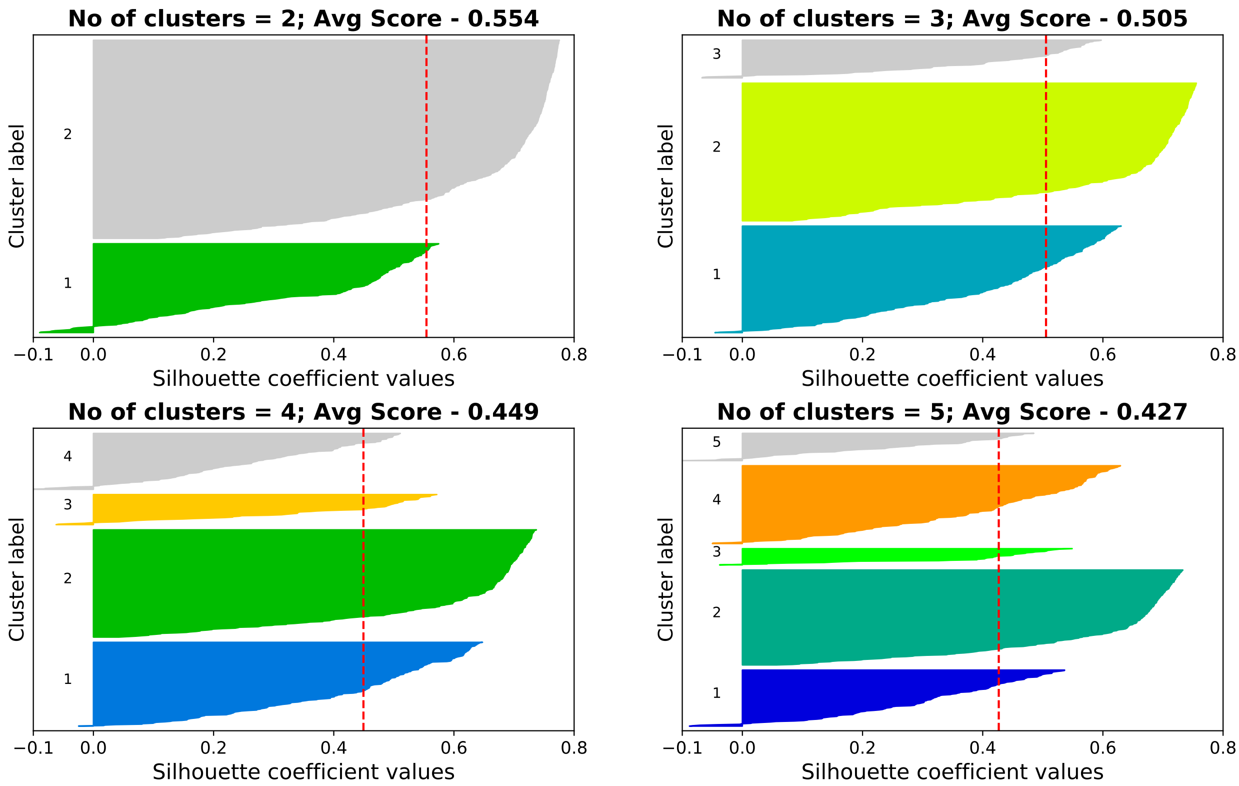


Figure 7: Silhouette plots for clustering with 2-5 clusters : Asset ownership

2.2 Elbow plots

Th elbow plots also point towards a choice of k = 3 as unit distortion on the y axis is below 1 for all the variables for k = 3.

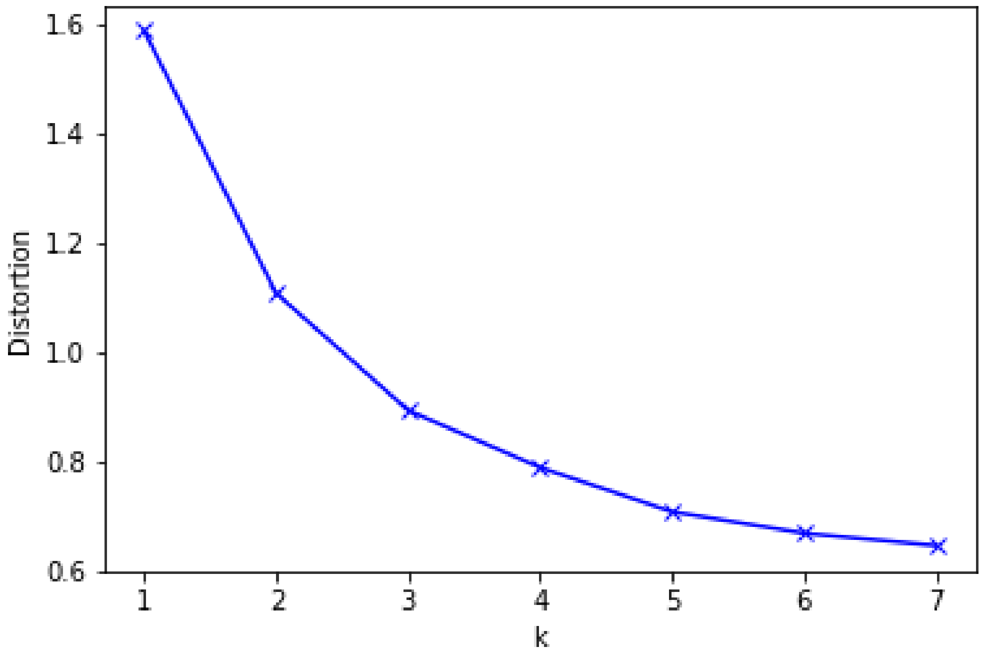


Figure 8: Elbow plot showing optimal k : Asset Ownership

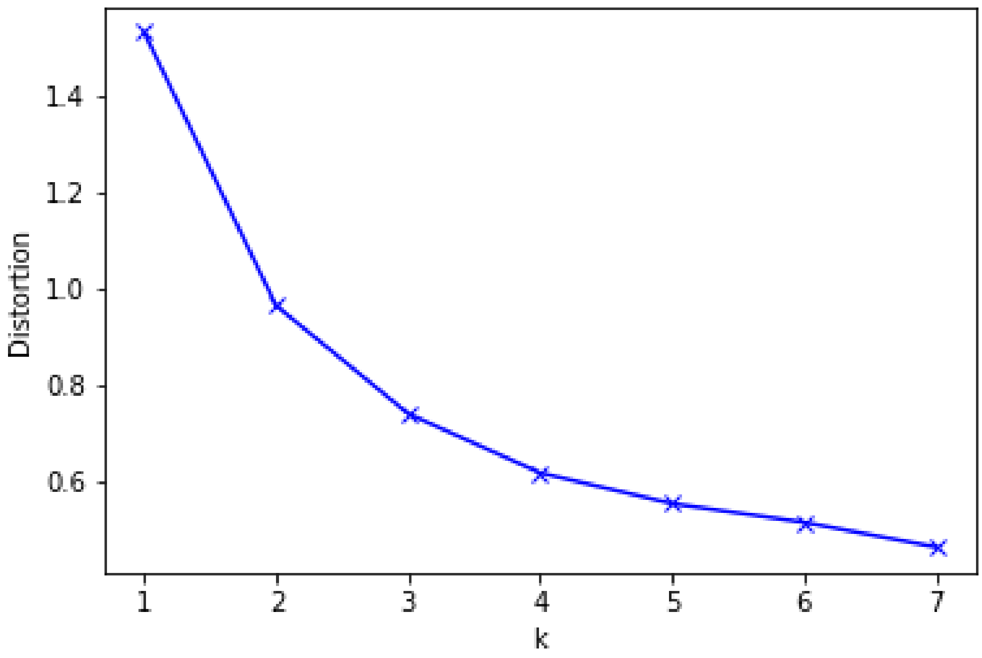


Figure 9: Elbow plot showing optimal k : Bathroom facility

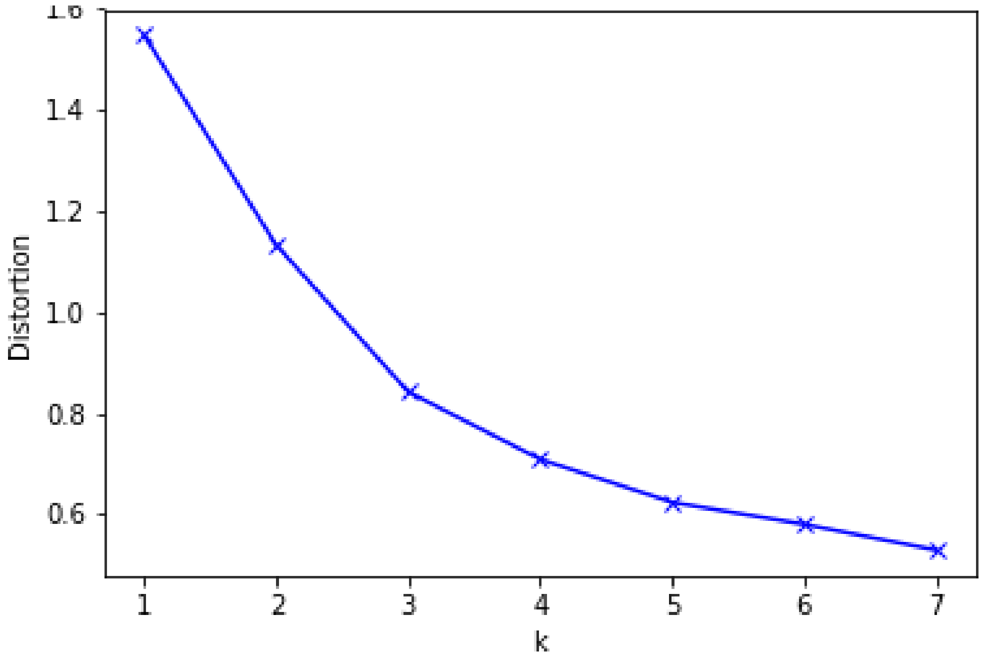


Figure 10: Elbow plot showing optimal k : Fuel for cooking

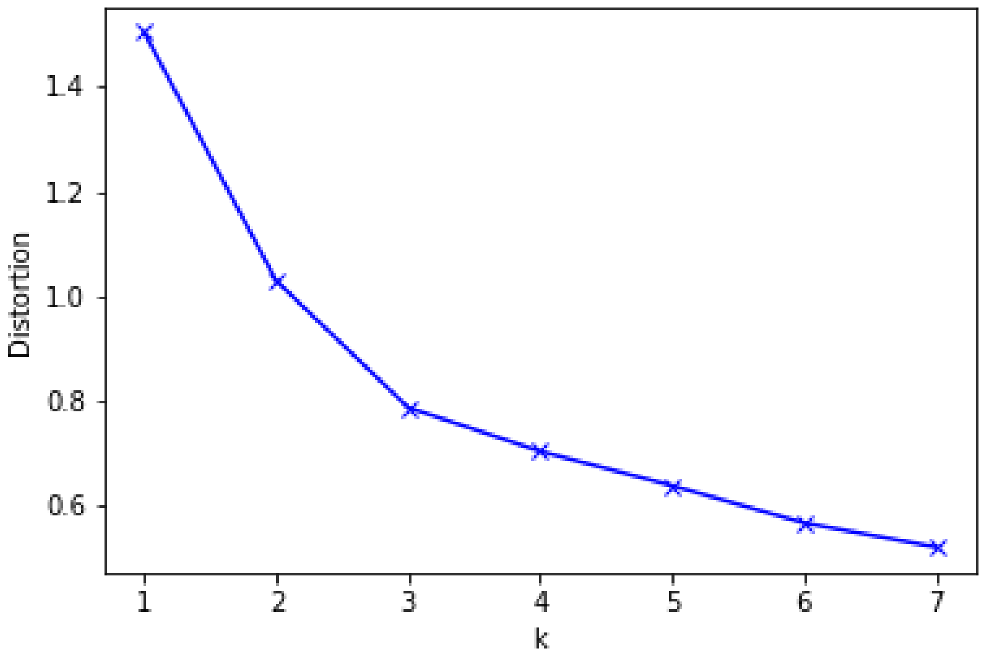


Figure 11: Elbow plot showing optimal k : Condition of household

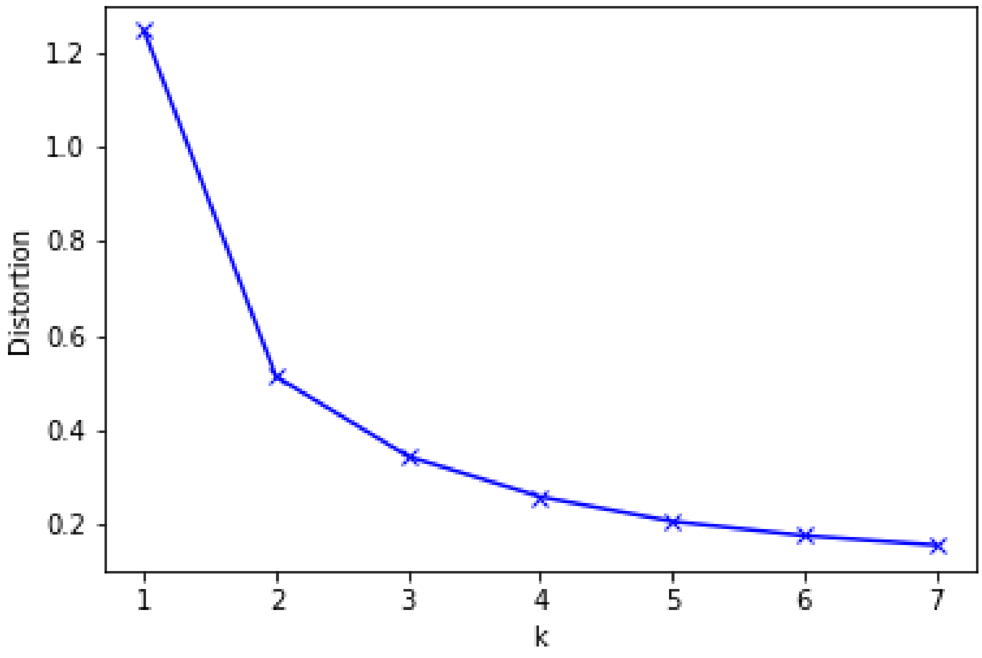


Figure 12: Elbow plot showing optimal k : Main source of light

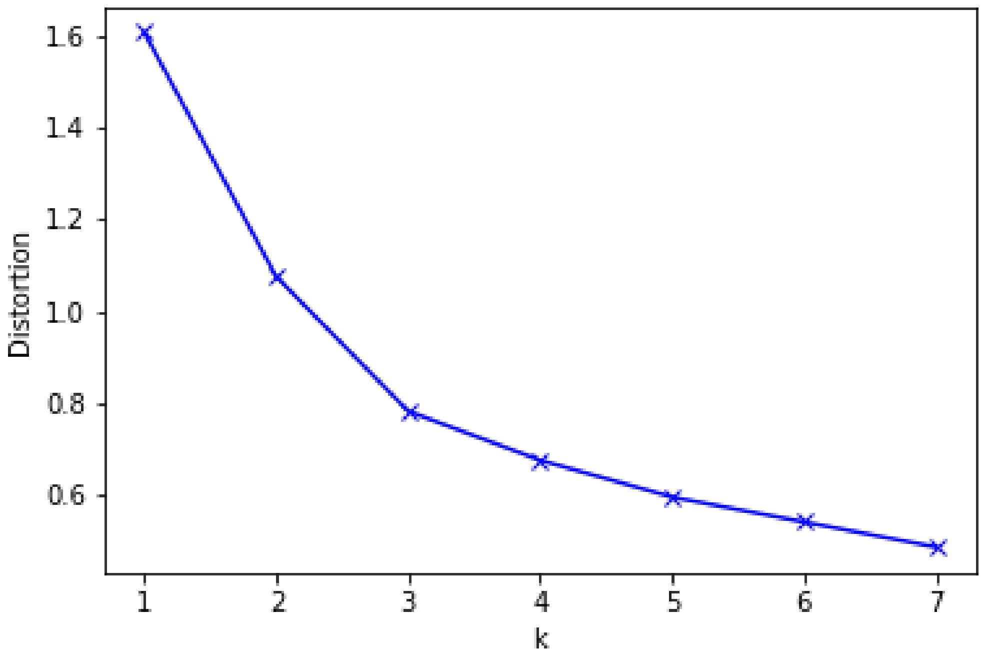


Figure 13: Elbow plot showing optimal k : Main source of water

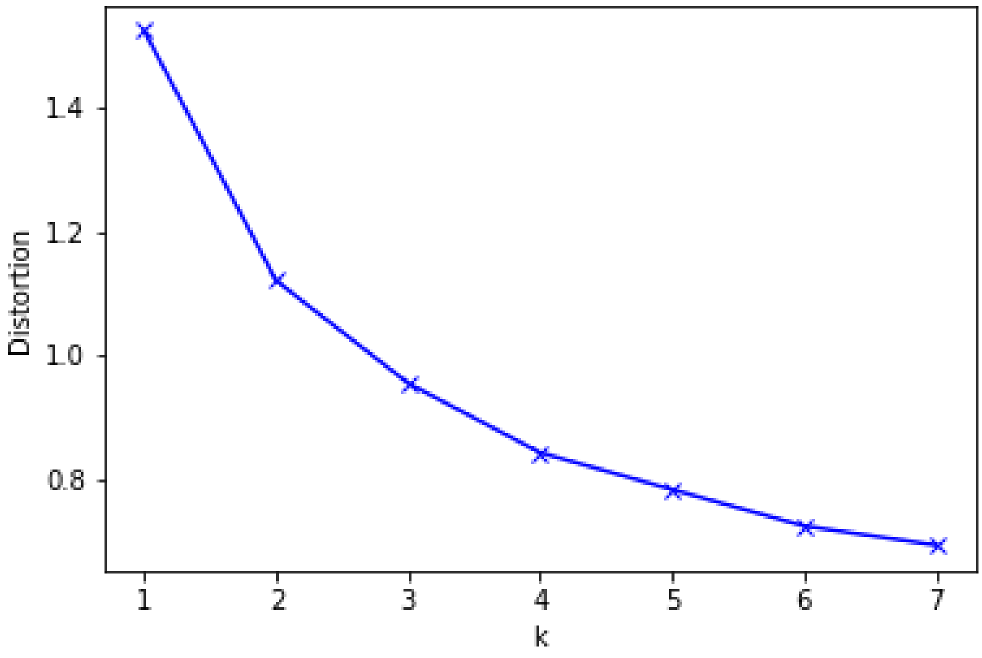


Figure 14: Elbow plot showing optimal k : Type of Employment

2.3 Sensitivity analysis using k = 4

(Refers to Table 4 of Section 4 which explains change in indicators based on the type of employment)

An analysis of the relevant hypothesis was done by using k = 4 as well. Our results are consistent with what has been reported in the paper with k = 3. It shows that the findings are not sensitive to the choice of k. Table 1 shows the change probabilities for k = 4.

Variable	**Existing Status**	**Non Agricultural**	**Agricultural**	**High Unemployment**	**Total**
Asset Ownership	Level-1	0.909	0.592	0.74	0.697
	Level-2	1	0	0	
	Level-3	0.851	0.417	0.917	
Bathroom Facility	Level-1	0.8	0.246	0.206	0.279
	Level-2	0.574	0.444	0.179	
	Level-3	0.647	0.184	0.023	
Fuel for Cooking	Level-1	0.704	0.209	0.138	0.186
	Level-2	0.429	0.143	0.059	
	Level-3	0.417	0	0.059	
Condition of Household	Level-1	0.545	0.364	0.19	0.381
	Level-2	0.733	0.455	0.167	
	Level-3	0.569	0.433	0.357	
Main Source of Light	Level-1	1	0.328	0.316	0.539
	Level-2	0.714	0.686	0.442	
	Level-3	0.821	0.68	0.529	
Main Source of Water	Level-1	0.233	0.5	0.471	0.242
	Level-2	0.556	0.027	0.139	
	Level-3	0.36	0.159	0.14	

Table 1: Change in indicators based on the type of employment for k = 4

- Hypothesis 1: As seen for k = 3, even with k = 4 all the indicators except the main source of water have the highest probability for growth in non agricultural districts.

- Hypothesis 2: Asset ownership shows the highest positive change (0.697), consistent with our findings for k = 3.

- Hypothesis 3: We find that the main source of light has improved more than main source of water ((0.539 as compared with 0.242), consistent with our findings for k = 3.

3 Statistical significance of hypothesis tests

(Refers to Sections 4.1, 4.2, 4.3, 4.5 and 4.6)

We carry out one-tailed z-tests to establish the statistical significance of the various hypotheses. They reinforce the hypotheses quite convincingly.

All the tests are concerned with comparing the *population proportions* corresponding to 2 different groups, p_1 and p_2. Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions corresponding to the 2 different groups, and n_1 and n_2 be the corresponding sample sizes.

The null-hypothesis we want to test, and the corresponding alternate-hypothesis are:

$$H_0 : p_1 = p_2; H_A : p_1 > p_2$$

The Z statistic for testing that the hypothesis is -

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

y_1 and y_2 are the number of positive samples (that do not invalidate the hypothesis) corresponding to the respective groups.

For all the tests, the confidence level is 95% (or p-value is 0.05).

3.1 Hypothesis 1

Non-agricultural districts see the greatest improvement in all indicators.

Here we compare the probability of a positive change given the employment category of the district (High-unemployment, Agricultural, Non-agricultural), with that corresponding to a different employment category. We concluded in the paper that non-agricultural districts show the greatest improvement in all indicators, and that is validated in the Z-score and p-values shown in Table 2. Except the main source of water, all p-values are less than 10^{-24}. We also saw that improvements in the main source of water for non-agricultural districts was not significantly better than in other types of districts, and that is also validated from the p-values (which are both greater than 0.05).

	$p_1 =$ **NAL,** $p_2 =$ **AL**		$p_1 =$ **NAL,** $p_2 =$ **UN**		$p_1 =$ **AL,** $p_2 =$ **UN**	
	Z	**p-value**	**Z**	**p-value**	**Z**	**p-value**
BF	19.78	2.22E-87	19.45	1.52E-84	0.27	0.392
FC	10.49	4.98E-26	8.52	8.24E-18	-1.42	0.078
CHH	10.27	4.67E-25	12.42	1.03E-35	3.45	2.83E-04
MSL	14.37	4E-47	10.47	6.18E-26	6.66	1.4E-11
MSW	-0.52	0.302	1.47	0.071	2.78	0.003
ASSET	20.76	4.66E-96	25.75	1.70E-146	-1.29	0.097

Table 2: Z-score and p-values corresponding to Hypothesis 1. Probabilities of change in different socio-economic indicators compared for districts at different employment levels: NAL = Non-agricultural, AL = agricultural, UN = High-unemployment. Acronyms used for variables: BF = Bathroom facility, FC = Fuel for cooking, CHH = Condition of household, MSL = Main source of lighting, MSW = Main source of water

3.2 Hypothesis 2

Households prefer to invest in assets first, followed by investment in other indicators which they can influence through their own choices.

Here we compare the probability of a positive change in one discretionary variable, with that of another. We concluded in the paper that people invest in assets first, followed by other variables. We can clearly see in Table 3 and 4 that the Z-scores corresponding to the comparison of change in assets with other variables, is significantly higher than 1.96. This is also reflected in the small p-values.

	BF	**FC**	**CHH**
Asset	9.114	15.415	8.872

Table 3: Z-score corresponding to tests between discretionary variables for Hypothesis 2

	BF	**FC**	**CHH**
Asset	3.96E-20	6.46E-54	3.58E-19

Table 4: p-values corresponding to tests between discretionary variables for Hypothesis 2. Row corresponds to p_1, column corresponds to p_2

3.3 Hypothesis 3

Government has prioritized electrification and lighting over other indicators that depend upon government support.

Here we compare the probability of a positive change in the main source of light, with the probability of a positive change in main source of water. The extremely low p-value reinforces our conclusion that government has prioritized electrification over other indicators dependent upon government support.

Z	**p-value**
5.97	1.14E-09

Table 5: Statistical test corresponding to Hypothesis 3. Main source of light corresponds to p_1, main source of water corresponds to p_2

3.4 Hypothesis 5

Districts with more manufacturing and services industries end up developing faster.

A district's industrial presence can be of the following types: Type-4 (High Services), Type-3 (High Manufacturing), Type-2 (Moderate Industrial Presence), or Type-1 (Low Industrial Presence). Consider a discretionary variable such as BF (Bathroom Facility). Let p_1 correspond to the probability of a positive change with respect to BF, given a district's industry is of Type-4 or Type-3. Let p_2 be defined in a similar manner if a district's industry is of Type-2 or Type-1. We perform a statistical test comparing the two, and similarly perform tests for all socio-economic indicators. The results are shown in Table 6. Table 6 shows extremely high Z values (and correspondingly extremely low p-values; significantly lower than 0.05) for all indicators except MSW. This statistically confirms that the improvement in Type-3 and Type-4 districts is more than that in Type-1 and Type-2, which validates the findings in the paper.

	p_1 = **Type-4 or Type-3,** p_2 = **Type-2 or Type-1**	
	Z	**p-value**
BF	18.04	4.59E-73
FC	8.85	4.24E-19
CHH	10.07	3.69E-24
MSL	8.39	2E-17
MSW	0.35	0.362
ASSET	14.61	1.20E-48

Table 6: Z-score and p-values for statistical tests corresponding to Hypothesis-5. The acronyms are the same as stated in the description of Table 2.

3.5 Hypothesis 6

Female participation in the workforce has decreased, primarily with a reduction in marginal employment.

The statistical tests for the last hypothesis are performed in a different manner. For illustration, consider female marginal employment. Every district will be either at Level-1, Level-2, or Level-3 with respect to this variable. Let p_1 denote the probability that a district's female marginal employment is at Level-1 in 2011, and similarly let p_2 denote that probability for 2001. We conduct a statistical test between these two probabilities. A significantly high Z-score would mean that $p_1 > p_2$, ie. more districts are at Level-1 in 2011 as compared to 2001. We conduct the above Z-test for Level-1, Level-2, and Level-3 districts, for both female marginal and female main employment. The results are shown in Table 7.

	Fem Marg Emp		**Fem Main Emp**	
	Z	**p-value**	**Z**	**p-value**
Level-1	7.562	1.98E-14	-0.859	0.195
Level-2	5.671	7.07E-09	-0.427	0.334
Level-3	-11.441	1.31E-30	1.313	0.094

Table 7: Z-scores and p-values for statistical tests corresponding to Hypothesis-6. Refer to the description above for the exact formulation of the statistical tests

In the above table we can see that the Z-scores corresponding to Level-1 and Level-2 female marginal employment are very high. This points to the conclusion that more districts are at Level-1 and Level-2 in 2011 as compared to 2001. The Z-score corresponding to Level-3 female marginal employment is extremely negative. That indicates that less districts in 2011 are at Level-3 female marginal employment as compared to 2001. Since Level-1 corresponds to low marginal employment, and Level-3 corresponds to high marginal employment, we can safely conclude that female marginal employment has significantly reduced. However, if we look at the table corresponding to female main employment, we cannot make any such conclusions. The Z-scores for Level-1 and Level-2 female main employment are negative, which indicate a small increase in female main employment; however, they are not negative enough to reject the null hypothesis. Similarly, the Z-score corresponding to Level-3 is positive, which indicates a greater number of districts at Level-3, ie. an increase in female main employment. However, the p-value is 0.09 which is not significant enough for us to reject the null hypothesis. Therefore, through the statistical tests we cannot conclude that main female employment is increasing, which validates the findings presented in the main paper.

4 Calculation of mutual information

(Refers to Sections 4.4)

The respective tables given below have been used to calculate the mutual information between the four factors of interest in hypothesis 4 (literacy, formal employment, current status, and government support for social infrastructure), and change in each of the four discretionary variables (asset ownership, bathroom facility, fuel for cooking, and condition of household).

	Non Agricultural		**Agricultural**		**High Unemployment**	
Literacy	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.010	0.015	0.192	0.142	0.132	0.125
Level2	0.030	0.064	0.027	0.064	0.029	0.047
Level3	0.039	0.046	0.002	0.024	0.002	0.012
Formal Employment	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.013	0.008	0.128	0.037	0.091	0.051
Level 2	0.005	0.000	0.067	0.125	0.064	0.094
Level 3	0.061	0.116	0.025	0.067	0.007	0.039
Current Status	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.007	0.057	0.214	0.212	0.159	0.142
Level 2	0.015	0.067	0.007	0.017	0.002	0.042
Level 3	0.057	0.000	0.000	0.000	0.002	0.000
Investment in MSL	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.074	0.088	0.152	0.133	0.120	0.130
No Change	0.005	0.037	0.069	0.096	0.042	0.054
Investment in MSW	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.069	0.110	0.170	0.191	0.135	0.157
No Change	0.010	0.015	0.051	0.039	0.027	0.027

Table 8: Probability of (+ve Change/No Change) in Asset ownership based on Type of Employment with respective variables

Literacy	Non Agricultural		Agricultural		High Unemployment	
	No Change	+ Change	No Change	+ Change	No Change	+ Change
Level1	Rs. 0.008	Rs. 0.017	Rs. 0.310	Rs. 0.024	Rs. 0.221	Rs. 0.035
Level2	Rs. 0.025	Rs. 0.069	Rs. 0.061	Rs. 0.030	Rs. 0.046	Rs. 0.030
Level3	Rs. 0.057	Rs. 0.027	Rs. 0.013	Rs. 0.012	Rs. 0.007	Rs. 0.007
Formal Employment	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	Rs. 0.010	Rs. 0.012	Rs. 0.159	Rs. 0.007	Rs. 0.132	Rs. 0.010
Level 2	Rs. 0.000	Rs. 0.005	Rs. 0.164	Rs. 0.029	Rs. 0.118	Rs. 0.040
Level 3	Rs. 0.081	Rs. 0.096	Rs. 0.062	Rs. 0.030	Rs. 0.024	Rs. 0.022
Current status	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	Rs. 0.013	Rs. 0.039	Rs. 0.327	Rs. 0.054	Rs. 0.211	Rs. 0.044
Level 2	Rs. 0.019	Rs. 0.074	Rs. 0.057	Rs. 0.012	Rs. 0.057	Rs. 0.029
Level 3	Rs. 0.059	Rs. 0.000	Rs. 0.000	Rs. 0.000	Rs. 0.005	Rs. 0.000

Investment in MSL	No Change	+ Change	No Change	+ Change	No Change	+ Change
+ Change	Rs. 0.064	Rs. 0.098	Rs. 0.243	Rs. 0.042	Rs. 0.196	Rs. 0.054
No Change	Rs. 0.027	Rs. 0.015	Rs. 0.142	Rs. 0.024	Rs. 0.078	Rs. 0.019

Investment in MSW	No Change	+ Change	No Change	+ Change	No Change	+ Change
+ Change	Rs. 0.305	Rs. 0.056	Rs. 0.305	Rs. 0.056	Rs. 0.231	Rs. 0.061
No Change	Rs. 0.079	Rs. 0.010	Rs. 0.079	Rs. 0.010	Rs. 0.042	Rs. 0.012

Table 9: Probability of (+ve Change/No Change) in Bathroom facility based on Type of Employment with respective variables

	Non Agricultural		Agricultural		High Unemployment	
Literacy	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.019	0.007	0.319	0.015	0.241	0.015
Level2	0.074	0.020	0.083	0.008	0.064	0.012
Level3	0.071	0.013	0.020	0.005	0.012	0.002
Formal Employment	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.020	0.002	0.162	0.003	0.137	0.005
Level 2	0.003	0.002	0.182	0.010	0.148	0.010
Level 3	0.140	0.037	0.078	0.015	0.032	0.013

Current Status	No Change	+ Change	No Change	+ Change	No Change	+ Change
Level1	0.047	0.022	0.349	0.029	0.140	0.012
Level 2	0.008	0.019	0.067	0.000	0.164	0.017
Level 3	0.108	0.000	0.005	0.000	0.013	0.000
Investment in MSL	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.125	0.037	0.263	0.022	0.228	0.022
No Change	0.039	0.003	0.159	0.007	0.089	0.007
Investment in MSW	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.140	0.039	0.337	0.024	0.270	0.022
No Change	0.024	0.002	0.084	0.005	0.047	0.007

Table 10: Probability of (+ve Change/No Change) in Fuel for cooking based on Type of Employment with respective variables

	Non Agricultural		**Agricultural**		**High Unemployment**	
Literacy	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.013	0.012	0.265	0.069	0.224	0.032
Level2	0.062	0.032	0.059	0.032	0.059	0.017
Level3	0.037	0.047	0.015	0.010	0.008	0.005
Formal Employment	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.010	0.012	0.142	0.024	0.132	0.010
Level 2	0.003	0.002	0.140	0.052	0.130	0.029
Level 3	0.099	0.078	0.057	0.035	0.030	0.015
Current Status	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
Level1	0.010	0.025	0.108	0.061	0.137	0.032
Level 2	0.046	0.066	0.165	0.051	0.130	0.022
Level 3	0.057	0.000	0.066	0.000	0.025	0.000
Investment in MSL	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.089	0.073	0.219	0.066	0.218	0.032
No Change	0.024	0.019	0.120	0.046	0.074	0.022
Investment in MSW	**No Change**	**+ Change**	**No Change**	**+ Change**	**No Change**	**+ Change**
+ Change	0.094	0.084	0.270	0.091	0.248	0.044
No Change	0.019	0.007	0.069	0.020	0.044	0.010

Table 11: Probability of (+ve Change/No Change) in Condition of household based on Type of Employment with respective variables

5 Prediction Of Change In Discretionary Variable

(Refers to Sections 4.4 of main paper)

We created two classification models to see if we can predict the change in discretionary variables. Since we wanted to train a model for a response variable that is dichotomous *positive change and non-positive change in the discretionary variables*, we used a logistic regression model to predict the two classes. In the first model, we used the current status of all six socio-economic variables as the features to predict the outcome. In the second model, we also added variables for formal employment and literacy. The data consisting of the entire set of districts was split into an 80:20 ratio for training and testing, with a 5-fold cross-validation. We use the SMOTE (Synthetic Minority Oversampling Technique) method [Chawla, Bowyer, Hall, and KegelmeyerChawla et al.2002] on the training dataset to address class imbalance issues. SMOTE creates new minority class instances (synthetic) between existing (real) minority instances. Table 12 shows the results for both the models. The second model which used the variables for literacy and formal employment, showed much better performance. In fact, the performance to predict change in asset ownership, bathroom facilities, and condition of household, is quite respectable in comparison to a baseline for majority prediction, and further points towards the consistency being followed in social development and economic growth models in the country.

Variable	Baseline Model		Model 1		Model 2	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
Asset Ownership	0.53	0.35	0.7	0.7	0.83	0.82
Bathroom Facility	0.72	0.42	0.73	0.77	0.74	0.78
Fuel for Cooking	0.9	0.41	0.69	0.56	0.72	0.62
Condition of Household	0.72	0.42	0.64	0.62	0.8	0.76

Table 12: Accuracy and F1-scores for Change prediction

References

[Chawla, Bowyer, Hall, and KegelmeyerChawla et al.2002] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research* 16 (2002), 321–357.