Towards Building a District Development Model for India Using Census Data

A thesis submitted in fulfillment of the requirements for

M.TECH MAJOR PROJECT
by

Dibyajyoti Goswami

Entry No. 2017MCS2873

Shyam Bihari Tripathi

Entry No. 2017MCS2832

Under the guidance of

Dr. Aaditeshwar Seth

Department of Computer Science and Engineering,
Indian Institute of Technology Delhi.
July 2019.

Certificate

This is to certify that the thesis titled **Towards Building a District Development Model for India Using Census Data** being submitted by **Dibyajyoti Goswami** and **Shyam Bihari Tripathi** for the award of **Master of Technology** in **Computer Science & Engineering** is a record of a bonafide work carried out by them under my guidance and supervision at the **Department of Computer Science & Engineering**. The work presented in this thesis has not been submitted elsewhere either in part or full, for the award of any other degree or diploma.

Dr. AADITESHWAR SETH
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi

Abstract

Models for socio-economic development are useful for planners to build appropriate policies. Such models are ideally constructed based on empirical data, and we take up the problem of working towards a district development model for India by using two waves of census data. As per 2011, India had almost six hundred districts, diverse in terms of their social and economic development. This presents a unique natural experiment to understand how social and economic factors interplay with one another. In this work, we present some interesting observations we are able to make from the analysis of census data from the years 2001 and 2011, and also raise some questions calling for the need for additional ethnographic and other surveys to be able to understand the underlying mechanisms that would have led to the observed patterns.

This would present a novel way for policy makers to make policy decisions on governance and allocation of resources paving the way to a better district development model, a framework to make informed decisions and ultimately help the national agenda of development.

Acknowledgments

This project has been of immense learning value as it introduced us to the world of Data Science and Computational Social Science. We now as students better appreciate how science can be used for the benefit of the mankind. Foremost, we would like to express our sincere gratitude to our supervisor *Dr Aaditeshwar Seth* for his valuable advice and support throughout the project. His ideas and constant guidance has shaped this study into the present form. It has been a learning and motivating experience working under him. It would be impossible to produce this work without his enthusiasm, and unfathomable knowledge. We thank him for giving us the freedom to explore multiple facets of the problem.

We are also grateful to Shivam Pathak and Sansiddh Jain who have been collaborating and contributing to various ideas, implementations and experiments. In addition, we would like to thank IIT Delhi for providing excellent facilities to support the project.

We would also like to thank our families and friends for their love and unflinching support.

Dibyajyoti Goswami
Shyam Bihari Tripathi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Objective

The intent has been to complete the following tasks:

- To **download data from web** to create a comprehensive table of variables based on **Census 2001 & Census 2011** for analysis.
- To **discretize the continuous variables** by using **clustering algorithms** with an aim to keep the labels consistent over the two Census and to enable a simple comparison of the socio-economic development states over the years.
- To study the **pattern of change** between **Census 2001 & Census 2011** and finally use probability analysis to **formulate hypothesis** based on observed patterns and test them using statistical tests.
- To use **Association Rule Mining and Bayesian Network** to study the prediction capability of census labels.
- To study **inequality** among districts in terms of socio-economic development states of villages for each variable.
- To seek answers to questions like -
  - *Does economic development leads to social development or vice versa?*
  - *Does government favour one approach over the other?*
  - *Can relationships be formed using data collected by government organisations viz. Census?*
  - *What is the prediction capability of these data?*

We feel that such a platform will help reveal richer insights to inform the debate between economic and social development, and eventually contribute towards building development models that have explicit values embedded within them. We feel that unless these values are not explicitly debated, mere statistics alone will get interpreted in different ways to suit different approaches.

1.2 Motivation & Related Work

Development in a country is measured as a sum total of economic and social development. Policy makers in India have traditionally relied on planned resource allocation towards economic and social development goals [9]. Over the years many researchers have tried to answer the question, 'Whether economic development leads to social development?' and vice versa. Our work is an empirical effort that can aid in national planning activities by drawing out observations on how different types of districts in India have evolved so far, what gaps remain to be addressed, patterns that warrant further investigation, and outcomes likely to arise if the same model of development is followed. Researchers have used a similar empirical approach relying on the use of satellite data to observe growth in inequality across different regions [14]. Extensive empirical work has also been conducted using five yearly sample surveys in India to understand patterns of change in female employment [26] and labour force participation [29]. Census data has been used to study the impact of district level administrative boundaries on access to household amenities [3]. A spatial concentration in economic activities is also evident using data from the Annual Survey of Industries [17, 19].

At the country level, models for emerging economies to transition from an agricultural economy to a non-agricultural one (manufacturing or services), has seen considerable discussion including an IMF working paper on the model followed by India [16]. However, models have not been proposed at the district level, especially in the context of India. Our work combined with such similar work can help further analyze the underlying dynamics behind these observations, and lead to value-based criteria in building policies. Our eventual goal is to build such a district development model.

1.3 Thesis Overview

1. **Data Set** - We have explained the pre-processing of census data to extract required information and finally create a comprehensive table of variables. The challenges and methods adopted while collecting and integrating data have been explained in detail.

2. **Methodology** - A complete overview of the system to discretize the continuous variables by using clustering algorithms with an aim to keep the labels consistent over the two Census has been provided. The motivation to do this and the justifications for the choice of no of clusters is given here.

3. **Analysis of Change**- Here we have described how change has happpened between 2001 and 2011 and the various patterns of development.

4. **Hypothesis Testing**- We describe here the formulation and testing of seven hypotheses with certain observations on why the hypothesis statements are actually true.

5. **Change Prediction**- Here we have described how we model classifiers to predict change through diffrent approaches.

6. **Study of Inequality**- Here we have described how we have made a modification of the ginin coefficient to calculate inequality in development states of villages in a district.

7. **Conclusion and Future Work** - Various ways by which the present work can be expanded are listed here along with the conclusion of the work in hand.

Chapter 2

Data Set

2.1 Introduction to Census India

The Government of India conducts a population census every ten years.The Census of India has been conducted total of 15 times, as of 2011. It was started in 1872 and since then conducted every ten years, the first complete census was conducted in 1881. The last census was conducted in 2011.The census is conducted broadly under two categories, House listing and People enumeration.

- **House Listing census** - This census collects information about the all buildings. It consists of 35 questions which talk about the house holds like condition of household, material of wall, material of roof , fuel for cooking, availability of bathroom facility, main source of light ,main source of water and asset ownership etc. This data is available under houselisting housing data.
- **People Enumeration** - This census collects information about the population ,the data is available under primary abstract table. It consists of 30 questions like literacy, type of employment, category of economic activity etc. This census gives the exact figure of population for each administrated unit.

We used data from the 2001 and 2011 censuses, available from the official website [5]. The Houselisting Housing data and Primary Census Abstract data table for both the years has been used in our study. The Primary census abstract table contain 90 categories about population enumeration while Houselisting Housing table contains data about 140 amenities and assets in the household. The number of households in each spatial unit (village, district, state), belonging to these different categories were used to create our own variables for the study. For Census 2011, data was collected for every state which had values for each village which were merged per district to get a single file for India. The data types for each column was made consistent. The 2001 data was available only at the district level, hence we did our analysis at the district level. Between 2001 and 2011, 47 districts split into smaller districts due to administrative changes, and for our analysis we grouped them back into the original 593 districts that existed as of 2001.

2.1.1 Census 2011

The 2011 Census was conducted April 2010 onwards. Spread across 29 states and 7 union territories, the census covered 640 districts, 5,924 sub-districts, 7,935 towns and more than 600,000 villages. A total of 2.7 million officials visited households in 7,935 towns and 600,000 villages, classifying the population according to gender, religion, education and occupation. The census collected information on various aspects. For our analysis we selected 6 socio economic indicators. The subcategories on which the data was collected for each indicator is shown in Table 2.1.

Indicator	**Parameters**
Asset ownership	Radio/Transistor/Television/ computer/laptop/Telephone /Mobile/Bicycle/Two wheeer/Four wheeler/Nonw
Bathroom facility	Piped sewer sytem/Septic tank/Pit latrine/Disposed into open drain/ Removed by human/Serviced by animal/No latrine
Fuel for cooking	Fire-wood/Crop residue/Cow dung/Coal charcoal/ kerosene/LPG/PNG/Electricity/Biogas/other /No cooking
Condition of House hold	Good/Livable/Dilapidated
Main source of light	Electricity/Kerosene/Solar energy/Other oil /Anyother/No lighting
Main source of drinking water	Tapwater from treated source/Untreated source/Covered Well /Uncovered well/Handpump/Tubewell/Borehole/Spring/ River/Tank/Pond/Lake/Other sources

Table 2.1: Parameters on which information was collected for each indicator

2.1.2 Census 2001

2001 Census was also conducted broadly under two categories house listing and people enumeration. However there are some differences in Census 2001 and 2011 in terms of parameters on which data was collected. For example, in 2001, mobile phones and internet had not proliferated much, therefore this detail was not collected in 2001 census. Other than asset ownership, other indicators on which our analysis is based did not have mismatch in the parameters. One of the major differences between data available for 2001 Census and 2011 Census is that 2011 census data is available at village level for all indicators but 2001 census data is available at village level for people enumeration only, the data for house listing is available for sub district level only. The six socio-economic indicators used for analysis are from house listing census table, and therefore we did our major analysis at district level.

2.2 Mapping of districts between 2001 and 2011 Census

In 2001 census the total number of districts in India were 593 and in 2011 it increased to 640. 47 new districts were created from the existing districts. In order to make the analysis uniform we merged the newly created districts with their parents , so the geographical area under consideration is same. The census districts code were also different in 2001 and 2011. In 2011 uniform coding was used , which means the district code in 2011 census was from 1 to 640 but in 2001 census the districts code were assigned as per the state, so every state has district code starting from 1. In our analysis we have used the 2011 census district code. We reassigned the district code in 2001 census as per 2011 census to make them uniform. The list of newly created 47 districts along with their parent districts is given in Table 2.2.

New District Name (2011)	Census Code	Separated From (2001)	Census Code
Bandipore	9	Baramulla	8
ganderbal	11	Srinagar	10
Shopian	13	Anantnag	14
Kulgam	15	Anantnag	14
Ramban	17	Doda	16
Kishtwar	18	Doda	16
Reasi	20	Udhampur	19
Samba	22	jammu	21
Taran taran	50	Amritsar	49
SAS nagar	52	Ropad , Patiala	51
Barnala	54	Sangrur	53
Mewat	87	Gurgaon	86
Palwal	89	Faridabad	88
Pratapgarh	131	Chittorgarh, Udaipur ,banswara	130
Kashi ram nagar/Kasganj	202	Etah	201
Arwal	240	Jehanabad	239
Kurung Kumey	256	Lower Subansiri	255
Lower dibang valley	258	Dibang Vlley	257
Anjaw	260	lohit	259
Longleng	268	Tuensang	267
Kiphire	269	Tuensang	267
peren	271	Kohima	270
Chirang	320	Bongaigaon	319

kamrup Metripoliton	322	Kamrup	321
baksa	324	Barpeta, Nalbari,kamrup	323
Udalgiri	326	Darrang	325
Purba Medinipur	345	medinipur	344
latehar	359	palamu	358
Ramgarh	361	Hazaribag	360
jamatra	363	Dumka	362
khunti	365	Ranchi	364
Simdega	367	Gumla	366
saraikela kharsawan	369	West singhbhoom	368
narayanpur	415	bastar	414
Bijapur	417	Dantewada	416
Ashok nagar	459	Guna	458
Anuppur	461	Shahdol	460
Singrauli	463	Sidhi	462
Alirajpur	465	Jhabua	464
Burhanpur	467	East Nimar	466
Tapi	493	Surat	492
yadgir	580	Gulbarga	579
Chikkaballapura	582	Kolar	581
Ramnagra	584	Bengaluru Rural	583
Krishnagiri	631	Dharmapuri	630
Tiruppur	633	Coimbatore ,Erode	632
South Andman	640	Andman	639

Table 2.2: List of newly created districts between 2001 and 2011

2.3 Pre-processing of Data

We first selected parameters for our study primarily focusing on the impact they have on the social and economic development. Census Data for the year 2001 and 2011 have been structured in a different way. Though most of the parameters match, we first made them consistent by categorizing them into three broad classes. For Employment we made three variables viz. % of population employed in Agricultural activities, % of population employed in Non-Agricultural activities and % of population Unemployed. For all other variables we divided them into variables Rudimentary (using primitive/old methods), Intermediate (better methods than rudimentary) and Advanced (developed methods).

Variable	**Broad Class**	**Using/Access to**
Bathroom Facility	Rudimentary	No Latrine facility/Disposed in open
	Intermediate	Pit Latrine
	Advanced	Piped Sewer/Septic Tank
Fuel for cooking	Rudimentary	Firewood/Crop residue
	Intermediate	Cow Dung/Kersosene/Coal/Charcoal
	Advanced	LPG/PNG/Bio gas/Electricity
Condition on Household	Rudimentary	Dilapidated House
	Intermediate	Livable House
	Advanced	Good House
Main Source of Light	Rudimentary	No source of light
	Intermediate	Kerosene oil/Other oil/Other Sources
	Advanced	Electricity/Solar Light
Main Source of Water	Rudimentary	Well/Spring/River/Pond/Lake
	Intermediate	Hand Pump/Tube Well/Borehole
	Advanced	Tap Water/Treated water

Table 2.3: List of variables aggregated as Advanced, Intermediate and Rudimentary for each Indicator

These have been described in a Table 2.3. We did not aggregate the parameters of the asset ownership because each parameter can be bought at different price thus indicating the prosperity , for example the asset mobile phone can be bought as cheap as ₹ 1000 or as expensive as ₹ 70,000, so possession of mobile phone does not indicate whether it is rudimentary or advanced. We did the clustering of the asset ownership directly on the raw data of four type of asset radio/transistor, television, two wheeler and four wheeler. In 2001, the availability of computer and internet was not much so the census did not collect the data for computer and internet in 2001 census. Accordingly we did not consider the possession of computer and internet in our asset ownership clustering to make the data consistent.

2.4 Correlation between input Variables

In the previous section we have defined the input parameters (Advanced, Intermediate and Rudimentary) for each indicator. Before moving to analysis of change happened between 2001 and 2011 the study of relation between input parameters is very important. The relation between input parameters will tell whether development is happening across all indicators or it is restricted to selected indicators. It will also give answers to questions like if the development in one indicators is positively correlated with development in other indicators. To find answer to such questions we calculated the correlation between input variables which is shown in Figure **??**.

By observing Figure **??** we can see that advanced features are positively correlated with each other and advanced features are negatively correlated with rudimentary features. This correlation diagram indicates that development in any one indicator is not happening in isolation, generally the districts are developing in all the indicators, although the magnitude of development will differ from indicator to indicator. This analysis seems to follow the general convention that development happens across all the indicators. For example if a district has good water distributed system where most of the households are getting treated tap water and a electricity connection in majority of households then it is very likely that they will have good bathroom facility also.

In indicators for living conditions like bathroom facility, fuel for cooking, main source of lighting and main source of water, advanced indicators have weak and negative correlation to intermediate ones which in turn have weak and negative correlation to rudimentary features. Thus, development in any particular parameter are present in districts as a combination of advanced, intermediate and rudimentary and cannot be studied using any single indicator. There exist states of development in each indicator individually and the combination of these states of development across different parameters results in the classes of regions with different levels of socioeconomic development.We therefore used unsupervised methods to label the districts.

Chapter 3

Methodology

The diversity of India presents a natural experimental setup to study how different districts have evolved over time. We use census data for two years separated by a decade, and select a mix of six social development and economic growth indicators, to examine the patterns of change over this decade. Around these six indicators, we construct seven hypotheses and test these hypotheses. All these hypotheses are largely about comparing the pace of change of an indicator with respect to other intervening variables, to get a sense of which indicators move faster under which conditions. These are precisely the kind of empirical insights that can inform the Bhagwati-Sen debate about when social development happens viz-a-viz economic growth, and vice-versa. Our approach at a high-level is similar to association rule mining [1], but we choose to use a simpler method of comparing the mean rate of change in different indicators (at a statistically significant level) to come up with easily interpretable patterns. We develop a unique discretization method for real-valued data across many categories, to make the hypotheses more interpretable. Subsequently we use association rule mining to determine the relationship between the various socio-economic indicators and construct a Bayesian Network to predict change.

3.1 Discretization of Variables

Each category variable provided by the census may have multiple parameters, and reports the number of households in a district for each parameter. For example, for the variable regarding the primary type of fuel used for cooking by households, the census reports separately the number of households in a district using firewood, kerosene, LPG (Liquid Petroleum Gas), PNG (Piped Natural Gas), biogas, etc. For our analysis, we wanted to reduce these to a single value for each variable and we developed the following procedure. We first group the mutually exclusive parameters within a variable into three broad parameter types of rudimentary (Rud), intermediate (Int), and advanced (Adv). For example, firewood is considered as a rudimentary type of fuel for cooking, kerosene and cow dung are grouped together as an intermediate type, and PNG, LPG and bio gas are grouped together as advanced types of fuel for cooking. This grouping is showing in Table 3.1.

Variable	**Using/Access to**	**Level - 1**	**Level - 2**	**Level - 3**
Bathroom Facility	Rud: No Latrine facility	65-82	20-40	18-40
	Int:Pit Latrine	0-5	30-45	0-10
	Adv :Piped Sewer/Septic Tank	15-28	25-40	50-70
Fuel for Cooking	Rud:Firewood	60-80	30-50	20-40
	Int:Cow Dung/Kersosene	5-15	40-60	5-20
	Adv:LPG/PNG/Bio gas	15-35	5-20	45-65
Condition of Household	Rud:Dilapidated House	5-10	0-5	0-5
	Int:Livable House	55-65	40-50	25-35
	Adv:Good House	30-40	45-55	65-75
Main Source of Light	Rud:No source of light	0-5	0-5	0-5
	Int:Kerosene oil/Other oils	70-80	30-50	5-15
	Adv:Electricity/Solar Light	20-30	50-70	85-95
Main Source of Water	Rud:Well/Spring/River	40-70	2-20	5-15
	Int:Hand Pump/Tube Well	2-25	55-80	10-28
	Adv:Tap Water/Treated water	20-40	10-28	60-85
Asset Ownership	TV	15-30	30-50	60-85
	Telephone	35-55	40-60	50-60
	2-Wheeler	5-12	5-18	20-40
	4-Wheeler	0-2	0-5	2-12

Table 3.1: Census variables to classify districts in terms of levels for different indicators: Shown is the % of households within various indicators

We then do a k-means clustering on the districts based on the percentage of households in each district that use different types of fuel: rudimentary, intermediate, and advanced. Figure 3.1 shows a box-plot for the distribution of districts across three levels (k = 3) in terms of their use of different types of fuel for cooking. The other box-plots are shown in subsequent figures.

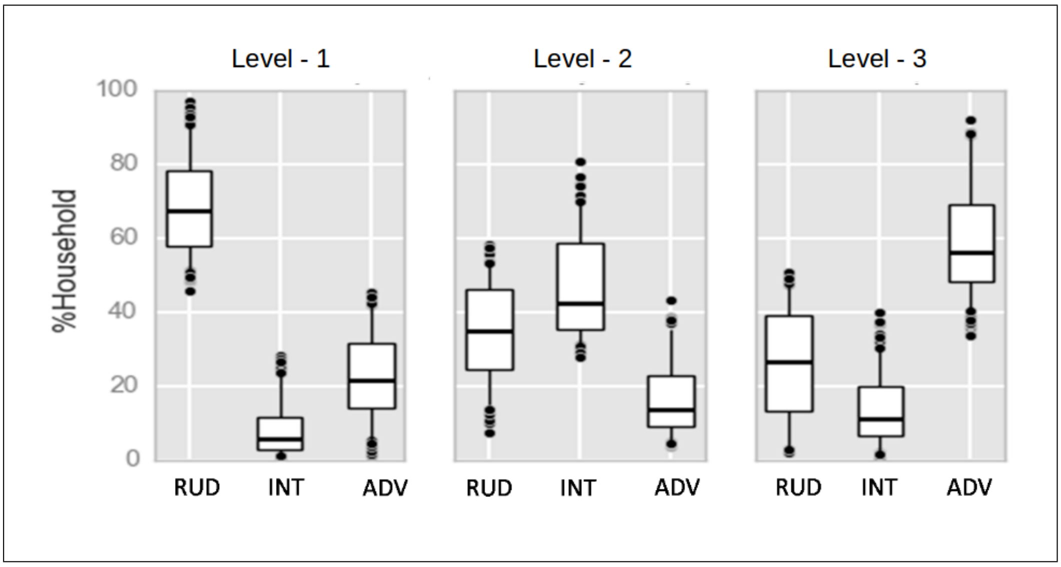

Figure 3.1: Discretizing the level of development of districts in terms of the fuel used for cooking in the district households. Shown is the distribution of the % of households using different types of fuel (rudimentary, intermediate, or advanced) across three district clusters. These clusters are used to label the level of a district for the fuel for cooking indicator

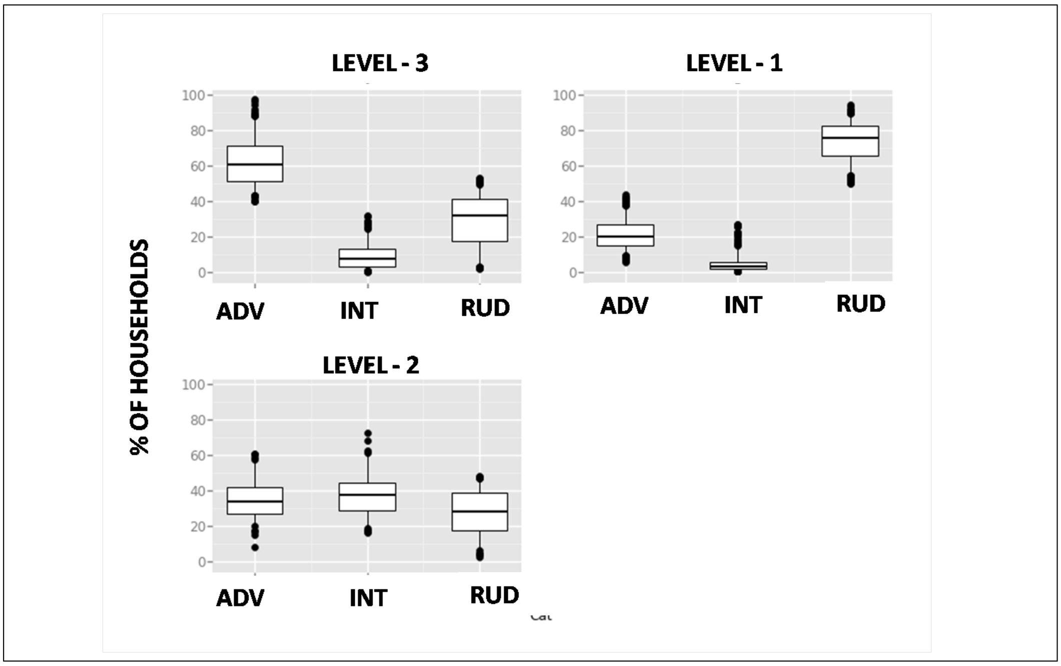

Figure 3.2: Discretizing the level of development of districts in terms of bathroom facility in the district households. These clusters are used to label the level of a district for bathroom facility

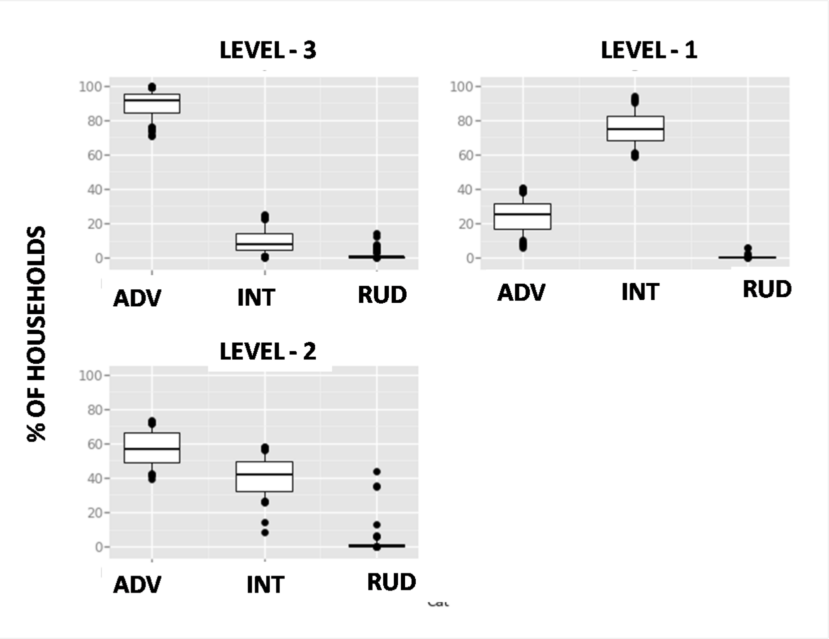

Figure 3.3: Discretizing the level of development of districts in terms of the main source of light. These clusters are used to label the level of a district for the main source of light

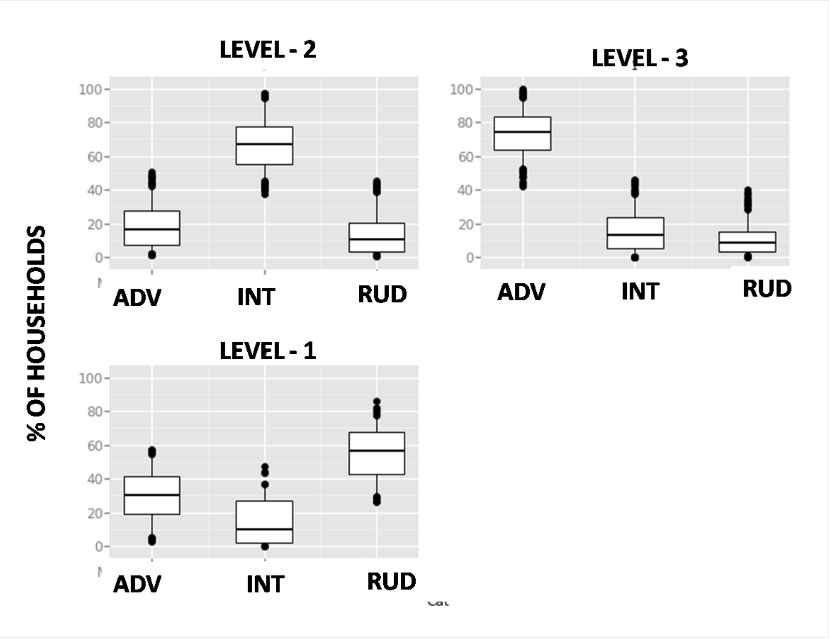

Figure 3.4: Discretizing the level of development of districts in terms of main source of water in the district households. These clusters are used to label the level of a district for main source of water

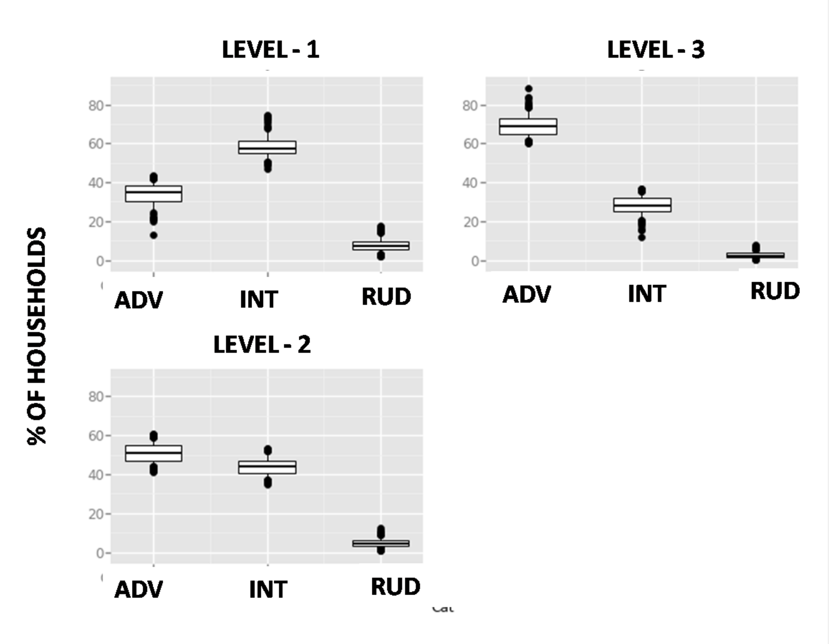

Figure 3.5: Discretizing the level of development of districts in terms of the condition of house hold in the district households. These clusters are used to label the level of a district for the condition of house hold

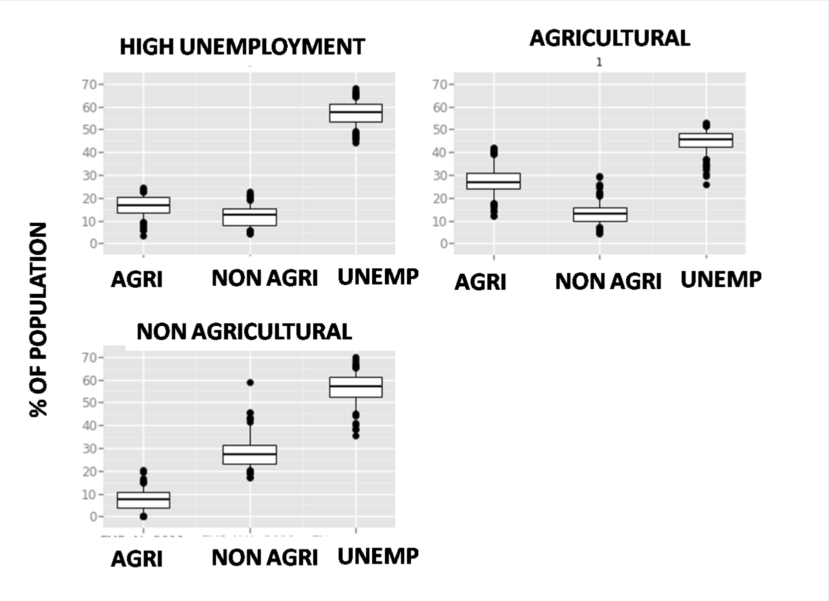

Figure 3.6: Discretizing the level of development of districts in terms of type of employment. Shown is the distribution of the % of population employed in non agricultural , agricultural and unemployment activities. These clusters are used to label the level of a district for type of employment

This allows us to label each district as a level-1/2/3 district: Level-1 districts predominantly use rudimentary types of fuel for cooking, level-2 districts primarily use intermediate types of fuel, and level-3 districts predominantly use advanced types of fuel for cooking. We follow the same method for other indicators also. Table 3.1 indicates the percentage of households having access to rudimentary, intermediate, and advanced measures of the indicators for each level. This method therefore allows us to map each district to a single coarse value for each variable.

We experimented with different values of k for different variables in terms of the quality of clusters obtained, and eventually settled on k = 3 as a reasonable and uniform mapping of districts for all the variables. We justify this choice of k in a subsequent subsection.

We apply the same method to also classify districts in terms of the dominant type of employment: agricultural, non-agricultural, or high unemployment as given in Table 3.2. Continuous variables such as female employment and literacy are retained as such since they did not have multiple internal parameters.

Variable	**% of population employed in**	**District Type**		
		Non Agricultural (in %)	**Agricultural (in %)**	**High Unemployment (in %)**
Employment	Unemployed	50-60	40-50	55-65
	Agricultural labour	5-10	25-35	15-20
	Non Agricultural work	22-35	10-15	8-15

Table 3.2: Census variable to classify districts in terms of type of employment: Shown is the % of population in different types of employment

3.2 Motivation for Discretization

Its always difficult to interpret the overall change in socio-economic conditions specially when each socio-economic parameter has multiple variables representing it. Here we present a novel idea of discretization which is useful for several reasons. First, as shown in the book *Factfulness* by Hans Rosling

[25], who used a similar 4-level mapping for different stages of development of countries and regions, such a coarse mapping is easy for people to interpret and to easily compare different districts with one another. Second, it reduces the variables to a single quantity without assigning arbitrary weights to club together multiple parameters for each variable. Third, as we show next, it allows us to compare different variables with one another using simple probabilistic analysis to determine broad patterns, instead of more complex regression methods which may be hard to interpret and to determine significant relationships. This is the key method we use in this paper to study the relationships between different variables. Note that each district may be marked as belonging to a different level for different variables, for example, a district could be at level-1 in terms of its dominant use of fuel for cooking, but at level-3 in terms of asset ownership, and so on. Seeing how districts move from one level to the other across different variables, allows us to determine broad patterns about which variables tend to move first before others, and any inter-dependencies that might exist between the variables.

3.3 Justification for choice of k

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.

3.3.1 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.

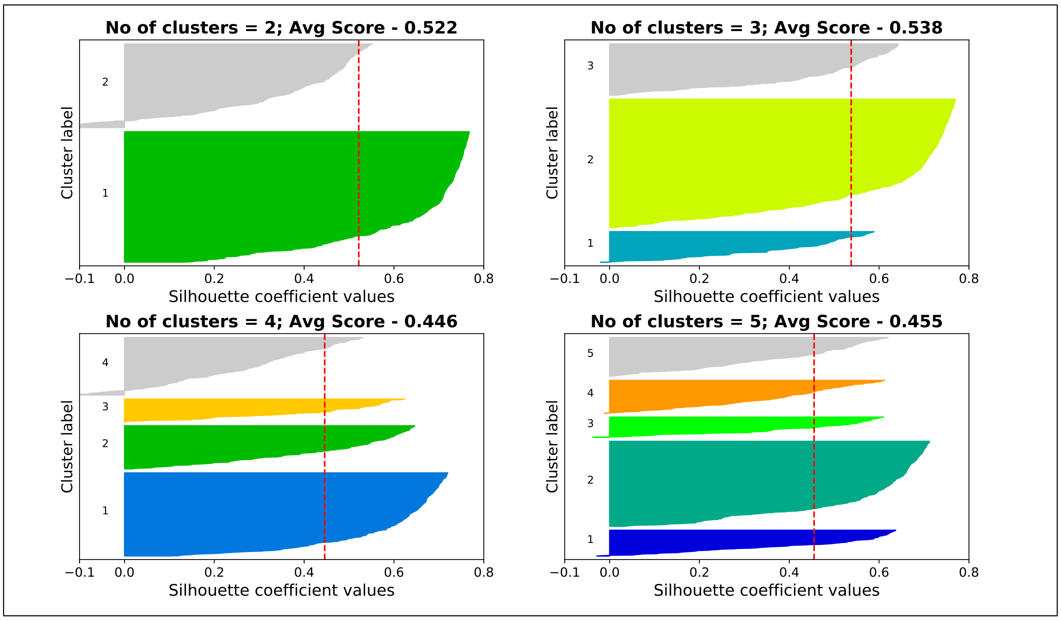

Figure 3.7: Silhouette plots for clustering with 2-5 clusters : Fuel for Cooking

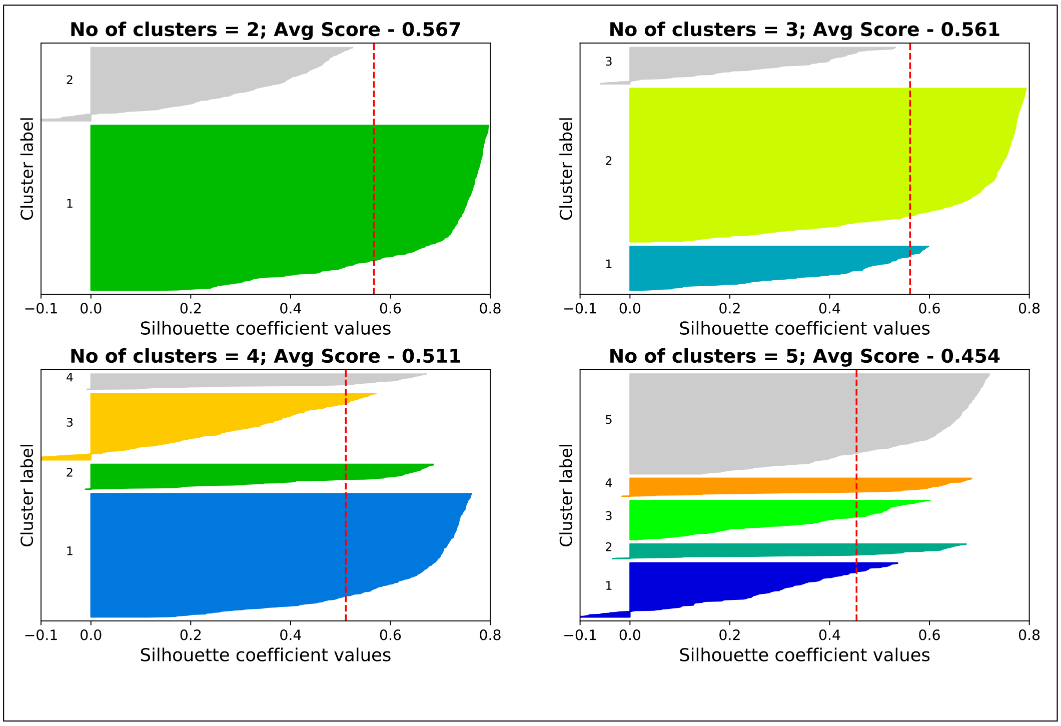

Figure 3.8: Silhouette plots for clustering with 2-5 clusters : Bathroom facility

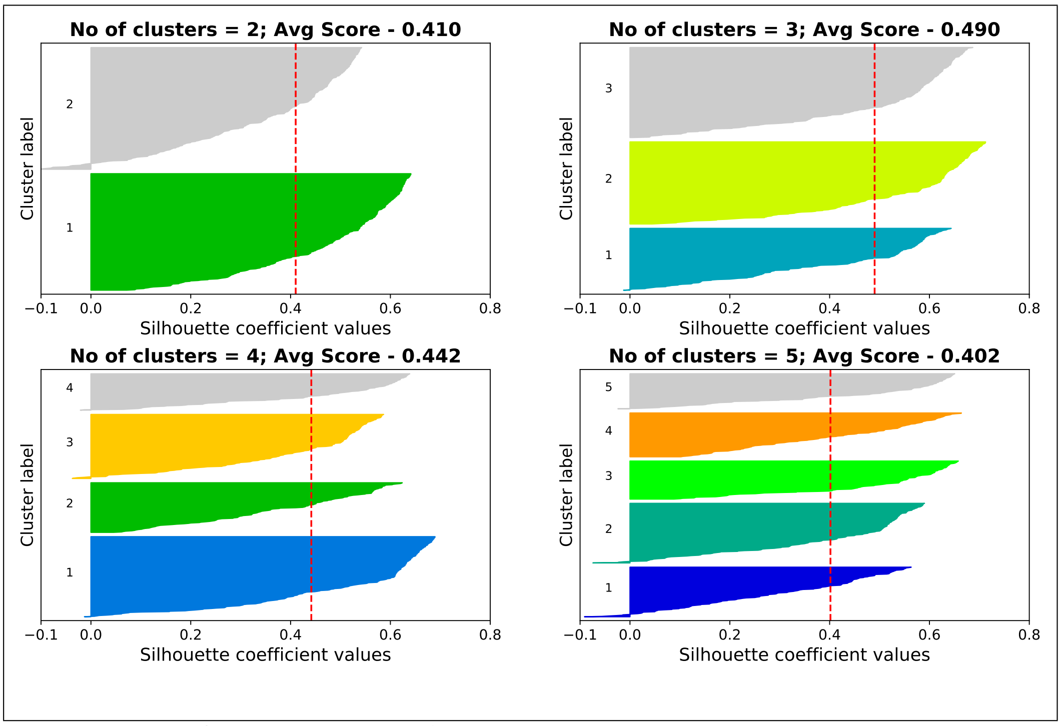

Figure 3.9: Silhouette plots for clustering with 2-5 clusters : Main source of water

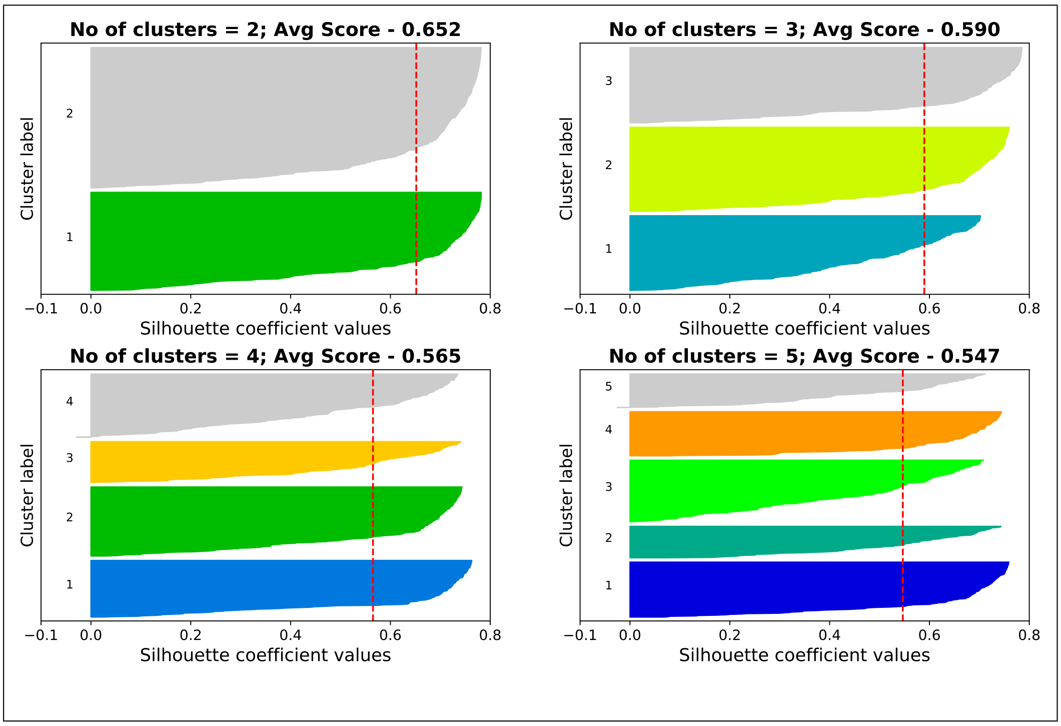

Figure 3.10: Silhouette plots for clustering with 2-5 clusters : Main source of light

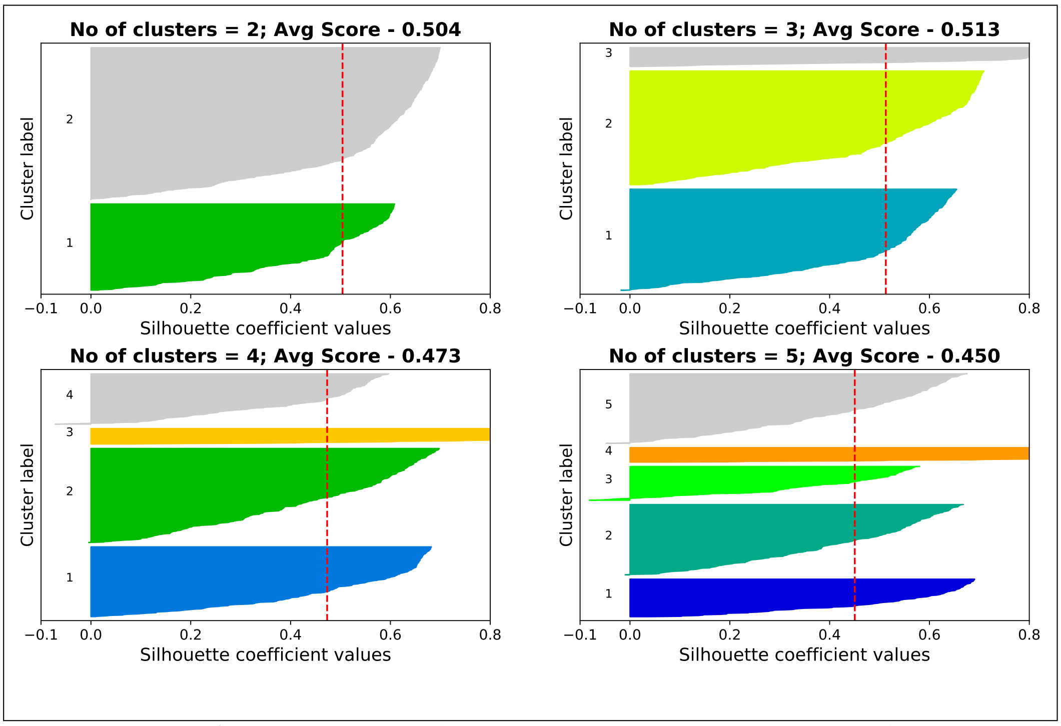

Figure 3.11: Silhouette plots for clustering with 2-5 clusters : Condition of Household

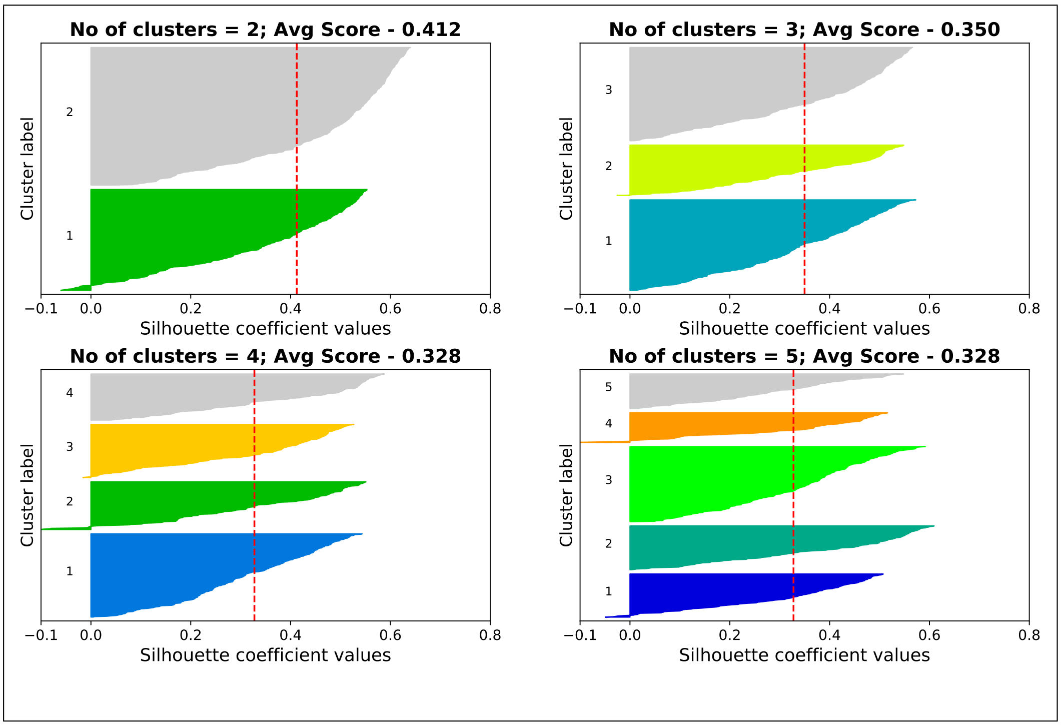

Figure 3.12: Silhouette plots for clustering with 2-5 clusters : Employment

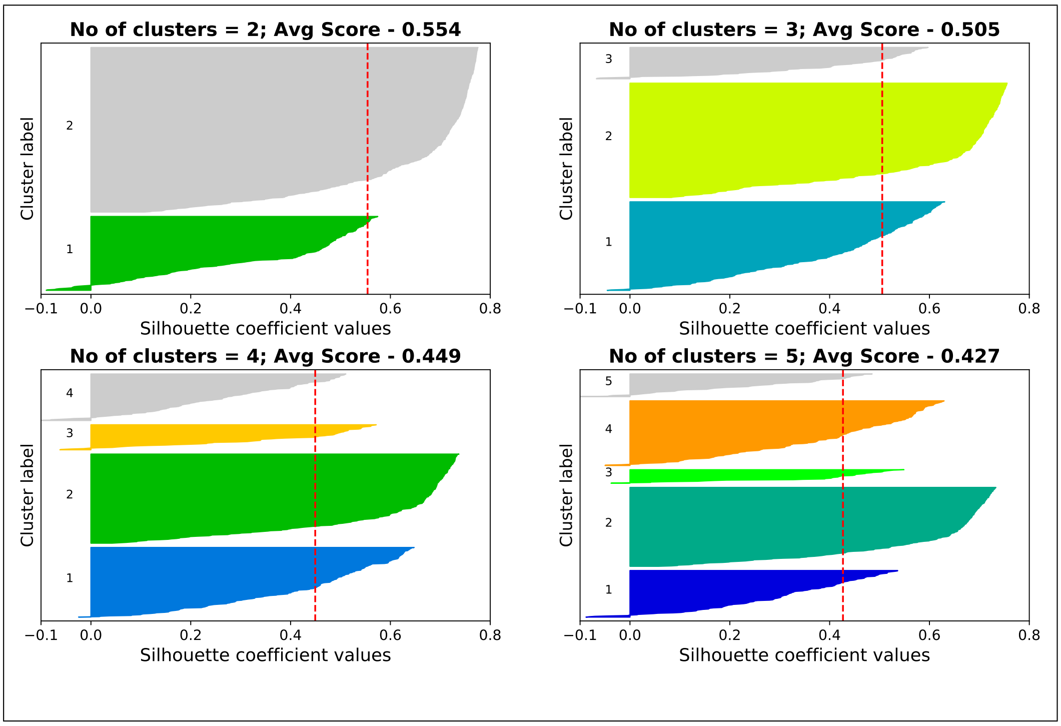

Figure 3.13: Silhouette plots for clustering with 2-5 clusters : Asset ownership

3.3.2 Elbow plots

The 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.

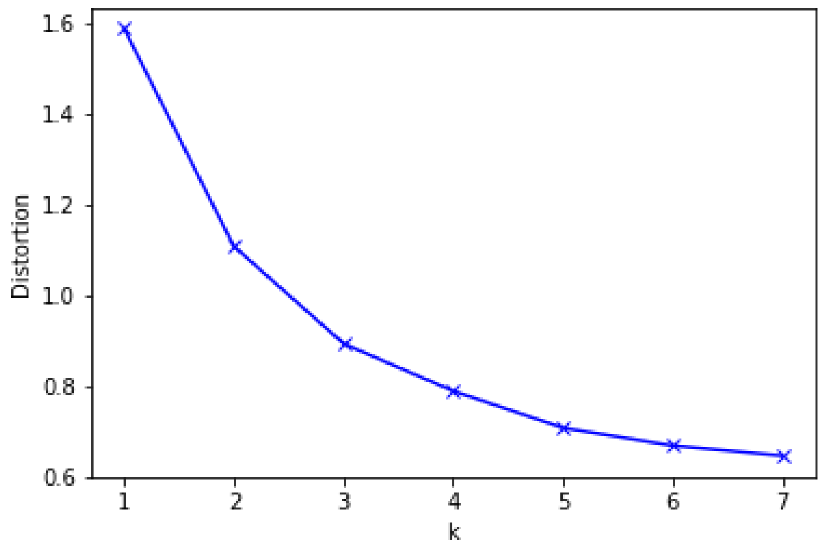

Figure 3.14: Elbow plot showing optimal k : Asset Ownership

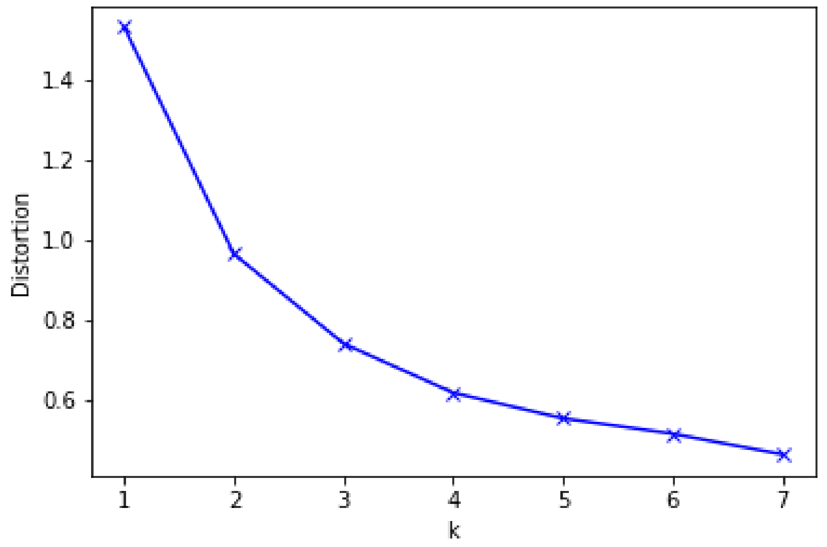

Figure 3.15: Elbow plot showing optimal k : Bathroom facility

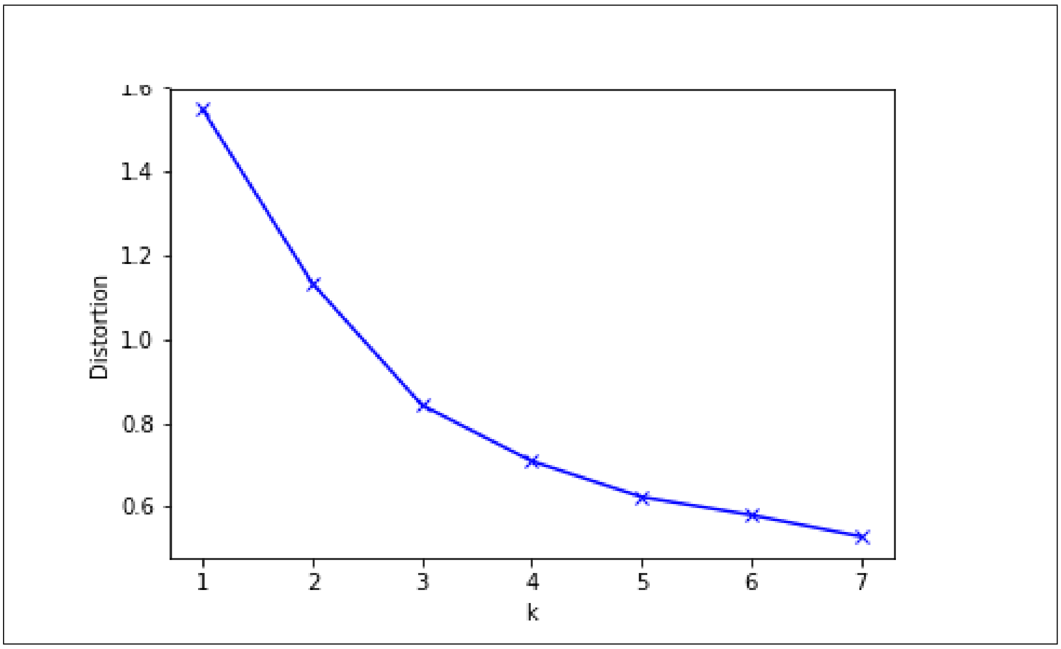

Figure 3.16: Elbow plot showing optimal k : Fuel for cooking

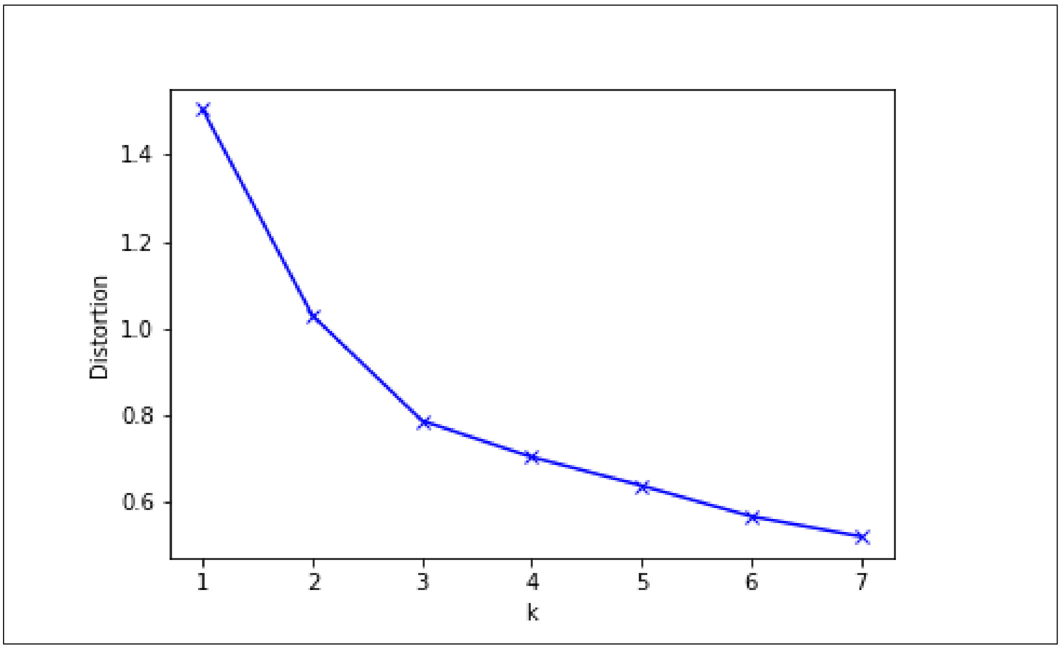

Figure 3.17: Elbow plot showing optimal k : Condition of household

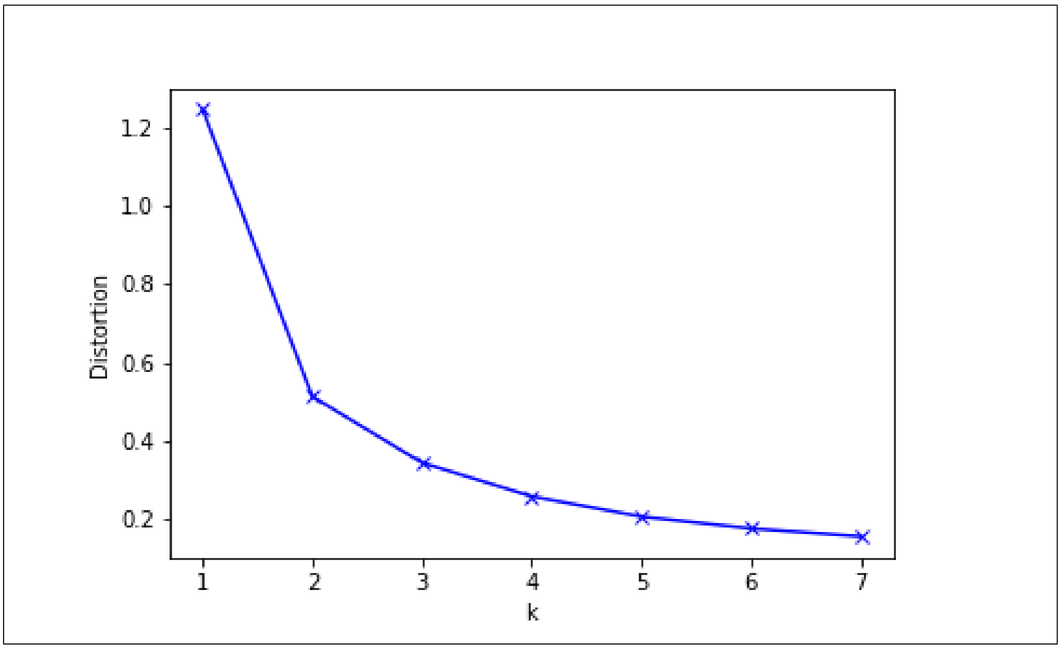

Figure 3.18: Elbow plot showing optimal k : Main source of light

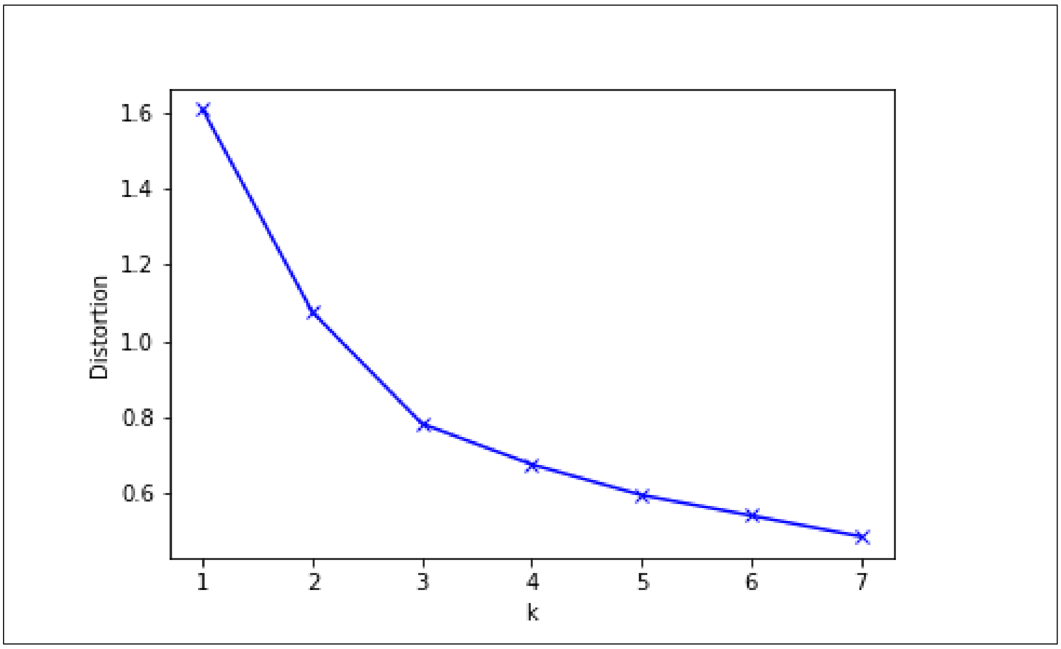

Figure 3.19: Elbow plot showing optimal k : Main source of water

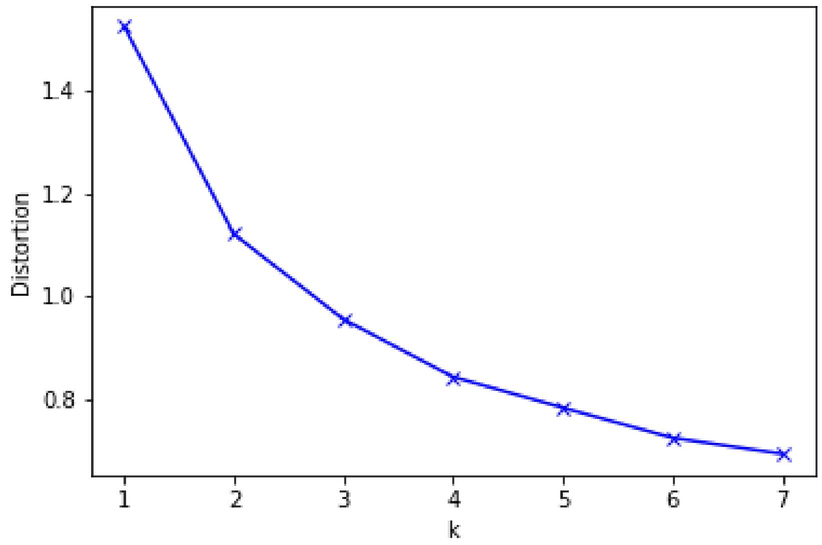

Figure 3.20: Elbow plot showing optimal k : Type of Employment

3.3.3 Sensitivity analysis using k = 4

We test 6 hypotheses based on our observations from the change analysis. An analysis of the relevant hypotheses was done by using k = 4 as well. The results are consistent with what we subsequently show in Chapter 5 for k = 3 as well. It shows that the findings are not sensitive to the choice of k. Table 3.3 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 3.3: Change in indicators based on the type of employment for k = 4

Chapter 4

Analysis of Change

We need to allow for comparison of a district over the two census years, to determine whether the district moved from one level to another, for different variables. To do this, we first obtain the clusters using 2011 data since it showed more diversity in all the variables. For each variable, we then calculate the centroid for every cluster and determine the level of a district in 2001 by seeing which centroid is the closest for the district. This allows us to obtain Table 4.1 which shows the percentage of districts that moved from a lower level to a higher level between 2001 to 2011, stayed the same, or even dropped from a higher level to a lower level. This change is shown separately for districts based on their level as of 2001.

Variable	**Level 1 (values in %)**			**Level 2 (values in %)**			**Level 3 (values in %)**		
	+ve Growth	**-ve Growth**	**No Change**	**+ve Growth**	**-ve Growth**	**No Change**	**+ve Growth**	**-ve Growth**	**No Change**
Asset Ownership	52.03	0.00	47.97	84.27	0.00	15.73	0.00	0.00	100.00
Bathroom Facility	19.85	0.00	80.15	46.26	6.80	46.94	0.00	10.53	89.47
Fuel for Cooking	10.42	0.00	89.58	12.88	14.11	73.01	0.00	0.00	100.00
Condition of Household	31.67	0.00	68.33	28.87	7.75	63.38	0.00	15.91	84.09
Main Source of Light	29.90	0.00	70.10	62.56	1.54	35.90	0.00	2.94	97.06
Main Source of Water	48.75	0.00	51.25	9.61	0.00	90.39	0.00	4.90	95.10
Female Emp (Main)	23.08	0.00	76.92	16.75	13.88	69.38	0.00	9.09	90.91
Female Emp (Marginal)	10.47	0.00	89.53	7.08	57.52	35.40	0.00	52.28	47.72

Table 4.1: Shown is the percentage of districts at each level, that have shown a positive movement to a higher level, or a negative movement, or no change from their current level. The table shows the percentage values for all the indicators

4.1 Change between 2001 and 2011

There are several interesting patterns to notice from Table 4.1. Most level-3 districts for any variable show no-change (last column in Table 4.1), which is understandable because these districts were already at the highest level of development for that variable. There is a negative growth in some districts though, which we believe is due to significant inbound migration into these districts that caused an increase in the population living in slum areas and suburbs due to people coming in search of better employment opportunities to districts placed at higher levels of development. We also see that there has been an overall drop in female employment. Interestingly, there has been a significant drop for females in marginal employment (less than six months of work in a year), indicating either an increase in formalization in female employment or a decrease in overall female participation in the workforce. We will get back to this topic later in the paper.

Getting back to Table 4.1, we see that many districts have improved for some variables at level-1 and level-2, but fewer districts have improved for some other variables. Variables that require government investment in infrastructure, such as electrification to alter the main source of light used by households, or water pipelines and handpumps to alter the main source of water, show different degrees of change. This possibly indicates that different priorities may have been placed by the government on these two variables. In the same way, discretionary variables determined more by household decision making, also show different degrees of change. Districts at level-2 in asset ownership have improved significantly, but not districts at level-2 on the condition of household, for instance. In the following sections, we examine such patterns in more detail.

4.2 Change in Employment Profile

Figure 4.1 highlights the changes in the dominant type of employment in districts between 2001 and 2011. Only a few districts have actually changed over the decade on this variable.

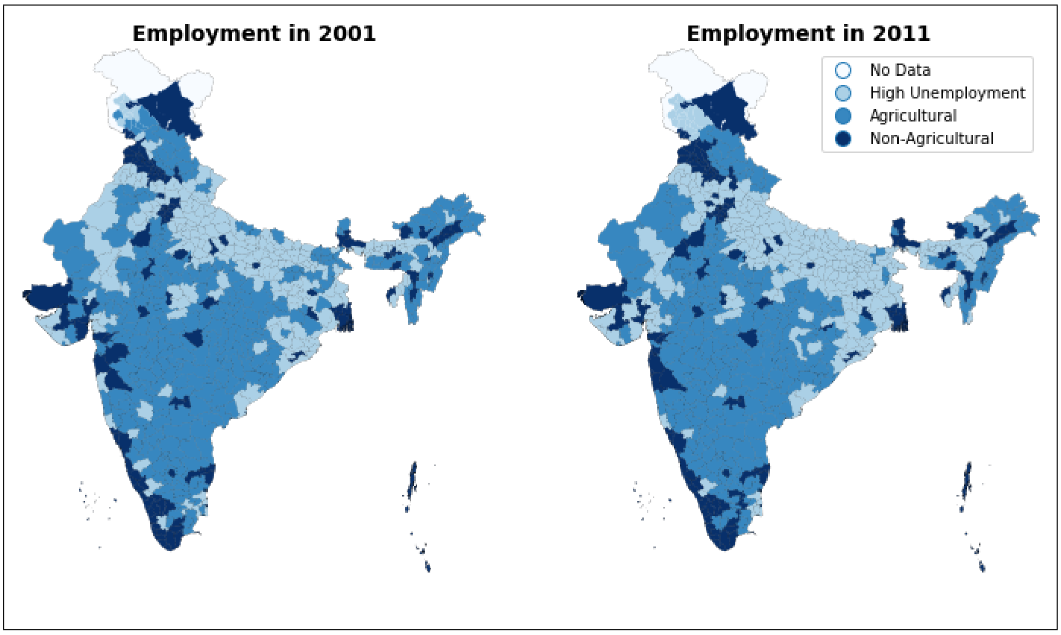

Figure 4.1: Districts are colour-coded on their dominant type of employment in 2001 and 2011

Only 23 districts which had a high degree of unemployment in 2001 were able to move to a predominantly agricultural or non-agricultural type of employment. Similarly, only 5 districts changed from a predominantly agricultural employment profile to a non-agricultural employment profile. A few districts even moved from a non-agricultural to agricultural profile, possibly indicating that there was little growth in non-agricultural employment in these districts causing the growing number of households there to fall back on their agricultural inheritance instead of being able to find employment in other sectors or even migrate outside the district. Most districts (87%) stayed the same in terms of their employment profile.

4.3 Patterns of Development

Variable	Existing Status	Non Agricultural	Agricultural	High Unemployment	Total
Asset Ownership	Level-1	0.895	0.498	0.472	0.571
	Level-2	0.816	0.714	0.962	
	Total	0.851	0.509	0.534	
Bathroom Facility	Level-1	0.742	0.142	0.172	0.268
	Level-2	0.8	0.171	0.333	
	Total	0.779	0.146	0.213	
Fuel for Cooking	Level-1	0.317	0.076	0.078	0.111
	Level-2	0.688	0	0.093	
	Total	0.421	0.064	0.086	
Main Source of Light	Level-1	0.833	0.315	0.261	0.462
	Level-2	0.741	0.636	0.54	
	Total	0.758	0.513	0.345	
Condition of Household	Level-1	0.714	0.36	0.19	0.3
	Level-2	0.591	0.234	0.144	
	Total	0.621	0.289	0.168	
Main Source of Water	Level-1	0.212	0.588	0.511	0.257
	Level-2	0.333	0.072	0.066	
	Total	0.263	0.325	0.189	

Table 4.2: Shown is the probability of positive change in indicators for districts having different types of employment. This is dis-aggregated further into the probability of positive change based on the existing status of the districts. The last column shows the overall probability of positive change

Table 4.2 shows the probability of positive change in an indicator given the type of employment of a district, ie. P(positive change | type of employment). A movement from level-1 to level-2 or level-3, or a movement from level-2 to level-3, is considered as a positive change. All other changes are considered as non-positive changes. For example, considering the indicator for asset ownership, the last column shows P(positive change) for asset ownership = 0.571, calculated across all districts. P(positive change | type of employment) in asset ownership is shown in the last row as 0.851, 0.509, and 0.534, for the changes in non-agricultural, agricultural, and high unemployment districts respectively. A further dis-aggregation is shown for P(positive change | type of employment, current status) considering the current status of a district of being at level-1 or level-2 in asset ownership. The P(positive change) for all other variables was calculated in a similar manner. We analyze this table carefully to determine several broad patterns of development, as explained next. Note that level-3 districts are not shown in the table because they are already at the most developed level, and very few negative movements were recorded from this level.

Variable	**Non Agricultural**	**Agricultural**	**High Unemployment**
Asset Ownership	64.38	15.75	19.86
Bathroom Facility	64.07	17.37	18.56
Fuel for Cooking	72.66	11.72	15.63
Condition of Household	48.52	39.05	12.43
Main Source of Light	37.69	43.30	19.00
Main Source of Water	34.47	46.38	19.15

Table 4.3: Shown in the percentage of districts at level-3 (as of 2011) for the three types of districts

We also check whether non-agricultural districts are more likely to be at level-3 for various socio-economic indicators. Table 4.3 shows the % of districts at level-3 (as of 2011) for each variable. We see that for four out of six variables, non-agricultural districts have a higher share of level-3 districts.

In the next chapter we test seven hypothesis which we formulate through our observations in the patterns of development.

Chapter 5

Hypothesis Testing

5.1 Hypothesis 1

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

5.1.1 Probability Analysis of Hypothesis 1

Observing Table 4.2, we can see that out of the six indicators, non-agricultural districts see the highest probability for positive change in five of them. Further dis-aggregating based on the current status of the districts, shows that this hypothesis is true irrespective of the current status. Only for the main source of water, do we see that agricultural districts have a greater tendency for positive change, which is also predominantly due to level-1 agricultural districts having progressed rapidly. There were 53 such districts, many of them from central India, indicating that the hypothesis was invalidated in only a few districts which potentially saw special policy attention being given on them.

5.1.2 Statistical Test of Hypothesis 1

In order to statistically test the hypothesis, we prepared two groups for the test and then conducted T-tests and one way ANOVA to test the hypothesis. For preparing the groups to test the hypothesis we took samples (80 percent) for each type of district (e.g. non agricultural, agricultural and high unemployment) and calculated the percentage of districts which have seen positive growth (for each indicator separately). We did this sampling 20 times and then conducted the T-test and One Way ANOVA for each indicator. The null hypothesis is that there is no significant difference between the groups and alternate hypothesis is that there is significant difference between the groups. T-test can compare only two groups at a time so in order to test our hypothesis we did pair wise T-test for all the groups. In one way ANOVA test more than two groups can be compared so we conducted ANOVA test for all the groups in one go. The ANOVA test established the fact that there is significant difference between the groups but it does not tell where the difference lies. In order to test the relative performance of the groups, in order to find the group with the maximum mean we conducted Tukey's pair wise HSD test. The result of T-test between Non Agricultural and Agricultural districts is shown in Table 5.1.The result of T-test between Non Agricultural and High Unemployment districts is shown in Table 5.2. From both the tests it can be seen that extremely low p-value strongly invalidate the null hypothesis and support the alternate hypothesis that Non agricultural district see the greatest improvement in all indicators except main source of water.

Indicator	**t statistic**	**p value**
MSL_Change	17.67965855	8.21E-10
MSW_Change	-4.681807961	0.000473
FC_Change	34.25305616	6.58E-12
BF_Change	68.5802514	1.04E-14
CHH_Change	24.50319135	7.17E-11
Asset_Change	50.42330692	1.04E-17

Table 5.1: t test for each indicator between Non-Agricultural and Agricultural districts

Indicator	**t statistic**	**p value**
MSL_Change	33.28462995	2.32E-13
MSW_Change	5.019960137	0.000316
FC_Change	28.24321421	1.66E-10
BF_Change	51.0519094	2.22E-16
CHH_Change	43.96621411	2.60E-17
Asset_Change	52.24780108	5.20E-19

Table 5.2: t test for each indicator between Non-Agricultural and High Unemployment districts

Results of one way ANOVA test is shown in Table 5.3 and post hoc test is shown in Table 5.4. The ANOVA test establishes that there is significant difference between the agricultural , non agricultural and high unemployment groups and the post hoc tests shows that Non Agricultural districts have seen maximum improvement in Asset Ownership. Same results are found in other indicators except main source of water where agricultural districts have seen maximum growth.

Indicator	F score	P-value	F crit
Asset	2784.993322	1.44E-57	3.158843
BF	9135.657166	3.49E-72	3.158843
FC	3336.113212	8.82E-60	3.158843
CHH	2629.45069	7.30E-57	3.158843
MSL	1525.585053	3.21E-50	3.158843
MSW	169.3387049	1.04E-24	3.158843

Table 5.3: One Way ANOVA test results for each indicator between Non-Agricultural, Agricultural and High Unemployment districts

Group1	Group2	Mean difference
Non Agricultural	Agricultural	0.3454
Non Agricultural	Unemployment	0.3149
Agricultural	Unemployment	-0.0305

Table 5.4: Post hoc test for each asset ownership between Non-Agricultural, Agricultural and High Unemployment districts

5.1.3 Observation and Analysis

This observation is not surprising considering that as per the 2018 India Wage Report by the International Labour Organization (ILO), there were large wage disparities between workers based on occupation, literacy, and gender [23]. The average daily wages of regular workers in the primary sector (agriculture and allied activities) was only ₹ 192, while the wages in the secondary and tertiary sectors were ₹ 357 and ₹ 424 respectively [23]. This substantial wage disparity would provide higher disposable income to households engaged in non-agricultural activities, to improve their indicators faster.

Combined with the observation in Table 4.3 that non-agricultural districts also tend to be at a higher level on various indicators than agricultural and high-unemployment districts, this shows a trend towards increasing inequality. The better-off districts are moving more rapidly towards an improved life than less well-off districts, highlighting that less well-off districts may require policy attention so that they are not left behind for long. This phenomenon of rising inequality has been explained by the economist Simon Kuznets [18, 12] who states that as an economy develops, market forces first increase and then decrease economic inequality, and is depicted by the inverted U-shaped Kuznets curve. This happens because when an economy moves from agricultural to non-agricultural employment, an influx of cheap rural labour leads to diminished wages and rising inequality. As the economy industrializes further to absorb surplus labour, and aided also by social welfare mechanisms, the inequality is expected to decrease. Based on our data analysis, during 2001 to 2011 India seems to have been on the rising part of the Kuznets curve of increasing inequality. As we show later, this inequality is now observable across districts because industrialization has not spread in a spatially equitable manner in India and has remained concentrated in the same geographical regions.

5.2 Hypothesis 2

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

5.2.1 Probability Analysis of Hypothesis 2

Out of the six indicators, four of them (asset ownership, bathroom facility, fuel for cooking, and condition of household) are likely to be governed by choices made by households of where to invest their disposable income: Should we invest in purchasing assets, or improve sanitation facilities at our home, or use a better fuel for cooking? We call these discretionary variables because they seem to be more about household preferences than limited by government investments in social infrastructure. Although the choice of bathroom facility and fuel for cooking can be influenced by government assistance, such as by providing sewer connections for bathrooms or distributing LPG for cooking fuel, but irrespective of this government support an individual household can still upgrade itself if needed. For example, pit latrines, covered slabs, or septic tanks can be used instead of piped sewers, or instead of an LPG connection healthier alternatives for fuel for cooking can be used such as coal, or kerosene.

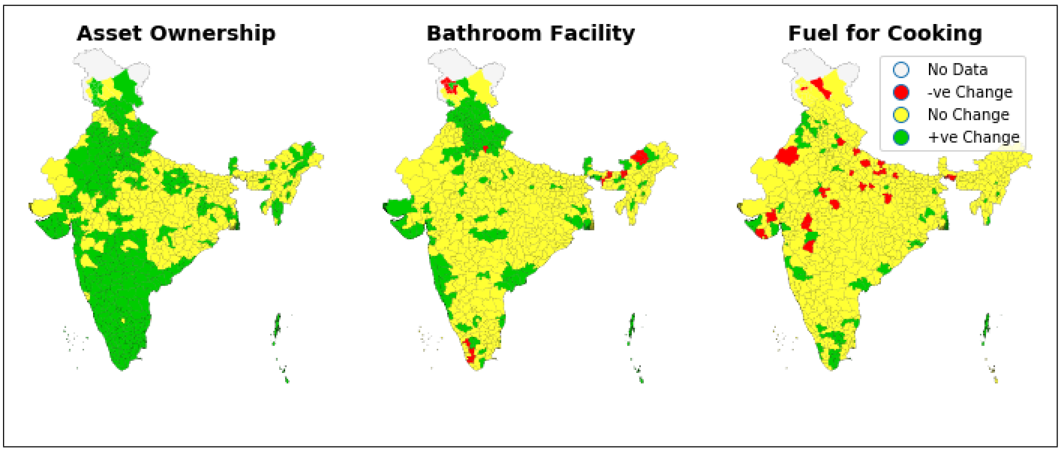

Figure 5.1: Change in the levels of districts between 2001 and 2011, for the discretionary variables of asset ownership, bathroom facility, and fuel for cooking.

To find out the relative investment preferences among these four variables, we again review Table 4.2. We can clearly see that asset ownership seems to be changing the fastest irrespective of the type of the district. This remains consistent even when dis-aggregated based on the current status of the variables. People seem to invest more readily for assets than for other amenities that would lead to a healthier lifestyle for the household. We visualize this in Figure 5.1 which shows on a map how asset ownership shows more positive change (most amount of green) than bathroom facilities and fuel for cooking (least amount of green).

5.2.2 Statistical Test of Hypothesis 2

To test this hypothesis we created four groups (Asset ownership, bathroom facility , fuel for cooking and condition of household) . The method of creating groups is same as explained in Section 5.1.2 where we have taken random samples of 80 percent of districts for each group and then counted the fraction of districts which have seen positive change. The result of T-test is shown in Table 5.5. T-test shows that change is asset is more than other indicators

Indicator 1	Indicator 2	T statistic	p-value
Asset	BF	87.5	5.36E-38
Asset	FC	138	3.09E-40
Asset	CHH	72	5.83E-40

Table 5.5: Results of T-test between Asset and remaining 3 indicators

The result of ANOVA test between four discretionary variables shows the values F = 7682.935, p-value= 3.03E-94, F-Critical = 2.724944. Higher F score indicates that there is significant difference between indicators. To find out the best performing indicator we conducted post hoc test the results of which are shown in Table 5.6. They clearly point towards the fact that Asset Ownership is the fastest growing indicator.

Group1	**Group2**	**Mean difference**
Asset	BF	0.3033
Asset	CHH	0.273
Asset	FC	0.461
BF	CHH	-0.03
BF	FC	0.157
CHH	FC	0.188

Table 5.6: Post hoc tests between all four discretionary variables

5.2.3 Observation and Analysis

This choice can be influenced by many factors. Consumer behaviour research shows that personal, interpersonal, and cultural effects can shape asset acquisition and usage, and in particular assets are often seen as status symbols

[20]. Alternately, assets such as mobile phones could be economically useful as well, as reported in the context of fishermen and wholesalers in South India [15]. Further, the amount of disposable income may also pay a role: A cheap mobile phone would cost much less than installing a simple water filter, for example. This raises the need for further studies to understand the reasons behind their preferences on such expenditure.

Another pattern which emerges from Table 4.2 is that districts which are already at level-2 have better chances of improving their level. Consider asset ownership for which the P(positive change | agricultural districts) at level-2 (0.714) is higher than P(positive change | agricultural districts) at level-1 (0.498), and similarly for bathroom facility and other indicators where in general level-2 districts have a higher probability to improve than level-1 districts. This again points towards growing inequality, which augmented with the observation that households prefer to invest in assets over other amenities that might be more important for a healthier life, suggests that suitable policies should be created to not leave the less well-off districts behind and to further nudge households towards making more appropriate investments [28].

5.3 Hypothesis 3

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

5.3.1 Probability Analysis of Hypothesis 3

In the previous hypothesis we considered indicators which can be upgraded based on a household's discretion. We next consider variables which require significant government support, and call them social infrastructure variables. This includes the main source of light, where electrification is predominantly provided by government utility companies, and the main source of water which requires tap water distribution networks or handpumps and wells funded by the government.

Examining P(positive change | type of employment) from Table 4.2, we observe that both these indicators have improved considerably over the years, demonstrating government support. However, between the two, the main source of light has improved much more, with non-agricultural districts showing the strongest change, but not any significant differences based on the current status of a district. This indicates a consistent effort in infrastructure provisioning by the government irrespective of the current status of a district, but with a bias towards electrification over drinking water provisioning. Figure 5.2 also shows this visually, about the extent of change in development levels for the main source of light and the main source of water.

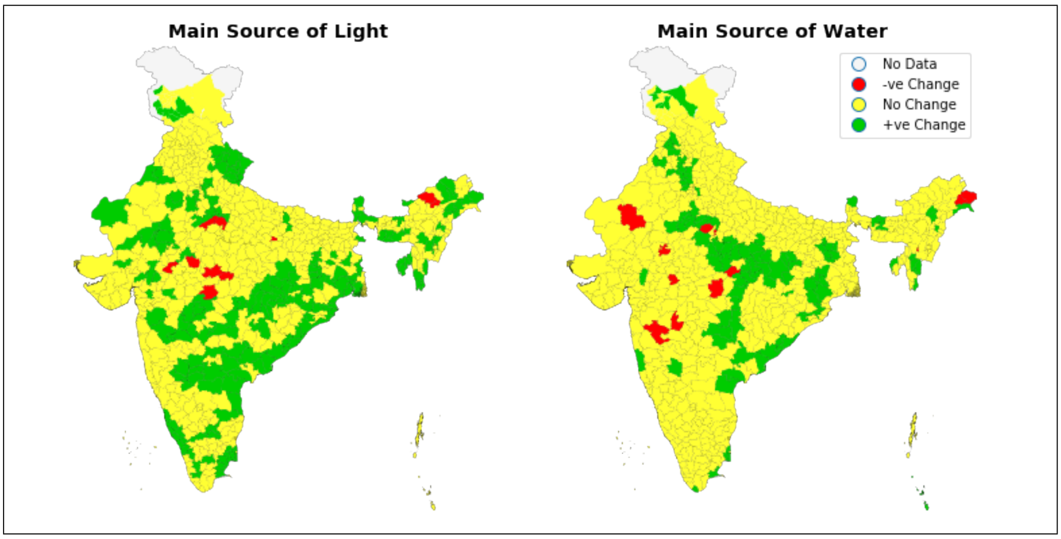

Figure 5.2: Change in levels of districts between 2001 and 2011, for the social infrastructure variables of main source of light and the main source of water

5.3.2 Statistical Test of Hypothesis 3

To test this hypothesis we created two groups (main source of light and main source of water). The method of creating groups is same as explained in earlier sections. The t-test gave a T-statistic of 61.4 and p-value of $2.11 * 10^{-36}$, which clearly validates the hypothesis that main source of light has seen more development.

The results of one way ANOVA test with F = 3770.643 and p-value = $1.24 * 10^{-39}$ also clearly indicates that main source of light has seen more development. Post hoc test also gives the mean difference of 0.2055 between the group of main source of electricity and main source of water, thus validating the hypothesis.

5.3.3 Observation and Analysis

Historical reports indeed point towards this gap. The introduction of the Electricity Act of 2003 encouraged participation of the private sector in electricity production, and the Rajiv Gandhi Grameen Vidyutikaran Yojana (RGGVY) launched in 2005 was aimed at creating rural electrification infrastructure so as to electrify all villages and give electricity connections free of charge to families living below the poverty line. The evaluation gave a 93.3% success rate to RGGVY in meeting its targets [10], and access to electricity rose from 59 % of the population in 2000 to 74 % in 2010 [24]. In contrast, the government also launched the Bharat Nirman Program in 2005 with an emphasis on providing drinking water especially to habitations affected by poor water quality, but achieved limited goals with six states reporting less than 50% achievement against targets, and an average achievement of 80.4% [22].

5.4 Hypothesis 4

Discretionary spending by households is closely related to literacy and formal employment and does not seem to be affected by social infrastructure provisioning by the government.

5.4.1 Calculation of Mutual Information for Analysis of Hypothesis 4

We next check what factors may affect the discretionary spending by households, or rather since we cannot make any causality claims, we check which factors may be stronger predictors of discretionary spending. We check for four factors: literacy, formal employment, current status, and government support for social infrastructure. Variables literacy and formal employment were also discretized into 3 levels of development. Since our variables are discretized into multiple ordinal levels, instead of studying a correlation between pairs of different variables we choose to examine the mutual information between the pairs to form an assessment of the strength of relationship between them.

Literacy	**Asset Ownership**					
	Non Agricultural		**Agricultural**		**High Unemployment**	
	No Change	+ve change	No Change	+ve change	No Change	+ve change
Level-1	0.0101	0.0152	0.1922	0.1417	0.1315	0.1248
Level-2	0.0304	0.0641	0.027	0.0641	0.0287	0.0472
Level-3	0.0388	0.0455	0.0017	0.0236	0.0017	0.0118

Table 5.7: Shown is the relationship between change in levels of asset ownership with literacy, sliced by the type of employment of the district. This is an example to study the relationship between changes in discretionary variables with factors such as literacy that might predict the change

We first create tables by calculating probabilities of +ve change/no change, such as the one shown for the relationship between literacy and change in asset ownership, in Table 5.7 for the three types of districts. The sum of the probabilities in each table is equal to 1. We then calculate the mutual information between the two variables:

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$$

where $p(x,y)$ is the joint probability function of x and y, and $p(x)$ and $p(y)$ are the marginal probabilities of x and y respectively. In Table 5.7, x ranges over the six classes of asset ownership and y ranges over the three levels of literacy. The same method is used to calculate mutual information between each of the four factors of interest in this hypothesis (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). The other tables used for calculation of mutual information are given in the Appendix.

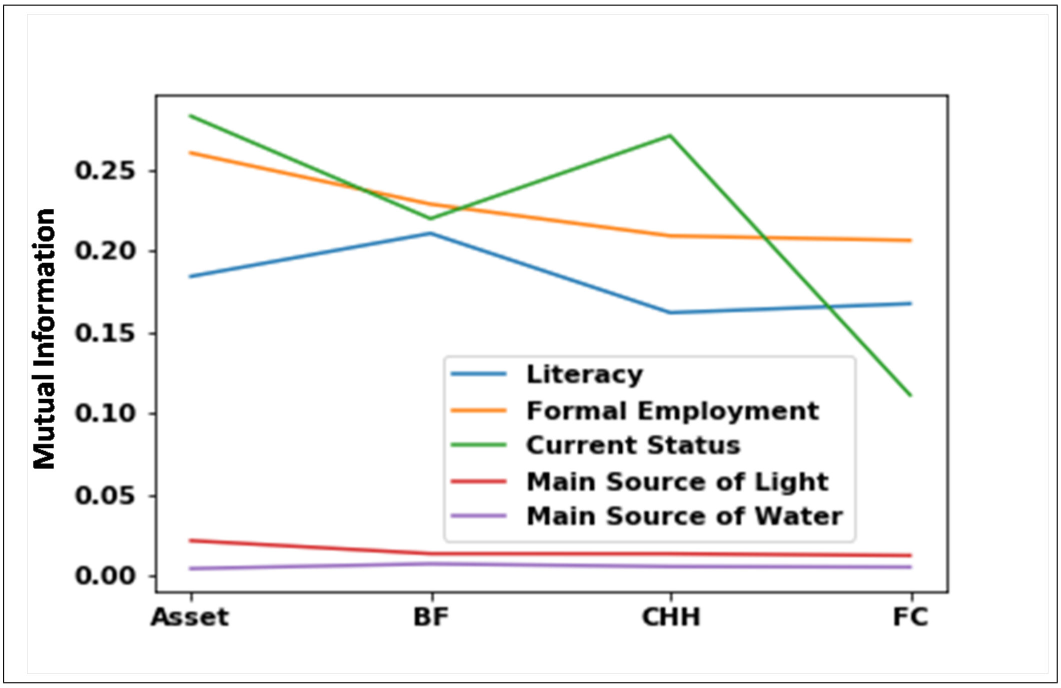

Figure 5.3: Mutual information derived between change in discretionary variables (x-axis) and factors that might predict the change

Figure 5.3 shows the mutual information calculated on these tables. We can see that literacy, formal employment, and the current status are all more predictive of change in discretionary variables as compared to government support on the main source of lighting and main source of water.

5.4.2 Observation and Analysis

Similar gaps exist in wages for different levels of education, and also between educated women in formal and unorganized sector employment. A deeper analysis indeed validates that districts with high formal employment in general are more likely to change positively in discretionary indicators. A similar pattern is observed for literacy. Formal employment and literacy are also correlated with each other, with a Pearson correlation coefficient of 0.61. This validates the observations that formal employment and literacy go hand in hand, and formal employment leads to more disposable income that can be spent on discretionary variables.

An interesting insight however is that government spending towards social infrastructure provisioning does not seem to be related to discretionary spending, emphasizing even more on the need to increase disposable income to bring about positive changes in living conditions for people.

5.5 Hypothesis 5

Districts with more manufacturing and services industries end up developing faster. However, the presence of these industries has not spread geographically and has remained spatially concentrated in the same regions over the years.

5.5.1 Probability Analysis of Hypothesis 5

We observed earlier that non-agricultural districts saw the fastest improvement in all indicators, other than the main source of water. Non-agricultural employment can be further divided into several industries such as manufacturing, retail, real estate, banking, etc. We aggregated these industry sectors into four categories shown in Table 5.8, and as before we then did a k-means clustering based on the percentage of population employed in each of these categories.

(a)	**Business activity**: Includes the sectors of retail, wholesale, transportation , storage, hotel and restaurant etc.
(b)	**Services**: Includes services sectors like electricity, gas, water supply, defence forces, administration, social security, education, health and social work etc
(c)	**Construction and mining** sector
(d)	**Manufacturing** sector

Table 5.8: Industry sectors aggregated into four broad categories

We were thus able to label each district in terms of the dominant nature of employment in the district. A good clustering was obtained for k = 4, and a boxplot for these district-types is shown in Figure 5.4, labeled for a high presence of manufacturing industries (Type-4), services industries (Type-3), a moderate industry presence (Type-2), and low industry presence (Type-1).

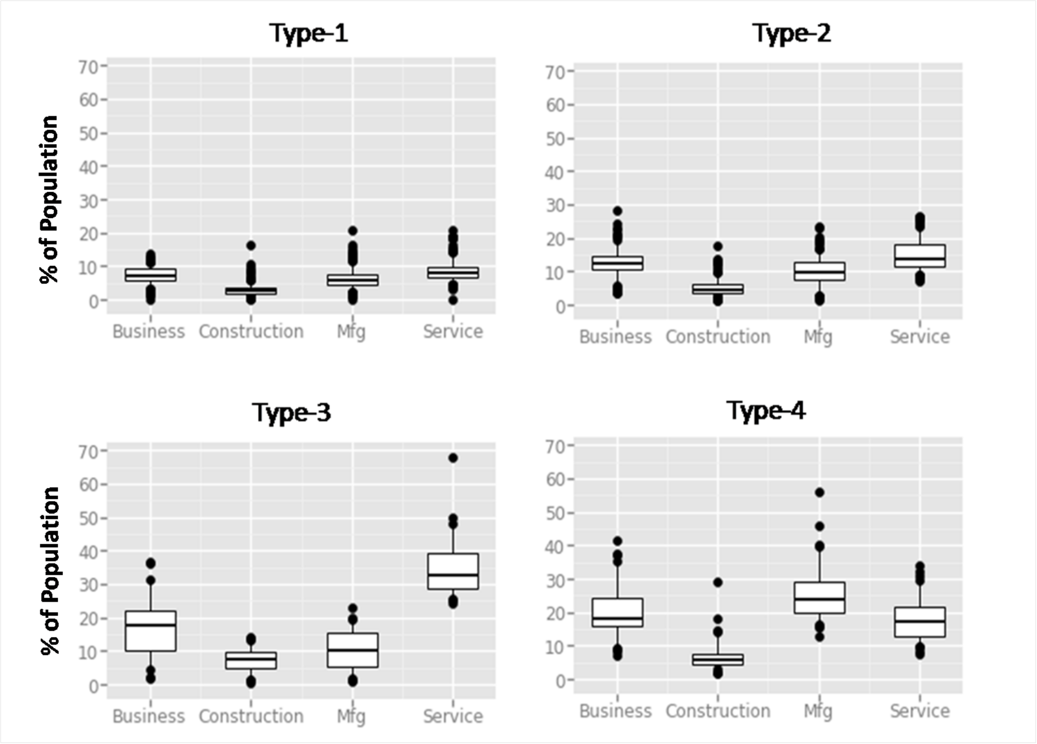

Figure 5.4: Shown is the distribution of population employed in four industrial categories (business, construction, manufacturing, and services), across four district clusters. The clusters are used to label districts in terms of their dominant type of industrial employment

Table 5.9 shows the change in various discretionary and social infrastructure variables, with respect to the type of industries in the district. We can make out a clear pattern that type-3 and type-4 districts (having services and manufacturing industries) are more likely to grow faster in all the variables except the main source of water. These industries also seem to favour formal employment, as seen with a correlation coefficient of 0.77.

Variable	Existing Status	Type1	Type2	Type 3	Type 4	Total
Asset Ownership	Level-1	0.385	0.678	0.714	0.867	0.571
	Level-2	0.667	0.872	0.875	0.783	
	Total	0.388	0.726	0.784	0.83	
Bathroom Facility	Level-1	0.066	0.369	0.4	0.517	0.269
	Level-2	0.121	0.383	0.788	0.714	
	Total	0.072	0.374	0.737	0.6	
Fuel for Cooking	Level-1	0.032	0.132	0.19	0.478	0.111
	Level-2	0	0.236	0.5	0.389	
	Total	0.022	0.165	0.217	0.439	
Main Source of Light	Level-1	0.164	0.574	1	0.8	0.462
	Level-2	0.586	0.662	0.692	0.667	
	Total	0.343	0.622	0.714	0.696	
Main Source of Water	Level-1	0.592	0.5	0.25	0.267	0.25
	Level-2	0.05	0.141	0.333	0.238	
	Total	0.231	0.305	0.261	0.25	
Condition of Household	Level-1	0.224	0.375	0.75	0.636	0.3
	Level-2	0.107	0.361	0.607	0.447	
	Total	0.169	0.366	0.65	0.49	

Table 5.9: Shown is the probability of positive change in indicators for districts having different types of industrial employment. This is also disaggregated into the probability of positive change based on the existing status of districts. The overall probability of positive change for an indicator is given in the last column

5.5.2 Statistical Test of Hypothesis 5

We created four groups based on type of industry sector (Type-1, Type-2, Type-3 and Type -4) for each indicator. The method of creation of groups is same as mentioned in previous hypothesis tests. The results of T-test between different type of districts base on Industry sector for asset ownership is shown in Table 5.10. The tests show that for asset ownership, districts with more manufacturing (Type-4) and services (Type-3) industries end up developing faster. Similar results are obtained for all other indicators except main source of water where moderate industry presence (Type-2) is growing faster.

Sector 1	**Sector 2**	**T statistcs**	**p value**
Type-4	Type3	3.87	0.000453789
Type-4	Type-2	16.524	2.11E-14
Type-4	Type1	65.653	2.51E-34
Type-3	Type-2	7.5407	1.92E-07
Type-3	Type1	45.857	6.55E-26
Type-2	Type2	86.859	1.67E-38

Table 5.10: Results of T -test between different type of districts based on Industry sector for Asset Ownership

We also conducted one way ANOVA test and post hoc tests to test the hypothesis. The result of ANOVA test between type of districts based on Industry sector is shown in Table 5.11 and the post hoc results for asset ownership is shown in Table 5.12. ANOVA combined with post hoc tests indicate that for asset ownership districts with more manufacturing(Type-4) and services(Type-3) industries end up developing faster. Similar results are obtained for all other indicators except main source of water where moderate industry presence (Type-2) is growing faster.

Indicator	**F Score**	**P-value**	**F crit**
Asset	1415.12591	1.45E-66	2.724944
BF	2169.774453	1.62E-73	2.724944
FC	493.1408912	1.04E-49	2.724944
CHH	950.2484403	3.88E-60	2.724944
MSL	389.5600671	4.92E-46	2.724944
MSW	16.67235966	2.04E-08	2.724944

Table 5.11: ANOVA test between type of districts based on Industry sectors

Group1	**Group2**	**Mean difference**
Type-4	Type-3	0.038
Type-4	Type-2	0.1005
Type-4	Type-1	0.4363
Type-3	Type-2	0.062
Type-3	Type-1	0.398
Type-2	Type-1	0.338

Table 5.12: Results of post hoc tests for Asset Ownership

5.5.3 Observation and Analysis

Put together with the earlier hypotheses, this shows that discretionary variables tend to improve with improvements in formal employment in the manufacturing and services sectors by generating disposable income. These sectors are likely to employ more literate people and hence we see a strong correlation between literacy and formal employment. Among discretionary variables people seem to choose to invest in assets before other essential amenities, possibly pointing to the need for policies to nudge behavior towards making appropriate household investments. Finally, all of these variables seem to improve faster in level-2 districts than level-1 districts, which raises a concern about growing inequality. Given the importance therefore of formal employment in the manufacturing and services sectors in the overall development process, we go on to investigate patterns in the growth of these sectors over the years between 2001 to 2011.

Figure 5.5 plots the dominant industrial type in each district, for the years of 2001 and 2011. There has clearly not been much change over an entire decade. This shows that the manufacturing and services sectors have not expanded to other geographies, an observation also made in other studies [17, 19].

A few regions in Maharashtra (close to Mumbai) and Tamil Nadu (garments and textile industry) have seen a spatial expansion of manufacturing and services industries, likely due to a spillover to neighbouring districts as a consequence

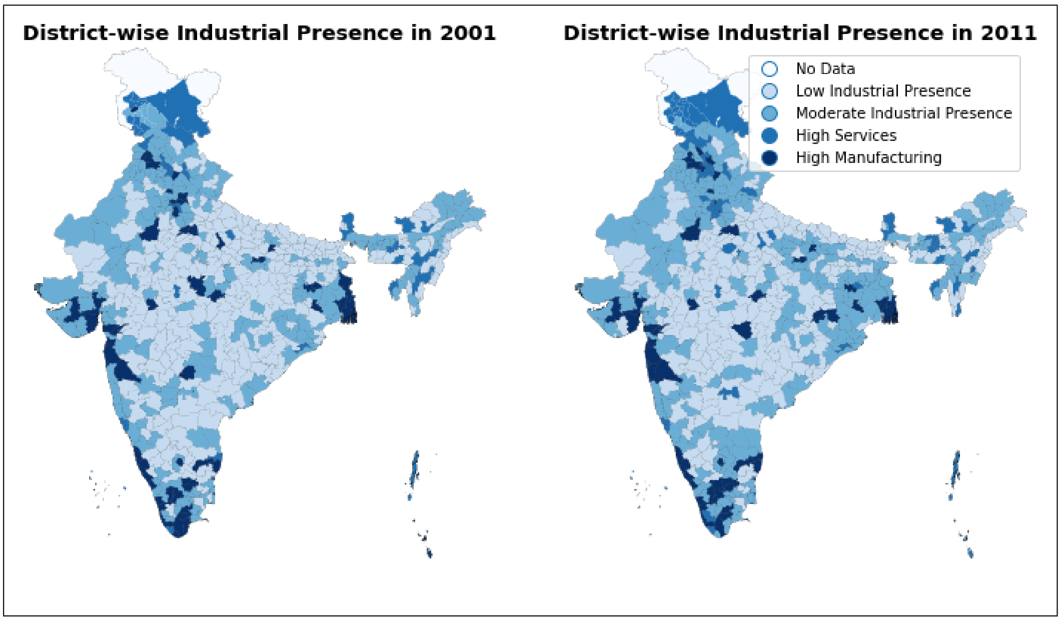
District-wise Industrial Presence in 2001
District-wise Industrial Presence in 2011
○ No Data
● Low Industrial Presence
● Moderate Industrial Presence
● High Services
● High Manufacturing

Figure 5.5: Districts are colour coded based on their type of industrial employment in 2001 and 2011

of growth in the industries. Some new hubs seem to have emerged in Maharashtra (Nagpur) and Orissa (Sundargarh), both of which are known to have strong local industries. An increase is also seen from a low industrial presence to a medium presence in large areas of Tamil Nadu, Orissa, Eastern Uttar Pradesh, and Bihar. Interestingly regions in West Bengal (Murshidabad) and Madhya Pradesh (Sagar) have actually seen a decrease. Overall however, this observation points towards the need for policies to create more widespread non-agricultural employment. This does not however answer questions about the consequence of this spatially concentrated growth in non-agricultural employment. Other studies show that it has led to an increase in rural-urban migration, which tends to be exploitative, and diminishes the ability to distribute economic growth equitably across the country [11].

5.6 Hypothesis 6

Female participation in the workforce has decreased, primarily with a reduction in marginal employment that has not been compensated with an equivalent increase in female main employment.

5.6.1 Probability Analysis of Hypothesis 6

Earlier in Table 4.1 we observed a sharp fall for female employment as marginal workers (less than six months of employment in a year). This could be either due to an increase in formalization that saw more women moving from marginal workers to main workers (more than six months of employment in a year), or an overall decrease in female employment itself.

Female Employment main	**Female Employment marginal**		
	+ve change	-ve change	No change
+ve change	7	33	43
-ve change	1	28	16
No change	9	210	246

Table 5.13: Shown is the number of districts that have changed positively, negatively, or not changed in their level for the variable of female main employment, against changes in their level for the variable of female marginal employment

Table 5.13 shows a confusion matrix for the change in levels for main and marginal female workers between 2001 and 2011. We can see that there are 210 districts which saw a fall in female marginal employment but with no change in main employment, and an additional 28 districts which saw a fall even in the main employment of women. Only 83 districts have actually shown an increase in their levels for female main employment. This seems to indicate that although there is a positive movement from marginal to main employment in some districts, but overwhelmingly more often women are actually falling out from the workforce, especially in marginal employment.

Type of District	**Female Employment Main**			**Female Employment Marginal**		
	+ve Change	-ve Change	No Change	+ve Change	-ve Change	No Change
Non Agricultural	0.1	0.12	0.78	0.04	0.41	0.55
Agricultural	0.16	0.05	0.79	0.03	0.37	0.61
High Unemployment	0.14	0.08	0.78	0.02	0.6	0.38

Table 5.14: Shown is the number of districts that have changed positively, negatively, or not changed in their level for the variable of male employment(main and marginal) for each type of district (Non-Agricultural, Agricultural and High Unemployment)

Table 5.14 shows the change in levels for main and marginal female workers between 2001 and 2011 in each type of district. Observing female employment patterns for different types of districts, we find that agricultural and non-agricultural districts are not very different from each other. Both have seen a large fraction of districts (37% and 41% respective) reduce in female marginal employment, and a small fraction of districts (16% and 10% respective) increase in female main employment. This seems counter-intuitive because non-agricultural districts should expect to see a stronger movement to main employment than agricultural districts; the phenomenon therefore needs to be monitored and further studies should be conducted to check whether the transition from marginal to main employment for women is happening in a robust manner as predicted by the U-shaped hypothesis.

Type of Industry	**Female Employment Main**			**Female Employment Marginal**		
	+ve Change	-ve Change	No Change	+ve Change	-ve Change	No Change
Type-1	0.18	0.04	0.78	0.02	0.43	0.55
Type-2	0.11	0.13	0.76	0.03	0.54	0.43
Type-3	0.12	0.1	0.79	0.06	0.31	0.63
Type-4	0.07	0.07	0.85	0.04	0.46	0.49

Table 5.15: Shown is the number of districts that have changed positively, negatively, or not changed in their level for the variable of female employment(main and marginal) for each type of Industry

Table 5.15 shows the change in levels for main and marginal female workers between 2001 and 2011 in different types of districts based on industries. We observe that Type-1 districts have seen the best positive change while it has decreased the most in Type-2 districts as per female main employ-

Level	T Statistic	p-value
Level-1	36.23	3.47E-17
Level-2	35.71	3.70E-18
Level-3	-68.23	1.11E-22

Table 5.16: Statistical tests corresponding to Hypothesis-6: Shown are the t-scores and p-values for changes between 2001 and 2011 in the levels of female marginal employment

ment. Marginal employment for females has decreased throughout all type of industries with negligible increase in some districts.

5.6.2 Statistical Test of Hypothesis 6

To test the hypothesis we created two groups (one for 2001 and one for 2011) for each level of female marginal employment districts. Similarly to other hypothesis testing we randomly selected 80 percent of districts and then found out the fraction of districts at respective level of female marginal employment in 2001 and 2011 respectively. Then we conducted t test between these two groups for each level. The result of the t test are shown Table 5.16. The high negative t statistic for level 3 indicates that huge number of districts have moved from level -3 to lower level in female marginal employment. The high positive t statistic for level-1 and level -2 also indicates that districts have moved to level 1 and level 2 from higher level between 2001 and 2011.

5.6.3 Observation and Analysis

Several studies try to explain the falling out of women from the workforce [27, 26] and cite reasons such as poor working conditions and wage disparities between men and women, which demotivate women to take up work, especially as women are getting more educated. Some models of development that relate economic growth with gender equality suggest that this is actually expected [13, 21]. As household income increases at the same time as the economy moves from agricultural to non-agricultural employment, women who earlier were involved in agricultural work begin to move out from the workforce. Eventually as education and formal employment increases, women begin to enter the workforce again, leading to a U-shaped function for female employment. According to our data analysis, India seems to have been on the falling part of the curve during 2001 to 2011.

5.7 Hypothesis 7

Male participation in the workforce has increased, primarily with a considerable increase in marginal employment and a slight reduction in main employment.

5.7.1 Probability Analysis of Hypothesis 7

In the previous hypothesis we observed a sharp fall for female employment as the number of marginal workers reduced. These characteristics are completely different when we evaluate male employment. Table 5.17 shows a confusion matrix for the change in levels for main and marginal male workers between 2001 and 2011.

Male Employment Main	**Male Employment Marginal**		
	+ve Change	-ve Change	No Change
+ve Change	29	1	18
-ve Change	38	7	42
No Change	204	9	245

Table 5.17: Shown is the number of districts that have changed positively, negatively, or not changed in their level for the variable of male main employment, against changes in their level for the variable of male marginal employment

We can see that there are a total of 87 districts which saw a fall in male main employment while only 17 districts saw a fall in male marginal employment. A large number of districts (271) saw a positive change in male marginal employment and 48 districts saw a rise in the main employment of males. This seems to indicate that although there has been a fall in main employment, marginal employment has seen a huge increase.

Type of District	Male Employment Main			Male Employment Marginal		
	+ve Change	-ve Change	No Change	+ve Change	-ve Change	No Change
Non Agricultural	0.13	0.11	0.76	0.41	0.04	0.55
Agricultural	0.06	0.16	0.78	0.36	0.03	0.61
High Unemployment	0.08	0.15	0.77	0.6	0.02	0.37

Table 5.18: Shown is the number of districts that have changed positively, negatively, or not changed in their level for the variable of male employment(main and marginal) for each type of district (Non-Agricultural, Agricultural and High Unemployment)

Table 5.18 gives the distribution of male employment in terms of type of districts. Increase in main employment for men is seen in non-agricultural districts. While increase in marginal employment is seen across all types of districts, it is the highest in high unemployment districts. This phenomenon indicates towards governments efforts to provide marginal employment in the high unemployment districts.

5.7.2 Statistical Test of Hypothesis 7

To test the hypothesis we created two groups (one for 2001 and one for 2011) for each level of female marginal employment districts. Similarly to other hypothesis testing we randomly selected 80 percent of districts and then found out the fraction of districts at respective level of female marginal employment in 2001 and 2011 respectively. Then we conducted t test between these two groups for each level. The result of the t test are shown Table 5.19. The high negative t statistic for level 1 indicates that huge number of districts have moved from level 1 to higher level in male marginal employment. The high positive t statistic for level-2 and level -3 also indicates that districts have moved from level 1 and level 2 to higher level between 2001 and 2011.

5.7.3 Observation and Analysis

While there has been an overall increase in male employment in contrast to female employment which has decreased, it is mainly because of the increase in

Level	**T Statistic**	**p-value**
Level-1	-66.111	1.01E-22
Level-2	35.71	3.40E-17
Level-3	35.85	3.43E-18

Table 5.19: Statistical tests corresponding to Hypothesis-7: Shown are the t-scores and p-values for changes between 2001 and 2011 in the levels of male marginal employment

marginal employment (employment less than 6 months). This seems largely an impact of the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) which was launched in February 2006 [8] and aimed at enhancing the livelihood security of the people in rural areas by guaranteeing hundred days of wage employment in a financial year, to a rural household whose members volunteer to do unskilled manual work. In Phase-I it was introduced in 200 of the most backward districts of the country and in an additional 130 districts in Phase II in 2007-2008. With a total budget outlay of ₹ 11300 crores in financial year 2006-2007 which increased to ₹ 40100 crores in financial year 2010-2011 it definitely has impacted a large number of rural households.

Chapter 6

Change Prediction

In this chapter we look at various methods to predict the change in socio-economic variables. We first see the performance of a logistic regression model followed by the formulation of a Bayesian Network. An additional method to see the use of Akaike Information Criterion has also been discussed.

6.1 Change prediction classifier and its result

We created two classification models to see if we can predict the change in discretionary variables between 2001 and 2011. 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 (labels as per 2001) as the features to predict the outcome (change in 2011). 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 [6] on the training dataset to address class imbalance issues. SMOTE creates new minority class instances (synthetic) between existing (real) minority instances.

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 6.1: Accuracy and F1-scores for Change prediction

Table 6.1 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.

6.2 Akaike Information Criterion to evaluate Socio-economic variables

The change in any indicator can depend on various parameters, so the question arises that which parameters are more suited to predict the change. In our analysis we have seen that spending on discretionary variables is related to various parameters. We also built a logistic regression model to predict the change in discretionary variables. Different models give different results. In terms of current state of a variable for a given discretionary variable we want to check whether the current state of economic development is a stronger predictor of change compared to the investment in social infrastructure? This is similar to Section 5.4 where we have already shown the high mutual information for a current state of the discretionary variable. To answer this question we compared the Akaike Information Criterion (AIC) of different models. The independent feature in first model was current state of that indicator. In the second model change in main source of lighting was the independent feature while in the third model it was change in main source of water. We then tried to predict the change in the label of each of the four discretionary variables. We compared the AIC scores of each model to find the indicator which is best suited to predict the change. Table 6.2 shows the AIC scores which clearly depicts that information loss is least when we try to predict the change from the current state of discretionary variables.

Discretionary variable	**AIC: Prediction of change in Economic Indicator**		
	Current state	**Change in MSL**	**Change in MSW**
Asset	700	811	758
Fuel for cooking	345	570	568
Bathroom facility	631	716	707

Table 6.2: AIC score for prediction of change in discretionary variable from current state, change in main source of light and change in main source of water

6.3 Predictions through Bayesian Network

A Bayesian model/network is a probabilistic directed acyclic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph. The bayesian network can be used to predict the values of any output variable. Given the evidences, the bayesian network can predict the likely output of the test variable. We constructed bayesian networks to predict the change in discretionary variables. The method of construction of bayesian network for our model is explained in subsequent paragraphs.

6.3.1 Association Rule Mining

We used Association Rule Mining(ARM) to construct the bayesian network by identifying the interdependence between the input variables. For our analysis we tried to predict the change in discretionary variables based on the parameters used in Hypothesis 4 i.e. formal employment , literacy, current state of the variable and change in social infrastructure variable (main source of light and main source of water). We used Apriori algorithm to find the frequent item set. For our analysis the items sets are labels of input parameters. The results ofARM were used to identify significant dependency between variables, and thus lead to construction of the bayesian network.

6.3.2 Bayesian Network Created from Census Labels

Here we explain the construction of bayesian model to predict the change in asset ownership. As per our belief we considered four factors that can influence the change in asset ownership : Formal employment , Literacy, current state and change in social infrastructure (main source of light and main source of water). We created an initial bayesian network as shown in Figure 6.1 which may contain some unnecessary nodes/edges.

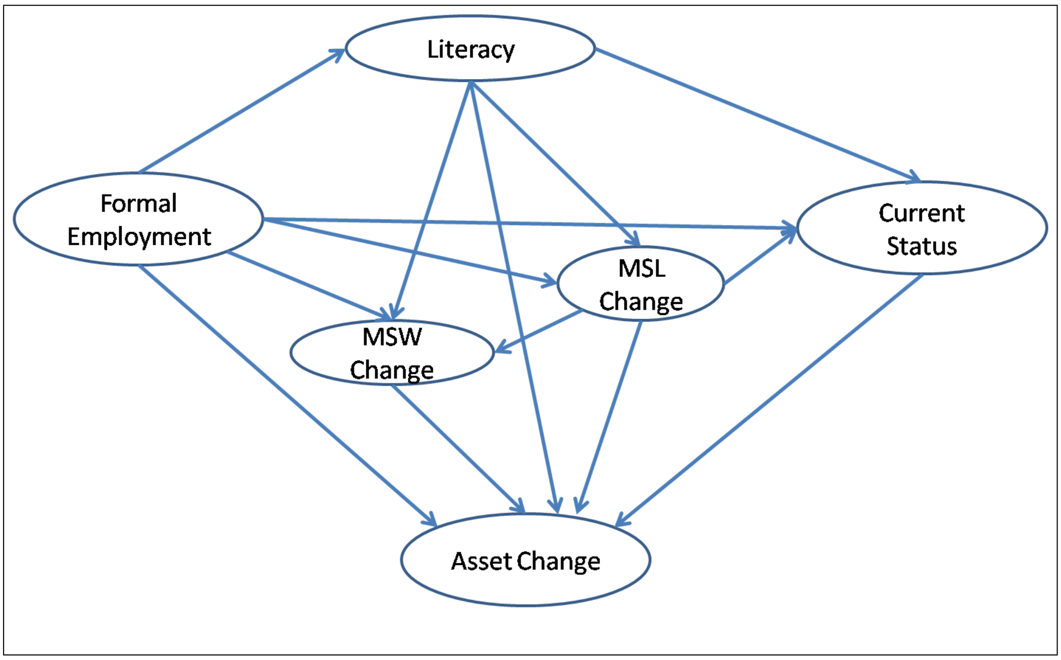

Figure 6.1: Bayesian network based on initial belief to predict the change in asset ownership

The network was subsequently updated by adding/deleting the nodes/edges based on the results of the ARM. With ARM, interesting rules were determined by the confidence and support measure. Interestingness thresholds for support and confidence measures are taken as 0.15 and 0.9 respectively, however we found that there were very less rules above the threshold containing asset change, so we relaxed the threshold only for the rules containing asset change to 0.09 and 0.7 for support and confidence respectively. We deleted the node MSW Change along with incoming and outgoing edges as there were very less number of rules having MSW in antecedents. The updated bayesian network is shown in Figure 6.2. Similarly, the bayesian network for prediction of change in bathroom facility, fuel for cooking and condition of household is shown in Figure 6.3, Figure 6.4 and Figure 6.5 respectively.

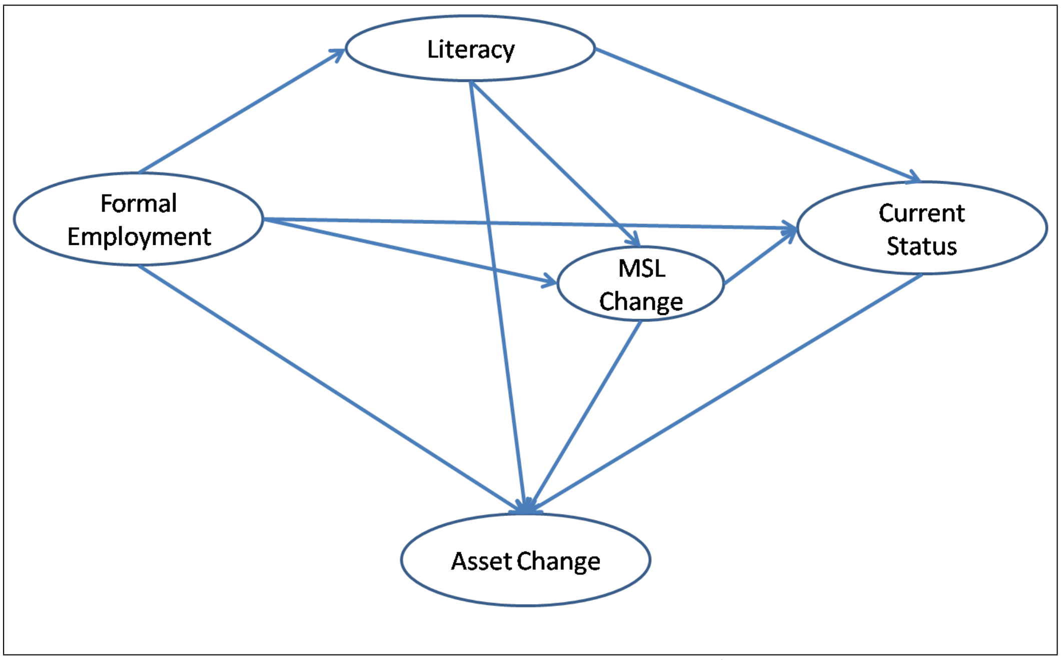

Figure 6.2: Updated Bayesian network based on Association Rule Mining to predict the change in Asset Ownership

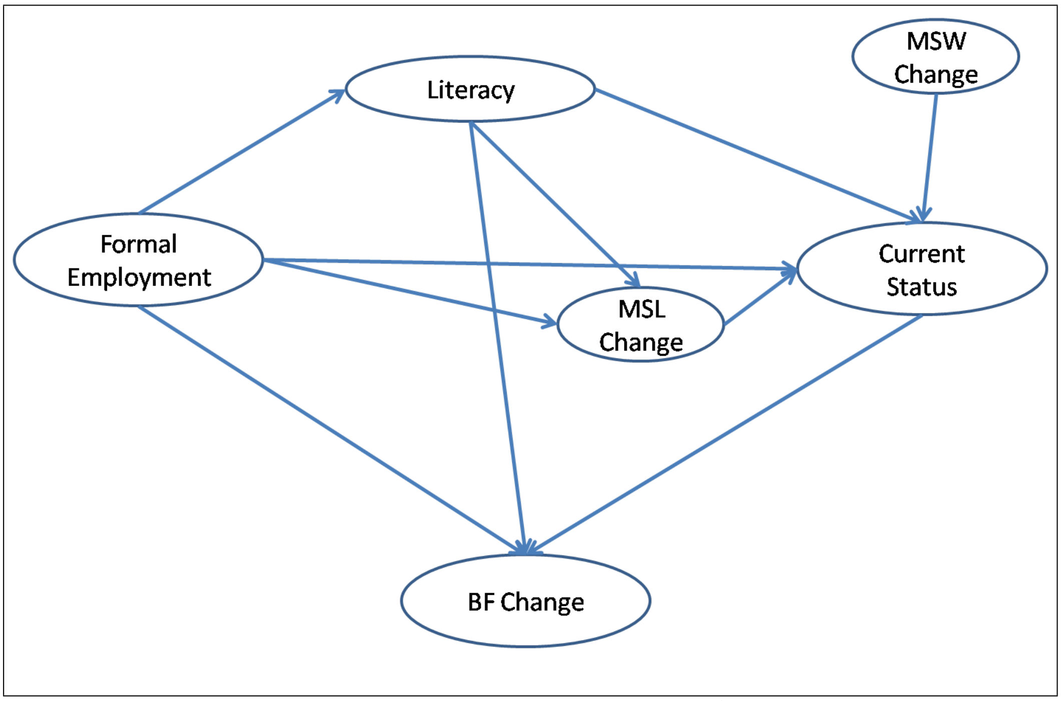

Figure 6.3: Updated Bayesian network based on Association Rule Mining to predict the change in Bathroom Facility

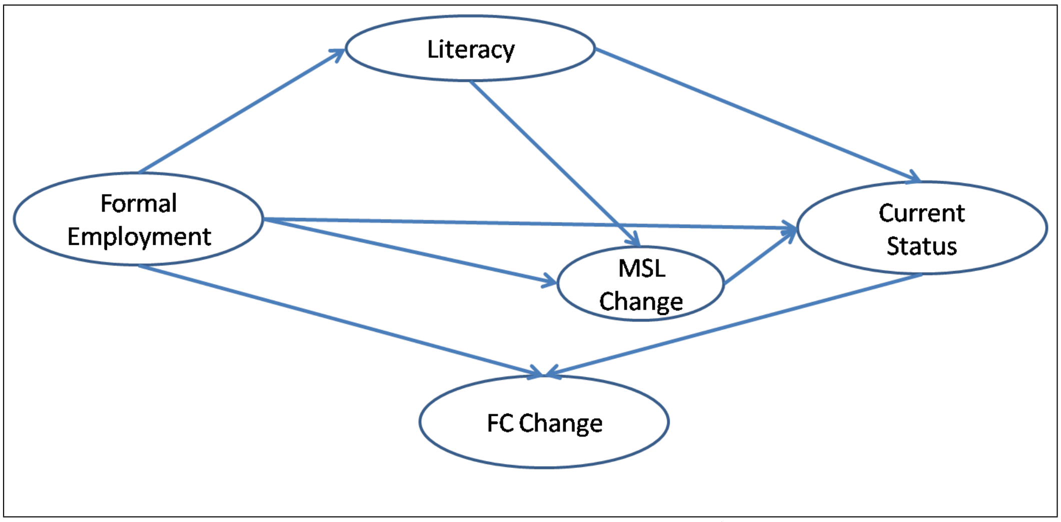

Figure 6.4: Updated Bayesian network based on Association Rule Mining to predict the change in Fuel for Cooking

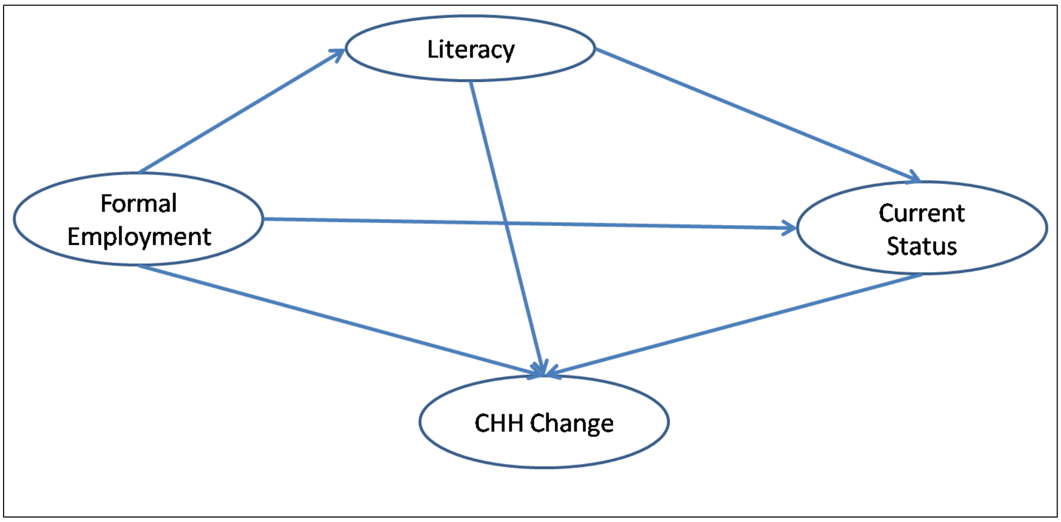

Figure 6.5: Updated Bayesian network based on Association Rule Mining to predict the change in Condition of Household

6.3.3 Prediction of change through Bayesian Network

We used the bayesian model to predict the change in four discretionary variables. We run queries for the change in respective discretionary variables while passing the other variables in the nodes as evidences. If the probability for change is higher than that of no change we label it as 1 (label change over the two census years) and vice versa if we get a higher probability for no change. We carried out a five fold cross validation and compared the results with the results of the logistic regression model explained in Section 6. The comparison of both the models is shown in Table 6.3. Except for fuel for cooking the logistic regression model seems to be giving better results.

Variable	**Log Reg Model**		**Bayesian Model**	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
Asset ownership	0.83	0.82	0.57	0.58
Bathroom facility	0.74	0.78	0.77	0.71
Fuel for cooking	0.72	0.67	0.93	0.72
Condition of household	0.8	0.76	0.68	0.56

Table 6.3: Comparison of logistic regression and bayesian models for prediction of change in discretionary variables

Chapter 7

Study of Inequality

The study of inequality in our research refers to how the socio-economic variables are distributed among groups in a population. There are multiple instances with respect to inequality here. Women consistently work less in the labor market and earn lower wages than men while economic empowerment of women is an important objective for sustained economic growth. Similarly having an equitable distribution of development in a district in terms of socio-economic parameters is important.

7.1 Method to calculate Inequality

We present here a variation of the gini coefficient way of calculating inequality in districts with respect to the development states of the villages in respective districts. We first create vectors for each district containing the labels of villages in the district. We then calculate how many pairs of villages are actually having same labels out of the total $\binom{n}{2}$ combinations where n is the number of villages in the district. We substract this number from 1 to get a score for inequality between 0 and 1. As an example consider the vector for district A with labels for 5 villages

$$DistrictA = \begin{bmatrix} 1 & 1 & 1 & 1 & 3 \end{bmatrix}$$

There are a total of $\binom{5}{2} = 10$ pairs of villages ($\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 3 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 3 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 3 \end{bmatrix}$, $\begin{bmatrix} 1 & 3 \end{bmatrix}$) out of which 6 pairs have same labels. So here the inequality score or gini coefficient would be

$$1 - (6/10) = 0.4$$

Please note that in case all labels are equal then all the pairs will have same label and the inequality score or gini coefficient would be

$$1 - (10/10) = 0$$

which means that there is no inequality among the villages in that district. This method was used to calculate the gini coefficient for each socio-economic variable.

7.2 Analysis of Inequality in Districts

After categorizing each the inequality scores as high, medium and low for each variable we checked there characteristics with respect to the level of development in each district and also with respect to the type of employment in a district.

Type	**Asset**	**BF**	**FC**	**MSL**	**MSW**	**CHH**
Non Agricultural	Low	High	High	Low	Low	Low
Agricultural	High	Moderate	Low	High	High	High
Unemployment	Moderate	Low	Moderate	Moderate	Moderate	Moderate

Table 7.1: Comparison of inequality in development level of villages in a district based on type of employment

Table 7.1 gives the comparison of inequality in development level of villages in a district based on type of employment. We can see that there is high inequality in agricultural districts for asset ownership, main source of light, main source of water and the condition of household while it is high in non-agricultural districts for bathroom facility and fuel for cooking.

Existing Level	**Asset**	**BF**	**FC**	**MSL**	**MSW**	**CHH**
Level-1	Moderate	Low	Low	Moderate	High	Moderate
Level-2	High	High	High	High	Low	High
Level-3	Low	Moderate	Moderate	low	moderate	low

Table 7.2: Comparison of inequality in development level of villages in a district based on levels of development

Table 7.2 gives the comparison of inequality in development level of villages in a district based on levels of development. We observe that there is high inequality in level-2 districts with respect to all the socio-economic variables except main source of water. This is another indication of the fact that during 2001 to 2011 India seems to have been on the rising part of the Kuznets curve of increasing inequality [18].

Chapter 8

Conclusion and Future Scope

8.1 Conclusion

We were able to make several interesting observations from analysis of the census data between 2001 and 2011. We saw a clear link between non-agricultural employment in the manufacturing and services sectors, with literacy and household spending for assets. We saw that these industrial sectors have not expanded to other geographies but have remained spatially concentrated in their pre-existing locations. We saw that households prefer to spend on assets before they spend on other amenities such as the fuel for cooking, bathroom facilities, and physical condition of their households. We saw that the government provides support on social infrastructure such as electrification and drinking water, but puts more preference on electrification. We saw that this government support however does not seem to influence discretionary spending by households. We saw that female participation in the workforce has reduced, and we are able to associate it with a reduction in marginal employment that has not been compensated with an increase in main employment as yet, irrespective of the dominant type of employment in a district. The underlying mechanisms of why these patterns emerge and what consequences they lead to, need to be investigated further.

8.2 Challenges and Future Scope

Many studies have investigated these patterns but several questions remain unanswered. This leads us to suggest that data analysis of sources like census data, including also much work happening with the use of big-data sources such as satellite data, commodity prices, and cellphone call records, can certainly reveal interesting observations. However, these observations need to be investigated further through ethnographic and survey studies to understand the dynamics that might be leading to the observations and resulting from them. We feel that newspaper reports, social media, and other participatory media networks where public conversations take place between people, may give hints about these underlying mechanisms. So far, studies have used social media and mass media data to build indicators such as for unemployment [2], economic uncertainty [4], and trade and retail [7]. However, studies have not been done to explain observations that are being noticed through a purely statistical analysis of different data sources. We feel that this can be a rich area for future research, where observations made through especially big-data sources about interesting anomalies and correlations between variables, can trigger specific analysis in qualitative data from media networks that publish information about the lives of people. Combining the mechanisms highlighted through such qualitative data analysis, with the observations made through quantitative big-data analysis, can help build a comprehensive district development model that can be explicitly evaluated on values deemed to be more important by policy makers.

8.3 ACM COMPASS' 2019

The work in this thesis was submitted as a research paper and has been selected to be included and presented at the second annual ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS 2019) to be held at Accra, Ghana from 3rd to 5th July 2019.

Appendix A

Tables for calculation of Mutual Information

The respective tables given below have been used to calculate the mutual information between the four factors of interest in Section 5.4.

	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 A.1: 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 A.2: 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 A.3: 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 A.4: Probability of (+ve Change/No Change) in Condition of household based on Type of Employment with respective variables

Bibliography

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] Dolan Antenucci, Michael Cafarella, Margaret Levenstein, Christopher Ré, and Matthew D Shapiro. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research, 2014.

[3] Sam Asher, Karan Nagpal, and Paul Novosad. The cost of distance: Geography and governance in rural india. 2017.

[4] Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016.

[5] C Chandramouli and Registrar General. Census of india, 2011. *Provisional Population Totals. New Delhi: Government of India*, 2011.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[7] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88:2–9, 2012.

[8] Planning Commission. Report of the working group on mgnrega. `https://bit.ly/2RBZitb`, 2011.

[9] Planning Commission et al. India human development report 2011: Towards social inclusion. *Institute of Applied Manpower Research, Government of India Google Scholar*, 2011.

[10] Planning Commission et al. Evaluation report on rajiv gandhi grameen vidyutikaran yojana (rggvy). 2014.

[11] Arjan De Haan. *Inclusive Growth?: Labour Migration and Poverty in India.* International Institute of Social Studies, 2011.

[12] Amit Basole et al. State of working india 2018. `https://bit.ly/2Fb4TC6`, 2018.

[13] Claudia Goldin. The u-shaped female labor force function in economic development and economic history. Technical report, National Bureau of Economic Research, 1994.

[14] J Vernon Henderson, Adam Storeygard, and David N Weil. Measuring economic growth from outer space. *American economic review*, 102(2):994–1028, 2012.

[15] Robert Jensen. The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *The quarterly journal of economics*, 122(3):879–924, 2007.

[16] Kalpana Kochhar, Utsav Kumar, Raghuram Rajan, Arvind Subramanian, and Ioannis Tokatlidis. India's pattern of development: What happened, what follows? *Journal of Monetary Economics*, 53(5):981–1019, 2006.

[17] Anirudh Krishna. *The broken ladder: The paradox and potential of India's one-billion.* Cambridge University Press, 2017.

[18] Simon Kuznets and John Thomas Murphy. *Modern economic growth: Rate, structure, and spread*, volume 2. Yale University Press New Haven, 1966.

[19] Somik Vinay Lall and Sanjoy Chakravorty. Industrial location and spatial inequality: Theory and evidence from india. *Review of Development Economics*, 9(1):47–68, 2005.

[20] Peter Ling, Steven D'Alessandro, and Hume Winzar. *Consumer Behaviour in Action.* Oxford University Press Oxford, 2015.

[21] Kristin Mammen and Christina Paxson. Women's work and economic development. *Journal of economic perspectives*, 14(4):141–164, 2000.

[22] Ministry of Drinking Water and Government of India Sanitation. Report of the working group on rural domestic water and sanitation. `https://bit.ly/2JyR7g8`, 2011.

[23] International Labour Organistaion. India wage report, wage policies for decent work and inclusive growth. `https://bit.ly/2wjhQ8O`, 2018.

[24] Sheoli Pargal and Sudeshna Ghosh Banerjee. *More power to India: The challenge of electricity distribution.* The World Bank, 2014.

[25] Hans Rosling. With ola rosling and anna rosling rönnlund. 2018. *Factfulness: Ten Reasons We're Wrong About the World And Why Things Are Better Than You Think.*

[26] Prachi Salve. Why rural women are falling out of india's workforce at faster rates than urban women. `https://bit.ly/2FeZc5d`, 2019.

[27] Sunita Sanghi, A Srija, and Shirke Shrinivas Vijay. Decline in rural female labour force participation in india: A relook into the causes. *Vikalpa*, 40(3):255–268, 2015.

[28] Richard H Thaler and Cass R Sunstein. *Nudge: Improving decisions about health, wealth and happiness.* Penguin, 2009.

[29] Jayan Jose Thomas, MP Jayesh, et al. Changes in india's rural labour market in the 2000s: Evidence from the census of india and the national sample survey. *Journal*, 6(1):81–115, 2016.