ESSAYS ON BIG DATA AND DEVELOPMENT

by

Sachin Garg

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Public Policy

Committee:

_____	Philip E. Auerswald, Chair
_____	Siona R. Listokin
_____	Aditya Johri
	T. Haque, External Reader
_____	Sita N. Slavov, Program Director
_____	Mark J. Rozell, Dean
Date: _____	Summer Semester 2017 George Mason University Fairfax, VA

Essays on Big Data and Development

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Sachin Garg
Master of Engineering
University of Allahabad, 1996
Bachelor of Engineering
Sambalpur University, 1994

Director: Philip E. Auerswald, Professor
Schar School of Policy and Government

Summer Semester 2017
George Mason University
Fairfax, VA

DEDICATION

Dedicated to the loving memory of my father
Dr. Shyam Behari Lal Garg

ACKNOWLEDGMENTS

As my doctoral journey draws to a close, it is time to reflect and sincerely thank some of the those who made this journey possible.

At the outset, I would like to thank my chair, Professor Philip E. Auerswald who steered me towards researching an area that is both impactful and contemporary, as well as one I could relate to due to past experience and training. I am also thankful to the other committee members — Professors Siona Listokin and Aditya Johri, who have been part of this journey for the greater part of three years, starting from the fields stage. My committee members helped me flesh out numerous research ideas and enabled me to focus on the specifics while not losing sight of the big picture. A special word of thanks to Dr. Tajamul Haque for agreeing to serve on my committee as an external reader and providing valuable feedback.

On the personal front, this endeavor could not have taken shape and brought to fruition without the unflinching support of my wife Kshma who gave up a beautiful and comfortable house and moved to a foreign land to support my dream. Not only did she provide moral support, edit various drafts of this dissertation but also took upon herself the responsibility of paying the bills. My daughter Sameeksha cheerfully took the moves to change four schools in five years in her stride.

My mother has been a source of constant love and encouragement, while my eldest brother, Salil has been a father figure throughout. I owe a special debt of gratitude to my elder brother, Swapnil who helped me flesh out the nuances of my research and provided constant feedback and support on the writing. My father-in-law, Late Sh. Ved Prakash Agarwal, my brother-in-law, Piyush, sisters-in-law and the large extended family have been a source of constant encouragement, love and affection.

I especially want to thank Dr. Len Nichols for his support to me by way of a research position that gave me an opportunity to play with big data. I found a very supportive faculty at the Schar School in the form of Dean Mark Rozell,

Dr. Jack Goldstone, Dr. Hilton Root, Dr. Janine Wedel, and Dr. Jonathan Gifford. Elizabeth Eck was instrumental in finding and providing financial support, while Shannon Williams ensured that I remain on track.

Among my fellow graduate students, I had the good fortune to have met Amit Patel who taught me a lot, Lokesh Dani for being a sounding board for ideas and Nabuhiko Daito & Ammar Malik for being wonderful and caring friends. I am thankful to my college friend Anupam Jaju and his wife Rishita who helped me acclimatize to the USA and supported me during my initial time in the USA.

I would like to thank Mr. Aniruddhe Mukherjee, IAS Government of Madhya Pradesh and Mr. Mukund Sinha Officer on Special Duty (Urban Transport), Ministry of Urban Development, Government of India who were instrumental in helping me collect data for my research. Mr. Pranab Choudhury, VP, NRMC introduced me to the larger community involved with land rights, while Mr. Dinesh Singh, Secretary, Department of Land Resources, Ministry of Rural Development, Government of India provided deep experiential insights that significantly improved the research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CLR Computerisation of Land Records

CMO Chief Minister Office

CoLR (MP) Commissioner, Land Records and Settlement (MP)

CSV Comma Separated Values

DGPS Differential Geographical Positioning System

DILRMP Digital India Land Records Modernisation Programme

DoLR Department of Land Resources

ETS Electronic Total Station

GCP Ground Control Point

GIS Geographic Information System

GoHR Government of Haryana

GoI Government of India

GoMP Government of Madhya Pradesh

GPS Geographical Positioning System

GSDP Gross State Domestic Product

HDI Human Development Index

HTML Hypertext Markup Language

ICT Information and Communication Technology

ILMS Integrated Land Management System

ISRO Indian Space Research Organisation

IT Information Technology

MGNREGA Mahatma Gandhi National Rural Employment Guarantee Act

MIS Management Information System

MoRD Ministry of Rural Development, Government of India

MoUD Ministry of Urban Development, Government of India

MP Madhya Pradesh

MPC Multi-Purpose Cadastre

MPCE Monthly per-Capita Consumption Expenditure

MRR Modern Record Room

NASSCOM National Association of Software and Services Companies

NGO Non-Governmental Organization

NIC National Informatics Centre

NLRMP National Land Records Modernisation Programme

NRSC National Remote Sensing Centre

NSDI National Spatial Data Infrastructure

NSDP Net State Domestic Product

OSD Officer on Special Duty

PAN Permanent Account Number

PCA Principal Component Analysis

PDF Adobe Portable Document Format

PII Personally Identifiable Information

PWD Public Works Department

RDBMS Relational Database Management System

RFP Request for Proposal

RoR Record of Rights

SLR Superintendent (Land Records)

SoI Survey of India

SRA & ULR Strengthening of Revenue Administration & Updating of Land Records

SRO Sub-Registrar Office

UAV Unmanned Aerial Vehicle

UGC University Grants Commission

UT union territory

WBLPC World Bank Land and Poverty Conference

XLS Microsoft Excel format

ABSTRACT

ESSAYS ON BIG DATA AND DEVELOPMENT

Sachin Garg, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Philip E. Auerswald

The world today is in the midst of a "data deluge". Thanks to the rise of Information and Communication Technologies (ICTs) and the mainstreaming of Machine Learning and Artificial Intelligence techniques, it is now possible to link disparate data sources and analyze this big data to gain deeper insights into human behavior than ever before. Examples abound in the field of economic development of how data sourced from mobile phone records can and are being used to identify disease patterns, socioeconomic status or identify new public transit routes. However, questions arise about the wider availability and accessibility of such data as they often form part of the core business assets of private corporations, and a source of comparative advantage. These essays on big data for development take the view that the public sector need not depend solely on the private sector for its big data needs, but should tap into its own existing data. Governments possess a wealth of administrative data, gathered through the normal process of governing. These administrative data can be linked together to create big data, which can then be used for decision making. As governments embrace

e-governance, most of the newly created administrative data will be digital (for example, the Indian government's Unique ID project (*Aadhar*) contains digital identities of more than a billion people). However, especially in the emerging economies, many times the existing legacy data have not yet been converted into a format suitable for linking to create big data. This dissertation examines related, but distinct aspects in the creation of big data and its use for development and the challenges encountered on the way, by focusing on land administration in India.

Land records are a prime example of a legacy data source. Land is economically, politically and socially important, and often the main cause of human conflict. Significant populations the world over still do not have equitable access to land, and various land reform programs have attempted to provide such access. However, these land reforms rely on the land records to accurately reflect the true situation. But, in many emerging economies, it is often the case as the land records do not mirror the ground situation, and need correction. Information and Communication Technologies (ICTs) are being harnessed to efficiently and effectively create and correct the land administration data. India started work on its land records computerization in the nineteen eighties, and the work continues.

The first essay — *Land Administration in India: A Big Data Perspective* is an exploratory study that seeks to identify the reasons behind the paradox of why a country that has both a largely agrarian society (a need for land data) and is also considered to be an Information Technology powerhouse (the means to create it) remains deficient in good quality digital land data. The study is based upon interviews with land administrators and other stakeholders in India. These include officials from both the federal government and those working for the state of Madhya Pradesh, as well as

members of civil society organizations working on land issues.

The second essay—*Diffusion of Data Policies: A Sub-national Study across India* is an empirical investigation of the state level proliferation of a Government of India program, the National Land Records Modernisation Programme (NLRMP). This program provides financial and technical support to the Indian states for their land records modernization activities. This essay applies the policy adoption / diffusion framework to a novel data set on Indian states' and districts' adoption of the NLRMP, to identify the main factors that impact adoption of data creation policies. Hypotheses based on the challenges identified by land administrators in the earlier essay are tested here to see if they are unique to Madhya Pradesh or can be generalized across the country.

The third essay in this dissertation—*Big Data Paradigm Applied to Land Administration* makes the case that as land data is big data, it should be treated as such. It uses specific land administration use-cases to demonstrate the need for a big data paradigm for land administration. It proposes a model for a flexible, adaptive and resilient land administration system built around data. Applying this big data paradigm to land administration ensures that the issues identified during the use of traditional land administration practices in the big data era are resolved. The essay emphasizes that applying the big data paradigm requires a supportive policy environment and the key elements of such an environment are identified.

The dissertation concludes by discussing, and expanding upon, the policy implications that emerge from each of the three main essays.

CHAPTER 1: BIG DATA FOR DEVELOPMENT: AN INTRODUCTION

ABSTRACT

Economic development continues to be a key concern for policy makers, and possibly will be for times to come. We are in the midst of a big data deluge, leading to a euphoria about how big data can help bring about development. Digging into this euphoria, this chapter identifies that so far, big data has been used for development on a piecemeal basis. A comprehensive usage of big data for development requires asking and answering a broader set of questions. This chapter sets the stage for this larger inquiry. It provides the context on big data for development, while developing the set of questions that need to be answered when using big data for development. It identifies land administration data as an example of administrative big data. It gives an overview of dissertation chapters — 2, 3 and 4 which answer the big data for development questions in the context of land administration big data.

I Introduction

Evidence is required for effective policy implementation and analysis so as to understand what does and does not work (Jug, 2014; Secretary-General's Independent Expert Advisory Group, 2014). This evidence is provided by survey and administrative data gathered from multiple sources[1]. These sources include various surveys like census, health, economic, or transportation to name a few, or administrative records which include land records, tax data, government documents, vehicle registration etc. Due to a lack of data "infrastructure", the emerging economies lack high quality data (Devarajan, 2013; Jerven, 2013; Jug, 2014; Round, 2014). However, with the increasing digitization of the world around us, new data sources (in the form of mobile phones and social media) are emerging, which provides a ray of hope. Data from these newly emerging multiple sources can be linked together to create big data, which can then be used for evidence based policy analysis.

We find numerous innovative examples of how this big data has been used for development. For example, data sourced from mobile phone records was combined with other data (like census data and socioeconomic indicators), and used to identify disease patterns, socioeconomic status or identify new public transit routes (Cukier & Mayer-Schöenberger, 2013; Forum, 2013, April 7; Frias-Martinez & Virseda, 2013; Kirkpatrick, 2013; Mehndiratta & Alvim, 2014, December 30; Taylor, Cowls, Schroeder, & Meyer, 2014; UN Global Pulse, 2012; UN Stats, 2013, February 22). Despite a number of such uses of these data, issues persist on how such usage can be made, especially in a replicable manner. This is because (a) many sources of these data and the algorithms used are

owned and controlled by the private sector, leading to questions about their wider accessibility and replicability (Kirkpatrick, 2013; Lazer, Kennedy, King, & Vespignani, 2014; Taylor & Broeders, 2015), (b) due to the manner in which the data have been created, these data sources have inherent biases about who is included or excluded from them (for example, not everyone is on social media, and cellular coverage continues to be uneven) (boyd & Crawford, 2012; Taylor et al., 2014; Taylor & Schroeder, 2014), (c) such data are often anonymized, and the anonymization process adds uncertainty and thus may not reflect the true population (Daries et al., 2014), (d) data needs to analyzed in specific contexts and multiple layers of pre-processing often make them lose the context (Taylor, 2014), and (e) the data are often created for purposes other than social science research and thus may lack the needed accuracy, veracity and fitness for public policy analysis purposes (Barocas, 2012; boyd & Crawford, 2012; Hilbert, 2016; Taylor, 2016b; Taylor & Broeders, 2015; Taylor et al., 2014).

However, governments are responsible for administration and carrying out development projects. During the process of governing, the government agencies gather huge amounts of "administrative data". These administrative data can be linked together to create big data, which can then be used for decision making. As jurisdictions embrace e-governance, much of the newly created administrative data will be digital[2]. However, in many instances, especially in the emerging economies, legacy data are not yet available in formats that can be linked to create big data for use in development.

This dissertation, set in the context of land administration in India, examines three related, but distinct aspects of creating big data for development from

such legacy data.

Land records are a prime example of a legacy data source. Land is economically, politically and socially important, and often the main cause of human conflict. Significant populations still do not have equitable access to land, and land reform programs have attempted to provide such access. To be successful, these land reform programs require land records. However, in many emerging economies, due to various reasons, land records either do not exist or, if available, do not mirror the ground situation. To effectively and efficiently create and update these land records, governments have turned to Information and Communication Technologies (ICTs).

Land administration data has a dynamic character which varies across both spatial and temporal dimensions. Land gets consolidated or divided and its usage and ownership changes over time. This information about its different aspects is spread across multiple agencies, each of which is tasked with a specific function (Dale & Mclaughlin, 1999; van der Molen, 2002). Hence, land data are quintessential big data, and a big data perspective is required to use it for intelligent decision making.

The three essays of this dissertation explore the data and big data aspects of land administration. We first perform an exploratory study of the land administration and land data creation processes in a large state in India. The key findings of this study are then empirically tested to see if they are generalizable. Finally, a paradigm-shifting architecture of a land management system based on the big data perspective is proposed.

Big data is a nebulous concept that is variously interpreted by different audiences. Therefore, in the next section, we define big data in the current discourse and identify its salience to human development. The dissertation structure and the salient features of each essay are then presented.

2 What is Big Data?

Increasing connectivity and digitization of the world has created new sources of exponentially growing data, resulting in a "data tsunami" (Decker, 2014). Many of these data are generated passively as human beings go about their daily lives and has been called "digital exhaust" (UN Global Pulse, 2012) or the more colorful "perspiration of the digital age" (Solove, 2004). Data from these various sources are often structured and formatted differently, resulting in data which possess volume, velocity and variety. The literature considers these to be essential traits of big data and refers to them as the "3Vs" (Volume, Velocity and Variety) (Borne, 2013; Diebold, 2012; Kitchin & McArdle, 2016).

2.1 Big Data Definitions

Big data has been defined in many "ambiguous and often contradictory" ways (Ward & Barker, 2013). Some definitions assert its size, complexity or technology dimensions (Manyika et al., 2011a; UN Global Pulse, 2012; Ward & Barker, 2013), while others emphasize its capabilities (Borne, 2013; Cukier & Mayer-Schöenberger, 2013). In the context of public affairs, Mergel, Rethemeyer, and Isett (2016) treat big data as data created through combining "structured" ad-

ministrative data with other (public or private) structured and / or unstructured data that may include "digital exhaust". All these definitions highlight that big data is created by linking together structured or unstructured data coming from varied sources (Hilbert, 2016; Taylor et al., 2014). It is this "linked aspect" of big data that we primarily concern ourselves with here. Analyzing this "linked data" using the emerging techniques of data mining and machine learning can help uncover patterns that throw new light on human society, help public policy evaluation and / or analysis and thus potentially help human development. Thus, closely linked with the notion of big data as "data", is the implicit assumption that modern analytical techniques will be used to draw inferences from the data. Thus, big data is not purely data, but rather should be considered a process that enables deployment of new analytical techniques (Hilbert, 2016; Taylor et al., 2014).

2.2 Sources of Big Data

A key characteristic of big data is that it comes in many forms and sizes, and from many sources. Some of the data sources used to create big data include social media, mobile phones, digitally mediated transactions, online news media and administrative records (Taylor et al., 2014). However, much of the literature around big data concerns itself with data sourced from social media, mobile telephony or Internet transactions (Ansolabehere & Hersh, 2012; Aragón, Kappler, Kaltenbrunner, Laniado, & Volkovich, 2013; Barocas, 2012; T. D. Cook, 2014; Crawford & Finn, 2014; Golder & Macy, 2014; González-Bailón, 2013; Kramer, Guillory, & Hancock, 2014; Lazer et al., 2014), all of which are examples of data

"born digital" (PCAST, 2014).

An often overlooked data source in this conversation around big data is the already existing "non-digital" data, that is data "born analog" (PCAST, 2014). Such "born analog" data includes voice, video and other real-world artifacts where explicit conversions are required to make them digital. It should be understood that "digitizing" such data does not make them immediately amenable to processing that can extract semantic meaning from the data. An example of such data are administrative records that are often available only in hard copy. Even when documents have been "scanned" into their digital facsimiles, they may not be usable as big data, if they lack a textual layer that is key to inferring semantic meaning. Such challenges mean that a closer look is needed at the data creation policies and practices, especially if the data is going to be "re-purposed, reprocessed, retrofitted, and reinterpreted" (Schintler & Kulkarni, 2014). We next discuss the role that big data plays and can play in human development.

3 Big Data for Development

There are multiple ways in which big data can help in development. The UN Global Pulse envisages using such data to monitor and mitigate the effects of exogenous shocks on vulnerable populations (UN Global Pulse, 2012). Big data can help in controlling the spread of disease by helping understand migration patterns (2010 Haiti earthquake and subsequent cholera outbreak (Taylor et al., 2014; Taylor & Schroeder, 2014)), analyzing the impact of socio-economic factors on cell phone usage (Frias-Martinez & Virseda, 2013), to create food

security indices (Decuyper et al., 2014, November 22), or identifying emerging macro-economic trends before traditional indicators (MIT's Billion Prices Project (Taylor & Schroeder, 2014)). It can also be used to gather information from citizens and close citizen feedback loops. Examples of these include Boston Speed Bump (PCAST, 2014) and the Indian government's "*Meri Sadak*" mobile phone application that allows users to provide geo-tagged feedback (along with pictures) on roads being built under a specified government program[3]. The application also provides feedback to the users when the issues are resolved.

Despite the promises of big data for development, significant challenges emerge. Most of these have to do with the data themselves. One of the most widely used sources of big data is social media, and cellphones; both of which are largely controlled by private entities and the data are often a source of competitive advantage. Moreover, there are significant biases in how the data are created (Taylor, 2016b; Taylor & Broeders, 2015). There also exist capacity and skills constraints, both in data creation and use (Hilbert, 2016).

These challenges of accessibility to data can be obviated by relying on administrative data. Because administrative data are produced and owned by the government, there are significantly fewer barriers to the government to access them. Moreover, the issues related to data veracity and accuracy also diminish as these are data that the government uses for its day-to-day functioning. However, even in the case of administrative data, certain questions, which are discussed next, need to be answered.

3.1 Big Data for Development Questions

Starting from first principles, three questions need answers when creating administrative big data and using it for development. These questions rest on an assumption that issues related to inter-agency information sharing, privacy and confidentiality etc., have been resolved, and that existing or future policies do not preclude linking administrative data to create big data.

Q1. Do the data exist? This question deals with the existence of the data themselves. Are the required data even being collected? For example, France and Rwanda do not collect statistics on religion and ethnicity, which may result in making "invisible" societal cleavages with regard to religion or ethnicity (Taylor et al., 2014). Also, statistics are often considered a public good, and users always demand more (Round, 2014), without understanding the costs involved in generating such statistics. Data collection has significant costs associated with it, especially for explicitly collected data like survey data[4]. However, it is expected that administrative data which is collected and generated during routine government operations should be available for developmental purposes. But, as discussed in the specific context of land records, this is not necessarily the case (section 4).

Q2. Are the data in a digital format? Digital data is a prerequisite to using big data tools and techniques. Although, it is commonly assumed that storage of data on digital media make the data digital, a distinction needs to be made between data that is "born digital" and that "born analog" (PCAST, 2014). Data

that is “born analog” needs to be transformed into a form that is amenable to digital processing. An example are the legacy public sector data, particularly administrative records that still exist as hard copies. To analyze such records, it is necessary to convert the records into a textual representation. This conversion is essentially a two step process that starts with scanning the document(s) to create its facsimile which is a digital (series of zeros and ones) representation of the image and does not necessarily capture the semantics and meaning of the document. Semantic analysis of the document to infer meaning from it requires it to have some sort of textual representation. This “textual” representation is often added as another layer and can be created using manual labor[5], Optical Character Recognition (OCR) techniques or more often a combination of both.

Though often assumed that an available scanned document contains such text layers, it may not always be true. For example, the clerk’s office of Fairfax County, Virginia, USA has computerized the land deeds and other documents. All land documents going back to the eighteenth century are available in the digital archives, albeit as simple images, without any associated textual layers. Hence, these deeds are not amenable to automated text processing. The situation of a lack of digital data is not just limited to legacy data, but the new data being created is also being uploaded as scanned images and thus will not be amenable to automated processing. Hence, policies need to be in place to enable the creation of digital data from legacy as well as new data sources. The challenge of a lack of textual layers is greatly amplified in emerging economies, especially those having a multitude of languages which use non-Roman scripts for which OCR techniques are not as well developed as for the Latin alphabet.

Q3. How can big data be used for development? For resources to be expended in the creation and linking of digital data, it is necessary to have use cases to justify expending such resources. Big data is being used for development in myriad ways (Hilbert, 2016; Taylor et al., 2014; UN Global Pulse, 2012), and the public sector can also use big data to improve its programmatic outcomes (Desouza & Jacob, 2014). Specific use cases of land big data are discussed in section 4.

We now discuss the specific context in which this study on big data for development is set, that is land administration data.

4 Land Administration Data as Big Data

We study the creation of administrative big data in the context of the data created and used for administering land.

As land is a key input to economic activity, its administration, and policies regulating its use play an important role in development (Banerjee & Iyer, 2005; Besley & Burgess, 2000; Dale, 1997; Deininger, Jin, & Nagarajan, 2009; Feder & Feeny, 1991; Feder & Nishio, 1998). The infrastructure required to implement land policies is provided by the land administration function (Williamson, 2001). A major component of the land administration infrastructure are land records, which define the roles and responsibilities of the stakeholders (Bennett, Wallace, & Williamson, 2008; Wallace & Williamson, 2006; Williamson, 2001). Given the centrality of land to society, this information needs to be freely available and accessible to all members of society, thus giving land records the character of a public good. These factors have led scholars to propose treating

land administration as a public good, and part of the national critical infrastructure (Bennett, Rajabifard, Williamson, & Wallace, 2012; Bennett, Tambuwala, Rajabifard, Wallace, & Williamson, 2013).

However, as land often has high economic value, information about it is also equally valuable, making its provision (or non-provision) lucrative and a potential source of corruption (Bussell, 2012; Goyal, 2012). Land administration is also a dynamic process with unique spatio-temporal characteristics (van der Molen, 2002). This facet adds to existing complexities. Land changes hands over time as it gets sold to multiple parties. Land also gets partitioned when family assets are divided when passed on through generations. All these events are required to be accurately recorded in the cadastres[6]. However, due to administrative inefficiencies and/or vested economic interests, many times this recordation does not occur. This missing information leads to disputes and an inability to unlock the value of real property, which poses an obstacle to economic development. (Deininger & Goyal, 2012; Narasappa & Vidyasagar, 2016; Venkataraman, 2014).

Hence, it is imperative for development that the land records accurately reflect the ground position and contain all information necessary to comprehensively manage the land. Towards this end, governments the world over have turned to ICTs and taken numerous digital initiatives to update and maintain their land administration systems(Habibullah & Ahuja, 2005; Lang, 1981; Lemmen & van Oosterom, 2001; Maggs, 1973; McCormack, 1992; Navratil & Frank, 2004).

This land administration data possesses *variety* as it comes from different sources in various forms and sizes. Some of these sources include the land registry

system, the financial system for information on mortgages etc., the court system to identify any disputes etc. Integration with a Geographic Information System (GIS) is critical because of the need to geo-reference the land parcels. It also possesses significant *volume*[7]. Owing to the dynamic nature of land administration, the data also has *velocity*.

Thus, land data possesses the three attributes, or the 3Vs associated with big data, making it *quintessential* big data. Because of the centrality of land data to development, and its inherent big data character, it is chosen to be an example of a legacy administrative big data whose creation and use is studied.

5 Structure of the Dissertation

This dissertation follows a three essay structure. Each of these essays deals with a specific aspect of land data administration.

5.1 Essay 1: Land Administration in India: A Big Data Perspective

This exploratory study seeks to understand some of the institutional challenges present in the creation of big data that can be used for development. The study uses land administration as an example source of big data. As a largely agrarian society, land administration and land reforms are extremely important to poverty alleviation in India. Information and Communication Technologies (ICTs) are being used the world over to enable effective and efficient land administration. India is also largely considered to be an ICT powerhouse in the world[8]. However,

despite the need for high quality land data for development, and the capacity to create it using ICTs, availability of high quality digital land administration data continues to be scarce. This exploratory study attempts to find why this paradox exists by taking a deep look at land administration practices in India by focusing on a national level program — the NLRMP. This program provides support to Indian states for modernizing their land administration system.

This essay is set in the central Indian state of Madhya Pradesh where the federally supported national program — the NLRMP is underway. Spread over more than three hundred thousand square kilometers, Madhya Pradesh is one of India's largest states. With a largely agrarian economy, it is home to over seventy three million people. The state's land administration system has a unique history as the state was formed by merging five different regions having very different land administration systems.

Primary data was collected by interviewing key officials involved in land administration in India as well as members of civil society organizations working on land issues. This primary data was triangulated with information from published and other records made available to us.

This study finds that data creation is impacted by myriad factors that include: (a) historical legacies, (b) the level of administrative support, (c) the existing extent of economic development, and (d) the policy design. The interviewees also highlighted the crucial role of politicians and senior bureaucrats in enabling policy adoption.

5.2 Essay 2: Diffusion of Data Policies: A Sub-national Study across India

Building on the first essay, the second essay takes a multilevel perspective. In this essay we seek to empirically study whether the insights on the issues about data creation are unique to the state of Madhya Pradesh or can be nationally generalized. Using the same context of land administration and focusing on policy adoption aimed at the creation of land records, we seek to identify the issues determining the adoption of the National Land Records Modernisation Programme (NLRMP) across the country. We also observe that while some states have taken the lead to implement the program throughout the state, others are implementing it only a few districts. These variations in the country wide proliferation of the NLRMP are put to an empirical test using the policy adoption/diffusion framework (F. S. Berry & Berry, 2014). As this is a multilevel problem, the analysis is performed at two-levels (state and district) to identify the key factors impacting policy adoption at the state level, and the selection criteria for district level implementation.

This empirical analysis required a novel dataset to be created. We created this by combining data sourced from the NLRMP Management Information System (MIS) with other indicators sourced from multiple statistical and administrative data sources. As there existed no comprehensive and composite indicator of development at the district level to capture the district level development, we also developed a composite indicator. This indicator of district development was created by statistically combining multiple indicators.

Our analysis finds strong support for the extent of development and implemen-

tation complexity at both the state and district levels impacting policy adoption. Our indicators for issue salience, or the perceived importance of the issue differ at the state and district levels. At the state level, issue salience is indicated by the legacy tenure type which impacts the granularity of land records available for land administration. The district level measures of issue salience are the size of the district's rural area and concentration of the workforce dependent on agriculture. The issue salience hypothesis finds support at the state level, but not at the district level. We also do not find support for our hypothesis that additional federal funding leads to a greater level of policy adoption.

5.3 Essay 3: Big Data Paradigm Applied to Land Administration

The third essay in the dissertation revisits the big data literature to relate our findings to it. This essay builds the case for land data as big data and identifies ways and means of how this can be done. More specifically, the essay provides a policy perspective on big data for development.

To build this perspective, this essay identifies some of the issues in land administration and demonstrates how the traditional land administration approaches fail to resolve them. Identifying that land data is quintessential big data, it argues for adoption of a big data paradigm for land administration. It develops a set of framework elements needed for a big data approach to land administration. It proposes an architecture for a *flexible*, *adaptive* and *resilient* Multi-Purpose Cadastre (MPC) based upon the big data paradigm. This vision of the MPC puts data at its core, moving all transactions to the periphery. Re-evaluation of the identified issues in the context of the big data land administration paradigm

bolster the claim for adopting this approach as these issues are found not to exist any more.

The essay further identifies the need for, and the areas where, public policy must evolve. A supportive policy environment will make this vision of a flexible, adaptive and resilient MPC, which enables an *efficient*, *effective* and *near real-time* land administration a reality.

6 Conclusion

We have briefly touched on the critical role that data plays in public policy evaluation and analysis. With the increasing digitization of the world today, new sources of data are coming online. By linking together these new data sources, we can get snapshots of human activity at extremely fine granularity that can better inform the policy process, and potentially improve programmatic outcomes and policy design (Taylor et al., 2014). However, there are significant challenges to what data can be used for this purpose, and how it can be used, considering that many times this data is either privately owned or there are strict restrictions on its re-use.

Nonetheless, the government owns a treasure trove of data in the form of "administrative records", which are created in the normal course of governing. This administrative data can be linked together to create administrative big data, which can then be used for intelligent decision making. But, linking together these administrative records is easier said than done. In many situations, the data exists, but is not in a form easily amenable to linking, for example it is still

only available in hard copy. Even when documents have been "scanned" into their digital facsimiles, they may not be usable because they lack a text layer that allows inferring semantic meaning.

The three essays in this dissertation look at various aspects of how administrative big data can be created from land records in an emerging country context. Initially, the data creation process is explored by talking to key stakeholders to gain their perspective. This is followed by empirically testing the key propositions emerging from the exploratory research. These results are then used to make the case for applying a "big data perspective" for building a flexible, adaptive and resilient Multi-Purpose Cadastre. By allowing land administration to be an efficient, effective and in near real-time, such an MPC will contribute to development.

Notes

[1]Main data sources include survey data and administrative records. See: The OECD Glossary of Statistical Terms http://stats.oecd.org/glossary/detail.asp?ID=7045

[2]An example of newly minted digital data is the Indian government's Unique ID project (Aadhar) that contains digital identities of more than a billion people.

[3]See: https://smartnet.niua.org/sites/default/files/webform/SMART%20CITY%20BROCHURE%20-%20C-DAC.pdf. Retrieved: 23 April, 2017

[4]Morten Jerven estimates a spend of $254 billion to provide data in support of the new post-2015 development targets, which is almost twice the annual spend on Official Development Assistance (Jerven, 2015). Also see Kitchin and McArdle (2016).

[5]The Indian government has created a crowd-sourcing platform "Digitize India" (https://digitizeindia.gov.in/) to help digitize government documents. Retrieved: April 23, 2017.

[6]According to the International Federation of Surveyors (FIG), a cadastre is a "parcel based, and up-to-date land information system containing a record of interests" (FIG, 1995).

[7]The magnitude of the data can be gauged from the size of the land databases of the Indian state of Madhya Pradesh is of the order of a few terabytes. This data excludes historical land records, the geospatial data and data from the deeds registries.

[8]According to the NASSCOM, the Indian Information Technology companies' association, the ICT industry contributes around 7.7% of India's GDP for FY2017. Source: http://www.nasscom.in/knowledge-center/publications/it-bpm-industry-india-2017-strategic-review. Retrieved: June 19, 2017.

CHAPTER 2: LAND ADMINISTRATION IN INDIA: A BIG DATA PERSPECTIVE

ABSTRACT

This exploratory study seeks to understand some of the institutional challenges present in the creation of big data that can be used for development. The study uses land administration as an example source of big data and focuses on a national level program — the NLRMP. This program provides support to Indian states for modernizing their land administration system.

This study is set in the central Indian state of Madhya Pradesh. Spread over more than three hundred thousand square kilometers, Madhya Pradesh is one of India's largest states. With a largely agrarian economy, it is home to over seventy three million people. The state's land administration system has a unique history as the state was formed by merging five different regions having very different land administration systems.

This mixed methods research analyzes primary data collected through unstructured interviews of various land administration officials, as well as civil society members involved with land. The primary data was triangulated with information from published data and records made available to us.

This study finds that data creation is impacted by myriad factors that include: (a) historical legacies, (b) the level of administrative support, (c) the existing extent of economic development, and (d) the policy design. The interviewees also highlighted the crucial role of politicians and senior bureaucrats in enabling policy adoption.

1 Introduction

Land is a key factor of production and thus an extremely important input to economic activity. Land plays an even greater role in societies that are largely agrarian as it supports a large fraction of the population. Because of this centrality of land to development, its administration is also crucial for development (Banerjee & Iyer, 2005; Besley & Burgess, 2000; Dale, 1997; Deininger et al., 2009; Feder & Feeny, 1991; Feder & Nishio, 1998). For example, it is land administration that facilitates the transfer of real property, and has been included as measuring "business friendliness" in the World Bank's Doing Business Index[1]. However, land administration is a dynamic process with unique institutional and spatio-temporal characteristics. Land changes hands over time as it gets sold or partitioned when family assets are divided and passed on (van der Molen, 2002). Land also has strong social and cultural connotations, especially in largely rural and agrarian societies. Further, multiple and specialized agencies are involved in performing the distinct land administration functions — juridical, regulatory, fiscal and information management (Dale & Mclaughlin, 1999). These myriad factors tend to make land administration a complex exercise (Dale & Mclaughlin, 1999; Williamson & Ting, 2001).

This complexity often, especially in emerging economies, leads to information asymmetries, for example, in the lack of information about land ownership and usage patterns. This lack of information also leads to disputes and an inability to unlock the value of real property, which again poses an obstacle to economic development. (Deininger & Goyal, 2012; Narasappa & Vidyasagar, 2016; Venkataraman, 2014). Land records contain this information and are

therefore key to providing tenure security. Security of tenure is an important factor in combating poverty and leads to economic development[2]. However, land ownership and usage varies across societies (Payne, 2001; Törhönen, 2004), which make the land administration challenges unique. The use of ICTs has been proposed as one way in which land administration can be improved by providing timely and accurate information about land parcels. This information can then be used for various purposes, especially to bring in reforms that reduce the inequities in access to land[3].

This exploratory study seeks to understand some of the institutional challenges encountered when administering land in a large emerging economy with a rich legacy of land administration, that is India. India is a largely agrarian society, making land administration and land reforms extremely important to poverty alleviation. However, a long history of land reforms has not had the desired impact (DoLR, 2009a, December 24; Habibullah & Ahuja, 2005; Mishra, 2016). Furthermore, India has had a mixed experience in using computers for land administration. Thus, we are faced with a paradox — while almost half the population depends on agriculture[4] and the country is largely considered to be an ICT powerhouse in the world[5], why is high quality digital land administration data scarce? This exploratory study attempts to find why this paradox exists by taking a deep look at the land administration practices as they pertain to the use of ICTs.

India has a federal form of government, with the state governments being largely responsible for land related matters. However, as land is such an important subject, the central government also provides certain policy directions as well

as financial and technical support to the states. Hence, this study looks at land administration in India at two levels — nationally and state level. The state that has been chosen is the central Indian state of Madhya Pradesh.

This study finds that myriad factors impact the creation of legacy data. Some of these factors are the historical development trajectory, the administrative capacity and resources available to implement the project as well as the extent of higher level political support. Design of the policy itself is also extremely important to on-ground policy implementation.

The next section briefly describes the methodology and data sources used in the study. After that, in section 3, the context of the study is set out. The national context is discussed first, followed by the state specific context. This section also briefly discusses the role of ICTs in land administration and introduces the National Land Records Modernisation Programme (NLRMP). This is followed by a description of the NLRMP implementation in section 4 on page 42. Section 5 on page 60, identifies and discusses the main findings from this evidence. We conclude in section 6.

2 Methodology

To build the deep insight into land administration necessary to explain the paradox — why a country with a largely agrarian society as well as an Information Technology world leader continues to have a scarcity of high quality digital land administration data, this study uses mixed methods. From various government documents, especially those published by the Department of Land Resources

(DoLR) of the Ministry of Rural Development, Government of India (MoRD), India[6], the NLRMP was identified as the program of interest. This program supports states in improving their land administration systems by providing financial and technical assistance for creating the underlying data. However, to understand the context in which this program is being implemented, it was found necessary to go beyond the published evidence and collect primary data on implementation of the NLRMP.

The primary data was collected by interviewing key officials involved in land administration in India, both in the central government, as well as at the state level. To get an outside perspective on the program, other stakeholders especially civil society members working on land issues were also interviewed. These unstructured interviews took place at various times and places over the course of almost two years. Details of interviewees are provided in Table 2.1 on page 26.

The selection of interviewees followed a snowball method. Key officials and stakeholders were identified based on published evidence. I reached out to certain civil servants through my network, asking them to connect me with people in certain positions. They introduced me to some career bureaucrats, who further offered to connect me with others who were more knowledgeable or had first hand knowledge of the subject.

Some of the interviewees were concerned about being identified individually as they were largely discussing their own experiences in their line of work and were unsure about how their superior officers would react. Further, some interviewees put forward their views in informal settings and it was considered prudent not to "officially" attribute comments to them. Keeping this in mind, it

was decided to hide the identity of the interviewees as is the practice in such research[7]. The conversations were hand transcribed and the transcripts verified with the interviewees. The findings from the interviews were triangulated with information from other sources, which included published sources, archival records as well as documents provided by the interviewees.

One key point that emerged from the interviews with central government officials, was the fact that the state level perspective was key to understanding land administration. Towards this end, the central Indian state of Madhya Pradesh (MP) was chosen. The choice of MP is purposive. It is one of the larger states of India[8], lies in its heartland and has traditionally been underdeveloped. It used to be a part of the "*bimaru*" (hindi for sick) quartet of states, along with the neighboring states of Bihar, Uttar Pradesh, and Rajasthan. However, in the last few years, MP has seen an uptick in its development[9]. MP also has a unique and complex land administration history and has taken up implementation of the NLRMP in earnest. Action plan and tender documents created by the state for its NLRMP implementation have been hailed as exemplary by the DoLR[10].

3 Background

To effectively analyze the data sourced from interviews and documents, a deep understanding of land administration in India is required. While chapter 1 has provided a general outline of land administration, this section discusses the specifics particular to India, especially a historical perspective on why the current land administration practices are what they are. This is then followed

Table 2.1: Interviews

Sl.	Interviewee	Position	General Job Description	Date & Location	Duration
1	A	Senior Official, MoUD	Managing all land assets of Delhi	28 Jul 2015 (New Delhi office)	60m
2	B	OSD, CMO MP	Managing Customer Service Centers that provide copies of Record of Rightss (RoRs)	20 Sep 2016 (Bhopal, MP office)	90m
3	C	Senior Official, CoLR (MP)	Responsible for NLRMP since 4 years	26 Sep, 2016 (Gwalior, MP office)	300m
4	D	NIC Official CoLR (MP)	Technical representative of NIC	26 Sep 2016 (Gwalior, MP office)	90m
5	E	Official (DoLR MoRD)	Officer in charge of NLRMP	9 Dec 2015 (Telephonic)	30m (Phone)
6	F	Advisor, , Niti Aayog (GoI)	Land Policy Specialist	Mar 2016, 28 Sep 2016 (Washington, DC (Sidelines of WBLPC, 2016) & New Delhi office)	120m
7	G	Senior civil servant (GoMP)	Managing state finances	20 Sep 2016 (Bhopal, MP office)	30m
8	H	Senior civil servant (DoLR, MoRD)	Highest ranking official looking after land records.	23/24 Mar 2017 (Washington, DC (Sidelines of WBLPC 2017))	120m
9	I	Senior civil servant (GoHR)	Town & Country planning	20 Mar 2017 (Washington, DC (Sidelines of WBLPC 2017))	30m
10	J	Executive (Indian NGO)	Stakeholder working on crowdsourced data collection in Odisha state	March 2016, 28 Sep 2016, Mar 2017 (Washington, DC (Sidelines of WBLPC 2016 & 2017) & New Delhi)	120m
11	K	Executive (US NGO)	NGO working on ground-truthing land records data	Mar 2016 ((Washington, DC (Sidelines of WBLPC 2016))	30m

by a deeper dive into the land administration system of the state of Madhya Pradesh (MP).

3.1 Historical Context

India has an extremely intricate and complex land administration system that goes back several millennia. The land administration practices varied across the country, primarily due to the presence of multiple sub-cultures and a lack of political unity. Despite the rise of several large empires, none of them actually controlled the entire country at any point of time. This led to varying land administration practices which significantly changed over time. Hence, a brief historical perspective is necessary to understand the present day situation. Scholars have divided India's land administration system in three main phase before India's independence from the British in 1947.

Ancient India to the First Millennium AD India is a continuously settled civilization for almost five millennia and thus has a long history of land administration. Although, not much is known about the land administration system before the Mughal period (that is before the sixteenth century), ancient Hindu texts do mention the obligation of the cultivator(s) to pay part of their produce to the king (Mookerjee, 1919; Wingfield, 1869). The *Arthashaastra*, a treatise of governance written by Vishnugupta Chanakya (circa 300 BC) details various aspects of land tenure and revenue systems (cf. Chanakya)). The king's portion was normally fixed to be a sixth of the produce, rising up to a quarter in terms of war (Dowson & Sheppard, 1956; Wingfield, 1869). Neale (1962) de-

scribes an elaborate system by which the produce was shared among the various village-folk.

Regarding, whether the king or the cultivator held proprietary rights in the land, Mookerjee (1919) cites evidence that the cultivator always had proprietary right in the land, a view also concurred with by other scholars (George, 1970; Neale, 1962; Wingfield, 1869). Thus, neither did India have communal ownership for land (other than the common land) (George, 1970), nor did a feudal system develop as it had in Britain during the middle ages (Dowson & Sheppard, 1956).

Muslim Period Significant changes occurred in the systems of land revenue collection with the muslim invasions starting in the eleventh century. The local population was not dispossessed of its lands, but the revenue share was enhanced and now ranged from one-third to whatever the cultivators could bear (George, 1970; Maddison, 1971). The manner of assessment and collection of the land revenue changed from a fixed proportion of the produce to a fixed assessment (Mookerjee, 1919). Also, in lieu of cash salaries, state officials started getting non-hereditary grants of villages and lands (*jagirs*), creating the *jagirdari* system (Mookerjee, 1919). Such officials were called "*jagirdars*". This period also saw the emergence of the landed aristocracy or *zamindars* as the earlier minor rulers who submitted to the conquerors were allowed to keep their lands and pay tribute (Mookerjee, 1919).

In the sixteenth century, the Mughal emperor Akbar brought in significant changes in the land revenue system. A "fixed money" rate was substituted for the customary produce share (Mookerjee, 1919). For this calculation, the land

was divided into four categories and the average yield of each category assessed by repeated trial reaping and weighing (Dowson & Sheppard, 1956; Mookerjee, 1919). The state's share was fixed one-third of the yield (with necessary deductions made for fallow lands and adverse circumstances) and converted into a cash value using the average grain price of the last nineteen years. These dues were not related to the actual crop grown (Mookerjee, 1919). This settlement was made directly with the cultivators ("*raiyats*") and did not involve any intermediaries (like *zamindars*) who may have had any right to collect the revenue. Irrespective of how the revenue was collected — directly by government officers, through the village headman or through *zamindars*; the dues from each peasant (*raiyat*) were fixed and formed the basis of all calculations, all the way up to the district and state levels (Gottschalk, 2013; Mookerjee, 1919). Thus, village accountant (*patwaris*) and their supervisors (*kanungos* or Revenue Inspectors) started playing an important role. The *patwaris* were responsible for periodically measuring the fields, ascertaining the amount of produce and accounting for the dues, while the *kanungos* supervised the *patwaris* of a group of villages (Dowson & Sheppard, 1956; Gottschalk, 2013).

Decline of Mughal Rule and Advent of English East India Company Many of the *zamindars'* who were allowed to collect the land revenue were erstwhile rulers who had hereditary claims and collected the revenue in their own right as compensation for rendering military service. Their past hereditary claims, coupled with their revenue collection authority led to them becoming extremely powerful against both the state and the cultivators. With the decline of the central government, their hereditary interests in the revenue collection started to

be recognized. They started paying a fixed sum (that bore no relation to the actual assessments) to the state. They also started to encroach upon cultivators' rights, becoming proprietors themselves, and appropriating as much as possible. However, custom prevented them from increasing the rent *per se*, and the additional amounts were added on as "cesses". This led to the *zamindar* becoming a "landlord" in his relation to the cultivator, and as a tenant when the state was concerned (Mookerjee, 1919; Neale, 1962; Wingfield, 1869).

With Aurangzeb's (the last strong Mughal emperor) death in 1707, the situation deteriorated further. His successors leased out the revenue collection to the highest bidder(s), who in turn gave contracts to others. In the process, revenue collection became disorganized, cultivators' rights disregarded and the cultivators squeezed to the maximum possible (Mookerjee, 1919; Wingfield, 1869).

When the British East India Company was granted rights to collect revenue for the provinces of Bengal, Bihar and Orissa in 1765, it found the land system to be in disarray (George, 1970). Its officials were unacquainted with the rich history of the Indian tenure system and finding the *zamindar* the most important and powerful person, they modeled him on the English landlord. Thus, the *zamindar* became the landlord and the *raiyat* (or the cultivator), a mere tenant (Mookerjee, 1919; Rothermund, 1969; Wingfield, 1869). The company's attempts to manage the land revenue across the country resulted in multiple land tenure systems, which are discussed next.

3.2 Types of Tenure

There are three main types of tenure systems that were developed during British rule. These are (a) *zamindari*, or the landlord system, (b) *raiyatwari*, or the cultivator system, and (c) *mahalwari*, or the village system.

Zamindari or the landlord system resulted from the failure of the company's earlier attempts to supervise and use the existing revenue staff to collect revenue. In 1793, Lord Cornwallis, the then Governor-General decided to permanently settle the land revenue by way of the "Permanent Settlement of Bengal". The underlying thought behind a permanent settlement was to encourage the *zamindars* to improve their lands by providing them a secure legal position with a heritable and transferable estate. This settlement used the existing valuations as the basis, without any detailed land valuations, obviating the need to survey the land. It was also felt that conducting a full survey would have been difficult and possibly elicit a distrust of the *zamindars* (Baden-Powell, 1907; Wingfield, 1869). During this process, significant changes were made to the amounts the *zamindars* could collect, and a number of the previously charged cesses were abolished. This permanent settlement had long-term sociopolitical repercussions across the country which will be discussed in section 3.3.2 on page 34.

The *Mahalwari* System As the Company learned its lessons from Bengal, it eschewed permanent settlement. While taking over new territories, the Company realized that the tenure systems were different. In 1822, the "*Mahalwari*" system

was developed for the territories that lay north and west of Bengal. In this system, an estate called a "*mahal*" was the unit of assessment. Such a *mahal* could comprise multiple villages, or there could be many *mahals* in a village. As opposed to the zamindari (or landlord) system, in the *mahalwari* system, the onus of paying the revenue could fall on a single owner, but more frequently, it was a community or a group, mostly represented by the village headman or *lambardar* with whom the settlement was entered into (Baden-Powell, 1907, 1892c). The revenue assessment was not permanent, but for a period of twenty to thirty years.

The other significant difference between this and the *zamindari* system in Bengal was that the village lands were thoroughly surveyed. A village map (*shajra*) showing each field was built and an index of fields (*khasra*) was used to identify the owners of respective fields. In case of disputes relating to boundary or ownership, the settlement officer could decide possession or refer the matter to arbitration (Baden-Powell, 1907, 1892c).

The *Raiyatwari* System This system of dealing with individual cultivators was first proposed in south India (the Madras Presidency) in the early nineteenth century, but was only widely adopted around 1855. This system was also based on accurate surveys, much like the *mahalwari* system, with one important difference. The record was not a record of the titles, but just of the land (Baden-Powell, 1907; Rothermund, 1971). In this system, the settlement officers did not bother about who owned the land, but only with who was in current possession of the same, and thus liable to pay the revenue. Thus, the taxes were payable *in rem*, rather than *in personam* (Baden-Powell, 1907; Kent, 1988). In this scheme, after

the village boundaries have been demarcated, each field is assigned a unique and fixed "survey number". The main document, which is the "settlement register" contains the details of the holdings in terms of the survey number and its occupant (*raiyat*). One field could be having multiple smaller occupants and this is all noted. From this register, a per-*raiyat* account (*chitta*) is created, which lists the details of all fields (or fraction of field) held by the *raiyat*, along with the assessment. The assessment rates consider different factors like irrigation, type of soil etc.

3.3 The Land Administration System

This section provides a brief overview of the land administration system in practice, identifying the role of the various officials and the long-term impact of the various tenure systems on land administration.

3.3.1 Land Administration Officials

In India, the district is the fundamental unit for land and related administration. The chief district officer is the district magistrate, but as this official was also responsible for the land revenue, (s)he has been called the "collector", a nomenclature that persists to this day. The districts are divided into *tehsils*[11], under the control of a *Tehsildar*, who is assisted by a deputy, the *naib-tehsildar*. Revenue inspectors, or *kanungos* report into the *tehsildar* (Baden-Powell, 1907). As an example, the modern day organization chart of the MP Revenue Department up to the district level is shown in Figure 2.1 on page 41.

Duties of the *patwari* The most important village level official is the village accountant, or the *patwari*[12]. The *patwari* performs multiple duties (Baden-Powell, 1907) which include (a) managing the village accounts, (b) serving as the official in-charge of village maps and other land records, (c) providing extracts from land records as and when needed, (d) keeping the land records updated, (e) performing inspections and provide agricultural statistics, and (f) noting all changes occurring in the land ownership (mutation). In the modern era, the *patwari'*s duties have expanded to include various activities required for provision of crop insurance and to serve as the village level interface for the *tehsil* office. Thus, we note that the patwari performs a gamut of activities that impact land administration as well as the larger policy environment.

3.3.2 Long Term Impacts of Different Types of Tenure

The different tenure types have had various long-term repercussions. For example, although land revenue had emerged as the largest single item in the Government of India's budget by the mid-nineteenth century, the permanent settlement of Bengal prevented its revision to meet changing circumstances, which led to stressed government finances (Rothermund, 1969). Further, the office of the *patwari*, was subsumed by the *zamindars,* and there were no revenue inspectors. This led to the record of rights, which had hitherto been public, becoming private. Thus, the rights in the land below the *zamindars* were not known to the revenue administration. This led to protracted disputes whose settlement was left to the courts, resulting in much litigation (Rothermund, 1971). Also, the entire administration suffered from a lack of agricultural statistics.

Similar issues existed in the *raiyatwari* areas. In these areas, the record kept were of land, and not of rights. The rationale behind this was that the government was really only interested in who was liable for paying the land revenue, and the actual owner did not matter. A consequence of this was that changes in property ownership, especially land transferred to moneylenders would continue to be registered in the farmer's name as the moneylender did not want his name in the list, and the farmer hoped to get the land back (Rothermund, 1971). This led to the creation of what is called *benami* or anonymous property.

Further, as both the *zamindari* and the *raiyatwari* systems did not have records of rights that showed ownership, these records could only be used as presumptive proof of ownership, and not conclusive. Rothermund (1971) notes that there were some proposals from the *raiyatwari* states to have the record of rights serve as conclusive proof if such a record was to be made from the normal settlement operations. However, for various reasons, these plans were shot down, and the land records continue to have only presumptive value. To provide some security, a two-fold system of deeds registration and continuous updating of the land records is followed (Rothermund, 1971).

Thus, we note that the various tenure types had their own issues related to the management of land records. These issues have severely impacted development, despite several land reforms programs. Neale (1962) points out that around the time of independence, many *patwaris* fudged records to show that *zamindars* were cultivating certain tracts of land so that they could get to keep them even after abolition of *zamindari*. The lack of land records has also impacted the implementation of reforms like the land ceiling act where the government did

not have a clear idea of the actual holdings. Further, the role of land revenue diminished in state finances, leading to a decay in the land administration mechanisms (Rothermund, 1971).

3.4 Use of Information and Communication Technologies (ICTs) in Land Administration

Land administration is seeing an increasing use of ICTs, primarily in the management of land records. This usage is not new and computers had started being for land records management fifty years ago,that is shortly after their commercial availability. This automation of land records was pioneered by the USA and Australia in the early seventies (Fiflis, 1968; Jensen, 1973; Lang, 1981; Maggs, 1973; Moyer & Fisher, 1973). By the eighties Austria, Ontario province of Canada and eight districts in India had started their projects (Habibullah & Ahuja, 2005; McCormack, 1992).

In India, the planning commission emphasized the role of well-maintained land records for administration. In 1985, the conference of state revenue ministers advocated launching a pilot project for the computerization of land records data. In 1988, the central government launched the Computerisation of Land Records (CLR) scheme across eight districts of the country. The central government provided 100% financial support to this scheme through the Ministry of Rural Development, Government of India (MoRD). The main objectives of this scheme were to (a) create a database of the basic records, (b) facilitate issuing of copies of records, (c) reduce paper work, (d) minimize the manipulation of land records, and (e) create a land management information system (Habibullah & Ahuja,

2005). The scheme was further extended in 1997 to provide RoRs to landowners *on demand.* After almost a decade of operation, it was decided to also integrate the spatial data in the form of cadastral maps and funding for thirty-two pilot projects across twenty-one states was sanctioned in 1998.

Along with the Computerisation of Land Records (CLR), another centrally sponsored scheme, the Strengthening of Revenue Administration & Updating of Land Records (SRA & ULR) scheme, was also launched in 1988. The purpose of this scheme was to help the states and union territories (UTs) in developing capacities and capabilities for land administration. This was to be done by (a) setting up and strengthening the survey and settlement organizations and revenue infrastructure, and (b) modernizing the survey and settlement operations. Thus, while the CLR was concerned with the Information Technology aspect of land administration, the SRA & ULR concerned itself with the organizational and infrastructural challenges.

In 2008, these two schemes (CLR and SRA & ULR) were merged to create the NLRMP, which has (since early 2016) been renamed as the Digital India Land Records Modernisation Programme (DILRMP). The NLRMP differs from the earlier CLR and SRA & ULR schemes in that its aim is to usher in a system of conclusive, or "Torrens" titling and provide title guarantee in India (National e-Governance Division, 2011, February 22, pp 147–152). Providing conclusive titling was never an aim of the earlier schemes, as they were aimed at strengthening the revenue administration, with conclusive titling being an afterthought. The selection of activities to perform under the schemes was left to the states, and most of the activities chosen had little to do with moving towards conclusive

titling, for example, construction of housing for revenue staff or provision of computerized offices. The activities chosen were not necessarily interconnected, and each had set a goal for itself, rather than part of a systematic process to reach the end-goal of "conclusive titling". Also, the schemes were formulated in a manner that no time-frame for the end-goal of conclusive titling could be set. Further, technology options for survey were not indicated and the schemes also excluded inter-connectivity, GIS based mapping and connectivity with financial and legal institutions. The NLRMP was launched with certain specific features to solve these shortcomings of the CLR and SRA & ULR.

3.4.1 National Land Records Modernisation Programme (NLRMP) Project

The NLRMP was conceived to usher in conclusive titling by undertaking the following set of activities:

- Completing the computerization of the Record of Rightss (RoRs),
- Digitizing the maps and integrating them with updated land records,
- Survey/re-survey using multiple technologies including Geographical Positioning System (GPS), aerial photography and remote sensing,
- Computerization of the land registration process,
- Automatic generation of mutation notices,
- Training of the revenue officials and field staff,

- Inter-connect the land records and registration offices, and

- Building modern record rooms/land records management centers at jurisdictional level.

The state governments are provided financial and technical assistance under the NLRMP by the central government to perform the following activities:

- For the computerization of land records including digitization of cadastral maps, integration of textual and spatial data, data centers at *tehsil*, Sub-division, District and State level, inter-connectivity among revenue offices, the central government provides full funding,

- For the survey/re-survey and updating the survey & settlement records (including ground control network and ground truthing) using modern technology options, the central government provides up to 90% funding in case of Special Category States and half for the other states,

- for computerization of Registration including connectivity to Sub-Registrar Offices (SROs) with revenue offices, the central government provides the Special Category States up to 90% of required funding, and a quarter to the other states),

- the central government provides 90% to funding to Special Category States to setup Modern Record Rooms/land records management centers at the *tehsil* level, while other states are provided up to half the required funding,

- training & capacity building is fully funded by the central government, and

- the Core Geographic Information System (GIS) is also fully funded by the central government.

Thus, we see that the NLRMP is a project that aims to create land records data and improve land governance. We use the NLRMP as an example project, to explore the challenges faced by the land administrators on the ground when attempting to manage the land data. This study is situated in the central Indian state of Madhya Pradesh (MP). The next sections provides a brief about the state of MP and its land administration system.

3.5 Madhya Pradesh State

Madhya Pradesh state, situated in central India is the second largest state by area and the fifth largest by population. The Madhya Pradesh Revenue Department comprises of five major departments looking after various activities. Figure 2.1 on the next page shows the organization chart of the MP Revenue Department up to the district level. The maintenance and update of land records is done by the office of the Commissioner, Land Records and Settlement (MP) (CoLR (MP)) which is based in Gwalior. The CoLR (MP) office looks after six major works for the entire state that include updating of land records, providing agricultural statistics, performing survey and settlements and implementation of land reforms and other land related policies. It is the designated agency for execution of the NLRMP project, other than for land registration (Mishra, 2016). The office has a sanctioned staff strength of almost 19,000 personnel out of which almost 12,000 are the village accountants (*patwari*) responsible for all activities related to land records management. The *patwari* reports to a

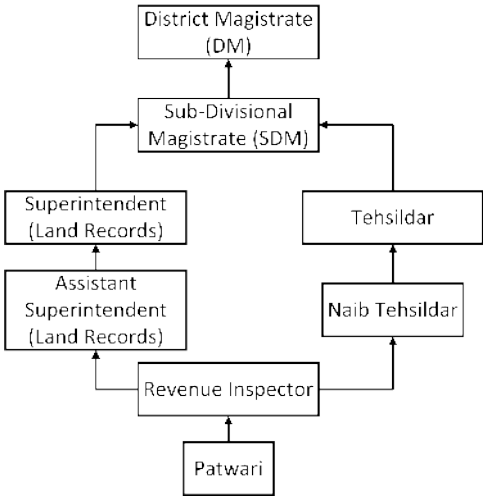


Figure 2.1: MP Government Land Administration Organization Chart

Revenue Inspector (RI) who in turn reports into the Assistant Superintendent and Superintendent Land Records who then report into the district level officials.

There is a system of checks and balances which ensure that any one individual does not have the sole ability to alter land records, but the requests are actually approved by the *Tehsildar*, who is a district officer not reporting into the CoLR (MP) office. The system has also been designed such that no one can directly alter the database itself and all such requests have to be processed through the front-end using proper authentication mechanisms.

3.5.1 Land Administration Challenges in Madhya Pradesh

The state of MP was created by merging different regions and princely states. Historically, each of these regions and princely states had their own systems of revenue administration, resulting in a multiplicity of revenue systems. In 1956, a common revenue code was enacted with a view to usher in land reforms (Mishra, 2016). These land reforms included putting an upper limit (ceiling) on the amount of land that could be owned by an individual. The idea was to provide the landless agriculturists with land by re-distributing it. However, differing measurement systems and survey chain lengths made implementing the revenue code and reform process extremely difficult. This was sought to be obviated by starting a fresh survey and settlement operation in 1975 (for 26 districts) using the metric system and prepare maps at the scale of 1:4000. However, this exercise was halted owing to local opposition and most of the cadastral records date from 1920s–30s (Mishra, 2016, Table 13, pp51). Under the aegis of the NLRMP, a fresh survey/re-survey has been initiated in 2014.

4 National Land Records Modernisation Programme (NLRMP) implementation

To understand how the NLRMP implementation has been proceeding, I initially contacted Interviewee E, senior official in the DoLR, MoRD for his views on the inter-state variations. According to him, the DoLR is only responsible for

Table 2.2: Selected Anecdotal Evidence

Serial	Evidence	Source	Core Idea
1	States are unable to/not wanting to implement. Would have been implemented if a fully central scheme.	Interviewee E	Interstate Variations
2	Interstate variations are due to state legacy. Example: (a) Haryana able to do it due to regular plot shapes, which are because of recent consolidation exercise, while states like UP have irregular plots due to multiple subdivisions; (b) Northeastern states have different tenure mechanisms.	Interviewee E	Institutional History
3	Program specifications are unable to take care of inter-state variations and thus they have to be malleable	Interviewee F	Rigid Specifications
4	Lohardaga (Jharkhand) effort to integrate textual and spatial data to provide an online cadastre was abandoned as it was not a top priority.	Interviewee B	Bureaucratic/ Political Support
5	Property records both in revenue offices and sub-registrar offices (conveyance deeds) generally agree on the physical boundaries — there is no dispute regarding the same in modern records. However, a lack of GIS based systems is an issue. This makes urban property records more challenging due to lack of convergence between the rural records (based on *khasra* /*khatouni* /*khata*) and the urban records (based on plot numbers).	Interviewees A & H	Institutional History

setting the specifications and releasing funds, while the actual implementation is the job of the states. He was of the view that had this been a project being implemented wholly by the central government, there would lesser hurdles (point 1 in Table 2.2 on the preceding page). Sud (2014) also notes a similar viewpoint of DoLR officials. It is also possible that not all states are equally enthused about the project or are willing to provide funds for the same. For example,there are huge variations in the per-unit costs of the survey/re-survey activity between the states[13].

The official also pointed out that some states are able to proceed faster, especially in the survey / re-survey activities owing to their fields being of a regular shape and size (point 2 in Table 2.2 on the previous page), for example the states of Punjab and Haryana. This aspect points to a legacy issue as concurred by Interviewee I who hails from the state of Haryana. He told me that this regularity in fields is due to the fact that land consolidation happened much later (in the 1960s) in the states of Punjab and Haryana.

Another interviewee (F) pointed out that the specifications are extremely rigid and not enough leeway is given to the states to take care of their local circumstances (point 3 in Table 2.2 on the preceding page), a view also concurred with by other interviewees (B & C) in the state of MP. Interviewee E pointed to Lohardaga in Jharkhand state (item 4 in Table 2.2 on the previous page) as an example for the need of widespread political support in project implementation. In Lohardaga, a pilot project for an integrated land information system had been started but it has now been abandoned. Interviewees A and H mentioned that the challenges in creating urban records stem primarily from a lack of conver-

gence between the rural and urban identifiers (item 5 in Table 2.2 on page 43). Interviewee H, the top ranking official in the DoLR, MoRD provided a detailed exposition on the linkages between the land records and land registry systems. He was of the view that most of the time the land owners and land records generally agree on the broad boundaries and boundary disputes are rare. Most land disputes relate either to transactions or inheritances/partitioning etc. that have not been properly recorded.

Based on these aspects, a deeper dive was done regarding the project's implementation and land administration practices in the state of MP. The following sub-sections provide a detailed account of this as gleaned from the interviews and triangulated through analyzing the primary and secondary data sources. Each of the following sub-sections analyzes one key part of the NLRMP in terms of its goals and the challenges faced in MP.

4.1 Core Geographic Information System (GIS)

The core GIS is the foundation underlying modern land administration (Williamson, Enemark, Wallace, & Rajabifard, 2010) that allows the integration and provision of multiple services. It consists of geo-referenced satellite imagery of the village index base maps and integrates three layers of data, viz (a) cadastral maps from revenue records, (b) spatial data from aerial photograph or high-resolution satellite imagery, and (c) Survey of India and Forest Survey of India maps (DoLR, 2009b, April 17, 2008, August 21). However, for this core GIS to be made available, the underlying data needs to be created. Various activities of the NLRMP contribute to creation of the core GIS. The "computerization of

land records" activity provide support to creating the cadastral maps from revenue records (Habibullah & Ahuja, 2005, Chapter 14) while the survey/re-survey exercise results in the creation of geo-referenced satellite imagery.

4.2 Computerization of Land Records

This activity has four main components: (a) data entry aspects, (b) digitization of cadastral maps and integration of textual and spatial data, (c) creation of data centers at district, sub-division and tehsil level, and (d) providing interconnectivity among the revenue offices. In the state of MP, a sum of almost one hundred and twenty million Indian Rupees (₹) was spent on the digitization of the Record of Rights (RoR) (Interviewee C). According to my source (Interviewee C) this was done without any attempts to perform a verification of the details, that is get to the ground-truth, and he felt it was a waste of money. This lack of ground-truthing is not unique to the state of MP, but a pervasive problem in most states as concurred with by Interviewee K who leads an international organization working worldwide on land rights. The major issue with the lack of ground-truthing is that since the data does not match the ground reality, the system faces the problem of "garbage in, garbage out", a term Interviewee I used in relation to the much-lauded *Bhoomi* project implemented in the southern state of Karnataka (Chawla & Bhatnagar, 2004; Habibullah & Ahuja, 2005). Ground truthing requires ears on the ground, and Interviewee K's organization is performing a ground-truthing exercise of land records in Warangal district (in the southern state of Telangana) by going door to door to check the veracity of the RoRs.

Interviewee C was of the opinion that in MP the initial computerization of land records took five years (completed only in 2013) instead of the stipulated eighteen months due to the initial lack of planning and delays in the digitization of cadastral maps. These were due to an underestimation in both the cost of digitization, as well as the volume of work. The initial cost estimates provided by the DoLR in the NLRMP guidelines were off by ten to twenty percent. While the program guidelines provided an estimate for digitizing an A3-sized record of ₹ 1,060 (DoLR, 2009b, April 17), in reality it ranged from ₹ 1,160 to ₹ 1,260. Further, the scanning of the maps sheets has been a huge exercise—almost 134,000 map-sheets spread across 53,480 villages.

In addition to these flaws in policy design, he highlighted two other reasons for the delay in digitization of the cadastral maps: (a) bureaucratic apathy, and (b) vendor incompetence.

4.2.1 Bureaucratic Apathy

The digitization of maps in MP had been outsourced to multiple vendors as allowed by the program guidelines. My source (Interviewee C) told me that during this digitization process, some of the vendors complained about a lack of co-operation from district level officials, especially the Superintendent (Land Records) (SLR). The officials were either withholding or delaying providing the maps to be digitized to the vendors. To get around this, the progress of the records digitization was tied to the performance appraisal of the district's Superintendent (Land Records).

Further, some of the village maps had been listed as either "missing" (two hundred and ten) or in a dilapidated condition (five hundred and eighty-one). These maps are stored at three different places: (a) *tehsildar* office, (b) *patwari* office, and (c) record room. Apart from this, the irrigation department also has certain village maps available. Management of village maps has a long history in the Indian land administration system (Baden-Powell, 1907) and the possibility of maps disappearing from all three places, resulting in "mapless villages" is remote. Considering this, the Commissioner, Land Records and Settlement (MP) suspected bureaucratic apathy and vested interests to be a major reason behind this non-availability of maps. To investigate whether the maps were really missing, or the officials were simply being lax in their work, a police complaint was filed against the custodians of maps, that is the Record Room In-charge, the *Tehsildar* and the *Patwari*. Due to this ingenious solution, one hundred and forty village maps were found, which reduced the number of "mapless" villages to seventy. For these seventy cases, the maps had gone missing over a period of time or were in extremely poor condition. In some of the cases, it was attempted to use land records maps available with the irrigation department to fill the void. For the villages still missing maps, a survey and settlement which requires going door to door to gather the actual ground position and tallying with the previous settlement was ordered.

4.2.2 Vendor Incompetence

The digitization of maps was a new and fresh exercise in MP and not all five vendors were equally competent. While the program provided guidelines for

the process and the National Informatics Centre (NIC) also provided technical assistance, two vendors were incompetent and this led to the digitization work for some of the districts being hampered. According to my source (Interviewee C), the soft copy provided by one vendor (who had been paid eighty percent of the amount upfront) was missing metadata. Because of this missing metadata, a "mosaic" tehsil map could not be created by combining the digital maps of the various villages, which was requirement of the NLRMP (DoLR, 2009b, April 17).

To build capacity, the program guidelines stipulate that vendors provide training to the revenue officials in use of the software. In the case of MP, the contract stipulated that the vendors train five revenue personnel (Interviewee C), who would train further staff and so on. However, many vendors did not fulfill this requirement which has resulted in a lack of trained resources. To fix this, my source told me that a proposal is underway to float a short term tender and get the requisite training provided.

To the question as to whether the CoLR (MP) office found any lack of clarity in the specifications or funds from the central government to be impediments in implementing the NLRMP, interviewee C answered in the negative. This was also concurred with by other sources (interviewees F and H).

The state of MP had already setup its own data centers and built a wide area network to interconnect the revenue offices, and therefore did not require any funds for the same as documented in the project proposals and concurred by my source (Interviewee C).

4.3 Survey/Re-survey and Updating of the Survey & Settlement Records

The aim of the NLRMP is to usher in "Torrens" titling in India (DoLR, 2008, August 21; National e-Governance Division, 2011, February 22, pp 147–152). This registration system is founded on the "mirror principle"—that is the maps have to mirror the ground realities so as to not require going through the chain of documents conveying title (McCormack, 1992; Zasloff, 2011). Regular surveying of the land is a means to ensuring that the maps accurately reflect the ground position and this is critical to the development of an integrated land management system that can be utilized to get real-time information on the land (Williamson et al., 2010). As discussed by (Mishra, 2016) and corroborated by talking to multiple people (Interviewees B, H, I and J), the last large-scale surveys were done before Indian gained independence (1947). In MP, owing to the presence of multiple map scales and areas of measurement, the first state-wide survey operation was initiated only in 1975. However, even this survey was stopped after covering only seventeen of the fifty one districts, citing objection from the local population (Mishra, 2016). Thus, almost two-thirds of the state is still working with maps that are almost a century old.

Since then, the population has increased manifold, leading to the village residential area (*aabaadi kshetra*) also increasing and changes in other artifacts like canals and roads. Thus, the extant cadastral maps do not match up when recent satellite imagery is superimposed on them. Also, many times, especially near large urban centers, rural land has been converted to urban use, and records of *Nazul* land (land acquired by local developmental authorities and provided on lease) are largely unavailable, leading to widespread encroachment of govern-

ment land (interviewees A, C, F, H, I and J). The current map scales are 1:4000, which although adequate for rural land management is grossly inadequate for dense urban areas that requires maps to a finer scale of 1:1000 or less like 1:500.

MP has been at the forefront of implementing modern survey practices using Electronic Total Station (ETS) and the machines have been made available to all districts and *tehsils* from state funds (Mishra, 2016). The revenue inspectors are trained in the usage of ETS only, while the *patwaris* are also trained in surveying using the traditional chain method. As of September 2016, the survey/re-survey exercise has been completed in twenty districts (Interviewee C).

The central government provides half of the funding for the survey/re-survey activity. The program allows for either a pure ground method (using Electronic Total Station (ETS), Differential Geographical Positioning System (DGPS)), or a "hybrid" approach that combines the ground method with aerial imaging and satellite remote sensing (DoLR, 2009b, April 17). MP has chosen the satellite imagery based hybrid approach. In this approach, processed satellite imagery is used to capture the land parcel data and this is compared with the existing land records to demarcate the land parcels. A sample is cross-checked with ground surveys using ETS to verify the accuracy of the satellite imagery and existing records.

According to the program guidelines, after the parcel data has been verified and the maps updated, the vendors update the land records data with additional personal information of the owner[14] and the other details needed to prepare the RoRs. These draft land parcel maps are first delivered to the administration for checking and then to the landowners for further verification and comments.

There are public interactions between the landowners and the administration which also involves the vendors to refine the database as well as to resolve issues as far as possible. As per law and the NLRMP guidelines, if the owner(s) do not raise objections to the new survey boundaries, the area and other details recorded in the RoR, this record is finalized. However, if the owner(s) finds mismatches between the new and old data it is marked as disputed (DoLR, 2009b, April 17). In MP, Interviewee C told me that a solution to the disputes has to be suggested by the administration within three days of hearing all the parties. If this is accepted, then the dispute is marked as "closed", else it is kept pending and moves up the administrative chain. In order to make this hybrid approach of survey/re-survey work, two items: Ground Control Points (GCPs) and satellite imagery are crucial.

4.3.1 Establishment of Ground Control Points (GCPs)

A Ground Control Point (GCP) is a point on the ground that has well-known co-ordinates, usually based on the Geographical Positioning System (GPS). A number of such points are required to correlate the satellite imagery and actual ground positions. The NLRMP guidelines propose three types of GCPs to be setup (DoLR, 2009b, April 17). The entire state of MP has been divided into three grids (primary, secondary and tertiary) of different dimensions, which are (a) 16 km × 16 km grid for the Primary GCPs, (b) 4 km × 4 km grid for the Secondary GCPs, and (c) 1 km × 1 km grid for the Tertiary GCPs. The program guidelines specify that each GCP be conspicuously marked ("monumented") such that it is clearly and unambiguously visible in the satellite imagery. The

guidelines mandate that the GCPs be calibrated using GPS and specifies the procedure (DoLR, 2009b, April 17). While the primary GCP is calibrated over a 72-hour continuous observation (to a ten digit precision), the secondary and tertiary GCPs are observed for 3 hours and forty-five minutes respectively.

Interviewee C told me that in some districts, the vendors were unwilling to go the field owing to the poor law and order situation. He recounted an anecdote about GPS equipment worth around US$70,000 being stolen in Morena district while a primary control point was being established. However, he emphasized that despite these issues, the work of setting up the GCPs has been completed in all the fifty-one districts.

4.3.2 Satellite Imagery

For survey/re-survey using satellites, and to adequately build the cadastral maps, it is necessary to get specific high resolution stereo imagery. Such images should not have any clouds, norshould the fields be sown. Further, the images have to be of recent vintage (after setup of GCPs) to ensure proper geo-referencing. This means that every district would have a specific and differing time window[15]. The NLRMP guidelines propose that states utilize the services of the National Remote Sensing Centre (NRSC) of Indian Space Research Organisation (ISRO), which is the nodal agency in India for providing satellite imagery (DoLR, 2009b, April 17). However, according to Interviewee C, while the CoLR (MP) wanted the NRSC to play the role of a full-fledged remote sensing consultant, it was serving primarily as an agent selling satellite imagery from various sources. Thus, the entire burden of negotiations (including technical specifications) was

done by officials of the CoLR (MP) itself. As per Interviewee C, the NRSC quoted an initial price of US$ 43 per sq km,suggesting that apart from the imagery, one additional item (not included in the Request for Proposal (RFP)) was also being provided free of charge. But, when the CoLR (MP) officials started the negotiation process, it was found that the extra item (that was not really needed) was actually being provided at a cost and the CoLR (MP) officials negotiated the price down to US$30 per sq km. These negotiations resulted in estimated savings of nearly seventy million Indian Rupees for the more than three hundred thousand sq km of the state for which imagery was to be bought. The satellite imagery was initially procured for eighteen districts, but in 2016 a further seventeen districts have been added.

The survey exercise in MP differs from that in many other states in that the administration is also attempting to create detailed maps of village residential areas (*aabaadi kshetra*) and urban areas, apart from the farmland. However, maps of urban areas are required to be at a much greater scale (1:1000 or 1:500), which leads to different issues. Adding to this is the need to accommodate multi-storied buildings, multi-owner apartment complexes or commercial spaces. Interviewee C told me about a pilot project underway in Dabra *tehsil*, Gwalior district for urban survey. For this project there is a thinking around the use of Unmanned Aerial Vehicles (UAVs) or "drones" for the surveying and mapping activities. However, there is a lack of policy clarity in the use of UAVs as indicated by Interviewee Iand corroborated by Interviewee H. According to Interviewee H, his department is working with other agencies to develop appropriate policies that can allow the usage of UAVs for large scale and cost-effective mapping activities.

4.4 Computerization of Registration

India has a deeds registration system and a "presumptive titling" scheme (Dowson & Sheppard, 1956; Rothermund, 1971). In such a system, the legal cadastre is maintained by registering the property deeds. A registered document becomes part of the public record and is used to show priority. In the "presumptive titling" scheme, a chain of registered documents is used to show the conveyance of property from one person to another. Therefore, it is critical to integrate the registration process with management of the RoR. When a property transfer occurs, the entries in the RoR are changed to indicate the new owner(s) through a process known as "mutation" (Interviewees H and J). The NLRMP provides support to certain activities to integrate the registration process and RoR maintenance. These activities include (a) computerization of the Sub-Registrar Offices (SROs), (b) data entry of details of property valuation, (c) scanning and preservation of old documents, and (d) providing inter-connectivity among the SROs. According to my sources (Interviewees C and D), MP started the computerization of registration without central funding support and using state funds. The property valuation is updated every year and the process of entering legacy encumbrance data is in process. Interviewee D told me that the registration department in MP consists of more than two hundred Sub-Registrar Offices (SROs) and e-registration facility has been made available in all districts. Using this facility, users may complete the initial registration formalities from anywhere and book an appointment to complete the process which requires capture of biometrics of all transacting parties. However, the registration system is still not integrated with the RoRs and thus online mutation is not possible (Mishra, 2016).

4.5 Modern Record Rooms

A key task of a land management system is document management. It is necessary to keep the old and legacy records in physical form for both historical and legal reasons[16]. The NLRMP has provided funding to upgrade the existing *tehsil* level record rooms to Modern Record Rooms (MRRs) that will both preserve the existing documents, as well as allow the documents to be retrieved electronically on an as needed basis. The NLRMP guidelines specify how documents will be managed (DoLR, 2009b, April 17). Each document is cataloged using barcodes and scanned into the system with the necessary metadata. The scanned map sheets have to be geo-referenced using the existing Geological Survey of India reference points. The originals are laminated and kept in climate controlled record rooms with compact shelving to minimize space requirements. Creation of the MRRs involves both civil as well as Information Technology (IT) activities. However, the NLRMP guidelines forbid any new construction and the states' (through the district administration) have to provide a room no larger than 1200 square feet (DoLR, 2009b, April 17). The Modern Record Room is created by renovating this space at the maximum allowable rate of ₹ 288 per sq ft. This renovation includes provisioning of false ceiling, air conditioning, IT systems (computers, scanners and printers), fire suppression systems and a heavy duty vault door. Based on these figures, and the assumption that there are an average of sixty-six thousand revenue records per *tehsil,* the program provides full funding for the MRR at the rate oftwo and a half million Indian Rupees per *tehsil*. However, the guidelines state that maintenance of theMRR has to be done through the regular funding available to the government's Public Works Department (PWD) (DoLR, 2009b, April 17).

According to Interviewee C, the specifications of the door make it extremely heavy (unsuitable for many walls), and very expensive (₹ 90,000 or up to one-fifth of the total budget). Further, the room size is strictly to be adhered to and no payment was to be made for anything exceeding 1200 sq ft in a *tehsil*.

The guidelines stipulate that document scanning start after the room has been setup with all the equipment. In MP, the Superintendent (Land Records) (SLR) was responsible for providing the records to be scanned to the vendors on time. Interviewer C told me that many times the SLRs did not provide documents on time leading to missed deadlines. Another challenge was the gross underestimation of the number of records. While the budget for the MRR was estimated based on there being around sixty-six thousand revenue records per *tehsil*, the actual number ranged from a minimum of one hundred and fifty thousand up to one and a quarter million per *tehsil*, which meant the potential of enormous cost overruns.

Interviewee C recounted that midway through the document scanning (in 2013–14), the state's Chief Secretary decided that not only revenue records, but also other non-revenue district records be digitized. For this purpose, additional funding to the tune of 190 million Indian Rupees was provided by the state government. However, due to the initial wrong estimates, the revenue records themselves were not being fully digitized, and the additional funds lapsed. Further, although the NLRMP provided funds only for twenty seven districts, the MP government created MRRs for the entire state by re-appropriating from other heads like data centers and connectivity. My source also indicated that they were able to keep cost escalations under control by judicious vendor management and building

personal relationships with the vendors.

4.6 Training & Capacity Building

Training and capacity building is extremely important to ensure that the project is successful in the long run. As discussed in section 4.2.2 on page 48, a mechanism of trainees training others had been designed to ensure that all officials were well-versed in the use of the software. However, at the ground level, the key functionary is the *patwari*, who as discussed in section 3.3.1 on page 33 performs a number of duties and is key to effective and efficient land administration.

In MP, my source (Interviewee C) told me that the jurisdictions of the *patwari* (called *halka*) and the village local self-government (*panchayat*) were different. Till around 2014, MP had 11,622 *patwari halke* spread over twenty-three thousand *panchayat*s covering more than fifty thousand villages. Thus, each patwari handled around two panchayats, covering around four villages. The *panchayat*s were often created due to various political reasons and some of them were at considerable distance from the each other and/or the *patwari* office. As the patwari had to be present at the different *panchayat* offices throughout the week, this led to significant challenges. Further, there was no differentiation between the large and small *panchayat*s leading to an uneven workload on the *patwaris*.

Considering all of this, it was decided to align the *patwari*'s jurisdiction (*halka*) with the *panchayat* based on four criteria: (a) *panchayat* should not be subdivided to create the *halka*, (b) a large *panchayat* can have two *halka*, while small *panchayat*s can be combined to form a *halka*, (c) a *halka* should have

between 2500 to 4000 land records, and (d) the distance between the *panchayat* headquarters and the *halka* should not be more than two kilometers (one and a half miles). This alignment of the *patwari*'s jurisdiction with the village *panchayat* led to the proposal to create an additional 7,398 patwari positions. Adding these to the backlog of the 1,700 *patwari* positions already lying vacant means that MP is short by more than 9,000 *patwari*s. This was highlighted in the report presented to the Chief Minister, a copy of which was provided to me.

This huge number of vacancies at the *patwari* level is going to present significant training and capacity building issues. It has been recognized that continuous training and capacity building exercises are key to ensuring long term success of the NLRMP. The program has provisions to build NLRMP Cells that provide training and continuously update the skills of the staff. Towards this end two such cells have been setup in the state. These cells augment the training provided by the nine *patwari* training centers, two revenue inspector training schools and one state training institute.

Enhancing the entry level qualifications will also help in mitigating the skills gap. Earlier, the *patwari* only needed to have completed twelve years of schooling. But now, Interviewee C told me that the minimum qualification of a *patwari* has been enhanced to be a three year diploma in computer applications recognized by the University Grants Commission (UGC). Apart from training on administrative aspects, the *patwaris* are also trained in surveying using both the traditional chain survey and also the use of ETS. Interviewee C filled me on the intricacies of land administration at the ground level. He told me that after five years, the *patwari* is eligible for promotion as a "revenue inspector". Every year, half of

the revenue inspector vacancies are filled by fresh civil engineering graduates, and the remainder by promoted *patwaris*.The eligible *patwaris* have to take a pre-examination. Selected candidates then undergo a nine month revenue inspector training after which they take the revenue inspector examination. This procedure ensures that there is a healthy mix of fresh and experienced personnel at the revenue inspector level and that the *patwaris* have a clear career path.

5 Discussion

5.1 Advantages of Computerized Land Records

Interviewee H told me that computerization of land records has improved the land administration system in many ways across the country. In MP it has reduced the drudgery of manual paper work and increased the efficiency of the CoLR (MP) staff as identified by interviewees B,C and D. Another aspect much appreciated by the CoLR (MP) officials has been its capability to reduce fraud. Interviewee C recounted two examples by which the system is helping to detect and prevent fraud.

Land transfer fraud was perpetrated by a *patwari* in Gwalior district of MP. As custodians of land records, the *patwaris* generate the mutation requests required for land to be transferred. However, as discussed in section 3.5 on page 40, the *tehsildar* is the official responsible for approving the requests.

The officials access their respective role of generation and approval of mutation

notices by logging into the system using their unique credentials. However, in Gwalior, where the fraud happened, the *tehsildar* was not technology-savvy and had shared his credentials with the *patwari*s. One of the *patwari*s took advantage of the situation and transferred 90% of the land belonging to a government trust to him[17]. Next, using the *tehsildar*'s credentials, he immediately approved the mutation request.

The fraud was detected due to multiple anomalies. Firstly, the transaction happened in the evening hours, and secondly, the time difference between initiation and approval of the mutation request was much less than normal. In a non-computerized system, it would have been extremely difficult to detect such a fraud.

Prevention of multiple sanctions for the same project. Many "*zilla panchayat*" (district level self-government) officials used to get funds for the *same* civil works from *different* funding sources. For example, grants are sought simultaneously from various departments—rural development, irrigation, and agricultural for digging the same tube-well. This fraudulent practice was stopped by the Panchayat and Social Justice Department of MP by linking sanction orders and grants with the GIS and maintaining geo-referenced records, ensuring that each project was identified in space.

5.2 Lessons Learned from NLRMP Implementation

Autonomous society to route funds. Interviewee C told me that MP has created an autonomous, empowered committee that is registered as a society (under the Societies Act). This society is headed by the Chief Secretary and has a fourteen member executive committee headed by Principal Secretary (Revenue). All funds for the NLRMP implementation are routed through this society, which is the final decision maker obviating the need for any further sanctions. The presence of this committee allows funds to be seamlessly re-appropriated as needed, for example the data-centers had already been created using state level funds and thus available funds could be used for other activities, for example creation of MRRs.

Legacy matters in implementation of the NLRMP. The impact of history and legacy in MP has been most felt in the survey/re-survey exercise. As pointed out in section 3.5.1, the state had multiple tenure systems before 1956, and most land surveys are almost a hundred years old. These issues have resulted in the cadastral records being old and having differing scales and thus lacking accuracy.

Another challenge has been the ambivalence[18] of the central government in asking the states to necessarily align their new surveys to the existing national geographical grid of the Survey of India[19]. States are at liberty to co-ordinate with the Survey of India for setting up of GCPs which may result in duplicated efforts or worse, non-alignment with the national spatial infrastructure.

Development matters in project implementation. Those districts of MP that are less developed lack public support for the various NLRMP activities (Interviewee C). Incidentally, some of these districts are in an area that has historically been crime prone areas, leading to issues like robbery of the GPS equipment (section 4.3.1 on page 52) which leads to further project delay.

Administrative support at all levels is crucial for the project to succeed. As discussed earlier, the *patwaris* have a key role to play in the entire land administration process. However, the *patwari* wears many hats and plays a variety of roles. Among the land administration fraternity, it is common knowledge that many times the *patwaris* do not carry out physical verification (Interviewees B, C, F, H and I). This lack of physical verification leads to huge delays in records update as well as the possibility of disputes and fraud. Another example for the need of administrative support was the lack of co-operation from district level officials in providing maps for digitization (section 4.2.1 on page 47) which led to cost and time overruns in MP. This particular challenge was overcome by linking progress in land records computerization to the officials' annual performance appraisal. Similarly, to fix the problem of missing maps, the route of filing a police complaint and forcing a police investigation resulted in many maps becoming available. The interviewees also identified that some times the core activities got derailed because of conflicting priorities. An example of such "mission creep" is found in the decision to also digitize non-revenue records, in parallel, without completing digitizing of the revenue records (section 4.5 on page 56). The aspect of inter-agency co-operation has been pointed out in the case of satellite imagery, where the central government agency was behaving more as a vendor than as

a partner (section 4.3.2 on page 53). Coupled to these challenges has been the sheer size of the endeavor, which was much under-estimated, for example in the case of MRR (section 4.5 on page 56). As already discussed, the large number of staff vacancies also contributes to delays in project implementation. It is also possible that some states may not be able to provide funds to maintain the MRR in the long run.

Rigid project specifications created centrally have also been pointed to for leading to implementation issues. The case of MRR in section 4.5 on page 56 highlights the issues with strict, "one size fits all" specifications and wrong estimates. The door of the record room was specified so that it took up over a fifth of the budget. The size of the MRR was strictly specified and no new construction allowed. This meant that any additional funds had to be diverted from different heads. Further, the number of records was grossly underestimated which led to significant budget overruns. While officials of the DoLR, MoRD were not willing to accept that the specifications were rigid, other interviewees (C, F and I) agreed that states need to have certain leeway in implementation.

6 Conclusion

Effective land administration has been identified to be crucial for development. Given that India is largely an agrarian society, the reduction of inequities of access to land make its administration even more important. Land administration revolves around land records, making proper land record maintenance key to effective land administration. Land administration in India has developed

over the course of millennia and has undergone significant changes through the ages, making it extremely complex. These complexities of land administration are reflected in the manner in which land records are managed.

Starting in the nineteen eighties, the Government of India, in line with efforts the world over, started pilot projects to computerize the land records system. Land administration in India is under the purview of the state governments and in many states these pilot projects have morphed into full-blown implementations. However, significant challenges remain in land records management, especially those related to adjudicating the rights and resolving disputes. Towards this end, in 2008, the central government decided that the ultimate aim of the land records modernization programs should be to aim for conclusive titling that would not only clarify the titles, but also provide title guarantee. The National Land Records Modernisation Programme (NLRMP) was launched that had a number of activities that would ultimately lead to conclusive titling. However, there are significant variations in the adoption of the NLRMP and the availability of high quality digital land data. This study set out to understand why this is the case for a country where land is so important and which is largely considered to be an Information Technology powerhouse.

To unravel this paradox and considering the complexity of land administration, it was imperative to talk to the concerned officials and stakeholders to understand the various challenges and how these challenges are mitigated in the course of program implementation. I initially talked to officials in the Ministry of Rural Development, Government of India (MoRD) who manage the scheme. Their main refrain was that they only lay down the broad parameters and provide

funds, while the ground level implementation is managed by the states.

At the state level, I selected the state of MP in central India. Traditionally, an underdeveloped state, MP has grown significantly on the back of its strides in agriculture and has also taken up implementation of the NLRMP in earnest. The state government's action plan and tender documents for the NLRMP have also been hailed as exemplary by the DoLR.

The findings that emerged out of this study are that project success is largely dependent upon the support of the top bureaucracy as well as the political establishment. MP has been able to implement the NLRMP largely because the state's political executive has been treating the program at priority leading to a buy-in from the senior bureaucrats. This can be observed from the manner in which attempts by some lower level bureaucrat's to sabotage the project by not providing maps and other necessary artifacts on time, or by calling them as "missing" were resolved. However, public support for the project varies with the level of development of the district. The population of the less developed districts has been less supportive of the project compared to those in the more developed districts. The impact of multiple legacy tenure systems and a lack of modern surveys is being felt and is hindering swift completion of the program.

Another issue that emerged was that of co-ordination between different agencies as evidenced in the purchase of satellite imagery. While the state administration wanted the National Remote Sensing Centre (NRSC) to be a guide in its purchase of satellite imagery, their perspective was that of a vendor agent. This underlines the need to have in-house technical expertise in the field of remote sensing similar to the one in the state of Haryana.

An underlying challenge that was faced across the board was due to the extremely rigid project specifications, which is a serious lacuna in the program design. The program specifications are centrally-managed and driven top-down by the DoLR in the MoRD. This one-size fits all approach does not distinguish between the large and small states or those with varying degrees of complexity in the land administration system as can be seen in the extremely low estimate of sixty-six thousand records per *tehsil*, with the real world figures being much higher. If the MP government had strictly followed norms, it was possible that significant parts of the project would not have been implemented. However, the senior bureaucrats stepped in and provided support and funds (from the state budget, over and above the funds committed) to ensure project success.

This study identifies that to be successful, a land administration modernization program should be designed such that the unique aspects of the states and their varying needs can be accommodated. Further, the projects can only be implemented if it is supported by a committed bureaucracy as well as the political class.

Notes

[1]Source: http://www.doingbusiness.org. Retrieved April 28 , 2017

[2]"Why Secure Land Rights Matter" World Bank Group (2017, March 24).

[3]See note 2

[4]Source: https://www.cia.gov/library/publications/the-world-factbook/geos/in.html. Retrieved: June 19, 2017

[5]According to National Association of Software and Services Companies (NASSCOM), the trade association of Indian Information Technology companies, the ICTs industry contributes around 7.7% of India's GDP for FY2017. Source: http://www.nasscom.in/knowledge-center/publications/it-bpm-industry-india-2017-strategic-review. Retrieved: June 19, 2017.

[6]Department of Land Resources (DoLR) website at: http://www.dolr.nic.in/. Retrieved April 30, 2017.

[7]See for example Sud (2014). The list of interviewees and other details are available from the author.

[8]MP was India's largest state by area before the state of Chattisgarh was carved out in 2000.

[9]See for example, newsreports at http://timesofindia.indiatimes.com/good-governance/madhya-pradesh/Madhya-Pradeshs-growth-story-impresses-UK-investors/articleshow/54561359.cms and http://www.businesstoday.in/magazine/features/how-agriculture-growth-has-boosted-madhya-pradesh-gdp/story/217695.html. Retrieved: April 30, 2017

[10]Example documents at http://www.dolr.nic.in/dolr/downloads/pdfs/NLRMP_Tenders/cadastral_tender_mp.pdf and http://www.dolr.nic.in/dolr/downloads/pdfs/Madhya%20Pradesh's%20NLRMP%20Quarterly%20Action%20Plan%20for%202013-14.pdf. Retrieved: April 30, 2017.

[11]The term *"tahsil"* means "place of collecting" and is used in north and central India. In the south and west India, the corresponding term is a *"taluka"* and the official called *mamlatdar* (Baden-Powell, 1907).

[12]In western India, this official is called a *kulkarni* (if hereditary) or *talati* (if appointed) (Baden-Powell, 1907)

[13]The unit cost norm for survey/re-survey was set at ₹ 15,000/km^2. However, states proposed rates that varied between ₹ 15,000 to 40,000 per km^2 (Andhra Pradesh). The hill states of Himachal Pradesh and Sikkim proposed rates of₹ 35,562/km^2 and between ₹ 46,500 and ₹ 56,000 per km^2 respectively. An outlier was the state of Kerala, which proposed a rate of ₹ 2,70,000/km^2 (exclusive of surveyor wages) citing "major area of the state is under thick tree cover, cost of land is very high and heavily fragmented, most of the area is urban/semi-urban and labor charges are on higher side". See "Minutes of 2nd meeting of the Committee…" (DoLR, 2010, December 23)

[14]This information varies from state to state (Interviewee H). In MP this information includes photograph, Aadhar id etc. (Interviewee C).

[15]The challenges of creating cadastral maps from satellite and aerial imagery were highlighted

during various sessions at the World Bank Land and Poverty Conference (WBLPC) 2016. See Lakshmanappa and Singh (2017, March 21) for an Indian context.

[16] The judicial system requires authenticated originals and does not accept scanned copies.

[17]By leaving a part of the land with the trustees, he ensured that the original record was not deleted, thus preventing immediate detection.

[18]Program guidelines (DoLR, 2009b, April 17, pp 119)

[19]Interviewee I pointed out that years of neglect has led to the Survey of India (SoI) monuments disappearing and thus part of the exercise is also to setup the new monuments.

CHAPTER 3: DIFFUSION OF DATA POLICIES: A SUB-NATIONAL STUDY ACROSS INDIA

ABSTRACT

The how, why and when of policy proliferation are important questions that need to be answered while framing and evaluating policy. By understanding the reasons behind, and the mechanisms of policy adoption, policy makers can design policies that can be tailored to fit varying contexts and purposes, thereby ensuring wider adoption. The policy diffusion literature has identified multiple factors that affect policy adoption at the sub-national level. With most studies focusing on the United States, there is a lack of empirical studies on policy adoption in emerging country contexts. These countries' political and bureaucratic systems are significantly different from the US, preventing direct application of the learnings from US policy experiments. Policy design and implementation needs to be tempered with local knowledge, necessitating an understanding of the factors leading to policy adoption. This study is a modest attempt to fill the void created by the lack of empirical studies on policy diffusion in the Indian context.

Using a novel data set, we analyze the adoption of a land reform policy, the "National Land Records Modernisation Program" (NLRMP) aimed at modernizing (computerizing) land records and land administration in India. Despite the federal government's financial and technical support to the program, its adoption varies significantly across Indian states. We hypothesize that policy salience, the relative level of socio-economic development, the complexity in policy adoption and the level of federal support impacts policy adoption in this context. These hypotheses are tested on our dataset using binary logistic regression.We find mixed support for these factors, with some caveats. The policy implications and scope for future work is discussed.

1 Introduction

> "In today's interconnected world, understanding policy diffusion is crucial to understanding policy advocacy and policy change more broadly" —Shipan and Volden (2012)

Policy diffusion studies the how, why and when of policy adoption across jurisdictions. It is well accepted that not all policies are new "inventions", but that jurisdictions learn about policies from each other. There are numerous factors that can lead to a policy being adopted or not. It is only by understanding the mechanisms and reasons behind policy adoption, that policies can be tailored to make them contextual, thus leading to wider adoption.

Most policy diffusion research at the sub-national level has looked at how policies proliferate across the states of the United States of America. This literature has identified that policy adoption rests on a complex interplay of economic, social and political factors as well as policy context and salience. Successful, widely adopted policies are contextual and tempered with local knowledge.

The United States has a unique, strongly federal polity, which is significantly different from the political systems and bureaucratic structures that exist in many emerging economies. This means that US policy learnings are not directly transferable to emerging economies. However, a lack of studies in emerging economies leads to a proliferation of policies that are either directly imported from the developed world, or simply mimic such policies. These policies may not mesh well with the local context, resulting in policy failures. Therefore, it is imperative to study the reasons behind policy (non)adoption in specific, emerging

country contexts to understand what works and what doesn't.

We study a land reforms policy in India that seeks to create digital data from land records. This policy is supported by the federal government, but has seen uneven adoption across the Indian states. Using a novel dataset, we seek to uncover some of the determinants of policy adoption in an emerging country context, that is India. We hypothesize that policy salience, relative level of socio-economic development, the complexity in policy adoption and the quantum of federal support impacts policy adoption in this context. These hypotheses are then tested on our dataset using binary logistic regression.We find mixed support for these factors.

The next section lays out the theoretical framework underlying policy adoption. Section 3 lays out the importance of land administration and its challenges in an emerging economy. This is followed by the research question and the hypotheses in section 4. Section 5 details the data sources, the variables and the testable hypotheses. The statistical analyses are discussed in section 6, followed by the results and the limitations in section 7. We conclude by discussing the policy implications and scope for future work in section 8.

2 Policy Adoption

Why does a policy get adopted? Or, why does a seemingly wonderful policy not have any takers? Public policy analysis requires a clear understanding of not only how policies work but also why they were (or were not) adopted in the first instance. These studies are even more important in a federal system, where

states serve as the laboratories of policy, developing new policies that diffuse across jurisdictions (Gray, 1994). Studying the mechanisms of policy adoption allows policy makers to identify factors that make certain policies amenable to adoption, and hinder the adoption of others (Shipan & Volden, 2012).

Policy innovation has been defined as an idea that is new to the jurisdiction adopting it, as opposed to a completely new policy "invention" (Gray, 1994; J. L. Walker, 1969). Variations have been observed in the adoption of policies across jurisdictions, with two main explanations to account for this variance (F. S. Berry & Berry, 2014). The first explanation discusses influencing factors that are unique to the adopting jurisdiction (or to the policy itself) called the internal determinants. The second explanation considers the influence of other jurisdictions in the policy adoption process, that is policies "diffuse" across jurisdictions (F. S. Berry & Berry, 2014; Graham, Shipan, & Volden, 2013, 03; Gray, 1994, 1973; Karch, 2007, 2006; Makse & Volden, 2011; Nicholson-Crotty, 2009; Nicholson-Crotty & Carley, 2016; Shipan & Volden, 2012; J. L. Walker, 1969). Scholars have acknowledged that by itself, either explanation is insufficient to explain policy adoption. For better understanding the policy process, it is necessary to disentangle the effects of both explanations (F. S. Berry & Berry, 2014). The diffusion explanation is discussed first in section 2.1 followed by the internal determinants explanation in section 2.2.

2.1 Policy Diffusion

The field of policy diffusion is anchored in the studies of innovation diffusion, propounded by Rogers (2003). He defined diffusion as "the process by which

an innovation is communicated through certain channels over time among the members of a social system" (Rogers, 2003, p. 5). When studying policy diffusion, the innovations are specific policies "new" to the adopting jurisdiction(s), and the "social system" consists of all the jurisdictions that participate in the innovation process (J. L. Walker, 1969). These could include countries, regions, states, or municipalities depending on the level of study.

2.1.1 Policy Diffusion Mechanisms

Policy diffusion can be said to occur whenever a jurisdiction adopts a policy influenced by a similar policy in another jurisdiction (F. S. Berry & Berry, 2014). Karch (2007) identified three main mechanism of policy diffusion that he called "imitation, emulation, and competition". Shipan and Volden (2008) identified a fourth — "coercion", while F. S. Berry and Berry (2014) added the "normative pressure" mechanism to the mix. The main aspects of these mechanisms are outlined below.

Imitation is when jurisdiction (A) simply copies a policy from another jurisdiction (B) so as to look like B (Shipan & Volden, 2008). It occurs because policy-makers in A perceive B to be worthy of emulation, making them adopt any policy that B does, without evaluating its effectiveness or attributes (F. S. Berry & Berry, 2014). Although similar to "learning" (described below), it is different because in imitation, the focus is primarily on on the other jurisdiction (the actor), rather than on the action (the policy) (Shipan & Volden, 2008).

Learning, or emulation is a special case of imitation (Karch, 2007). It differs from imitation in not only being driven by what the jurisdictions have in common but the adopted policy's perceived success is also key. The aim of late adopters is to "equal or surpass the positive achievements of early adopters" (Karch, 2007). This is the process that has led states to be called the "laboratories of democracy" (Graham et al., 2013, 03; Karch, 2007; Shipan & Volden, 2008). F. S. Berry and Berry (2014) distinguish between "complete" and "incomplete" information in policy learning. If the learning is complete, every government has full information about the successes and failures of the policy in every jurisdiction it has been adopted in. In practice, it may be costly to gather complete information about the policy's success and failures, which forces policy-makers into taking shortcuts or processing information only from a subset of prior adopters. Interpreting the broader policy adoption without abandonment is an example of such a shortcut (Shipan & Volden, 2008). Early policy diffusion scholars considered geographical proximity key to learning (J. L. Walker, 1969), but now with improved communications other jurisdictions — those considered to be "leaders" in their field or peers with shared values are also becoming important (F. S. Berry & Berry, 2014; Karch, 2007; Shipan & Volden, 2008). According to Karch (2007), it is challenging to test this explanation owing to a lack of objective criteria and changing political conditions which may lead to changes in the criteria themselves over time.

Normative Pressure is a policy diffusion mechanism where a jurisdiction adopts a policy, not because it learns about it or imitates another jurisdiction, but simply because the policy has being widely adopted by jurisdictions with

whom it has shared norms (F. S. Berry & Berry, 2014; Sugiyama, 2012). These shared norms can be shaped by sharing of experiences through membership of professional organizations or via non-governmental organizations supported by international donors (Sugiyama, 2012).

Competition occurs when a jurisdiction adopts a policy to either gain an economic advantage over other jurisdictions, or to pre-empt them from doing so. It differs from the learning mechanism in that the adopting jurisdiction makes strategic policy choices to "shift the goalposts" (F. S. Berry & Berry, 2014). F. S. Berry and Berry (2014) categorize this mechanism as either "location-choice" or "spillover-induced". Location choice competition occurs when jurisdictions adopt certain policies to either entice firms or individuals to source from the jurisdiction goods and services whose provision is beneficial to it, or to discourage them from obtaining goods and services that are costly for it to provide. Examples of the former includes setting up of industries that provide employment to residents and tax revenue to the jurisdiction, while the latter includes welfare payments and subsidies (F. S. Berry & Berry, 2014; Shipan & Volden, 2012). A "spillover-induced" policy adoption occurs because another jurisdiction adopts policies that have either positive or negative spillover effects on the jurisdiction (F. S. Berry & Berry, 2014; Shipan & Volden, 2008). Creation of uniform infrastructure has been cited as an example of this mechanism "at work" by Shipan and Volden (2008). However, Shipan and Volden (2012) caution about exaggerating the impact of competition on policy diffusion by pointing out that the evidence for this mechanism is mixed.

Policy Coercion occurs when a jurisdiction *coerces* another into adopting its preferred policies by using force, threats or incentives (F. S. Berry & Berry, 2014; Graham et al., 2013, 03; Shipan & Volden, 2012, 2008). Economic sanctions are an example of such "sticks" in international politics (Graham et al., 2013, 03; Shipan & Volden, 2012, 2008). The federal government can force states into adopting its preferred policies by mandating certain actions, or by creating financial motivations via grants-in-aid (F. S. Berry & Berry, 2014; Eyestone, 1977; Gray, 1973). National financial incentives have been shown to influence policy adoption (Allen, Pettus, & Haider-Markel, 2004; F. S. Berry & Berry, 2014; Eyestone, 1977; Shipan & Volden, 2012, 2008; Welch & Thompson, 1980).

More than one of these mechanisms may affect policy diffusion and the mechanisms can also vary over time (F. S. Berry & Berry, 2014; Shipan & Volden, 2008). Attributes of the policies themselves also impact their diffusion (Gray, 1973; Makse & Volden, 2011).

2.1.2 Models for Policy Diffusion

Based on the above diffusion mechanisms, F. S. Berry and Berry (2014) have proposed three main policy diffusion models.

The National Integration Model is based on communication theorists' view of the diffusion process and assumes that the adopters are spread out nationally. It posits an S-shaped curve of adoption over time, which can be explained as few innovations being adopted initially . However, as adopters

come in contact with each other, or knowledge about these adoptions gets circulated, the adoptions increase, finally tapering off as the pool of potential adopters saturates (F. S. Berry & Berry, 2014). This model was used for earlier studies (Gray, 1973), but its utility is limited. It assumes that all potential adopters have similar characteristics and interact randomly, thus failing to account for either their inherent propensity to innovate or the real-life, non-random interactions amongst them (F. S. Berry & Berry, 2014).

The Regional Diffusion Model states that a jurisdiction is more likely to be influenced by its geographical neighbors. These models can be either neighbor-models, where the assumption is that influences only work across shared borders, or fixed-region models which consider jurisdictions as regional groupings (F. S. Berry & Berry, 2014; Karch, 2007). Karch (2007) posits that policy diffusion may be influenced by geographical proximity in a number of ways — networks amongst policy makers, shared media markets or cultural and demographic similarities. However, he contends that in light of modern communication technologies and existence of national and international professional networks the impact of geography is diminished. He further points out that empiricists have found limited support for this model (Karch, 2007).

The Leader-Laggard Model is compatible with a few different mechanisms of policy diffusion. Pioneer (or leader) jurisdictions adopt policies, which are then learned and adopted by the laggards. If the laggards are only interested in looking like the leaders, then this model becomes compatible

with the imitation mechanism. However, the challenge of this model is its inability to *a priori* identify either the leader jurisdictions, or an expected order of policy adoption. These lacunae make the model non-testable (F. S. Berry & Berry, 2014).

We see that all the three diffusion models discussed above have weaknesses that prevent them, when used in isolation, from adequately identifying the factors impacting policy adoption. As discussed earlier, scholars have identified the important role of internal determinants in policy adoption. These internal determinants are now discussed.

2.2 Internal Determinants

The internal determinants models assume that once a policy is known to a jurisdiction, the internal characteristics of the jurisdiction are the primary factors that impact adoption (F. S. Berry & Berry, 2014). The theory underlying these models is drawn largely from research on innovation at the individual and firm levels (F. S. Berry & Berry, 2014; Karch, 2006). Mohr (1969) proposed that the probability to innovate is directly related to the motivation and availability of resources to innovate, while being inversely related to the strength of obstacles hindering innovation. We discuss these three factors below.

2.2.1 Motivation to Innovate

Mohr (1969) had operationalized motivation to innovate in terms of attributes like "activism" and "ideology". In the context of policy adoption, the motivation can be translated into getting a problem solved for possibly winning re-election (F. S. Berry & Berry, 2014; Karch, 2006). If the problem appears to be severe enough on the policy makers' agenda to warrant a solution, there will be a greater motivation to innovate and adopt relevant policy, if other factors permit.

However, before a problem can be solved, it needs identification and appearance on the policy maker's radar (Bardach, 2012). Many factors impact this process of problem identification and its movement onto the agendas of policy makers (Bardach, 2012; Kingdon, 2011). Issue salience is one such factor. An issue is considered to be salient if it is important to a large part of the population because it either impacts them directly, or it is an issue which they care a lot about (Nicholson-Crotty, 2009). Problem salience itself is influenced by multiple factors which could be political, economic, social, historical or technical (F. S. Berry & Berry, 2014; Gray, 1973). As discussed by Nicholson-Crotty (2009), increased problem salience increases the political incentives for involvement and thus possibly the adoption of relevant policies.

2.2.2 Availability of Resources

Mohr (1969) had hypothesized that innovations requires resources and looked at organizational expenditures as a proxy for such resources. Policy implementa-

tion requires adequate financial and managerial resources (F. S. Berry & Berry, 2014; Sabatier & Mazmanian, 1979). In the context of policy adoption, J. L. Walker (1969) had pointed to the need of having "slack" (both monetary as well as human) resources to draw on. He also pointed out that such "slack" resources allow the "luxury of experiment". Gray (1973) had shown that providing additional resources through federal grants to welfare programs hastens their adoption, which was corroborated by Welch and Thompson (1980). Another resource possibly enabling the adoption of policies is the presence of a professional legislature that can afford the time and effort needed to gather knowledge about proposed policies (Karch, 2006; Shipan & Volden, 2006, 2012). Scholars have also treated the "policy entrepreneur" as a case of resource availability (F. S. Berry & Berry, 2014). Gray (1994) pointed out that of all these factors, the economic ones were the most important followed by political factors for policy adoption, while social factors showed policy-specific, mixed effects.

2.2.3 Barriers to Innovate

Closely aligned with the availability of resources, is the presence of barriers to policy adoption. Mohr (1969) provided examples of such obstacles to innovation at both the community and organization levels.These included worker attitudes, resistance to change, lack of information and the presence of rigid and mechanistic decision structures. An example of resistance to change inhibiting innovation is provided by Rogers (2003, pp 8–11) when he discusses the market failure of the technically superior DVORAK typewriter keyboard over the now-ubiquitous QWERTY layout. David (1985) discusses this as a case of "dependence" and

"locking" into existing technology.

The obstacles to adoption can change over time (Savage, 1985). They can also be due to existing policies, or be unintended effects of new policies, as Karch (2006) found with the passing of the Health Insurance Portability and Accountability Act (HIPAA). HIPAA restricted the conditions under which Medical Savings Accounts (MSA) could be setup, thus creating obstacles to the states' adoption of Medical Savings Accounts (MSA) legislation. F. S. Berry and Berry (1990) considered information uncertainty and public opinion as obstacles to the adoption of lottery policies.

2.2.4 Need for a Unified Model

Although Mohr's three correlates of innovation (discussed above) are crucial in explaining policy adoption, in isolation, they cannot explain all policy adoption. Specific policy content as well as diffusion mechanisms like learning or coercion play a key role in policy adoption (F. S. Berry & Berry, 2014; Clark, 1985; Gray, 1973; Makse & Volden, 2011). Makse and Volden (2011) tested the role of specific attributes in criminal justice policies along five dimensions and found that all dimensions mattered. Federal interventions have similarly been found to significantly affect policy adoption (Allen et al., 2004; Eyestone, 1977; Gray, 1973; Karch, 2006; Nicholson-Crotty, 2009; Welch & Thompson, 1980). Factoring in these determinants requires including variables that may be "*ad hoc*" in the context of innovation theory, but crucial in explaining adoption of the policy in question (F. S. Berry & Berry, 2014, pg 322). As an example, F. S. Berry and Berry (2014, pg 322) point out that the presence of religious fundamentalists

doesn't reduce policy adoption *per se*, but reduces the probability of adopting policies that go against issues central to the group's religious beliefs.

Therefore, to better understand policy adoption, it is necessary to include all these multiple sources simultaneously in the model, which requires a unified model of policy adoption.

2.3 Policy Adoption Model

Multiple factors impact policy adoption and their effects cannot be modeled using the existing diffusion models discussed in section 2.1.2. To get around the limitations of these diffusion models, F. S. Berry and Berry (2014) proposed a unified model of policy adoption, which can be written as:

$$P_{adopt} = f(M, R, O, E) \quad \text{where}$$

P_{adopt} : Probability of Policy Adoption
M : Motivation to Adopt
R : Resources or Obstacles
O : Other Policies
E : External Determinants

The dependent variable in this model is the probability of policy adoption and the study focuses on a single policy, rather than a set of policies (as done in the earlier policy diffusion studies (Gray, 1973; J. L. Walker, 1969)). This unified model is able to accommodate all the factors identified earlier.

The innovation correlates identified by Mohr (1969) are captured by variables that operationalize (a) the motivation to adopt, or problem severity *(M)* , and (b) the barriers for adoption, or the resources available to surmount them *(R)*. The

external determinants like policy salience, federal support (or lack of it) etc. are captured by the E factors. By operationalizing O, the model also considers the presence (or absence) of competing / complementary policies .

As Shipan and Volden (2012) have noted, it is crucial to understand the hows and whys of policy adoption so as to be able to tailor policies as needed and ensure their wider adoption. However, most of the empirical studies have concerned themselves with policy proliferation in the fifty states of the United States of America, with not many studies done in emerging economy contexts. The need for such studies and the current research contribution is discussed next.

2.4 Policy Adoption in Emerging Economies

Successful, widely adopted policies are contextual and tempered with local knowledge. Designing such policies requires studying the reasons behind the adoption or non-adoption of a policy to understand what does and does not work.

The policy diffusion literature is largely anchored in the policy processes of the United States of America (USA), with its unique, strongly federal polity. The USA also has a unique "separation of powers" governance system[1], which is significantly different from the political systems and bureaucratic structures existing in many countries. This means that policy learnings from the US system are not directly transferable to other nations with different political and social systems, especially the emerging economies.

Few have studied policy diffusion in emerging economies and / or other political systems. The handful of non-US studies include Kim, Kim, and Moon's 2014

study of South Korean local governments' support to multicultural families, and a 2008 study by Sugiyama on the diffusion of social reform policies across Brazil's largest cities. Both these countries are examples of multi-party, federal, presidential governments.

This lack of empirical studies on policy adoption in emerging country and different political contexts means that we do not know enough about the local factors that encourage or impede policy adoption. Hence, often policies are either imported wholesale from the developed world or slightly tweaked to fit specific contexts. However, as these policies miss local context, they often do not mesh well with the local context resulting in failures (cf. Dolowitz & Marsh, 2000; Heeks, 2002).

The current research attempts to fill this void by studying the factors that impact the adoption of policies in India — a large, multiparty, Westminster styled federal democracy as well as an emerging economy. The policy being studied here is a central government sponsored program that aims to modernize land administration across Indian states. This program provides technical and financial support to states to help them modernize their land records data.

The next section provides a brief context on land administration and its importance to development, especially the need to create and manage land administration data. This is followed by a brief introduction to the central government sponsored program aiming to create land data.

3 Land Administration

Land being a key input to economic activity, its administration and policies regulating land use play an important role in development (Banerjee & Iyer, 2005; Besley & Burgess, 2000; Dale, 1997; Deininger et al., 2009; Feder & Feeny, 1991; Feder & Nishio, 1998). The land administration function provides the infrastructure required to implement land policies (Williamson, 2001). A major component of the land administration infrastructure are land records, which define the roles and responsibilities of the stakeholders (Bennett et al., 2008; Wallace & Williamson, 2006; Williamson, 2001). Given the centrality of land to society, this information needs to be freely available and accessible to all members of society, thus giving land records the character of a public good. These factors have led scholars to propose treating land administration as a public good, and part of the national critical infrastructure (Bennett et al., 2012; Bennett et al., 2013). However, as land often has high economic value, information about it is also equally valuable, making its provision (or non-provision) lucrative and a potential source of corruption (Bussell, 2012; Goyal, 2012).

By virtue of being one of the oldest administrative functions, land administration is also impacted by extant social and cultural norms, making it extremely complex, rooted in tradition, and with significant variations across geographies. For example, the way in which land is held ("tenure") varies drastically across societies (Payne, 2004; Törhönen, 2004). The information relating to land tenure is recorded in registers called "cadastres". Both, the content of the record, as well as the recording method(s), vary significantly across geographies.

Further, land administration is a dynamic process with unique spatio-temporal characteristics (van der Molen, 2002). This adds to the existing complexities. Land changes hands over time as it gets sold to multiple parties. Land also gets partitioned when family assets are divided or when passed on through generations. All these events are required to be accurately recorded in the cadastres. However, due to administrative inefficiencies and / or vested economic interests, many times this recordation does not occur.

This missing / ambiguous / old information about the land leads to disputes and consequent loss of productivity. According to a report by the McKinsey Global Institute, most land parcels in India were under dispute, and land market distortions accounted for more than one and a quarter percent of lost growth annually (MGI, 2001). Robinson (2013) found land disputes to be almost a tenth of the Indian Supreme Court's workload. Given that land is a state subject under the Indian constitution, and thus the Supreme Court does not have original jurisdiction on land matters, this figure is alarming. A recent survey of the lower judiciary (district and state high courts) found land and property related matters to be two-thirds of civil litigation, a figure consistent across income levels (Narasappa & Vidyasagar, 2016). These pervasive land disputes impact everyone, rich and poor alike.

Hence, it is imperative for development that the land records accurately reflect the ground position and contain all information necessary to comprehensively manage the land. Towards this end, governments the world over have turned to ICTs and taken numerous digital initiatives to update and maintain their land administration systems(Habibullah & Ahuja, 2005; Lang, 1981; Lemmen &

van Oosterom, 2001; Maggs, 1973; McCormack, 1992; Navratil & Frank, 2004). In India, as land is a state subject, it is ultimately the responsibility of the state government(s) to drive changes to get their land administration systems in order. However, considering the importance of land administration and its vital role in economic development (Besley & Burgess, 2000; Dale, 1997; Feder & Feeny, 1991; Feder & Nishio, 1998), the central government has been supporting the states in strengthening their land administration practices and processes. One of the avenues of providing such support is the "DILRMP" of the Department of Land Resources (part of the Ministry of Rural Development). The next section provides a brief description of the program.

3.1 The Digital India Land Records Modernisation Programme (DILRMP)

As discussed in Chapter 2, the DILRMP is a recently (2016) modified version of an earlier program called the NLRMP. The NLRMP was started with the ultimate objective of providing land title guarantee in India by moving the land titling system from the current "presumptive" titling to a conclusive (or "Torrens") titling regime (National e-Governance Division, 2011, February 22, pp 147–152). In a conclusive titling scheme, the state guarantees title to the land and all liens are recorded on the certificate, which is maintained by the "Registrar of Titles" (Kent, 1988; Törhönen, 2004). For more on the features and benefits of a conclusive titling system see Chapter 2. The features of the Torrens system can lead to a reduction in, and, a faster resolution of any land disputes (Bostick, 1987; Goldner, 1982; McCormack, 1992; Wadhwa, 2002). This in turn can help unlock land value, providing consequential developmental benefits (Bhidé, 2008;

Deininger & Goyal, 2012; Galiani & Schargrodsky, 2010; Venkataraman, 2014).

However, as the state stands guarantee, it is liable to provide compensation to the injured party in any case of fraud or error in the land transaction(s) (Kent, 1988; Risk, 1971; Szypszak, 2003). This implies that the land administration function requires a near real-time, integrated spatio-temporal view of the land resources. ICTs can provide this spatio-temporal view by linking together various related systems, which may include — (a) a GIS to uniquely identify the land parcel(s) in space, (b) the land registration system to validate the antecedents of both land and transacting parties and maintain temporal integrity, (c) banks and mortgage providers to get clarity on any liens on the property, and (d) the legal system to flag any disputes etc. (R. N. Cook, 1969; McCormack, 1992). The aim of the NLRMP / DILRMP is to provide the necessary infrastructure and support, allowing such an integrated system to be built[2]. The uptake of the program(s) is being studied between 2008 (inception of NLRMP) till 2014 (when a new dispensation took charge at the center), and hence in this study, we will refer to the program as NLRMP.

The NLRMP was started in 2008 by merging two centrally-sponsored programs (started in 1988) — the IT-centric "CLR" and the transaction oriented "SRA & ULR". However, as these programs did not aim at conclusive titling, the NLRMP was conceived to enable the long-term goal of conclusive titling by fixing the shortcomings in the CLR and SRA & ULR schemes.

Under the NLRMP, the central Ministry of Rural Development provides a specified quantum of financial and technical assistance to the state governments, as given in Table 3.1 on the following page. The unit of implementation of the

Table 3.1: Central Assistance provided under the NLRMP

Activity	**Quantum of Central Assistance**	
	Special Category State	**General Category State**
Computerization of land records, which includes (a) digitization of cadastral maps, (b) integration of textual and spatial data, (c) data centers at *tehsil*, Sub-division, District and State level, (d) interconnectivity among revenue offices	100%	
Survey/resurvey and updating the survey & settlement records (including ground control network and ground-truthing) using modern technology options	90%	50%
Computerization of Registration including connectivity to Sub-Registrar Offices (SROs) with revenue offices	90%	25%
Modern Record Rooms/land records management centers at the *tehsil* level	90%	50%
Core Geographic Information System (GIS)	100%	
Training and capacity building	100%	

NLRMP is the district.

In the Indian administrative structure, the country is divided into twenty nine states and seven UTs[3]. The central government in India devolves a share of its revenue to the states in accordance with a certain formula (Ministry of Finance, Government of India, 2013). For this purpose of revenue sharing, the states have been grouped into two categories—"general" and "special". The "special

category" states are those that have traditionally been underdeveloped, have rugged terrain, or are on the international border. These special category states get a larger share of the central resources (Ministry of Finance, Government of India, 2013). As Table 3.1 indicates, the quantum of central assistance varies by state category in the case of the NLRMP also.

The states and UTs are further subdivided into districts, whose administrative sub-divisions are called *tehsils*, which comprise villages. According to the 2011 Census of India there are 640 districts in the country.

The state governments are responsible for selecting the districts, and creating the required proposals to seek funds for implementing the NLRMP. These proposals are then approved by the central Ministry of Rural Development, and the funds released in a phased manner. The quantum of central assistance to be provided is as given in Table 3.1 on the previous page.

However, across the states, the program's uptake has been uneven. There is a significant variation among the states that are part of the program. This study seeks to understand the factors contributing to the adoption of NLRMP using the policy adoption framework discussed in Section 2. The next section discusses the research question and the hypotheses being presented.

4 Research Questions & Hypotheses

Modernization of land records is an important development imperative. Land records help in providing tenure security which has been shown to be an impor-

tant factor in combating poverty and thus for economic development[4]. However, despite numerous attempts at both the central and state government levels, there continues to be considerable heterogeneity in the condition of land records. This heterogeneity exists despite the recognition that though the ultimate objective is to move towards "conclusive" titling[5], numerous intermediate benefits also manifest. Even in the current framework of "presumptive" titles, the computerization of land records and modernization of related infrastructure, can lead to many positive developmental and social outcomes (Deininger & Goyal, 2012; Venkataraman, 2014).

This heterogeneity in land records can be attributed to the the differences in adopting the policies aimed at computerizing and modernizing the land records. Evidence of the heterogeneity in adoption of the land reforms policy (NLRMP) in India is shown in the map of Figure 3.1 on the following page as well as the numbers in Table 3.2 on page 94. This variation exists at both the state and district levels. Hence, we state the main research question as: "*If land reforms are universally recognized as being key to development, why is there a variation in the adoption of land policies across states and across districts?*". We investigate this question using the policy adoption framework, at both the state and district levels, to identify the potential factors bringing about this variation.

The adoption of land reforms is a complex social phenomenon that is determined by an interplay of diverse factors. These may be viewed as internal determinants, external determinants and specific policy attributes (Gray, 1994, 1973). Our review of policy adoption and diffusion processes in section 2 allows us to categorize these factors as falling into four domains — (a) policy salience, or the

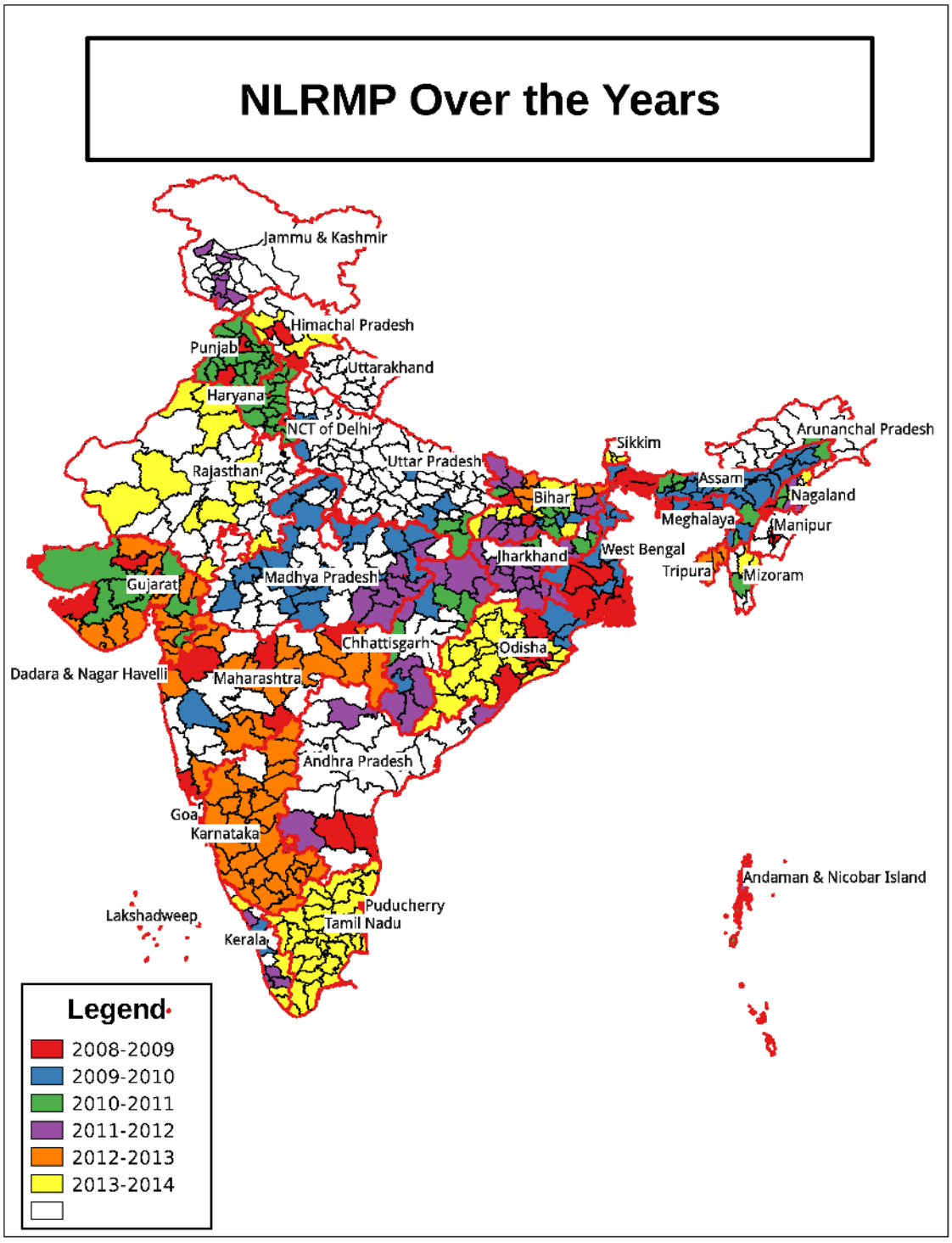


Figure 3.1: NLRMP Proliferation over the years (2008–14)

Table 3.2: NLRMP Proliferation over the years

Year	**Number of**	
	States ($N = 29$)	**Districts ($N = 613$)**
2009	17	64
2010	10	68
2011	12	67
2012	9	59
2013	5	64
2014	7	78
Total	**26**	**400**

motivation to adopt (Nicholson-Crotty, 2009), (b) resources to adopt, or the availability of slack resources to adopt and implement policy (Gray, 1973; Tolbert, Mossberger, & McNeal, 2008; J. L. Walker, 1969), (c) implementation complexity, or the barriers to adoption (Nicholson-Crotty, 2009; Nicholson-Crotty & Carley, 2016; Sabatier & Mazmanian, 1979; Sapat, 2004; Tolbert et al., 2008; R. M. Walker, 2014), and (d) external factors like federal support (Eyestone, 1977; Gray, 1973; Karch, 2006; Welch & Thompson, 1980).

4.1 Policy Salience

In terms of its policy attributes and salience, the land policy program in India that we study (NLRMP) is a unique program. *Prima facie*, it is an administrative reform. However, under the hood it is a digital data creation project and a program that is seeking to usher in e-governance in land administration. It

is salient as it deals with the highly complex and emotive issue of land; in a country that is overwhelmingly rural, driven by an agricultural economy, and possesses high illiteracy rates. These facets of the program move it from the realm of a mere routine administrative reform (McNeal, Tolbert, Mossberger, & Dotterweich, 2003), to a multi-hued program impacted by diverse factors.

The (NLRMP) deals with the creation of land records (or cadastral) data. Creating land records data is a long drawn, tedious effort which doesn't yield immediate results (see Chapter 2 for a brief history of land records computerization around the world). Hence, it can be safely assumed that the severity of the problem depends upon whether the land records exist in a usable form or not. Jurisdictions that have usable land records (whether digital or not) will be less inclined to adopt such a program as it may not yield immediate results (F. S. Berry & Berry, 2014, pg 325). On the other hand, jurisdictions that do not have clarity of land records, mainly due to historical reasons, will adopt the program.

As discussed in section 3, cadastres record land tenure information. Thus the structure and information of the cadastre is defined by the characteristics of the land tenure. This means that the type of tenure largely determines what land records are available in a jurisdiction.

India has a long history of land administration, which in some cases goes back to more than two thousand years (Chapter 2). When the English East India Company started to collect land revenue (ca. 18^{th} century), they faced a multiplicity of land revenue systems in operation around the country (Baden-Powell, 1907, 1892a). They classified these into three main systems, namely the (a) *raiyatwari* (cultivator), (b) *mahalwari* (village), and (c) *zamindari* (landlord). Variants of

these tenure systems were in operation across the country as can be seen in column 6 ("Tenure Type") of Table 3.3 on page 103. These tenure details and their histories have been detailed by Baden-Powell in his three volume work "The Land-systems of British India" (1892a). The salient features of the three tenure systems and their impact on land records are given below (Baden-Powell, 1907, 1892a; Mishra, 2016; Rothermund, 1971).

***Raiyatwari*,** or cultivator system. Here, the individual cultivators hold ownership of the land and are liable to pay the land revenue. To ensure efficient revenue collection, the administration records many details about the land including who owns or rents it. Hence, in this case the administrative records have full knowledge of who owns or tills the land.

***Mahalwari*,** or the village system, where a village, or group of villages was liable to pay the revenue to the government. Every cultivator/tenant's name was recorded by the village accountant (or *patwari*) so as to have a clear idea of liabilities. Hence, in this case also, the administration has knowledge of who owns what and what are their rights. Further, in many places where such systems were prevalent, complete cadastral surveys had been performed (Baden-Powell, 1892c), which was not the case with the *raiyatwari* system (Rothermund, 1971).

***Zamindari*,** or landlord system was largely prevalent in the eastern part of the country, especially in Bengal. The English East India Company modified the erstwhile land tenancy systems and permanently settled the revenue with landlords[6]. In this landlord system, the actual tiller and his/her rights were hidden from the government administrative machinery. The

government was only interested in collecting its share of the land revenue and the landlord was its single point of contact. Thus, as far as the Company revenue officials were concerned, the only records that mattered were the extent of the fields and which landlord(s) were liable for the land revenue. The landlords were the land owners and free to rent it out to the highest bidder. Thus, the record of rights were all in the landlord's names, with only the landlord knowing the antecedents of the actual tenant. This led to significant administrative challenges. Post-independence (1947), this system was abolished and limits placed on the amount of land an individual could own. However, in the period leading up to the abolition, the village accountants (*patwaris*) practiced corruption by falsifying records to benefit the erstwhile *zamindars* and themselves (Neale, 1962, pp 245). This led to the land records not mirroring the actual ground situation.

Because of this unique history, the areas under the erstwhile *zamindari* systems do not have land records of the granularity that are available in the *raiyatwari* and *mahalwari* areas.

Based on the foregoing, we hypothesize that policy adoption will vary according to the historical tenure types. Specifically:

H1: *A state with a Zamindari type of land tenure will show a greater propensity for adopting the policy compared to either the Raiyatwari or Mahalwari states*.

4.2 Resources to Adopt

Policy adoption and its implementation requires resources to be allocated and dispensed. The "slack" resources hypothesis stipulates that only a jurisdiction with sufficient available resources would be able to commit to a policy adoption (J. L. Walker, 1969). These resources include not only money, but also the availability of skilled professional staff as well as various socio-economic factors (J. L. Walker, 1969). Variegated resources are especially needed when the policy encompasses two major activities — land administration and digital data creation (McNeal et al., 2003; Tolbert et al., 2008). For parts of the NLRMP, the states are expected to provide funds (see Table 3.1 on page 90), and in such cases the availability of slack resources with the states becomes important. Deployment of funds is also not automatic, as it requires administrative frameworks and capable manpower, making institutional capacities essential for policy adoption and its implementation (Sabatier & Mazmanian, 1979; Sapat, 2004; Tolbert et al., 2008). The availability of funds and institutional capabilities gets reflected in the extent of existing development levels of the state. Hence, we hypothesize:

H2: *Controlling for tenure type, a more developed jurisdiction will have a greater propensity for adopting the policy.*

4.3 Implementation Complexity

Clear and unambiguous land records that mirror the actual ground situation provide tenure security. However, the creation and management of land records data is a complex exercise cutting across administrative functions and requires

significant amount of financial and technical capabilities. Besides this inherent complexity, in emerging economies other factors, such as administrative inefficiencies, corruption, and lack of resources, make the task even more challenging. It has also been observed that diffusion of administrative reforms is often left out, or is the last to be taken up in contrast to major re-distributive or economic development policies, as they are largely technical and not "value-laden" (McNeal et al., 2003). Further, they generally impact the public officials, rather than the populace, and thus professional networks may matter more in adoption of administrative reforms (McNeal et al., 2003). With limited implementation challenges, simpler policies tend to get adopted and implemented with ease. Complex policies require learning and thus spread more slowly (Makse & Volden, 2011; Nicholson-Crotty, 2009).

Institutional capacities matter in implementation of e-governance initiatives (Tolbert et al., 2008). These institutional capacities lie in the bureaucracy, which is a key constituency for implementing administrative reforms. Thus, the capacities and capabilities of the bureaucracy in negotiating the implementation challenges of complex policies will impact policy adoption (McNeal et al., 2003). Hence, we hypothesize:

H3: *Controlling for tenure type and the level of existing development; an increase in the extent of implementation complexity leads to a reduced propensity for adopting the policy.*

4.4 External Factors

Highlighting the need of economic resources for policy adoption, it has been found that policies that have federal support diffuse more rapidly than those which do not enjoy such support (Eyestone, 1977; Gray, 1973; Karch, 2006; Shipan & Volden, 2012, 2008; Welch & Thompson, 1980).

The land reforms program that we are studying (NLRMP) provides financial and technical support from the Union Ministry of Rural Development to the Indian states as detailed in Table 3.1 on page 90. This quantum of support varies depending on whether the state is a "General Category State", or a "Special Category State" (see Table 3.1 on page 90). The special category states have been indicated with a "Y" in the "SCS (2)" column in Table 3.3 on page 103. Hence, we can use the state category as a proxy for federal support. We hypothesize:

H4: *Controlling for tenure type, the level of existing development, and implementation complexity; a special category state will have a greater propensity for adopting the policy.*

Based on the foregoing, we outline our data, methods and testable hypotheses in section 5. The analysis is conducted at two levels — the state, and the district. The state level analysis looks at the factors that impact adoption of the policy. The district level analysis will try to identify why certain districts are chosen for policy implementation.

At the state level, tenure type, resource availability, administrative capacities, and external factors are expected to vary. At the district level we expect policy

salience, resources and administrative capacity to vary. While analyzing at the district level, we control for the factors identified at the state level.

5 Data and Methods

We use a novel dataset created by cross-linking data provided by the Management Information System (MIS) of Digital India Land Records Modernisation Programme (DILRMP) with other datasets including the national census, national agricultural census, and other state and national indicators[7]. The hypotheses offered in sections 4.1–4.4 are tested on this dataset using logistic regression (Agresti & Finlay, 2009; James, Witten, Hastie, & Tibshirani, 2013; Kabacoff, 2015; R Core Team, 2017) as the statistical technique (section 6 on page 118).

We first define the sample space. This is followed by a description of the variables used and the testable hypotheses at both the state and district levels.

5.1 Sample Space

India has twenty nine states and seven Union Territories. All of these should find a place in our analysis at the state level and be the unit of analysis.

The Union Territories have been excluded from the analysis due to their lack of administrative autonomy and lack of data[8].

Of the twenty nine Indian states, Telangana has been excluded as it came into

existence in 2014 (at the end of the study period). Additionally, five states which were not part of British India are excluded because their tenure systems are incompatible with either *raiyatwari*, *mahalwari* or *zamindari*[9]. This brings the sample space down to twenty three states ($N = 23$).

At the district level, the canonical source of data is the Census of India, 2011. The census lists 640 districts. The excluded seven UTs and five states comprise 56 districts. Further, the agriculture census does not have data for six districts, which also have to be excluded. This gives us an $N = 578$ in the district dataset.

The data sourced from the MIS of the DILRMP covers the period 2008 till date (the NLRMP program started in 2008). In 2014, a significant political regime change took place at the center. To prevent this exogenous shock from confounding the results, we decided to restrict our analysis from 2008 to 2014. A summary of the data sourced from the NLRMP MIS has been given in Table 3.2 on page 94.

5.2 Dependent Variables

5.2.1 State Level

At the state level, we use the "proportion of districts adopting the policy" (PropAdoption) as the dependent variable. This value for every state is listed in the "Prop Adopt" column (5) in Table 3.3 on the next page.

Table 3.3: Salient Characteristics of the Indian States

State	SCS	Dist Tot	Dist Adopt	Prop Adopt	Tenure Type	State Dev Index	Num Marg Hold
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Andhra Pradesh	0	23	5	0.22	R	0.46	0.95
Arunachal Pradesh	1	16	1	0.06	O	0.26	0.29
Assam	1	27	27	1.00	R	0.29	1
Bihar	0	38	38	1.00	Z	0.24	1.36
Chattisgarh	0	18	13	0.72	R	0.26	0.87
Goa	0	2	0	0.00	O	0.95	1.15
Gujarat	0	26	26	1.00	R	0.5	0.55
Himachal Pradesh	1	12	7	0.58	M	0.58	1.04
Haryana	0	21	20	0.95	M	0.57	0.72
Jharkhand	0	24	20	0.83	Z	0.26	1.02
Jammu & Kashmir	1	22	7	0.32	R	0.47	1.24
Karnataka	0	30	29	0.97	R	0.52	0.73
Kerala	0	14	10	0.71	R	0.85	1.44
Maharashtra	0	35	24	0.69	R	0.63	0.73
Meghalaya	1	7	5	0.71	R	0.3	0.73
Manipur	1	9	4	0.44	O	0.42	0.76
Madhya Pradesh	0	50	27	0.54	M	0.24	0.65
Mizoram	1	8	4	0.5	R	0.48	0.81
Nagaland	1	11	9	0.82	R	0.43	0.05
Odisha	0	30	30	1.00	Z	0.21	1.08
Punjab	0	20	20	1.00	M	0.61	0.23
Rajasthan	0	33	11	0.33	M	0.35	0.54
Sikkim	1	4	4	1.00	O	0.59	0.79
Tamil Nadu	0	32	32	1.00	R	0.64	1.15
Tripura	1	4	4	1.00	O	0.53	1.29
Uttar Pradesh	0	71	8	0.11	M	0.35	1.18
Uttarakhand	1	13	0	0.00	R	0.61	1.1
West Bengal	0	19	19	1.00	Z	0.44	1.22

5.2.2 District Level

The DILRMP MIS[10] provides information about the proliferation of the program amongst the various districts, which is used as the key dependent variable. This information for the years 2008–2014 has been pooled together to get the list of states and districts which adopted the NLRMP. Based on this data, the dependent variable PolicyAdoption has been developed as a dichotomous indicator of the *probability* of policy adoption[11]. If a district adopted the policy between 2008–2014, then PolicyAdoption is set to '1', else PolicyAdoption is set to '0'.

5.3 State Level Independent Variables and Testable Hypotheses

At the state level, we identify four testable hypotheses. These have to do with policy salience (or the motivation to adopt), resource availability, obstacles to implementation and vertical diffusion (federal support). Ready indicator variables for the different hypothesized factors impacting policy adoption could not be universally found. While some of these variables were available, others had to be proxied. The four hypotheses, along with their associated variables are described next.

5.3.1 Policy Salience (State)

As discussed in section 4.1, the motivation to adopt the policy, or its salience can be proxied by the tenure type prevalent during the British (pre-independence) era. We have noted that owing to the unique administrative structure, the

zamindari tenures had a lack of land records data compared to the either the *mahalwari* or the *raiyatwari* systems. We assume for the purpose of this study that tenure types were largely homogeneous within the states[12].

As discussed in section 5.1, this study only considers the twenty three states that had either the *zamindari, mahalwari* or *raiyatwari* tenure type. These tenure types are represented as "Z", "M" and "R", respectively in column 6 ("Tenure Type") of Table 3.3 on page 103. The variable is called: "$\mathsf{TenureType}_{\mathrm{STATE}}$". The testable hypothesis at the state level is:

TH1_STATE: *The probability of adoption of the NLRMP at the state level depends upon the tenure type prevalent in the British era (* $\mathsf{TenureType}_{\mathrm{STATE}}$*).*

To test this hypothesis, $\mathsf{TenureType}_{\mathrm{STATE}}$ is treated as a nominal categorical variable. Forthe regression analysis, *raiyatwari* ("R") is the base against which the two tenure types ("M" and "Z") are compared.

5.3.2 Resource Availability (State)

We hypothesize that the extent of existing development plays a *positive* role in policy adoption. The availability of "slack" resources (Gray, 1973; McNeal et al., 2003; Mohr, 1969; Sapat, 2004; Tolbert et al., 2008; J. L. Walker, 1969) and the extent of development are viewed as proxies for one another in the policy diffusion literature.

J. L. Walker (1969) used a mix of measures — percentage of urban population, average per capita income, value added by manufacturing, average value of farm-

land, percentage of illiterate population and median school years completed. F. S. Berry and Berry (1990) proxied resource availability using state revenue to expenditure ratio, and real per capita state income in their analysis of state lottery adoptions. McNeal et al. (2003) use a number of measures like, total state revenue per capita, state income per capita, as well as education and urbanization levels as development proxies. Sapat (2004) used per capita personal income and agency staffing levels to proxy slack resources. Resources were proxied by Daley and Garand (2005) using real per capita income and they also added education as measured by the percentage of state residents with bachelor's degrees. Karch (2006)'s proxy for state resources was state per capita income,while Tolbert et al. (2008) use total state revenues per capita to measure this aspect. Similarly, Bhatti, Olsen, and Pedersen (2011) in their study of the proliferation of customer service centers in Denmark used tax base per capita, controlled with per capita expenses as their "slack" resources proxy. Lyson (2016) in her study of farm to school programs in the US used the mean income of people in the state as a proxy for this aspect.

Thus we see that researchers have used various measures to proxy the availability of resources to adopt or implement policy. These variables have often been selected based on availability and best fit to the policy explanation(s).

We argue that land policy is a multi-faceted policy and it not only involves deployment of "slack" financial resources, but also demands a larger institutional capacity for implementation. Further, there is a "demand side" aspect to this particular policy as it seeks to usher in e-governance in land administration (Tolbert et al., 2008). Therefore, a single measure of resources like per capita state

gross domestic product, or even a set of measures like education etc. would not be able to capture the wide gamut of capabilities required in this case[13]. This study uses a state development index based on the Raghuram Rajan committee's 2013 *Report of the Committee for Evolving a Composite Development Index of States* (Ministry of Finance, Government of India, 2013)[14]. This report proposed a composite index that takes values between 0 and 1 as a measure of a state's "under-development". A *higher* value of this *under*-development index meant a *lesser* extent of development. However, we need an index where *higher* values mean a *greater* extent of development. To get this state "development" index, the Rajan committee's *under*-development index is subtracted from '1'. More details on the Rajan committee's index are given in appendix B. This variable is indicated as $\mathsf{DevIDX}_{\mathsf{STATE}}$and its values for the states given in column 7 ("State Dev Index") of Table 3.3 on page 103. Our testable hypothesis at the state level is:

TH2_STATE: *Controlling for tenure type, the probability of adoption of the NL-RMP at the state level is directly proportional to the state's development level (* $\mathsf{DevIDX}_{\mathsf{STATE}}$*)*

5.3.3 Obstacles to Implementation (State)

We have hypothesized earlier that more complex implementations lead to a decrease in the propensity of policy adoption (section 4.3). The literature has differing interpretations of policy complexity. According to Nicholson-Crotty (2009), complex policies require specific technical expertise, while for Makse and Volden (2011) policy complexity is related to its ability to be legislated. Neither

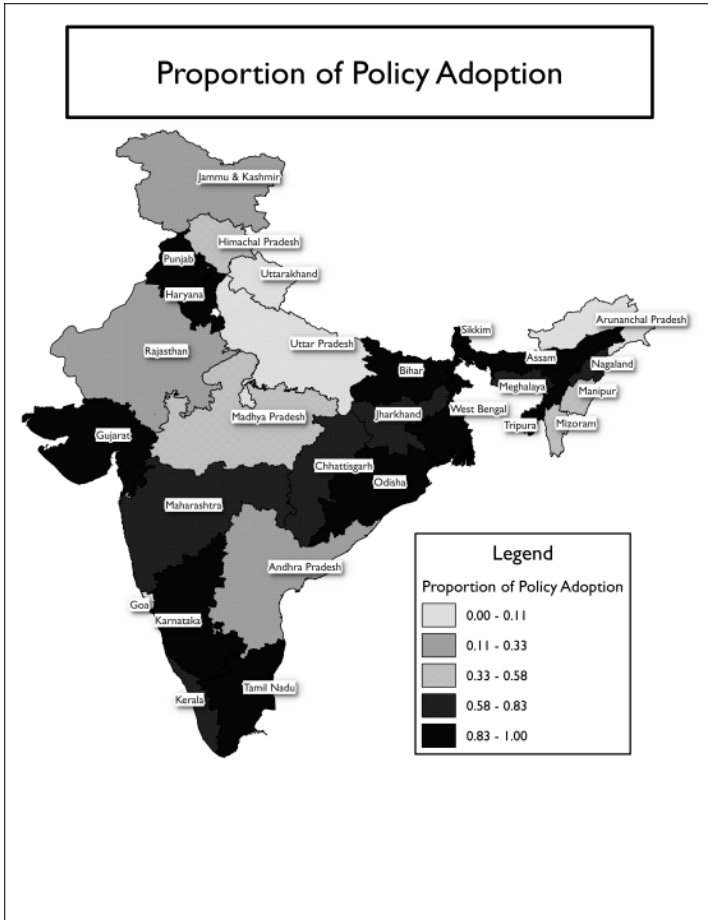


Figure 3.2: Proportion of Policy Adoption

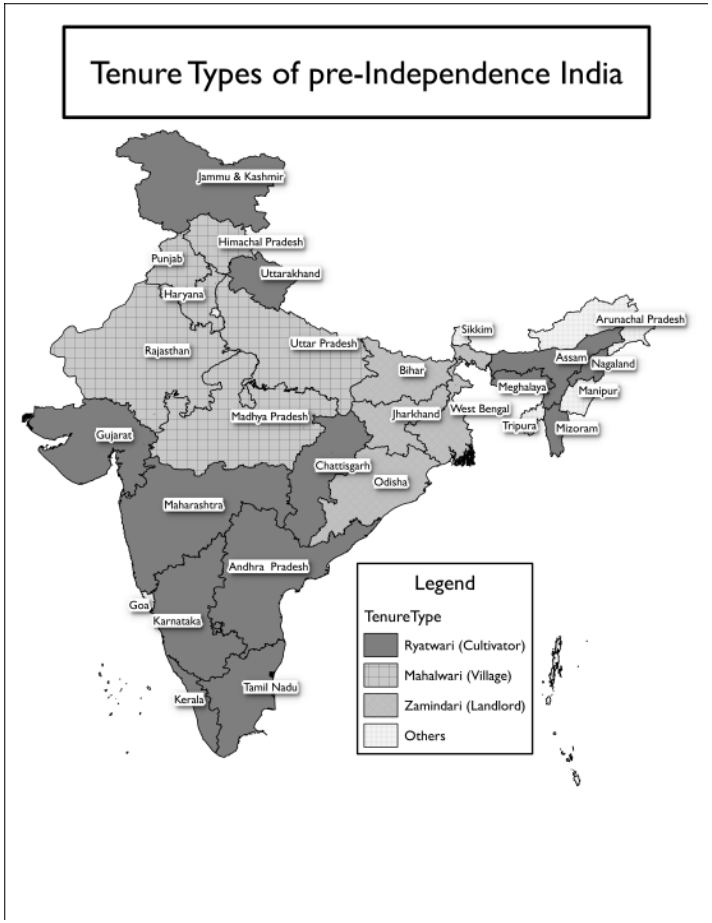


Figure 3.3: Pre-independence India tenure types

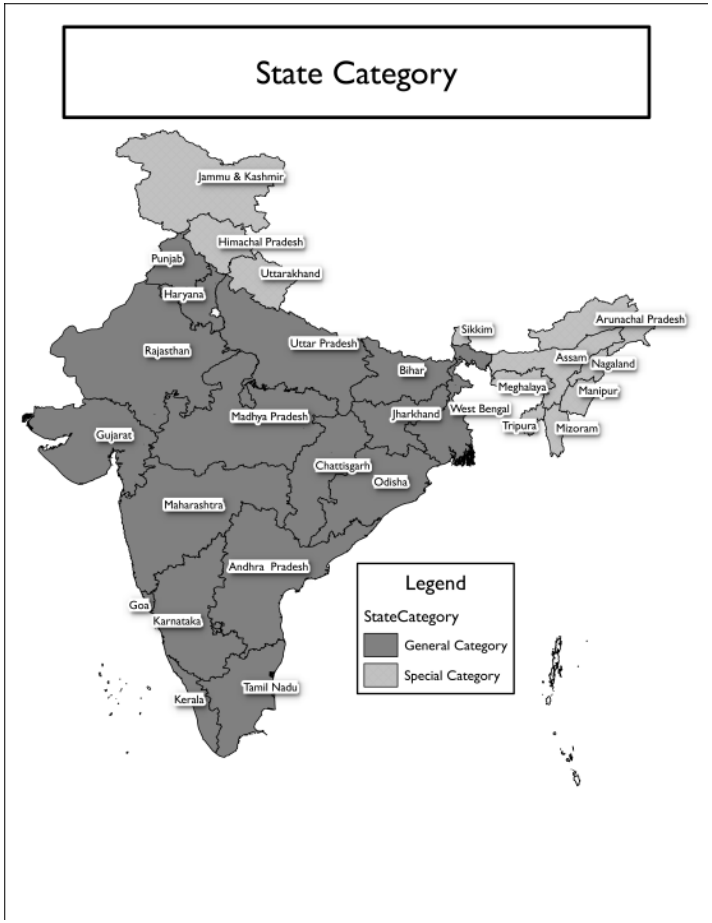


Figure 3.4: State Category

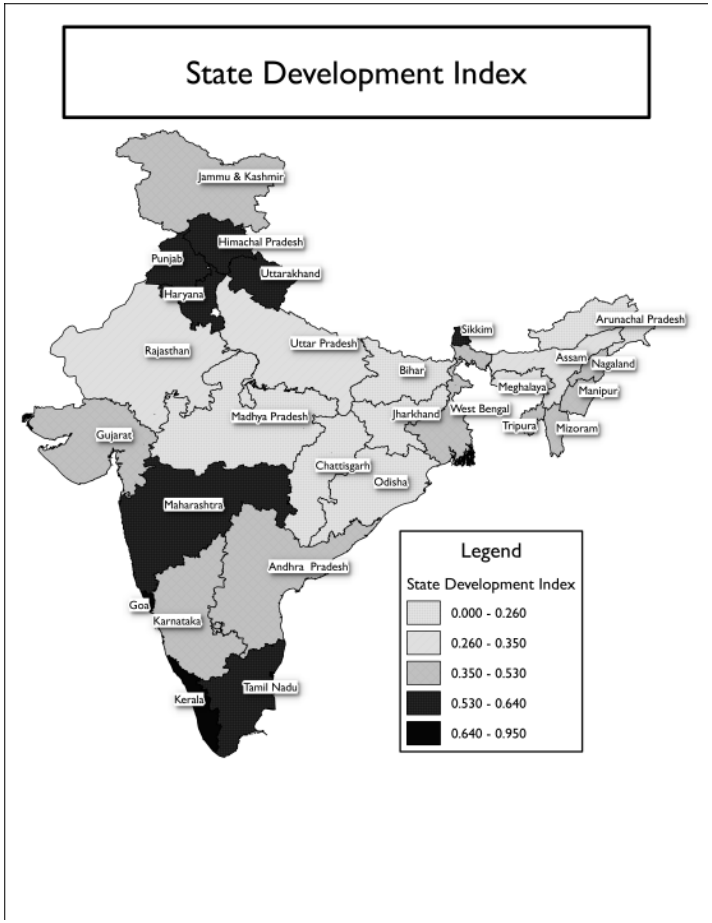


Figure 3.5: State Development Index

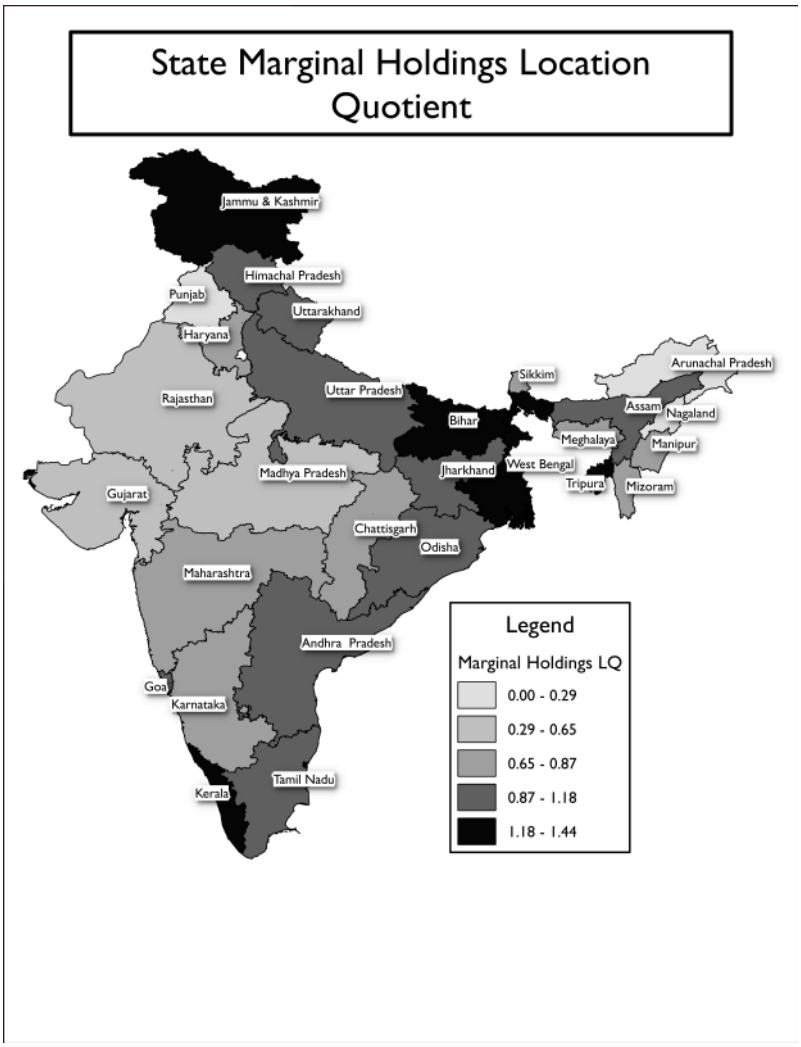


Figure 3.6: State marginal Holdings Location Quotient

of these measures take note of complexities in policy implementation, which is important according to McNeal et al. (2003) and Sabatier and Mazmanian (1979). McNeal et al. (2003) used measures for administrative complexity (legislative professionalism and professional networks). However, these measures are not transferable to the Indian context due to its vastly differing political and bureaucratic systems.

As discussed earlier in section 3, land administration is complex, and the challenges get exacerbated due to the sheer size of the endeavor. For example, the central Indian state of Madhya Pradesh (MP) is divided into fifty one districts having more than fifty thousand villages. These villages have around nine million operational holdings (Agricultural Census Division, 2014, February 28, Table 20), and the number of revenue holdings[15] (those that are directly linked to land records) is even larger — forty districts have more than twenty two million land records[16]. The administration of such large numbers is extremely labor intensive which can be gauged from the manpower employed for the work. MP's "Commissioner, Land Records and Settlement (MP) (CoLR (MP))" office employs nineteen thousand personnel, of which twelve thousand are village level functionaries. Further, there exist ten thousand vacancies that are required to be urgently filled[16].

Considering these peculiarities of the study context, there emerges the need for a contextual measure that directly proxies administrative complexity. The 2010-11 Agricultural Census (Agricultural Census Division, 2014, February 28) is a good source for this information. The Agricultural Census is a five-yearly exercise that gathers data on land holdings — their number, size, and

distribution patterns at the district level and aggregates it at the state level.

The 2010–11 Agricultural Census (Agricultural Census Division, 2014, February 28) provides such a measure of complexity in the form of the "number of marginal holdings". A marginal holding is defined as an operational holding of less than 1 hectare in size. We find this to be a good proxy for administrative complexity because a larger number of marginal holdings in a district translates into more work for the land administration staff. Also, as an "operational holding" may be comprised of multiple "revenue holdings", the complexity increases manifold. Instead of using the value directly, we identify the degree to which a state has more marginal holdings compared to the entire country. To do so, we use "location quotients", which are a measure developed and used by regional economists to measure such concentration by localities (Higgins & Savoie, 1997, pp 156).

We define the location quotient of the number of marginal holdings in the state: $\text{NumMarginal\_LQ}_{\text{STATE}}$ as:

Number Marginal_Holdings Location Quotient ($\text{NumMarginal\_LQ}_{\text{STATE}}$) is the state's *proportion* of marginal holdings *normalized* to the national *proportion* of marginal holdings. This identifies the intensity of marginal holdings in the state's total land holdings, compared nationally.

A value of unity means that the state has the same proportion of marginal holdings as the country, while a number greater than unity means that the state has a larger proportion of marginal holdings than the national norm. A value less than unity signifies that marginal holdings make up a smaller fraction of the state's land holdings. Mathematically:

$$NumMarginal\_LQ_{STATE} = \frac{State\,Num\,Marginal\,Holdings/State\,Num\,All\,Holdings}{National\,Num\,Marginal\,Holdings/National\,Num\,All\,Holdings}$$

The values of the location quotient of the number of marginal holdings are given in column 8 ("Num Marg Hold") of Table 3.3 on page 103. Our testable hypothesis at the state level is:

TH3_STATE: *Controlling for tenure type and state development, the probability of adoption of the NLRMP at the state level is inversely proportional to the state's number of marginal holdings expressed as a Location Quotient (* $NumMarginal\_LQ_{STATE}$ *).*

5.3.4 Vertical Diffusion (State)

One of the policy diffusion mechanisms, namely coercion (section 2.1.1) discusses the presence of incentives or dis-incentives to policy adoption. Scholars have found that federal policies or the presence of "vertical" diffusion can impact policy adoption (Allen et al., 2004; Eyestone, 1977; Gray, 1973; Karch, 2006; Lyson, 2016; Nicholson-Crotty, 2009; Shipan & Volden, 2012; Welch & Thompson, 1980). These external effects may be manifested through the presence of mandates or provision of financial support to the states if they adopt certain policies. Allen et al. (2004) looked at whether states received federal financial incentives to create certain policies, while Lyson (2016) looked at the amount of federal funding per

student in her study of farm to school programs.

The NLRMP provides us a natural indicator for such external influence in the form of additional funding for "special category states". The concept of "special category states" has been explained in sections 3.1 on page 88 and 4.4. In the case of the NLRMP, the central government provides up to half the funding for general category states, but in case of the special category states, the funding goes up to ninety percent. Thus, this aspect of the policy can help us uncover the effects of federal support.

This variable is given in column 2 ("SCS") of Table 3.3 on page 103. For analytical purposes, it is coded as a dichotomous variable ("Y" (yes): 1, "N" (no): 0). Our testable hypothesis at the state level is:

TH4_STATE: *Controlling for tenure type, state development and administrative complexity, the probability of adoption of the NLRMP at the state level is more if the state is a special category state, compared to a general category state.*

5.4 District Level Independent Variables and Testable Hypotheses

At the district level, we are interested in identifying the factors that impact selection of a district for program adoption. We control for the state level factors while doing so. Three testable hypotheses are identified at the district level. These have to do with policy salience (or the motivation to adopt), resource availability and obstacles to implementation. These three hypotheses, along with their associated variables are now described.

5.4.1 Policy Salience (District)

As discussed in section 2.2.1, if an issue is important to a large section of people or if it impacts them directly, then there is a greater motivation to solve this. The policy under investigation (the NLRMP) deals with the highly emotive issue of land reforms which impacts a large section of the population, and is targeted at rural landholdings, in a largely agrarian society. It is expected that the policy will bring transparency in land administration and reduce the disputes and inefficiencies in the system due to the lack of proper record-keeping. Thus, the beneficiaries of this land records modernization policy will be largely rural, especially the agricultural workforce. Thus, *ceteris paribus*, *more* rural districts with a *higher* agricultural workforce should have a higher propensity to adopt the policy.

The salience of the policy at the district level is proxied by capturing these two aspects: (a) its rural area ($\mathsf{AreaRural}_{\mathrm{DISTRICT}}$), and (b) its agricultural workforce (TWFRAGRI_LQ_Tot). The 2011 census provides this information.

Using $\mathsf{AreaRural}_{\mathrm{DISTRICT}}$, a testable hypothesis is proposed at the district level:

TH1(A)_DISTRICT: *Controlling for state level factors, the probability that a district will be selected for adoption of the NLRMP is directly proportional to the district's rural area (*$\mathsf{AreaRural}_{\mathrm{DISTRICT}}$*).*

For the agricultural workforce, we calculate a location quotient that is defined as:

Total Agricultural Workforce Location Quotient (TWFRAGRI_LQ_Tot$_{\mathrm{DISTRICT}}$)

is the district's *proportion* of total agricultural workforce *normalized* to the state's *proportion* of agricultural workforce. This captures the intensity of agricultural workforce in the district, compared statewide. The Agricultural Workforce includes both cultivators (who own their land) and agriculture labor (who work on other people's lands). Further, both full-time workers and marginal workers (who work six months or more) are included. Mathematically:

$$TWFRAGRI\_LQ\_Tot_{DISTRICT} = \frac{District\ Total\ Agricultural\ Workforce / District\ Total\ Workforce}{State\ Total\ Agricultural\ Workforce / State\ Total\ Workforce}$$

The testable hypothesis for agricultural workforce at the district level is:

TH1(B)_DISTRICT: *Controlling for state level factors, the probability that a district will be selected for adoption of the NLRMP is directly proportional to the district's total agricultural workforce(* $TWFRAGRI\_LQ\_Tot_{DISTRICT}$ *) .*

Next, both these variables are combined and tested together:

TH1(C)_DISTRICT: *Controlling for state level factors, the probability that a district will be selected for adoption of the NLRMP is directly proportional to the district's rural area (*$AreaRural_{DISTRICT}$*) and the district's total agricultural workforce (* $TWFRAGRI\_LQ\_Tot_{DISTRICT}$ *).*

5.4.2 Resource Availability (District)

This hypothesis is similar to the one at the state level (TH2_STATE in section 5.3.2 on page 105). However, no district level index akin to the composite index

(provided by the Raghuram Rajan Committee (Ministry of Finance, Government of India, 2013)) is available. We compute such a district level development index, called "$\text{DevIDX}_{\text{DISTRICT}}$" by relying on the format used by the committee, but using sub-component scores calculated at the district level. These district level sub-component scores are calculated using data from multiple sources. These sources include the 2011 census that provides data on the district's population, area, demographics and workforce characteristics. For district level health, education and other infrastructure data we use tables compiled by the Niti Aayog. More details on these data sources and the process of creating the district development index ($\text{DevIDX}_{\text{DISTRICT}}$) are given in appendix C.

The testable hypothesis for the district level development is:

TH2_DISTRICT: *Controlling for state level factors, and district level policy salience, the probability that a district will be selected for adoption of the NLRMP is directly proportional to its extent of development (* $\text{DevIDX}_{\text{DISTRICT}}$ *).*

5.4.3 Obstacles to Implementation (District)

For proxying implementation challenges at the district level, we again turn to the 2010-11 Agricultural Census to provide a contextual measure. We use the "number of operational holdings" from the 2010 Agricultural Census (Agricultural Census Division, 2014, February 28)[17]. This variable "$\text{NumberHoldings}_{\text{DISTRICT}}$" at the district level is used as a measure of implementation complexity. The testable hypothesis for implementation complexity at the district level is:

TH3_DISTRICT: *Controlling for state level factors, district level policy salience*

and the extent of development at the district level, the probability that a district will be selected for adoption of the NLRMP is inversely proportional to the number of landholdings in the district ($\mathsf{NumberHoldings}_{\mathsf{DISTRICT}}$ *).*

We next discuss the statistical analysis performed at the state and district levels.

6 The Statistical Analysis

We have two dependent variables, one for each level of analysis — state and district. At the state level, we use the proportion of the state that has adopted the policy (PropAdoption) as the dependent variable:

$$PropAdoption = \frac{Number\,of\,Districts\,Adopting\,the\,Policy}{Number\,of\,Districts\,in\,the\,State}$$

At the district level, the dependent variable (PolicyAdoption) is binary ('0' or '1'), signifying whether the district was selected for implementation ('1') or not ('0'). Both these dependent variables are bounded to the interval $[0, 1]$.

Ordinary Least Squares (OLS) regression techniques are discouraged when the dependent variable is binary or a proportion. Logit or probit regression methods are preferred (Agresti & Finlay, 2009; James et al., 2013; Kabacoff, 2015) in such cases. For this analysis, binary logistic regression is used as the preferred technique.

As discussed in section 5.1, our period of analysis is 2008–14. The data used for the state level analyses is shown in Table 3.3 on page 103. The district level

analysis is run on data pooled from the NLRMP MIS for the analysis time period. All statistical analyses are done using the R statistical package (James et al., 2013; Kabacoff, 2015; R Core Team, 2017).

6.1 State Level Analyses

The summary statistics of the state level variables are given in Table 3.4, and Table 3.5 on the following page shows their correlations. We note from Table 3.4 that the average adoption of the policy is seventy percent, with variations from no-adoption in Uttarakhand to full-adoption in nine states. Figure 3.7 shows the adoption proportion by states.

Table 3.4: Summmary Statistics (State)

Statistic	N	Mean	St. Dev.	Min	Max
Proportion Adoption	23	0.696	0.318	0.000	1.000
Special Category State	23	0.304	0.470	0	1
State Dev Idx	23	0.447	0.166	0.210	0.850
Marg Holding Conc	23	0.887	0.343	0.050	1.440

The state development index varies from a low of 0.21 to a high of 0.85 ($mean = 0.45,\ sd = 0.17$). The distribution of the state development index is shown in

Table 3.5: Correlation table (State)

	Proportion Adoption	Special Category State	State Dev Idx	Marg Holding Conc
Proportion Adoption	1.000			
Special Category State	-0.285	1.000		
State Dev Idx	-0.055	0.016	1.000	
Marg Holding Conc	-0.154	-0.066	0.044	1.000

the density histogram and kernel density plot in figure 3.8a on page 122. The location quotient of the number of marginal holdings varies from 0.05 (state has a very small proportion of marginal holdings copared nationally) to 1.44 (state's proportion of marginal holdings is almost one and a half times the national proportion) ($mean = 0.89$, $sd = 0.34$). Figure 3.8b on page 122 shows the density histogram of the marginal holdings location quotient overlaid with its kernel density plot. No significant correlation (> 0.6) between our variables is discernible in Table 3.5.

A cross-tabulation of the state category with the system of tenure (Table 3.6

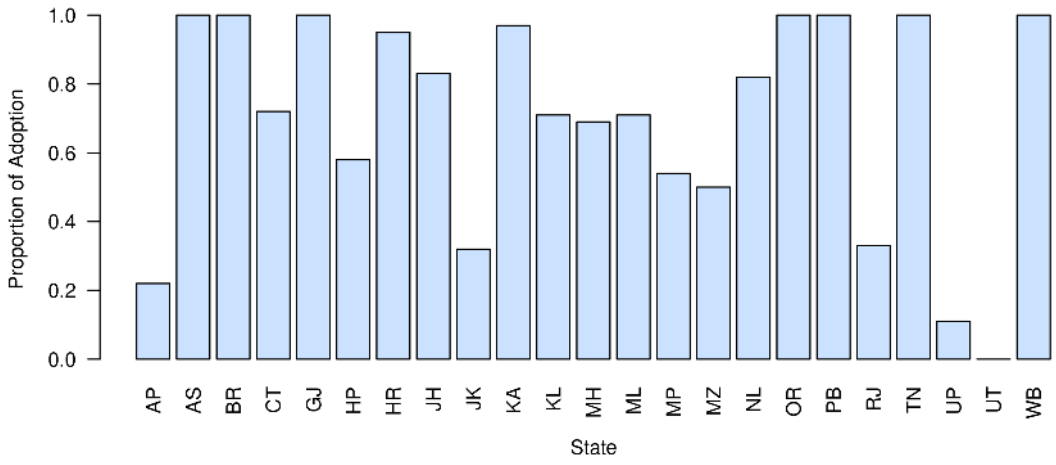
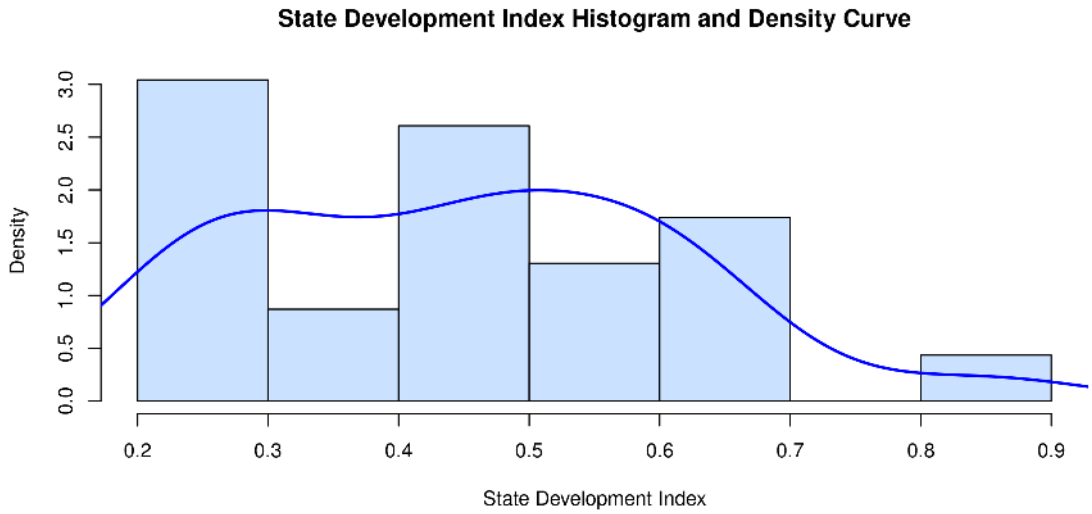


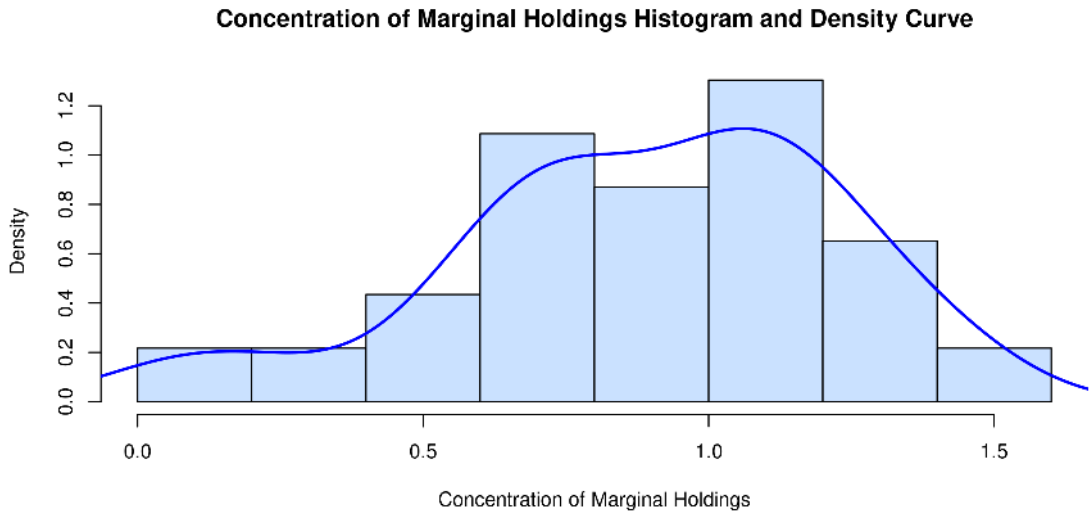
Figure 3.7: State-wise Adoption Proportion (2008–14)

Table 3.6: Cross-tabulation of State Category with type of tenure

	State Category	Tenure Type			Row Total
		raiyatwari	Mahalwari	Zamindari	
Count	General	7	5	4	16
Row Percent		43.75%	31.25%	25.00%	
Column Percent		53.85%	83.33%	100.00%	
Total Percent		30.43%	21.74%	17.39%	69.57%
Count	Special	6	1	0	7
Row Percent		85.71%	14.29%	0.00%	
Column Percent		46.15%	16.67%	0.00%	
Total Percent		26.09%	4.35%	0.00%	30.43%
Column Total		13	6	4	23
Column Percent		56.52%	26.09%	17.39%	



(a) Histogram and Kernel Density Plot of State Development Index



(b) Histogram and Kernel Density Plot of State Concentration of Marginal Holdings

Figure 3.8: Density histograms & kernel density plots of state development index and marginal holdings

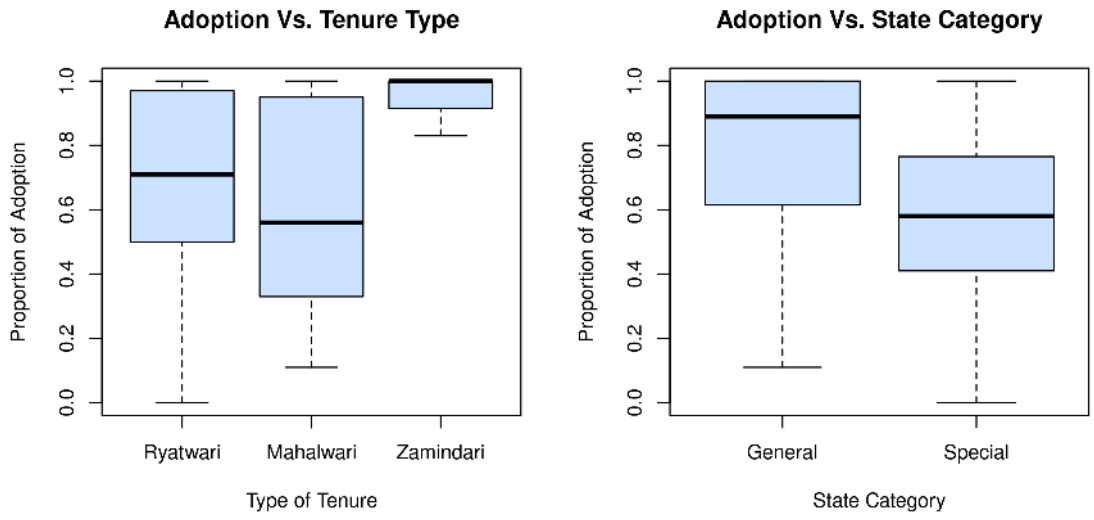


Figure 3.9: Boxplots of Proportion of Adoption Vs. Tenure Type and State Category

on page 121) provides an idea of how the states are distributed along these parameters. It shows that we have 7 special category states and 16 general category states. Of these, 13 states had the *raiyatwari*, 6 had the *mahalwari* and 4 had the *zamindari* system of land tenure. All states with the *zamindari* system belong to the general category, while only one special category state had the *mahalwari* system (Himachal Pradesh). Boxplots of adoption proportion against both the categorical variables (*TenureType* and *StateCategoryStatus*) are shown in Figure 3.9 on page 123.

6.1.1 Does Policy Adoption Depend only on State Category?

A question that could arise is whether policy adoption is *solely* dependent upon the category of the state? If this is the case, that is,the state's category is the *sole* predictor of adoption, then other factors may not be as important. A Welch's two-sample t-test is used to see if the means of the adoption proportion varies between the general ($mean\ adoption = 0.754$) and special ($mean\ adoption = 0.561$) category states (Kabacoff, 2015). The results of the test show that the null hypothesis of *no significant difference* between the means **cannot** be rejected (two sample t-test, $t = 1.32$, $df = 10.7$, $p = 0.215$). This rules out the category of the state being the *sole* determinant for policy adoption.

We next test each of the hypotheses TH1_STATE, TH2_STATE, TH3_STATE, and TH4_STATE in order. The results of the logistic regressions are shown in Table 3.7 on page 126 and discussed below. In these regressions, all the continuous variables have been normalized.

6.1.2 TH1_STATE: State Tenure Type ($TenureType_{STATE}$)

The model used to test this hypothesis is given in Equation 3.1. The results of this logistic regression are shown in Table 3.7, model 1.

$$PolicyAdoption = \beta_0 + \beta_1 TenureType_{STATE} \tag{3.1}$$

We see that state's tenure type for *mahalwari* system is statistically significant

at the 10% level, but negative. This means that the odds of adopting the policy are 0.32 (or around 32%) compared to a similar state that had a *raiyatwari* system.

6.1.3 TH2_STATE: State Development Index ($\text{DevIDX}_{\text{STATE}}$) controlling for $\text{TenureType}_{\text{STATE}}$

We add the state's development index to the model in equation 3.1 on the preceding page, giving us the one in equation 3.2. The results of this logistic regression are shown as model 2 in Table 3.7.

$$PolicyAdoption = \beta_0 + \beta_1 TenureType_{STATE} + \beta_2 DevIDX_{STATE} \tag{3.2}$$

All the coefficients except the intercept lose their significance, however the magnitudes are roughly similar to those of model 1.

6.1.4 TH3_STATE: State Concentration of Marginal Holdings ($\text{NumMarginal}_{\text{STATE}}$) controlling for $\text{TenureType}_{\text{STATE}}$ and $\text{DevIDX}_{\text{STATE}}$

The relative concentration of the number of marginal holdings is used as a proxy for implementation complexity. This variable, $\text{NumMarginal\_LQ}_{\text{STATE}}$ is added as a predictor to the model in equation 3.2, giving us equation 3.3. Model 3 in Table 3.7 on the following page shows the results of this logistic regression.

Table 3.7: The Logistic Models (State)

	Proportion of Policy Adoption				
	(1)	(2)	(3)	(4)	(5)
Mahalwari	−1.150*	−0.901	−1.200	−1.260	−0.547
	(0.652)	(0.711)	(0.738)	(0.813)	(0.873)
Zamindari	2.330	2.850	3.740*	3.620*	4.530**
	(1.740)	(1.850)	(1.790)	(1.910)	(1.900)
State Dev Idx		0.370	0.451	0.431	0.923*
		(0.406)	(0.380)	(0.401)	(0.481)
Special Category State				−0.187	0.085
				(0.891)	(0.989)
State Marg Holding Conc			−0.943**	−0.922*	−0.900*
			(0.435)	(0.453)	(0.451)
State Dev X Category					−2.660*
					(1.400)
Constant	0.939*	0.828*	0.919*	0.991	0.631
	(0.455)	(0.476)	(0.470)	(0.593)	(0.588)
N	23	23	23	23	23

*p < .1; **p < .05; ***p < .01

$$PolicyAdoption = \beta_0 + \beta_1 TenureType_{STATE} + \beta_2 DevIDX_{STATE} + \beta_3 NumMarginal\_LQ_{STATE} \quad (3.3)$$

This model is interesting for two reasons. First, the state's concentration of marginal holdings is statistically significant at the 5% percent level, with the coefficient being negative, as expected. This can be interpreted as "for every 1 standard deviation increase in the concentration of marginal holdings from the mean, the odds of policy adoption reduce 39% from the mean". Secondly, the *zamindari* tenure type now becomes statistically significant at the 10% level. The can be interpreted as "compared to a *raiyatwari* system, a state that had a *zamindari* tenure is almost 42 times more likely to adopt the policy".

6.1.5 TH4_STATE: Federal Support Effect (SCS) controlling for Tenure $Type_{STATE}$, $DevIDX_{STATE}$ and $NumMarginal_{STATE}$

As discussed in section 5.3.4, we use the state's category as a proxy for federal support. This model is shown in equation 3.4 and the regression results in column 4 of Table 3.7 on the previous page. These results are quite similar to model 3 (equation 3.3), except for slight differences in the magnitudes of the coefficients. However, the constant (intercept term) loses its statistical significance.

$$PolicyAdoption = \beta_0 + \beta_1 TenureType_{STATE} + \beta_2 DevIDX_{STATE} + \beta_3 NumMarginal_{STATE} + \beta_4 SCS \quad (3.4)$$

However, considering that the state's category is also a measure of development, we interact SCS with $DevIDX_{STATE}$ to disentangle the effects of state level development and federal grants:

$$PolicyAdoption = \beta_0 + \beta_1 TenureType_{STATE} + \beta_2 DevIDX_{STATE} + \beta_3 NumMarginal_{STATE} + \beta_4 SCS + \beta_5 DevIDX_{STATE} \times SCS \quad (3.5)$$

The results of this model (equation 3.5) are shown under column 5 of Table 3.7 on page 126. This model is discussed further in section 6.1.7. Before that, we compare the different models and identify the best one in section 6.1.6.

6.1.6 Comparing the State Level Models

The models in equations 3.1, 3.2, 3.3, 3.4, and 3.5 are nested models. To identify if any model with more variables is better than one with lesser variables, we resort to a χ^2 ANOVA test between these models. The results of this test are shown in Table 3.8 on the following page. From the table, we note that the fifth

Table 3.8: χ^2 ANOVA test between all the State Logistic Models

Model	Residual Df	Residual Dev	Df	Deviance	Pr(>Chi)
1	20	258			
2	19	248	1	9.700	0.303
3	18	192	1	56.100	0.013**
4	17	191	1	0.455	0.823
5	16	145	1	46.600	0.024**

*p < .1; **p < .05; ***p < .01

model (equation 3.5) (involving an interaction between $\mathsf{DevIDX}_{\mathrm{STATE}}$ and SCS) is the best. It has a residual deviance of 145 and the χ^2 test is statistically significant at the 5% level. Therefore, we select the fifth model (3.5) to discuss further.

6.1.7 Discussing the State Models

As discussed earlier, the model proposed for testing the TH4_State hypothesis is the best in terms of its fit. In this section, we discuss the model coefficients and their statistical significance and tie it with the hypotheses outlined earlier.

1. The *zamindari* tenure type is now statistically significant at the 5% level. Its magnitude is also significantly increased, with the odds-ratio now increasing to 93 is to 1. Thus, a state that had a *zamindari* tenure is almost ninety three times more likely to adopt the policy than a state that had the *raiyatwari* system. This *supports* the "policy salience" hypothesis.

2. The state's development index is statistically significant at the 10% level. For every 1 standard deviation increase in the state's development, the odds-ratio of policy adoption increase to 2.5 is to 1, *supporting* the "resource availability" hypothesis.

3. The state's proportion of marginal holdings is also statistically significant at the 10% level and in the expected direction. For every one standard deviation increase in the concentration of marginal holdings, the odds of policy adoption drop to almost 41%. This supports the "*implementation complexity*" hypothesis.

4. The category of the state by itself is not a statistically significant predictor of adoption. However, when the category is interacted with the State Development Index, we find statistically significance at the 10% level. We can interpret this as follows:

    a) For a General Category State, SCS is 0. So, the interaction term vanishes, and the effect of the Development Index is as before — *log odds* of 0.923, translating into an odds ratio of 2.5 : 1. Propensity to adopt policy *increases* with increased development.

    b) For a Special Category State, SCS is 1. For such states, the total effect of the Development Index comes to $0.923 - 2.660 = -1.740$. This large negative *log odds* translates to an *odds ratio* of 0.176. For every one standard deviation increase in the state's development index, the odds of policy adoption drop to 17%. This means that in case of Special Category States, the propensity to adopt policy *reduces* with increased

development.

We thus do not find support for the presence of "vertical diffusion". This aspect is discussed further in section 7.

5. The intercept term is positive, but not statistically significant.

This model (equation 3.5 on page 128) is now used as the base model when testing for factors that impact district selection for policy adoption.

6.2 District Level Analyses

The summary statistics of the district level variables are given in Table 3.9, while Table 3.10 on page 135 shows their correlations. From Table 3.9, we note the average adoption of the policy is sixty seven percent.

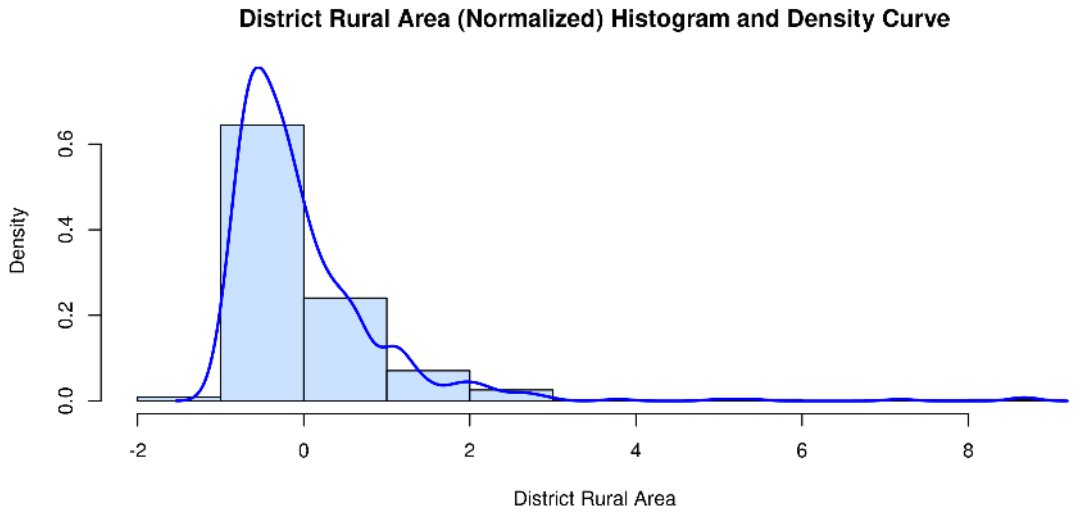
Table 3.9: Summary Statistics (District)

Statistic	N	Mean	St. Dev.	Min	Max
Policy Adopted	578	0.670	0.471	0	1
Dist Dev Idx	578	0.460	0.222	0.000	0.988
Dist Num Holdings (number)	578	236,719	187,021	4,120	982,314
Dist Rural Area (km^2)	578	5,065	4,632	234	45,382
Dist Tot Agri Worker Conc	578	1.070	0.308	0.090	2.780

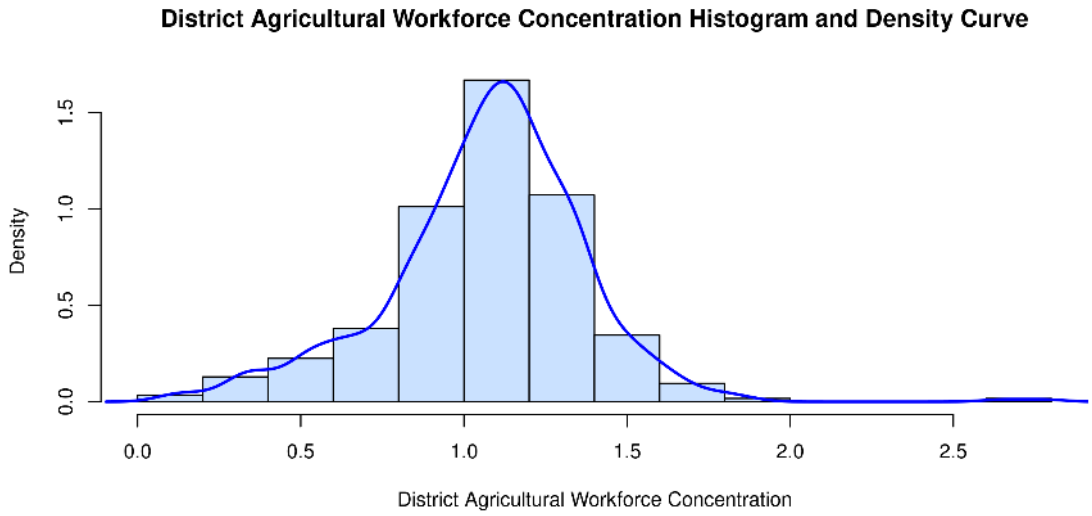
The rural areas of the districts show a left skewed distribution ($mean = 5,065,\ sd = 4,632,\ min = 234,\ max = 45,382$). This skew can also be seen in the density histogram and kernel density plot of the (normalized) district rural area in Figure 3.10a on the next page. The concentration of agricultural workers varies from 0.09 (virtually non-existent) to 2.78 (highly agricultural) ($mean = 1.070,\ sd = 0.308$). From Figure 3.10b on the following page (which shows the density histogram and kernel density plot), we note that this follows a roughly normal distribution. The mean development index of the district is 0.46, with a standard deviation of 0.22. Figure 3.11a on page 134 shows the density histogram and kernel density plot of the district development index. The number of holdings is highly left skewed ($mean = 236,719,\ sd = 187,021,\ min = 4,120,\ max = 982,314$) as can be seen from the density plots in Figure 3.11b on page 134.

The correlation table (Table 3.10 on page 135) does not show any significant correlation (> 0.6) except between the State Development Index ($\text{DevIDX}_{\text{STATE}}$) and the District Development Index ($\text{DevIDX}_{\text{DISTRICT}}$). However, this correlation is to be expected and is discussed in section 7.

We next test each of the hypotheses: TH1(A, B, C)_DISTRICT, TH2_DISTRICT, and TH3_DISTRICT in order. The results of the logistic regressions shown in Table 3.11 on page 137 are discussed. In these regressions, all the continuous variables have also been normalized.

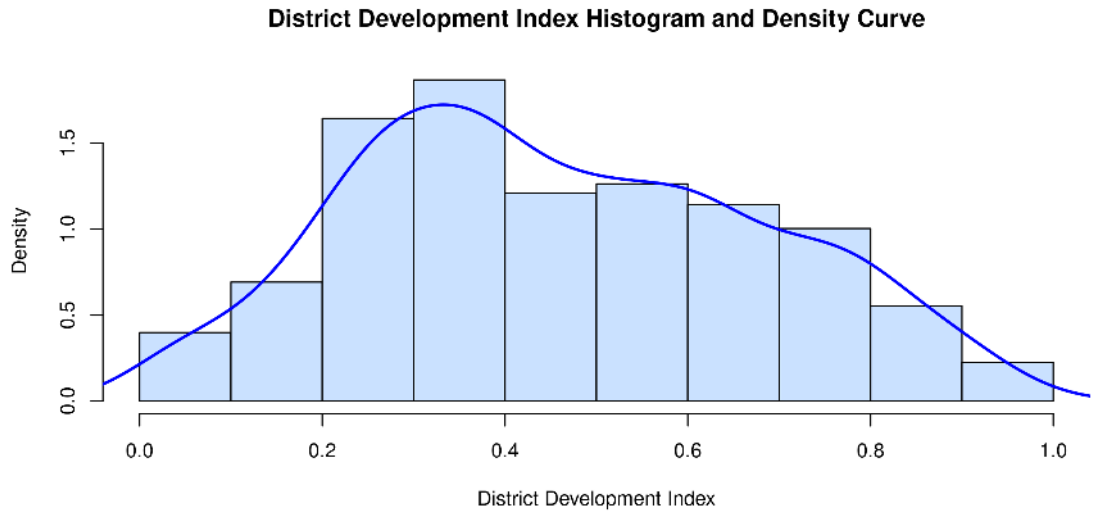


(a) Histogram and Kernel Density Plot of District Rural Area

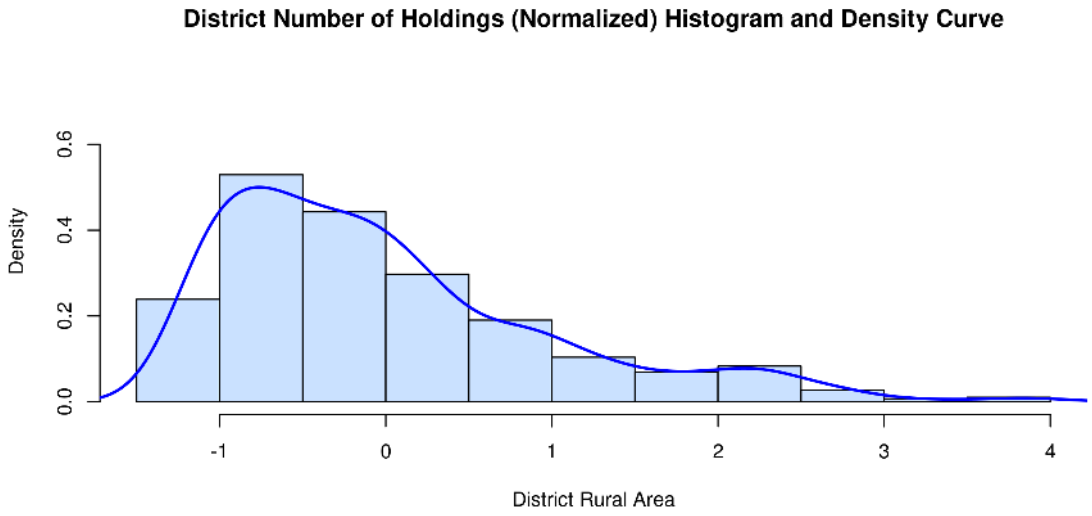


(b) Histogram and Kernel Density Plot of District Agricultural Worker Concentration

Figure 3.10: Density histograms & kernel density plots of district rural area and agricultural worker concentration



(a) Histogram and Kernel Density Plot of District Development Index



(b) Histogram and Kernel Density Plot of District's Number of Holdings

Figure 3.11: Density histograms & kernel density plots of district development index and number of holdings

Table 3.10: Correlation table (District)

	Policy Adopted	Special Category State	State Dev Idx	Marg Holding Conc (State)	Dist Dev Idx	Dist Num Holdings	Dist Rural Area	Dist Tot Agri Workforce
Policy Adopted	1.000							
Special Category State	-0.077	1.000						
State Dev Idx	0.052	0.058	1.000					
Marg Holding Conc (State)	-0.096	0.023	-0.089	1.000				
Dist Dev Idx	0.056	0.000	0.763	-0.316	1.000			
Dist Num Holdings	-0.106	-0.420	0.098	0.331	-0.061	1.000		
Dist Rural Area	-0.103	-0.160	0.017	-0.196	0.045	0.300	1.000	
Dist Tot Agri Workforce	-0.024	0.002	0.142	-0.023	-0.266	0.046	0.053	1.000

6.2.1 Running State Model on District Data set

Initially, we run the state level interaction model (TH4_STATE) discussed in section 6.1 (equation 3.5 on page 128) on the district level data. The results of this are shown in column 1 of Table 3.11 on the following page. This essentially confirms the model and has similar magnitude. However, there are a few points worth noting.

1. The *mahalwari* tenure type is now statistically significant at the 5% level. Its coefficient is negative which can be interpreted as: the odds of states that had the *mahalwari* type of tenure adopting the policy are 0.56 to 1, as compared to the *raiyatwari* states..

2. The statistical significance levels of the *zamindari* tenure, the state's development index and the marginal holdings concentration now increases to 1%.

3. While the category of the state continues to be statistically insignificant by itself, its interaction with the state development index is statistically significant at the 1% level.

4. The McFadden R^2 (a measure of model fit) is almost 0.3.

From now on, we will refer to this state level model as $Model_{STATE}$.

Table 3.11: The Logistic Models (District)

	Policy Adoption					
	(1)	(2)	(3)	(4)	(5)	(6)
Mahalwari	−0.576**	−0.774**	−0.615**	−0.810***	−0.898***	−1.080***
	(0.291)	(0.309)	(0.293)	(0.311)	(0.316)	(0.329)
Zamindari	4.520***	4.280***	4.510***	4.270***	4.390***	4.520***
	(0.633)	(0.644)	(0.634)	(0.644)	(0.649)	(0.677)
State Dev Idx	0.875***	0.810***	0.900***	0.838***	0.422**	0.531**
	(0.153)	(0.157)	(0.155)	(0.160)	(0.211)	(0.219)
Special Category State	0.538	0.312	0.524	0.304	0.312	−0.454
	(0.355)	(0.375)	(0.356)	(0.375)	(0.374)	(0.423)
Marg Holding Conc State	−0.823***	−0.908***	−0.831***	−0.915***	−0.796***	−0.570***
	(0.137)	(0.146)	(0.138)	(0.147)	(0.149)	(0.155)
District Rural Area		−0.264**		−0.264**	−0.251**	−0.080
		(0.114)		(0.113)	(0.114)	(0.115)
District Agricultural Workforce			−0.579	−0.590		
			(0.364)	(0.370)		
District Dev Idx					0.536***	0.494**
					(0.200)	(0.203)
District Num Oper Holdings						−0.596***
						(0.154)
State Dev X Category	−2.520***	−2.370***	−2.530***	−2.380***	−2.450***	−2.610***
	(0.441)	(0.444)	(0.443)	(0.446)	(0.452)	(0.454)
Constant	0.418**	0.584***	1.060**	1.240***	0.622***	0.863***
	(0.210)	(0.225)	(0.458)	(0.470)	(0.228)	(0.243)
McFadden R^2	0.297	0.305	0.301	0.309	0.315	0.336
N	578	578	578	578	578	578
Log Likelihood	−258.000	−255.000	−256.000	−254.000	−251.000	−243.000
AIC	529.000	526.000	529.000	525.000	520.000	507.000

*p < .1; **p < .05; ***p < .01

6.2.2 TH1({A, B,C})_DISTRICT: Rural Measures ($\text{AreaRural}_{\text{DISTRICT}}$ and $\text{TWFRAGRI\_LQ\_Tot}_{\text{DISTRICT}}$) controlling for State Level Factors

TH1(A)_DISTRICT This model can be written as:

$$PolicyAdoption = Model_{STATE} + \beta_6 AreaRural_{DISTRICT}$$
$$where \qquad Model_{STATE}: \quad \text{State Level Model of eq. 3.5} \qquad (3.6)$$

The results of this model are shown in column 2 of Table 3.11 on the previous page. We note that the coefficient of $\text{AreaRural}_{\text{DISTRICT}}$ is statistically significant at the 5% level and negative (opposite the hypothesized direction). Its interpretation is that for 1 standard deviation increase of the district's rural area, the odds of selecting the district for policy adoption drop by almost 23%. This result will be discussed further in section 7.

TH1(B)_DISTRICT This model incorporates the district's agricultural workforce ($\text{TWFRAGRI\_LQ\_Tot}_{\text{DISTRICT}}$) and its equation is:

$$PolicyAdoption = Model_{STATE} + \beta_7 TWFRAGRI\_LQ\_Tot_{DISTRICT} \quad (3.7)$$

The results of this model are shown under column 3 of Table 3.11 on the preceding page. The district's agriculture workforce concentration has a

negative, but not statistically significant coefficient. This result will also be discussed further in section 7.

TH1(C)_DISTRICT This model incorporates both the area and the workforce variables and can be specified as:

$$PolicyAdoption = Model_{STATE} + \beta_6 AreaRural_{DISTRICT} + \beta_7 TWFRAGRI\_LQ\_TOT_{DISTRICT} \quad (3.8)$$

These results shown in column 4 of Table 3.11 on page 137 are virtually the same as those from the models of equations (3.6) and 3.7, that is the coefficients of both $\text{AreaRural}_{\text{DISTRICT}}$ and $\text{TWFRAGRI\_LQ\_Tot}_{\text{DISTRICT}}$ are negative. The coefficient of $\text{AreaRural}_{\text{DISTRICT}}$ is statistically significant at the 5% level, while that for $\text{TWFRAGRI\_LQ\_Tot}_{\text{DISTRICT}}$ is statistically insignificant at the conventional levels.

Comparing the TH1({A,B,C})_DISTRICT models: The models listed above (equations 3.6, 3.7 and 3.8) are compared to identify whether the concentration of agricultural workforce ($\text{TWFRAGRI\_LQ\_TOT}_{\text{DISTRICT}}$) adds any value. All three models have similar values for the McFadden R^2, as well as the AIC (a goodness of fit measure based on information theory (James et al., 2013; Kabacoff, 2015)). Therefore, we perform a χ^2 ANOVA test between these models. This test is performed between the full model (eq. 3.8) and each of the reduced models (equations 3.6 & 3.7) separately. The results of the test between the models

Table 3.12: χ^2 ANOVA test between the District Logistic Models 2 and 4 (Rural Area vs Rural Area & Workforce)

Model	Residual Df	Residual Dev	Df	Deviance	Pr(>Chi)
2	570	510			
4	569	507	1	2.560	0.109

*p < .1; **p < .05; ***p < .01

Table 3.13: χ^2 ANOVA test between the District Logistic Models 3 and 4 (Workforce vs Rural Area & Workforce)

Model	Residual Df	Residual Dev	Df	Deviance	Pr(>Chi)
3	570	513			
4	569	507	1	5.860	0.015**

*p < .1; **p < .05; ***p < .01

of TH1(A)_DISTRICT and TH1(C)_DISTRICT is shown in Table 3.12 on the preceding page, which shows that the additional variable does not add value. Table 3.13 on the previous page shows the test results between TH1A_DIST and TH1(C)_DIST. From this table, we note that model 4 (TH1(C)_DISTRICT) (eq. 3.8 on page 139) is better than model 3 (TH1(B)_DISTRICT specified by eq. 3.7 on page 138) as the χ^2 test is statistically significant at the 5% level. The only difference between these two models is the presence of the $\text{AreaRural}_{\text{DISTRICT}}$ variable. Thus, we keep the model of equation 3.6 ($\text{AreaRural}_{\text{DISTRICT}}$) only for future analysis.

6.2.3 TH2_DISTRICT: District Development Index ($\text{DevIDX}_{\text{DISTRICT}}$) controlling for $\text{AreaRural}_{\text{DISTRICT}}$ and State Level Factors

This model builds off equation 3.6 on page 138 by accounting for the district's development index. Model specification is given in equation 3.9:

$$\begin{aligned} PolicyAdoption = Model_{STATE} + \beta_6 AreaRural_{DISTRICT} \\ + \beta_8 DevIDX_{DISTRICT} \end{aligned} \tag{3.9}$$

The results of this model are given in column 5 of Table 3.11 on page 137. We note that the district's development index ($\text{DevIDX}_{\text{DISTRICT}}$) is statistically significant at the 1% level and positive as expected by the hypothesis. Interpreting the coefficient: a one standard deviation increase in the district's development index results in the odds of selecting the district for policy adoption increasing

by 71%.

6.2.4 TH3_DISTRICT: Obstacles to Implementation ($NumberHoldings_{DISTRICT}$) controlling for $DevIDX_{DISTRICT}$, $AreaRural_{DISTRICT}$, and State Level Factors

This model builds off equation 3.9 by accounting for the number of land holdings in the district, which is a proxy for implementation complexity. The model is specified in equation 3.10:

$$PolicyAdoption = Model_{STATE} + \beta_6 AreaRural_{DISTRICT} + \beta_8 DevIDX_{DISTRICT} + \beta_9 NumberHoldings_{DISTRICT} \quad (3.10)$$

The results of this model are given in column 6 of Table 3.11 on page 137. The district's number of operational holdings ($NumberHoldings_{DISTRICT}$) is statistically significant at the 1% level and negative as expected by the hypothesis. The log odds coefficient of -0.596 can be interpreted in terms of odds ratio as: a one standard deviation increase in the district's number of holdings results in the odds of selecting the district for policy adoption reducing by 45%. This model is discussed further in section 6.2.6. Next, we compare the different models to identify the best one in section 6.2.5.

Table 3.14: χ^2 ANOVA test between all the District Logistic Models

Model	Residual Df	Residual Dev	Df	Deviance	Pr(>Chi)
1	571	515			
2	570	510	1	5.840	0.016**
5	569	502	1	7.420	0.006***
6	568	487	1	15.300	0.0001***

*p < .1; **p < .05; ***p < .01

6.2.5 Comparing the District Level Models

The models specified by the equations 3.6, 3.9, and 3.10 are nested models. To identify the best model amongst these we have the following ways: (a) compare their McFadden R^2 values (Agresti & Finlay, 2009), (b) compare their AIC (Akaike Information Criterion) (James et al., 2013; Kabacoff, 2015), or (c) compare the models using a χ^2 ANOVA (Kabacoff, 2015) A larger McFadden R^2 value indicates a better fit, while in case of AIC, lower is better. When comparing using χ^2 ANOVA, we want the difference between the model deviances to be statistically significant for a small change in the degrees of freedom.

McFadden R^2 of the models in Table 3.11 varies from slightly less than 0.30 (column 1) to around 0.34 (for column 6). Thus, the model specified by equation 3.10 seems to be the best by this criterion.

AIC of the models in Table 3.11 varies from 529 to 507. Since smaller AIC indi-

cates better fit, we can again choose the model specified by equation 3.10.

χ^2 **ANOVA** test results are given in Table 3.14. We only compare the state level model and those specified by equations 3.6 on page 138, 3.9 on page 141, and 3.10 on page 142. Although each of the models is seen to be better than the preceding one (also validated by the increasing McFadden R^2 and reducing AIC values), the model specified by equation 3.10 is the best. The value of the χ^2 test is statistically significant at the less than 1% level ($p = 0.0001$). We select the fifth model (3.10) and discuss it further in section 6.2.6.

6.2.6 Discussing the District Models

As discussed in section 6.2.5, the model specified by equation 3.10 on page 142 for testing the TH3_DISTRICT hypothesis is the best. In this section, we discuss the model coefficients and their statistical significance, tying it with the hypotheses outlined earlier.

1. The *mahalwari* tenure type is now statistically significant at the 1% level. Its coefficient has increased in magnitude, while still being negative. Thus, the odds of states that had the *mahalwari* type of tenure adopting the policy are further reduced by almost 66%, as compared to the *raiyatwari* states.

2. The *zamindari* tenure keeps its statistical significance at 1% and its effect is almost the same. Compared to the *raiyatwari* states, the odds that states with the *zamindari* tenure will adopt the policy are almost 92 is to 1.

3. The state development in index is statistically significant at the 5% level and positive. However, compared to the state level model (column 1) of Table 3.11 on page 137, it magnitude has reduced. Now one standard deviation increase in the state's development increases the odds of policy adoption by 70% (odds ratio of 1.7 : 1) as against 140% (odds ratio of 2.4 : 1) earlier.

4. The category of the state changes sign but continues to be statistically insignificant by itself. However, its interaction with the state development index continues to be statistically significant at the 1% level and of similar magnitude.

5. The proportion of marginal holdings in the state keeps its statistical significance at 1%, but its magnitude reduces, while continuing to be negative. Now, a one standard deviation increase in the proportion reduces the odds of policy adoption by 43% as against 56% earlier.

6. The district rural area loses its statistical significance, as well being substantially diminished in magnitude.

7. The district development index is now reduced in magnitude and statistically significant at the 5% level (it was at the 1% level in the model specified by equation 3.9 on page 141). One standard deviation increase in the district's development increases the odds of policy adoption by 64% as against 71% earlier.

8. The number of holdings in the district is statistically significant at the 1% level and is negative in sign, as expected. A one standard deviation

increase in the number of holdings in the district reduces the odds of policy adoption by 45%.

9. As the state level model tells us, the category of the state by itself is not a statistically significant predictor of adoption. However, when the category is interacted with the State Development Index, we find it statistically significant at the 1% level (the state level model showed statistical significance at the 10% level). As before, this can be interpreted as:

    a) For a General Category State, SCS is 0. So, the interaction term vanishes, and the effect of the Development Index is as before — log odds of 0.531, translating into an odds ratio of 1.7:1. Propensity to adopt policy *increases* with increased development.

    b) For a Special Category State, SCS is 1. For such states, the total effect of the Development Index comes to 0.531 – 2.610 = –2.08. This large negative log odds translates to an odds ratio of 0.125. For every one standard deviation increase in the state's development index, the odds of policy adoption drop to around 13%. This means that in case of Special Category States, the propensity to adopt policy *reduces* with increased development.

10. The McFadden R^2 (a measure of model fit) is almost 0.3.

We next discuss these results are discussed in the context of the hypotheses laid out for policy adoption factors.

7 Discussion

7.1 The Results

This study attempts to find the factors that impact adoption of a land reforms policy (the NLRMP) in India. It was hypothesized in section 4 that policy adoption depends on four main factors: (a) policy salience, (b) resource availability, (c) presence of obstacles, and (d) external factors. These hypotheses were tested at two levels — the state and the district. The state level hypotheses were first tested using a state-level dataset ($N = 23$), and then the state-level full model was run on the district level dataset ($N = 578$). The district level hypotheses were tested on the district level dataset. The empirical analyses provided in section 6 largely confirm these hypotheses. Each hypothesis is discussed further below.

7.1.1 Policy Salience

We use different proxies for policy salience at the state and district levels. At the state level, the proxy for policy salience is the land tenure type existing in pre-independence India, while at the district level we use measures of district rural area and concentration of agricultural labor in the workforce.

At the state level, the hypothesis is confirmed. For the state level data, the tenure type variable for the *zamindari* (or landlord) tenure is statistically significant at the 5% level with a log-odds value of 4.53, which translates into an odds-ratio of $93 : 1$. When the same model is run on the district level data, the odds-ratio is

similar, while the statistical significance for the *zamindari* tenure increases to 1%. On the district level dataset, the log-odds for the *mahalwari* tenure varies from -0.576 to -1.080 (odds-ratios from $1:1.8$ to $1:3$) (models 1–6 in Table 3.11 on page 137). This tenure is also statistically significant (5% to 1%). Thus, the tenure type is an important predictor of which states will adopt the policy.

However, at the district level, we do not find support for this hypothesis. The district's rural area ($\text{AreaRural}_{\text{DISTRICT}}$) is used as a predictor in models 2, 4 and 6 (Table 3.11 on page 137). In models 2 and 4, we find the variable to be statistically significant at the 5% level. However, its direction is negative, implying that an increased rural area reduces the propensity of policy adoption. This result is the opposite of what had been hypothesized. However, in model 6, when we also add the district's number of operational holdings ($\text{NumberHoldings}_{\text{DISTRICT}}$) (our proxy for implementation complexity), the coefficient on $\text{AreaRural}_{\text{DISTRICT}}$ becomes statistically insignificant at the conventional levels with a much reduced magnitude (-0.080 from -0.264). Our other proxy for policy salience, $\text{TWFRAGRI\_LQ\_Tot}_{\text{DISTRICT}}$ also has negative magnitude, however it is also not statistically significant (model 3 in Table 3.11 on page 137).

It is possible that the $\text{AreaRural}_{\text{DISTRICT}}$ instead of proxying policy salience is working as a (weak) proxy for implementation capability. The two variables: $\text{AreaRural}_{\text{DISTRICT}}$ and $\text{NumberHoldings}_{\text{DISTRICT}}$ are not substitutes for each other as can be seen by the rather weak correlation (0.30) (Table 3.10 on page 135) between them, and this is possibly why $\text{AreaRural}_{\text{DISTRICT}}$ loses statistical significance when $\text{NumberHoldings}_{\text{DISTRICT}}$ is added to the model.

7.1.2 Resource Availability

We use state and district level development indicators to proxy for the availability of resources at both the state and district levels respectively. On the state level data, the state's development index is statistically significant at the 10% level, with a log-odds value of 0.923 (odds-ratio of 2.5 : 1) (model 5, Table 3.7 on page 126). On the district level data, it is statistically significant at the 1% level with similar log-odds (model 1, Table 3.11 on page 137). However, as more variables are added, its magnitude reduces.

We find the district development index also to be a statistically significant predictor of which district will be adopted, in line with the specified hypotheses. This variable is statistically significant at the 1% level (odds-ratio of 1.7 : 1) (model 5, Table 3.11 on page 137), which drops to 5% (odds-ratio of 1.6 : 1) (model 6, Table 3.11 on page 137 as other variables are added.

There is a high correlation (0.76) (Table 3.10 on page 135) between these two indices, which could potentially result in biased estimates. We can detect multi-collinearity using the Variance Inflation Factors (VIF) statistic. If the VIF is greater than 2 or 2.5, then we need to be concerned about multi-collinearity (Kabacoff, 2015). The VIFs of our models incorporating the state and district level development indices are much less than 2 and hence we can rule out multi-collinearity.

7.1.3 Presence of Obstacles

The hypothesis for presence of obstacles is that *more* the obstacles, *lesser* the propensity to adopt. Complexity of implementation is considered as the prime barrier to adoption. Proxies for implementation complexity include the (a) statewide concentration of marginal (less than 1 hectare in size) holdings (for state level hypothesis), and (b) the number of holdings in the district (district level hypothesis). From Table 3.10 on page 135, we note that these variables have a correlation coefficient of 0.33. On the state level data, the statewide concentration of marginal holdings is statistically significant at the 10% level, with a log-odds value of -0.9 (odds-ratio of $1 : 2.5$) (model 5, Table 3.7 on page 126). On the district level data, it is statistically significant at the 1% level with similar log-odds (model 1, Table 3.11 on page 137). However, as more variables are added, its magnitude reduces. At the district level also, the hypothesis is confirmed. We find the district's number of holdings to be statistically significant at the 1% level, with a log-odds value of -0.596 (odds-ratio of $1 : 1.8$) (model 5, Table 3.11 on page 137). These results confirm our hypothesis at both the state and district levels.

7.1.4 External Factors

We proposed that policy adoption depends on external factors, and at the state level we hypothesized that teh additional funding available to special category states would prompt them to adopt the program. This hypothesis was not proposed at the district level.

The results (models 4 and 5, Table 3.7 on page 126 and models 1–6, Table 3.11 on page 137) do not show any support for this hypothesis using the state's category as the predictor. We find that the coefficients are statistically insignificant with large standard errors and frequently changing sign.

However, the interaction of the category with the state's development index is statistically significant at the 10% level (odds-ratio 1 : 14) (model 5, Table 3.7 on page 126) and the statistical significance increases to 1% in the district level models with a similar magnitude. This can be interpreted as:

1. For a General Category State, SCS is 0. So, the interaction term vanishes, and the effect is that of the Development Index, which is positive in our models. Propensity to adopt policy *increases* with increased development.

2. For a Special Category State, SCS is 1. For such states, the magnitude of the the interaction term's coefficient is subtracted from the Development Index to get the total effect. However, interaction term has a much larger magnitude than the coefficient on the state development index, resulting in a net effect that is negative. This means that in case of Special Category States, the propensity to adopt policy *reduces* with increased development.

We find mixed support for our hypotheses of policy adoption as well as uncovering new avenues for research. Next, we discuss the limitations of this work followed by future research avenues.

7.2 Limitations

One of the limitations of this research is the lack of a suitable proxy for policy salience at the district level. As discussed in section 7.1.1, rural area could be a weak proxy for implementation complexity, rather than policy salience, while proportion of agricultural workforce is not-statistically significant. Also, the coefficients of both these variables have signs opposite to that hypothesized.

Although the literature has indicated a strong positive effect of federal support ("vertical diffusion") on policy adoption (Karch, 2006; Shipan & Volden, 2012, 2008; Welch & Thompson, 1980), we do not see this effect in our case. Possibly, state category is not the right instrument, or the special category states do not feel that this particular policy is salient to them. This could be a case because we see from Table 3.6 that most of the special category states had the *raiyatwari* form of tenure, and none had the *zamindari* system which is most strongly associated with policy adoption. These aspects will need further study.

There is a data limitation in the form of the lack of availability of high quality health indicators at the district level. A socio-economic development indicator should include health indicators, as the state level index does with Infant Mortality Rate (IMR). However, no such information is available at the district level. Another indicator that could possibly help explain the factors behind policy adoption is corruption perception (Bussell, 2012). However, the challenge of including is is that it is not available for individual districts.

Another data limitation is the absence of ICT indicators that are able to gauge the real status of the availability of ICTs. Most such indicators use the availability

of telephone connections (landlines) to proxy ICT penetration. However, this metric fails to account for the much wider penetration of ICTs due to mobile telephony having "leapfrogged" traditional landlines. Hence, the effect of the "demand"-side on policy adoption may be underestimated.

The methodological limitation of this study is that as districts are contained within states, our dataset has a problem of endogeneity. This may result in the coefficients not being unbiased. Although a VIF test did not show the effect of multi-collinearity, running a hierarchical linear regression model to separate out the fixed and random effects of the states and the districts may uncover new findings.

These limitations result in the following open questions: (a) what could be the potential indicators of policy salience at the district level? (b) is there any effect of the political clout of a district on program adoption? (c) why are more developed, special category states less prone to adopt the policy?, and (d) why does additional funding to special category states not result in greater adoption?

8 Conclusion

This study is one of the first studies on policy diffusion in an Indian context. It attempts to identify the main factors impacting policy adoption across Indian states. The specific policy under study is the National Land Records Modernisation Programme (NLRMP), a program supporting Indian states in modernizing and digitizing their land administration systems. The program provides financial and technical assistance to the states for various activities that lead towards

deployment of a digital land administration system. However, this program has seen uneven adoption and we seek to understand why?

The program's adoption is analyzed at two levels (state and district) using the policy diffusion/adoption framework. The study hypothesizes that diffusion of land data creation policies depends on four main factors, namely (a) policy salience, (b) resource availability, (c) implementation complexity, and (d) external factors. We find support for policy salience, resource availability and implementation complexity at the state level, while the impact of external factors is not supported. At the district level, resource availability and presence of obstacles find support, while the policy salience hypothesis does not find support.

8.1 Policy Implications

The implications of this study for policy makers and analysts include:

1. Policy salience is key. The size of the problem motivates innovation and policy adoption to occur as can be seen by the huge odds ratio of $92:1$ in the propensity of policy adoption in the *zamindari* states compared to the *raiyatwari* states.

2. Resources matter, not only in financial terms but also human capacities as evidenced by our use of state level development indicators, and the state's category. While the state's category can be considered as a measure of funds availability, we find that its effect is unstable and statistically insignificant, while the development index has a consistent and statistically significant positive effect.

3. Implementation matters. If the challenges in implementation are larger than the perceived benefits, the administrations would be unwilling to adopt the program, preferring to apply resources elsewhere. However, given the importance and centrality of land records to development, there is a need to reduce the challenges in the implementation of land reforms programs. This can be done by building institutional capacities and the use of modern technologies, as well as equipping local youth with skills that can mitigate the workload of the land administration bureaucracy (para-surveyors, customer service centers etc.).

4. We find that the agricultural workforce concentration has a negative, but not statistically significant coefficient. A possible reason is a fear that impacted populations may not react positively to land records modernization, given the historical opaqueness surrounding government programs in India (Benjamin, Bhuvaneswari, Rajan, & Manjunatha, 2007; Nayak, 2013). This requires studying the attitudes of the local populace with regards to land records modernization, as well as opening up communication channels to educate everyone on the program and its perceived benefits.

8.2 Scope for Future Work

We find that interacting the state development index with state category gives an unexpected result of reduced propensity to adopt the policy if the special category state is more developed. This aspect needs further investigation.

The policy salience hypothesis at the district level looks at the proportion of rural

population in the district and its rural area. Another way of looking at this could be the to identify the presence of "powerful, vested interests". One indicator for this could be a measure of the in-equality in land holdings. However, this again is a chicken-and-egg problem, because that information itself would need to come from land records that are themselves not accurate.

Another aspect that could be relevant is the political positions of the various actors. However, the Indian political system is a multiparty democracy, with no party having strict ideological positions, which makes modeling this behavior challenging.

This study paves the way for future studies on policy adoption in emerging economies, especially the Indian context. It identifies a few open questions that can help in developing a better understanding of the policy process in emerging country contexts.

Notes

[1]The uniquely American way of governing is summed up by Jones (2005): "[T]he president is not the presidency. The presidency is not the government. Ours is not a presidential system".

[2]More details on these two schemes can be found in Chapter 2.

[3]The UTs are directly administered by the central (union) government.

[4]World Bank Blog "Why Secure Land Rights Matter" http://www.worldbank.org/en/news/feature/2017/03/24/why-secure-land-rights-matter. Retrieved April 1, 2017.

[5]There are some scholars who hold that the move to conclusive titling may not be the right thing, given the state of India's land records and development. See for example Zasloff (2011)

and Gupta (2010–2011).

[6]The "Permanent Settlement of Bengal" was entered into by Lord Cornwallis in 1793 (cf. Baden-Powell, 1892b).

[7]A brief overview of the data collection and cleansing process is given in appendix A.

[8]Only two out of the seven union territories have their own legislature, and they are administered primarily by the union government. Hence, policy adoption is more dependent on administrative, rather than political reasons. Their tenure types are also not available owing to their unique histories. There is also a lack of reliable data regarding the Union Territories, which has been acknowledged by the Raghuram Rajan Committee (Ministry of Finance, Government of India, 2013, pp 26).

[9]These states are Arunachal Pradesh, Goa, Manipur, Sikkim and Tripura.

[10]Website: http://dilrmp.nic.in. Retrieved 1 April 2017.

[11]As per F. S. Berry and Berry (2014, pg 321), this makes sense when studying a single policy.

[12]This may not be exactly accurate as present day state boundaries do not align with the states and territories of the British era. However, there is no dataset available that can map today's state and district boundaries with those in the British era. We do not anticipate that the results will be significantly different.

[13]We did not find the widely used composite measure like the Human Development Index (HDI) to be suitable as it does not include critical components. It was also not available for all the states separately, nor for the time period under study.

[14]Appendix B contains details on how this index is being used in this study.

[15]A single person/household may have multiple revenue holdings (which have separate land records) spread across the village. The agricultural census consolidates all these into a single operational holding. Thus, the number of revenue holdings is much larger than the number of operational holdings.

[16]Personal communication with official in the Madhya Pradesh Commissioner, Land Records and Settlement (MP).

[17]The district level data is available separately on the web at http://agcensus.dacnet.nic.in.

CHAPTER 4: BIG DATA PARADIGM APPLIED TO LAND ADMINISTRATION

ABSTRACT

High quality land administration requires a comprehensive view of land assets in near real time. Land data, provided by manifold data sources is key to this comprehensive view, and forms the linchpin of land administration. This land data is often dispersed across geographies, across agencies, and in various formats. This makes getting a comprehensive view easier said than done. A lack of such a view leads to significant land administration challenges, especially in emerging economies, with their weak administrative capacities. We demonstrate some of these land administration challenges with the aid of land administration use cases, and identify the key issues.

Land data is quintessential big data, and a solution to the land administration challenges lies in taking a big data perspective on land administration. The main framework elements needed to build a big data based land administration system are identified herein and an architecture proposed for the same.

At the core of this architecture lies a virtual data lake. This allows building a land system that is flexible enough to incorporate multiple information sources, while being resilient and adaptive to changing circumstances. The land administration use cases are re-evaluated in the context of the big data land administration system to find that the problems are resolved.

Taking the big data perspective on land administration necessitates having a supportive policy environment which cuts across multiple domains. This essay concludes with identifying the main areas where policies will need to be formulated and suggests the key aspects to be addressed.

1 Introduction

Effective land administration requires data which comes from various spatial and non-spatial sources. These data sources are dispersed across multiple geographies and administrative agencies at different (local/state/national) levels. Agencies differ in their missions, visions, goals and objectives. These inter-agency differences are reflected in their data collection, collation and publishing practices.

Policy making and policy analysis require an integrated view of this scattered land data. This integrated view can be provided by a Multi-Purpose Cadastre (MPC) which combines the pre-processed data from each source and presents it as a set of layers, with a Geographic Information System (GIS) serving as the base. This layered approach to the MPC has certain shortcomings, as neither can new data sources, nor new analyses be readily added. Thus, a fresh approach to the design of the MPC is needed.

This fresh approach is provided by treating the land data as big data and designing the MPC using a big data paradigm. Land data possesses the characteristics of big data, namely Volume, Variety and Velocity, thus making it quintessential big data. The big data paradigm puts data at the core, while moving the transactions to the periphery. This allows building a flexible, adaptive and resilient MPC. Building an MPC using the big data paradigm requires a new systems architecture. In this architecture, a "virtual data lake" mediates and controls access to the data sources by enforcing well-defined data governance policies. In order to function, this big data MPC also requires a supportive policy ecosystem.

This essay identifies the framework elements needed for building a big data MPC and proposes a possible architecture for the same. We demonstrate how the big data MPC is able to solve some of the issues occurring in land administration, while mitigating the risks associated with big data systems. We also present the key policy aspects that are necessary to develop and deploy this system.

The next section expands upon the role of data in land administration and motivates the need for a MPC by suggesting example use-cases. It further discusses the current layered approach to MPC construction. Section 3 discusses why land data should be treated as big data and outlines the need for a big data paradigm for using big data. This is followed by developing a set of use cases that highlight the issues in land administration which an MPC based on the layered architecture is unable to solve. The big data based land administration system is introduced in section 5, which identifies the framework elements for the system, a proposed architecture and how the big data land administration system can solve the land administration issues highlighted earlier. Section 6 discusses the necessary policy ecosystem before concluding.

2 Land Administration and Data

A nation's land policies are implemented using the infrastructure provided by its land administration system. It is generally accepted that effective land administration is key to achieving all-round, or sustainable development of an economy (Dale & Mclaughlin, 1999; Williamson et al., 2010; Williamson & Ting, 2001). Land administration systems have evolved in specific socio-

cultural contexts and thus vary widely around the world. Thus, the manner in which land administration can help in development also varies across contexts. Some economies want to manage their rapid urbanization and urban sprawl, while others might need land administration systems to support sustainable agriculture, emergency management or economic decision making (Williamson et al., 2010).

Land administration can support these goals by (a) helping in identifying land with specific attributes, (b) preventing and detecting fraud, (c) generating and analyzing agricultural statistics, and (d) clarifying the complex financial aspects of real estate. Each of these goals would require answering multiple questions like: where is this land parcel located?, who owns it?, what is it being used for?, is this land disputed? etc. Getting answers to such questions requires data from various sources.

The cadastre[1] is the core data for land administration and it provides the basic information on the land parcel, as well as how various people relate to the specific land parcel and/or any buildings etc. on it (Williamson et al., 2010). However, land administration is a complex activity[2], performed by different administratively and possibly geographically dispersed agencies. These agencies often have very different mission and vision goals.

The inter-agency differences get reflected in the data management processes resulting in varying data structures and data collection methods. This variety in the data, as well as its geographically and/or administratively dispersed nature, is a major challenge to effective land administration (Dale & Mclaughlin, 1999; Williamson et al., 2010). Land administrators have long sought a comprehensive

and continuously updated view of the land assets in near real-time (R. N. Cook, 1969; National Research Council, 1980). This comprehensive view has been termed as a MPC, also known as an Integrated Land Management System (ILMS).

2.1 Need for Multi-Purpose Cadastre (MPC)

A few examples are given below as to how a well-designed MPC can help support the land administration goals.

Identifying land with specific attributes for purposes of its acquisition and development for various purposes. In 2008, the government of Gujarat state in India used geo-spatial data to identify where to site an automobile manufacturing unit[3]. By integrating various data sources, identifying the owners of such land, as well as the rights and responsibilities associated with it can lead to a fast and transparent process.

Preventing and detecting fraud by geo-tagging assets built with public funds. One of the interviewees in the Indian state of MP recounted how multiple sanctions for the same project were averted by tagging the geographic location of the projects. A nationwide exercise is being conducted by the ISRO by in geo-tagging satellite images of physical assets created under the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) scheme[4], which allows verification and monitoring of the assets.

Generation and analysis of agricultural statistics is vital to policy making and analysis. Harvest volumes are estimated using a method known as “crop

cutting experiment", which collects crop samples from various village fields. These estimates are then aggregated at the state and country level to get an idea of the crop coming in at the end of the harvest season. These estimates are then used to plan policy regarding the various agricultural products. An integrated system would allow a policy maker to drill down and identify areas where crop failure or a glut could potentially occur and take the necessary measures.

Clarify complex financial aspects of real estate. Land markets have become exceedingly complex. In developed economies, the secondary mortgage market is huge and the presence of players like MERS[5] obfuscates the holding patterns. Added to this has been the creation of complex financial products like Collaterized Debt Obligations (CDOs) which were held largely responsible for the 2008 mortgage crisis that snowballed into a worldwide financial crisis. If all interests in a land parcel are linked together and the information made available on a near real-time basis, CDOs could become transparent. This can help avert financial crises like that in 2008-09 could be averted (Buhler & Cowen, 2010).

Linking multiple databases is not only useful for tackling the complex products, but also helps in detecting the proceeds of ill-gotten wealth, which is often parked in "*benami*" or anonymous properties in India.

Thus we note that integrating various systems can help land administration achieve the larger development goals. We next discuss how the MPC has been proposed to be built.

2.2 Building the Multi-Purpose Cadastre (MPC)

The current conception of the MPC is as a set of data layers stacked atop each other (figure 4.1 on the following page). This conception owes its origin to National Research Council (1980) who considered it analogous to "a registered set of transparencies that were manually registered with a set of pins" (National Research Council, 2007). In this system, each layer is often treated independently.

Over a period of time, this austere view of the MPC has given way to a richer vision that attempts to incorporate a multitude of data sources (figure 4.2 on page 166) within the same layered architecture. These multiple data sources could be maintained at different administrative levels, or dispersed across geographies, but the integrated system should be treated as a uniform national resource in line with the recommendations of the International Federation of Surveyors (UN-FIG)[6].

The basis of the system is a core Geographic Information System (GIS), which could be linked to a National Spatial Data Infrastructure (NSDI), if available (National Research Council, 2007; Williamson & Ting, 2001). This core GIS is overlaid with data from various other systems, for example cadastral data, data on administrative boundaries, soil data, watershed data, mortgage and financial data, identity information etc. (National Research Council, 1995, pp

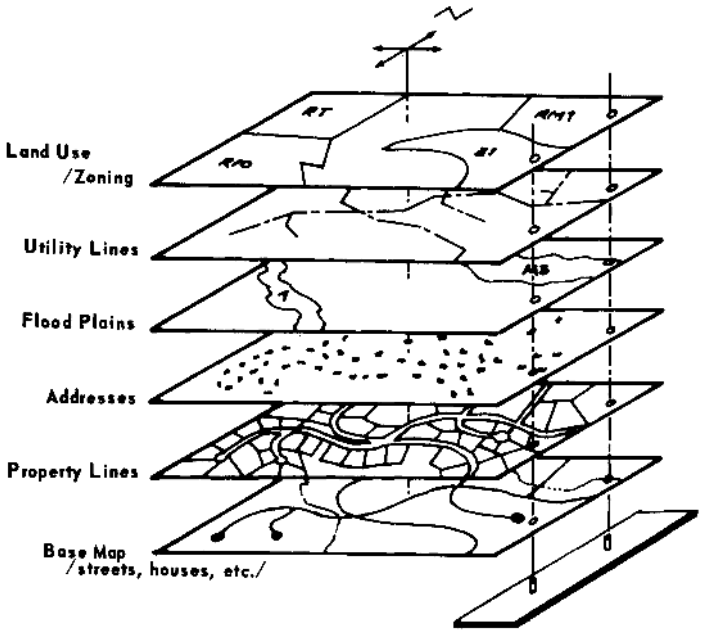


Figure 4.1: Stacked Layer Diagram. (Source: National Research Council (1980, pg. 42))

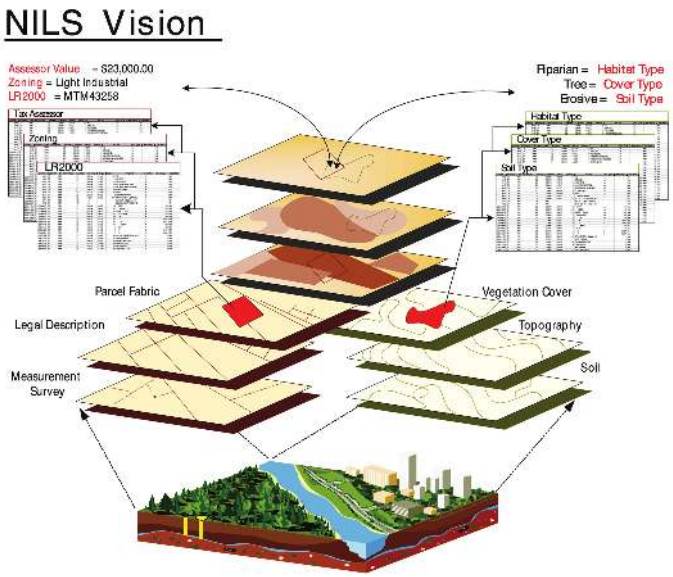


Figure 4.2: A Vision of an Integrated Land System. (Source: National Research Council (2007))

28). It should be emphasized that each system provides a processed view of its data—the raw underlying data is not provided to any other application.

In order to integrate these various data sources to get a comprehensive view of the land, the data must be harmonized and attached to the underlying GIS. This harmonization of the multiple sources requires *a priori* judgment of both the supply and demand sides of the data (National Research Council, 1995, pp 25).

2.3 Deficiencies of the Layered Architecture Multi-Purpose Cadastre (MPC)

A fatal flaw in the layered approach to the MPC, (outlined earlier) is the requirement to "harmonize" the data. As these "harmonized" data structures are often designed to be applicable to very broad categories, the system is incapable of accepting new datasets that do not conform to its pre-existing notions of what the data looks like. Thus, it cannot incorporate data structures that are specific (and critical) to certain use cases, for example, the emerging concept of 3D cadastres used for capturing information on multi-storied buildings. This limits their usefulness and potentially leads to a proliferation of data standards[7].

Systems designed using the layered architecture often take a "single source of truth" view, and this may result in their having a single point of failure. They are unable to collate data from multiple sources (that might provide incomplete and/or "messy" information) and triangulate it to get a reasonably close approximation of the truth.

Using the traditional layered frameworks, it is not possible to manage the wide variety of restrictions and responsibilities that affect land (figure 4.3 on the following page). This is because the layered approach not only limits the data variety, but also as it simply overlays data from different sub-systems. It does not provide any ability to fuse together different datasets to gain insights. Any such fusion has to happen separately, meaning that these systems are incapable of providing near real-time information that has become critical with the advent of "complex commodities"[8] in the land market (Wallace & Williamson, 2006).

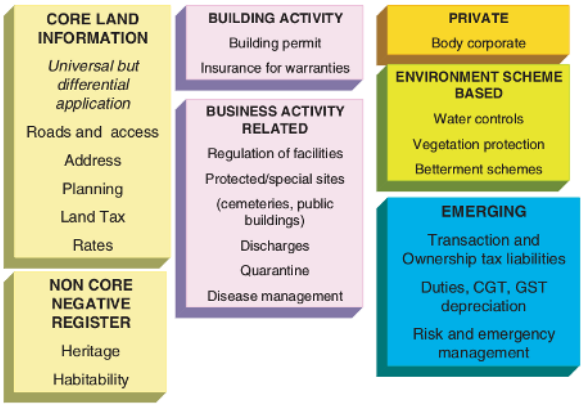


Figure 4.3: Restrictions and Responsibilities that affect land. Source: Wallace and Williamson (2006)

Thus, the MPC designed as a set of layers is unable to fulfill the vision that Williamson and Ting (2001) have for a modern land administration system: "a standardised, complete, nation-wide, current, on-line land information system in real time, which is efficient, economically justified and compatible with other information systems". The key to realizing this vision is to treat land data as the big data it is, and build a *flexible*, *adaptive* and *resilient* land administration system using the big data perspective. We next discuss why land data should be considered big data and how the big data perspective can be applied.

3 Land Data as "Big Data"

Big data has no single, concrete definition. It has been defined in various, sometimes ambiguous and mutually contradicting ways (Cukier & Mayer-Schöenberger, 2013; Manyika et al., 2011b; Mergel et al., 2016; UN Global Pulse, 2012; Ward & Barker, 2013). However, most definitions agree on big data possessing three key attributes — Volume, Variety and Velocity, often referred to its 3Vs (Diebold, 2012; Laney, 2001). These 3Vs are what distinguishes big data from "lots" of data (Borne, 2013). Big data for public policy and public affairs possesses the 3Vs, and is created by linking together public sector data (often administrative data) as well as private sector data (Mergel et al., 2016; Pirog, 2014). This data can be created both actively as well as passively, the so called "digital exhaust" (Mergel et al., 2016; UN Global Pulse, 2012).

Focusing on only the size, diversity and speed of big data, often leads to overlooking the one critical aspect of big data that distinguishes it from lots of data from multiple, high speed sources. This is that by combining, or fusing together disparate data sources, allows a level of analysis that was hitherto not possible. These analyses are made possible due to significant advances in the computing fields of machine learning and artificial intelligence. The use of machine learning techniques, makes it possible to create "actionable intelligence" (Hilbert, 2013). Machine learning allows the models to change as new data becomes available, making it feasible to extract information in a dynamic environment, which enhances decision making. Thus, when we talk about big data, we have to necessarily include the analytical aspects in addition to large sized, rapidly changing data from multiple sources.

3.1 Land Data is Quintessential Big Data

Land data is quintessential big data. Land data is created by combining manifold spatial[9] and non-spatial data sources. This data is dynamic in nature across both time and space. This linked data possesses all the three attributes — Volume, Variety and Velocity that define big data.

Volume is due to the sheer amount of data. According to my interviews in the Indian state of MP, the data for the state, excluding the spatial part is of the order of terabytes. This data does not include the old historical records as the status was captured only at the time of computerization. It also does not include the data that is captured by the deeds registration systems or the financial records. Adding all this data will increase the size of the dataset many times over.

Variety is due to the data being sourced from multiple spatial and non-spatial sources, which may be structured or unstructured. Figures 4.1 on page 165 and 4.2 on page 166 show a small sampling of the various spatial sources. Apart from these, non-spatial sources may include "textual" information garnered from the record of rights, deeds registry, financial databases as well as databases that serve to identify individuals.

Velocity is due to the dynamic nature of the land data (van der Molen, 2002). Land gets alienated, divided or merged and changes ownership over time. All these activities contribute to the data possessing some velocity. There is also a variation in the speed at which the various data change. The land parcel information may change swiftly in the time domain, but slowly in

the spatial domain. Other spatial information (watersheds, soils etc.) may not change for generations. The financial records may change as mortgages are paid off or the land re-mortgaged.

Data Analytics is the core required to extract intelligence from the data. The examples highlighted in section 2.1 (page 162) and the use cases discussed in section 4 (page 174) all need data analytics.

Thus, land data is a perfect example of big data. Apart from the above mentioned characteristics, land data also has to have veracity so that it can be used for administrative and policy purposes. The data also needs to be accompanied by a clear chain of provenance that indicates its sources as well as all the intermediate processing steps its has undergone. By treating it as big data, these attributes can be included as metadata to ensure that they remain integral to the data throughout the data processing and analysis phases.

3.2 Need for a Big Data Paradigm

Due to the unique characteristics of big data — big data is greater than the sum of its parts, it is imperative to treat it differently from just *any* bunch of data. The major issues identified with big data for public policy are of privacy, discrimination and a lack of control.

Privacy gets compromised when multiple datasets that have differing privacy/anonymity requirements are merged together without understanding the ramifications. Even when the data has been made available in an anonymized

form (stripped off any individual identifiers), it has been shown that it is possible to re-identify people in the dataset by combining it with other data (Barocas & Nissenbaum, 2014; Narayanan & Felten, 2014; Podesta, Pritzker, Moniz, Holdren, & Zients, 2014). Other than the possibility of re-identification, the anonymization process (a) also suppresses certain records[10], and (b) potentially strips the data of its ability to ensure provenance and accountability which could lead to wrong conclusions or mis-interpretations (Daries et al., 2014; Podesta et al., 2014).

Discrimination may occur with the use of big data in policy. This could be due to the training data[11] either being heavily biased towards a demographic, or not having enough representation (Barocas & Selbst, 2014). This results in the fitted model not matching the reality. One of the larger issues with this "algorithmic bias" is that it is extremely difficult to identify and even more difficult to prove and hold any entity responsible (Pasquale, 2015; Podesta et al., 2014).

Lack of control on data when it exits the organization. This means that it is impossible to identify who would be responsible if data de-anonymization occurs via combination with third party data that the organization does not control (Washington, 2014). As discussed by Mergel et al. (2016), transaction costs of gathering information in the pre-Internet data era were "nontrivial". Citing the example of land data, they identify that, earlier although the information was public, retrieving it necessitated a visit to a government office. With the advent of the Internet and the fact that much of this information is published online, the time and effort necessary to get it reduces significantly.

As a thought experiment, house construction plans are deemed to be public knowledge accessible to anyone[12]. This means that before Mr. Donald J. Trump became president of the United States of MAerica in 2016, the plans to his house in New York City (Trump Towers) were part of the public record. Thus, technically anyone could access them. However, due to the nontrivial transaction costs, only those with a valid reason would go to the local government office to access them. One can assume that the officials might play the role of a gatekeeper to ensure that only people with valid reasons get access. Post the 2016 US elections, since the building is now a high-security one, the local government might decide to classify those plans. What if the plans were earlier freely available on the Internet and multiple people had downloaded them? How can government officials redact already existing copies?

Given that these challenges occur when multiple data sources are fused together to create big data, it is imperative that a framework exists which can mitigate them. It has to be accepted that big data is here to stay and multiple data sources will continue to be fused together to gain "actionable intelligence". The solution lies in accepting the inevitable and working to minimize risks, rather than living in denial. While privacy concerns are justified, they have to be managed so that they do not impact bonafide data collection and research (Lane & Stodden, 2013).

Doing the above requires applying the big data perspective to minimize and mitigate risk. A case in point of inadvertent disclosure of Personally Identifiable Information (PII), in the context of land data is given in section 4.3. In section 5.3.3 we show how applying the big data paradigm can help in identifying the underlying issues and provides a solution.

4 Conceptualizing a Big Data Land Administration System

The layered approach that has been taken so far to view land assets comprehensively is to take an integrative view. Herein, each sub system of the complex land administration system is viewed as a sub-system. However, as land data is big data, the need for a big data paradigm exists.

This section presents a few use cases to bolster the arguments about land data being big data. These use cases are about some of the fundamental concerns of a developing economy, like low level frauds, large scale corruption and concerns about private information. Through these use cases the issues that remain unresolved by taking a layered system view are highlighted. I argue that these can be addressed by taking the big data paradigm. This paradigm moves beyond the recognition that land data is big data, to also highlight the requirement for policies that need to emerge for conceptualizing a big data administration system.

I posit that the core challenge that these issues identify lies in recognizing that land data is inherently big data. Hence, a potential solution lies in taking a big data perspective.

4.1 Land Fraud

During the study of land administration practices in India, a case of fraud was reported by one of the interviewees. This occurred in the city of Gwalior (in Madhya Pradesh state) a few years ago.

The existing land administration system: India follows a deeds registration system and property transfer is a multi-stage process in India. There are two main phases: (a) registration of the deed conveying the property, and (b) actual transfer (mutation) of the land parcel. In a deed registration system, the registered deed is evidence of the title, and not the title per se. It serves as evidence of a transaction with an intent to transfer interest in the property from one party to another. The actual ownership transfer does not happen until the names are changed in the Record of Rights, a process called "mutation". Thus no legal interest in the property is created until the mutation process is over.

Both the processes—deeds registration and mutation are important and significant. The deeds registration process ensures that the transaction is legal, and that the property is unencumbered by providing an opportunity to any existing lien holders to come forth and register their objections. On the other hand, the mutation process ensures that everyone on the ground is aware of an impending transaction and provides them an opportunity to make their objections, if any known. Such an objection could be related to the possible abridgment of someone's *de facto* (as opposed to *de jure*) rights. An example of a *de facto* right could be that of a tenant farmer who has an interest in the land and needs to be suitably compensated and/or notified. The *patwari* or village

accountant, by virtue of being the custodian of the village records, is the person on the ground responsible for ascertaining such information and initiating the mutation process.

A request for mutation is initiated by the deeds' registration office and is transmitted to the *patwari* in whose jurisdiction the property lies. After performing due diligence, the *patwari* performs the mutation in his/her records and sends the changes for approval to the superior officer, in this case the *tehsildar*. It should be noted that the *patwari* and *tehsildar* perform different administrative functions and do not have a direct superior-subordinate relationship. This serves to ensure checks and balances on the powers of both the officials, thus reducing corrupt and fraudulent activities.

With the use of computers in the mutation process, these approvals are now performed online. All officials, including the patwari and *tehsildar* have been assigned defined roles and issued distinct system credentials (login id and password) to access these roles. A patwari logs into the system with his/her id, performs a mutation request which is then routed to the *tehsildar*. The *tehsildar*, in turn, logs in using his/her id and approves or rejects the mutation request as the case may be.

The fraud event: When the fraud occurred, the *tehsildar* was not technology savvy and thus reluctant to approve the mutations online. He had shared his system credentials with the *patwaris*, who used it to perform the approval and other functions on his behalf.

Taking advantage of the situation, one of the *patwaris* fraudulently transferred nine-tenths of the land belonging to a government trust to himself[13]. However, he performed the transaction late in the evening (when not many transactions occur). The approval from the *tehsildar*[14] was given in the system almost instantaneously, which was not normal for such requests. This anomaly led to the fraud being detected.

Reflection: Had the patwari been more careful, then the fraud might not have come to light. In such a circumstance, a traditional, layered-architecture system would not have helped in identifying this fraud.

4.2 *Benami* (Anonymous) Property

Background: One of the challenges to development in India has been large scale corruption. One of the most sought after avenues of disposing off the ill-gotten gains has been to invest them in land. However, to avoid attention of the tax and other authorities, this property cannot be registered in the name of the person actually paying for it, but is done in the name of other (often fictitious) persons. Such property is called *"benami"* (literally: "without name") property. Due to these properties having shell ownership, the authorities are unable to identify the actual beneficiaries and take necessary legal action.

In 2016, the Government of India passed a law to prohibit the holding of such property. Identifying *benami* property is key to application of this law. Identifying *benami* property is not easy and requires copious amounts of local knowledge.

This local knowledge has to be pieced together with other evidence about the property holders and only then can a deal be suspected and further investigation initiated. The parliamentary standing committee looking into the bill has noted that digital land records could reduce the instance of *benami* transaction. It further recommended sharing of property registration data between the registration and tax authorities.

Controlling the corruption challenge: However, digital land records by themselves, and information exchange between the registration and tax authorities by itself cannot be used to identify potential *benami* properties. For large scale identification of such properties, integration of multiple administrative data sources, as well as linkage with certain external (possibly private sector) data (like social media) is required. Further, this data has to be enhanced with local knowledge. It is not possible to build these multi-dimensional linkages using the traditional layered architecture system.

Reflection: The long term goal of the law is to prevent creating new *benami* properties. Towards this end, the committee has recommended that all parties to a transaction be positively identified, and this identification recorded. However, recording of these identifiers could lead to a potential leakage of individual's Personally Identifiable Information (PII), and this aspect is discussed next.

4.3 Inadvertent disclosure of Personally Identifiable Information (PII)

Background: The Records of Rights and the registration documents need to uniquely identify the individuals in whose name the property is held, or those who are planning to transact in the same. Initially, these identifiers included the person's name, their parentage, address and age or date of birth. When most dealings were local, these parameters were sufficient to prevent some level of fraud. However, in today's mobile society with a large population these simple means no longer suffice.

To solve this problem, additional identifiers are used to uniquely identify the transacting parties. These identifiers could be some sort of national identification, but in absence of such an identifier, the person's photographs and/or biometrics (in form of finger prints) started to get used. Thus, in many places in India, a registration document has the photographs and biometrics of the transacting parties affixed to it.

Since 2009, India has embarked on an ambitious national identity database, called *Aadhar*. This system is based on biometrics (ten fingerprints and iris scans). An individual's biometrics are captured digitally and matched against all pre-existing entries in the database to see whether they are unique. If found to be unique, a twelve digit random number is assigned to the individual as his/her unique id. Till date, more than a billion such unique identities have been created.

This *Aadhar* number is unique to every individual, however it is not a secret or confidential number and thus can be quoted[15]. However, the law (*Aadhar*

Act, 2016), forbids divulging the biometrics of the *Aadhar* holder. To establish identity, the biometrics of the person are captured and sent to the authentication system (along with the *Aadhar* number) for verification. Based on whether the *Aadhar* number and the supplied biometrics matched or not, the authentication system returns a YES/NO answer to this authentication query. This process allows the *Aadhar* number to be recorded on the documents and used in lieu of the biometric identifiers to authenticate the transacting parties at the time of transaction as well as whenever required.

With the advent of *Aadhar*, the *Aadhar* number can and is replacing the biometrics that used to be a part of the registration documents[16] earlier, as shown in figure 4.4 on the next page. As discussed in section 4.2, the parliamentary standing committee on *benami* transactions has also recommended linking either the *Aadhar* or the Permanent Account Number (PAN)[17].

Privacy violation: However, this brings up an interesting problem. In the case of the deed registration based system that India follows, the seller has to provide a set of documents that provide evidence of his/her title to the property. This evidence takes the form of an unbroken chain of property conveyances that goes back some period of time and can therefore show that the seller has a legitimate interest in the property he/she is disposing of now.

From figure 4.4 on the following page, we note that when John Doe conveys his property xxx1, the two documents A and B have to be combined together to create the chain of evidence. However, doing so results in the biometrics and the *Aadhar* number getting inadvertently matched and available to anyone who has

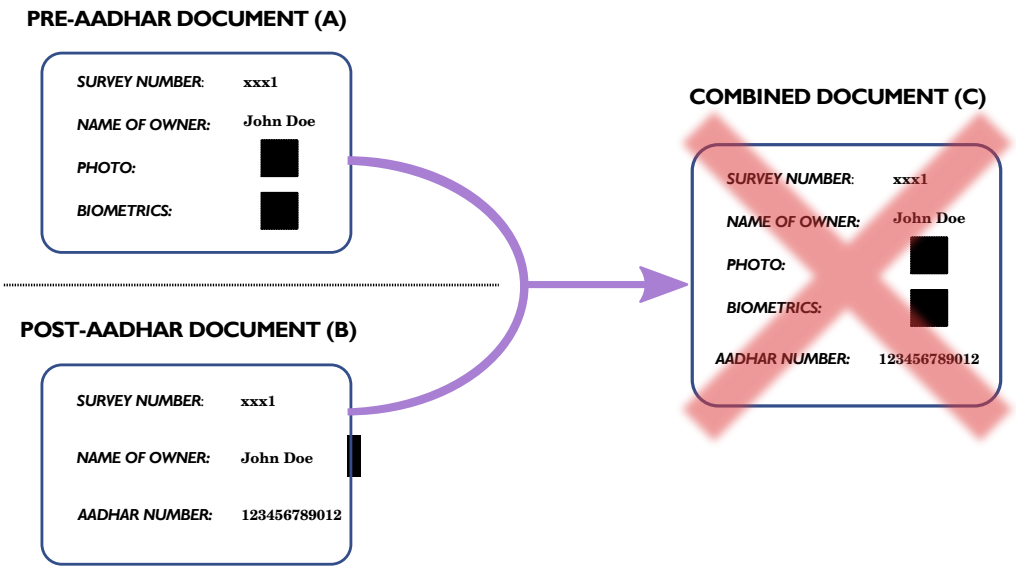


Figure 4.4: Document Structure before and after Aadhar

access to these documents (combined document C). It must be emphasized that these documents are part of the public record and thus available to anyone.

A naive solution to the issue of "leakage" of biometric information is to simply suppress the *Aadhar* layer in a layered-architecture system. However, doing so takes away the authentication advantages inherent to the *Aadhar* system and thus the source of trust reposed in the system by the transacting parties being sure of each other's identity. Further, as the *Aadhar* number is not secret, it can be found by various other means. As the biometrics form part of the document itself, and not a separate layer, there is no way to suppress them in a layered-architecture system. Hence, by keeping the biometrics available through the

registration document, and suppressing the *Aadhar*, we do not solve the leakage problem.

4.4 Sketch of Big Data Motivated Land Administration System

Land administration needs several varieties of data. These data are provided by various sources. An example is the spatial data that includes cadastral data, data on the administrative boundaries, soils and watershed data among others (National Research Council, 1995, pp 28). Other data used for land administration comes from non-spatial sources, which may include mortgage and financial data, in addition to data about the identities of the individuals involved. Integrating data from such multiple sources requires the development of small, adaptable and modular systems, instead of large, rigid monoliths van der Molen (2002).

An adaptive and flexible system would be able to integrate multiple data sources whose structures are not known *a priori*. A resilient land administration system would be able to deal with incomplete and "messy" data and be able to mediate between the data sources.

Not all data is pristine, or all sources equally trustworthy, nor can the provenance of all data be verified. Hence, it is necessary to relatively weigh these data sources, and score the evidence before integrating them. It is also important to provide the the rationale behind the scoring of the evidence, so that any biases present may be identified and taken care of, thus preventing the system from becoming a "black box" (Lazer et al., 2014; Pasquale, 2015).

A land administration system has to allow the seamless transfer of property when it is sold, inherited or otherwise transferred. During the transfer, the process has to ensure that due notice is provided to all concerned about an impending change of ownership. These notices are necessary as the land tenure is complex and there exist a multitude of overlapping rights and responsibilities which often lack clarity (Payne, 2004; Törhönen, 2004). However, much of this information on rights and responsibilities is often not available in the land records, necessitating information inputs from other sources. These information sources include people on the ground who possess local knowledge, thereby making context extremely critical to the processing of such information.

Big data has an innate capability to handle manifold data sources, and thus it can be self-describing. The data structures can be designed so as to carry their context along as they pass through the multiple processing steps. De-contextualization of big data (boyd & Crawford, 2012) is an artifact of the manner in which the data is collected and processed (Jagadish et al., 2014; Schintler & Kulkarni, 2014), and not inherent to big data..

Scientific databases ensure that this context is available by attaching the meta-data about the sources of information as part of the database . In the world of big data analytics, a data format known as Apache Avro[18] carries its schema definitions with it as it gets processed. Hence, using the big data paradigm, we can ensure that the land administration data is not de-contextualization by adding information on the context and ensuring that this information travels with the data through the processing pipeline. Similar mechanisms can also be used for management of the data provenance and ensuring its integrity.

The next section provides a high level architectural view of how a comprehensive land administration system using big data at its core can be built and how such a system can potentially solve some of the issues highlighted in the sections 4.1, 4.2 and 4.3.

5 Big Data Based Land Administration System

The key to building a *flexible*, *adaptive*, and *resilient* land administration system is to first identify its foundational elements. This is followed by defining a high-level conceptual architecture that can be implemented using available tools and technologies. This section concludes by providing example solutions to the issues highlighted in sections 4.1, 4.2 and 4.3.

5.1 Framework Elements

Four framework elements have been identified as key to the big data land administration system. These are— (a) stakeholder consultation, (b) incremental system design, (c) fit-for-purpose development, and (d) data governance. The role of each of these elements is briefly discussed.

5.1.1 Stakeholder Consultation

Land records are an abstraction (or a “model”) of the actual ground position. During my field research, the interviewees were of the view that community

members (on the ground) by and large agree on parcel boundaries[19], but the records often reflect a different story. A major reason behind this discrepancy is the dynamic nature of land ownership, which is often not captured by the administrative systems used (van der Molen, 2002). However, this has not always been the case. In India, prior to independence, taxes from land was the largest item in the budget (Rothermund, 1969). This ensured primacy of the land revenue administration, which was maintained by regular physical verification and audits (or "ground-truthing"). This created multiple touch-points between the administration and the land holders, ensuring that the records reflected the actual ground position(s).

However, post-independence, the importance of land revenue has dwindled, leading to laxity in the administration, which, in turn has led to increasing discrepancies between the land records and the actual ground position (Habibullah & Ahuja, 2005). This aspect was also highlighted during the various field interviews. The interviewees pointed out that traditionally, it has been the responsibility of the village accountant (*patwari*) to keep the records updated. However, over time, the *patwari's* duties have increased, without a concomitant increase in their numbers[20], which limits the state's administrative capacity. This mismatch between the recorded and the actual ground position is one of the sources of land disputes. During the interviews, it was emphasized that the land records in the state of Madhya Pradesh were digitized *as-is*, without any ground-truthing. Similar concerns were expressed about "Bhoomi", the flagship land records computerization project of the state of Karnataka[21]. The interviewees also observed that in many instances, the *patwaris* do not physically verify the land boundaries, often just signing off on paper. The entire burden of

verification is on the administrative machinery, with the citizens expected to be passive observers.

Continuous stakeholder consultation is one of the means to ensure that the on-record and the on-ground situations remain in sync. Crowdsourced mapping platforms provide one of the many mechanisms to incorporate local information into the official records. Examples of such crowdsourced mapping systems include US-AID's Mobile Application to Secure Tenure (MAST) implemented in Tanzania[22] and the tribal lands mapping project in Odisha, India (Choudhury, Rao, Kumar, Deo, & Dash, 2016, March 17). These projects use smartphones with GPS functionality to collect the data. These data are adjudicated, inconsistencies removed, and minor disputes resolved by holding village council meetings before being incorporated into the official records.

The creation of such crowdsourced data means that the number of data providers increases manifold. Earlier, the interactions were limited between the revenue staff and the land holder and his/her few neighbors, now everyone with a cellphone is a potential data provider to be interacted with. This results in creation of new categories of stakeholders, that is (a) data providers, who provide the data, (b) data aggregators, who aggregate the data sourced from multiple providers, and (c) data users, who are the end users of this data. The creation of these new categories has not only shifted the balance of power between these stakeholders, but also their relative transaction costs (Mergel et al., 2016). Earlier, the data user had to bear most of the costs, but now the individual data provider's costs have increased primarily because (s)he has to deal with multiple data aggregators and users. Therefore, any system that is based on big data

has to ensure against onerous transaction costs for the data provider, as well as against giving the data user a free pass at the expense of the data provider(s).

Such crowdsourced data are but one type of data becoming newly available. Many sources of data are yet to be discovered, along with the methods of integrating them. Thus, the capabilities of the data which combine to form big data is not known *a priori*, meaning that extra caution has to be exercised when merging multiple data sources.

5.1.2 Incremental System design

Traditional systems design freezes the specifications well in advance, only allowing for minimal system changes as the system is built. In software engineering parlance, this is the "waterfall model" of software development (Brooks, 1995). In this model, the entire design cycle is broken up into distinct phases like requirements analysis, system specification, system design, system implementation, testing and deployment. This design and development method tends to build large monolithic and tightly coupled systems where making changes is extremely difficult and expensive.

In a fast changing environment, it is impossible to know *a priori* the many uses to which a system will be put to, making it nearly impossible to completely specify the system requirements up front. Moreover, many times the specifications are developed for a model environment and do not always suit the actual context. Hence, the waterfall model of systems design is largely discredited today (Brooks, 1995).

Land administration systems are dynamic, and thus rigid monolithic systems designed using the waterfall model's design philosophy are clearly unsuited and the use of such systems and thinking leads to challenges in the field. An example was given regarding the creation of Modern Records Rooms in the state of MP while implementing the NLRMP. The program mandated a strict set of implementation guidelines, which included a significant underestimation of the amount of records to be digitized as well as an assumption that suitable physical space to setup the record rooms would be available in all districts and tehsils. Further, the specifications did not take care of inter-state differences which was often a hurdle in program implementation.

Putting the data as central to the design embraces the reality of changing specifications by allowing systems to be conceived of and built incrementally and iteratively. The big data paradigm allows for systems to evolve, machine learning to occur and new insights to be gleaned. This allows systemic changes to occur, while ensuring reproducibility by versioning both the data and algorithms together.

This approach eschews designing the "perfect" platform in favor of a conceptually clean, but incremental and modular system whose components are re-usable as well as replaceable.

5.1.3 Fit for Purpose Development

Another challenge closely associated with monolithic systems designed using a top-down approach is of "mission creep" that might occur while the system is

still being developed. As such systems have a long gestation period, often due to a changing circumstances, the system may be changed midway to perform additional tasks. An example of this was recounted by one of the interviewees. In the state of MP, while the revenue records were being digitized, it was also proposed to simultaneously digitize other (non-land revenue administration) records the using existing machinery. Unfortunately, this midway change in scope led to significant delays in the digitization of the land revenue records themselves due to underlying differences between the two tasks.

Another challenge is that many times, the magnitude of the issues is not known beforehand, and designers tend to look out for solutions. However, emerging economies have the capabilities to move to an advanced stage by bypassing (leapfrogging) intermediate stages, and thus systems built elsewhere may not be the right model. For example, many cadastral systems ask for extreme precision in measurement, which might not be suitable in all circumstances, as the context differs. At the World Bank Land and Poverty Conference, 2017, an administrator from India discussed how they were able to map an entire city using an UAV (drone) within a budget of under two hundred thousand dollars, including the cost of the drone.. At the same conference, one corporation was selling a “professional” land mapping UAV for more than half this amount.

Using a “fit-for-purpose” (Enemark, Bell, Lemmen, & McLaren, 2014) design philosophy helps solve these challenges. Systems built using the fit-for-purpose philosophy are designed to initially perform the minimum tasks required to achieve their goals, while being amenable to future functional enhancements. Big data practitioners acknowledge that all data sources are not equal, and ‘big

data" is inherently messy. Because the systems built using the 'fit for purpose" philosophy are designed to start small and grow they can use multiple, possibly "messy" data sources of varying resolutions. The data is cleaned or made "fit for use", at the point of use, which also allows fusion of higher fidelity data if it is needed or available.

An example of such "messy data" is redundant data, or data from multiple sources, which is "messy" in the traditional Relational Database Management System (RDBMS) view. In the RDBMS world, redundancy is shunned and system designers strive to design data schemas where every data is stored only once. However, as one of the interviewees recounted, in MP, maps existed with multiple agencies, and not all these maps were alike. These "messy" maps were extremely important in building up the cadastral records of villages whose maps had gone missing from the land administration offices. Thus, these maps helped avoid a full-blown and expensive land settlement exercise.

5.1.4 Data Governance

Data has a lifecycle—it gets created, used and after a period of time, the information content reduces and it simply adds to noise. Data that has outlived its life has to be culled or archived and new data takes its place. Traditionally, data has not been treated as a core asset, and data management has largely been concerned with operational issues like preventing unauthorized access or ensuring proper backup systems are in place, without acknowledging its lifecycle aspects. With the coming of big data, the data is core leading to a paradigm shift.

Data governance provides the set of rules that enable data management throughout the data lifecycle. Data governance requires a set of policies to ensure consistency, reliability, accountability and prevent ad-hocism. Adding to the challenge of big data governance is that "messy" (but not inaccurate or wrong) data is considered to be okay for big data systems. Hence, data governance is the foundation underpinning systems based on the big data paradigm. The main aspects of the data that governance policies have to manage are its (a) lifecycle, (b) quality, (c) provenance, and (d) access.

Data lifecycle. Data has a finite lifecycle, as it gets created, used and finally retired. The life of the data varies across sources and uses. For example, a high frequency sensor's output data has much lesser lifespan compared to census data which is collected every ten year. Further, even when the data is no longer in use, it needs to be archived and possibly made available as needed. Thus, the policies to manage this data will vary according to the context.

Data quality. Big data is inherently "messy", and all data has to be processed to make it amenable to processing. It is imperative to maintain a strict quality control on the data, possibly with the ability to trace it to its source if needed. Clear data quality policies allow systems and their users to understand the quantum of error or uncertainty that can exists in the results, ensuring that any decision making takes these aspects into account.

Data provenance is key to managing public sector big data. Big data comes from multiple sources and tracing its provenance is key to understanding its veracity and consequent usefulness. This requires that the metadata

(data about the data) be always available, as well as information about the intermediate processing that the data has undergone. This requires that some entity has overall control of the data and can verify that the data has not been tampered with.

Data access is a key aspect of public sector data, especially that intended to be in the public domain, like land records. However, as Mergel et al. (2016) mention, big data and ICTs have reduced (or eliminated) transaction costs, making that, which was once public, albeit inaccessible, now in easy reach, thus breaking down barriers to access. However, this also prevents agencies from performing vital gate-keeping functions that may be essential to preventing data misuse.

The data should be governed according to four foundational principles termed as "FAIR"–Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). Data governance is a continuous process, and the policies and rules will change over time as new lessons emerge. A challenge with big data is that the capabilities and dangers of mixing unknown data is not known beforehand. Therefore, whenever a new data source emerges, rigorous analyses need to be performed to understand which data can be allowed to be fused, and what data fusion is out of bounds.

5.2 Architecture of Big Data Based Land Administration System

Current approaches to the Multi-Purpose Cadastre (MPC) conceive a layered system (figure 4.2 on page 166). Such a system is designed as a conglomerate

of diverse systems, whose *processed, and not raw* data being meshed together. These systems use data whose sources are "silo"-ed across different agencies with varying mission and vision goals. Such systems require the data to be harmonized as discussed earlier in section 2.2. Hence, they cannot accommodate dynamic data as any change in the data structure(s) requires re-building the whole system.

Shifting the comprehensive land administration system's paradigm to big data puts *raw* data at the core, while moving the transactions to the periphery. Such systems do not require *a priori* data harmonization since the *raw* data is processed on an as-needed basis by the application(s) using it. This paradigm shift results in a *flexible*, *adaptive* and *resilient* land administration system.

The core concept used in building such a system is that of a "data lake". The "data lake" concept in the big data paradigm differs from the traditional data management systems in two main ways, which are its sources of strength. Firstly, while traditional data management systems, due to their harmonization requirements, force some sort of structure on the data, the data in a "data lake" exists in its native format. Thus, the raw data is available to any application to use as it deem fit, which allows the system to be flexible and adapt as needed. Secondly, big data is inherently messy and incomplete. Such data flows into the lake from a multitude of sources, each of which may have differing veracity. The onus of understanding what the data means and processing it is shared both by the data producers and the consumers, leading to a shared understanding of the data capabilities.

A comparison between the main attributes of a land administration system

Table 4.1: Layered Architecture Land Administration System versus Big Data Land Administration System

Layered Architecture	**Big Data (Virtual Data lake)**
Data from multiple sources is combined together	Data from multiple sources is combined together
Data remains with, and continues to be controlled by agency that collects it	Data remains with, and continues to be controlled by agency that collects it
Data is harmonized	No data harmonization. Data continues to be in native format.
Pre-defined usage scenarios	Usage defined at point of use.
Addition of new data sources requires significant re-engineering	New data sources can be plugged in and published

based on the layered architecture and one based on the big data paradigm is shown in Table 4.1.

The data lake has evolved from enterprise data warehousing systems and is like a “data dump”, where all the various sources dump in their data, although in a manner that allows individual datasets to be tagged so that they are Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al., 2016). However, this approach is not feasible for a resilient land administration system that has to scale nationally as not only does it create a single point of failure, but leads to severe data governance challenges. All data and data sources are not equal.

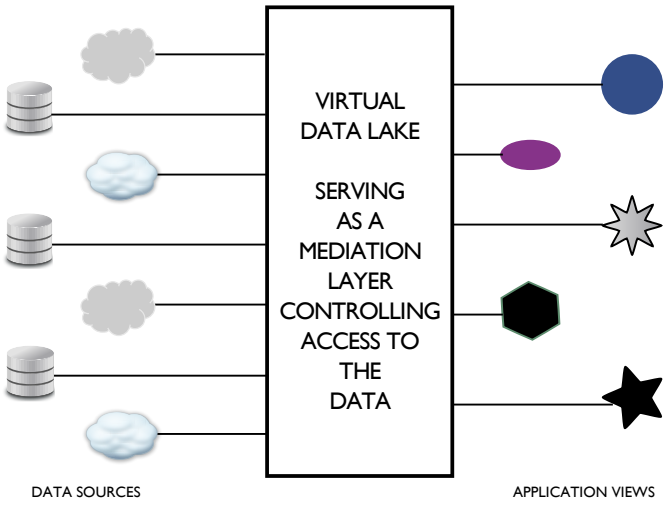


Figure 4.5: Architecture of Virtual Data Lake Based Comprehensive Land Administration System

Some of the data is sensitive data, while the reliability of other pieces of the data may be suspect. Also, centralizing the data leads to capacity issues in terms of processing, storage and network traffic. As the data is often fetched from the source system in batches, there is no assurance that the data is current. The alternative to using a "physical" data lake is a "virtual" data lake. The architecture of such a system is shown in figure 4.5.

As the figure illustrates, the virtual data lake serves to mediate and control access to the various data sources, based on the needs of specific applications. While, in the layered architecture model, each data source presented its own processed view of the data, in this system, each data source presents its data schema to the virtual data lake. It is the data lake's job to pick and choose what

elements of each source to use, based on a set of rules that define what sort of data can be fused on not. Thus, the data lake enforces data fusion policies. Further, the data lake mediates access to the data sources, preventing against data leakage. It can also serve as a secure enclave for sensitive data, ensuring that such data does not cross system boundaries, only the results of the analysis do. The next section explains how this architecture can help in solving the issues presented by the use cases outlined in sections 4.1, 4.2 and 4.3.

5.3 Solutions to Use Cases

5.3.1 Fraud Prevention

The fraud case discussed in section 4.1 on page 175 was allowed to be perpetrated due to the lack of three elements, namely, (a) backward and forward linkages between the deeds registry and the record of rights system, (b) mechanisms to identify and notify impacted parties of a transaction, and (c) defined process timelines.

Backward and forward linkages between the deeds registry and the record of rights would allow the record of rights system to query the deeds registry system to verify if the mutation being processed is part of an ongoing, valid property transfer. It would allow the mutation request to proceed only for valid transactions, else the request would be flagged as a possible fraud and appropriate authorities notified.

Mechanisms to identify and notify impacted parties would be able to prevent fraud by asking the concerned parties to verify the transaction(s). For this to work, the various land records databases could be queried for the contact details of the parties involved. By mining other databases — genealogical databases, court records, social networks, etc., this could also be used to pro-actively identify people who might have a potential interest in the land parcel In today's highly mobile society, the real impacted person(s) might have migrated and as the mutation process only provides notice to those who reside in the same geographical or administrative areas, fraud can be committed on absentee landlords. All such parties could be notified via multiple electronic or traditional means (text message, email, physical post etc.) and asked to verify the transaction.

Defined process timelines would prevent short-circuiting the land transfer process and ensure that due process of law is followed. It would also streamline the process by capping the maximum amount of time a request could be kept pending.

5.3.2 Identification and Prevention of *Benami* Property

The ability to analyze deed documents and mine multiple databases to identify persons having potential interests in a property can be extended to detect *benami* property and its beneficiaries. An example is that of the registration deeds. If the deeds are amenable to text processing[23], then artificial intelligence techniques could be used to perform automated analyses and identify the Rights, Restrictions and Responsibilities of the grantors and grantees. Maggs (1973)

had hinted at this possibility due to the limited vocabulary often used by legal documents, and Wouters, Meijerink, Vaandrager, and Zavrel (2010) have shown that this is feasible in the context of identifying listed encumbrances. Linking together multiple administrative databases like tax records with other sources (social media, financial transactions, on the ground knowledge etc.) will provide valuable insights in the nature of such holdings.

Thus, a big data based land administration system, using the virtual data lake concept can help in the endeavor of identifying *benami* property as well as prevent its further generation.

5.3.3 Prevent Leakage of Personally Identifiable Information (PII)

As discussed in section 4.3 on page 179, the naive solution to preventing leakage of personal sensitive information by suppressing the *Aadhar* layer will not work. However, a solution based on the big data paradigm can work if both the documents A and B in figure 4.4 on page 181 are available in a digital form.

This solution relies on the capabilities of big data analytic techniques (machine learning and artificial intelligence) to identify patterns, especially images. These techniques can identify the location of the biometric data in the documents. Once this location is identified, this biometric data can be redacted on versions that are presented publicly.

A more sophisticated version of the above could be to (a) identify where the biometrics are present (using image recognition techniques), and (b) capture images of these biometrics and upload these images (along with other information

- name, address, date of birth etc.) to the *Aadhar* database to find a potential match. If such a match is returned by the *Aadhar* database, the returned *Aadhar* number can be used to replace the biometrics when the document is presented in public, ensuring that the biometrics are not divulged to anyone.

Thus, we see that by building a land administration system using the big data paradigm which employs a virtual data lake, we can solve many of the challenges that occur in land administration. However, it should be emphasized again that these abilities of a big data based land administration system cannot be harnessed in a policy vacuum. Suitable policies have to be evolved that would create an enabling environment for such a system to thrive. The next section identifies the policies required to create such an ecosystem.

6 Policy Environment

The land administration system based on the big data paradigm requires an appropriate policy ecosystem to work. These policies have to be at multiple levels, and the exact policy would depend upon the context. However, the critical areas that would require policies to be formulated are (a) legal, (b) data governance, and (c) information systems. This will provide an accurate picture to the land administration which can be used to update the records. However, doing so requires development of policies that can set standards on what is acceptable when using such data collection tools and provide statutory backing to the collected data.

However, the fit-for-purpose approach should be the norm, rather than the

exception. Doing so requires the development of policies that encourage this approach, rather than penalizing its results. Such policies will cut across various domains—Information Systems, Data Governance, and Legal.

A supportive policy environment would be needed for identifying the set of admissible sources, the metadata (including provenance requirements), and defining the parameters for creating such a hierarchy.

This study uses a "big data" perspective to understand the data aspects of land administration. It is situated in the Indian state of Madhya Pradesh (MP). Using mixed methods—interviews and archival research, the major challenges in creation of land data are identified. It identifies that putting "big data" at the center allows the creation of a flexible, transparent and resilient Integrated Land Management System. This can help overcome many land administration challenges. However, such a system can only exist if appropriate policies for big data are in place to allow fusion of land data in a manner consonant with the larger public interest.

6.1 Legal

It is imperative to root the big data paradigm based land administration system within a well defined statutory framework. Such a statutory framework will have to take cognizance of all existing rules and regulations, and identify its points of intersection with the extant law.

One key element that needs to be developed are laws concerning privacy of the individual and identifying the conditions under which certain data can be fused.

Emerging countries have a significant lack of such regulations, which puts the data of entire populations at risk (Taylor, 2016a, 2016b; Taylor & Broeders, 2015). However, privacy is culturally specific (Capurro, 2005; Margulis, 2003), which means that the elements of what constitutes PII depend on cultural and social values and norms. It is also necessary not to conflate between privacy and anonymity, which according to Skopek (2014) work in different ways. While privacy removes information from circulation, anonymity removes identity to put information into circulation. From the policy perspective, the rules and regulations will have to be devised based on "how" the data is going to be used (Mundie, 2014; PCAST, 2014).

Other statutory elements will be related to how evidence from third-party sources can be integrated, especially those that can potentially change administrative data, for example community based mapping. Other regulations that could impact methods of data collection, for example usage of UAVs or the level of precision required to make certain data "fit-for-purpose" as against "unfit-for-purpose".

6.2 Data Governance

Data governance is the linchpin of big data and its impact is felt throughout the data lifecycle. Data governance (Khatri & Brown, 2010) is needed to maintain metadata, assure data provenance and ensure consistent usage. Data governance policies need to be cognizant of the multiple aspects surrounding data privacy, data anonymity and data ownership which are some of the biggest challenges in the broader use of big data. Data governance policies are needed to ad-

dress the data quality challenges, while ensuring that the data is “fit-for-purpose”. Well defined data governance policies would be needed to ensure that data fusion happens according to well defined rules and the appropriate data controls are exercised throughout the system.

6.3 Information Systems

Policies in the domain of information systems will define and set the parameters for how the information processing systems work, including how they are accessed and controlled. This will include policies governing the use of cloud computing, security and access control amongst others. Information systems policies will define how citizen engagement occurs — what information can be solicited from citizens, what processes can be used to solicit such information and the rules under which said information can be processed.

Table 4.2 provides a rough mapping of the key framework elements to the policy areas they would impact the most. However, we note from this table that most framework elements would require policies to be developed in all the three areas. These policy areas are also closely intertwined and policies in one area will impact how policies in the other areas evolve, necessitating looking at all three aspects simultaneously.

Table 4.2: Mapping the key aspects of a big data MPC to the policy environment

Sl.	Key Finding	Main aspects	Big Data Perspective	Policy Domains for Big Data
1	Stakeholder Consultation	**Problem:** Administration has limited capacity, which leads to records not being updated and thus disputes. **Solution:** Involve citizens in the data creation process. **Requires:** Citizen feedback policies, processes and mechanisms.	• Community mapping • Multiple touchpoints • Crowd sourcing	• Information Systems • Data Governance • Legal

Sl.	Key Finding	Main aspects	Big Data Perspective	Policy Domains for Big Data
2	Incremental System Design	**Problem:** Land administration is dynamic and requires flexible systems. Specifications frozen early on in the development process do not Land administration is dynamic and needs **Solution:** Allow systems to be conceived of and built incrementally and iteratively. **Requires:** Paradigmatic shift in system conception and design	• Data is central. • Adaptive systems that evolve with changed requirement. • Identify gaps and focus on plugging them, incrementally • Do *not* aim for Big Bang reforms	• Information Systems • Data Governance

Sl.	Key Finding	Main aspects	Big Data Perspective	Policy Domains for Big Data
3	Fit for Purpose	**Problem:** Systems are designed to solve present and anticipated *future* problems. This results in monolithic, over-engineered and expensive systems. **Solution:** A "fit-for-purpose" approach allows systems to be built to solve today's problems, while making them adaptable **Requires:** Develop policies that eschew "one size fits all" approach.	• Data as is - cleaning to be done at point of use	• Information Systems • Data Governance • Legal

Sl.	Key Finding	Main aspects	Big Data Perspective	Policy Domains for Big Data
4	Data Management	**Problem:** Data needs to be managed throughout its lifecycle. **Solution:** Data governance needs to be in place **Requires:** Data policies that care care of data quality, provenance, access and security	• Manage Data Lifecycle—creation, curation and archival of data • Data quality • Data provenance • Data accessibility	• Information Systems • Data Governance • Legal

7 Conclusion

Data is the linchpin of land administration. An effective and efficient land administration system is necessary for development. Land administration is complex due to its dynamism, spatio-temporal dimensions and by being embedded in varying socio-economic contexts. This complexity has hitherto been managed by apportioning it across different agencies. However, policy decisions require a comprehensive view of the land assets. As the land data is dispersed across different geographies and agencies with varied goals and objectives, this comprehensive view is not automatic. The concept of the Multi-Purpose Cadastre (MPC) has been mooted as a way to get this comprehensive view.

The extant thinking on the design of the MPC is a system integrating "layers" of data from the various agencies to provide the required comprehensive view. But, this data, defined according to agencies' own goals and objectives, differs in its forms and contents. Thus, data integration requires data harmonization that inevitably leads to loss of information and precludes addition of new data sources without extensive re-engineering. Hence, systems built using this layered architecture are static and rigid which is not aligned to the inherent nature of land administration. Thus, modern day land administration requires a paradigm shift.

This paradigm shift is effected by realizing that land data is quintessential big data. Big data, typified by its 3Vs — Volume, Variety and Velocity is created by fusing large structured and/or unstructured datasets from manifold data sources. The key aspects of big data is that the data sources have varying structures and formats and the primary analytical tools are from the fields of computer machine learning and artificial intelligence. Land data is big data, not only because it possesses these attributes, but also because it behaves like big data as demonstrated by a set of use cases. These use cases raise certain issues that are not easily resolvable using a land administration system built using the layered approach.

By treating land data as big data, we envisage building a *flexible*, *adaptive* and *resilient* land administration system that puts data at its core, while all transaction related intelligence is pushed to the periphery. Four framework elements namely, (a) stakeholder consultation, (b) incremental system design, (c) fit for purpose, and (d) data management are identified as essential to a big

data land administration system.

The core data storage and management concept is that of a “virtual data lake”. A data lake differs from traditional data storage systems in that it does not force any structure on the data, but rather the data exists in its native format. This data is managed using FAIR — Findable, Accessible, Interoperable and Reusable practices, which ensure that it can be used by multiple, distributed applications. Agencies’ autonomy and differing operational practices are respected by making the data lake “virtual”, which keeps the data under the owning agency’s control. This ensures that the data is always controlled by the respective agency and its data sharing policies, preventing “fracture of the control zone” Lagoze (2014). It also prevents issues due to unwarranted merging of data and obviates single points of failure. The data lake mediates access to the data and enforces policy regarding permissible accesses and uses of the data. Using the big data paradigm for land administration also resolves the problems posed by the example use cases.

However, for the big data land administration system to function, an appropriate, multi-domain, enabling policy ecosystem is required. The three main policy areas identified are (a) legal, (b) data governance, and (c) information systems. The big data land administration system requires an unambiguous statutory backing that lays down clear guidelines on what is permissible and what is not permissible with the data, including what data can be fused with what other data and this fusion is to proceed. A strong policy framework for data governance to support the legal requirements by controlling data over its entire lifecycle, while assuring data quality and control is imperative. Policies also

need to be devised for various operational aspects that include access and control mechanisms, data center siting, use of cloud technologies among others.

Taking a big data perspective on land data, coupled with an enabling policy environment will thus allow development and deployment of a flexible, adaptive and resilient land administration system.

Notes

[1]Cadastre Definition: "A Cadastre is normally a parcel based, and up-to-date land information system containing a record of interests in land (e.g. rights, restrictions and responsibilities). It usually includes a geometric description of land parcels linked to other records describing the nature of the interests, the ownership or control of those interests, and often the value of the parcel and its improvements. It may be established for fiscal purposes (e.g. valuation and equitable taxation), legal purposes (conveyancing), to assist in the management of land and land use (e.g. for planning and other administrative purposes), and enables sustainable development and environmental protection." Source: The International Federation of Surveyors (FIG). The FIG Statement on the Cadastre. *FIG PUBLICATION No 11, 1995*. Available at: http://www.fig.net/resources/publications/figpub/pub11/figpub11.asp. Retrieved May 1, 2017.

[2]Bathurst Declaration (1999) identified that the "range of rights, restrictions and responsibilities related to land is increasingly complex". http://www.fig.net/resources/publications/figpub/pub21/figpub21.asp. Accessed July, 15 2015.

[3]Source: http://timesofindia.indiatimes.com/city/ahmedabad/Nano-land-identified-through-remote-sensing/articleshow/3674129.cms. Retrieved May 13, 2017

[4]Sources: http://nrega.nic.in/Netnrega/WriteReaddata/Circulars/1674SOP_GIS_MGNREGA_27062016.pdf, http://isro.gov.in/mou-signed-between-isro-and-mord-geo-tagging-assets-of-mgnrega, http://bhuvan.nrsc.gov.in/governance/tools/nrega_v2.1/nrega_manual_v1.pdf. Retrieved May 13, 2017)

[5]In the US, Mortgage Electronic Registration Systems, Inc. or MERS is a huge player in the secondary mortgage market. MERS allows any of its members to hold a mortgage, while the recorded lien is held by the original mortgagor. This results in a difference between what

the public record shows and who really holds the mortgage. For more on this, refer to Dordan (2010), Kranz (2012).

[6]See note 2.

[7]As the author has personally experienced on multiple occasions, some data is manually ”mangled” to force-fit the notion of what the data should look like.

[8]An example of such complex commodities could be the Collaterized Debt Obligations (CDOs) held largely responsible for the 2008–09 financial crisis.

[9]According to an estimate, spatial data is almost eighty percent of all big data (Leszczynski & Crampton, 2016).

[10]One of the ways to anonymize data is to remove those outliers who are readily recognizable. However, one of the advantages of big data is that it pertains to almost the entire population and thus reduces the sampling bias (Welles, 2014). If the outliers and minorities are removed from the data, this utility vanishes.

[11]Machine learning models have to be trained on some initial data. The model can learn biases if the training data itself is biased. For more, refer to Mitchell (1999).

[12]Source: http://homeguides.sfgate.com/original-blueprints-house-8712.html. Retrieved May 19, 2017.

[13]The patwari knew that if the entire property was transferred, it would lead to a deletion of the trustees name from the records, possibly leading to immediate detection. Hence, by leaving a part of the land with the trustees, the original record was not deleted, and the fraud was not detected immediately.

[14]Actually, the *patwari* using the *tehsildar's* credentials.

[15]Blog posting by Telecom Regulatory Authority of India Chairman, Mr. Ram Sevak Sharma at http://blogs.economictimes.indiatimes.com/et-commentary/there-has-been-no-aadhaar-data-leak/ (Retrieved May 9, 2017)

[16]However, it seems that the processes of using *Aadhar* with biometrics has not been clearly laid down, and both are being used, leading to the identified problem being present even now.

[17]The Permanent Account Number (PAN) is a number allotted to individuals and corporates for purposes of depositing taxes and filing tax returns.

[18]More details on the Apache Avro format can be found at https://avro.apache.org. Retrieved: May 9, 2017.

[19]Larger agreement of the community members on the parcel boundaries without recourse to formal documentation is well known (cf. Baden-Powell (1892c, pp 33)).

[20]The *patwari* (or village accountant) is a village official whose duties cut across multiple departments, not just land revenue. For example, the *patwari* is responsible for performing the crop-cutting experiments (for yield estimations) for the agriculture department. Also, many of these positions are vacant as indicated by one of the interviewees.

[21]It has been referred to as a "Garbage In, Garbage Out" scheme due to its failure to consider the actual ownership. Rothermund (1971) also discusses the tensions between revenue officials in states (like Karnataka) having the *raiyatwari* system and those from North India, which had a meticulous system of record keeping. Landesa (http://www.landesa.org), a charity organization working in the area of land rights, was also using its resources to ground-truth land ownership in the southern India state of Telangana.

[22]Source: https://www.usaidlandtenure.net/project/mobile-application-to-secure-tenure-tanzania/. Retrieved: May 15, 2017. See: Neyman, Linkow, and Kijazi (2016, March 17)

[23]To be "amenable to text processing", means that one can perform free-form text searches on the document content, and not just its metadata. For this to happen, the documents need to have a text layer associated with them, often using Optical Character Recognition (OCR) techniques. Although having text layers seems to be obvious, there are significant costs associated with the process and many times this step is skipped. An example is the County Clerk's office of Fairfax County, Virginia, USA. While, the office has computerized its deeds, free-form text searches are not possible as these documents as saved as images without any textual representation of the information. Unfortunately, even newly filed documents continue to be uploaded as images.

CHAPTER 5: CONCLUSIONS AND POLICY IMPLICATIONS

1 Public Policy and Big Data

With the rapid rise in Information and Communication Technologies, the world has seen an exponential increase in the amount of "born digital" (PCAST, 2014) data being created. This has led to what Decker (2014) has called a"data deluge". These new data sources differ from extant sources in their ability to be linked together, thanks to the mainstreaming of Machine Learning and Artificial Intelligence techniques coupled with the continuously reducing costs of data processing, storage and transfer. This linking of manifold data sources along with the application of modern analytic techniques has led to the creation of what is called "big data". Big data is a whole much greater than the sum of its parts. Big data can provide much deeper insights into human behavior than was possible earlier, leading to "actionable intelligence" (Hilbert, 2013) and has been called the "new oil"[1].

Data and its analysis is the core of social science (research), and therefore of policy analysis. Without data, policy analysts would be unable to understand

the impact of how extant policies nor develop new policies. The information content of big data is richer than ever before, thus allowing building models with greater accuracy and predictive power. These accurate models can thus help in developing targeted policies, and the near real-time feedback possible can allow the policies to be tweaked as needed. This increased efficiency and effectiveness in policy making makes big data extremely desirable to the policy world, especially to those in the field of international developmental (Taylor et al., 2014; Taylor & Schroeder, 2014; UN Global Pulse, 2012). However, big data is a "double-edged sword"[2] and the use of big data in policy comes with its own set of challenges.

The major issues identified with big data for public policy are of privacy, discrimination and a lack of control. Privacy gets compromised when multiple datasets having differing privacy/anonymity requirements are merged together without understanding the ramifications. Discrimination in the use of big data in policy may occur due to the training data[3] either being heavily biased towards a demographic, or not having enough representation (Barocas & Selbst, 2014). This results in the fitted model not matching reality. A challenge, not directly related to big data is the lack of control on the data when it exits the organization, or as Lagoze (2014) says "fracturing of the control zone". As discussed by Mergel et al. (2016), transaction costs of gathering information in the pre-Internet data era were "nontrivial", which ensured some level of control on the data. A question arises on who is liable if if data de-anonymization occurs via combination with third party data that the organization does not control (Washington, 2014). Adding to these big data challenges is a larger one which impacts emerging economies who don't have the necessary data collection infrastructures in place.

The sources of big data is the "global south" are largely in the hands of private players running the social media and telecommunication companies (Taylor & Broeders, 2015), which leads to a potential "digital divide" (boyd & Crawford, 2012). Thus, not only do policy analysts have to figure out how to use big data for policy, but at the same time policies for big data are required.

However, in this fascination for new sources of digital data, what is oft forgotten is that the public sector collects huge amounts of data during the normal process of governing. Thus, the government agencies can create administrative big data by linking together data sources that already exist and use this big data for public policy, especially for developmental purposes. But, before this administrative data can be linked together, it has to first exist and be in a digital format so that big data can be created.

One rich source of such administrative data pertains to land records. Land is important to human society and plays a vital role in human development. However, in many parts of the world, especially emerging economies, access to land is not equitable. Good quality land data can go a long way in identifying such inequities and framing appropriate policies to resolve the same. The essays in this dissertation have looked at distinct, but related aspects of land big data in an emerging economy — India.

The first two essays in this dissertation have tried to understand and identify the challenges that exist in the creation of digital data from legacy, physical sources using both qualitative and empirical research techniques. The third essay looked at how land big data can help in development, proposed a paradigm shift in the treatment of land data by arguing that as land data is big data, it

ought to be treated as such and made the case for such treatment.

2 Findings

Essay 1: Land Administration in India has been an exploration into the processes of land administration and land data creation. Land in India is under the purview of respective state governments. However, as land is so important to development, the central (federal) government funds various activities that improve land administration. One of these is the NLRMP that provides funding and technical support to the state governments to modernize their land records by undertaking a predefined set of activities. The NLRMP is studied as an example of a program that aims to create digital administrative data.

I identified the key challenges faced in project implementation by talking to various stakeholders, who included both central and state government officials, and triangulating their experiences with documentary evidence. These challenges are (a) historical legacy, (b) existing level of economic development, (c) level of administrative support, and (d) policy design. India has a long and complex history of land administration which is manifested in a multiplicity of land tenure systems. These different land tenure systems have led to a variety in both the land records and the administrative processes. The Indian states vary widely in their geographical and socioeconomic characteristics as well as administrative capacity. Thus, every state has its own nuances and challenges in implementing the NLRMP. However, the program as designed lays down extremely rigid specifications which are largely the same for all states. This uniformity and rigidity

has led to significant challenges in program implementation.

Essay 2: Diffusion of Data Policies: a Sub-National Study investigated if the state level variations in adoption of the NLRMP could be explained by the challenges identified in the first essay. The data from the NLRMP MIS was combined with data from multiple sources to create a novel dataset which was analyzed using a policy diffusion framework. This analysis was performed at both the state and district levels. At the state level, it was hypothesized that four factors, namely (a) policy salience (proxied by tenure type), (b) level of socio-economic development, (c) complexity of policy implementation, and (d) level of federal support (proxied by the state's category) mattered. At the district level, three factors were hypothesized to impact choice of district. These were (a) policy salience (proxied by district's rural area and proportion of agricultural workforce), (b) level of socio-economic development, and (c) complexity of policy implementation. The binary logistic regression finds mixed support for these factors. At the state level, policy salience is statistically significant and in the expected direction, while the hypothesis is not supported at the district level. The level of socioeconomic development is both statistically significant and as expected at both the state and district levels. Similar results are obtained for policy implementation complexity. However, at the state level, no support is found for the hypothesis that the level of federal support for the policy impacts adoption.

Essay 3: Big Data Paradigm Applied to Land Administration made the case that land data is big data, and thus it should be treated as such. Using spe-

cific land administration use-cases, it demonstrated how current systems are amenable to fraud, leakage of PII and unable to support various law enforcement requirements. It proposed a new model for developing a MPC, one that puts data at its core and ensures that all data access passes through a mediating layer, thus ensuring that only certain types of data are fused together. This essay further identified the key domains where policies have to develop to make this big data land administration system a success.

3 Policy Implications

This work has identified that big data creation and usage from administrative data requires an appreciation of the problem, and an enabling policy environment adequately supported by financial and administrative resources. Thus, policies for the creation and usage of big data have to be at two levels, namely (a) a broader and generic data perspective, and (b) a set of domain specific policies.

3.1 Policy Environment for Big Data

As identified in chapter 4, the policies needed for the creation and use of big data cut across three policy domains—Information Systems, Data Governance, and Legal. However, it should be noted that these policy areas are not exclusive, but closely intertwined. Therefore, it is necessary to develop policies simultaneously in all the three areas.

Information Systems policies will define and set the parameters for how the information processing systems work, including how they are accessed and controlled. This will include policies governing the use of cloud computing, security and access control amongst others. Especially relevant to this are policies governing the trans-border flow of information and its potential impact on national security and competitiveness.

An allied area is how these policies impact citizen engagement, by defining what information can be solicited from citizens, what processes can be used to solicit such information and the rules under which said information can be processed. It also needs to consider the provision and solicitation of information to/from from the disabled.

Data governance is the linchpin of big data as its impact is felt throughout the data lifecycle. It is needed to maintain metadata, assure data provenance and ensure consistent usage (Khatri & Brown, 2010). These policies have to be broad, as well as deep to address the multiple aspects which include data privacy, data anonymity, data ownership and data quality. Well defined data governance policies would be needed to ensure that data fusion happens according to well defined rules and the appropriate data controls are exercised throughout the system.

A legal framework for big data is imperative to ensure that citizens' privacy and security are not abridged while creating and using big data. The statutory framework has to take cognizance of all existing rules and regulations, and

identify its points of intersection with the extant law.

Laws concerning privacy of the individual and identifying the conditions under which certain data can be fused need to be developed. This is of special concern in emerging economies that lack such regulations, putting the data of entire populations at risk (Taylor, 2016a, 2016b; Taylor & Broeders, 2015). However, as privacy is culturally specific (Capurro, 2005; Margulis, 2003), what constitutes PII depends on cultural and social values and norms. These rules and regulations will have to be devised based on "how" the data is going to be used (Mundie, 2014; PCAST, 2014).

The statutory framework also has to contend with how to integrate emerging data sources like UAVs, as well as third-party data sources.

3.2 Land Data Policies

GISs are a major constituent of land data. An efficient and effective land administration system needs to be geo-referenced, preferably to the National Spatial Data Infrastructure (NSDI) (Williamson et al., 2010, Ch 9). However, India lacks a clear national geospatial policy. Although there have been several attempts to build such policies, they have been thwarted by the conflation of geospatial data with national security. The unfortunate consequence of this has been the mushrooming of ad hoc solutions, which may not integrate, scale and actually cause harm to national security[4]. Therefore, it is imperative to create a national geospatial policy and ensure that all geospatial products in India adhere to said policy.

Significant interstate variations exist in the availability and type of land data among the Indian states. These variations are due to historical legacies and different development pathways, and thus the states' land administration processes, needs and capabilities differ significantly. Further, there are differences within the state itself, especially in the larger states. This means that any such policy needs to be flexible to accommodate the on-the-ground variations.

The National Land Records Modernisation Programme (NLRMP) has a major lacuna in the form of its rigid specifications, on one hand, while being ambiguous on certain other aspects. The scheme needs to re-look at some of its underlying assumptions and possibly re-design, for example its strict guidelines and estimates on setting up of the MRR. These specifications call for a heavy, fire-proof door which may not be supported by all existing buildings, while disallowing any fresh construction. There is also a strict limitation on the size of the room. But, as events have shown, the number of records has been grossly underestimated, which necessitates a larger record room.

On the other hand, there is an ambivalence about the usage and creation of geographic data. The states have been asked to directly negotiate with the national mapping and surveying authorities like the NRSC and the SoI. However, this has led to the national agencies thinking of themselves as vendors and treating the states as a customer, instead of as partners in a national endeavor. The DoLR could have taken a lead role in engaging the central mapping and surveying agencies and facilitated the states' engagement with these agencies, rather than asking the states to co-ordinate on their own on an ad-hoc basis as in the case of integration with the national spatial grid. As one of the interviewees

noted, out of the expected thirty-five SoI benchmarks for the city, they were only able to find eleven,as the rest had disappeared owing to years of neglect. Higher level co-ordination between the agencies and the states, mediated by the DoLR can lead to the states and the SoI to work in tandem to replace the missing benchmarks and setting up new ones.

It needs to be seen if the recent change which apart from renaming the program, makes it fully centrally funded has addressed these issues or not[5].

4 Future Directions

Creating of land big data requires coordination of efforts between the various stakeholders with an adequate understanding and appreciation of the issues. On the specific aspect of the NLRMP, that fact that the program has not been systematically evaluated till date precludes identifying and fixing its shortcomings. Once the program is evaluated, the future directions could involve re-jigging the areas of support and rationalizing the quantum of central government support.

On the broader subject of big data, national policies and legal frameworks need to be developed that can help obviate the many challenges thrown up by indiscriminate data fusion, while ensuring that the benefits of such data fusion can be tapped for development. The architecture of the big data MPC outlined in chapter 4 is a possible framework that amalgamates ideas of privacy preserving data mining (Vaidya, Clifton, & Michael, 2006), reproducibility (Stodden, 2014), data repositories (King, 2011) and data enclaves (Abowd & Lane, 2004) using the virtual data lake approach.

Notes

[1]"Data Is the New Oil of the Digital Economy". Source: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/. Retrieved: May 26, 2017

[2]"The rise of big data: A double-edged sword". Source: http://www.dailyherald.com/article/20140419/business/140418227/. Retrieved: May 26, 2017

[3]Machine learning models have to be trained on some initial data. The model can learn biases if the training data itself is biased. For more, refer to Mitchell (1999).

[4]An example is the depiction of national boundaries. India has specific boundary disputes with its neighbors. While the official maps of India show these to be a part of India, the maps produced by private players often do not. Interestingly, the geographical boundaries differ depending upon where the map is being viewed from.

[5]In January 2016, the NLRMP moved from being a centrally sponsored scheme to a fully central scheme with the new name — DILRMP.

APPENDIX A: CREATING THE DATASET

As discussed in chapter 3, a novel dataset was required to empirically analyze policy adoption at the sub-national level. This dataset was created by combining together information provided by the National Land Records Modernisation Programme (NLRMP) Management Information System (MIS) with other datasets created by combining and collating various socioeconomic indicators.

A.1 Data from the National Land Records Modernisation Programme Management Information System

The data from the NLRMP MIS website were scraped using a custom web-scraper and the HTML code then converted into a format suitable for analysis. This required analyzing the Hypertext Markup Language (HTML) code to create a set of values in Comma Separated Values (CSV) format. However, these data required further processing, primarily as the districts in the NLRMP MIS do not exactly match the districts in the 2011 Census of India. This is due to variant spellings or misspellings, as well as the creation of new states[1] and districts after the 2011 census enumeration.

A.2 Socioeconomic Indicators

Data on socioeconomic indicators was scattered around and in diverse formats (for example in Adobe Portable Document Format (PDF), Microsoft Excel format (XLS) or Comma Separated Values (CSV)), which all needed to be matched and integrated into a singular dataset. These data were sourced from multiple organizations which resulted in a lack of congruence between the individual datasets leading to the need to triangulate and manually clean parts of these data. The various data sources used are given in section A.2.1:

A.2.1 Data Sources

1. Census of India, 2011 abstracts available at http://www.censusindia.gov.in/. These abstracts were used to get a canonical list of districts and a program written in Python to match the state and district names in other data sources.

2. State and district level statistics collated by the Niti Aayog available at http://niti.gov.in/content/state-stats.php and http://niti.gov.in/content/district_wise_statistics.php[2]. These data were cleaned and processed using custom computer programs. The list of indices is given in Table A.1

3. Agricultural Census of India, 2010. The data are available at http://agcensus.nic.in/ and http://agcensus.dacnet.nic.in in Adobe Portable Document Format (PDF) and Microsoft Excel format (XLS) formats. A custom web-scraper was written to automatically download the various tables,

which were then combined together using a program written in Python.

4. Raghuram Rajan Report, 2013 or the *Report of the Committee for Evolving a Composite Development Index of States* (Ministry of Finance, Government of India, 2013). This is used for the state level development indices as discussed in Appendix B.

Table A.1: District Level Development Indicators and their Sources. Indicator identifiers (A.1, A.2, B.1…E.3 are given in the "Indicators" column)

Sl. No.	Category	Data Source	Indicators
A	Health	District Census Handbook, Census of India (2011)	A.1: Sex Ratio (Number of females per 1000 males)
		District Level Health Survey, 4th Round (DLHS-4) (2012-13)	A.2–A.25: 24 Health Indicators
B	Education	District Census Handbook, Census of India (2011)	B.1: Literacy Rate
		District Information System for Education (DISE) (2013-14)	B.2–B.13: 12 Education and Schooling Indicators
C	Electricity	District Level Health Survey, 4th Round (DLHS-4) (2012-13)	C.1: Households with electricity (%)
		House-listing and Housing Census Data Tables- District Level (HH-7), Census of India (2011)	C.2–C.5: 4 Household Level Indicators of Electricity
D	Water	District Level Health Survey, 4th Round (DLHS-4) (2012-13)	D.1: Households with improved source of drinking water (%)
		House-listing and Housing Census Data Tables- District Level (HH-6), Census of India (2011)	D.2–D.5: 4 Household Level Indicators of Water
E	Tele-communications	District Census Handbook, Census of India (2011)	E.1–E.3: 3 Indicators of Information and Communication Technologies Penetration

Notes

[1]The state of Telangana was formed in 2014 by dividing the state of Andhra Pradesh and in this study, it is treated to be a part of Andhra Pradesh.

[2]As of February 20, 2017 the location of the Niti Aayog district data has changed to http://niti.gov.in/best-practices/district-wise-statistics and it is available as a set of Microsoft Excel spreadsheets (one for each district).

APPENDIX B: STATE DEVELOPMENT INDEX

This study uses a State Development Index based on the composite State *Under*-Development Index developed by the 2013 committee headed by Dr. Raghuram Rajan (Ministry of Finance, Government of India, 2013). This committee created a simple (*under*) development index by averaging ten sub-components, namely: (a) Monthly per-Capita Consumption Expenditure (MPCE), (b) education, (c) health, (d) household amenities, (e) poverty rate, (f) female literacy, (g) percent of SC-ST population, (h) urbanization rate, (i) financial inclusion, and (j) connectivity. This index is normalized to have a value between 0 and 1, with *higher values signifying a greater level of deprivation*.

Although, there has been certain criticism of the index developed by the committee on its choice of sub-components as well as on methodological grounds[1], this is the best index currently available for sub-national deprivation in India. Considering the criticism by one of the committee members, the report lists out two different under-development indices, one using Monthly per-Capita Consumption Expenditure (MPCE) and the other using the Net State Domestic Product (NSDP)[2].

For the purpose of this study, the need is for a "development" index, or an index

where a *higher value signifies a greater level of development*. This *development index* is got by subtracting the *under*-development index from 1.0. We use the NSDP based index in this study as it is better captures states' development capacities and capabilities[3].

Table B.1 lists the under-development and development indices along with state rankings. The various columns of this table are:

$IDX_{UNDERDEV}$: NSDP based *Under*-Development Index.

IDX_{DEV}: NSDP based Development Index (referred to as $\mathsf{DevIDX}_{\mathsf{STATE}}$ in chapter 3). $IDX_{DEV} = 1.0 - IDX_{UNDERDEV}$

$RANK_{IDX\_DEV}$: the state's ranking based on the IDX_{DEV}

Table B.1: State Under-Development and Development Indices and Rankings based on the Raghuram Rajan Committee Report Ministry of Finance, Government of India (2013)

STATE	**$IDX_{UNDERDEV}$**	**IDX_{DEV}**	**$RANK_{IDX\_DEV}$**
Andhra Pradesh	0.54	0.46	15
Arunachal Pradesh	0.74	0.26	23
Assam	0.71	0.29	22
Bihar	0.75	0.24	26
Chattisgarh	0.74	0.26	24
Goa	0.05	0.95	1
Gujarat	0.50	0.50	12
Haryana	0.43	0.57	9
Himachal Pradesh	0.42	0.58	8
Jammu & Kashmir	0.53	0.47	14
Jharkhand	0.74	0.26	25
Karnataka	0.48	0.52	11
Kerala	0.15	0.85	2
Madhya Pradesh	0.76	0.24	27
Maharashtra	0.37	0.63	4
Manipur	0.58	0.42	18
Meghalaya	0.70	0.30	21
Mizoram	0.52	0.48	13
Nagaland	0.57	0.43	17
Odisha	0.79	0.21	28
Punjab	0.39	0.61	5
Rajasthan	0.65	0.35	19
Sikkim	0.41	0.59	7
Tamil Nadu	0.36	0.64	3
Tripura	0.47	0.53	10
Uttar Pradesh	0.65	0.35	20
Uttarakhand	0.39	0.61	6
West Bengal	0.56	0.44	16

Notes

[1]A dissent note of one of the committee members (Mr. Shaibal Gupta) is available as pages 40–49 of the report, while Debroy (2013, October 27) and Singh (2014, December 1) provide a broader criticism of the index.

[2]The use of the Net State Domestic Product (NSDP) instead of Gross State Domestic Product (GSDP) was also criticized by Mr. Shaibal Gupta. See note 1.

[3]See note 1.

APPENDIX C: DISTRICT DEVELOPMENT INDEX

For this study, it was necessary to quantify the level of development at the district level in the form of a composite District Development Index similar to the state level development index discussed earlier in appendix B. However, no such index is available at the national level for all the districts of the country, necessitating development of such an index. The a District Development Index has been created relying on the format used by the 2013 Raghuram Rajan Committee's *Report of the Committee for Evolving a Composite Development Index of States* (Ministry of Finance, Government of India, 2013). The Raghuram Rajan Committee used the statistical technique of Principal Component Analysis (PCA) for creating its index. In the literature (Krishnan, 2010; Nardo et al., 2005, August 9), Principal Component Analysis (PCA) has been commonly used as the statistical method to create indices and the same technique is also used here.

This index is based on the list of district level indicators shown in table C.1, which is based on data provided by the Niti Aayog (see appendix A). The four main steps in creating this index are: (a) dataset preparation, (b) indicator selection, (c) data transformation (if needed), and (d) sub-index creation and index creation.

Table C.1: District Level Development Indicators and their Sources. Indicator identifiers (A.1, A.2, B.1…E.3 are given in the "Indicators" column)

Sl. No.	Category	Data Source	Indicators
A	Health	District Census Handbook, Census of India (2011)	A.1: Sex Ratio (Number of females per 1000 males)
		District Level Health Survey, 4th Round (DLHS-4) (2012-13)	A.2–A.25: 24 Health Indicators
B	Education	District Census Handbook, Census of India (2011)	B.1: Literacy Rate
		District Information System for Education (DISE) (2013-14)	B.2–B.13: 12 Education and Schooling Indicators
C	Electricity	District Level Health Survey, 4th Round (DLHS-4) (2012-13)	C.1: Households with electricity (%)
		House-listing and Housing Census Data Tables- District Level (HH-7), Census of India (2011)	C.2–C.5: 4 Household Level Indicators of Electricity
D	Water	District Level Health Survey, 4th Round (DLHS-4) (2012-13)	D.1: Households with improved source of drinking water (%)
		House-listing and Housing Census Data Tables- District Level (HH-6), Census of India (2011)	D.2–D.5: 4 Household Level Indicators of Water
E	Telecommunications	District Census Handbook, Census of India (2011)	E.1–E.3: 3 Indicators of Information and Communication Technologies Penetration

C.1 Dataset Preparation

Prior to performing any analysis, it has to ensured that the values in the dataset make sense and that missing values have been adequately taken care of. Further, as the sub-index and index creation will use the statistical technique of PCA, the values have to be standardized so as to lie on the same scale.

Table C.2: Missing values in the initial dataset

Statistic	N	Mean	St. Dev.	Median	Min	Max
Count of Not Available Values	141	33.17	37.09	11.95	0.00	100.00

C.1.1 Identify and Fix Missing Values

From Table C.2, we note that on average, 33% of the observations (across 141 variables) are missing data, with half the variables missing almost 12% observations. Therefore, in order to get meaningful data values, it is necessary to drop variables that have a high number of missing values. The threshold used for dropping variables from consideration is 7% — that is we drop variables that

Table C.3: Missing values in the dataset after dropping variables where Not Available > 7%

Statistic	N	Mean	St. Dev.	Median	Min	Max
Count of Not Available Values	61	0.68	0.53	0.48	0.00	1.78

are missing data for more than 7% of the observations. This reduces the field to 61 variables, that miss an average of 0.68% of the values, with the maximum being 1.78% (Table C.3). A the side-effect of this exercise is that all the 24 health indicators sourced from the District Level Health Survey, 4th Round (DLHS-4) conducted between 2012-13 have to be dropped from consideration.

C.1.2 Impute Missing Values

For PCA to work, there cannot be any missing values in the dataset. This means that values have to be imputed to the the observations where values are not available. We choose to impute missing values as the state level median value for the missing observation (Kabacoff, 2015). Using the state level observation(s) instead of the national level ensures that varying state conditions are taken care of. Further, as it is mainly the less developed states where data are missing, this ensures that the values are not inflated which they could be if the national media was used.

C.2 Indicator Selection

The available indicators are grouped into categories that they best represent. As discussed in appendix B, the Raghuram Rajan Committee used ten sub-indices. However, the Niti Aayog district level data does not have data for all these sub-indices, neither are all the sub-indices equally relevant for a development index[1]. The sub-indices chosen for creating the district development index are (a) health, (b) education, (c) education infrastructure, (d) electricity, (e) water, and (f) telecommunications (or ICT). However, as mentioned in section C.1.1, the cleaned dataset is missing health data, leaving us with five sub-indices. These sub-indices and the indicators for the sub-index are given in table C.4. As the unit of analyses for creating this index is the district, only the district-level aggregate indicators are used[2]. However, before these data can be used to create the sub-indices and the index, certain transformations are required to be applied which are discussed next.

Table C.4: Sub-Indices and nineteen indicators forming part of the District Development Index

Sub-Index	**Indicator(s)**
Education	Literacy Rate (district level); Net and Gross Enrollment rates at the primary and upper-primary school levels
Education Infrastructure	Primary school pupil-teacher and student-classroom ratio; new primary government schools since 2003; primary schools with boys/girls toilets, drinking water facility and electricity;
Electricity	Households with electricity as main light source or no lighting (district level).
Water	Households with main source of drinking water within in premises and those receiving treated water in premises (district level).
Telecommunications	Households with internet, landline and mobile phone connections (district level).

C.3 Data Transformations

C.3.1 Monotonic Indicators

For calculation of the development index, all indicators need to be monotonic. A set of indicators is monotonic if either “higher is better”, or “lower is better” for all indicators, either larger values will signify more development, or smaller values will indicate more development. The signify consistent in whether we need to have indicators where either “higher is better”, or “lower is better” for all

of them. However, this is not the case with our data. Although "higher is better" for most of the indicators, the converse is true for a select set of indicators. These indicators have to be transformed to get values where higher continues to be better. These transformations are discussed next.

Pupil-Teacher Ratio (PTR) The PTR is transformed into an index that benchmarks the districts performance in terms of the 2008 Right to Education Act target PTR of 30:1 at the primary level[3]. This index is calculated as:

$$PTR_{idx} = \frac{PTR_{target} - PTR_{district}}{PTR_{target}} \tag{C.1}$$

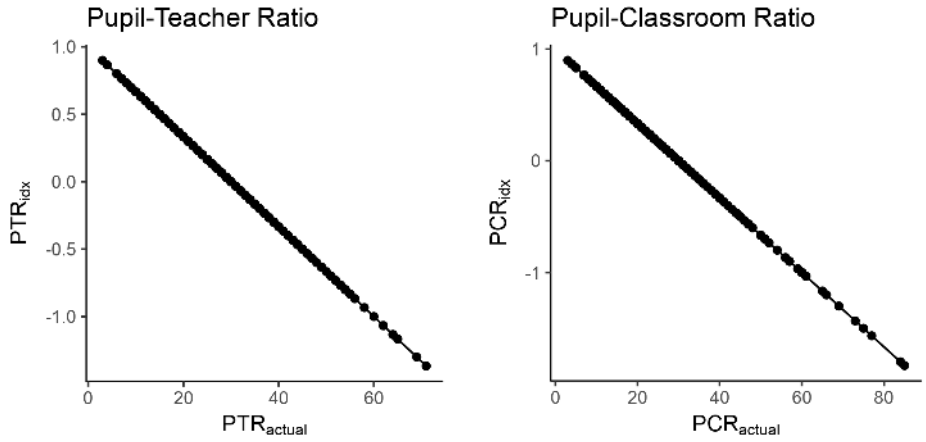
where $PTR_{target} = 30$. The linear nature of this transformation can be seen in Figure C.1a which plots the PTR_{idx} against PTR_{actual}.

Pupil-Classroom Ratio (PCR) The PCR is transformed into an index using a formula similar to the one in Equation C.1. Specifically:

$$PCR_{idx} = \frac{PCR_{target} - PCR_{district}}{PCR_{target}} \tag{C.2}$$

where $PCR_{target} = PTR_{target} = 30$. There is no target PCR specified and the target PTR is used as a proxy, assuming that the number of classrooms will be equal to the number of teachers. Again, Figure C.1b (plot of PCR_{idx} against PCR_{actual}) shows the linear nature of this transformation.

Percentage Households with NO Lighting is used as a measure of the electricity scenario of a district. This is an indicator where "lower is better", and it is transformed into a "higher is better" indicator by multiplying with -1.



(a) Pupil Teacher Ratio (Index vs Actual)

(b) Pupil Classroom Ratio (Index vs Actual)

Figure C.1: Plots of Pupil-Teacher and Pupil-Classroom Ratio Indices against the actual values showing that the transformation is linear.

C.3.2 Outlier Management

The presence of outliers is an issue as they tend to shrink the variation between most of the values when they are constrained to lie between a fixed interval $[0, 1]$ as in the case of index creation. Figure C.2 on page 241 is a boxplot of a few selected indicators, namely (a) Upper Primary Gross Enrollment Ratio (GER_UP) (b) Index of Primary level Pupil-Teacher Ratio (PTR_P_IDX) (c) Percentage of households with Internet connected computer (Computer) (d) Percentage of households with no lighting (NoLighting) (e) Percentage of households receiving treated tap water within premises (TreatedTapWater). From the figure, we note how the presence of outliers distorts the scale, while figure C.3 on the next

page shows how top-coding the outliers (as discussed below) brings all values to the same scale. Stem and Leaf plots of the selected indicators are shown in figures C.4, C.6, C.8, C.10, and C.12 on pages 242–246 and indicate how the presence of outliers tends to compress most of the values into an extremely tiny range.

Outliers are managed by top-coding (assigning one value to all data whose value is *above* a threshold) and bottom-coding (assigning one value to all data whose value is *below* a threshold). In this case, the lower threshold is the 5^{th} percentile, while the 95^{th} percentile has been taken to be the upper threshold. This means that any values that are greater than the 95^{th} percentile are replaced with the value of the 95^{th} percentile, while all values smaller than the 5^{th} percentile are replaced with the 5^{th} percentile.

Figures C.5, C.7, C.9, C.11, and C.13 on pages 242–246 are stem and leaf plots of the selected indicators after the outliers have been trimmed. Comparing the "*before* trimming" and "*after* trimming" plots shows how top- and bottom-coding the outliers tends to evenly spread out the values, allowing for broader variation among the values.

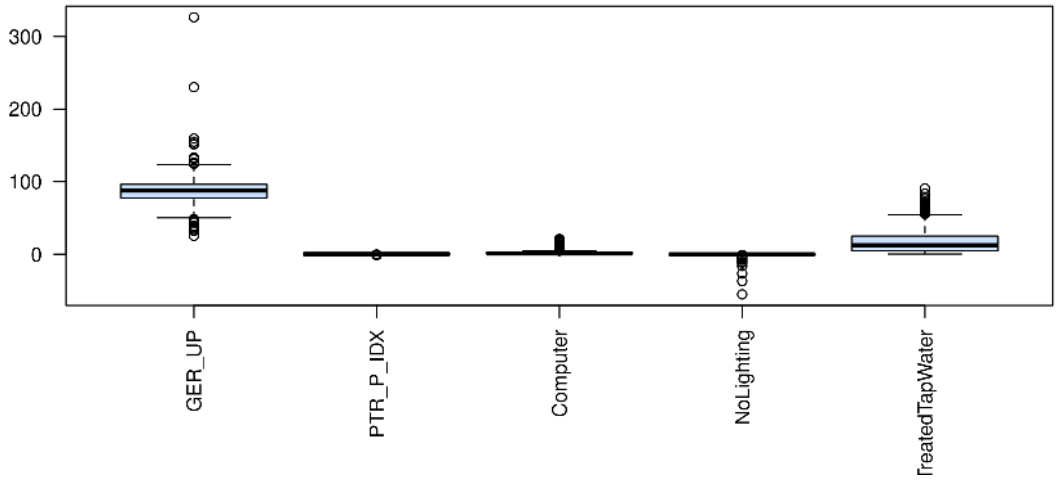


Figure C.2: Boxplot of selected indicators before trimming outliers

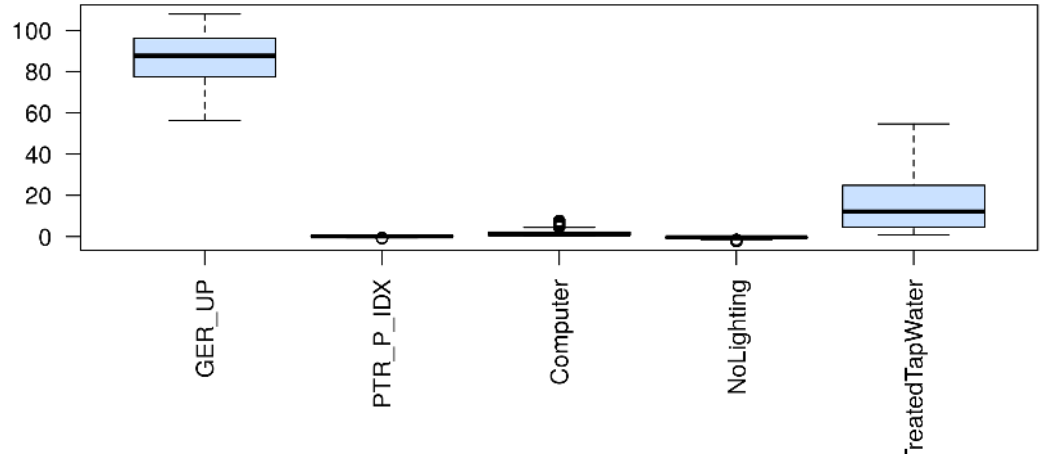


Figure C.3: Boxplot of selected indicators after trimming outliers

Figure C.4: Stem and leaf plot of Upper Primary Gross Enrolment Ratio (before trimming)

```
The decimal point is 1 digit(s) to the left of the |

0 | 02334556677888889999999
1 | 000000000011111122222222223333333333444444444444444444555555555555+139
2 | 000000000000000000000000000000000000000000000000000011111111111111111+281
3 | 122223456
4 | 2235
5 |
6 | 8
7 |
8 |
9 |
10 | 0
```

Figure C.5: Stem and leaf plot of Upper Primary Gross Enrolment Ratio (after trimming)

```
The decimal point is 1 digit(s) to the left of the |

0 | 00000000000000000000000000000000124677778888899
1 | 022333444677899
2 | 00111122333344444555566677799
3 | 000000001111122223333444445555555555566677788899999
4 | 0000000011111112222233334444444555555555566666677777888888888999999
5 | 000000111111111222222222233333334444445555555555666666677777777888+4
6 | 000000000001111111111112222333333333444444444455555566666677777777888+2
7 | 00000001111111111111112222222233333333333344444444444444444444555555+20
8 | 000000000000000000011111122222223333333333444444444445555556667788 8
9 | 000011111112222222233334455 6799
10 | 00000000000000000000000000000000
```

Figure C.6: Stem and leaf plot of PC ownership (before trimming)

The decimal point is 1 digit(s) to the left of the |

```
0 | 000111111111111111111111111111111111111111111112222222222222222222222222+371
1 | 000000000000000000011111111111111111111222222222222223333333334444445+18
2 | 00000000000011122233556688899
3 | 01112255678889
4 | 000115788
5 | 034489
6 | 238
7 |
8 | 0117
9 | 145
10 | 0
```

Figure C.7: Stem and leaf plot of PC ownership (after trimming)

The decimal point is 1 digit(s) to the left of the |

```
0 | 00000000000000000000000000000000000000000011111111111111111111222222+220
1 | 00000000000000000000011111111111111111222222222222222223333333333334+44
2 | 000000111111111222223333334444444555555566666777777778888889999999
3 | 00000011222223333344566677788899
4 | 0112244566668999
5 | 011112222233334444444556689
6 | 0134
7 | 00116789
8 | 1156688
9 | 67
10 | 00000000000000000000000000000000
```

Figure C.8: Stem and leaf plot of households without lighting (before trimming)

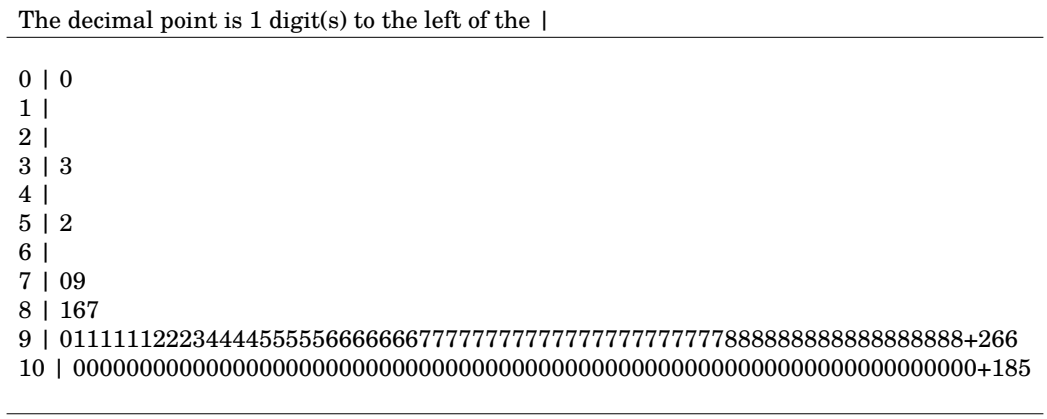


Figure C.9: Stem and leaf plot of households without lighting (after trimming)

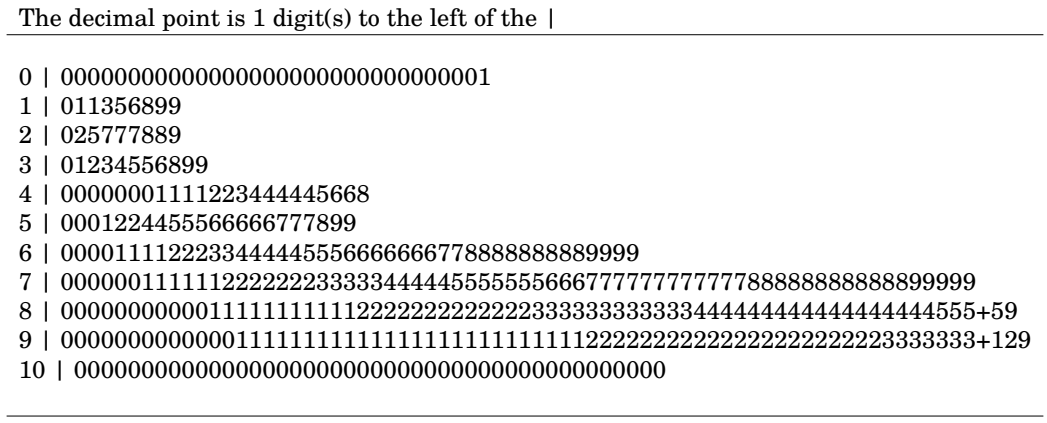


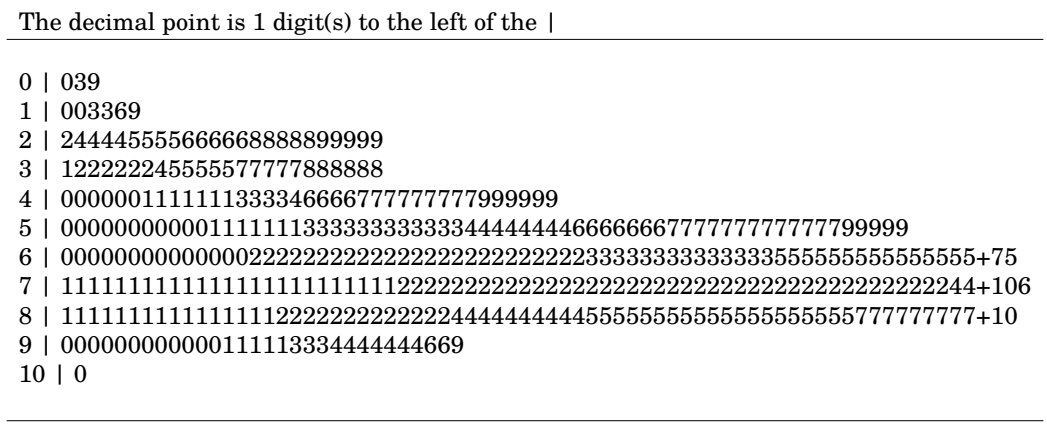
Figure C.10: Stem and leaf plot of households receiving treated tap water (before trimming)

```
The decimal point is 1 digit(s) to the left of the |

0 | 00000000000000001111111111111111111111111111111111111111111111111222+164
1 | 00000000000000000001111111111111112222222222222222223333333333333333+71
2 | 00000000001111111111122222222333344444444455555555555566666677777+8
3 | 00000011122222233333333334444555566677777788888899
4 | 01112233344445556677777889999
5 | 01111111123334556678999
6 | 011222333557899
7 | 001223445569
8 | 047
9 | 2
10 | 0
```

Figure C.11: Stem and leaf plot of households receiving treated tap water (after trimming)

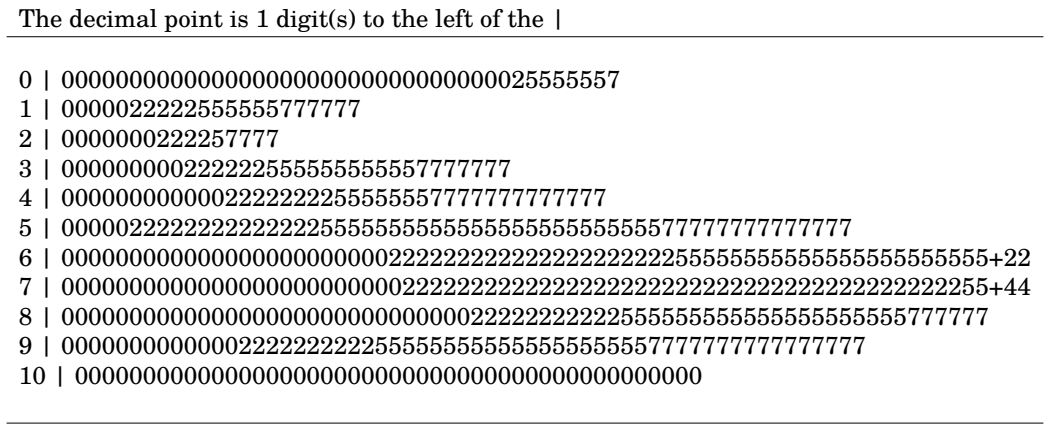
```
The decimal point is 1 digit(s) to the left of the |

0 | 000000000000000000000000000000000000000000001111111111111111111+107
1 | 00000111111111111112222222222233333333344444444444444455555555556+28
2 | 00000000111111111111222222222223333333334444444455666666677777777+1
3 | 0000000001111111122222222222334444444555556666678888999999999
4 | 00000111111112223333334445555556666777777778889999
5 | 000112233334444444456667777799
6 | 00011112222333446788 9
7 | 0112233355566788889
8 | 001112444555556789
9 | 0113346788
10 | 00000000000000000000000000000000
```

Figure C.12: Stem and leaf plot of pupil teacher ratio index (before trimming)



```
The decimal point is 1 digit(s) to the left of the |

0 | 039
1 | 003369
2 | 24444555666668888899999
3 | 12222224555577777888888
4 | 000000111111333346666777777777999999
5 | 000000000001111111333333333333444444446666666777777777777799999
6 | 000000000000002222222222222222222222223333333333333335555555555555+75
7 | 11111111111111111111111112222222222222222222222222222222222222244+106
8 | 11111111111111112222222222224444444445555555555555555555777777777+10
9 | 00000000000011111334444444669
10 | 0
```

Figure C.13: Stem and leaf plot of pupil teacher ratio index (after trimming)



```
The decimal point is 1 digit(s) to the left of the |

0 | 00000000000000000000000000000025555557
1 | 0000022222555555777777
2 | 0000000222257777
3 | 000000002222225555555557777777
4 | 00000000000222222225555557777777777777
5 | 000002222222222222255555555555555555555555577777777777777
6 | 0000000000000000000000022222222222222222222555555555555555555555+22
7 | 00000000000000000000000022222222222222222222222222222222222222255+44
8 | 00000000000000000000000000000222222222225555555555555555555777777
9 | 00000000000022222222255555555555555555557777777777777777
10 | 0000000000000000000000000000000000000000000000
```

C.3.3 Data Standardization

As discussed earlier, it is suggested that the data used for PCA be standardized so as to lie along the same scale. The standardization of indicators is done by computing their $z-scores$ (Agresti & Finlay, 2009; Nardo et al., 2005, August 9) so as to have a mean $\bar{x} = 0$ and standard deviation $\sigma_x = 1$. For an indicator x , the $z-score$ $(\hat{x})$ for an indicator is given by:

$$\hat{x} = \frac{x - \bar{x}}{\sigma_x}$$

where: $\bar{x}$: mean of indicator x, and σ_x is its standard deviation.

C.4 Sub-Index and Index Creation

The nineteen indicators that have been identified earlier (table C.4 on page 237) are combined together to create five sub-indices. The biggest challenge in combining multiple indicators into a single number is to identify the relative weights of each indicator in the final number. Multiple methods have been proposed in the literature from giving equal weights to all indicators (a simple mathematical average) (Ministry of Finance, Government of India, 2013) to using the statistical technique of Principal Component Analysis (PCA) to have a statistical basis for identifying the appropriate weights (Krishnan, 2010; Nardo et al., 2005, August 9)[4]. Owing to the nature of the indicators, it is not considered appropriate to weight them equally, and thus for creating the District Development Index, the PCA technique is used to identify the appropriate weights. The PCA

is performed using a process based on Kabacoff (2015) using the psych package in R (Revelle, 2016).

As discussed in section §C.2, we select nineteen indicators and perform a PCA to create five sub-indices, one each for (a) education (b) education infrastructure, (c) electricity, (d) water, and (e) telecommunications (or ICT). These sub-indices are further combined into a composite index by performing another PCA on them which results in a *Non-standardized index* (Krishnan, 2010) called the *Non-standardized District Development Index* ($\mathsf{DEVIDX}_{\mathsf{DIST}_{\mathsf{nsi}}}$) . This $\mathsf{DEVIDX}_{\mathsf{DIST}_{\mathsf{nsi}}}$ is based on *z-scores* (section C.3.3) and has to be converted into a *Standardized District Development Index* ($\mathsf{DEVIDX}_{\mathsf{DIST}}$) which lies in the interval $[0, 1]$. The normalization process to create the $\mathsf{DEVIDX}_{\mathsf{DIST}}$ from the $\mathsf{DEVIDX}_{\mathsf{DIST}_{\mathsf{nsi}}}$is:

$$DEVIDX_{DIST} = \frac{DEVIDX_{DIST_{nsi}} - \min(DEVIDX_{DIST_{nsi}})}{\max(DEVIDX_{DIST_{nsi}}) - \min(DEVIDX_{DIST_{nsi}})}$$

C.5 Conclusion

The above steps result in a District Development Index whose summary statistics are provided in Table table C.5

Table C.5: Summary Statistics of $DEVIDX_{DIST}$

Statistic	N	Mean	St. Dev.	Median	Min	Max
$DEVIDX_{DIST}$	619	0.47	0.22	0.44	0	1

Notes

[1]See note 1 in appendix B.

[2]The Niti Aayog data set also contains data separately for the rural and urban parts of the district.

[3]Azim Premji Foundation, "Pupil-Teacher Ratios in Schools and their Implications", February 2014 available at: http://www.azimpremjifoundation.org/pdf/PTR%20report.pdf. Retrieved: 21 February, 2017.

[4]The Raghuram Rajan Committee (Ministry of Finance, Government of India, 2013) created its ten sub-indices using Principal Component Analysis (PCA) and weighted the indicators within each sub-index based on the output of the PCA. However, it eschewed the PCA when it came to combining the sub-indices to form the final composite index and decided to go with an arithmetic mean instead, despite having done a Principal Component Analysis (PCA). Further, it should be noted that other methods of constructing indices, for example the geometric mean (used in calculating the Human Development Index (HDI)) cannot be used in this case as a single indicator with a value of 0 will make the entire index 0.

REFERENCES

Abowd, J. M. & Lane, J. (2004). New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in statistical databases* (Vol. 3050, pp. 282–289). Lecture Notes in Computer Science. Springer Berlin Heidelberg. doi:10.1007/978-3-540-25955-8_22

Agresti, A. & Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Pearson Education. Pearson Prentice Hall.

Agricultural Census Division. (2014, February 28). *Agriculture Census, 2010–11*. Department of Agriculture & Co-operation, Ministry of Agriculture, Government of India. Retrieved from http://agcensus.nic.in/document/agcensus2010/CompleteReport.pdf

Allen, M. D., Pettus, C., & Haider-Markel, D. P. (2004). Making the national local: specifying the conditions for national government influence on state policymaking. *State Politics & Policy Quarterly*, *4*(3), 318–344. doi:10.1177/153244000400400304. eprint: http://spa.sagepub.com/content/4/3/318.full.pdf+html

Ansolabehere, S. & Hersh, E. (2012). Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate. *Political Analysis*, *20*(4), 437–459. doi:10.1093/pan/mps023. eprint: http://pan.oxfordjournals.org/content/20/4/437.full.pdf+html

Aragón, P., Kappler, K. E., Kaltenbrunner, A., Laniado, D., & Volkovich, Y. (2013). Communication dynamics in twitter during political campaigns: the case of the 2011 Spanish national election. *Policy & Internet*, *5*(2), 183–206. doi:10.1002/1944-2866.POI327

Baden-Powell, B. H. (1907). *A Short Account of the Land Revenue and Its Administration in British India: With a Sketch of the Land Tenures* (Second) (T. W. Holderness, Ed.). Oxford. Retrieved from https://books.google.com/books?id=3U0ZAAAAYAAJ

Baden-Powell, B. H. (1892a). *The Land-systems of British India*. The Land-systems of British India: Being a Manual of the Land-tenures and of the Systems of Land-revenue Administration Prevalent in the Several Provinces. Oxford.

Baden-Powell, B. H. (1892b). *The Land-Systems of British India: Book I. General. Book II. Bengal*. The Land-systems of British India: Being a Manual of the Land-tenures and of the Systems of Land-revenue Administration Prevalent in the Several Provinces. Oxford. Retrieved from https://books.google.com/books?id=qfRKAAAAYAAJ

Baden-Powell, B. H. (1892c). *The Land-Systems of British India Vol. II. Book III. The System of Village or Mahál Settlements*. The Land-systems of British India: Being a Manual of the Land-tenures and of the Systems of Land-revenue Administration Prevalent in the Several Provinces. Oxford. Retrieved from https://books.google.com/books?id=6vVKAAAAYAAJ

Banerjee, A. & Iyer, L. (2005). History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India. *The American Economic Review*, *95*(4), 1190–1213. Retrieved from http://www.jstor.org/stable/4132711

Bardach, E. (2012). *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. C Q Press.

Barocas, S. (2012). Big data are made by (and not just a resource for) social science and policy-making. Retrieved from http://ipp.oii.ox.ac.uk/2012/programme-2012/track-c-data-methods/panel-1c-what-is-big-data/solon-barocas-big-data-are-made-by-and

Barocas, S. & Nissenbaum, H. (2014). Big Data'a End Run around Anonymity and Consent. In *Privacy, Big Data, and the Public Good* (pp. 44–75). Cambridge University.

Barocas, S. & Selbst, A. D. (2014). Big Data's Disparate Impact. *Available at SSRN: (October 19, 2014)*. Retrieved from http://ssrn.com/abstract=2477899

Benjamin, D. S., Bhuvaneswari, R., Rajan, P., & Manjunatha. (2007). Bhoomi: 'E-Governance', or an Anti-Politics Machine Necessary to Globalize Bangalore? Retrieved from https://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf

Bennett, R., Rajabifard, A., Williamson, I., & Wallace, J. (2012). On the need for national land administration infrastructures. *Land Use Policy*, *29*(1), 208–219. doi:http://dx.doi.org/10.1016/j.landusepol.2011.06.008

Bennett, R., Tambuwala, N., Rajabifard, A., Wallace, J., & Williamson, I. (2013). On recognizing land administration as critical, public good infrastructure. *Land Use Policy*, *30*(1), 84–93. doi:http://dx.doi.org/10.1016/j.landusepol.2012.02.004

Bennett, R., Wallace, J., & Williamson, I. (2008). Organising land information for sustainable land administration. *Land Use Policy*, *25*(1), 126–138. doi:http://dx.doi.org/10.1016/j.landusepol.2007.03.006

Berry, F. S. & Berry, W. D. (2014). Innovation and diffusion models in policy research. In P. A. Sabatier & C. M. Weible (Eds.), *Theories of the policy process* (Third, Chap. 9, pp. 307–359). Westview Press.

Berry, F. S. & Berry, W. D. (1990). State lottery adoptions as policy innovations: an event history analysis. *The American Political Science Review*, *84*(2), 395–415. Retrieved from http://www.jstor.org/stable/1963526

Besley, T. & Burgess, R. (2000). Land Reform, Poverty Reduction, and Growth: Evidence from India. *The Quarterly Journal of Economics*, *115*(2), 389–430. Retrieved from http://www.jstor.org/stable/2586998

Bhatti, Y., Olsen, A. L., & Pedersen, L. H. (2011). Administrative professionals and the diffusion of innovations: the case of citizen service centres. *Public Administration*, *89*(2), 577–594. doi:10.1111/j.1467-9299.2010.01882.x

Bhidé, A. (2008). What holds back bangalore businesses?. *Asian Economic Papers*, *7*(1), 120–153.

Borne, K. D. (2013). What is Data Science and Why is it Needed? Learning From Data, Big and Small. Retrieved from http://complex.gmu.edu/www-phys/colloquium/Fall_2013/Fall%5C%202013%5C%20abstracts/kborne-SPACS-2013nov14.pdf

Bostick, C. D. (1987). Land Title Registration: An English Solution to an American Problem. *Indiana Law Journal*, *63*(1), 55–112. Retrieved from http://heinonline.org/HOL/Page?handle=hein.journals/indana63%5C&id=65

boyd, d. & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878. eprint: http://www.tandfonline.com/doi/pdf/10.1080/1369118X.2012.678878

Brooks, F. (1995). *The mythical man-month : essays on software engineering*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Buhler, D. & Cowen, D. J. (2010). The United States Mortgage Crisis and Cadastral Data. In *XXIV FIG International Congress*. Sydney, Australia: 11– 16 April, 2010. Retrieved from http://www.fig.net/resources/proceedings/fig_proceedings/fig2010/papers/ts07a/ts07a_cowen_buhler_4022.pdf

Bussell, J. (2012). *Corruption and reform in india : public services in the digital age*. Cambridge: Cambridge University Press.

Capurro, R. (2005). Privacy. An Intercultural Perspective. *Ethics and Information Technology*, *7*(1), 37–47. doi:10.1007/s10676-005-4407-4

Chanakya, V. (n.d.). Kautilya's arthashaastra. Retrieved from https://ia802703.us.archive.org/13/items/Arthasastra_English_Translation/Arthashastra_of_Chanakya_-_English.pdf

Chawla, R. & Bhatnagar, S. (2004). Online Delivery of Land Titles to Rural Farmers in Karnataka, India: A Global Learning Process and Conference. In *Scaling up poverty reduction*: *Case studies in scaling up poverty reduction*. The World Bank. Shanghai.

Choudhury, P. R., Rao, G. V., Kumar, K., Deo, B., & Dash, T. (2016, March 17). Community Mapping through An Android based GIS application: An attempt towards Inclusive, Transparent and Participatory mapping

of Community Forest Rights in India, March 17, 2016. The World Bank. Annual World Bank Conference on Land and Poverty. Washington, DC. Retrieved from https://www.conftool.com/landandpoverty2016/index.php/Choudhury-664-664_paper.pdf?page=downloadPaper&filename=Choudhury-664-664_paper.pdf&form_id=664&form_version=final

Clark, J. (1985). Policy diffusion and program scope: research directions. *Publius*, *15*(4), 61–70. Retrieved from http://www.jstor.org/stable/3330042

Cook, R. N. (1969). Land Law Reform: A modern Computerized System of Land Records. *University of Cincinnati Law Review*, *38*(3), 385–448. Retrieved from http://heinonline.org/HOL/Page?public=false%5C&handle=hein.journals/ucinlr38%5C&id=401

Cook, T. D. (2014). "BIG DATA" IN RESEARCH ON SOCIAL POLICY. *Journal of Policy Analysis and Management*, *33*(2), 544–547. doi:10.1002/pam.21751

Crawford, K. & Finn, M. (2014). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, 1–12. doi:10.1007/s10708-014-9597-z

Cukier, K. N. & Mayer-Schöenberger, V. (2013). Rise of Big Data: How it's Changing the Way We Think about the World, The. *Foreign Aff.* *92*, 28. Retrieved from http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data

Dale, P. (1997). Land Tenure Issues in Economic Development. *Urban Studies*, *34*(10), 1621–1633. doi:10.1080/0042098975376. eprint: http://usj.sagepub.com/content/34/10/1621.full.pdf+html

Dale, P. & Mclaughlin, J. D. (1999). *Land Administration*. New York: Oxford University Press.

Daley, D. M. & Garand, J. C. (2005). Horizontal diffusion, vertical diffusion, and internal pressure in state environmental policymaking, 1989-1998. *American Politics Research*, *33*(5), 615–644. doi:10.1177/1532673X04273416. eprint: http://dx.doi.org/10.1177/1532673X04273416

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., … Chuang, I. (2014). Privacy, Anonymity, and Big Data in the Social Sciences. *Commun. ACM*, *57*(9), 56–63. doi:10.1145/2643132

David, P. A. (1985). Clio and the economics of qwerty. *The American Economic Review*, *75*(2), 332–337. Retrieved from http://www.jstor.org/stable/1805621

Debroy, B. (2013, October 27). Why Raghuram Rajan Ranked Gujarat Low: His report on states lacks rigour and is flawed. *Business Today*.

Decker, P. T. (2014). Presidential Address: False Choices, Policy Framing, and the Promise of "Big Data". *Journal of Policy Analysis and Management*, *33*(2), 252–262. doi:10.1002/pam.21755

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J.-M., Krings, G., Gutierrez, T., … Luengo-Oroz, M. A. (2014, November 22). Estimating Food Consumption and Poverty Indices with Mobile Phone Data. *preprint arXiv:1305.3212*. arXiv: http://arxiv.org/abs/1412.2595v1 [cs.CY, physics.soc-ph]

Deininger, K. & Goyal, A. (2012). Going digital: credit effects of land registry computerization in India. *Journal of Development Economics*, *99*(2), 236–243. doi:10.1016/j.jdeveco.2012.02.007

Deininger, K., Jin, S., & Nagarajan, H. K. (2009). Land Reforms, Poverty Reduction, and Economic Growth: Evidence from India. *The Journal of Development Studies*, *45*(4), 496–521. doi:10.1080/00220380902725670. eprint: http://dx.doi.org/10.1080/00220380902725670

Desouza, K. C. & Jacob, B. (2014). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*. doi:10.1177/0095399714555751. eprint: http://aas.sagepub.com/content/early/2014/11/06/0095399714555751.full.pdf+html

Devarajan, S. (2013). Africa's statistical tragedy. *Review of Income and Wealth*, *59*, S9–S15. doi:10.1111/roiw.12013

Diebold, F. (2012). On the origin (s) and development of the term "big data". *Penn Institute for Economic Research, Pier Working Paper*, 12–37. Retrieved from http://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf

Dolowitz, D. P. & Marsh, D. (2000). Learning from abroad: the role of policy transfer in contemporary policy-making. *Governance*, *13*(1), 5–23. doi:10.1111/0952-1895.00121

DoLR. (2010, December 23). Minutes of 2nd meeting of the Committee, constituted to consider the unit cost for survey/resurvey and updating of survey & settlement records (including ground control network and ground truthing) under the National Land Records Modernization Programme (NLRMP), held on 23rd December, 2010 at 11.30 AM in Committee Room of the Department of Land Resources in NBO Building, Nirman Bhawan, New Delhi. Retrieved November 10, 2016, from http://dolr.nic.in/nlrmp/Report%20of%20the%20Committee%20on%20unit%20cost%20for%20survey-resurvey%20-final.doc

DoLR. (2009a, December 24). *Report of the Committee on State Agrarian Relations and the Unfinished Task in Land Reforms*. Department of Land Resources, Ministry of Rural Development, Government of India. Retrieved from http://dolr.nic.in/Committee%20Report.doc

DoLR. (2009b, April 17). *The National Land Records Modernization Programme (NLRMP): Guidelines, Technical Manuals and MIS*. Department of Land Resources, Ministry of Rural Development, Government of India. Retrieved from http://dolr.nic.in/Guidelines%20NLRMP%2017.4.2009.pdf

DoLR. (2008, August 21). *The National Land Records Modernization Programme (NLRMP) 2008 Cabinet Note*. Department of Land Resources, Ministry of Rural Development, Government of India. Retrieved from http://dolr.nic.in/NLRMP-2008.pdf

Dordan, R. E. (2010). Mortgage Electronic Registration Systems (MERS), Its Recent Legal Battles, and the Chance for a Peaceful Existence. *Loyola Journal of Public Interest Law*, *12*(1), 177–208. Retrieved from http://heinonline.org/HOL/Page?public=false%5C&handle=hein.journals/loyjpubil12%5C&id=181

Dowson, S. E. M. & Sheppard, V. L. O. (1956). *Land Registration* (Second). Colonial Research Publications. Her Majesty's Stationery Office, London.

Enemark, S., Bell, K. C., Lemmen, C., & McLaren, R. (2014). *Fit-For-Purpose Land Administration*. FIG/World Bank. Retrieved from https://www.fig.net/resources/publications/figpub/pub60/Figpub60.pdf

Eyestone, R. (1977). Confusion, diffusion, and innovation. *The American Political Science Review*, *71*(2), 441–447. Retrieved from http://www.jstor.org/stable/1978339

Feder, G. & Feeny, D. (1991). Land Tenure and Property Rights: Theory and Implications for Development Policy. *The World Bank Economic Review*, *5*(1), pp. 135–153. Retrieved from http://www.jstor.org/stable/3989973

Feder, G. & Nishio, A. (1998). The benefits of land registration and titling: economic and social perspectives. *Land Use Policy*, *15*(1), 25–43. doi:10.1016/S0264-8377(97)00039-2

Fiflis, T. J. (1968). Security and Economy in Land Transactions: Some Suggestions from Scotland and England. *Hastings Law Journal*, *20*(1), 171–216.

FIG. (1995). FIG Statement on the Cadastre. Retrieved May 1, 2017, from http://fig.net/resources/publications/figpub/pub11/figpub11.asp

Big Impact: New Possibilities for International Development. (2013, April 7). Retrieved from http://www3.%20weforum,%20org/docs/WEF-TC-MFS-BigData%20Biglmpact%5C_Briefing.%20%5C_2012.%20pdf

Frias-Martinez, V. & Virseda, J. (2013). Cell Phone Analytics: Scaling Human Behavior Studies into the Millions. *Information Technologies & International Development*, *9*(2), 35–50. Retrieved from http://www.itidjournal.org/index.php/itid/article/view/1051

Galiani, S. & Schargrodsky, E. (2010). Property rights for the poor: effects of land titling. *Journal of Public Economics*, *94*(9–10), 700–729. doi:10.1016/j.jpubeco.2010.06.002

George, P. T. (1970). The evolution of land tenures in India. *Artha Vijnana*, *12*(1–2), 1–15. Retrieved from http://dspace.gipe.ac.in/xmlui/handle/10973/27389

Golder, S. A. & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, *40*(1).

Goldner, B. (1982). The Torrens System of Title Registration: A New Proposal for Effective Implementation. *UCLA Law Review*, *29*(3), 661–710. Retrieved from http://heinonline.org/HOL/Page?public=false%5C&handle=hein.journals/uclalr29%5C&id=675

González-Bailón, S. (2013). Social Science in the Era of Big Data. *Policy & Internet*, *5*(2), 147–160. doi:10.1002/1944-2866.POI328

Gottschalk, P. (2013). *Religion, science, and empire : classifying Hinduism and Islam in British India*. New York: Oxford University Press.

Goyal, A. (2012). The value of land administration information for financial development. Retrieved from http://www.ideasforindia.in/Article.aspx?article_id=39

Graham, E. R., Shipan, C. R., & Volden, C. (2013). The diffusion of policy diffusion research in political science. *British Journal of Political Science*, *43*, 673–701. doi:10.1017/S0007123412000415

Gray, V. (1994). Competition, emulation, and policy innovation. In L. C. Dodd & C. Jillson (Eds.), *New perspectives on american politics* (pp. 230–248). CQ Press Washington, DC.

Gray, V. (1973). Innovation in the states: a diffusion study. *The American Political Science Review*, *67*(4), 1174–1185. Retrieved from http://www.jstor.org/stable/1956539

Gupta, P. S. (2010–2011). Ending finders, keepers: the use of title insurance to alleviate uncertainty in land holdings in india. *U.C. Davis Journal of International Law & Policy*, *17*, 63. Retrieved from https://jilp.law.ucdavis.edu/issues/volume-17-1/63-109.pdf

Habibullah, W. & Ahuja, M. (2005). *Computerisation of Land Records*. In Land Reforms in India. SAGE Publications India Pvt., Ltd.

Heeks, R. (2002). Information systems and developing countries: failure, success, and local improvisations. *The Information Society*, *18*(2), 101–112. doi:10.1080/01972240290075039. eprint: http://dx.doi.org/10.1080/01972240290075039

Higgins, B. & Savoie, D. J. (1997). *Regional Development Theories & Their Application*. New Brunswick, U.S.A: Transaction Publishers.

Hilbert, M. (2016). Big data for development: a review of promises and challenges. *Development Policy Review*, *34*(1), 135–174. doi:10.1111/dpr.12142

Hilbert, M. (2013). Big Data for Development: From Information-to-Knowledge Societies. *Available at SSRN 2205145*. doi:10.2139/ssrn.2205145

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and Its Technical Challenges. *Communications of the ACM*, *57*(7), 86–94. doi:10.1145/2611567

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer.

Jensen, J. E. (1973). Computerization of Land Records by the Title Industry. *American University Law Review*, *22*(2), 393–406.

Jerven, M. (2013). *Poor numbers : how we are misled by African development statistics and what to do about it*. Ithaca: Cornell University Press.

Jerven, M. (2015). Post-2015 Consensus: Data for Development Assessment. Retrieved April 23, 2017, from http://www.copenhagenconsensus.com/publication/post-2015-consensus-data-development-assessment-jerven

Jones, C. O. (2005). *The presidency in a separated system* (2nd). Brookings Institution Press.

Jug, M. (2014). Information Revolution - From Data to Policy Action in Low-Income Countries: How Can Innovation Help? Retrieved from http://www.paris21.org/sites/default/files/PARIS21-DiscussionPaper3-Innovation.pdf

Kabacoff, R. (2015). *R in Action: Data Analysis and Graphics with R*. Data/Statistics/Programming. Manning.

Karch, A. (2007). Emerging issues and future directions in state policy diffusion research. *State Politics & Policy Quarterly*, *7*(1), 54–80. doi:10.1177/153244000700700104. eprint: http://spa.sagepub.com/content/7/1/54.full.pdf+html

Karch, A. (2006). National intervention and the diffusion of policy innovations. *American Politics Research*, *34*(4), 403–426. Journal earlier called "American Politics Quarterly". Article cited in Shipan & Volden (2008) AJPS 52(4) pp 840–857. doi:10.1177/1532673X06288202. eprint: http://apr.sagepub.com/content/34/4/403.full.pdf+html

Kent, R. B. (1988). Property tax administration in developing countries: alternatives for land registration and cadastral mapping. *Public Administration & Development*, *8*(1), 99–113.

Khatri, V. & Brown, C. V. (2010). Designing data governance. *Commun. ACM*, *53*(1), 148–152. doi:10.1145/1629175.1629210

Kim, H. J., Kim, P. S., & Moon, K. (2014). Policy diffusion and its determinants: the case of the multicultural family support ordinance in south korean local governments. *Philippine Political Science Journal*, *35*(2), 158–184. doi:10.1080/01154451.2014.964793. eprint: http://dx.doi.org/10.1080/01154451.2014.964793

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872. eprint: http://www.sciencemag.org/content/331/6018/719.full.pdf

Kingdon, J. (2011). *Agendas, Alternatives, and Public Policies*. Boston: Longman.

Kirkpatrick, R. (2013). Big Data for Development. *Big Data*, *1*(1). doi:10.1089/big.2012.1502.

Kitchin, R. & McArdle, G. (2016). What makes big data, big data? exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130. doi:10.1177/2053951716631130. eprint: http://dx.doi.org/10.1177/2053951716631130

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. doi:10.1073/pnas.1320040111. eprint: http://www.pnas.org/content/111/24/8788.full.pdf+html

Kranz, J. K. (2012). Expedition E-Recording, First Stop URPERA: How Universal E-Recording Under URPERA Could Revolutionize Real Estate Recording in the United States and Why it Should. *Minnesota Journal of Law Science & Technology*, *13*(1), 383–406.

Krishnan, V. (2010). Constructing an Area-based Socioeconomic Index: A Principal Components Analysis Approach. Retrieved July 22, 2017, from http://www.cup.ualberta.ca/wp-content/uploads/2013/04/SEICUPWebsite_10April13.pdf

Lagoze, C. (2014). Big data, data integrity, and the fracturing of the control zone. *Big Data and Society*, *1*(2). doi:10.1177/2053951714558281. eprint: http://bds.sagepub.com/content/1/2/2053951714558281.full.pdf+html

Lakshmanappa, S. T. & Singh, D. (2017, March 21). A Viable Approach to Establish Conclusive Land Title in India, March 21, 2017. The World Bank. Annual World Bank Conference on Land and Poverty. Washington, DC. Retrieved from https://www.conftool.com/landandpoverty2017/index.php/01-12-Timmiah_Lakshmanappa-715_paper.pdf?page=downloadPaper&filename=01-12-Timmiah_Lakshmanappa-715_paper.pdf&form_id=715&form_version=final

Lane, J. & Stodden, V. (2013). What? Me Worry? What to Do About Privacy, Big Data, and Statistical Research. Retrieved from http://magazine.amstat.org/blog/2013/12/01/bigdatastatresearch/

Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved from http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/

Lang, A. G. (1981). Computerised Land Title and Land Information. *Journal of Law and Information Science*, *1*(2), 230–255.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506. eprint: http://www.sciencemag.org/content/343/6176/1203.full.pdf

Lemmen, C. & van Oosterom, P. (2001). Cadastral Systems. *Computers, Environment and Urban Systems*, *25*(4–5), 319–324. doi:10.1016/S0198-9715(00)00043-0

Leszczynski, A. & Crampton, J. (2016). Introduction: spatial big data and everyday life. *Big Data & Society*, *3*(2). doi:10.1177/2053951716661366. eprint: http://bds.sagepub.com/content/3/2/2053951716661366.full.pdf

Lyson, H. C. (2016). National policy and state dynamics: a state-level analysis of the factors influencing the prevalence of farm to school programs in the united states. *Food Policy*, *63*, 23–35. doi:http://doi.org/10.1016/j.foodpol.2016.06.008

Maddison, A. (1971). *Class Structure and Economic Growth: India and Pakistan since the Moghuls*. W. W. Norton & Company, Inc. New York.

Maggs, P. B. (1973). Automation of the Land Title System. *American University Law Review*, *22*(2), 275–331.

Makse, T. & Volden, C. (2011). The role of policy attributes in the diffusion of innovations. *Journal of Politics*, *73*(1), 108–124.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011a). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011b). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Margulis, S. T. (2003). Privacy as a Social Issue and Behavioral Concept. *Journal of Social Issues*, *59*(2), 243–261. doi:10.1111/1540-4560.00063

McCormack, J. L. (1992). Torrens and Recording: Land Title Assurance in the Computer Age. *William Mitchell Law Review*, *18*(1), 61–129. Retrieved from http://heinonline.org/HOL/Page?handle=hein.journals/wmitch18%5C&id=75

McNeal, R. S., Tolbert, C. J., Mossberger, K., & Dotterweich, L. J. (2003). Innovating in Digital Government in the American States. *Social Science Quarterly*, *84*(1), 52–70. doi:10.1111/1540-6237.00140

Mehndiratta, S. & Alvim, B. (2014, December 30). Big Data comes to transport planning: how your mobile phone helps plan that rail line. Retrieved December 30, 2014, from http://blogs.worldbank.org/transport/big-data-comes-transport-planning-how-your-mobile-phone-helps-plan-rail-line

Mergel, I., Rethemeyer, R. K., & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, *76*(6), 928–937. doi:10.1111/puar.12625

MGI. (2001). *India: The growth imperative*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/india/growth_imperative_for_india

Ministry of Finance, Government of India. (2013). *Report of the Committee for Evolving a Composite Development Index of States*. Retrieved from http://finmin.nic.in/reports/Report_CompDevState.pdf

Mishra, S. (2016). *Identifying Existing Capacities to Execute the National Land Records Modernization Programme in Madhya Pradesh: An Appraisal*. Centre for Rural Studies, Lal Bahadur Shastri National Academy of Ad-

ministration. Retrieved from http://centre.lbsnaa.gov.in/crs/admin/upload/Publication_Int/Madhya_Pradesh_Study2016.pdf

Mitchell, T. M. (1999). Machine Learning and Data Mining. *Communications of the ACM*, *42*(11), 30–36. doi:10.1145/319382.319388

Mohr, L. B. (1969). Determinants of innovation in organizations. *The American Political Science Review*, *63*(1), 111–126. doi:10.2307/1954288

Mookerjee, R. (1919). *Occupancy Right: Its History and Incidents*. Calcutta: University of Calcutta. Retrieved from http://hdl.handle.net/2027/coo1.ark:/13960/t4jm2tv7b

Moyer, D. D. & Fisher, K. P. (1973). *Land Parcel Identifiers for Information Systems*. Conatins papers from the CLIPPP (Compatible Land Identifiers—the Problems, Prospects, and Payoffs) conference (Atlanta, GA Jan 20–22, 1972). American Bar Foundation.

Mundie, C. (2014). Privacy pragmatism; focus on data use, not data collection. *Foreign Aff. 93*, 28.

Narasappa, H. & Vidyasagar, S. (2016). *State of the Indian Judiciary*. Daksh. Retrieved March 1, 2016, from http://dakshindia.org/state-of-the-indian-judiciary/00_cover.html

Narayanan, A. & Felten, E. W. (2014). No silver bullet: De-identification still doesn't work. Retrieved from http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005, August 9). Handbook on Constructing Composite Indicators:

Methodology and User Guide. *OECD Statistics Working Papers,* (2005/3). doi:10.1787/533411815016

National e-Governance Division. (2011, February 22). *Saaransh: A compendium of Mission Mode Projects under NeGP*. The Department of Information Technology, Ministry of Communications and Information Technology, Government of India. Retrieved from http://deity.gov.in/sites/upload_files/dit/files/Compendium_FINAL_Version_220211.pdf

National Research Council. (2007). *National Land Parcel Data: A Vision for the Future* (Committee on Land Parcel Databases, Ed.). ISBN: 0-309-11031-9, 172 pages, 6 x 9, (2007). Washington, DC: The National Academies Press. doi:10.17226/11978

National Research Council. (1980). *Need for a Multipurpose Cadastre*. Washington, DC: The National Academies Press. doi:10.17226/10989

National Research Council. (1995). *Procedures and Standards for a Multipurpose Cadastre*. Washington, DC: The National Academies Press. doi:10.17226/9083

Navratil, G. & Frank, A. U. (2004). Processes in a cadastre. *Computers, Environment and Urban Systems*, *28*(5), 471–486. Cadastral Systems {III}. doi:10.1016/j.compenvurbsys.2003.11.003

Nayak, P. (2013). Policy Shifts in Land Records Management. *Economic and Political Weekly*, *48*(24), 71–75.

Neale, W. C. (1962). *Economic Change in Rural India: Land Tenure and Reform in Uttar Pradesh, 1800–1955*. Yale University Press.

Neyman, Y., Linkow, B., & Kijazi, M. (2016, March 17). USAID's Mobile Application to Secure Tenure (MAST): Preliminary Results of Crowd-Sourced Community Mapping, March 17, 2016. The World Bank. Annual World Bank Conference on Land and Poverty. Washington, DC. Retrieved from https://www.conftool.com/landandpoverty2016/index.php/Neyman-673-673_paper.docx?page=downloadPaper&filename=Neyman-673-673_paper.docx&form_id=673&form_version=final

Nicholson-Crotty, S. (2009). The politics of diffusion: public policy in the american states. *Journal of Politics*, *71*(1), 192–205.

Nicholson-Crotty, S. & Carley, S. (2016). Effectiveness, implementation, and policy diffusion: or "can we make that work for us?" *State Politics & Policy Quarterly*, *16*(1), 78–97. doi:10.1177/1532440015588764. eprint: http://spa.sagepub.com/content/16/1/78.full.pdf+html

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Payne, G. (2004). Land tenure and property rights: an introduction. *Habitat International*, *28*(2), 167–179. Land Tenure and Property Rights. doi:10.1016/S0197-3975(03)00066-3

Payne, G. (2001). Urban land tenure policy options: titles or rights? *Habitat International*, *25*(3), 415–429. doi:10.1016/S0197-3975(01)00014-5

PCAST. (2014). *Big Data and Privacy: A Technological Perspective*. Executive Office of the President. The President's Council of Advisors on Science and Technology. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Pirog, M. A. (2014). Data will drive innovation in public policy and management research in the next decade. *Journal of Policy Analysis and Management*, *33*(2), 537–543. doi:10.1002/pam.21752

Podesta, J., Pritzker, P., Moniz, E., Holdren, J., & Zients, J. (2014). *Big Data: Seizing Opportunities, Preserving Values*. Executive Office of the President. Retrieved from http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Revelle, W. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.6.12. Northwestern University. Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych

Risk, R. C. B. (1971). The Records of Title to Land: A Plea for Reform. *The University of Toronto Law Journal*, *21*(4), 465–497. doi:10.2307/825245

Robinson, N. (2013). A Quantitative Analysis of the Indian Supreme Court's Workload. *Journal of Empirical Legal Studies*, *10*(3), 570–601. doi:10.1111/jels.12020

Rogers, E. M. (2003). *Diffusion of Innovations*. Free Press.

Rothermund, D. (1969). Government, Landlord and Tenant in India, 1875–1900. *Indian Economic & Social History Review*, *6*(4), 351–367. doi:10.1177/001946466900600402. eprint: http://ier.sagepub.com/content/6/4/351.full.pdf+html

Rothermund, D. (1971). The Record of Rights in British India. *Indian Economic & Social History Review*, *8*(4), 443–461. doi:10.1177/001946467100800405. eprint: http://ier.sagepub.com/content/8/4/443.full.pdf+html

Round, J. I. (2014). Assessing the demand and supply of statistics in the developing world: some critical factors. Retrieved from http://www.paris21.org/sites/default/files/PARIS21-DiscussionPaper4-Demand.pdf

Sabatier, P. & Mazmanian, D. (1979). The conditions of effective implementation: a guide to accomplishing policy objectives. *Policy Analysis*, *5*(4), 481–504. Retrieved from http://www.jstor.org/stable/42783358

Sapat, A. (2004). Devolution and innovation: the adoption of state environmental policy innovations by administrative agencies. *Public Administration Review*, *64*(2), 141–151. doi:10.1111/j.1540-6210.2004.00356.x

Savage, R. L. (1985). Diffusion research traditions and the spread of policy innovations in a federal system. *Publius*, *15*(4), 1–27. Retrieved from http://www.jstor.org/stable/3330039

Schintler, L. A. & Kulkarni, R. (2014). Big Data for Policy Analysis: The Good, The Bad, and The Ugly. *Review of Policy Research*, *31*(4), 343–348. doi:10.1111/ropr.12079

Secretary-General's Independent Expert Advisory Group, T. (2014). A World That Counts: Mobilising The Data Revolution for Sustainable Development. Retrieved from http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf

Shipan, C. R. & Volden, C. (2006). Bottom-up federalism: the diffusion of antismoking policies from u.s. cities to states. *American Journal of Political Science*, *50*(4), 825–843. doi:10.1111/j.1540-5907.2006.00218.x

Shipan, C. R. & Volden, C. (2012). Policy diffusion: seven lessons for scholars and practitioners. *Public Administration Review*, *72*(6), 788–796.

Shipan, C. R. & Volden, C. (2008). The mechanisms of policy diffusion. *American Journal of Political Science*, *52*(4), 840–857.

Singh, A. K. (2014, December 1). Evolving a Composite Development Index of States: A Critique. *Journal of Regional Development and Planning*, *3*.

Skopek, J. M. (2014). Anonymity, the Production of Goods, and Institutional Design. *Fordham L. Rev. 82*(4), 1751–1809. Retrieved from http://ir.lawnet.fordham.edu/flr/vol82/iss4/4

Solove, D. J. (2004). *the digital person: Technology and Privacy in the Information Age*. New York: New York University Press.

Stodden, V. (2014). Enabling Reproducibility in Big Data Research: Balancing Confidentiality and Scientific Transparency. In *Privacy, Big Data, and the Public Good* (pp. 112–132). Cambridge University.

Sud, N. (2014). Governing India's Land. *World Development*, *60*, 43–56. doi:http://dx.doi.org/10.1016/j.worlddev.2014.03.015

Sugiyama, N. B. (2012). Bottom-up policy diffusion: national emulation of a conditional cash transfer program in brazil. *Publius: The Journal of Federalism*, *42*(1), 25–51. Retrieved from http://mutex.gmu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=poh&AN=67382497&site=ehost-live

Sugiyama, N. B. (2008). Theories of Policy Diffusion: Social Sector Reform in Brazil. *Comparative Political Studies*, *41*(2), 193–216. doi:10.1177/

0010414007300916. eprint: http://cps.sagepub.com/content/41/2/193.full.pdf+html

Szypszak, C. (2003). Public Registries and Private Solutions: An Evolving American Real Estate Conveyance Regime. *Whittier Law Review*, *24*(3), 663–706. Retrieved from http://heinonline.org/HOL/Page?handle=hein.journals/whitlr24%5C&id=689

Taylor, L. (2014). A god's eye view? Big data, ground truth and the D4D challenge. Retrieved from http://www.academia.edu/8079377/A_god_s_eye_view_Big_data_ground_truth_and_the_D4D_challenge

Taylor, L. (2016a). Data subjects or data citizens? addressing the global regulatory challenge of bigdata: The philosophy of law meets the philosophy of technology. In M. Hildebrandt & B. van den Berg (Eds.), *Information, Freedom and Property* (1st ed., Chap. 4, pp. 81–105). Routledge.

Taylor, L. (2016b). No place to hide? the ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, *34*(2), 319–336. doi:10.1177/0263775815608851

Taylor, L. & Broeders, D. (2015). In the name of Development: power, profit and the datafication of the global South. *Geoforum*, *64*, 229–237. doi:http://dx.doi.org/10.1016/j.geoforum.2015.07.002

Taylor, L., Cowls, J., Schroeder, R., & Meyer, E. T. (2014). Big Data and Positive Change in the Developing World. *Policy & Internet*, *6*(4), 418–444. doi:10.1002/1944-2866.POI378

Taylor, L. & Schroeder, R. (2014). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal*, 1–16. doi:10.1007/s10708-014-9603-5

Tolbert, C. J., Mossberger, K., & McNeal, R. (2008). Institutions, policy innovation, and e-government in the american states. *Public Administration Review*, *68*(3), 549–563. doi:10.1111/j.1540-6210.2008.00890.x

Törhönen, M.-P. (2004). Sustainable land tenure and land registration in developing countries, including a historical comparison with an industrialised country. *Computers, Environment and Urban Systems*, *28*(5), 545–586. Cadastral Systems {III}. doi:10.1016/j.compenvurbsys.2003.11.007

UN Global Pulse. (2012). Big Data for Development: Opportunities & Challenges. Retrieved from http://www.unglobalpulse.org/projects/BigDataforDevelopment

UN Stats. (2013, February 22). Big Data for Policy, Development and Official Statistics - Concept Note. Retrieved from http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/concept_note.pdf

Vaidya, J., Clifton, C. W., & Michael, Z. Y. (2006). *Privacy Preserving Data Mining*. New York: Springer.

van der Molen, P. (2002). The dynamic aspect of land administration: an often-forgotten component in system design. *Computers, Environment and Urban Systems*, *26*(5), 361–381. doi:10.1016/S0198-9715(02)00009-1

Venkataraman, M. (2014). What is Title Guarantee Worth in Land Markets? Evidence from Bengaluru, India. *SSRN (December 2, 2014)*. IIM Bangalore Research Paper No. 473. doi:10.2139/ssrn.2532874

Wadhwa, D. C. (2002). Guaranteeing Title to Land. *Economic and Political Weekly*, *37*(47), 4699–4722. Retrieved from http://www.jstor.org/stable/4412872

Walker, J. L. (1969). The Diffusion of Innovations among the American States. *The American Political Science Review*, *63*(3), 880–899. Retrieved from http://www.jstor.org/stable/1954434

Walker, R. M. (2014). Internal and external antecedents of process innovation: a review and extension. *Public Management Review*, *16*(1), 21–44. doi:10.1080/14719037.2013.771698. eprint: http://dx.doi.org/10.1080/14719037.2013.771698

Wallace, J. & Williamson, I. P. (2006). Developing cadastres to service complex property markets. *Computers, Environment and Urban Systems*, *30*(5), 614–626. Cadastral Systems {IV}. doi:10.1016/j.compenvurbsys.2005.08.007

Ward, J. S. & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions. Retrieved from http://arxiv.org/pdf/1309.5821.pdf

Washington, A. L. (2014). Government Information Policy in the Era of Big Data. *Review of Policy Research*, *31*(4), 319–325. doi:10.1111/ropr.12081

Welch, S. & Thompson, K. (1980). The impact of federal incentives on state policy innovation. *American Journal of Political Science*, *24*(4), 715.

Welles, B. F. (2014). On minorities and outliers: the case for making big data small. *Big Data & Society*, *1*(1). doi:10.1177/2053951714540613. eprint: http://bds.sagepub.com/content/1/1/2053951714540613.full.pdf+html

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. doi:10.1038/sdata.2016.18

Williamson, I. P. (2001). Land administration ”best practice” providing the infrastructure for land policy implementation. *Land Use Policy*, *18*(4), 297–307. doi:10.1016/S0264-8377(01)00021-7

Williamson, I. P., Enemark, S., Wallace, J., & Rajabifard, A. (2010). *Land Administration for Sustainable Development*. Redlands, California: ESRI Press Academic. Retrieved from http://www.esri.com/landing-pages/industries/land-administration/e-book

Williamson, I. P. & Ting, L. (2001). Land administration and cadastral trends—a framework for re-engineering. *Computers, Environment and Urban Systems*, *25*(4–5), 339–366. doi:10.1016/S0198-9715(00)00053-3

Wingfield, C. (1869). *Observations on land tenures and tenant right in india*. London: W.H. Allen. Retrieved from http://hdl.handle.net/2027/hvd.hl43s0

Why Secure Land Rights Matter. (2017, March 24). Retrieved April 1, 2017, from http://www.worldbank.org/en/news/feature/2017/03/24/why-secure-land-rights-matter

Wouters, R., Meijerink, G., Vaandrager, R., & Zavrel, J. (2010). Extracting Information from Deeds by OCR and Text Interpretation. In *XXIV FIG International Congress*. Sydney, Australia: 11– 16 April, 2010. Retrieved from http://www.fig.net/resources/proceedings/fig_proceedings/fig2010/papers/ts02h/ts02h_wouters_3841.pdf

Zasloff, J. (2011). India’s Land Title Crisis: The Unanswered Questions. *Jindal Global Law Review*, *3*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1923903

BIOGRAPHY

Sachin Garg holds a Master of Engineering in Computer Science from the Motilal Nehru National Institute of Technology, Allahabad (Uttar Pradesh, India) (degree granted by University of Allahabad) and a Bachelor of Engineering in Applied Electronics and Instrumentation Engineering from the National Institute of Technology, Rourkela (Odisha, India) (degree granted by Sambalpur University).

His research interests lie at the juncture of technology and public policy. He is especially interested in understanding the role that public policy plays, or can play in the use of technologies for human development.

A hands-on technologist and opensource evangelist, he started his fifteen year long stint in the Information Technology sector with India's Center for Development of Advanced Computing, where he developed systems software for supercomputers. During his industry tenure, he held various leadership responsibilities where he architected, designed, developed and help maintain systems ranging from large (web scale) to small (embedded devices). His last position prior to joining the Schar School of Policy and Government was as a DevOps architect at Yahoo! Inc.