SOCIAL MONITORING AND CORRUPTION IN DEVELOPING COUNTRIES

Robert M. Gonzalez

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics.

Chapel Hill
2016

Approved by:

Klara Peter

Patrick Conway

Erica Field

David Guilkey

Tiago Pires

ABSTRACT

ROBERT M. GONZALEZ: Social Monitoring and Corruption in Developing Countries.
(Under the direction of Klara Peter)

This dissertation studies the impact of *social monitoring*–monitoring of government officials by ordinary citizens–on institutional corruption. Specifically, I study a monitoring initiative in which citizens used cell phones to report instances of fraud during the 2009 Afghan presidential election. Since implementation of the program required cell phone coverage, I combine coverage maps with unique data on the geographic location and fraud levels of polling centers across Afghanistan to determine: (i) the effect of coverage on fraud, and (ii) whether social monitoring is the main corruption-deterring mechanism among several competing channels. Using a spatial regression discontinuity (RD) design along the cell phone coverage boundary, I find considerable evidence that cell phone-based participation deters corrupt behavior. Polling centers inside cell phone coverage areas report up to a 26 percent drop in the share of fraudulent votes relative to centers outside. Analyses of the effect of coverage on election-related violence and the tribal composition of villages suggest that the observed declines in fraud cannot be attributed to these alternative channels. From a policy perspective, these results illustrate how a widespread technology, namely cell phones, can exert a positive externality on institutional development via corruption deterrence.

ACKNOWLEDGMENTS

This dissertation was made possible by an incredible group of mentors: Patrick Conway, Erica Field, David Guilkey, Klara Peter, and Tiago Pires. I owe my deepest gratitude to my advisor Klara Peter for her unconditional support and guidance. Emulating her dedication and work ethic has made me a better researcher. I thank Patrick Conway for always providing a thoughtful critique. I am deeply indebted to Erica Field for believing in me and for instilling in me, through her teachings and collaboration, a profound passion for the field of Development Economics. David Guilkey is one of the best teachers I have had in my life. I thank him for laying the foundation of my knowledge of Econometrics. Lastly, I thank Tiago Pires for dedicating countless times to read and review my work, but above all, for being a friend and always being there to listen.

I am also grateful to have found great support from individuals outside of my committee. I specially thank Donna Gilleskie for her always thoughtful advice and for making me a better writer and economist. I thank Helen Tauchen for both, insightful comments on my work and for helping me navigate the stressful graduate school process. I am indebted to faculty and fellow students in the UNC Applied Micro group for listening and commenting on my work from its very early stages.

Most importantly, this dissertation would not have been possible without the love and patience of my wife Yailin. I dedicate this dissertation to her and to my parents for being a constant source of love and inspiration. Finally, as an immigrant, I thank my parents for leaving everything behind (an established professional career, family, etc.) in order to provide me a better life in this great country. This dissertation is also the product of that sacrifice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2G	Second Generation
ACTE	Averaged Conditional Treatment Effects
AIMS	Afghanistan Information Management Services
CTE	Conditional Treatment Effects
ECC	Electoral Complaints Commission
GIS	Geographic Information System
GPS	Global Positioning System
GSM	Global System for Mobile Communications
IEC	Independent Election Commission
MISTI	Measuring Impacts of Stabilization Initiatives
NASA	National Aeronautics and Space Administration
RD	Regression Discontinuity
SRTM	Shuttle Radar Topography Mission

CHAPTER 1

INTRODUCTION

Corruption is widely presumed to have a large detrimental effect on economic growth. There is evidence that corruption decreases competition and investment, dampens government revenue, and obstructs the delivery of government services (Mauro (1995), Svensson (2005), Olken and Pande (2012)).[1] One theory is that corruption flourishes in settings in which citizens are particularly disenfranchised. This has motivated policy interventions to strengthen community efforts to engage in monitoring of public officials as a means of combating corruption.

In theory, bolstering grassroots monitoring may be more effective in reducing corruption than increasing external monitoring efforts, particularly at very high levels of corruption where auditors can readily be bought out.[2] Meanwhile, social monitoring may have little potential when community members lack a means of directly punishing corrupt officials or face immediate threats of retribution. Thus far, the empirical evidence has been mixed. Olken (2007) and Banerjee, Banerji, Duflo, Glennerster, and Khemani (2010) find little support for the idea that increasing community monitoring results in better behavior of public officials, while Bjorkman and Svensson (2009) find that government performance improves significantly when community members engage in active surveillance. One key ambiguity is that the intervention found to be effective took a very intensive (and expensive) approach to fostering community participation, leaving open the question of whether scalable community monitoring efforts have the potential to make a difference.

This paper provides evidence that they do. I show that a simple approach to strengthening

---

[1]Corruption in Afghanistan, the country studied in this paper, is so pervasive that in 2013, Transparency International ranked it as the world's most corrupt country (Transparency International 2013), while American authorities argue that corruption, and not the Taliban, is the main existential threat to this country (Riechmann 2014).

[2]Throughout the paper I conceptualize social monitoring as a mechanism by which ordinary citizens can better observe or report instances of corruption. In contrast, external monitoring relies on actors hired by the government or part of the government itself to conduct investigations and report corrupt activities.

social monitoring capacity via cell phone hotlines can be highly effective in reducing corruption. Hotlines create a direct means of reporting fraud through a widely available medium – mobile phones.[3] The advent of cell phones in the developing world make this approach particularly germane. Mobile connectivity rates have increased exponentially in the developing world over the past decade, giving rise to a host of potential interventions that rely on cell phones as the primary medium for citizens to monitor and report corrupt behavior.[4]

I investigate the impact of social monitoring on corrupt behavior in the context of a United Nations (U.N.)-led monitoring initiative that created election fraud hotlines to facilitate fraud reporting by ordinary citizens during the 2009 Afghan presidential election. Since implementation required cell phone coverage, I identify the causal effect of social monitoring by exploiting geographic variation in the areas where social monitoring was feasible based on cell phone coverage availability. Using a spatial regression discontinuity (RD) design that compares polling centers within a close distance of the cell phone coverage boundary, I examine the effect of mobile phone coverage – and hence the potential threat to a misbehaving official induced by greater social monitoring – on election fraud.

The empirical analysis employs several novel data sources, including (1) detailed coverage maps based on the location of cell phone towers of the two largest mobile service providers in Afghanistan;[5] (2) data on the precise location of polling centers collected by International Security Assistant Force (ISAF) inspection teams shortly after the election;[6] and (3) polling center level data on various measures of election fraud collected by a U.N.-sponsored audit shortly after the election.

My results indicate that cell phone coverage reduces corruption, and social monitoring is the

[3]According to the International Telecommunication Union (ITU), the cell phone penetration rate in Afghanistan was close to 40 subscribers per 100 people in 2009. By 2013, this number had risen to 71 (ITU 2015).

[4]For a detailed description of the history and growth of ICT-based electoral monitoring refer to Schuler (2008).

[5]I obtain these data from *Collins Bartholomew*, which represents the GSMA, an association of major GSM mobile service providers around the world. Cell phone service providers supply coverage data directly to the GSMA.

[6]Geolocation data were provided by the Afghan Independent Election Commission (http://www.iec.org.af)

primary mechanism through which this occurs. For polling centers within a 5 to 6 kilometer bandwidth around the coverage boundary, the share of fraudulent votes in centers inside coverage areas is about 26 percent lower relative to centers outside. The results are robust to several choices of bandwidth and polynomial order. To assess the validity of the RD design, I investigate whether other polling center characteristics change discontinuously at the boundary. In particular, I compare 28 electoral, geographic, socioeconomic, and demographic indicators for villages and settlements where polling centers are located. The results indicate a smooth transition across the coverage boundary and thus little evidence that changes in fraud at the boundary are explained by changes in these indicators. To the best of my knowledge, this is the first study to provide evidence on the impact of cell phone coverage on institutional corruption.

To explore the spatial heterogeneity of the results, I modify and implement a recently-developed boundary RD design that estimates treatment effects at various points along the two-dimensional coverage boundary.[7] Since the method recovers a distribution of treatment effects along the boundary, one can determine the areas or sections of the boundary where fraud is more responsive to social monitoring. The results suggest significant spatial heterogeneity in the effect both across and within regions of Afghanistan. Economically and statistically significant drops in fraud at the coverage boundary are present in the eastern and southern regions of the country whereas average impacts in the northern and western regions are close to zero in magnitude and statistically insignificant.

Fraud may respond to cell phone coverage for reasons other than social monitoring. I use an illustrative theoretical model to motivate which channels to explore. In particular, I specify a classical supply and demand model where votes can be bought legally (via advertising, campaign promises, etc.) or illegally. In the case of illegal/fraudulent votes, corruption takes the form of collusion between polling center managers (suppliers of votes) and the corrupt candidates (demanders of votes). The price of fraudulent votes is a function of social monitoring.[8] Given the

[7]I employ a modification of the boundary RD method proposed by Imbens and Zajonc (2011).

[8]Social monitoring is modeled as the probability that the center is audited as a result of complaints.

high incidence of election-related violence, the price of legal votes is a function of both voters' affinity towards the candidates and the likelihood that the polling center is attacked by insurgents. Higher levels of expected violence require a higher price to guarantee that constituents vote. Coverage enters the model as a shifter in the probability of social monitoring and, hence, a shifter in the equilibrium quantity of fraudulent votes. However, the price of legal votes, and hence the equilibrium level of fraud, may also shift with coverage if the likelihood of insurgent attacks or voters' affinity depends on coverage.[9] Therefore, these two channels may lead to changes in the equilibrium level of fraud that mimic the social monitoring effect.

First, I examine whether political violence by insurgent groups, which is strongly related to both cell phone coverage (Shapiro and Weidmann 2013) and electoral fraud (Callen and Weidmann 2013), confounds the social monitoring effect. I replicate the spatial RD analysis using recently declassified data on daily insurgent and IED attacks, along with data on civilian and military casualties around election day.[10] Using the boundary RD design, I test whether the violence outcomes change discretely at points in the coverage boundary where I observe significant changes in fraud as well. Results suggest that the treatment effects of coverage on violence are generally small in magnitude and statistically insignificant for most points in the boundary. This has key implications for the identification of the social monitoring effect since drops in fraud at the boundary cannot be explained by significant drops in violence outcomes. With this in mind, a secondary, yet important contribution of this paper is to advance our understanding of the relationship between cell phone access and insurgent violence.

Second, given the importance of tribal loyalty in Afghan society, I also test directly whether the boundary effects are confounded by discrete changes in the tribal composition of villages, a strong predictor of party affiliation.[11] To explore this possibility, I georeference detailed tribal

[9]For instance, if the likelihood of violence or voters' affinity drops with coverage, legal votes become less costly. Thus the candidate substitutes fraudulent with legal votes.

[10]Data on IED attacks are obtained from Shaver and Wright (2015) and refer to SIGACTs or Significant Actions. These data are collected directly by the military and constitute the official database of insurgent attacks. Data on civilian and military casualties are provided by the Worldwide Incident Tracking System (WITS 2009).

[11]This might be the result, for instance, of cell phone providers giving preference to certain ethnic groups by

maps collected by the Culture and Conflicts Studies program containing information on the geographic distribution of more than 50 tribes and ethnic groups across Southeastern Afghanistan.[12] I then combine the georeferenced maps with village location data from the Measuring Impacts of Stabilization Initiatives project (MISTI 2013) to construct village-level indicators of primary tribe and tribal confederation for almost 18,000 villages. I replicate the boundary RD analysis for portions of the boundary where there seem to be changes in the number of villages belonging to the same tribal confederation as the main candidates. I show that, while there is some evidence of changes in the tribal structure of villages at certain points along the boundary, these changes cannot explain the observed drops in fraud as there is no substantial overlap in the boundary points where both tribal affiliation and fraud change sharply.

This paper contributes to a growing effort to understand the effectiveness of grassroots monitoring on illegal behavior within the realm of election fraud. Callen and Long (2015) implement a field experiment where individuals record photographs of the total vote tally at randomly selected polling centers during the 2010 Afghan parliamentary election. This monitoring technology, however, is conceptually different from the one I study in this paper as it does not necessarily rely on cell phone coverage. Further, the monitoring is performed by a select group of individuals rather than all voters. Aker, Collier, and Vicente (2014) explore the impact of an SMS hotline during the 2009 Mozambican election, a monitoring technology that is very similar to the Afghan setting. However, while they show convincing evidence that the hotline lowers fraud levels, it is not clear whether this result would hold in a fragile security environment like Afghanistan. Prior to the election, the Taliban issued several warnings targeting polling centers and voters while on election day the number of attacks exceeded the 2009 daily average by a factor of seven.[13] In such cases, election-related violence hampers monitoring incentives as individuals fear retaliation or are simply unable to witness fraud if not present at the polling centers. The fact that I find significant drops

expanding coverage into their locations.

[12]The Culture and Conflicts Studies program is part of the Naval Postgraduate School.

[13]Figures calculated by author using SIGACTs data obtained from Shaver and Wright (2015). The average number of daily attacks in 2009, excluding election day, was 55.18 attacks. The number of attacks on election day was 422.

in fraud at the coverage boundary suggests that monitoring technologies that offer some degree of plausible deniability to potential whistleblowers can be effective even in settings characterized by extreme political violence.[14]

This study is also novel in providing rigorous evidence on the role of information and communications technologies (ICT) on improving information transfer and social monitoring capacity. This is a particularly important contribution given the rapid expansion in mobile services experienced in the developing world throughout the last decade.[15] More generally, the results in this paper show how commonly available communication devices, such as cell phones, can exert a positive externality on institutional development. In that sense, the results add to a rapidly advancing literature on the effectiveness of ICT-based policies on improving transparency and economic development outcomes in general (e.g., Jensen (2007), Aker (2010), and Aker et al. (2014)).

Lastly, from an empirical standpoint, this paper contributes to the literature on spatial and, more specifically, geographic RD (e.g., Imbens and Zajonc (2011), Keele and Titiunik (2013)). A literature that is rapidly growing as advances in GPS technology are increasing the availability of micro-level geospatial data. More importantly, however, is that it illustrates an empirical framework that can be used by other studies trying to uncover heterogeneous effects of mobile phone coverage on any outcome variable using a spatial framework. Further, the possibility of obtaining spatially heterogeneous effects along a geographic boundary has key policy implications as it can guide agencies in the design of localized anti-corruption policies.

The paper is structured as follows: Chapter 2 provides a background of the Afghan 2009 presidential election as well as the nationwide audit that followed shortly afterwards. Chapter 3 presents the theoretical framework. Chapter 4 describes the empirical method and reports results of the effect of coverage on fraud. Chapter 5 explores alternative channels of fraud. Lastly, Chapter 6 concludes.

[14]See Chassang and Padro-i-Miquel (2014) for a detailed treatment on the importance of plausible deniability to incentivize monitoring and to avoid side-contracting between monitors and misbehaving agents.

[15]In the case of Afghanistan, for instance, the number of mobile subscribers rose from 1.7 million in 2006 to around 17.1 million in 2012. This translates into a mobile penetration (number of subscribers per 100 inhabitants) of 6.27 in 2006 and 63.3 in 2012

CHAPTER 2

BACKGROUND

This chapter provides a detailed description of the 2009 Afghan presidential election and of the audit and recount process that took place shortly after the election.

2.1 The 2009 Afghan Election

The 2009 Afghan presidential election marked the second election after the toppling of the Taliban regime in 2001. Fraud allegations during the 2004 presidential election led to the creation of the Electoral Complaints Commission (ECC) precisely to investigate and adjudicate fraud related complaints for the upcoming 2009 presidential election. This constituted the first time in Afghan history that a formal channel for individuals to report electoral complaints was created. In addition to adjudication of complaints, the ECC was given the power to issue audits, recounts, and runoff elections if necessary (Electoral Complaints Commission 2010). To improve transparency and guarantee independence from the executive power, three of the five appointed ECC commissioners (including the chairman) were international experts directly appointed by the United Nations Representative of the Secretary General. The two Afghan commissioners were selected from the Afghanistan Independent Human Rights Commission and the Supreme Court(National Democratic Institute 2010).

According to ECC guidelines, individuals and organizations could file election-related complaints within 72 hours of the incident. Typical types of complaints included claims of bribery, intimidation, counting errors, and theft and manipulation of electoral documents. The ECC provided two hotlines for people to report and find the information needed to file a claim (where to file it, how to file it, deadlines, etc.). In addition, the ECC created local offices at each of the 34 provinces for individuals desiring to report fraud in person via a form. To guarantee some degree of accountability, individuals filing complaints were required to provide their names and addresses to the ECC in case a follow-up investigation would take place. The ECC, however, guaranteed this information was not to be disclosed. These hotlines were widely publicized through a public outreach program that included television and several radio advertisements in both Pashto and Dari.[1] Similarly, instances of violence, intimidation at the polling center, and corruption in general could be reported to the 119 Afghan corruption hotline led by the European Union Police Mission in Afghanistan (EUPOL) and relatively known by the Afghan population.[2] Private organizations also encouraged the use of cell phones to report instances of fraud. For example, in the weeks prior to the election, Pajhwok News, a major independent news agency in the country, along with other international NGOs, enabled several hotlines. In addition the agency deployed around 80 reporters around the country who were instructed to use their mobile phones to text and call in incidents of violence and fraud (Himelfarb 2010).

Allegations of fraud during the 2009 Afghan election were widespread.[3] According to ECC's chairman Grant Kippen, the agency received more than 3,300 complaints with close to 80 percent of these complaints received during the polling and counting period (Electoral Complaints Commission 2010). In terms of the types of complaints, most complaints–about 47 percent–dealt with polling and counting irregularities, followed by complaints on intimidation, and violence at the center (about 26 percent). The remaining types of complaints were distributed between: access to stations (11 percent), missing election materials at the center (4 percent), and other types (12 percent). This degree of citizen participation represented a major improvement from the 2004 presidential election when no formal channel to file claims existed. A direct implication of this was the implementation of a nationwide audit that is discussed in detail in the following section.

---

[1]More information on the hotlines as well as the public outreach program can be found at the ECC's official website www.ecc.org.af

[2]According to a 2012 UNDP survey cited by the Ministry of Interior Affairs (MOIA), about 80-90 percent of the Afghan population has some familiarity with the police hotline. This information was obtained by the author through an interview with MOIA representatives.

[3]Refer to Panels (b), (c), and (d) in Appendix Figure C.8 for examples of typical fraudulent activities.

2.2 The Audit and Recount

Election day took place on August 20, 2009. Eighteen days after the tallying of votes began; the ECC ordered a nationwide audit of polling stations after initial investigations of the received complaints revealed clear evidence of widespread fraud.[4] The audit called for the investigation of polling stations reporting unusually high turnout and an unusually high majority of votes for a single candidate. Specifically, the audit-triggering criteria were: (1) stations in which 600 or more votes were cast, (2) stations in which one candidate received 95 percent or more of the total votes cast, and (3) stations satisfying both (1) and (2). The ECC referred to each of these categories as Category A, B, and C, respectively.[5]

The motivation for these criteria lied primarily in the particular design of the election and the unusual pattern of reported total votes per station. In particular, polling station managers were provided with a ballot book containing exactly 600 empty ballots (Electoral Complaints Commission 2010).[6] However, a significant number of stations reported totals of exactly 600 or more votes cast. This was particularly unusual given the overall low turnout across the country resulting from the fragile security environment (Khadhouri 2010). Such discrepancies in reported turnout can be clearly seen in Figure 2.1a which shows a histogram of total votes cast per station for the top two candidates. Notice the pronounced jump in the frequency of total votes cast at exactly 600 for candidate Hamid Karzai in particular.[7] The incidence of stations where a candidate obtained more than 95 percent of the total vote share was equally unusual. Note in Figure 2.1b that a substantially high number of stations (with more than 100 total votes cast) had exactly 100 percent vote share

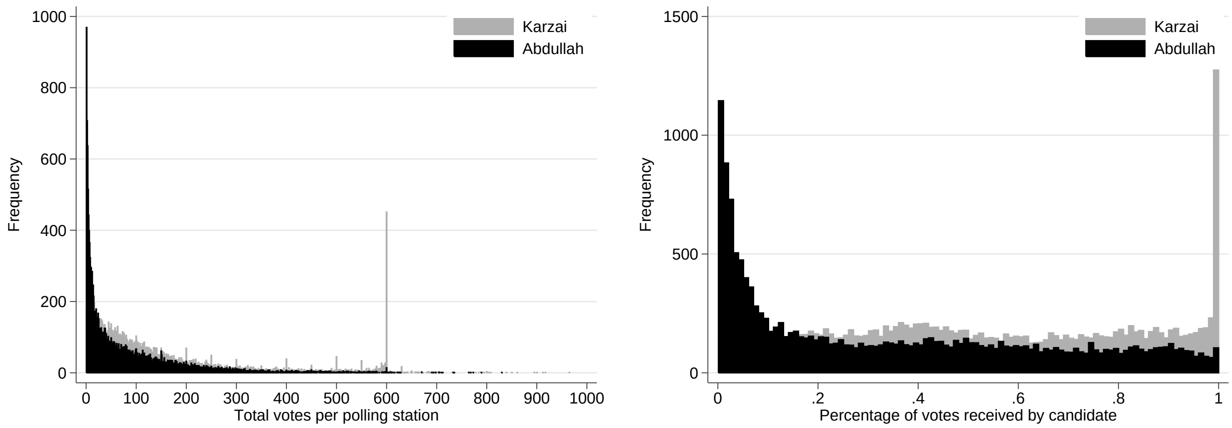
[4]For reference, a polling station is a physical location within a polling center. In the sample studied, the average number of polling stations per center is about 4 with some centers having up to 20 stations.

[5]To be more specific the ECC defined a total of six categories, however, given the similarity between some of the categories I reduce them to three aggregate categories. Refer to Appendix A for a more detailed explanation of the audit categories.

[6]Refer to Panel (a) of Appendix Figure C.8 for a sample ballot booklet.

[7]Also notice similar, although not as pronounced, peaks at various multiples of 50 starting with 200. See Beber and Scacco (2012) for a treatment on last digit-based measures of electoral fraud. In the case of the 2009 Afghan election, the relatively high number of stations with a last digit of zero in their total votes provide, according to Beber and Scacco (2012), a sign of electoral manipulation. For a more specific treatment of the 2009 Afghan election that looks precisely at the measures developed in Beber and Scacco (2012) refer to Callen and Weidmann (2013).

for a single candidate (particularly Karzai).



(a) Total votes (per polling station)

(b) Percentage votes received (per polling station)

Figure 2.1: Total Votes and Vote Percentage Received by Top Two Candidates

Notes: Frequency of total votes received by the top two candidates at the station level. Sample restricted to stations where candidates obtained more than zero votes. Bar width is 1. In panel (b), sample is restricted to stations where a candidate obtained a positive share and to stations were 100 or more total votes were cast. Bar width is 0.01.

The ECC classified 3,376 stations, or nearly 15 percent of all stations, as potentially fraudulent (i.e., falling in one of the three fraud categories mentioned above). Ultimately, the ECC performed a partial audit of all suspect stations given the need to determine, in a timely manner, whether a runoff election was needed. Particularly, 10 percent of the qualifying stations were randomly selected for a thorough investigation. From the inspected stations, the ECC created a "fraud coefficient" for each of the three categories described above. In essence, the fraud coefficients are the percentage of votes found to be fraudulent out of the total votes inspected within the category. Some indicators of fraud were: ballot boxes with broken or tampered seals, uniform markings in most ballots, discrepancies in tally sheets and box totals, etc.

On October 18, nearly two months after election day, the ECC released the results of the audit. Once suspect votes were eliminated from the count, Hamid Karzai's vote share dropped from 54.6 to 49.67 percent, while the vote share of his main challenger, Abdullah Abdullah, went from 27.8 to 30.59 percent. In lieu of the results, the ECC ordered an immediate runoff election. However, the runoff election did not take place as main challenger Abdullah withdrew from the race.

CHAPTER 3

THEORETICAL FRAMEWORK

The objective of this chapter is twofold: (i) to present a theoretical model that illustrates the link between cell phone coverage and electoral fraud, and (ii) to define channels, other than the social monitoring effect, that may equally affect fraud through coverage. I consider the problem of a candidate determining the purchase of fraudulent and legal votes at a polling center. In the case of fraudulent votes, I follow Callen and Long (2015) by assuming that the candidate pays an upfront price to an election official in charge of a polling center. This price takes into account the possibility that the center is audited as a result of complaints received from individuals at the polling center (i.e., social monitoring). With this in mind, social monitoring enters the model as a shifter in the price of fraudulent votes.[1] I refer to the individuals reporting fraud as transmitters. The price of legal votes takes into account the risk of violence that the voters face at the polling place. Higher levels of expected violence require a higher price. Lastly, given the prices of fraudulent and legal votes, the candidate then chooses an optimal level of each to purchase subject to his available campaign funds.

3.1 The Transmitter's Problem

Consider a polling center serving n voters. Furthermore assume that there exists a nationwide phone hotline to report electoral fraud. Given the widespread use of cell phones (as opposed to fixed-line phones) in the developing world and Afghanistan in particular, suppose individual i uses a cell phone if he decides to report fraud. Reporting fraud carries a physical cost $c(D)$ where D indicates the accessibility of the medium (cell phones) used to report fraud. In the context of this study D is an indicator for whether the polling center is located in an area with cell phone

[1] In section (3.3) I show that it also has a direct impact on the utility of the candidate as fraudulent votes are dropped if a center is audited which lowers the candidate's utility

coverage.[2] Specifically, let the cost of reporting fraud equal $\underline{c}$ if the center is on an area with cell phone coverage (i.e., $D = 1$) and $\bar{c}$ otherwise with $\bar{c} > \underline{c}$.[3]

Furthermore, assume that reporting fraud gives i a utility gain λ_i that can be interpreted as a "warm glow" parameter or i's satisfaction from his pro-social behavior. The individual's net payoff from reporting fraud is therefore given by $\lambda_i - c(D)$. He will then decide to report fraud if:

$$\lambda_i \geq c(D) \tag{3.1}$$

Assuming λ_i is distributed among voters at the center with probability function $G(\lambda)$ then the probability of an individual making a report given coverage status D is given by $\rho(D) = 1 - G(D)$.[4]

3.2 The Election Official's Problem

The candidate purchases fraudulent votes from an election official overseeing polling center j.[5] The price of these votes has to guarantee the official's compliance to sell them.[6] This price takes into account the probability that the center is audited as a result of reports by ordinary individuals (i.e., social monitoring). Assume the candidate and the official have perfect information over the distribution of λ_i and thus assess that the number of submitted fraud reports r follow a random process with probability function $H(r; n, \rho(D))$ that takes $\rho(D)$ and n as parameters.[7] Furthermore,

---

[2] In reality D should indicate whether there is coverage in the area where the individual decides to report fraud (the polling center, his house, etc.). This information is unavailable hence I only consider coverage at the polling center. This implicitly assumes that the call to report fraud is made right at the center. However, since centers were located within settlements, it is likely that any calls are made within the "catchment area" of the center so that using the center as a reference in determining coverage should not greatly affect the analysis of the model.

[3] Without loss of generality, I assume that the reporting cost on the non-coverage side $\bar{c}$ is constant, however, this cost might increase as polling center are further away from the coverage boundary. Refer to Appendix B for a discussion of an alternative specification of the reporting cost function that uses a smooth, non-linear function on the non-coverage side.

[4] $G(.)$ is actually a function of $c(D)$, which in turn, is a function of D. Refer to Appendix B for an extension of Equation 3.1 that considers the possibility of free-riding when reporting fraud.

[5] Although there are other methods for committing fraud, ballot stuffing and manipulation of total counts by officials seemed to be the most prevalent for of fraud during the 2009 Afghan election (see Callen and Weidmann (2013))

[6] The offered price must be so that the official's incentive compatibility constraint binds.

[7] A possible parameterization for $H(.)$ is a Poisson distribution with the mean rate given by $N\rho$. In such case, the assessed probability that center j is audited can be written as: $Pr(r \geq \bar{r}) = \pi(\bar{r}, \rho, N) = 1 - \sum_{r=0}^{\bar{r}-1} \frac{(N\rho)^r \exp(-N\rho)}{r!}$.

assume that the center is audited if the number of reports r exceeds a predetermined threshold $\bar{r}$ so that the probability that the center is audited is given by:

$$\pi(\bar{r}; n, D) = 1 - H(\bar{r}; n, \rho(D)) \tag{3.2}$$

Letting v_f, p_f, and F denote the number of fraudulent votes, their price, and a marginal fine respectively, I follow Callen and Long (2015) by assuming that the official faces a lottery in which he expects to be caught with probability π, receive fraud revenues $p_f v_f$ net of a total fine $v_f F$ or succeed with complementary probability $1 - \pi$ and pocket all fraud revenues instead.[8] Assuming that the official is an expected income maximizer then the minimum price per fraudulent vote that guarantees compliance to sell fraudulent votes is given by the expression:

$$\pi \left(p_f v_f - v_f F\right) + (1 - \pi) \, p_f v_f = 0$$
$$p_f = \pi F \tag{3.3}$$

where π is obtained from Equation 3.2 and it is assumed that the official receives an offer from only one of the candidates (i.e., payoff from non-compliance is zero).[9]

3.3 The Candidate's Problem

The candidate must decide how many votes (both legal and fraudulent) to buy from each center j. Assume that the auditing agency can differentiate between fraudulent and legal votes so that, once audited, any fraudulent votes are dropped and the candidate only receives legal votes v_l. In case where the center is not audited the candidate simply keeps all votes $v_l + v_f$. I consider the price of legal votes v_l to be a function of a parameter a that characterizes each village's affinity towards the candidate. Villages where the candidate is liked require a lower legal price per vote

[8] I assume that once a fraudulent center is audited, the candidate and polling center manager are penalized. Therefore I do not consider any "concealment technology" as in Cremer and Gahvari (1994)

[9] I rely on this assumption to simplify the analysis but also because the pattern observed in the data suggests that most fraud took place in areas where there was potentially a connection between the candidate and the official (e.g., same ethnicity or tribe), which might suggest that officials only received fraud offers from candidates of their liking otherwise we would expect to see fraud spread around different areas as well.

to entice constituents to vote.[10] Further, since elections in conflict zones are often characterized by violence, I also consider the price of legal votes to be a function of an exogenous probability δ that a violent event takes place at polling center j and as a result the village receives a negative payoff P. This consideration is particularly important in the Afghan context as the Taliban issued several warnings targeting polling centers and voters on election day (Gall (2009), Filkins (2009)). With this in mind, I define the price of legal votes as $p_l = f(\delta, P, a)$ with $\frac{\partial f(.)}{\partial \delta} > 0$, $\frac{\partial f(.)}{\partial P} > 0$, and $\frac{\partial f(.)}{\partial a} < 0$.[11]

Given the assessed probability of an audit in Equation 3.2 and assuming that the candidate has quasilinear preferences over votes, then the maximization problem of the candidate is given by:

$$\max_{v_l, v_f} \quad \pi v_l + (1-\pi)\left[v_l + v_f^{\alpha}\right]$$

$$\text{subject to} \quad p_f v_f + p_l v_l \leq E$$

where fraudulent votes enter non-linearly (with $\alpha \leq 1$) to capture the possibility that fraudulent and legal votes are not perfect substitutes and E is some campaign endowment of the candidate.[12] The solution to the problem above provides an optimal relationship between fraudulent votes and their price p_f.[13] Substituting the expressions for prices p_f and p_l in order to obtain the equilibrium level of fraud gives:

$$v_f = \left[\frac{\alpha(1-\pi) \cdot f(\delta, P, a)}{\pi F}\right]^{\frac{1}{1-\alpha}} \tag{3.4}$$

Given expressions 3.2 and 3.4 the main prediction of this model is that: Given an increase in the audit probability π due to coverage availability (i.e., $D = 1$ in Equation 3.2), then the equilibrium

[10]This legal price of votes can be interpreted as advertising costs, campaigning expenditures, etc.

[11]Refer to Appendix B.3 for an extension of the model that derives an expression for the legal price of votes.

[12]The quasilinear specification deviates from Callen and Long (2015) perfect substitutes specification. The appeal of the quasilinear specification is that it avoids a prediction where the candidate simply substitutes to all fraudulent or all legal votes as soon as the relative price deviates from 1. The studied sample shows a combination of fraudulent and legal votes for the most part not corner solutions like the ones obtained from a perfect substitutes specification.

[13]Recall that the quasilinear specification of the candidate's utility implies that there might exist a corner solution where the candidate only consumes fraudulent votes. More specifically, $v_f = \frac{E}{p_f}$ if $m \geq \alpha^{-1} p_f^{\alpha} E^{1-\alpha}$. I consider the interior solution only because polling centers with a share of fraudulent votes equaling 1 were rare.

fraud level v_f decreases. Notice, however, that the effect of social monitoring is one among others that explain fraud. To see this more clearly, I rewrite expression 3.4 by separating the different components of fraud:

$$v_f = \left[\alpha \cdot \underbrace{\frac{1-\pi}{\pi}}_{\substack{\text{Social}\\\text{monitoring}\\\text{effect}}} \cdot \underbrace{\frac{1}{F}}_{\substack{\text{Penalty}\\\text{effect}}} \cdot f(\underbrace{\delta, P}_{\substack{\text{Violence}\\\text{effect}}}, \underbrace{a}_{\substack{\text{Candidate}\\\text{affinity}\\\text{effect}}}) \right]^{\frac{1}{1-\alpha}} \tag{3.5}$$

I highlight three key results. First, fines lower fraud. Second, an increase in the likelihood or magnitude of violence (given by δ and P respectively) increases the price of legal votes and as a result increases fraud by making fraudulent votes less expensive relative to legal votes. I refer to this effect as the *violence effect*. Notice, however, that violence might also lead to fraud even in polling centers where the candidate is liked (i.e., areas with high a) since the price of legal votes might be too high (i.e., $f(\delta, P, a)$ is still high) and thus the candidate must substitute legal votes for fraudulent ones.[14] With this in mind, the main purpose of this paper is to empirically disentangle the social monitoring effect.

[14]This is a key result considering that fraud was widespread in areas were Karzai had strong support which were also the areas with the highest levels of violence.

CHAPTER 4

THE EFFECT OF COVERAGE ON FRAUD

This chapter describes the data and variables used in the empirical analysis. Further, it presents the empirical framework used to determine the effect of cell phone coverage on electoral fraud. The chapter concludes by presenting the results obtained from the estimation as well as several alternative specifications designed to assess the robustness of the results.

4.1 Data and Variables

I describe three key pieces of information: a measure of electoral fraud, a measure of accessibility to the medium used to report fraud, in this case a cell phone, and other variables used to assess the validity of the RD design.

4.1.1 Measures of fraud

I use the list of polling stations that were subject to the audit and the ECC fraud categories to define various measures of fraud. These measures constitute the primary outcome variables for most of the paper. I first aggregate the six fraud categories used by the ECC into three broader categories: *Category A* (stations with 600 or more votes cast), *Category B* (stations in which one candidate received 95 percent or more of the total votes cast), and *Category C* (stations satisfying categories *A* and *B* above).[1] For each of these categories, I define the primary measure of fraud as the polling center level vote share of stations qualifying in each of the categories. More specifically, given a polling center c with a total of s stations of which $n \leq s$ qualify for category $j \in \{A, B, C\}$ above, then the measure of type j fraud at center c is given by the total number of votes in the n suspect stations divided by the total votes cast in center c. To ease notation, I simply refer to them as Category j fraud for the remaining of the paper. Lastly, I use the data on stations that were disqualified, due to complaints, prior to the audit to define an additional fraud category: the vote

[1]Refer to Appendix A for a detailed description on the construction of the three categories.

share of both, disqualified stations and stations qualifying in *Category C*. I refer to this measure as *Category C+* fraud.

Note that although the measures defined above are referred to as “measures of fraud”, the fact that a station qualifies for one of the categories does not necessarily imply that fraud was committed in this station. One may have, for instance, stations with unusually high voter turnout rates or with unusually strong preferences for one specific candidate. With this in mind, one should interpret this measure as a signal or proxy of fraud. These proxies, however, provide a precise signal on actual fraud given the context and design of the election. For instance, in the case of *Category C* fraud, more than 96 percent of the ballots inspected in stations satisfying this category were actually found to be fraudulent (Electoral Complaints Commission 2010). For convenience, however, I will refer to the constructed measures above as fraud measures for the remainder of the paper.

4.1.2 Cell phone coverage

Cell phones are the primary medium of communication in Afghanistan since fixed line phones are relatively scarce.[2] To determine areas with cell phone coverage and, hence, where the monitoring initiative could be implemented in principle, I use GSM (2G) coverage maps directly provided by cell phone operators to the GSM Association and distributed by *Collins Bartholomew*.[3] The coverage maps indicate areas receiving 2G coverage based on the spatial distribution of cell phone towers across Afghanistan. Specifically, coverage data are in the form of a map raster or grid file indicating cells where signal strength is at least -100 decibel-milliwatts (dBm). This the typical *minimum received signal power* in GSM wireless networks, or broadly speaking, the threshold indicating the ability to make a call (Figueiras and Frattasi 2010). Figure 4.1a shows the 2G coverage raster file overlaid on a topographical map of Afghanistan. Shaded areas indicate areas with

---

[2]The number of mobile subscribers in Afghanistan rose from 1.7 million in 2006 to around 17.1 million in 2012. This translates into a mobile penetration (number of subscribers per 100 inhabitants) of 6.27 in 2006 and 63.3 in 2012. To put these numbers in perspective, the number of fixed line phone subscribers, for example, was only 110,000 in 2012, even less than the number of Internet users (Hamdard 2012).

[3]GSM is the type of cellular technology used by the Afghan cell phone companies in my sample. The GSM Association is a group comprising most GSM cell phone providers around the world. 2G stands for second-generation and this is the cellular network technology allowing mostly voice calls only (i.e., technology preceding smart phone technology or 3G). The dataset is called the *Collins Coverage Explorer* and more information can be found at: http://www.collinsbartholomew.com/

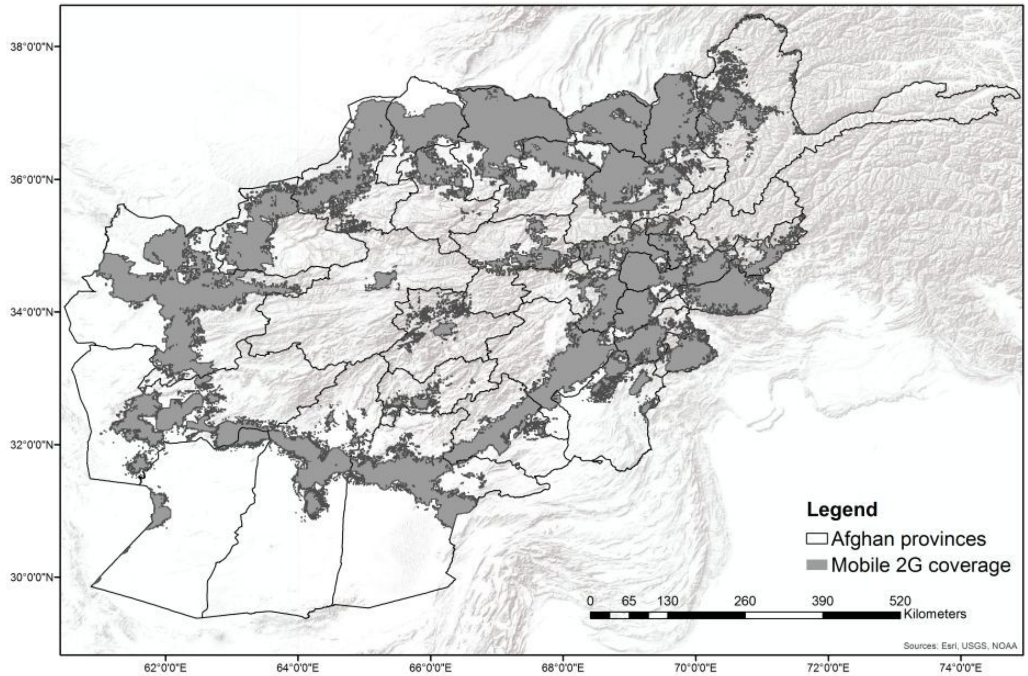
a signal strength of at least -100 dBm. Within coverage areas, however, *Collins Bartholomew* does not provide information on how the strength of coverage varies. In the case of Afghanistan, *Collins Bartholomew* provides information on two of the largest operators, MTN and Afghan Wireless (AWCC). These two operators encompass about 46 percent of all cell phone subscriptions in Afghanistan with more than 8 million subscriptions combined (Hamdard 2012).

The lack of data on other providers may be a source of concern since some non-covered areas may be wrongly classified as covered. However, a detailed inspection of cell phone tower locations in 2012 using maps provided by the Afghan Telecommunication Regulatory Authority (ATRA) suggests significant overlap in tower locations and service areas between MTN, AWCC and other operators. For the case of smaller operators (Etisalat, Wasel, and Afghan Telecom) the tower locations are entirely contained within the coverage areas of MTN and AWCC (Afghan Telecommunication Regulatory Authority 2012).[4] An additional source of concern may be the possibility of operators over-reporting coverage areas. This might result if operators desire to overstate their service areas for marketing purposes or to mislead competitors, for example. However, coverage data submissions by operators are considered to be a service to the GSM Association. Operators provide data at no cost which then the GSM Association uses to assess the state of the technology and to sell it as a way of raising funds for the agency to operate. With this in mind, the data are restricted to the general public and require a contractual agreement to purchase and use for research. Therefore, it seems unlikely that operators have an incentive to misreport coverage in such cases. Section 4.3 offers a detailed description on the implications of missing operators and over-reporting of coverage areas in terms of the empirical strategy.

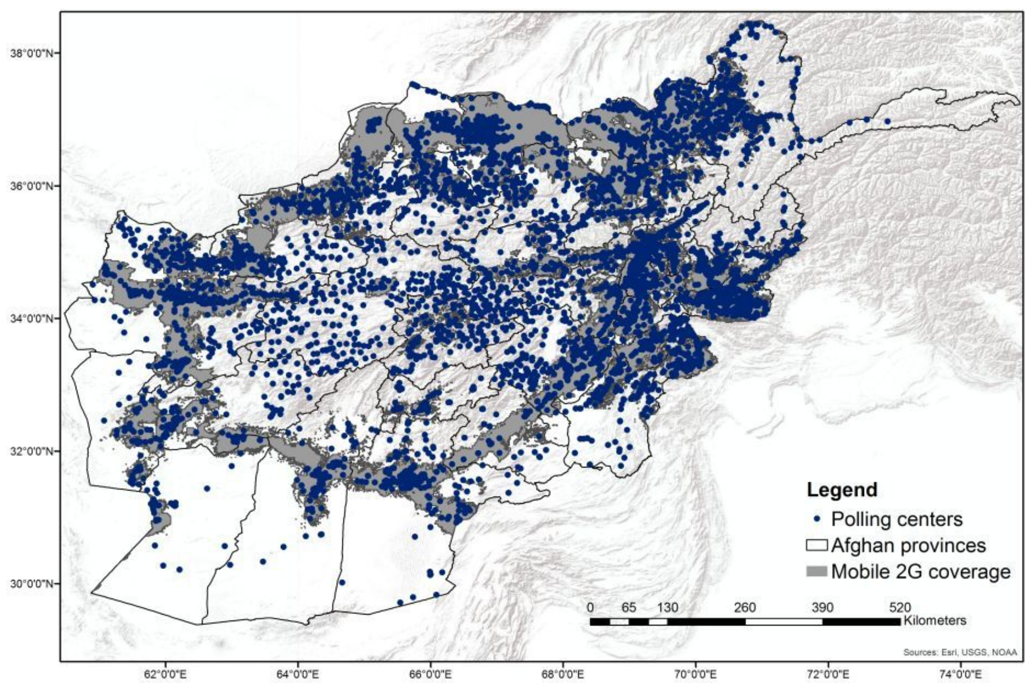
4.1.3 Polling Center Characteristics

To better assess the validity of the RD design, I collect data on electoral outcomes, polling centers' physical characteristics, and geographic, economic, and demographic indicators for areas around polling centers. Further, I obtain data on the latitude and longitude of polling centers from an IEC-led nationwide inspection of each polling center that took place less than a year after

[4] The earliest available information on cell phone tower locations is June, 2012.



(a) Mobile 2G Coverage, 2009



(b) Mobile 2G Coverage and Polling Centers

Figure 4.1: Mobile 2G Coverage and Polling Centers

Notes: Shaded areas represent availability of 2G GSM cell phone coverage for two largest cell phone providers in Afghanistan (MTN and AWCC) for the year 2009. Dots give the location of polling centers during the 2009 Afghan presidential election. Lines demarcate the provinces of Afghanistan. Map overlaid on USGS topographic basemap.

the 2009 presidential election. The purpose of the inspection was to assess the security status and accessibility of designated polling centers for the upcoming September, 2010 parliamentary election. The assessments were conducted jointly by ISAF and Afghan National Security Force teams. Each assessment contained four pieces of information: a polling center name and code, an MGRS grid providing the exact geographic location of the polling center, and a road accessibility status. Using the coordinates, I overlay the centers on the cell phone coverage map to determine each center's coverage status. Figure 4.1b depicts the spatial distribution of polling centers along the coverage areas.

To create a sample containing the fraud measures per center along with the geographic location of the centers, I merge the 2009 fraud data with the 2010 center assessment data described above. The data are merged based on the polling center code and name. In cases where the codes matched but the names did not (100 cases), the match was done based on the names only. The total sample consists of 6,160 polling center observations for which 5,904 (95.8 percent) have coordinates obtained directly from the 2010 assessment. For the remaining 256 centers coordinates were imputed as follows: 169 (2.7 percent) used the centroid coordinates of the village or settlement where the center was located, 81 (1.3 percent) used the coordinates of the center with the identifier code closest to it and lastly 6 (0.1 percent) simply used the coordinates of the district capital where the center was located. Appendix Table C.1 provides a detailed description of the sample and imputations used.

I use the released electoral results to obtain additional election-related outcomes: the number of expected voters prior to election day, the total votes cast at the center, the total number of stations per center, the voter turnout rate, and the percentage received by the two main candidates.[5] These data are complemented with pre-election data published by the IEC on polling center type (school, mosque, or other) along with the share of stations within a center designated to women and Kuchis.[6]

[5]The results data are publicly accessible at: www.iec.org.af/results_2009/

[6]Kuchis are a group of Pashtun nomads.

I use GIS resources to capture geographic and economic development characteristics of the area where each center is located. Information on exogenous geographic characteristics, namely polling center elevation and slope, is obtained from NASA's Shuttle Radar Topography Mission (SRTM30) (National Aeronautics and Space Administration and the National Geospatial Intelligence Agency 2000). I calculate distances from polling centers to primary and secondary roads, district hospitals, basic health centers, and primary, secondary, and seasonal rivers using vector files collected in 2005 by the Afghanistan Information Management Service (AIMS) and obtained from the Empirical Studies of Conflict Project (AIMS 2005).[7]

Lastly, demographic data on the population and ethnic composition around the location of the polling center comes from the Measuring Impacts of Stabilization Initiatives (MISTI) project sponsored by US Agency for International Development (USAID). The MISTI project (MISTI 2013) includes geographic coordinates and compiles demographic data from various data sources between the years 2012 and 2013 for more than 45,000 villages across Afghanistan.[8] Using these data, I create variables indicating the population size and the language spoken ("Pashto", "Dari", and "Other") in the village or settlement where the polling center is located.

4.2 Empirical Framework

This section presents the spatial RD framework used to estimate the effect of cell phone coverage on fraud. It also describes and tests the validity of the identifying assumptions.

4.2.1 Regression Discontinuity Design (RD)

Note from Figure 4.1a that: (1) cell phone coverage is a discontinuous function of latitude and longitude, and (2) changes from coverage to non-coverage areas define a two-dimensional boundary along the latitude-longitude space. With this in mind, I employ a spatial regression discontinuity (RD) design that takes advantage of the discontinuity in polling centers' cell phone access to estimate the effect of cell phone coverage on various election fraud outcomes. I present results using two approaches. First, I exploit the two-dimensional nature of the coverage boundary

[7] I consider river proximity to be a measure of development since a large portion of the Afghan population depends on agriculture as a mean of subsistence.

[8] See Appendix Figure C.5 for the distribution of villages across Afghanistan

to estimate *conditional treatment effects* at various points along the treatment boundary following Imbens and Zajonc (2011).[9] Second, I follow the usual approach in the literature by specifying a one-dimensional forcing variable, namely the distance to the closest point in the coverage boundary.[10] This is the equivalent of subtracting the cutoff value from the forcing variable in the one-dimensional design and then using this transformed forcing variable to estimate a single, boundary-wide average effect.

Broadly speaking, the first approach estimates treatment effects using observations within a neighborhood of a specific point in the treatment boundary. This exercise is then repeated for various points along this boundary thus providing a distribution of these effects along this dimension. However, since there are not enough observations within several neighborhoods to allow for consistent estimation of the conditional treatment effects, I propose a modification that uses all available observations. More specifically, let $\mathbb{C}$ and $\mathbb{B} = \text{bd}(\mathbb{C})$ denote the cell phone coverage area and its boundary respectively (i.e., shaded areas and the corresponding boundary between shaded and non-shaded areas in Figure 4.1a). Let $\mathbf{X}$ denote the latitude and longitude vector of a polling center. With this in mind, polling center j receives treatment assignment (i.e., covered) if its corresponding coordinate vector $\mathbf{x}_j = (\text{longitude}_j, \text{latitude}_j)$ falls within the coverage area $\mathbb{C}$. Let $\mathbf{b}_i$ with $i = 1, \ldots, I$ denote the coordinate vector of point i on the treatment boundary $\mathbb{B}$ (represented by the colored points in Figure 4.4). Furthermore, let $N_h(\mathbf{b}_i)$ denote a neighborhood of size h km around this point with $N_h^+(\mathbf{b}_i)$ and $N_h^-(\mathbf{b}_i)$ denoting the subset of this neighborhood that falls on the coverage and non-coverage sides of the boundary respectively. As shown in Imbens and Zajonc (2011), the conditional treatment effect at point $\mathbf{b}_i$, denoted as $\tau(\mathbf{b}_i)$ is therefore given by:

$$\tau(\mathbf{b}_i) = \lim_{\mathbf{X}\to\mathbf{b}_i} E\left[v_f|\mathbf{X} \in N_h^+(\mathbf{b}_i)\right] - \lim_{\mathbf{X}\to\mathbf{b}_i} E\left[v_f|\mathbf{X} \in N_h^-(\mathbf{b}_i)\right] \tag{4.1}$$

[9] Although there are multiple studies exploring RD methods with a multidimensional forcing variable (e.g., Reardon and Robinson (2010), Papay, Willett, and Murnane (2011), Wong, Steiner, and Cook (2010), Keele and Titiunik (2013)), we mostly follow the notation and terminology in Imbens and Zajonc (2011).

[10] See Holmes (1998), Black (1999), Kane, Riegg, and Staiger (2006), Lalive (2008), and Dell (2010) for examples of papers employing an RD design with distance to the treatment threshold as the forcing variable.

where v_f is a measure of electoral fraud. I estimate $\tau(\mathbf{b}_i)$ using Local Linear Regression, which has better boundary properties than other nonparametric estimators (Fan (1992), Fan and Gijbels (1996)) and has been shown to provide a consistent estimate of the treatment effect in an RD setup (Hahn, Todd, and van der Klaauw (2001), Porter (2003), discussed in Lee and Lemieux (2010)). More specifically, I estimate:

$$v_{f,ij} = \gamma + \beta D_{ij} + \mathbf{X}'_{ij}\alpha + D_{ij}\mathbf{X}'_{ij}\delta + \Omega_j + \epsilon_{ij} \tag{4.2}$$

for centers within h kilometers of the coverage boundary[11] and where $v_{f,ij}$ denotes a fraud measure for polling center j in neighborhood i, D_{ij} is an indicator equaling one if the center lies within the coverage area, $\mathbf{X}_{ij}$ is the geographic coordinate of center j in neighborhood i, and Ω_j is a neighborhood fixed effect. I choose h optimally as in Imbens and Kalyanaraman (2012). Lastly, to comply with the Boundary Positivity assumption discussed in Imbens and Zajonc (2011), I restrict the sample to only neighborhoods with at least one polling center on each side of the coverage boundary.[12]

From Equation 4.2, under certain conditions, a consistent estimator for $\tau(\mathbf{b}_i)$ (i.e., the causal effect of coverage on election fraud outcomes) is given by:

$$\hat{\tau}(\mathbf{b}_i) = \hat{\beta} + \mathbf{b}'_i\hat{\delta} \tag{4.3}$$

Such conditions are discussed in detail in the following section. In order to evaluate the treatment effect at various points in the boundary, I follow Imbens and Zajonc (2011) by choosing a number of evenly spaced boundary points $\mathbf{b}_i$ that cover the boundary reasonably well.

[11]This is the equivalent of using a rectangular kernel with bandwidth h. I rely on this simple kernel since Lee and Lemieux (2010) argue that kernel choice has little impact in practice therefore simple kernels (i.e., rectangular) can be used for convenience. The appeal of choosing a rectangular kernel is therefore that, since all observations receive a constant weight, the estimation simply reduces to an unweighted linear regression.

[12]Boundary Positivity requires the existence of observations near the boundary in order to identify the treatment effect in the multidimensional RD setting. More specifically, Boundary Positivity requires that for all $\mathbf{b}_i$ and $\epsilon > 0$, there are polling centers for which $P(\mathbf{x}_j \in N_h(\mathbf{b}_i)) > 0$.

I highlight two points regarding the modification proposed above: first, estimation of the conditional treatment effects follows from using the actual levels of the forcing variable (i.e., latitude and longitude) rather than the normalized levels (i.e., the distance to the boundary) as it is usually done in the literature. Simply put, this is the equivalent of *not* subtracting the threshold from the forcing variable in the one-dimensional case. From the estimation Equation 4.2, this guarantees that the RD polynomial $\mathbf{X}_{ij}'\alpha + D_{ij}\mathbf{X}_{ij}'\delta$ does not collapse to zero as the forcing variable converges to the treatment boundary. This, in turn, guarantees an estimate of the treatment effect that depends on a given value of the boundary (i.e, $\mathbf{b}_i$). Second, notice that Equation 4.2 uses all observations within a window h around the boundary rather than only the observations within a neighborhood of a chosen point $\mathbf{b}_i$, which in most applications might not yield a large enough sample size.

As discussed in Imbens and Zajonc (2011) and using the estimated conditional treatment effects from Equation 4.3, I estimate a boundary average effect, τ, as:

$$\hat{\tau} = \frac{\sum_{i=1}^{I} \hat{\tau}(\mathbf{b}_i) \cdot \hat{f}(\mathbf{b}_i)}{\sum_{i=1}^{I} \hat{f}(\mathbf{b}_i)} \tag{4.4}$$

where $\hat{f}(.)$ is the estimated bivariate density of polling centers' coordinate vectors evaluated at boundary points $\mathbf{b}_i$.[13] Following the notation described above, expression 4.4 provides an estimate of the average effect τ given by $\int_{\mathbf{x}\in\mathbb{B}} \tau(\mathbf{x}) f(\mathbf{x} \mid \mathbf{X} \in \mathbb{B}) d\mathbf{x} = \frac{\int_{\mathbf{x}\in\mathbb{B}} \tau(\mathbf{x}) \cdot f(\mathbf{x}) d\mathbf{x}}{\int_{\mathbf{x}\in\mathbb{B}} f(\mathbf{x}) d\mathbf{x}}$. In subsequent discussions of results, I refer to the estimate in Equation 4.4 as the *averaged conditional treatment effects*.

In the case of the one-dimensional approach, I estimate various specifications of the equation below:

$$v_{f,ij} = \gamma + \beta D_{ij} + g(\mathbf{X}_{ij}) + \Omega_j + \epsilon_{ij} \tag{4.5}$$

where the RD polynomial $g(\mathbf{X}_{ij})$ and sample restrictions vary with each specification. Specifically, I present results for three specifications. First, a Local Linear Regression with $g(\mathbf{X}_{ij}) =$

[13] Appendix Figure C.3 provides an illustration of estimates of $\hat{f}(.)$. I estimate the bivariate density via kernel density estimation using the *Kernel Density* tool in ArcGIS Spatial Analyst.

$\alpha \cdot dist_{ij} + \delta D_{ij} \times dist_{ij}$ where $dist_{ij}$ denotes the Euclidean distance between polling center j and the closest point on the coverage boundary and the estimation sample is restricted to polling centers falling within a bandwidth around the coverage boundary.[14] Following a parametric approach, the remaining specifications use all observations on either side of the coverage boundary, however, I allow a more flexible form for the RD polynomial by using higher order polynomials in distance to boundary and latitude and longitude, respectively. For instance, the RD polynomial of order K in the case where distance to the boundary is the forcing variable is given by $g(\mathbf{X}_{ij}) = \sum_{k=1}^{K} \alpha_k \cdot dist_{ij}^k + \delta_k D_{ij} \times dist_{ij}^k$. The optimal order of the chosen polynomial specification is determined using Akaike's criterion as in Black, Galdo, and Smith (2007) and suggested in Lee and Lemieux (2010). RD coefficient β gives the causal effect of cell phone coverage on fraud for areas in close proximity to the coverage boundary.

4.2.2 Validity of the RD Identifying Assumptions

Identification of $\tau(\mathbf{b}_i)$ requires a key assumption: potential outcome functions $E\left[v_f(1)|\mathbf{X}\right]$ and $E\left[v_f(0)|\mathbf{X}\right]$ must be continuous at point $\mathbf{b}_i$ in the treatment boundary.[15] Simply put, polling center characteristics (including unobservables) must transition smoothly across the treatment boundary. This assumption allows for centers in the non-coverage side to serve as a valid counterfactual for centers in the coverage side.

[14]The bandwidth is chosen optimally as in Imbens and Kalyanaraman (2012)

[15]One and zero denote assignment and non-assignment into treatment, respectively.

Table 4.1: Mean Comparison for Various Polling Center Characteristics

	Within 10 km of boundary			Within 5 km of boundary			RD estimates	
	Coverage (1)	No Coverage (2)	S.E. (3)	Coverage (4)	No Coverage (5)	S.E. (6)	RD coeff. (7)	S.E. (8)
Fraud outcomes (Category C+ fraud)								
All regions	0.08	0.11	(0.02)**	0.08	0.12	(0.02)**	-0.04	(0.019)**
East and South	0.14	0.20	(0.03)**	0.13	0.20	(0.03)**	-0.08	(0.032)**
North and West	0.01	0.01	(0.01)	0.01	0.01	(0.01)	0.00	(0.010)
Electoral outcomes								
No. of stations	4.09	3.73	(0.16)**	3.86	3.78	(0.18)	0.09	(0.175)
No. of expected voters	2194.00	1944.00	(94.04)***	2069.00	1979.00	(100.00)	94.33	(101.198)
Total votes	871.80	866.60	(56.34)	835.00	863.70	(62.41)	-28.22	(59.992)
Voter turnout	0.43	0.50	(0.02)***	0.45	0.49	(0.02)	-0.03	(0.023)
Vote share:								
Karzai	0.50	0.49	(0.03)	0.50	0.49	(0.03)	0.01	(0.030)
Abdullah	0.34	0.33	(0.03)	0.35	0.33	(0.03)	0.01	(0.031)
Polling center characteristics								
Polling center type:								
Mosque	0.24	0.26	(0.03)	0.25	0.26	(0.03)	-0.01	(0.029)
School	0.46	0.37	(0.03)***	0.44	0.37	(0.03)**	0.06	(0.034)*
Other type	0.30	0.37	(0.03)**	0.30	0.37	(0.03)*	-0.05	(0.032)
Polling center access (2010):								
Road access	0.76	0.78	(0.04)	0.73	0.78	(0.04)	-0.05	(0.039)
Limited access	0.08	0.07	(0.02)	0.10	0.07	(0.02)	0.03	(0.023)
Other access	0.16	0.15	(0.03)	0.17	0.15	(0.03)	0.02	(0.032)
Share female stations	0.44	0.45	(0.01)	0.44	0.45	(0.01)	-0.01	(0.012)
Share Kuchis stations	0.04	0.03	(0.01)	0.03	0.04	(0.01)	0.00	(0.010)

(Continues)

Table 4.1: Mean Comparison for Various Polling Center Characteristics - *Continued*

	Within 10 km of boundary			Within 5 km of boundary			RD estimates	
	Coverage (1)	No Coverage (2)	S.E. (3)	Coverage (4)	No Coverage (5)	S.E. (6)	RD coeff. (7)	S.E. (8)
Geographic characteristics								
Elevation (meters)	1570.00	1782.00	(58.06)***	1617.00	1756.00	(50.71)***	-128.72	(50.617)**
Slope (percent)	5.72	7.57	(0.53)***	6.46	7.66	(0.58)**	-1.01	(0.602)*
Economic development characteristics								
Distance (km) to:								
Primary road (2005)	35.30	48.62	(2.43)***	40.72	47.07	(2.31)***	-5.70	(2.290)**
Secondary road (2005)	44.65	52.60	(3.37)**	50.08	49.84	(3.12)	0.49	(3.056)
District hospital (2005)	37.56	45.42	(2.76)***	40.25	42.68	(2.73)	-2.34	(2.739)
Basic health center (2005)	20.12	24.31	(1.87)**	21.78	22.68	(1.74)	-0.38	(1.837)
Primary river	17.45	18.38	(1.40)	18.00	17.90	(1.29)	0.19	(1.319)
Secondary river	8.11	8.71	(1.12)	8.96	8.03	(1.10)	0.61	(1.040)
Seasonal river	12.66	11.05	(1.54)	13.16	10.30	(1.77)	2.70	(1.669)
District/Province capital	0.05	0.03	(0.01)**	0.04	0.03	(0.01)	0.01	(0.013)
Demographic characteristics (of closest settlement)								
Population (2012-2013)	1287.00	957.00	(223.69)	1004.00	1018.00	(145.37)	-51.25	(141.521)
Language spoken (2012-2013):								
Dari	0.43	0.46	(0.04)	0.43	0.43	(0.04)	0.00	(0.040)
Pashto	0.44	0.43	(0.04)	0.43	0.45	(0.04)	-0.01	(0.041)
Other	0.13	0.12	(0.03)	0.14	0.12	(0.03)	0.01	(0.026)
Observations	891	551		601	456		601	456

Notes: Columns (1), (2), (4), and (5) give the means of the corresponding variable. Columns (3), (6), and (8) give the clustered standard errors for the difference in means in parenthesis. Sample restricted to neighborhoods with at least one observation on each side of the boundary. *, **, and *** indicate 10, 5, and 1 percent significance respectively. "Coverage" refers to cell phone coverage. *No. of expected voters* is the number of voters predicted by the IEC prior to election day. *Total votes* cast is the actual number of votes tallied at the center. *Voter turnout* is defined as the number of votes cast at the center divided by the expected number of voters. *Vote share* is the share of votes received by the each of the two main candidates divided by total votes. I report *Total votes*, *Voter turnout*, and *Vote share* for centers without evidence of fraud. The remaining variables are defined in section (4.1) of the text. Year values in parenthesis indicate the year the data was collected. I do not specify the year for variables collected in 2009 or time-invariant variables.

Table 4.1 assesses the validity of the design by comparing electoral outcomes, geographic, economic, and demographic characteristics for centers on each side of the coverage boundary. In addition, it investigates how the primary fraud outcome measure varies across the boundary relative to other polling center characteristics. Columns (1) and (4) report the mean for polling centers within cell phone coverage areas for bandwidths of 10, and 5 kilometers, respectively. Columns (2) and (5) report the mean for centers in non-coverage areas. Columns (3) and (6) report the clustered standard error of the difference in means between covered and non-covered centers.[16] I highlight two important results. First, note that differences across the boundary for the fraud outcome variable remain economically and statistically significant as the bandwidth decreases.[17] Second, and most importantly in terms of design validity, notice that unlike the fraud measure, most differences in polling center characteristics become relatively small and statistically insignificant as the bandwidth decreases. To offer a more rigorous assessment, Column (7) presents the results from an RD analysis that estimates Equation 4.5 within a 5-kilometer bandwidth using each of the specified covariates in Table 4.1 as the outcome variable.[18] Similar to the mean difference results, the RD exercise shows that, unlike the fraud measure, center characteristics transition smoothly across the boundary for the most part. In all, 24 (out of the 28 baseline characteristics tested) result in statistically insignificant differences between covered and non-covered centers.

Polling center elevation, slope, and distance to the closest primary road are notable exceptions. Cell phone coverage depends on topographical features, thus it is plausible that coverage drops in areas with significant changes in elevation and slope. Similarly, primary road access is affected by the ruggedness of the terrain. In spite of these changes across the boundary, section 4.4 shows that the main RD results in section 4.3 are not sensitive to the inclusion of these covariates. For a better

---

[16]Standard errors are clustered at the boundary neighborhood level. Refer to section 4.3.3 for a detailed description of how boundary neighborhoods are defined. Additionally, Appendix Table C.2 shows the results in Table 4.4 using Conley (1999) standard errors. Note that the results do not differ greatly in terms of the type of clustering used.

[17]Note that the results exhibit a high degree of spatial variation with centers in the Southeast showing significant differences while differences in the Northwest region are indistinguishable from zero. I explore this result in more detail in Section 4.3

[18]More specifically, Column (6) uses a cubic polynomial in distance to the boundary as the specification of Equation 4.5. The bandwidth choice of 5 kilometers is to allow for comparability with the results from Columns (4) and (5).

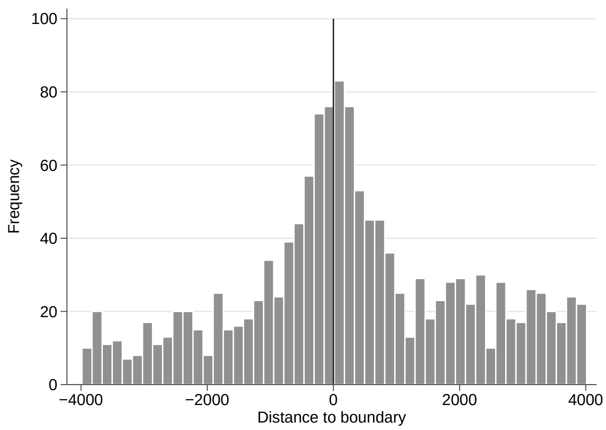
depiction of the continuity of baseline covariates across the coverage boundary, refer to Appendix Figure C.1 which presents RD plots for continuous 5-kilometer distance bins for all covariates in Table 4.1.

To further assess the validity of the identifying assumption, I perform McCrary (2008) test for breaks in the density of the forcing variable at the treatment boundary. A noticeable jump in the number of polling centers on only one side of the boundary may indicate endogenous assignment of polling centers which would invalidate the identifying assumption.[19] In the context of this study, however, endogenous sorting of centers close to the boundary is not a cause of concern since polling center locations were determined primarily by the location of settlements rather than cell phone coverage. In addition, locations were determined entirely by the U.N.-led Independent Elections Commission (IEC), thus, manipulation of the process by potentially corrupt candidates is unlikely.

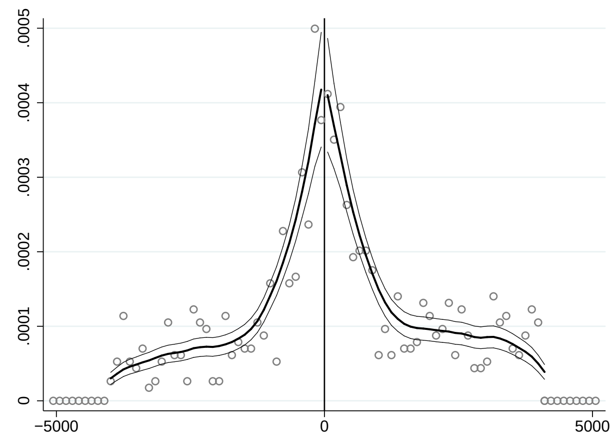
Figure 4.2a shows a histogram of the distance between polling centers and the closest point in the coverage boundary for a 4-kilometer window around this boundary. Figure 4.2b shows the results from McCrary (2008) test for discontinuities in the density of the forcing variable (distance to the boundary). "Negative" and "positive" distances denote distances for centers in non-coverage and coverage areas, respectively. A zero distance (represented by the solid vertical line) indicates the coverage boundary. Each bin has a width of 250 meters. Both figures clearly show that the density does not change discontinuously across the boundary suggesting that, for a narrow window around the coverage boundary, there seems to be no manipulation when locating polling centers. Figures 4.2c and 4.2d perform the same analysis for the distance between villages and the coverage boundary. This latter test is particularly important since the IEC located polling centers based on settlements and thus center locations may be endogenously selected (although indirectly) if the number of villages changes abruptly with coverage. Note, however, that similar to the polling center density, there is no evidence that the density of villages significantly jumps across the coverage boundary.

[19] Also referred to as "manipulation of the forcing variable" in the RD literature. See Lee (2008), McCrary (2008), or Lee and Lemieux (2010) for a treatment of this issue.

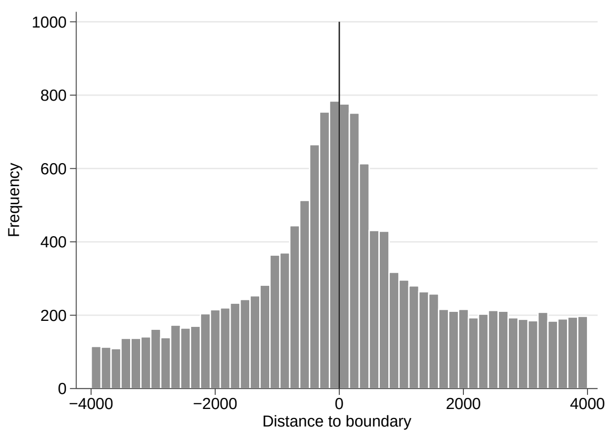
The absence of selective sorting of villages near the coverage boundary is institutionally plausible. Afghanistan experienced a period of rapid expansion in cell phone coverage throughout the second half of the 2000's. Mobile penetration (number of subscribers per 100 inhabitants) rose from 6.27 in 2006 to 63.3 in 2012 (Hamdard 2012). With this in mind, the incentives for households to move to a village that has coverage are very low when coverage might soon reach that household's village.



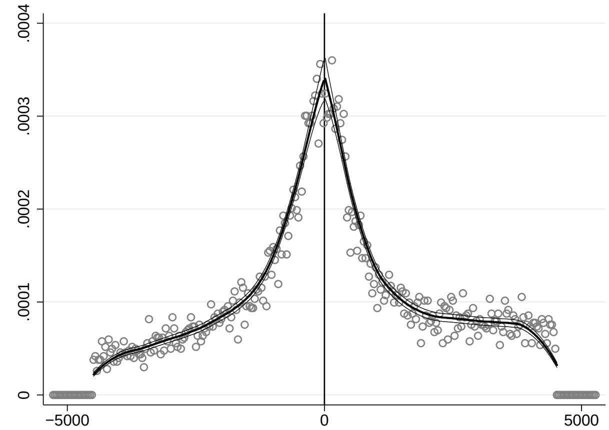
(a) Distance to coverage boundary (polling centers)



(b) McCrary's (2008) test (polling centers)



(c) Distance to coverage boundary (villages)



(d) McCrary's (2008) test (villages)

Figure 4.2: Histograms and Densities of the Forcing Variable

Notes: ”Distance to boundary” refers to the distance between a polling center (Panels a and b) and village (Panels c and d) to the closest point in the coverage boundary. Distance is measured in meters. Bin width of 160 meters. The distance to boundary (forcing variable) is normalized so that “negative” values of distance give the distance of polling centers/villages in non-coverage areas.

4.3 Results

This section begins by describing the results from a graphical analysis of the outcome variables. It then proceeds with a description of the results from the one-dimensional and boundary RD designs described in Section 4.2.1. Given the inherent differences across the Southeast and Northwest regions of Afghanistan, I present all results in this section separately by region.[20] Lastly, section 4.1 describes four alternative measures of fraud, however, to present results that are both parsimonious and informative, all results in this section use the polling center share of votes classified in Category C+ fraud as the outcome variable.[21]

4.3.1 Graphical Analysis

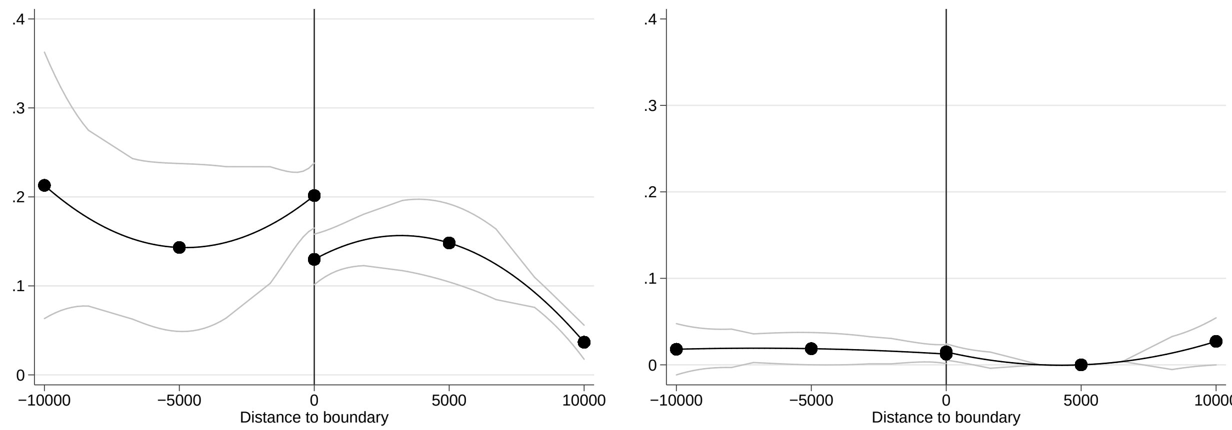
I begin by graphically analyzing the relationship between electoral fraud and cell phone coverage using RD plots of the outcome variable. Figures 4.3a and 4.3b plot the average share of Category C+ fraud for polling centers falling within 5-kilometer distance bins for the Southeast and Northwest regions, respectively. Solid dots represent the binned averages. Negative values of distance indicate polling centers in non-coverage areas. The solid line trends give the predicted values from a regression of the variable of interest on a second degree polynomial in distance to the boundary. The window of analysis is 15 kilometers on each side of the boundary and I estimate these regressions separately on each side.

Figure 4.3a shows that within a narrow window around the coverage threshold, the average level of fraud drops sharply for centers located on the coverage side. The average share of Category C+ fraud for centers within 5 kilometers of the boundary is around 7 to 8 percentage points lower on the coverage side. This compares to an average share of around 20 percent observed in centers on the non-coverage side within that same distance window. Further, average fraud levels are consistently higher on the non-coverage side and exhibit a declining trend on the coverage

[20] I define the regions based on the International Security Assistance Forces (ISAF) regional commands classification specified in the Measuring Impacts of Stabilization Initiatives (MISTI) dataset (MISTI 2013). ISAF divided Afghanistan into six regional command centers: Central (Kabul), East, North, South, Southwest, and West. I collapse the regions into Southeast (East, Central, South) and Northwest (North, West, and Southwest). Refer to Appendix Figure C.2 for a depiction of the two regions.

[21] Results using alternative measures of fraud are quantitatively similar and can be provided upon request.

side. Figure 4.3b presents the RD plot for centers in the Northwest region. Unlike centers in the Southeast, the average fraud share does not change significantly with coverage. However, this is primarily driven by the lower levels of fraud experienced in Northwest provinces in general. Notice that the average share of fraudulent votes for non-coverage centers within the 5-kilometer distance bin is about 1 percent and remains relatively low for higher distance bins.



(a) Category C+ (Southeast)

(b) Category C+ (Northwest)

Figure 4.3: Binned Averages for Category C+ fraud (RD plot)

Notes: Solid dots give the average share of votes classifying in Category C+ fraud for polling centers falling within 5000-meter distance bins. Refer to section 4.1 in the text for a detailed description of Category C+ fraud. Dots are plotted at the start of the bin (i.e., the dot representing the average for centers in the 0-5,000 meter bin is located at 0.). "Distance to boundary" refers to the distance between a polling center and the closest point in the cell phone coverage boundary. Distance is measured in meters. The distance to boundary (forcing variable) is normalized so that "negative" values of distance give the distance of polling centers/villages in non-coverage areas. The solid line trends give the predicted values from a regression of the outcome variable on a second degree polynomial in distance to the boundary that uses a rectangular kernel and a bandwidth of 15,000 meters.

4.3.2 One-Dimensional RD

In order to provide a more rigorous assessment, Table 4.2 presents the results from the one-dimensional RD design that estimates the causal impact of coverage on fraud using Equation 4.5. Results for the Southeast and Northwest regions are included in Panels A and B, respectively. Further, all specifications include boundary neighborhood fixed effects and standard errors clustered by neighborhood to account for potential spatial correlation of the error terms within neighborhoods.[22] For reference, Table 4.2 provides the mean of the dependent variable for centers in the non-coverage side.[23]

Columns (1) and (2) present results using distance to the coverage boundary as the forcing variable and the Local Linear Regression specification described in Section 4.2.1. Results in Panel A, Column (1) indicate a considerable drop in fraud levels for polling centers within the coverage area. In particular, for centers within a 5.7 kilometer bandwidth around the coverage boundary, the share of votes classifying as Category C+ fraud is about 5.2 percentage points lower in centers on the coverage side.[24] Given an average share of about 19.8 for centers on the non-coverage side, this represents about a 26 percent drop in the share of fraudulent votes. Note in Panel A Column (2) that although the estimate is not statistically significant at the given bandwidth, once it is slightly increased to about 6 kilometers, precision improves and the estimate becomes statistically significant. Furthermore, Figure 4.6a shows that the results remain statistically significant for higher bandwidths.[25]

Specifications in Columns (3) and (5) follow a parametric approach by using higher order polynomials of the forcing variable while using all observations in the sample. Table 4.2 reports the polynomial order. The order is chosen optimally using Akaike's criterion as suggested in Black et al. (2007). Panel A Column (3) presents the estimate of β in Equation 4.5 using a cubic polynomial in polling center distance to the boundary while Column (5) uses a quadratic polynomial in latitude and longitude. Similar to the results from the local linear regression specification, the share of fraudulent votes drops by about 6 to 8 percentage points depending on the specification. In relative terms, this translates into a 34 to 43 percent drop in fraud share levels from a baseline

[22]For simplicity, all subsequent specifications use neighborhood-level clustered standard errors instead of Conley (1999) standard errors given that there is no substantial difference in the magnitudes from either estimate. Neighborhoods around the boundary are defined as the group of polling centers that are closest to a specific boundary point $\mathbf{b}_i$. Section 4.3.3 provides a more detailed description

[23]Results presented in the remaining columns are discussed in Section 4.4.

[24]Recall from Section 4.2.1 that the bandwidth is chosen optimally as in Imbens and Kalyanaraman (2012)

[25]Section 4.4 provides a more detailed description of Figure 4.6 in a bandwidth sensitivity analysis setting.

average of about 18.3 in the non-coverage side.

Panel B Columns (1)-(3), and (5) document the estimation results for centers in the Northwest. Consistent with the results from the graphical analysis, centers in this region do not exhibit economically and statistically significant drops in fraud at the boundary. In fact, magnitudes are close to zero across all specifications. This is primarily driven by the low levels of fraud experienced in Northwest provinces. Notice that the average share of fraudulent votes for non-coverage centers within the bandwidth is slightly higher than 1 percent and less than 2 percent for the entire sample.

Table 4.2: Average Effect of Mobile Coverage on Category C+ Fraud (Scalar RD)

	LLR-Sharp RD		Polynomial in Distance		Polynomial in Lat-Lon	
	Optimal Bandwidth (1)	Wide Bandwidth (2)	Sharp RD (3)	Shifted Boundary (4)	Sharp RD (5)	Shifted Boundary (6)
Panel A. Eastern and Southern region						
Inside coverage	-0.052	-0.059*	-0.081**	-0.091*	-0.063**	-0.090**
	(0.036)	(0.035)	(0.036)	(0.052)	(0.029)	(0.038)
Observations	601	615	1,256	1,256	1,256	1,256
Mean outside coverage	0.198	0.196	0.183	0.183	0.183	0.183
Polynomial order	.	.	3	3	2	3
Bandwidth (km)	5.76	6.10	.	.	.	.
Neighborhoods	96	96	101	101	101	101
Panel B. Northern and Western region						
Inside coverage	0.013	0.009	-0.001	-0.002	0.001	0.000
	(0.014)	(0.011)	(0.012)	(0.013)	(0.012)	(0.014)
Observations	404	505	1,085	1,085	1,085	1,085
Mean outside coverage	0.013	0.012	0.018	0.018	0.018	0.018
Polynomial order	.	.	1	1	1	1
Bandwidth (km)	4.15	6.10	.	.	.	.
Neighborhoods	133	138	139	139	139	139

Notes: Optimal bandwidth chosen as in Imbens and Kalyanaraman (2012). LLR stands for Local Linear Regression. Wide bandwidth refers to the minimum bandwidth for which the coefficient estimate is statistically significant. Polynomial order determined using Akaike's criterion as suggested in Black et al. (2007). All specifications use neighborhood fixed effects and standard errors clustered at the neighborhood level. Refer to section 4.3.2 for a description of how boundary neighborhoods are created. Refer to section 4.3.2 for a description of how the Southeast and Northwest regions are defined. Columns (4) and (6) replicate the RD design in Columns (3) and (5), respectively, expanding the coverage boundary by 2 kilometers into the non-coverage side. ***, **, * indicate 10, 5, and 1 percent significance, respectively.

4.3.3 Boundary RD

To better assess the degree of spatial heterogeneity in the impact of cell phone coverage on electoral fraud, I estimate conditional treatment effects at various points along the coverage boundary using Equations 4.2 and 4.3. As suggested in Imbens and Zajonc (2011), I choose a random number of boundary points $\mathbf{b}_i$ that cover the entire boundary reasonably well. The points have a minimum distance of 50 kilometers between each other. This results in a total of 1,437 boundary points.[26] All specifications of equation 4.2 include boundary neighborhood fixed effects and standard errors clustered by neighborhood in order to account for spatial correlation of the error terms within neighborhoods. Neighborhoods are determined by first calculating the Euclidean distance between polling centers and boundary points $\mathbf{b}_i$. Centers that are closest to a given boundary point and within the specified bandwidth around the boundary define a neighborhood. To assess the statistical significance of the estimated conditional effects, I calculate the standard errors of the estimates of $\tau(\mathbf{b}_i)$ using the delta method described in Greene (2003). Lastly, as mentioned in the beginning of the section, all analysis is done separately by region and using the share of votes in Category C+ fraud as the primary outcome variable.

Figure 4.4 presents the estimated conditional treatment effects on a map of Afghanistan. The analysis uses bandwidths of 5.76 and 4.15 kilometers for the Southeastern and Northwestern areas, respectively. Shaded areas represent cell phone coverage. Dots indicate the location of the boundary points $\mathbf{b}_i$. The colors of the dots, presented in a mono-chromatic scale, give the magnitude of the estimated effects. Refer to the legend for specific cutoffs. Statistically significant effects are highlighted with hollow circles representing the one, five, and ten percent significance thresholds of the estimated p-values.

Similar to the results from the one-dimensional design, there is clear evidence that the share of fraudulent votes drops significantly at the coverage boundary. The magnitudes of the effects, however, are highly heterogenous both across and within regions of Afghanistan. Note that most

[26] I select the points using the *Create random points* tool in ArcGIS with a minimum allowed distance of 50 kilometers.

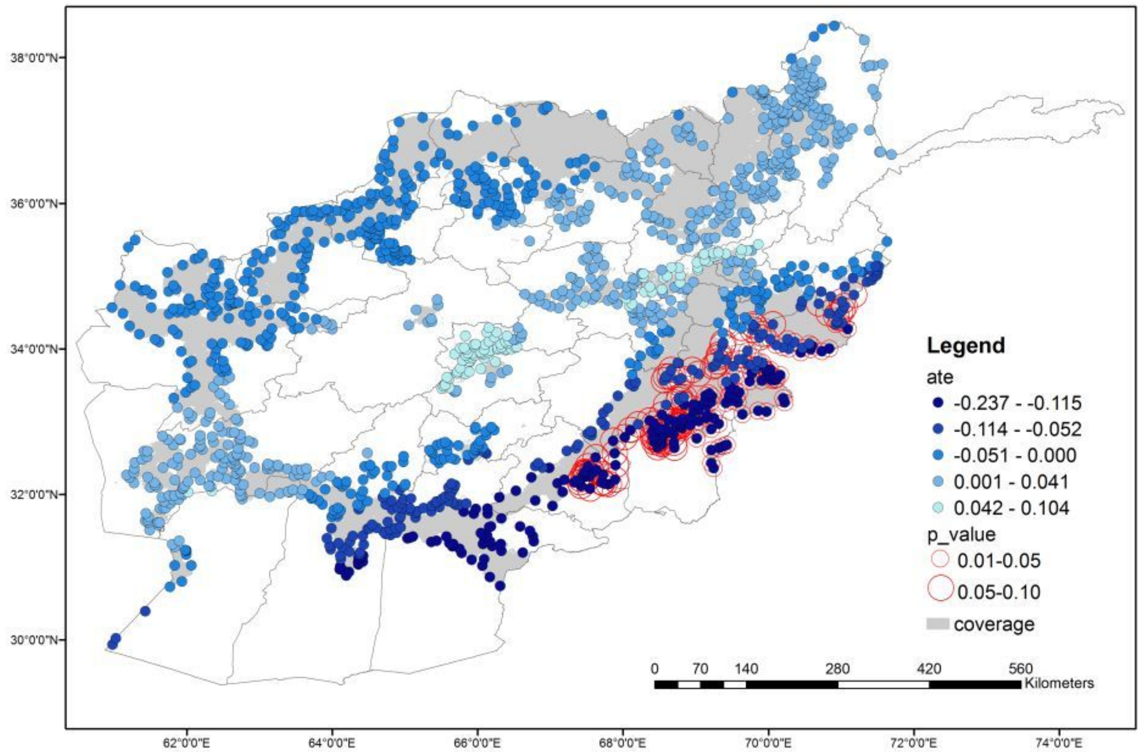


Figure 4.4: Spatial Distribution of Boundary Treatment Effects (Category C+ fraud)

Notes: Shaded areas represent cell phone coverage. Dots indicate the location of $\mathbf{b}_i$ evaluated in equation (4.3). The color of the dots represents the magnitude of the estimated effect. The effects are estimated using equation (4.3). Estimates include neighborhood fixed effects. Hollow circles of different size around the dots represent the p-values of the estimated effects. Standard errors are clustered by neighborhood. Refer to the legend for specific values.

of the economically significant effects are accumulated in the Eastern part of the country. The magnitude of the drop in fraud due to coverage in this area ranges between 11 to 24 percentage points in the eastern portion of the boundary. Similarly, most of the statistically significant effects appear in this area. Conditional treatment effects for other portions of the boundary within this region, although lower in magnitude, exhibit a negative sign. In all, about 76 percent (490 out of 642) of the boundary points evaluated in this area indicate a drop in fraud levels for centers within the coverage area relative to centers outside. Although some boundary points indicate a positive sign in the effect (and hence an increase in fraud due to coverage), none are statistically significant and show an average magnitude that is almost half the average of the conditional treatment effects with negative signs. Specifically, the average of the negative conditional treatment effects is about 9.2 percentage points whereas the average magnitude for positive conditional treatment effects is about 4.8 percentage points.[27]

The estimated effects for the Northwestern region offer a contrasting result. Although nearly half of the estimated conditional treatment effects are negative, their magnitudes are economically and statistically insignificant. This is not surprising, however, if one considers that the extent of fraud in these region is low relative to the Southeastern region. On average, less than two percent of the votes per center classify as potentially fraudulent (compared to almost 20 percent in the Southeastern region). Hence any estimates of differences in fraud across the boundary should indeed be relatively small.

Table 4.3 presents the averaged conditional treatment effects estimated from Equation 4.4. I estimate standard errors via bootstrap with 500 replications and resampling within districts. For reference, Table 4.3 presents the mean level of category C+ fraud for centers within the specified bandwidth and outside the coverage boundary. To facilitate comparisons between the one-dimensional and boundary RD results, Columns (3) and (4) include the average effects estimated from the one-dimensional design presented in Table 4.2. Average fraud levels drop by about 5.7 percentage points in the Southeastern region from a baseline average in non-coverage centers of about 19 percent. Similar to the one-dimensional RD results, centers in the Northwest region do not exhibit significant differences in fraud levels across the boundary. More importantly, note the results are robust to the estimation design used: average estimates from the one-dimensional and boundary RD designs are close in magnitude and precision.

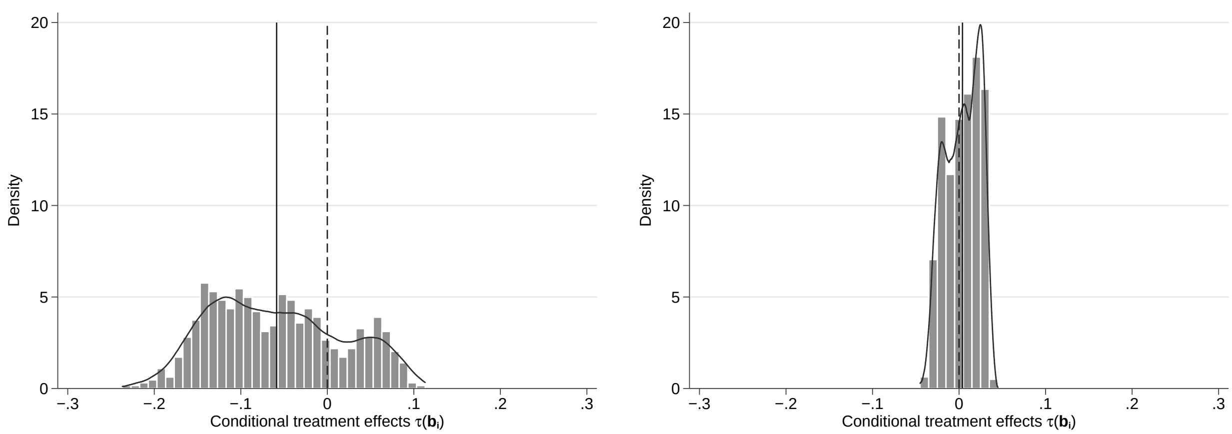
For a depiction of the variability of the effects, Figure 4.5 presents histograms of the estimates $\tau(\mathbf{b}_i)$ by region. The solid vertical line gives the estimated average conditional treatment effects discussed above. Notice that for the Southeastern region (Figure 4.5a) the estimated effects are largely negative but with significant variability. Estimates for the Northwest region, on the other hand, are mostly clustered around zero.

[27] The averages are not shown in the figure. However, refer to the end of this section for a detailed discussion on the averaged conditional treatment effects.

Table 4.3: Averaged Conditional Treatment Effects (Category C+ fraud)

	Boundary RD		Scalar RD	
	Eastern and Southern regions (1)	Western and Northern regions (2)	Eastern and Southern regions (3)	Western and Northern regions (4)
Inside coverage	-0.057**	0.003	-0.059*	0.009
	(0.029)	(0.012)	(0.035)	(0.011)
Boundary points	640	795	.	.
Mean outside coverage	0.198	0.013	0.196	0.012
Baseline regression characteristics				
Observations	601	404	615	505
Bandwidth (km)	5.76	4.15	6.10	6.10
Neighborhoods	96	133	96	138

Notes: Bootstrapped standard errors in parenthesis. Standard errors determined using 500 replications and re-sampling within districts. *Boundary points* refers to the total $\mathbf{b}_i$ evaluated within the specified region. Columns (3) and (4) use the wide bandwidth specification from Table 4.2. ***, **, * give 10, 5, and 1 percent significance, respectively.



(a) East/South

(b) North/West

Figure 4.5: Distribution of Boundary Treatment Effects (Category C+ fraud)

Notes: Bin width of 0.01. The solid vertical line gives the average of the estimated effects. Solid blue line represents the estimated density. The density estimate uses an Epanechnikov kernel function with bandwidth of 0.025.

4.4 Robustness Checks

This section assesses the robustness of the results to the possibility of non-compliers along the boundary, the inclusion of covariates in the main specification, and the choice of bandwidth and polynomial order in the local linear regression and parametric specifications, respectively.

4.4.1 Boundary Sensitivity

Under-smoothing of the RD plots in Figure 4.3 using narrower 2-kilometer distance bins reveals that centers falling within 2 kilometers of the boundary in the non-coverage side have, on average, similar fraud levels as their covered counterparts.[28] However, fraud levels jump sharply for centers that are further than 2 kilometers away from the boundary on the non-coverage side. This results from the cost of reporting fraud not being as sharp when an uncovered center is only within 2 kilometers of the boundary. In terms of the RD setting, this implies the existence of non-compliers near the boundary on the non-coverage side and hence a potential downward bias on the RD estimates presented in Columns (1), (3), and (5) of Table 4.2.

To examine this issue, Columns (4) and (6) of Table 4.2 replicate the RD analysis in Columns (3) and (5) shifting the boundary by 2 kilometers into the non-coverage side. Therefore, this exercise considers centers that are in close proximity to the boundary but on the non-coverage side as covered. Consistent with the evidence from the RD plots, the magnitude of the effect is about 1 to 3 percentage points higher using the shifted boundary in the Southeast region while not different from zero in the Northwest region.[29] This signals that the observed coverage boundary may not be sharp as individuals very close to the boundary on the non-coverage side may be able to report fraud (i.e., non-compliers). Although a Fuzzy RD design would be more appropriate in this setting, the lack of data on the exact location where fraud reports were made does not permit me to follow such strategy.[30] With this in mind, results using the observed coverage boundary as the main determinant for treatment assignment should be interpreted as an intent-to-treat effect.

---

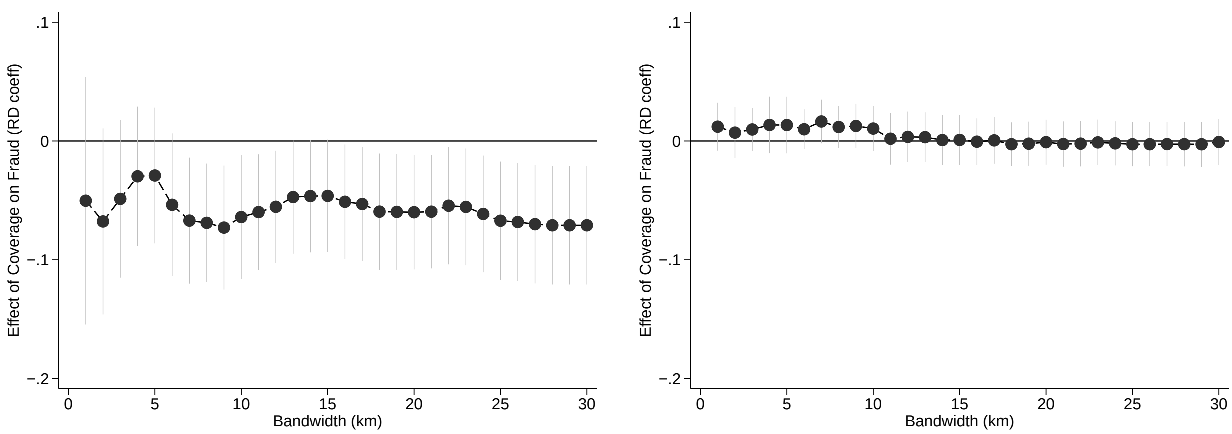
[28]RD plots using 2-kilometer distance bins are not shown but can be provided upon request

[29]The percentage point difference refers to 8.1 versus 9.1 points in Columns (3) and (4) and 6.3 versus 9.0 in Columns (5) and (6)

[30]A Fuzzy RD design would entail estimating a joint model where the probability that a center is treated (i.e., individuals are able to report from such center) is modeled as a function of the observed coverage status.

4.4.2 Bandwidth Choice and Polynomial Order

Figure 4.6 assesses the sensitivity of the Local Linear Regression results presented in Table 4.2 to the choice of bandwidth. Dots indicate the coefficient estimate using the bandwidth specified on the horizontal axis, while the vertical spikes give the 95 percent confidence interval of the estimate. Both regions show estimates that are robust to bandwidth choice. For polling centers in the Southeast, the share of fraudulent votes is consistently lower for centers in coverage areas regardless of the bandwidth chosen. Statistical significance also remains robust. Similarly, the magnitude of the effect for centers in the Northwest is non-distinguishable from zero for any bandwidth.



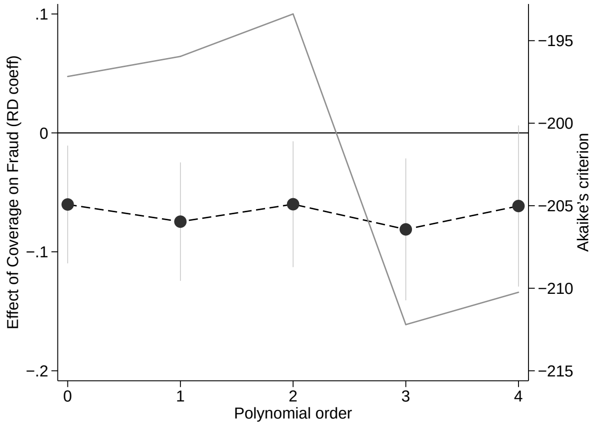
(a) Eastern and Southern region

(b) Northern and Western region

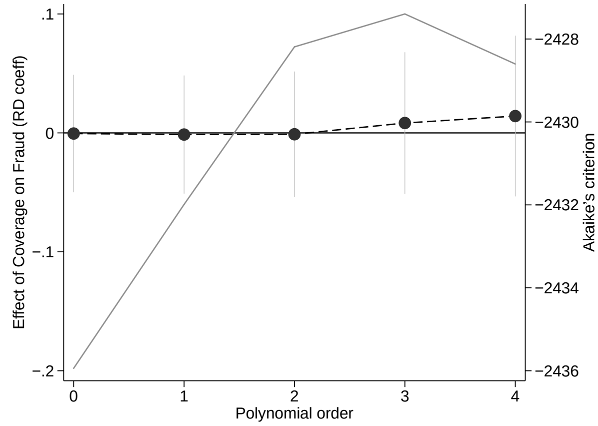
Figure 4.6: Sensitivity of Results to Bandwidth Choice (Category C+ fraud)

Notes: Each dot indicates the RD estimate using the specified bandwidth. Range spikes indicate 95% confidence intervals of the estimates.

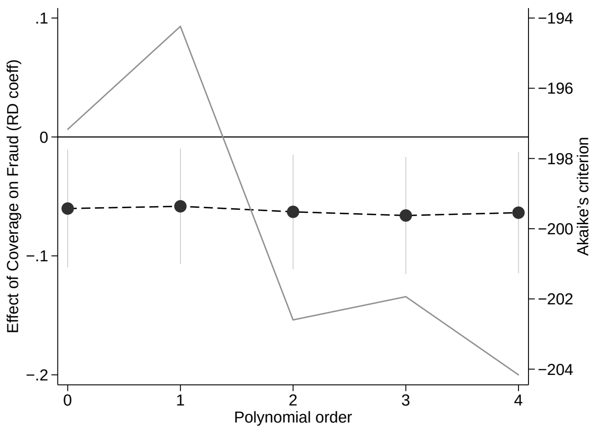
Figure 4.7 examines the sensitivity of results in Columns (3) and (5) of Table 4.2 to the polynomial order used. Figures 4.7a and 4.7b report the coefficient estimates and confidence intervals for the Southeast region while Figures 4.7c and 4.7d show results for the Northwest region. For reference, Figure 4.7 also reports the Akaike's criterion used to determine the optimal order in Table 4.2. Similar to the bandwidth exercise, the choice of polynomial order does not significantly affect the magnitude and statistical significance of the effects in either region. Average effects are negative in magnitude and statistically significant up to a cubic specification for the Southeast region while not significantly different from zero for the Northwest region.



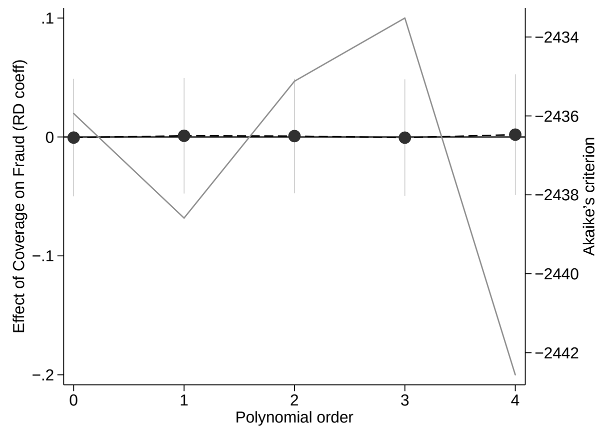
(a) Distance polynomial (SE region)



(b) Distance polynomial (NW region)



(c) Lat/Lon polynomial (SE region)



(d) Lat/Lon polynomial (NW region)

Figure 4.7: Sensitivity of Results to Polynomial Order (Category C+ fraud)

Notes: Each dot indicates the RD estimate using the specified order in the RD polynomial. Range spikes indicate 95% confidence intervals of the estimates. Gray line indicates the value of Akaike's criterion for each model (graphed on second axis)

4.4.3 Sensitivity of Results to Inclusion of Covariates

I explore the sensitivity of the results to the inclusion of baseline covariates in the one-dimensional specification. In a valid RD design, baseline covariates should be locally balanced at the boundary, thus their inclusion should not affect the consistency of the estimated causal impact. Furthermore, while baseline covariates are not needed for identification, they improve the precision of the estimates (e.g., Lee (2008), Imbens and Lemieux (2008)). Table 4.4 replicates the sharp RD results presented in Table 4.2 while adding a set of baseline covariates. All specifications include polling center level characteristics that capture voting outcomes, geographic characteristics, and economic development characteristics of the polling center. Specifically, the included covariates are: total polling stations, share of female stations, elevation, slope, distance to primary road and basic health clinic. For reference, Table 4.4 also reports the estimated effects without including covariates. Notice that, consistent with a valid RD design, the magnitude of the estimates does not change substantially while the precision improves to a small degree.

Table 4.4: Sensitivity of Results to the Addition of Baseline Covariates

	Local Linear Regression		Polynomial in Distance		Polynomial in Lat-Lon	
	No Controls (1)	Controls (2)	No Controls (3)	Controls (4)	No Controls (6)	Controls (7)
Panel A. Eastern and Southern region						
Inside coverage	-0.052	-0.048	-0.081**	-0.084**	-0.063**	-0.044
	(0.036)	(0.035)	(0.036)	(0.035)	(0.029)	(0.030)
Observations	601	615	1,256	1,256	1,256	1,256
Polynomial order	.	.	3	3	2	2
Bandwidth (km)	5.76	5.76	.	.	.	.
Neighborhoods	96	96	101	101	101	101
Panel B. Northern and Western region						
Inside coverage	0.013	0.016	-0.001	0.002	0.001	-0.001
	(0.014)	(0.014)	(0.012)	(0.013)	(0.012)	(0.013)
Observations	404	404	1,085	1,085	1,085	1,085
Polynomial order	.	.	1	1	1	1
Bandwidth (km)	4.15	4.15	.	.	.	.
Neighborhoods	133	133	139	139	139	139

Notes: Columns labeled "Controls" include a set of covariates in addition to the baseline specification. Baseline specification labeled "No Controls" and also reported in Table 4.2. The included covariates are: total polling stations, share of female stations, elevation, slope, distance to primary road and basic health clinic. Optimal bandwidth used in baseline regression obtained via Imbens and Kalyanaraman (2012). ***, **, * give 10, 5, and 1 percent significance respectively.

CHAPTER 5

ALTERNATIVE CHANNELS OF FRAUD

Social monitoring may be one among several mechanisms that affect fraud through coverage. Therefore estimates of $\tau(\mathbf{b}_i)$ may not solely estimate the social monitoring effect as other components in Equation 3.5 might equally lead to increases in fraud levels across the coverage boundary. Although several channels may be relevant, to provide an analysis that is both parsimonious and enlightening, this section explores two channels that based on the theoretical framework are important determinants of fraud and can thus confound the social monitoring effect. These are: election-related violence (δ and P in equation 3.5) and voter affinity (a in Equation 3.5). I isolate the social monitoring effect by testing whether these channels change discretely at points in the coverage boundary where I observe significant changes in fraud as well.[1]

5.1 Election-related Insurgent Violence

Findings from the literature on conflict and violence suggest a strong relationship between political violence and both, cell phone coverage (e.g., Shapiro and Weidmann (2013), Pierskalla and Hollenbach (2013)) and electoral fraud (e.g., Collier and Vicente (2012), Callen and Weidmann (2013)). Cell phone coverage may lead to surges in violence as insurgents can better coordinate attacks (e.g., Cordesman (2005), Strother (2007)). In contrast, collective action by citizens and cell phone tracking by counterinsurgency agencies might undermine terrorists' actions.

In the case of Afghanistan, political violence is a potentially important channel for two reasons: first, during the pre-election period the Taliban issued several warnings targeting polling centers and voters (Gall (2009), Filkins (2009)). This was followed by a sharp surge in violence on election

[1] Note from equation 3.5 that fines and penalties F associated with electoral fraud are also a key determinant of fraud, however, there is no evidence that penalty schedules changed with cell phone coverage.

day as depicted in Figure 5.2 which plots the number of daily attacks for the year 2009.[2] Secondly, media reports and anecdotal evidence suggest that the Taliban have a strong aversion to cell phone coverage and cell phone technology in general. For example, throughout the eastern and southern regions of the country, they have forced cell phone companies to regularly turn off their antennas at dusk to prevent villagers from informing coalition forces of their movements (**?**). Attacks to damage and destroy cell phone towers are also well documented (e.g., Lakshmanan (2010), Robinson (2013)).

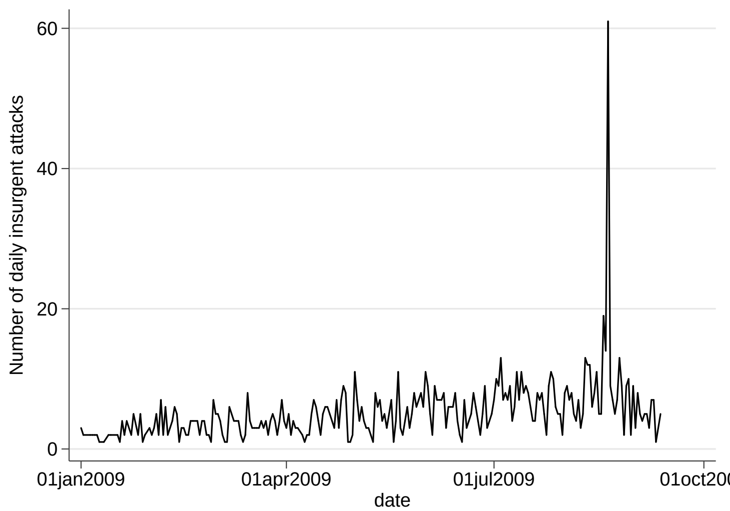


Figure 5.1: Insurgent Attacks on Election Day

Notes: Peak gives the number of insurgent attacks the on election day, August 20, 2009. Data collected by the Worldwide Incident Tracking System (WITS 2009) and obtained from Empirical Studies of Conflict (ESOC) project at Princeton University.

In terms of the theoretical model, if the candidate expects a drop in violence due to coverage (i.e., a drop in δ or P due to the Taliban preferring to operate in areas without coverage) then the price of legal votes $p_l = f(\delta, P, a)$ in covered centers drops relative to fraudulent votes which then leads the candidate to substitute fraudulent votes for legal votes. With this in mind, if violence declines with coverage then violence along with social monitoring might explain the effects of

[2]Data on attacks were compiled by the Worldwide Incident Tracking System (WITS 2009). The Worldwide Incident Tracking System (WITS 2009) is maintained by the National Counterterrorism Center and accessed via the Empirical Studies of Conflict (ESOC) project. Available at: http://www.nctc.gov/site/other/wits.html. A more detailed description of this dataset is provided later in this section.

coverage on fraud documented in section 4.3.

To assess whether violence changes significantly with coverage, I estimate conditional treatment effects using Equation 4.2 for various outcomes on insurgent violence using the same boundary points $\mathbf{b}_i$ used in the fraud results. The objective is to assess whether violence changes discontinuously at boundary points where fraud equally jumps. The analysis uses data on the location of villages and type of insurgent attack for the year 2009 and up to election day. Data on village location is obtained from the Measuring Impacts of Stabilization Initiatives project (MISTI 2013) while data on insurgent attacks, IED incidence, and civilian and military casualties is compiled by the Worldwide Incident Tracking System (WITS 2009) and the Empirical Studies of Conflict (ESOC) project.[3]

I define the violence outcome variables as follows: *Civilian* and *All Casualties* refer to the rate of civilian and combined casualties (i.e., civilian and military) within a 5-kilometer radius of the village while *Insurgent attack* and *IED* are indicators for whether there was an attack or an IED within a 5-kilometer radius of the village.[4] The definitions use data up to the election day. The purpose is to capture how individuals formed expectations on where violence would occur prior and up to the election day.[5] I am able to estimate conditional treatment effects for the Southeast region only where the majority of the attacks took place.[6]

Figure 5.2 presents the estimated conditional treatment effects for the four violence outcomes. Dots represent the boundary points $\mathbf{b}_i$ at which the treatment effect is evaluated. Color scale gives the magnitude of the estimated effects and statistically significant effects are highlighted in red. First, note in Figure 5.2a that there is substantial spatial variation in the effect of coverage on the

[3]IED refers to an Improvised Explosive Device.

[4]The rate for the civilian and combined casualties is per 10,000 inhabitants in the province where the attack took place. The rate could not be calculated over the population within the 5-kilometer radius because of lack of complete data on population at the village level. The choice of 5 kilometers comes from the average distance between villages and polling centers observed in the sample.

[5]Recall from section 3 that δ is a voter's assessed likelihood of an attack on the day of the election.

[6]This is due to the low levels of violence in the Northern and Western regions relative to the Southern and Eastern regions. Refer to Figure C.4 for the spatial distribution of the attacks.

likelihood of an insurgent attack. The signs of most conditional treatment effects on the eastern portion of the boundary are negative, suggesting a drop in attacks while the positive signs on the western side of the boundary indicate an increase in attacks as a result of coverage. However, about 55 percent of the boundary points report a negative coefficient suggesting a drop in the likelihood of an attack as a result of coverage. This is consistent with the evidence on the Taliban's aversion towards cell phone coverage mentioned previously. However, on aggregate terms, the magnitude of the effect is negligible and statistically insignificant. Specifically, the averaged conditional effect indicates about a 0.3 percent drop in the likelihood of an attack as a result of coverage.

Figure 5.2b presents the conditional treatment effects of coverage on the likelihood of an IED attack. Consistent with the evidence on the positive relationship between cell phone coverage and IEDs, all of the estimated conditional treatment effects are positive.[7] In terms of magnitude, cell phone coverage leads to a modest 2 to 3 percent increase in the likelihood of an IED attack with an averaged conditional treatment effect of about 2 percent. Notice that statistically significant effects are limited to a cluster of points on the upper section of the coverage boundary. Figures 5.2c and 5.2d report the effect of coverage on the civilian and combined death rate. The effects are consistently positive, although trivial in magnitude and statistically insignificant, throughout the boundary. This result is consistent with the high lethality of IEDs–which also increase with coverage–relative to other types of attacks.[8]

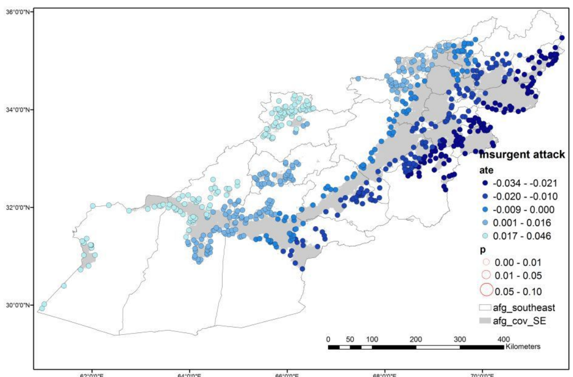
The findings in this section show that the conditional treatment effects of coverage on violence outcomes are generally small in magnitude and statistically insignificant. This has key implications for the identification of the social monitoring effect since drops in fraud at the boundary described in section 4.3.3 cannot be explained by significant drops in violence outcomes. In fact, the likelihood of an IED attack, and the rate of casualties increase with coverage.[9] In the case of insurgent

[7] See Shapiro and Weidmann (2013) for a description on the extensive use of cell phones as triggering devices for IEDs.

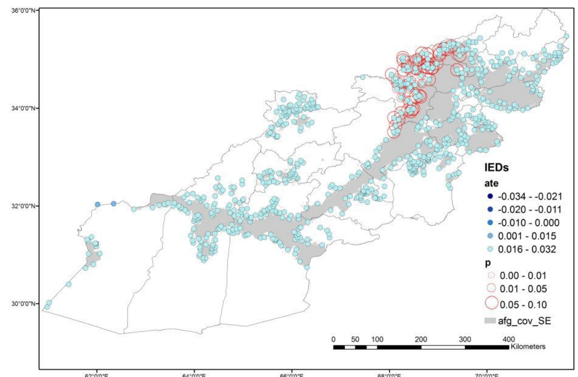
[8] The rate of civilian and combined casualties due to IEDs is more than 10 times higher than for other types of attacks.

[9] In terms of the theoretical framework, this implies that legal votes become more expensive with coverage (i.e., voters are less likely to vote if there is a higher likelihood that the center will be attacked) and hence fraud becomes a

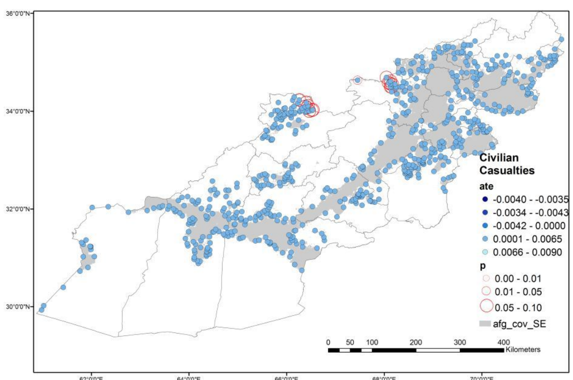
attacks, although there is evidence of a decline with coverage and hence a potential confounder of the social monitoring effect, the change is smooth and statistically indistinguishable from zero.



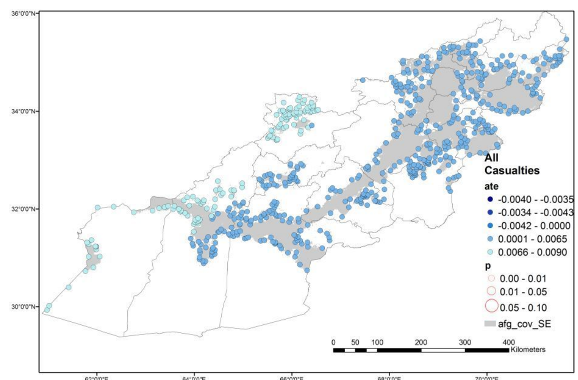
(a) Insurgent attack within 5km of village



(b) IEDs within 5km of village



(c) Civilian Casualties within 5km of village



(d) All Casualties within 5km of village

Figure 5.2: Spatial Distribution of Boundary Treatment Effects for Various Indicators of Violence
Notes: Shaded areas represent cell phone coverage. Dots indicate the location of $\mathbf{b}_i$ evaluated in equation (4.3) using the specified violence outcome. Monochromatic scale gives the magnitude of the estimated effect. The effects are estimated using equation (4.3). Estimates include neighborhood fixed effects. Red circles give the p-values of the estimated effects. Standard errors are clustered by neighborhood. Refer to legend for specific values.

5.2 Voter's Affinity

From the theoretical framework in section 3, note that if parameter a (i.e., a voter's affinity towards a candidate) significantly jumps with coverage, the price of legal votes $p_l = (\delta, P, a)$ in covered centers drops relative to fraudulent votes which then leads the candidate to substitute fraudulent votes for legal votes. Such case yields negative conditional treatment effects and thus the inability to disentangle the social monitoring effect from a voter's affinity effect.

---

more attractive choice for the candidate given the drop in the relative price of fraudulent votes.

Given the importance of tribal loyalty in Afghan society, ethnic and tribal identity are strong predictors of voter's affinity.[10] For instance, individuals with the same tribal affiliation as a candidate may be more willing to vote in spite of violence and exhibit a lower dislike towards fraudulent actions by this candidate. In this context, the social monitoring effect may be confounded if the tribal composition of villages changes sharply with coverage. This may result, for instance, if cell phone providers give preference to certain ethnic groups by expanding coverage into their locations. In such cases, changes in coverage may coincide with sharp changes in tribal composition and hence, voter's affinity towards candidates of specific ethnic backgrounds.

To examine the spatial distribution of ethnic groups and tribes relative to the coverage boundary, I georeference detailed tribal maps of the Southeastern region of Afghanistan collected by the Culture and Conflicts Studies program at the Naval Postgraduate School.[11] Georeferenced maps are then combined with village coordinate data from the Measuring Impacts of Stabilization Initiatives project (MISTI 2013) to construct village-level indicators of primary tribe for almost 18,000 villages in the Southeastern region of Afghanistan.[12] I aggregate the more than 50 tribes represented in the sample into eight tribal confederations using the Culture and Conflicts Studies program's definitions.[13] Confederations are typically formed by groups of tribes with common origin or historical alliances. Figure 5.3a presents the spatial distribution of each village's primary tribal confederation in Southeastern Afghanistan. Further, refer to Appendix Table C.3 for the list of tribes and confederations.

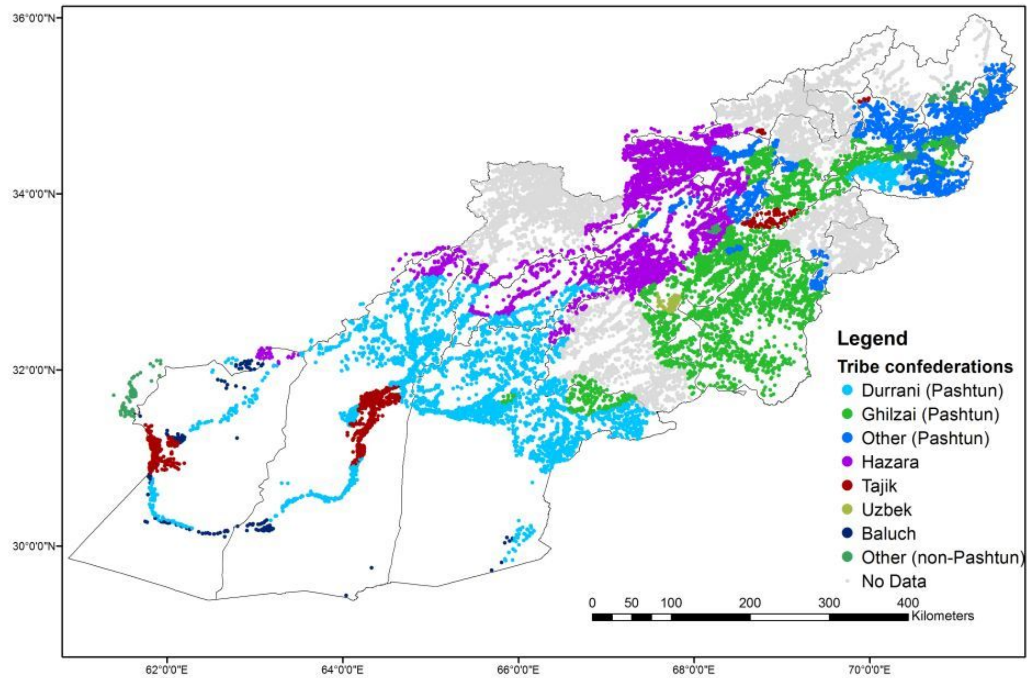
Figure 5.3 overlays the tribal structure data on the coverage maps. A detailed inspection suggests that potential issues may arise with the spatial distribution of Ghilzai Pashtun villages relative to the coverage boundary. Note that on parts of the western portion of the boundary the tribal composition of villages changes sharply from Ghilzai Pashtun to Hazara. Similarly, in the upper

[10] See (Tarzi and Lamb 2011) for a discussion of this topic related to the Pashtun ethnic group.

[11] Refer to Appendix Figure C.6 for an example of a tribal map for Kandahar province.

[12] Data on tribal composition by the Culture and Conflicts Studies program is only available for this region.

[13] The Culture and Conflicts Studies program tribal confederation definitions are, in turn, based on Tribal Hierarchy and Dictionary of Afghanistan (2007)



(a) Tribal confederations by village

Figure 5.3: Major Tribal Confederations, Southeast Afghanistan

Notes: Tribal confederations obtained from Conflict Studies program. Shaded areas represent availability of 2G GSM cell phone coverage for two largest cell phone providers in Afghanistan for the year 2009. Dots give the location of villages (MISTI 2013). Lines demarcate the provinces of Afghanistan.

portions of the boundary there appears to be significant changes in the distribution of Tajik villages relative to Ghilzai and other Pashtun villages. This is particularly concerning since candidate Hamid Karzai is an ethnic Pashtun and thus affinity may change discontinuously in these portions of the boundary. Therefore the analysis in this section focuses on the eastern area encompassing the mostly Ghilzai Pashtun villages.

To explore the possibility of jumps in the composition of Pashtun, and particularly Ghilzai Pashtun villages at the coverage boundary, I replicate the spatial RD design using an indicator for whether a village is majority Ghilzai Pashtun as the outcome of interest. Figure 5.4 presents the conditional treatment effects. Note that, although there is evidence of changes in tribal structure at some points along the coverage boundary, these changes cannot explain the observed drops in fraud as there is no substantial overlap in the boundary points where tribal affiliation and fraud (compare to Figure 4.4) change sharply.

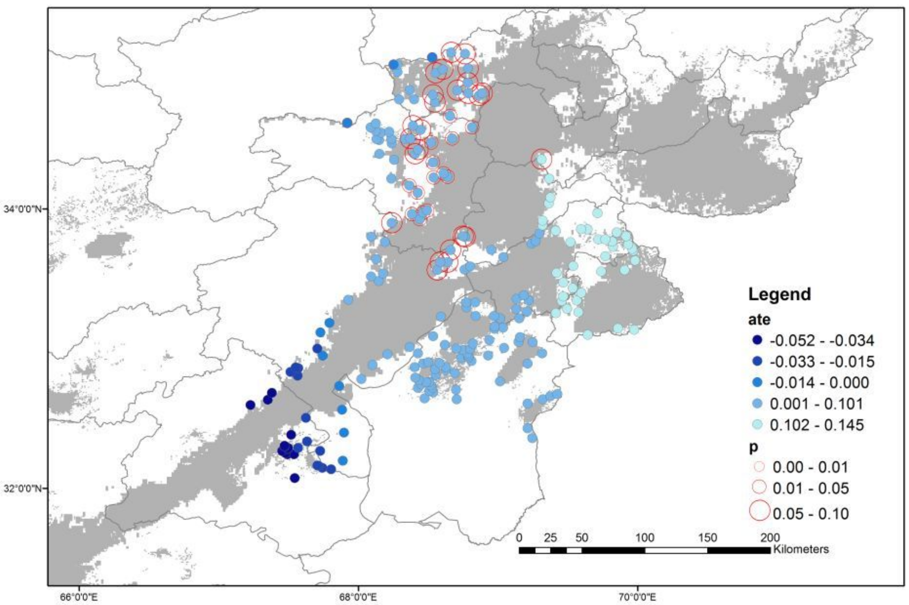


Figure 5.4: Spatial Distribution of Boundary Treatment Effects: Ghilzai Pashtun village

Notes: Shaded areas represent cell phone coverage. Dots indicate the location of $\mathbf{b}_i$ evaluated in equation (4.3) using whether a village is primarily Ghilzai Pashtun as the outcome. Monochromatic scale gives the magnitude of the estimated effect. The effects are estimated using equation (4.3). Estimates include neighborhood fixed effects. Red circles give the p-values of the estimated effects. Standard errors are clustered by neighborhood. Refer to legend for specific values.

CHAPTER 6

CONCLUSION AND DISCUSSION

The results in this paper provide considerable evidence that cell phone coverage–via citizen reporting–lowers electoral fraud. Specifically, up to 76 percent of the estimated conditional treatment effects along the coverage boundary are negative, implying a drop in fraud for centers just inside coverage areas. It is important to highlight that the estimated effects exhibit a considerable degree of spatial heterogeneity: average impacts in the South and East are economically and statistically significant while non-distinguishable from zero in other parts of the country. These results are potentially explained by a spatial pattern of election-related violence that strongly mimics the observed pattern of fraud.

Given the importance of political violence during the 2009 Afghan presidential election, I test empirically whether this metric may be a potential confounder of the social monitoring effect. With this in mind, a second set of results in this paper show that there is no clear evidence that changes in insurgent violence at the coverage boundary may explain the observed drops in fraud. The last set of results test whether changes at the coverage boundary in voter's affinity towards a candidate is the primary channel explaining the observed changes in fraud. Detailed analysis on changes in villages' tribal affiliation–a strong predictor of voter's affinity–shows evidence of changes in tribal structure at some points along the coverage boundary. However, these changes cannot explain the observed drops in fraud as there is no overlap in portions of the coverage boundary where tribal affiliation and fraud change sharply. Overall, the absence of significant changes at the boundary in both election-related violence and the tribal composition of villages suggests that citizen-based monitoring, and not these alternative channels, explains the observed drops in fraud.

From a policy perspective, this paper illustrates that the availability and expansion of cell phone usage along with citizen-based monitoring initiatives, can have positive externalities on institutional development via fraud deterrence and the mitigation of corrupt behavior in general.

APPENDIX A

THE AUDIT AND RECOUNT PROCESS

In its final report of the 2009 election (Electoral Complaints Commission 2010), the ECC reported that after receiving an increasing number of complaints on ballot stuffing and other irregularities, they decided on September 8, 2009 to conduct an audit of polling stations nationwide. To this end, they ordered the IEC to conduct an audit and recount of stations satisfying the following criteria:

- A1: Stations in which 600 or more valid votes were cast
- B1: Stations with more than 100 votes in which one candidate received 95 percent or more of the total votes cast
- C1: Stations satisfying both A1 and B1

After the initial samples were drawn, however, three additional categories were created due to a misunderstanding of the ECC orders by the IEC[1]. The three new categories expanded the scope of the audit. The categories were:

- A2: Stations with 600 or more votes cast (Excluding those in A1)
- B2: Stations in which a candidate received 95 percent or more of the total valid votes cast (Excluding those in B1)
- C2: Stations satisfying both A2 and B2

After the audit process, the IEC reported that 3,376 stations classified in at least one of these categories. Out of this sample, the ECC and IEC investigated 10 percent of the ballots within each category. Some of the physical indicators used to determine fraud were whether the ballot box was tampered, all required materials were included, visual inspection of the ballots, reviews of the tally results and the actual ballot counts, among others. For the purpose of this study I aggregate the six categories described above into three broader categories:

[1]The misunderstanding was mainly due to the definition used to classify votes as "valid".

- Category A: Stations with 600 or more votes cast. Defined as A1+A2+C1+C2 from the categories above.

- Category B: Stations in which one candidate received 95 percent or more of the total votes cast. Defined as B1+B2+C1+C2

- Category C: Stations satisfying Categories A and B above. Defined as C1+C2.

The number of polling stations within each category is 1,706 in category A (545 from A1 + 299 from A2 + 741 from C1 + 121 from C2), 2,532 in category B (1269 from B1 + 401 from B2 + 741 from C1 + 121 from C2), and 862 in category C (741 from C1 + 121 from C2).

APPENDIX B

MODEL EXTENSIONS

B.1 Nonlinear Reporting Cost Function

Since reporting fraud via cell phones increases in cost as polling centers are farther away from coverage areas, I consider the case where this function is nonlinear. Specifically, assume reporting fraud carries a physical cost $c(d)$ where d defines how accessible the medium (cell phones) used to report fraud is. In the spatial context of this study d can be interpreted as the shortest distance from the polling center where fraud takes place to a geographic boundary that defines whether there is cell phone coverage. Using this interpretation and assuming that $d < 0$ (i.e., a negative distance) arbitrarily defines a polling center located in the non-coverage side of the boundary one can specify a function $D = \mathbb{1}\{d \geq 0\}$ that indicates coverage such that the cost of reporting fraud is given by:

$$c(d) = D\underline{c} + (1 - D)h(d) \tag{B.1}$$

where $\underline{c}$ is the marginal cost of making a call when the center is in the coverage side. I assume this cost is equal for everyone in the coverage side. $h(d)$ is a smooth cost function faced by individuals on the non-coverage side with $h(d) > \underline{c}$ for all $d < 0$ and $h(0) = \bar{c}$ with $\bar{c} > \underline{c}$.[1] Refer to Appendix Figure (C.7) for a sample reporting cost function.

B.2 Free-riding and Fraud Reporting

I extend the *transmitter's problem* presented in section 3.1 by allowing for the possibility of "free-riding" in the reporting process. Free-riding can be a concern in this context if individuals assess that the probability that a fraudulent official is punished conditional on his report is trivial. If that is the case, then the probability of making a report does not change regardless of accessibility to the reporting medium, in this case, coverage status. I show that individuals have an incentive to report fraud, despite the free-riding problem, as long as there is some utility gain from the reporting process itself (i.e., the warm glow parameter specified in section 3.1)

[1] A possible parameterization for $c(d)$ could be $c(d) = \bar{c} \cdot \exp(-\beta d)$

More specifically, assume that since reporting fraud is costly, individual i assesses the likelihood that the center is actually audited (and hence the fraudulent candidate is penalized) as a result of his report.[2] More specifically, let ϕ_1 and ϕ_0, with $\phi_1 \geq \phi_0$ denote i's subjective assessment of the probability that the candidate will be punished given that he reports him and does not report him respectively. When the candidate is punished i gets a net utility value ξ_i that can be interpreted as a utility gain from the fraudulent candidate being punished net of any affinity or benefits that the individual might receive from non-punishment.[3] Additionally, assume as in section 3.1, that the act of reporting fraud gives i a utility gain λ_i.

The individual's net payoffs from reporting fraud are therefore given by $\xi_i + \lambda_i - c(d)$ when the fraudulent candidate is punished and $\lambda_i - c(d)$ otherwise. Lastly, if i decides *not* to report, he simply obtains ξ_i when the candidate is punished and zero otherwise. Assuming linear utility, he will then decide to report fraud if:

$$\begin{aligned} \phi_1[\xi_i + \lambda_i - c(d)] + (1 - \phi_1)[\lambda_i - c(d)] &\geq \phi_0 \xi_i \\ \lambda_i + (\phi_1 - \phi_0)\xi_i &\geq c(d) \end{aligned} \tag{B.2}$$

Notice that even when there is a "free-riding" problem (i.e., i believes that his report does not affect the probability that the candidate is punished ($\phi_1 = \phi_0$) and hence the decision rule above reduces to: Report if $\lambda_i \geq c(d)$) an individual i might still have an incentive to report fraud as he derives utility from doing this alone. Therefore "free-riding" will lower the willingness to report fraud but not eliminate it completely.

B.3 The Voter's Problem

An individual i considers a campaign promise and the possibility of a violent outcome on election day when deciding whether to vote. In the spirit of Dekel, Jackson, and Wolinsky (2008),

---

[2] I assume that once a fraudulent center is audited, the candidate and polling center manager are penalized. Therefore I do not consider any "concealment technology" as in Cremer and Gahvari (1994). In sections (3.2) and (3.3) I describe the penalties faced by the official and candidate respectively

[3] The idea is that the individual might obtain some "justice has been served" satisfaction while at the same time punishment to a candidate of his liking might bring some disutility.

assume each voter is characterized by parameter a_i representing the net utility the individual obtains from simply tendering his vote to the candidate. This can be interpreted as i's affinity towards the candidate. The candidate offers individual i a campaign promise $p_{l,i}$ regardless of the election result but conditional on i tendering the vote to him. Assuming the individual is an expected income maximizer then i's expected payoff from voting is given by $V_i^1 = a_i + p_{l,i}$ while i's expected payoff from not voting is simply $V_i^0 = 0$. Since elections in conflict zones are often characterized by violence, the individual takes into account an exogenous probability δ_j that a violent event takes place at polling center j and as a result receives a negative payoff P. This consideration is particularly important in the Afghan context as the Taliban issued several warnings targeting polling centers and voters on election day (Gall (2009), Filkins (2009)). The individual therefore decides to vote if $\delta(V_i^1 - P) + (1 - \delta)V_i^1 \geq V_i^0$.[4]

Given i's payoffs, the minimum price per legal vote (i.e., the campaign promise) that guarantees i's vote is therefore given by:

$$p_{l,i} = \delta_j P - a_i \tag{B.3}$$

where $p_{l,i} = 0$ if the affinity parameter a_i is sufficiently large as to offset the negative payoff of violence (i.e., $a_i > \delta_j P$).[5]

B.4 Two Candidates

This section introduces the possibility of two candidates in the *voter's problem*. A key distinction from the model presented in section B.3 is that the voter not only has to decide whether to vote but also for whom to vote taking into account each candidate's campaign promise. In terms of the *candidate's problem*, the level of fraud in equilibrium changes slightly when considering a second candidate. In essence, this introduces an additional channel of fraud, namely what I refer to as a *challenger effect*. Broadly speaking, in order for a candidate to entice voters to vote for him, he has to pay a legal price for their vote that matches the highest value between the expected net

---

[4]For simplicity, this specification of the model considers only the problem of whether to vote but not the problem of for whom to vote. Refer to Appendix B for an extension of the "Voter's Problem" that considers two candidates and hence the decision becomes whether to vote and for whom to vote.

[5]Alternatively, one can specify the reservation campaign promise $p_{l,i}$ as equal to the $max\ \{0, \delta_j P - a_i\}$

payoff from violence and the opposing candidates' campaign offer to voters.

The Voter's Problem Suppose voters have to decide between two candidates indexed by k. As in Dekel et al. (2008), assume that each voter i is characterized by parameters U_i^k and a_i^k that represent the utility the individual obtains from k's victory and from simply tendering his vote to k, respectively. Candidate k offers individual i a campaign promise $p_{l,i}^k$ regardless of the election result but conditional on i tendering his vote to k[6]. Letting $\psi_{k|l}$ denote the probability that candidate k wins given that i tendered his vote to candidate $l \in \{X, Y\}$ and letting the individual be an expected income maximizer then i's expected payoff from voting for k is given by:

$$V_i^k = \psi_{k|k}(U_i^k + a_i^k + p_{l,i}^k) + \psi_{\sim k|k}(U_i^{\sim k} + a_i^k + p_{l,i}^k) \tag{B.4}$$

where $\sim k$ denotes "not" k. Similarly, letting $\psi_{k|0}$ denote the probability that candidate k wins given that i did not vote, then i's expected payoff from not voting is simply:

$$V_i^0 = \psi_{k|0}U_i^k + \psi_{\sim k|0}U_i^{\sim k} \tag{B.5}$$

Since elections in conflict zones are often characterized by violence, the individual takes into account an exogenous probability δ_j that a violent event takes place at the polling center and as a result receives a very negative payoff P. The individual therefore decides to vote if:

$$\delta(V_i^k - P) + (1 - \delta)V_i^k \geq V_i^0 \tag{B.6}$$

Given expressions (B.4), (B.5) and (B.6), the individual will decide to vote and vote for k if the

[6]This campaign promise can be interpreted in either of two ways: Voting is not secret so that candidate k knows which individuals voted for him and hence he pays them the campaign promise $p_{l,i}^k$, or voting is secret but once a voter commits *a priori* to tender the vote to k he does not change his vote the day of the election.

following two conditions hold.

$$p_{l,i}^{k}+a_{i}^{k}+\psi_{k|k}U_{i}^{k}+\psi_{\sim k|k}U^{\sim k}\geq\delta P+\psi_{k|0}U_{i}^{k}+\psi_{\sim k|0}U^{\sim k} \tag{B.7}$$

$$p_{l,i}^{k}+a_{i}^{k}+\psi_{k|k}U_{i}^{k}+\psi_{\sim k|k}U^{\sim k}\geq p_{l,i}^{\sim k}+\psi_{k|\sim k}U_{i}^{k}+\psi_{\sim k|\sim k}U^{\sim k} \tag{B.8}$$

Assuming, for simplicity, that the individual believes that his vote is "non-pivotal"[7], then the two expressions above simply reduce to:

$$p_{l,i}^{k}+a_{i}^{k}\geq\delta P \tag{B.9}$$

$$p_{l,i}^{k}+a_{i}^{k}\geq p_{l,i}^{\sim k}+a_{i}^{\sim k} \tag{B.10}$$

The minimum price per legal vote (i.e., the campaign promise) that candidate k must pay is therefore given by:

$$p_{l,i}^{k}=max\left\{0,\delta P-a_{i}^{k},p_{l,i}^{\sim k}+a_{i}^{\sim k}-a_{i}^{k}\right\} \tag{B.11}$$

The Candidate's Problem Candidate k must decide how many votes (both legal and fraudulent) to buy from each center j. Assume that the auditing agency can differentiate between fraudulent and legal votes so that, once audited, any fraudulent votes are dropped and the candidate only receives legal votes v_l^k. In case where the center is not audited the candidate simply keeps all votes $v_l^k+v_f^k$. Given that the assessed probability of an audit is given by (3.2) and assuming that the candidate has quasilinear preferences over votes, then the maximization problem of the candidate is given by:

$$\max_{v_l^k,v_f^k} \quad \pi v_l^k+(1-\pi)[v_l^k+(v_f^k)^\alpha]$$

$$\text{subject to} \quad p_f v_f^k+\sum_{i=1}^{v_l^k}p_{l,i}^k\leq E^k$$

where fraudulent votes enter non-linearly (with $\alpha\leq 1$) to capture the possibility that fraudulent and

[7]This assumption simply states that the individual believes that tendering the vote to k will not alter the probability that k wins. More specifically, $\psi_{k|k}=\psi_{k|\sim k}=\psi_{k|0}$ for $k\in\{X,Y\}$.

legal votes are not perfect substitutes as specified in Callen and Weidmann (2013) and E^k is some campaign endowment of candidate k. To simplify the analysis, assume that the candidate does not observe the affinity parameters a_i^k but knows their distribution among the voters in center j so that $\sum_{i=1}^{v_l^k} p_{l,i}^k$ is simply given by $p_l^k v_l^k$ where p_l^k uses the expected value of these affinity parameters. The solution to the problem above provides an optimal relationship between fraudulent votes and their price p_f that is given by:

$$v_f^k = \left[\frac{\alpha(1-\pi)p_l^k}{p_f}\right]^{\frac{1}{1-\alpha}} \tag{B.12}$$

Substituting the expressions for prices p_f and p_l^k in order to obtain the equilibrium level of fraud gives:

$$v_f^k = \left[\frac{\alpha(1-\pi)\cdot max\left\{0, \delta P - a^k, p_l^{\sim k} + a^{\sim k} - a^k\right\}}{\pi F}\right]^{\frac{1}{1-\alpha}} \tag{B.13}$$

From this expression it is clear that the probability π that the center is audited (i.e., the level of social monitoring) decreases the equilibrium fraud level in the center, however, notice also that the "social monitoring effect" is one among others that explain fraud. To see this more clearly, I rewrite expression (B.13) by separating the different components of fraud:

$$v_f^k = \left[\alpha \cdot \underbrace{\frac{1-\pi}{\pi}}_{\substack{\text{Social}\\\text{monitoring}\\\text{effect}}} \cdot \underbrace{\frac{1}{F}}_{\substack{\text{Fine}\\\text{effect}}} \cdot max\left\{0, \underbrace{\delta P - a^k}_{\substack{\text{Violence}\\\text{effect}}}, \underbrace{p_l^{\sim k} + a^{\sim k} - a^k}_{\substack{\text{Challenger}\\\text{effect}}}\right\}\right]^{\frac{1}{1-\alpha}} \tag{B.14}$$

APPENDIX C

ADDITIONAL FIGURES AND TABLES

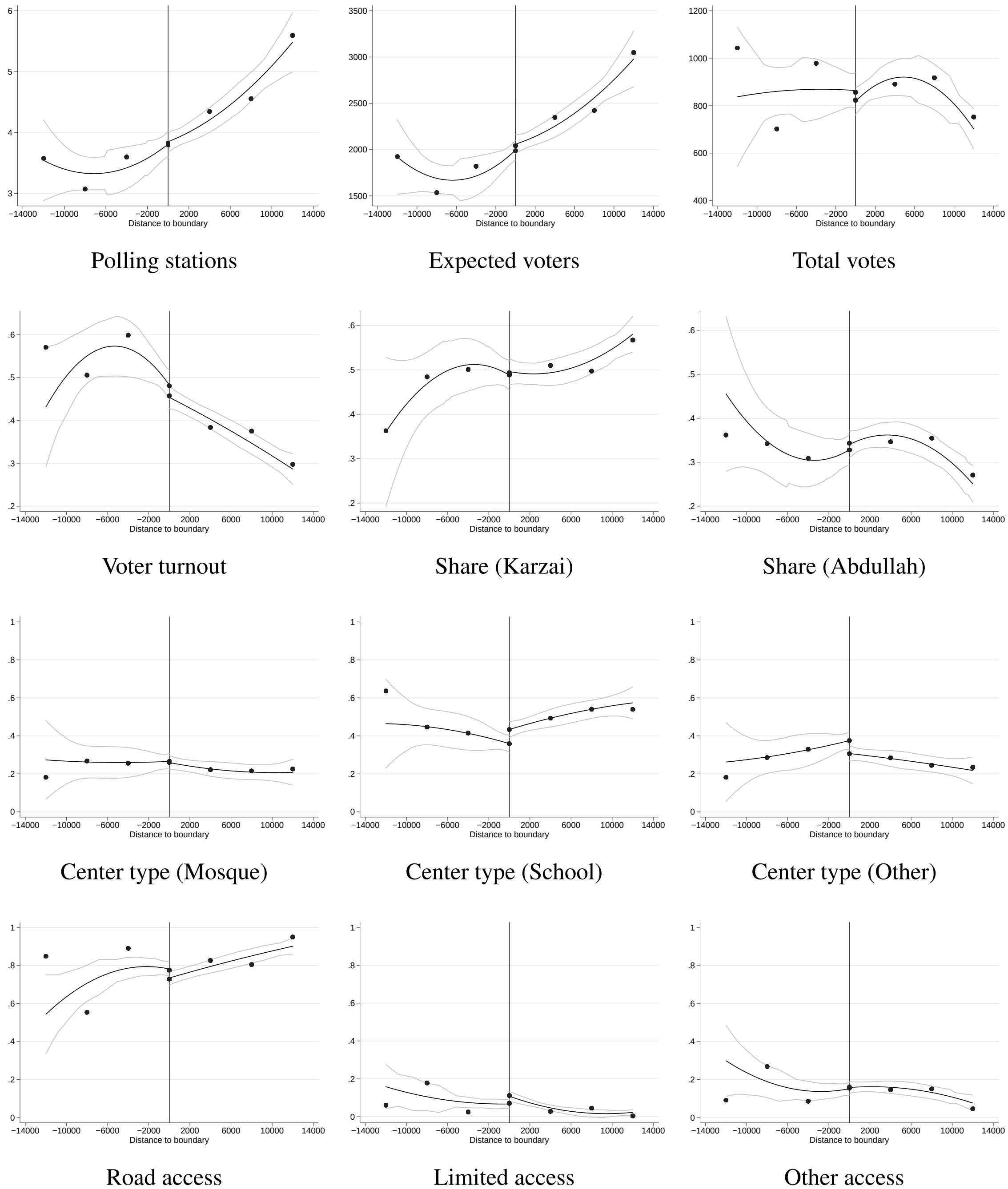


Figure C.1: Binned Averages for Various Covariates (Covariate RD Plots)

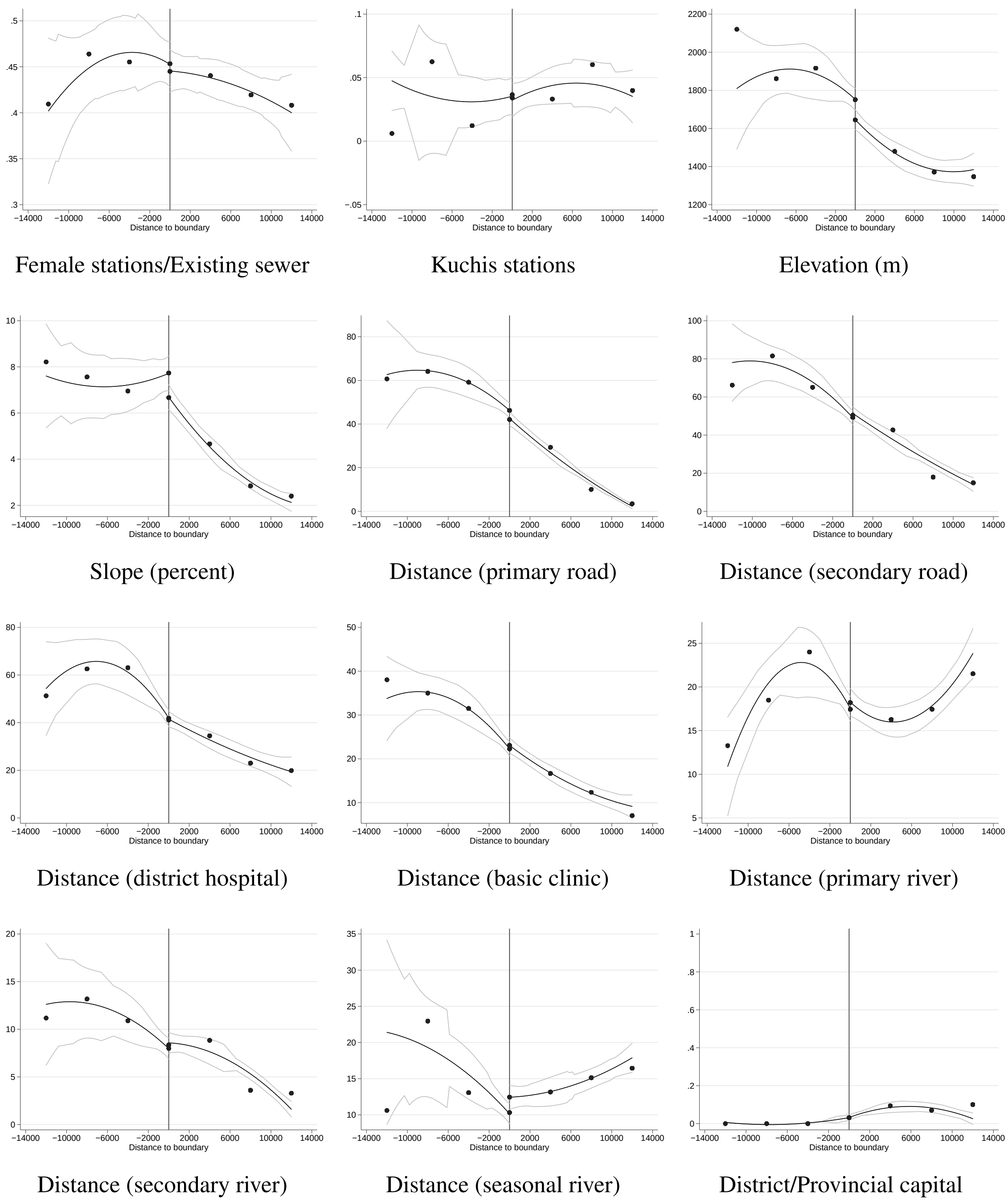
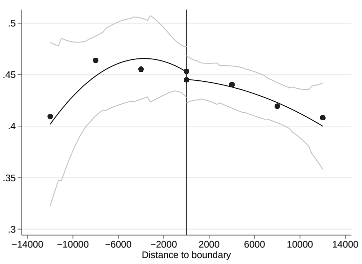
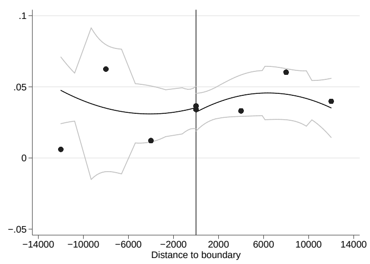


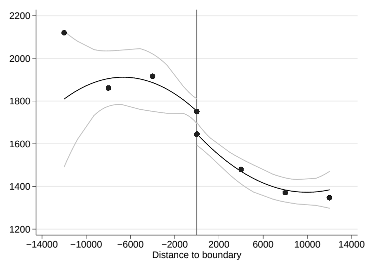
Figure C.1: Binned Averages for Various Covariates (Covariate RD Plots) - *Continues*



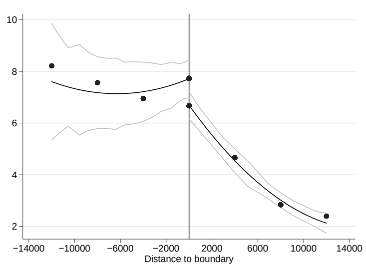
Population



Dari language



Pashto language



Other language

Figure C.1: Binned Averages for Various Covariates (Covariate RD Plots) - *Continued*

Notes: Solid dots give the average value of the specified covariate for polling centers falling within 5000-meter distance bins. Refer to section 4.1 in the text for a detailed description of each variable. Dots are plotted at the start of the bin (i.e., the dot representing the average for centers in the 0-5,000 meter bin is located at 0.). "Distance to boundary" refers to the normalized value of the forcing variable or distance between a polling center and the closest point in the cell phone coverage boundary. Distance is measured in meters. "Negative" values of distance give the distance of polling centers in non-coverage areas. The solid line trends give the predicted values from a regression of the outcome variable on a second degree polynomial in distance to the boundary that uses a rectangular kernel and a bandwidth of 15,000 meters.

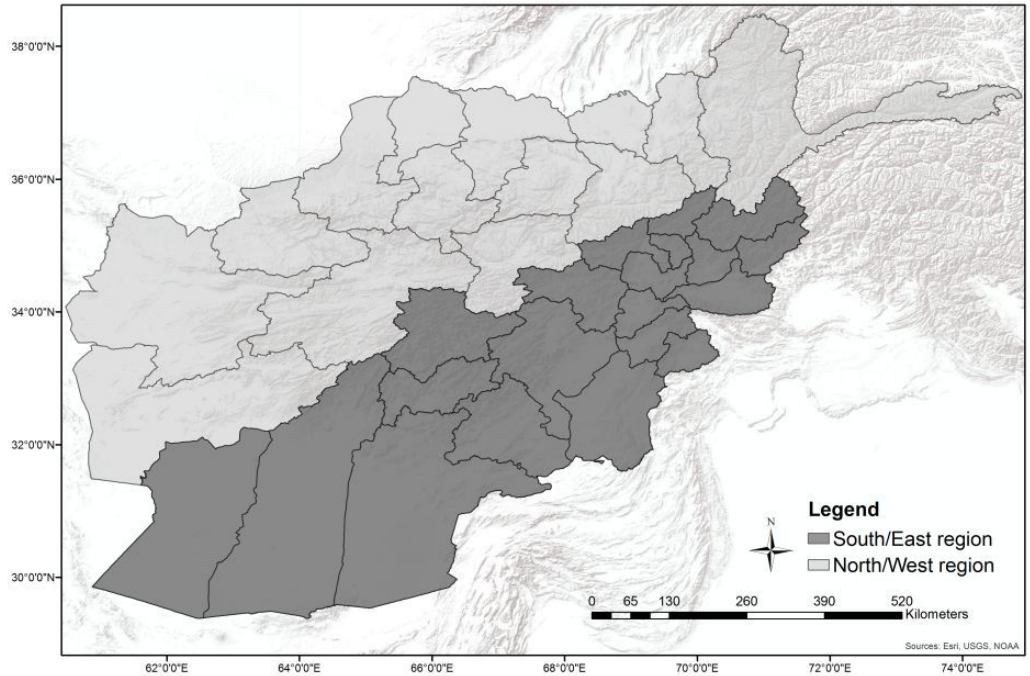


Figure C.2: Afghan Provinces and Regions

Notes: Regions of Afghanistan. Darker shade indicates the Southeastern provinces. Regions defined using International Security Assistance Forces (ISAF) regional command center definitions. Lines demarcate the provinces of Afghanistan. Map overlaid on USGS topographic basemap.

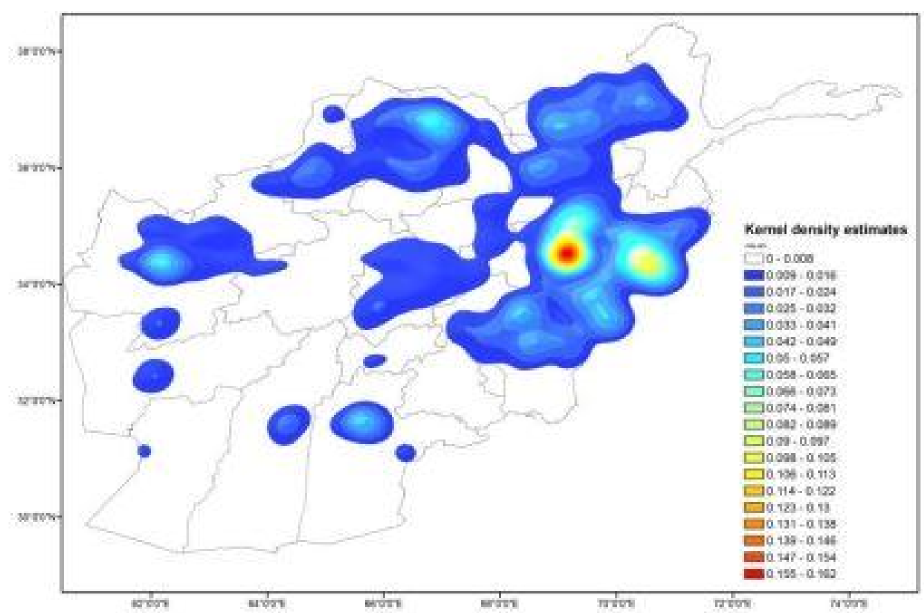


Figure C.3: Polling Center Density Estimates

Notes: Polling center density estimates obtained via kernel density estimation using a 50km bandwidth. Estimates obtained using the *Kernel Density* tool in *ArcGIS Spatial Analyst* package. Color scale gives the value of the density. Dots indicate the location of polling centers.

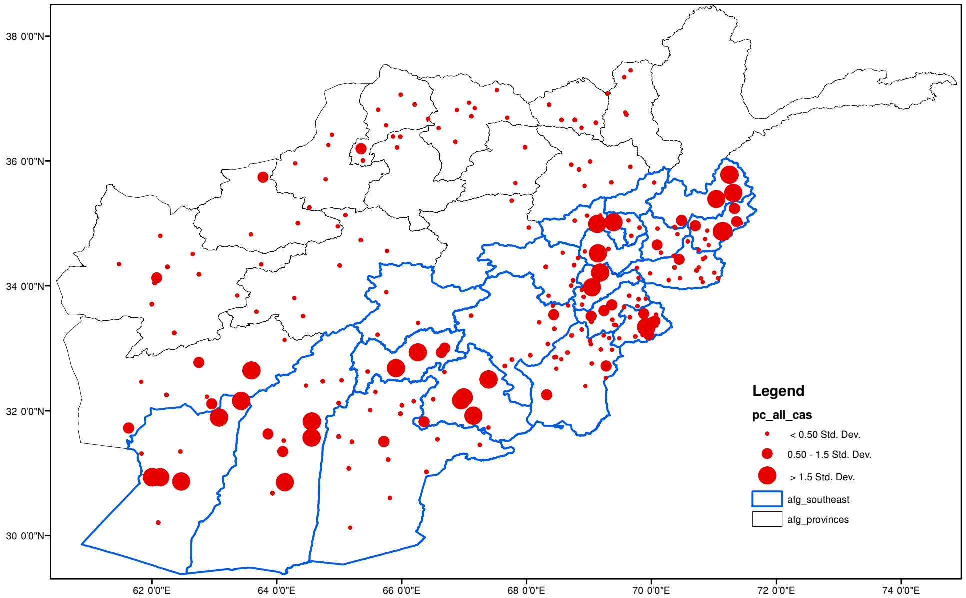


Figure C.4: All Casualties (Jan, 2009 - Sept, 2009)

Notes: Rate of combined casualties (military and civilian) per 10,000 population during 2009 (Election year)

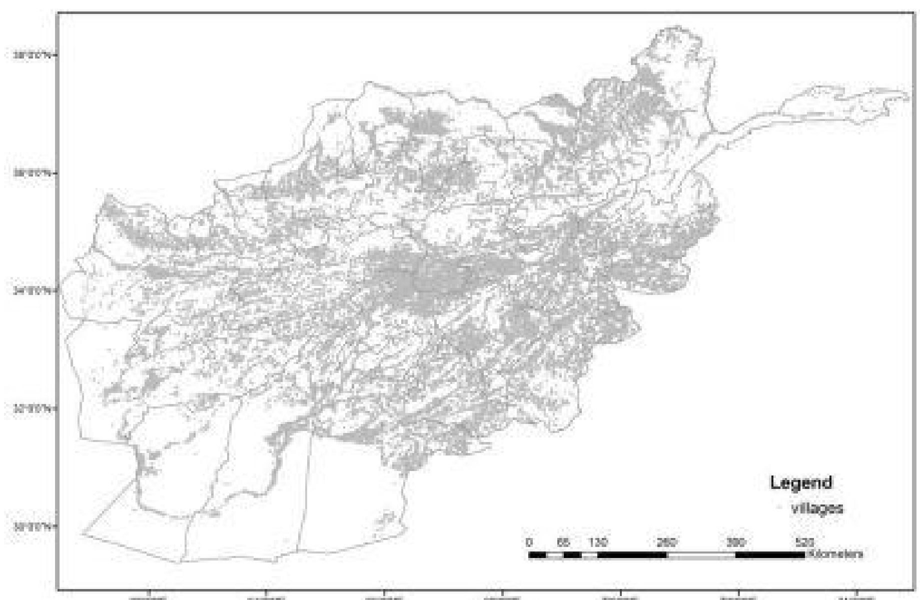


Figure C.5: Afghan villages, 2012

Notes: Afghan Villages (MISTI 2013)

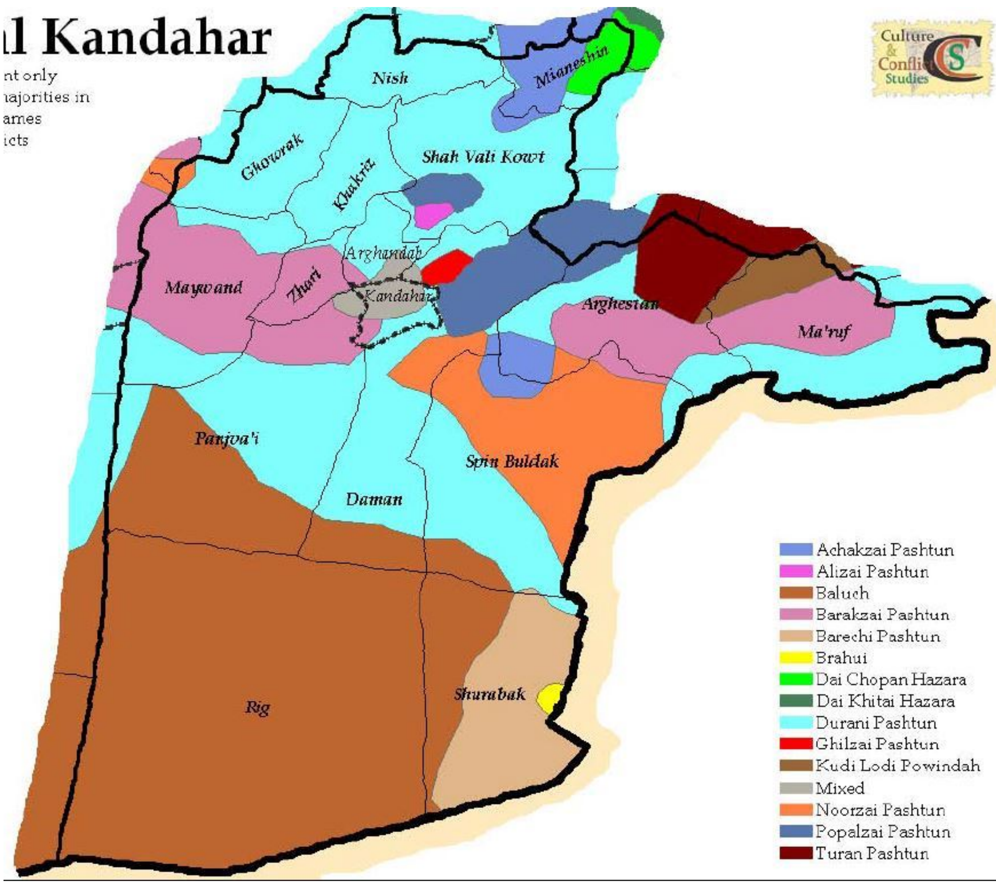


Figure C.6: Tribal Map, Kandahar Province

Notes: Tribal map for the province of Kandahar. Data obtained from the Culture and Conflict Studies program from the Naval Postgraduate School.

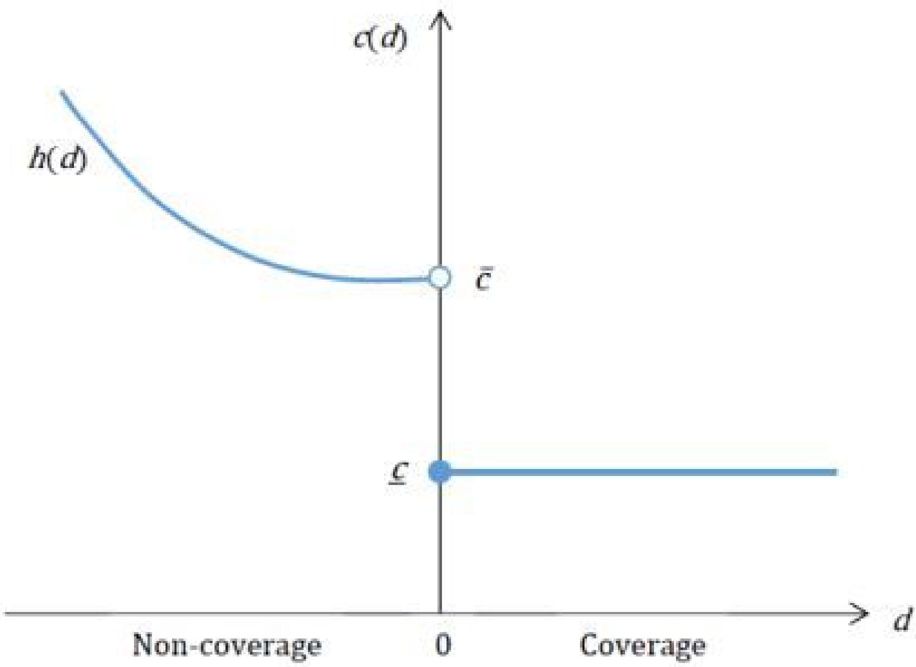
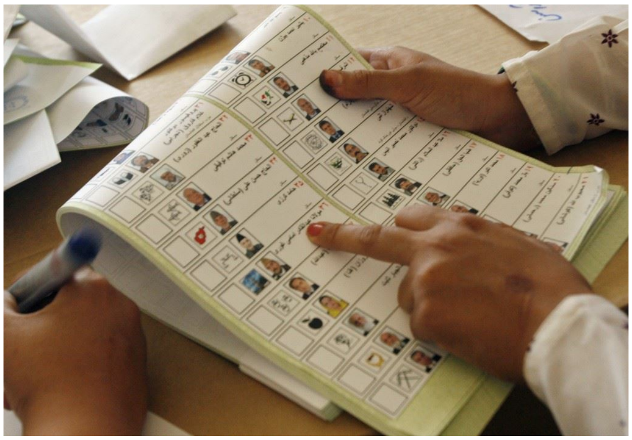


Figure C.7: Sample Reporting Cost Function



(a) Ballot

(b) Ballots tore in batches (Paktika province)

(c) Center manager filling ballots (Kandahar province)

(d) Purchased registration cards

Figure C.8: Sample Ballot and Examples of Fraud Captured by Local Media

Sources:
Panel A: Voice of America. Available at http://blogs.voanews.com
Panels B and C: Khadhouri (2010)
Panel D: Associated Press. Available at: http://www.rawa.org/temp/runews/2009/08/28

Table C.1: Sample and Imputations

Sample within:	Full sample		10 km of bound.		6 km of bound.		4 km of bound.	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
	Panel A. Unrestricted sample							
Not imputed	5,904	95.84	3,329	95.77	2,388	94.99	1,845	94.81
Imputed based on:								
Settlement	169	2.74	103	2.96	89	3.54	72	3.70
Nearest center	81	1.31	42	1.21	35	1.39	27	1.39
District capital	6	0.10	2	0.06	2	0.08	2	0.10
Total	6,160	100	3,476	100	2,514	100	1,946	100
	Panel B. Restricted sample							
Not imputed	2,331	96.04	1,377	95.49	1,106	95.18	912	95.10
Imputed based on:								
Settlement	67	2.76	49	3.40	41	3.53	36	3.75
Nearest center	26	1.07	15	1.04	14	1.20	10	1.04
District capital	3	0.12	1	0.07	1	0.09	1	0.10
Total	2,427	100	1,442	100	1,162	100	959	100

Notes: "Not imputed" refers to centers for which data were available after the merging of 2009 fraud data and 2010 geographic coordinate data. Imputations based on settlement give the polling center the coordinates of the village or settlement center where the polling center is located. Imputations based on nearest center give the polling center the coordinates of the polling center that, within the district, has the closest ID code to it. This is done because the assignment of ID codes followed a spatial order for the most part. Imputations based on district center simply give the polling center the coordinates of the district's capital where the center is located. *Restricted sample* refers to sample where at least one polling center is located on each side of a defined neighborhood. The restricted sample constitutes the main estimation sample.

Table C.2: Mean Comparison for Various Polling Center Characteristics

	Within 10 km of boundary			Within 5 km of boundary			RD estimates	
	Coverage (1)	No Coverage (2)	S.E. (3)	Coverage (4)	No Coverage (5)	S.E. (6)	RD coeff. (7)	S.E. (8)
Fraud outcomes (Category C+ fraud)								
All regions	0.08	0.11	(0.02)**	0.08	0.12	(0.02)**	-0.04	(0.019)**
			[0.02]**			[0.02]**		[0.020]**
East and South	0.14	0.20	(0.03)**	0.13	0.20	(0.03)**	-0.08	(0.032)**
			[0.03]**			[0.03]**		[0.033]**
North and West	0.01	0.01	(0.01)	0.01	0.01	(0.01)	0.00	(0.010)
			[0.01]			[0.01]		[0.010]
Electoral outcomes								
No. of stations	4.09	3.73	(0.16)**	3.86	3.78	(0.18)	0.09	(0.175)
			[0.16]**			[0.17]		[0.173]
No. of expected voters	2194.00	1944.00	(94.04)***	2069.00	1979.00	(100.00)	94.33	(101.198)
			[92.99]***			[99.35]		[100.161]
Total votes	871.80	866.60	(56.34)	835.00	863.70	(62.41)	-28.22	(59.992)
			[55.72]			[61.72]		[59.076]
Voter turnout	0.43	0.50	(0.02)***	0.45	0.49	(0.02)	-0.03	(0.023)
			[0.02]***			[0.02]		[0.023]
Vote share:								
Karzai	0.50	0.49	(0.03)	0.50	0.49	(0.03)	0.01	(0.030)
			[0.03]			[0.03]		[0.030]
Abdullah	0.34	0.33	(0.03)	0.35	0.33	(0.03)	0.01	(0.031)
			[0.03]			[0.03]		[0.030]
Polling center characteristics								
Polling center type:								
Mosque	0.24	0.26	(0.03)	0.25	0.26	(0.03)	-0.01	(0.029)
			[0.03]			[0.03]		[0.029]
School	0.46	0.37	(0.03)***	0.44	0.37	(0.03)**	0.06	(0.034)*
			[0.03]***			[0.03]**		[0.033]*
Other type	0.30	0.37	(0.03)**	0.30	0.37	(0.03)*	-0.05	(0.032)
			[0.03]**			[0.03]*		[0.032]
Polling center access (2010):								
Road access	0.76	0.78	(0.04)	0.73	0.78	(0.04)	-0.05	(0.039)
			[0.04]			[0.04]		[0.038]
Limited access	0.08	0.07	(0.02)	0.10	0.07	(0.02)	0.03	(0.023)
			[0.02]			[0.02]		[0.022]
Other access	0.16	0.15	(0.03)	0.17	0.15	(0.03)	0.02	(0.032)
			[0.03]			[0.03]		[0.032]
Share female stations	0.44	0.45	(0.01)	0.44	0.45	(0.01)	-0.01	(0.012)
			[0.01]			[0.01]		[0.012]
Share Kuchis stations	0.04	0.03	(0.01)	0.03	0.04	(0.01)	0.00	(0.010)
			[0.01]			[0.01]		[0.010]

(Continues)

Table C.2: Mean Comparison for Various Polling Center Characteristics - *Continued*

	Within 10 km of boundary			Within 5 km of boundary			RD estimates	
	Coverage (1)	No Coverage (2)	S.E. (3)	Coverage (4)	No Coverage (5)	S.E. (6)	RD coeff. (7)	S.E. (8)
Geographic characteristics								
Elevation (meters)	1570.00	1782.00	(58.06)***	1617.00	1756.00	(50.71)***	-128.72	(50.617)**
			[57.59]***			[51.89]***		[51.586]**
Slope (percent)	5.72	7.57	(0.53)***	6.46	7.66	(0.58)**	-1.01	(0.602)*
			[0.53]***			[0.58]**		[0.593]*
Economic development characteristics								
Distance (km) to:								
Primary road (2005)	35.30	48.62	(2.43)***	40.72	47.07	(2.31)***	-5.70	(2.290)**
			[2.51]***			[2.41]***		[2.370]**
Secondary road (2005)	44.65	52.60	(3.37)**	50.08	49.84	(3.12)	0.49	(3.056)
			[3.41]**			[3.19]		[3.110]
District hospital (2005)	37.56	45.42	(2.76)***	40.25	42.68	(2.73)	-2.34	(2.739)
			[2.86]***			[2.78]		[2.789]
Basic health center (2005)	20.12	24.31	(1.87)**	21.78	22.68	(1.74)	-0.38	(1.837)
			[1.83]**			[1.75]		[1.836]
Primary river	17.45	18.38	(1.40)	18.00	17.90	(1.29)	0.19	(1.319)
			[1.45]			[1.34]		[1.366]
Secondary river	8.11	8.71	(1.12)	8.96	8.03	(1.10)	0.61	(1.040)
			[1.13]			[1.13]		[1.053]
Seasonal river	12.66	11.05	(1.54)	13.16	10.30	(1.77)	2.70	(1.669)
			[1.62]			[1.83]		[1.715]
District/Province capital	0.05	0.03	(0.01)**	0.04	0.03	(0.01)	0.01	(0.013)
			[0.01]**			[0.01]		[0.013]
Demographic characteristics (of closest settlement)								
Population (2012-2013)	1287.00	957.00	(223.69)	1004.00	1018.00	(145.37)	-51.25	(141.521)
			[222.92]			[146.84]		[142.455]
Language spoken (2012-2013):								
Dari	0.43	0.46	(0.04)	0.43	0.43	(0.04)	0.00	(0.040)
			[0.04]			[0.04]		[0.040]
Pashto	0.44	0.43	(0.04)	0.43	0.45	(0.04)	-0.01	(0.041)
			[0.04]			[0.04]		[0.042]
Other	0.13	0.12	(0.03)	0.14	0.12	(0.03)	0.01	(0.026)
			[0.02]			[0.03]		[0.026]
Observations	891	551		601	456		601	456

Notes: Columns (1), (2), (4), and (5) give the means of the corresponding variable. Columns (3), (6), and (8) give the clustered standard errors for the difference in means in parenthesis and Conley (1999) standard errors in brackets for the difference in means. Conley (1999) standard errors use a distance cutoff of 50 kilometers and a Bartlett spatial weighting kernel. Refer to the notes in Table 4.4 for a definition of the variables.

Table C.3: Tribes and Tribal Confederations (Southeast Afghanistan)

Tribe	Confederation	Ethnic group
Alizai Pashtun	Durrani	Pashtun
Ashakzai Pashtun	Durrani	Pashtun
Barakzai Pashtun	Durrani	Pashtun
Durani Pahtun	Durrani	Pashtun
Mixed Durrani	Durrani	Pashtun
Noorzai Pashtun	Durrani	Pashtun
Panjpai Durani Pashtun	Durrani	Pashtun
Popalzai Pashtun	Durrani	Pashtun
Baezai Mohamand Powindah Pashtun	Ghilzai	Pashtun
Ghilzai Pashtun	Ghilzai	Pashtun
Ibrahim Ghilzai Pashtun	Ghilzai	Pashtun
Kudi Lodi Powindah	Ghilzai	Pashtun
Kukozai Mohamand Powindah Pashtun	Ghilzai	Pashtun
Mian Khel Powindah Pashtun	Ghilzai	Pashtun
Miani Powindah Pashtun	Ghilzai	Pashtun
Turan Pashtun	Ghilzai	Pashtun
Jadran Pashtun	Jadran	Pashtun
Kom	Kand	Pashtun
Mamund Kakazai	Kand	Pashtun
Salarzai	Kand	Pashtun
Wur	Kand	Pashtun
Dautani Pashtun	Lodi	Pashtun
Umar Khel Dautani	Lodi	Pashtun
Safi Pashtun	Safi	Pashtun
Alisher Khel Shinwari Pashtun	Shinwari	Pashtun
Manduzai Shinwari Pashtun	Shinwari	Pashtun
Sangu Shinwari Pashtun	Shinwari	Pashtun
Shinwari Pashtun	Shinwari	Pashtun
Wardak Pashtun	Wardak	Pashtun
Buto Khel Mohamand Pashtun	Other Pashtun	Pashtun
Kwhaezai Mohamand Pasthun	Other Pashtun	Pashtun
Pashtun	Pashtun undefined	Pashtun
Tajik	Tajik	Tajik
Besud Hazara	Hazara	Hazara
Chahar Dasta Hazara	Hazara	Hazara
Dai Chopan Hazara	Hazara	Hazara
Dai Khitai Hazara	Hazara	Hazara
Dai Kundi Hazara	Hazara	Hazara
Dai Zangi Hazara	Hazara	Hazara
Faoladi Hazara	Hazara	Hazara
Hazara	Hazara	Hazara
Jaghatus Hazara	Hazara	Hazara
Jaghuri Hazara	Hazara	Hazara
Khatai Hazara	Hazara	Hazara
Muhammad Kwaja Hazara	Hazara	Hazara
Polada Hazara	Hazara	Hazara
Uruzgani Hazara	Hazara	Hazara
Baluch	Baluch	Baluch
Kizilbash	Kizilbash	Kizilbash
Mixed	Mixed	Mixed
Tregami Nuristani	Nuristani	Nuristani
Gramsana	Other (non-Pashtun)	Other (non-Pashtun)
Kalasha	Other (non-Pashtun)	Other (non-Pashtun)
Sepah Mohamand	Other (non-Pashtun)	Other (non-Pashtun)
Tirahi	Tirahi	Tirahi
Uzbek	Uzbek	Uzbek

Notes: Tribal confederations created using Culture and Conflict Studies program's definitions. Definitions based on Tribal Hierarchy and Dictionary of Afghanistan (2007)

REFERENCES

Afghan Telecommunication Regulatory Authority (2012). Coverage footprint (2012).

AIMS (1997-2005). Roads of Afghanistan. *Afghanistan Information Management Services (AIMS)*.

Aker, J. (2010). Information from markets near and far: Mobile phones and agricultural markets in Niger. *American Economic Journal: Applied Economics*, 46–59.

Aker, J., P. Collier, and P. Vicente (2014). Is information power? using mobile phones and free newspapers during an election in Mozambique. *Working Paper*.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy 2*(1), 1–30.

Beber, B. and A. Scacco (2012). What the numbers say: A digit-based test for election fraud. *Political Analysis*, 211–234.

Bjorkman, M. and J. Svensson (2009). Power to the people: Evidence from a randomized field experiment on community-based monitoring in Uganda. *Quarterly Journal of Economics*.

Black, D. A., J. Galdo, and J. A. Smith (2007). Evaluating the regression discontinuity design using experimental data. *Working Paper, University of Michigan*.

Black, S. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly Journal of Economics*, 577–599.

Callen, M. and J. D. Long (2015). Institutional corruption and election fraud: Evidence from a field experiment in Afghanistan. *American Economic Review 105*(1), 354–381.

Callen, M. and N. B. Weidmann (2013). Violence and election fraud: Evidence from Afghanistan. *British Journal of Political Science*, 53–75.

Chassang, S. and G. Padro-i-Miquel (2014). Corruption, intimidation, and whistleblowing: A theory of inference from unverifiable reports. *Working Paper*.

Collier, P. and P. C. Vicente (2012). Violence, bribery, and fraud: the political economy of elections in Sub-Saharan Africa. *Public Choice 153*(1-2), 117–147.

Conley, T. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics 92*, 1–45.

Cordesman, A. (2005). Iraq's evolving insurgency. *Center for Strategic and International Studies. Washington, DC*.

Cremer, H. and F. Gahvari (1994). Tax evasion, concealment and the optimal linear income tax. *Scandinavian Journal of Economics*.

Dekel, E., M. O. Jackson, and A. Wolinsky (2008). Vote buying: General elections. *Journal of Political Economy 116*(2).

Dell, M. (2010). The persistent effect of Peru's mining mita. *Econometrica*.

Electoral Complaints Commission (2010). Final report. 2009 presidential and provincial council elections. Kabul.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 998–1004.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. CRC Press.

Figueiras, J. and S. Frattasi (2010). *Mobile Positioning and Tracking: From Conventional to Cooperative Techniques*. Wiley.

Filkins, D. (2009). Threats by Taliban may sway vote in Afghanistan. *The New York Times*.

Gall, C. (2009). Violence roils Afghanistan days before election. *The New York Times*.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall; 7 edition.

Hahn, J., P. Todd, and W. van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Hamdard, J. (2012). The state of telecommunications and internet in Afghanistan: Six years later 2006-2012. *Internews*.

Himelfarb, S. (2010). Can you help me now? Mobile phones and peacebuilding in Afghanistan. *United States Institute of Peace*.

Holmes, T. (1998). The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy 106*(4), 667–705.

Imbens, G. and K. Kalyanaraman (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies 79*(3), 933–959.

Imbens, G. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*(2), 615–635.

Imbens, G. and T. Zajonc (2011). Regression discontinuity design with multiple forcing variables. *Working Paper*.

ITU (2015). International telecommunication union (itu). (Available at: http://www.itu.int/en/Pages/default.aspx).

Jensen, R. (2007). The digital provide: Information (technology), market performance, and welfare in the south indian fisheries sector. *Quarterly Journal of Economics 122*, 879–924.

Kane, T., S. Riegg, and D. Staiger (2006). School quality, neighborhoods, and housing prices. *American Law and Economic Review 8*(2), 183–212.

Keele, L. and R. Titiunik (2013). Geographic boundaries as regression discontinuities. *Working Paper*.

Khadhouri, S. (2010). A review of suspected electoral fraud. 2009 Afghan presidential and provincial council elections. *Democracy International*.

Lakshmanan, I. A. (2010). Fighting the Taliban with cellphones. *The New York Times*.

Lalive, R. (2008). How do extended benefits affect unemployment duration? a regression discontinuity approach. *Journal of Econometrics 142*, 785–806.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics 142*(2), 675–697.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Mauro, P. (1995). Corruption and growth. *Quarterly Journal of Economics 110*, 681–712.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2).

MISTI (2013). Measuring impacts of stabilization initiatives (misti). (Available at: http://usaidmisti.com/gis-data).

National Aeronautics and Space Administration and the National Geospatial Intelligence Agency (2000). Retrieved from shuttle radar topography mission 30 arc second finished data.

National Democratic Institute (2010). The 2009 presidential and provincial council elections in Afghanistan. Washington DC, US.

Olken, B. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy 115*(2).

Olken, B. and R. Pande (2012). Corruption in developing countries. *Annual Review of Economics 4*, 479–509.

Papay, J. P., J. B. Willett, and R. J. Murnane (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics 161*(2), 203 – 207.

Pierskalla, J. H. and F. M. Hollenbach (2013, 5). Technology and collective action: The effect of cell phone coverage on political violence in Africa. *American Political Science Review 107*, 207–224.

Porter, J. (2003). Estimation in the regression discontinuity model. *Unpublished Manuscript*.

Reardon, S. and J. Robinson (2010). Regression discontinuity designs with multiple rating-score variables. *Working Paper*.

Riechmann, D. (2014). Us general: Corruption is top threat in Afghanistan. *Associated Press*, 411–439.

Robinson, F. (2013). Fewer cell towers are shut down in Afghanistan, minister says. *The Wall Street Journal*.

Schuler, I. (2008). National democratic institute: Sms as a tool in election observation. *Innovations 3*(2).

Shapiro, J. N. and N. B. Weidmann (2013). Is the phone mightier than the sword? cell phones and insurgent violence in Iraq. *Working Paper*.

Shaver, A. and A. Wright (2015). Are modern insurgencies predictable? New evidence from Afghanistan and Iraq. *Working Paper*.

Strother, T. (2007). Cell phone use by insurgents in Iraq. *Urban Warfare Analysis Center*.

Svensson, J. (2005). Eight questions about corruption. *Journal of Economic Perspectives 19*, 19–42.

Tarzi, A. and R. Lamb (2011). *Measuring Perceptions about the Pashtun People*. Center for Strategic and International Studies.

Transparency International (2013). Corruption perception index 2013.

Tribal Hierarchy and Dictionary of Afghanistan (2007). *Tribal hierarchy and dictionary of Afghanistan: a reference aid for analysts*. Courage Services Inc.

WITS (2005-2009). Worldwide Incidents Tracking System. *National Counterterrorism Center* (Available at: http://wits.nctc.gov).

Wong, V., P. Steiner, and T. Cook (2010). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics 38*(2), 107–141.