

Online Appendix to:
Hidden in Plain Sight? Irregularities on Statutory Forms
and Electoral Fraud

Zach Warner*

J. Andrew Harris[†]

Michelle Brown[‡]

Christian Arnold[§]

September 12, 2021

* Assistant Professor, Purdue University. email zachwarner@purdue.edu, web: <http://www.zachwarner.net>.

[†]Assistant Professor, New York University, Abu Dhabi. email andy.harris@nyu.edu, web: <http://www.jandrewharris.org/>.

[‡]Senior Advisor, Elections and Political Processes, National Democratic Institute.

[§]Senior Lecturer, Cardiff University. email: ArnoldC6@cardiff.ac.uk, web: <http://christianarnold.org/>.

Contents

A.1	Instructions for coding irregularities	3
A.2	Details about model estimation	20
A.2.1	Model architectures	20
A.2.2	Data augmentations	21
A.2.3	Tuning procedures	22
A.2.4	Best-performing models	22
A.3	Confusion matrices, classification model training	24
A.4	Computation of F_1 scores for Cantú (2019)	26
A.5	Empirical distribution of irregularities	27
A.6	Covariate data sources	28
A.7	Full results for Figure 3 (in main text)	29
A.8	Robustness to alternative definitions of strongholds	30
A.9	Full regression results, Table 3	33
A.10	Regression discontinuity analysis	37

A.1 Instructions for coding irregularities

The following pages present the instructions given to our research assistants for coding our training sample of statutory forms from 3,000 polling stations.

Project Instructions for: “Hidden in Plain Sight? Detecting Electoral Irregularities Using Statutory Results”

Last updated: April 30, 2021

This project requires you to determine whether documents from Kenya’s 2013 presidential election contain any irregularities. Each form is provided as a single, two-page PDF document. A master spreadsheet provides the names of all documents and provides a space for each potential irregularity to be coded. Your job is to go through each individual document, coding each potential problem. Your responses will always be 0 or 1, except in a few rare cases where the data to be coded cannot be discerned, in which case you may enter NA.

Each row of the spreadsheet corresponds to an individual document. The first five columns (A-E) of the master spreadsheet provide information about the document, and will already be filled in. You will enter data into the next nine columns (F-N).

The remainder of these instructions are as follows. The next two pages provide a complete example of the documents you will be coding. As this example shows, the first page typically has a black, red, and green border with the IEBC logo at the top, while the second page consists mostly of a large table. The remainder of this document then provides specific instructions for each irregularity to be coded.

(r.79(2)(a),83)

FORM 34
34 PR 00001444



INDEPENDENT ELECTORAL AND BOUNDARIES COMMISSION
DECLARATION OF PRESIDENTIAL ELECTION RESULTS AT A POLLING STATION

POLLING STATION: KHADIJA PRIMARY SCHOOL (001)
STREAM: 1
CONSTITUENCY: NYALI (004)

1.	Total number of registered voters for the polling station	696
2.	Number of spoilt ballot papers	006
3.	Total number of votes cast	487
4.	Number of rejected votes	003
5.	Number of disputed votes	000
6.	Number of rejected objected to votes	000
7.	Total number of valid votes cast (in figures and words)	484

The number of valid votes cast in favour of each candidate:

	Name of Candidate	No. of Valid Votes Cast
1	JAMES LEGILISHO KIYAPI	XX X 00 X
2	MARTHA WANGARI KARUA	XX X 01 X
3	MOHAMED ABDUBA DIDA	XX X 04 X
4	MUSALIA MUDAVADI	XX X 08 X
5	PAUL KIBUGI MUISTE	XX X 01 X
6	PETER KENNETH	XX 11 X
7	RAILA ODINGA	X X 356 X
8	UHURU KENYATTA	XX 103 X
9	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
10	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
11	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
12	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
13	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
14	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
15	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
16	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
17	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
18	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
19	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
20	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX
21	XXXXXXXXXXXXXXXXXXXXXX	XXXXXX

8. Declaration
We, the undersigned, being present when the results of the count were announced, do hereby declare that the results shown above are true and accurate count of the ballots in:
KHADIJA PRIMARY SCHOOL (001) Polling Station,
NYALI (004) Constituency.

(i) Presiding Officer: Pauline Ndiyo
Name: Pauline Ndiyo
ID No. / Passport: 10979160
Signature: Pauline Ndiyo Date: 4/3/13 March 2013



Figure 1: An example statutory result (page 1).

(ii) Deputy Presiding Officer:

Name: DANIEL MOI OSALA
ID No. / Passport: 23091740 / A1151305
Signature: *Daniel* Date: 4th March 2013

Candidates or Candidates' Agents

	Name	Signature	Reasons for refusal to sign
1.	FELIX OCHENGO ORIEND	<i>Ochenyo</i>	
2.	CHRISPUS KAVU	<i>Kavu</i>	
3.	SAMWELL N. MEST	<i>Samwell</i>	
4.	MIKAIL MOHAMED	<i>Mikail</i>	
5.	FREDRICKA JILWARA	<i>Jilwara</i>	
6.	ATHMAN JUMA	<i>Athman</i>	
7.	MUCA DIMITIA	<i>Muca</i>	
8.	IRENE ACHIENG	<i>Irene</i>	
9.	MOHAMED YUSUF	<i>Mohamed</i>	
10.			
11.			
12.			
13.			
14.			
15.			
16.			
17.			
18.			
19.			
20.			
21.			
22.			
23.			
24.			
25.			
26.			
27.			
28.			
29.			
30.			

9. Presiding Officer's Statutory Comments:

The voting started up this late but went on well and everything was successful.

Figure 2: An example statutory result (page 2).

Column F: “Does the form have a QR code?”

The QR code can be found in the top-left corner of the first page of the document, usually just beneath the serial number. The QR code is a unique code embedded in a square-shaped, black-and-white pattern.

Code this variable as 1 if the QR code is present and completely visible. Code this variable as 0 if the QR code is missing or obscured. Figure 3 provides three examples of forms which should be coded as 0 and Figure 4 provides three examples of forms that should be coded as 1.

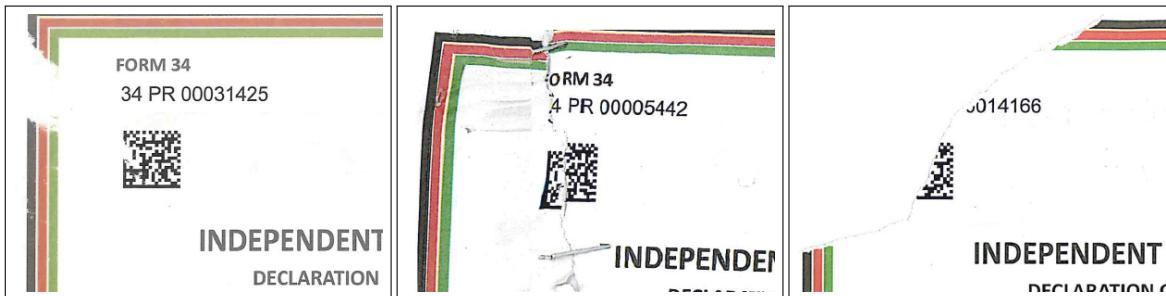


Figure 3: In these three examples, the QR codes are obscured or missing. Column F should be coded as 0 for these cases.



Figure 4: In these three examples, the QR codes are present and legible. Column F should be coded as 1 for these cases.

Column G: “Did the document scan correctly?”

Code this variable as 1 if there are no indications of an error while scanning, including stretched out text, miscoloration, or other visual artifacts. Code this variable as 0 if any such problems are apparent. Figure 5 provides two examples of forms which should be coded as 0 and Figure 6 provides two examples of forms that should be coded as 1.

The image contains two side-by-side scanned documents. The left document is a ballot form with a blue stamp that reads "I.E.B.C. PRESID OFFICE". The right document is a declaration form titled "INDEPENDENT ELE DECLARATION OF PR" with fields for "POLLING STATION: C.C.M MAC" and "STREAM: 1". Both documents appear to have been scanned correctly with clear text and images.

Figure 5: In these two examples, there is no evidence of problems with the document scan. Column G should be coded as 0 for these cases.

The image contains two side-by-side scanned documents. The left document is a ballot form with a blue stamp that reads "I.E.B.C.". The right document is a declaration form titled "INDEPENDENT ELE DECLARATION OF PR" with fields for "POLLING STATION: JILO MATA" and "STREAM: ". Both documents show significant scanning artifacts, including stretched-out text and distorted colors, indicating they did not scan correctly.

Figure 6: In these two examples, there is evidence that the document did not scan correctly. Column G should be coded as 1 for these cases.

Column H: “Is the first page stamped?”

The stamp is a circle with the words “I.E.B.C. PRESIDING OFFICER” or “IEBC RETURNING OFFICER” inside. It is usually placed near the bottom of the first page, but can be present anywhere on the first page, sometimes in multiple places.

Code this variable as 1 if the stamp is present on the first page, such that the words “presiding officer” (or “returning officer”) are legible. Code this variable as 0 if the stamp is missing. Figure 7 provides two examples of forms which should be coded as 0 and Figure 8 provides two examples of forms that should be coded as 1.

clare that the results shown above are true and accurate count of u AFIRISI NURSERY SCHOOL (026)	rsigned, being present when the results of the count were anno he results shown above are true and accurate count of the bal CHIEF'S CAMP - MACHAFUKO (010)
MU WEST (022))
Presiding Officer: Name: DICK CHARO ID No. / Passport: 2234608 Signature: 	Officer: RANDU ROSELYN SIDI Passport: 8522621 Date: 11.3.11

Figure 7: In these two examples, the stamp is missing or illegible on the first page. Column H should be coded as 0 for these cases.

above are true and accurate count of the ballots in: 1) STREAM 8 Pollin LIA MWADIME Cor 103118 Date: 5 th M	above are true and accurate count of the ballots in: 02) I.E.B.C. Polling Sta PRESIDING OFFICER Constitue HG ODINGA 3 Date: 4 th March 2
---	--

Figure 8: In these two examples, the stamp is present and legible on the first page. Column H should be coded as 1 for these cases.

Column I: “Is the presiding officer signature present?”

The presiding officer signature can be found at the bottom of the first page of the document. The signature is located just below the presiding officer’s name and ID, next to the date.

Code this variable as 1 if a signature is present. Code this variable as 0 if no signature is present. Figure 9 provides three examples of forms which should be coded as 0 and Figure 10 provides three examples of forms that should be coded as 1.

<p>8. Declaration</p> <p>We, the undersigned, being present when the results of the declare that the results shown above are true and accurate</p> <p>LIBEHIA PRIMARY SCHOOL (008)</p> <p>LAFEY (044)</p> <p>(i) Presiding Officer:</p> <p>Name: ABDIRASHID ABOOW ICA</p> <p>ID No. / Passport: 13330860</p> <p>Signature: _____</p> 	<p>8. Declaration</p> <p>We, the undersigned, being present when the results of the declare that the results shown above are true and accurate</p> <p>IFC CHURCH (007)</p> <p>KISAUNI (003)</p> <p>(i) Presiding Officer:</p> <p>Name: RUTH MEXU MAGAGO</p> <p>ID No. / Passport: 0232164</p> <p>Signature: _____</p>  
--	--

Figure 9: In these two examples, the presiding officer signature is missing. Column I should be coded as 0 for these cases.

<p>INARIA TRADING CENTER (005)</p> <p>TURKANA SOUTH (127)</p> <p>(i) Presiding Officer:</p> <p>Name: STEPHEN NAUR ETEE</p> <p>ID No. / Passport: 21952792</p> <p>Signature: _____</p> 	<p>IJARA (032)</p> <p>(i) Presiding Officer:</p> <p>Name: OMAR AHMED M</p> <p>ID No. / Passport: 21281920</p> <p>Signature: _____</p> 	<p>BURA (020)</p> <p>(i) Presiding Officer:</p> <p>Name: HASAN M. FAHIM</p> <p>ID No. / Passport: 20339238</p> <p>Signature: _____</p> 
---	---	--

Figure 10: In these three examples, the presiding officer signature is present. Column I should be coded as 1 for these cases.

Column J: “Is there at least one candidate/candidate agent listed?”

The candidate/candidate agent list is the large table that can be found in the middle of the second page of the document. Each row of the table provides the name of the candidate or agent, a space for their signature, and a space for a reason for refusing to sign.

Code this variable as 1 if there is at least one candidate or candidate agent listed in the left column of this table. Code this variable as 0 if no candidate or agent names are listed. Figure 11 provides two examples of forms which should be coded as 0 and Figure 12 provides two examples of forms that should be coded as 1.

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.		
2.		
3.		
4.		
5.		
6.		
7.		
Q		

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.		
2.		
3.		
4.		
5.		
6.		
7.		
Q		

Figure 11: In these two examples, there are no candidate or candidate agents listed. Column J should be coded as 0 for these cases.

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.	ISAIATH THOMAS AURA	
2.	BILLY IKUSYA MUVO	
3.	Mwakia, and mawacie	
4.	Kenya Kenya	
5.	SAID DENO MWANDO	
6.	Mwachyu AL	
7.	FELIX HOSA OGOLA	
Q		

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.	OMAR S. MAJANO	
2.	CLEMENT MWANGI	
3.	ESTHER MWAKA	
4.	TUMA M. MUNONY	
5.	ALNUOR S. ADAN	
6.	NATIMA K. HAMISI	
7.		
Q		

Figure 12: In these two examples, there are candidates or candidate agents listed. Column J should be coded as 1 for these cases.

Column K: “For each listed candidate/candidate agent, is there a signature?”

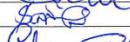
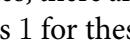
The candidate/candidate agent signatures can be found in the large table in the middle of the second page of the document. Each row of the table provides the name of the candidate or agent, a space for their signature, and a space for a reason for refusing to sign.

Code this variable as 1 if there is a signature in the middle column of this table for each candidate or candidate agent listed in the left column of this table. Code this variable as 0 if any candidate or agent listed does not have a corresponding signature. Figure 13 provides two examples of forms which should be coded as 0 and Figure 14 provides two examples of forms that should be coded as 1.

Candidates or Candidates' Agents		
	Name	Signature
1.	John Murage	
2.	John Kilatko m - Ghazi	
3.		
4.		
5.		
6.		

Candidates or Candidates' Agents		
	Name	Signature
1.	Oding END.	
2.	Catherine Mwitiya	
3.	Thoma mang'abi	
4.		
5.		
6.		

Figure 13: In these two examples, there is at least one candidate or candidate agent signature missing. Column K should be coded as 0 for these cases.

Candidates or Candidates' Agents		
	Name	Signature
1.	JACQUELINE AMALIA	
2.	NICAL LINET NGESA	
3.	SAMUEL ONTARI	
4.	CHARLES OMORI	
5.	DAVID NDUNGA	
6.	ASHA Mohamed	

Candidates or Candidates' Agents		
	Name	Signature
1.	MWEU KYENZE	
2.	CHRISTINE ADEERA	
3.	ANGELA SHANAY	
4.	JUMIA MBUL MAKUPISA	
5.	WANDERA-N ERNEST	
6.	DICHAZO MAINE KARIM	

Figure 14: In these two examples, there are no candidate or candidate agent signatures missing. Column K should be coded as 1 for these cases.

Column L: “Is there at least one signature of a candidate/candidate agent?”

The candidate/candidate agent signatures can be found in the large table in the middle of the second page of the document. Each row of the table provides the name of the candidate or agent, a space for their signature, and a space for a reason for refusing to sign.

Code this variable as 1 if there is at least one candidate or candidate agent signature in the middle column of this table. Code this variable as 0 if no candidate or agent signatures are present. Figure 15 provides two examples of forms which should be coded as 0 and Figure 16 provides two examples of forms that should be coded as 1.

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.	Ouring End.	
2.	Catherine Muthunga	
3.	Thoma mangala	
4.		
5.		
6.		

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.		
2.		
3.		
4.		
5.		
6.		

Figure 15: In these two examples, there are no candidate or candidate agent signatures. Column L should be coded as 0 for these cases.

<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.	Margaret chao	lloyd
2.	Muhammad Mujiua	MUJ
3.	Collins Otiende Omung'	BILL
4.	NELSON KIBARLU	YES
5.	HAFSIA SHAHAME	HAFSIA
6.	Beatrice Aoico	BEATRICE

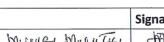
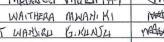
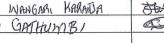
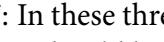
<u>Candidates or Candidates' Agents</u>		
	Name	Signature
1.	Caroline Amalva	Amalva
2.	ICAL LINET NGESA	ICAL LINET
3.	SAMREW ONTARI	SAMREW
4.	CHARLES OMORBI	CHARLES
5.	DAVIN NOUNGI	DAVIN
6.	Asha Mohamed	Mohamed

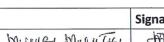
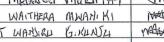
Figure 16: In these two examples, there is at least one candidate or candidate agent signature. Column L should be coded as 1 for these cases.

Column M: “Do the candidate/candidate agent signatures appear different?”

The candidate/candidate agent signatures can be found in the large table in the middle of the second page of the document. Each row of the table provides the name of the candidate or agent, a space for their signature, and a space for a reason for refusing to sign.

Code this variable as 1 if the signatures in the middle column of this table appear different from one another. Code this variable as 0 if the signatures appear similar. Figure 17 provides three examples of forms which should be coded as 0 and Figure 18 provides three examples of forms that should be coded as 1.

Candidates or Candidates' Agents		
	Name	Signature
1.	JOSÉPH MICHAEL MAKHETI	
2.	MARY WATHERA MUNGIKI	
3.	MILICENT WANDA GOMBAZA	
4.	JANE WANGOMA KOBAJA	
5.	LABATO GATHUMBI	
6.		

Candidates or Candidates' Agents		
	Name	Signature
1.	SALOME NDETA	
2.	ZIPORAH LUKIRINI	
3.		
4.		
5.		
6.		

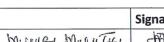
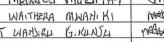
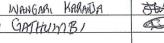
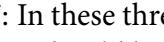
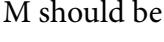
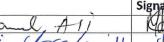
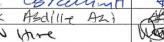
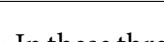
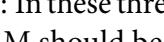
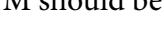
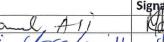
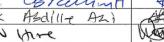
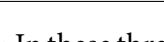
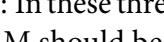
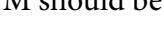
Candidates or Candidates' Agents		
	Name	Signature
1.	MICHAEL TUMITA	
2.	WANATE KOKATI	
3.	MINDA MACHKOP	
4.	STEPHEN TIERICE	
5.	James KARUOR	
6.	RONAL OLOKUNTA	

Figure 17: In these three examples, the candidate or candidate agent signatures appear similar. Column M should be coded as 0 for these cases.

Candidates or Candidates' Agents		
	Name	Signature
1.	Mohamed Ali	
2.	Nadeeb Ibrahim	
3.	Isaac Abdillahi Azi	
4.	Adow Kere	
5.	Abdulkadir Ahmed	
6.	Abdulhamid Mohamed	

Candidates or Candidates' Agents		
	Name	Signature
1.	PHILIP MUSOMI	
2.	KIYONDA MUSOMI	
3.	MOTAJI MUSOMI	
4.	PATRICIA X ASOU	
5.	ESTHER MUSOMI	
6.	SANET K. NOVITA	

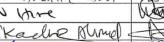
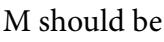
Candidates or Candidates' Agents		
	Name	Signature
1.	RAMESH K. KARAMJA	
2.	BABUA MUWAGIIDI	
3.	JAMES MUWABO	
4.	PETER K. MACHENGE	
5.	MARY Njeri Muriuki	
6.	ROBERT K. KARUA	

Figure 18: In these three examples, the candidate or candidate agent signatures appear different. Column M should be coded as 1 for these cases.

Column N: “Do the vote tallies or totals show any sign of crossing out or editing?”

Vote tallies can be found in the middle of the first page of the document, just beneath the bold text “The number of valid votes cast in favour of each candidate.” They are embedded in a table with 21 rows, where the left column consists of candidate names (e.g., “Uhuru Kenyatta”) and the right columns leave space for handwritten tallies (under “No. of Valid Votes Cast”). Vote totals can be found in the top-middle of the first page of the document, just beneath the polling station, stream, and constituency information. They are embedded in a table with 7 rows, where the left column consists of typed inputs (e.g., “Total number of votes cast”) and the right column leaves space for handwritten tallies.

Code this variable as 1 if the handwritten tallies or totals show any sign of having been crossed out or edited. Code this variable as 0 if none of the vote tallies or totals have any such signs. Figure 19 provides three examples of forms which should be coded as 1 because of bad tallies. Figure 20 provides three examples of forms which should be coded as 1 because of bad totals. Figure 21 provides three examples of good tallies, while Figure 22 provides three examples of good totals.

No. of Valid Votes Cast									
0									
0									
6									
6									
10									
11									
313									
128									
XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX

No. of Valid Votes Cast									
0									
2									
2									
8									
0									
7									
382									
160									
XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX

No. of Valid Votes Cast									
1									
2									
3	now								
5									
1									
4									
293									
63									
XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX

No. of Valid Votes Cast									
1									
0									
1									
2									
1									
5									
1									
9									
555									
XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX	XXXXXX

Figure 19: In these five examples, the vote tallies show signs of crossing out or editing. Column N should be coded as 1 for these cases.

1 545 - 01 140 439 edit 0 01 0 00 four hundred and forty eight (428)	682 1✓ 435 489 5✓ - Four hundred and forty eight four hundred and eighty four 484	ion 757 1 422 + 6 = 428 6 8 0 0 Four hundred and twenty two Three 426 e: Twenty six 426
---	--	--

Figure 20: In these three examples, the vote totals show signs of crossing out or editing. Column N should be coded as 1 for these cases.

No. of Valid Votes Cast									
									0
									0
									2
									7
									0
									0
									430
									68
No. of Valid Votes Cast									
									0
									1
									2
									6
									1
									5
									327
									119
No. of Valid Votes Cast									
									0
									1
									2
									7
									0
									2
									281
									145

Figure 21: In these five examples, the vote tallies show no signs of crossing out or editing. Column N should be coded as 0 for these cases if the vote totals also show no signs of crossing out or editing.

Example 1:

on	379
	1
	858
	0
	0
s)	(358) Three hundred and fifty eight
e:	

Example 2:

on	102
	0
	88
	2
	0
	0
	86 (Eighty six)

Example 3:

on	197
	0
	193
	1
	0
	0
)	One hundred and ninety two (192)
:	

Figure 22: In these three examples, the vote totals show no signs of crossing out or editing. Column N should be coded as 0 for these cases if the vote tallies also show no signs of crossing out or editing.

(End of instructions provided to research assistants)

A.2 Details about model estimation

A.2.1 Model architectures

We tune across five model architectures: {inception, average, wide, deep, hard}.¹ Note that throughout, we refer to a “block” as the combination of a 1-pixel empty border padding layer (which helps prevent the edges near the border getting ignored during training), a convolutional layer, a batch normalization, an exponential linear unit layer, and a pooling layer (in that order). The details of each model are as follows.

The inception model’s base model is the InceptionV3 model (Szegedy et al. 2016), trained on the ImageNet database (Deng et al. 2009). We take the top layer of this model (its output), add a pooling layer, a batch normalization layer, a dense layer with 1,024 exponential linear units (ELUs), and then a 50% dropout layer before the activation layer using a sigmoid function. We train the model such that all layers are trainable, including those in the underlying InceptionV3 model.²

The “average” model’s feature extraction step consists of three layers: a block in which the convolutional step has 32 filters, another block with 64 filters, and a third block with 128 filters. The prediction step is also grouped into three layers. In the first fully-connected layer we have a 10% dropout step, a flattening step, a dense step with 1,024 ELUs, a batch normalization, and an activation step. In the second fully-connected layer we have another 10% dropout step followed by a dense step with 512 ELUs, batch normalization, and then activation. The final layer is another 10% dropout, a dense step reducing to the final predictions, which are then batch normalized and activated using the sigmoid function.

The “wide” model builds on the average model. The first difference is that the feature extraction step includes a fourth block where the convolutional step has 256 filters. The second difference is that in the prediction step, the number of ELUs in the first fully-connected layer is 4,096 (still reducing down to 512).

The “deep” model differs from the average model in that the feature extraction step has 7 blocks, wherein the convolutional layers have 16, 16, 16, 16, 32, 32, and 64 filters. The prediction step has three fully-connected layers instead of the average model’s two, with 2,048, 1,024, and 256 ELUs, respectively.

Finally, the “hard” differs from the average model in that the feature extraction step has 5 blocks, wherein the convolutional layers have 32, 64, 64, 128, and 256 filters. The prediction step has three fully-connected layers instead of the average model’s two, with 4,096, 2,048, and 512 ELUs, respectively.

1 . The inception model is named for the base model we use; average because it is a reasonable starting point for a classification task of average difficulty; wide because the model appears wide if visualized vertically (due to the large number of ELUs in the first fully-connected layer); deep because the model appears tall if visualized vertically (due to the number of blocks in the feature extraction step); and hard because it is a complex model for difficult classification tasks.

2 . Usually only the top few layers of a transfer learning model need to be retrained, with the rest kept “frozen.” However, since our application is somewhat different than typical object-detection tasks—e.g., we are looking for blank spaces within a table instead of identifying faces—our models perform better when all layers are kept trainable.

A.2.2 Data augmentations

All images, irrespective of data augmentation, undergo the following preprocessing steps. First, images are cropped to the region of the form where the relevant feature is consistently found. For instance, the QR code is always in the top-left corner of the form, so we automatically zoom in on this region for the QR code irregularity. Second, the cropped images are resized to be square, so that the pixel width and pixel height are equal (the precise size is determined by the tuning parameters discussed below). This step increases computational stability and is standard in deep learning models. Third, all images are denoised using a median filter. This technique reduces noise while preserving edges better than other filters, e.g. a Gaussian filter, do. Finally, image pixel values are rescaled from RGB coordinates ranging over [0, 255] so that they are expressed as a proportion between 0 and 1.

Besides these preprocessing steps, we also explored data augmentation techniques. We emphasize that only one of the tuned models studied in the manuscript (“form not stamped”) uses data augmentation, and removing it does not substantially impact performance or our regression results; none of our findings depend on augmented data. However, since they are part of our tuning procedure, the details are provided here.

The `Python` module we use to fit our deep learning modules, `keras` (Chollet 2015), provides a number of built-in functions for augmentation through its `ImageDataGenerator` class. The augmentations relevant to our data and model are as follows:

1. `samplewise_center`: Set the global mean across all pixels to 0.
2. `samplewise_std_normalization`: Divide all pixels by the standard deviation across all pixels.
3. `horizontal_flip`: Flip the image horizontally with 50% probability.
4. `vertical_flip`: Flip the image vertically with 50% probability.
5. `brightness_range`: Shift the brightness of all pixels by a random value within the range given.
6. `zoom_range`: Zoom in or out of the image by a random value up to the value given.
7. `channel_shift_range`: Shift red, blue, and green channels of all pixels by a random value up to the value given.
8. `rotation_range`: Rotate the image in either direction by a random value up to the value given.
9. `width_shift_range`: Shift the image horizontally by a random value up to the value given.
10. `height_shift_range`: Shift the image vertically by a random value up to the value given.
11. `shear_range`: Shear the image (i.e., fix one axis and rotate the other axis) by a random value up to the value given.

For all augmentations that reveal pixels outside the existing boundaries (e.g., by shifting or rotating the image), empty pixel values are filled using the value of the nearest neighbor, the default setting in `keras`.

For the purposes of tuning our models, we use three settings: {none, low, high}. In the none condition, none of these data augmentations are applied. In the low setting, we turn on samplewise centering, samplewise scaling, horizontal flipping, and vertical flipping; set the brightness range and zoom range to [0.95, 1.05]; set the channel shift range to 5; and set the rotation range, width shift range, height shift range, and shear range to [-0.05, 0.05]. In the high setting, we turn on samplewise centering, samplewise scaling, horizontal flipping, and vertical flipping; set the brightness range and zoom range to [0.70, 1.30]; set the channel shift range to 30; and set the rotation range, width shift range, height shift range, and shear range to [-0.30, 0.30].

We note here that `keras`'s data augmentation is not replication-safe because it does not respect random seeds set at the beginning of a script call. We therefore implement a custom `ImageDataGenerator` class and a custom `Generator` sequence to call the class, bypassing this problem. This code can be found in our replication archive.

A.2.3 Tuning procedures

For each of our nine irregularities, we tune over the 60 models in a $\{5 \times 3 \times 2 \times 2\}$ grid. This grid consists of:

- Model architecture: inception, average, wide, deep, hard;
- Data augmentation: none, low, high;
- Image size: 256 x 256 pixels, 512 x 512 pixels; and
- Batch size: 16, 32.

These 540 models were run using `slurm` arrays on Cardiff's high-performance computing cluster, Supercomputing Wales, the national supercomputing research facility for Wales. Each model was allowed to run until either: its out-of-sample F_1 achieved 0.99, it reached 300 epochs, or it timed out after 48 hours (the maximum run-time allowed on the cluster).

A.2.4 Best-performing models

Table A1 provides the parameters that maximized predictive performance for each of the nine irregularities.

Table A1: Hyperparameters of the best-fitting models for each irregularity

Irregularity	Architecture	Augmentation	Image size	Batch size	Epochs
QR code missing	deep	none	256	16	20
Poor scan quality	deep	none	512	16	20
Forms not stamped	inception	little	256	32	30
Presiding officer did not sign	deep	none	512	16	30
No agents listed	deep	none	256	16	20
Any agent did not sign	inception	none	256	32	40
No agents signed	deep	none	512	32	20
Agent signatures identical	deep	none	512	32	20
Results edited	inception	none	512	16	30

A.3 Confusion matrices, classification model training

Below are confusion matrices from our image classification models. Each is constructed using the true test sample of approximately 500 images left out from model training. Note that some samples are slightly smaller than 500 due to data missingness (i.e., missing, partially missing, corrupted, or otherwise incomplete statutory forms).

Table A2: Confusion matrix, QR code missing

	True non-irregularity	True irregularity
Predicted non-irregularity	493	1
Predicted irregularity	1	5

Table A3: Confusion matrix, poor scan quality

	True non-irregularity	True irregularity
Predicted non-irregularity	497	0
Predicted irregularity	0	0

Table A4: Confusion matrix, form not stamped

	True non-irregularity	True irregularity
Predicted non-irregularity	476	8
Predicted irregularity	4	10

Table A5: Confusion matrix, presiding officer did not sign

	True non-irregularity	True irregularity
Predicted non-irregularity	491	3
Predicted irregularity	0	4

Table A6: Confusion matrix, no agents listed

	True non-irregularity	True irregularity
Predicted non-irregularity	484	2
Predicted irregularity	1	10

Table A7: Confusion matrix, any agent did not sign

	True non-irregularity	True irregularity
Predicted non-irregularity	466	5
Predicted irregularity	5	21

Table A8: Confusion matrix, no agents signed

	True non-irregularity	True irregularity
Predicted non-irregularity	483	0
Predicted irregularity	0	14

Table A9: Confusion matrix, agent signatures appear identical

	True non-irregularity	True irregularity
Predicted non-irregularity	481	1
Predicted irregularity	1	14

Table A10: Confusion matrix, results edited

	True non-irregularity	True irregularity
Predicted non-irregularity	321	16
Predicted irregularity	8	155

A.4 Computation of F_1 scores for Cantú (2019)

Table 1 in Cantú (2019) provides a confusion matrix with mean values across 20 samples, each of 150 images. We can approximate the F_1 score across this sample by multiplying each cell in the table by 3,000 (20×150). This gives 2,790 true negatives (top-left), 210 false positives (top-right), 450 false negatives (bottom-left), and 2,550 true positives (bottom-right). Applying the formula supplied in the main text (our paper), the approximate precision is 0.924, while the approximate recall is 0.850. F_1 is then given by the harmonic mean:

$$F_1 \text{ (Cantú)} = 2 \left[\frac{(0.924)(0.850)}{0.924 + 0.850} \right] \approx 0.885.$$

A.5 Empirical distribution of irregularities

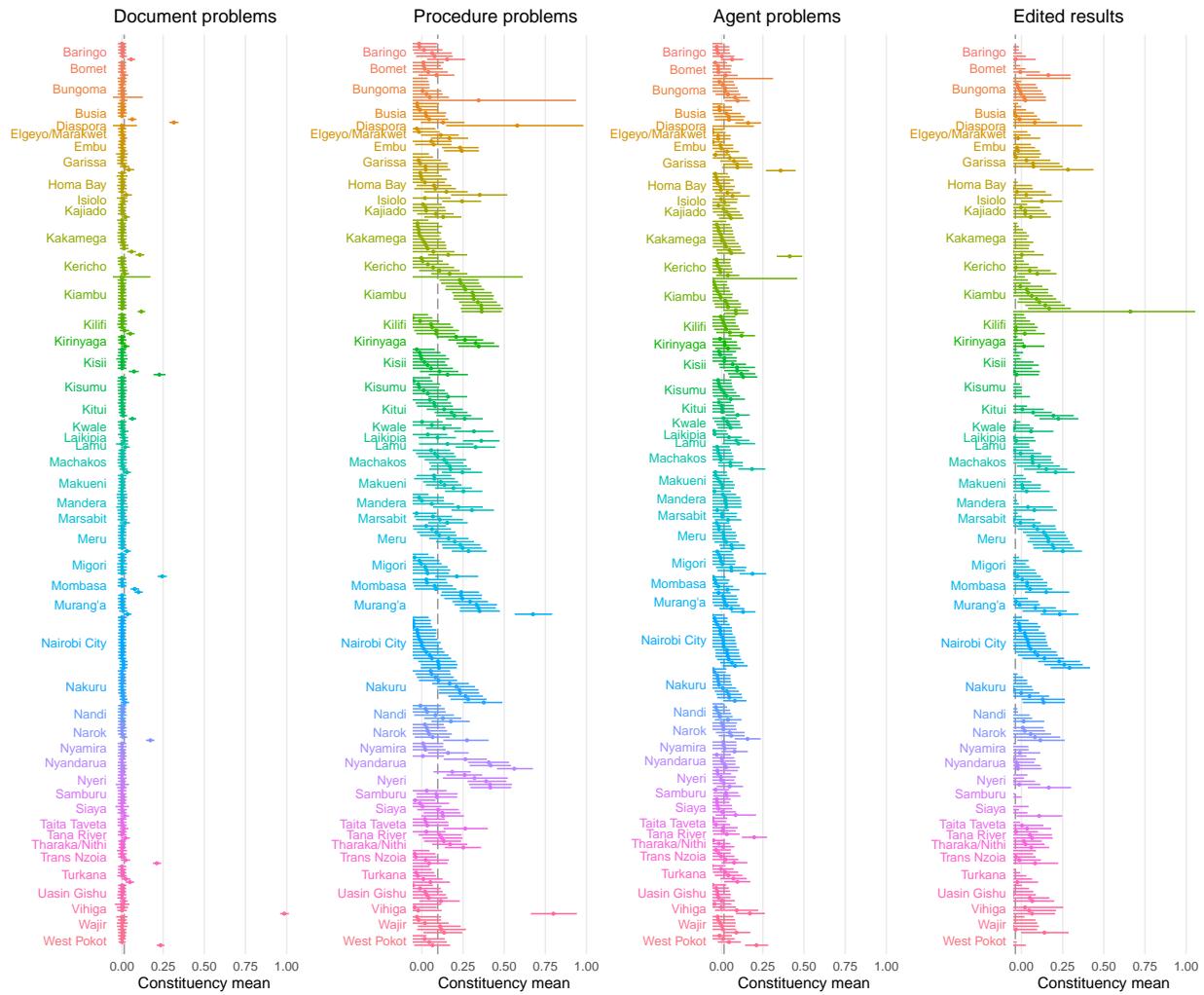


Figure A1: Frequency of electoral irregularities by constituency. Each pane refers to a type of irregularity: document problems, procedure problems, agent problems, and edited results. Each dot is the estimate for the constituency's fixed effect from a null model, with lines for 95% confidence intervals. Colors correspond to counties.

A.6 Covariate data sources

Population density data come from Linard et al. (2012), and is computed as the log plus one of raw population density. Data on terrain ruggedness are calculated from a digital elevation model and data provided by the United States Geological Survey (2004). Ethnic fractionalization is computed as one minus the sum of squares of the size of each ethnic group expressed as a proportion of registered voters; these proportions are estimated from the 2013 voter register using the method described in Harris (2015). Polling station isolation is computed as the minimum distance to the nearest polling center, weighted by the number of registered voters, using a method described in Doogan et al. (2018). Data on the poverty rate are described in Linard et al. (2012), night time lights data are from National Oceanic and Atmospheric Administration (2013), and literacy rate data are given in Bosco et al. (2017).

A.7 Full results for Figure 3 (in main text)

Table A11: Full results from Figure 3 in the main text

	Document prob.	Procedure prob.	Agent prob.	Results edited
Gov't stronghold	0.00 (0.00)	-0.04 (0.04)	0.01 (0.03)	-0.02 (0.05)
Opp. stronghold	0.00 (0.01)	0.15* (0.05)	0.05 (0.03)	-0.32* (0.05)
Population density	0.00* (0.00)	0.01 (0.01)	-0.00 (0.00)	0.03* (0.01)
Terrain ruggedness	-0.00 (0.00)	-0.01 (0.00)	-0.00 (0.00)	0.00 (0.00)
Ethnic frac.	-0.00 (0.00)	0.03 (0.02)	0.01 (0.01)	0.05* (0.02)
Poll. station isolation	-0.01 (0.02)	-0.11 (0.19)	0.17 (0.11)	0.08 (0.18)
Poverty rate	0.00 (0.00)	0.00 (0.01)	-0.01* (0.01)	-0.03* (0.01)
Literacy rate	-0.00 (0.00)	-0.01 (0.01)	0.00 (0.01)	-0.03* (0.01)
Night lights	0.00 (0.00)	-0.01 (0.01)	0.00 (0.00)	0.01 (0.01)

* $p < .05$. All models include constituency fixed effects and cluster standard errors by constituency.

A.8 Robustness to alternative definitions of strongholds

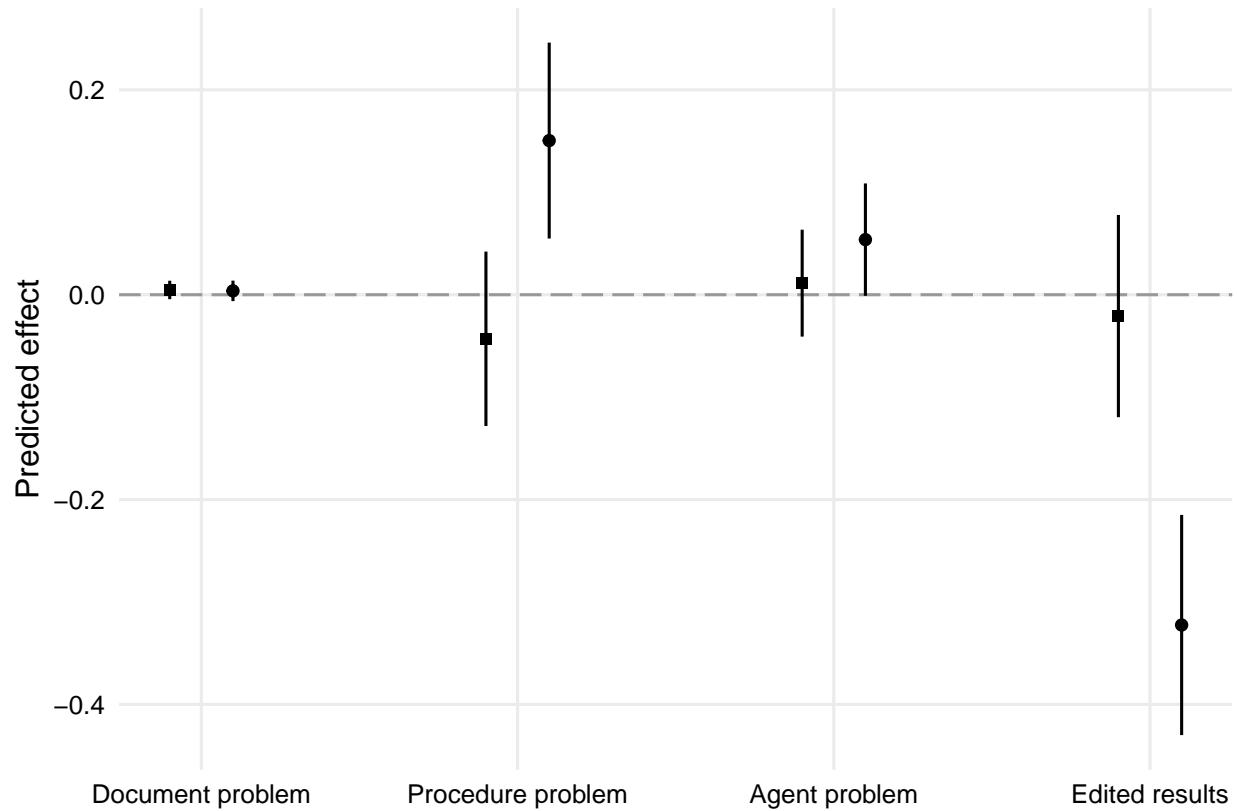


Figure A2: Frequency of electoral irregularities by polling station partisanship. Squares indicate government strongholds, with circles for opposition strongholds, relative to the baseline of competitive areas. Lines indicate 95% confidence intervals. Here, government and opposition strongholds are defined as polling stations in which at least 75% of the vote went to Kenyatta or Odinga, respectively.

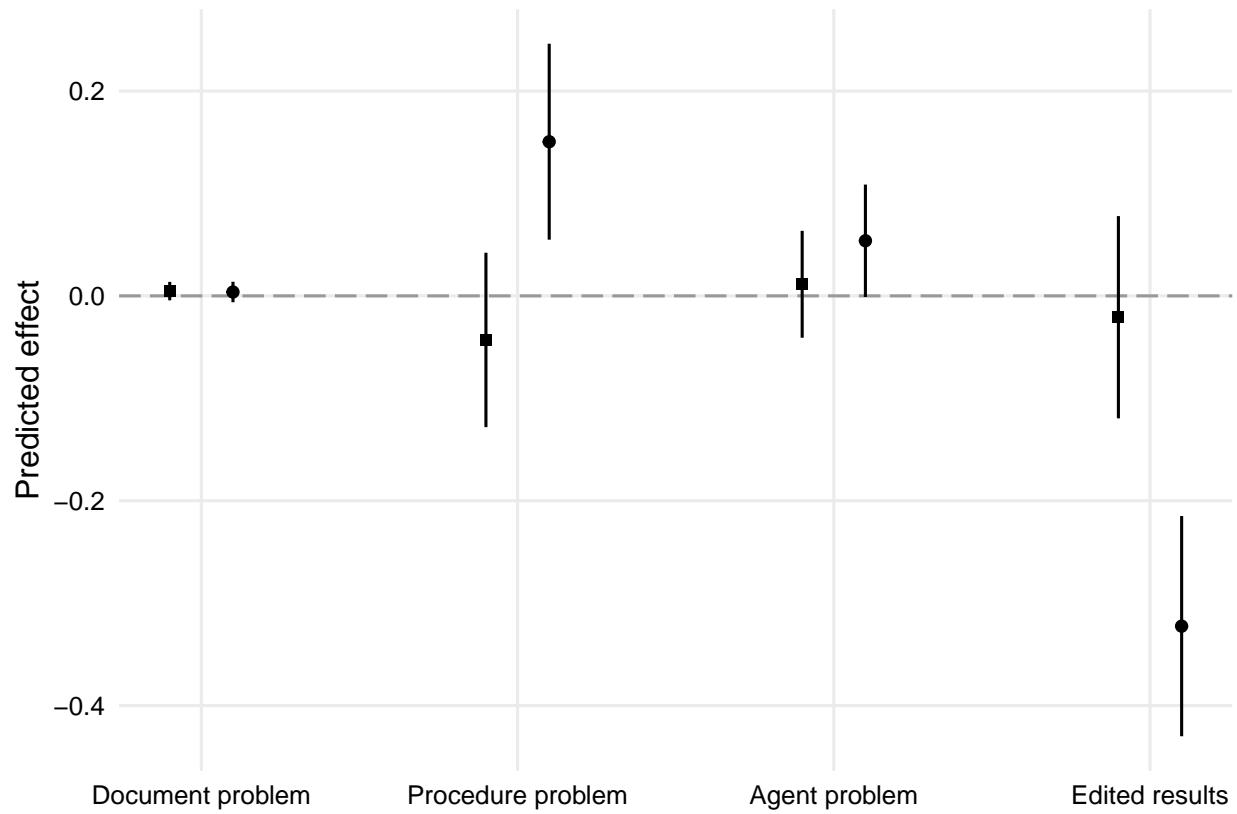


Figure A3: Frequency of electoral irregularities by polling station partisanship. Squares indicate government strongholds, with circles for opposition strongholds, relative to the baseline of competitive areas. Lines indicate 95% confidence intervals. Here, government and opposition strongholds are defined as polling stations in which at least 85% of the vote went to Kenyatta or Odinga, respectively.

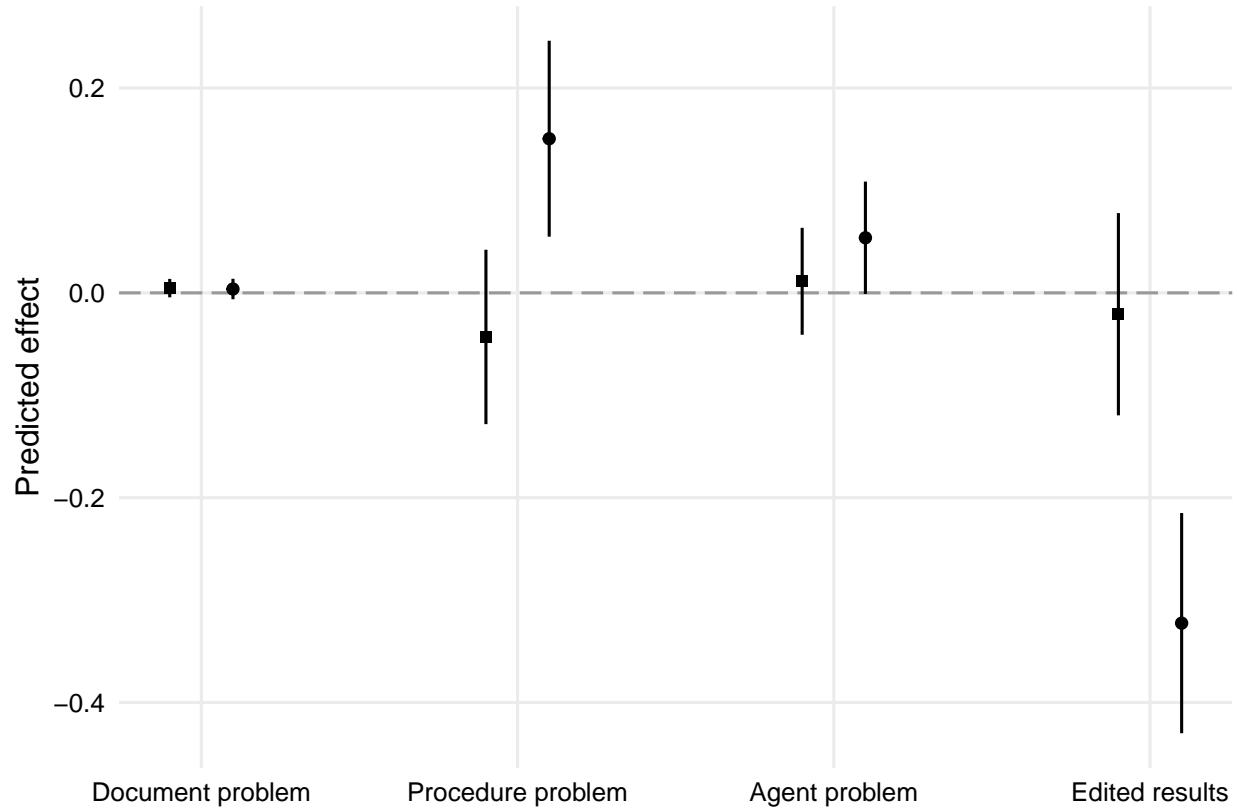


Figure A4: Frequency of electoral irregularities by polling station partisanship. Squares indicate government strongholds, with circles for opposition strongholds, relative to the baseline of competitive areas. Lines indicate 95% confidence intervals. Here, government and opposition strongholds are defined as polling stations in counties with at least 80% Kikuyu and Kalenjin or Luo and Kamba voters, respectively.

A.9 Full regression results, Table 3

Table A12: Full results for Table 3 in the main text, government strongholds

	Doc. prob.	Proc. prob.	Agent prob.	Results edited	Any prob.
Observed	-0.00*	0.01	-0.01	-0.03	-0.05
	(0.00)	(0.02)	(0.01)	(0.02)	(0.03)
Population density	0.00	0.03	-0.01	0.04	0.02
	(0.00)	(0.03)	(0.01)	(0.03)	(0.04)
Terrain ruggedness	-0.00	-0.00	0.00	0.02	0.00
	(0.00)	(0.02)	(0.01)	(0.02)	(0.02)
Ethnic frac.	-0.00	0.03	-0.03	0.06	0.04
	(0.00)	(0.07)	(0.03)	(0.07)	(0.08)
Poll. station isolation	-0.03	1.53	-0.83	-2.87*	-2.45
	(0.03)	(1.50)	(0.78)	(1.35)	(1.69)
Poverty rate	0.00	0.01	-0.01	0.02	0.03
	(0.00)	(0.04)	(0.03)	(0.04)	(0.05)
Night lights	0.00	0.02	-0.00	0.02	0.01
	(0.00)	(0.03)	(0.02)	(0.03)	(0.03)
Literacy rate	-0.00	-0.06	0.00	-0.02	-0.01
	(0.00)	(0.05)	(0.02)	(0.05)	(0.05)

* $p < .05$. All models include constituency fixed effects and cluster standard errors by constituency.

Table A13: Full results for Table 3 in the main text, opposition strongholds

	Doc. prob.	Proc. prob.	Agent prob.	Results edited	Any prob.
Observed	−0.00 (0.00)	0.02 (0.03)	0.02 (0.02)	0.01 (0.03)	0.02 (0.04)
Population density	0.00 (0.00)	0.07 (0.04)	−0.01 (0.03)	−0.03 (0.04)	−0.04 (0.05)
Terrain ruggedness	0.00 (0.00)	0.00 (0.02)	0.02* (0.01)	−0.02 (0.02)	−0.01 (0.02)
Ethnic frac.	−0.01 (0.01)	0.24* (0.09)	0.18 (0.11)	0.22 (0.12)	0.32* (0.11)
Poll. station isolation	0.03 (0.02)	2.14* (0.42)	−0.40 (0.33)	2.53* (0.51)	1.66* (0.56)
Poverty rate	−0.00 (0.00)	0.03 (0.04)	0.02 (0.03)	0.03 (0.04)	0.02 (0.05)
Night lights	0.00 (0.00)	−0.02 (0.04)	0.02 (0.04)	0.02 (0.04)	0.01 (0.04)
Literacy rate	−0.00 (0.00)	−0.01 (0.05)	0.01 (0.05)	−0.00 (0.06)	0.07 (0.05)

* $p < .05$. All models include constituency fixed effects and cluster standard errors by constituency.

Table A14: Full results for Table 3 in the main text, competitive areas

	Doc. prob.	Proc. prob.	Agent prob.	Results edited	Any prob.
Observed	0.00 (0.00)	0.02 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.00 (0.02)
Population density	0.01 (0.00)	0.04 (0.03)	-0.02 (0.02)	0.05 (0.03)	0.05 (0.03)
Terrain ruggedness	-0.00 (0.00)	0.01 (0.02)	-0.01 (0.01)	0.01 (0.02)	0.00 (0.03)
Ethnic frac.	0.01 (0.01)	0.08 (0.05)	-0.01 (0.04)	-0.03 (0.04)	-0.04 (0.06)
Poll. station isolation	-0.00 (0.02)	0.65 (0.65)	0.44 (0.38)	0.14 (0.72)	1.19 (0.79)
Poverty rate	0.00 (0.00)	-0.03 (0.04)	-0.02 (0.02)	-0.03 (0.04)	-0.09* (0.04)
Night lights	-0.00 (0.00)	-0.02 (0.03)	-0.00 (0.02)	-0.02 (0.03)	-0.03 (0.04)
Literacy rate	-0.00 (0.00)	-0.02 (0.02)	0.02* (0.01)	-0.06* (0.02)	-0.05 (0.03)

* $p < .05$. All models include constituency fixed effects and cluster standard errors by constituency.

Table A15: Full results for Table 3 in the main text, all polling stations

	Doc. prob.	Proc. prob.	Agent prob.	Results edited	Any prob.
Observed	-0.00 (0.00)	0.02 (0.01)	-0.00 (0.01)	-0.01 (0.01)	-0.01 (0.02)
Population density	0.00* (0.00)	0.04* (0.02)	-0.01 (0.01)	0.03 (0.02)	0.02 (0.02)
Terrain ruggedness	-0.00 (0.00)	0.00 (0.01)	0.00 (0.01)	-0.00 (0.01)	-0.01 (0.01)
Ethnic frac.	0.00 (0.00)	0.10* (0.04)	0.02 (0.03)	0.04 (0.04)	0.04 (0.05)
Poll. station isolation	0.00 (0.01)	1.26* (0.55)	0.08 (0.27)	0.72 (0.75)	1.08 (0.62)
Poverty rate	0.00 (0.00)	-0.00 (0.03)	-0.01 (0.01)	0.00 (0.02)	-0.02 (0.03)
Night lights	-0.00 (0.00)	0.00 (0.02)	0.00 (0.01)	0.01 (0.02)	0.00 (0.02)
Literacy rate	-0.00 (0.00)	-0.03 (0.02)	0.01 (0.01)	-0.04 (0.02)	-0.02 (0.02)

* $p < .05$. All models include constituency fixed effects and cluster standard errors by constituency.

A.10 Regression discontinuity analysis

In 2013, Kenya's election commission had a soft cap of 800 voters per polling station. We can use this administrative rule to estimate the (plausibly) causal effect of creating a new polling station on the likelihood of experiencing Form 34A irregularities via a regression discontinuity design. By comparing error rates in polling places with one station (and just fewer than 800 voters) to those with two stations (and just over 800 voters) we can examine whether changes in workload at the polling station affect the error rate at that polling station. A presiding officer (PO) in a polling station with 800 registered voters would face a more burdensome administrative task than a PO at a 401 registered voter polling station, and so we expect that this reduction in workload would lead to fewer errors.

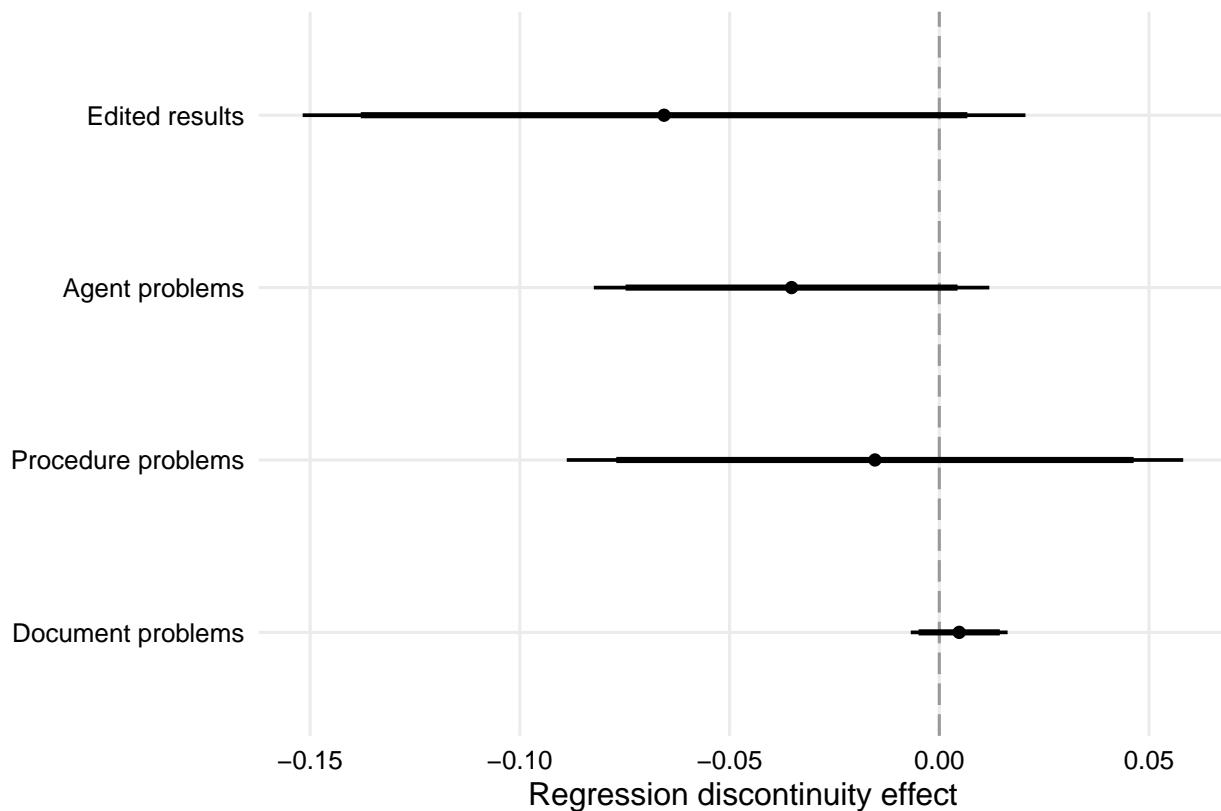


Figure A5: Regression discontinuity design analysis. Each dot represents the estimated effect of a polling station being just above the threshold of 800 voters for each type of irregularity. Thick lines represent 90% confidence intervals, with thin lines for 95% confidence intervals.

Estimates from these models are plotted in Figure A5. Consistent with the “fumble” hypothesis, our estimates suggest that edited results and agent problems are less likely in polling places with just over 800 voters, though the results just barely miss statistical significance. This evidence provides additional support for the results presented in the main analysis, which suggest that irregularities most typically arise through human error endemic in large-scale election administration tasks.

References

- Bosco, Claudio, Victor Alegana, Tomas Bird, Carla Pezzulo, Linus Bengtsson, Alessandro Sorichetta, Jessica Steele, et al. 2017. “Exploring the High-Resolution Mapping of Gender-Disaggregated Development Indicators.” *Journal of the Royal Society Interface* 14.
- Cantú, Francisco. 2019. “The Fingerprints of Fraud: Evidence from Mexico’s 1988 Presidential Election.” *American Political Science Review* 113 (3): 710–726.
- Chollet, Fran ois. 2015. Keras. <https://keras.io>.
- Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- Doogan, Nathan J., Megan E. Roberts, Mary Ellen Hewers, Erin R. Tanenbaum, Elizabeth A. Mumford, and Frances A. Stillman. 2018. “Validation of a New Continuous Geographic Isolation Scale: A Tool for Rural Health Disparities Research.” *Social Science & Medicine* 215:123–132.
- Harris, J. Andrew. 2015. “What’s in a Name? A Method for Extracting Information About Ethnicity From Names.” *Political Analysis* 23 (2): 212–224.
- Linard, Catherine, Marius Gilbert, Robert W. Snow, Abdisalan M. Noor, and Andrew J. Tatem. 2012. “Population Distribution, Settlement Patterns, and Accessibility across Africa in 2010.” *PLOS ONE* 7 (2).
- National Oceanic and Atmospheric Administration. 2013. DMSP-OLS Nighttime Lights Time Series Version 4, Stable Lights Product. Image and data processing by NOAA’s National Geophysical Data Center. DMSP data collected by US Air Force Weather Agency.
- Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. “Rethinking the Inception Architecture for Computer Vision.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.
- United States Geological Survey. 2004. USGS SRTM 30M Digital Elevation Model. Available at <https://www.usgs.gov>. Last accessed 11 July 2020.