An Introduction to Event Data

Zachary C. Steinert-Threlkeld

American Political Science Association

09.08.2020

Purpose

- Record dyadic interaction: Actor 1 → Event → Actor 2.
- Active area of development since $\approx$ 1970.
- Source is text, usually newspapers.

Types of Events

Event data are especially common in the study of contentious politics.

1. Violence: riots; armed attacks; lynching; police behavior.
2. Non-violence: protests; labor action; verbal statements.

Types of Events

Event data are especially common in the study of contentious politics.

1. Violence: riots; armed attacks; lynching; police behavior.
2. Non-violence: protests; labor action; verbal statements.

The main coding ontology, Conflict and Mediation Event Observations (CAMEO), includes 20 event types.

Outline

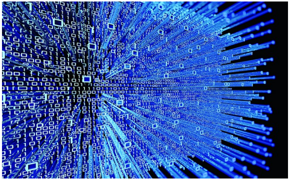
Approach	Example	Bias	Deduplication	Geolocation	Global Coverage	Magnitude	Protester Attributes
Text, Hand-coding	CCC, SCAD, NAVCO	✓	✓+	✓+	✓-	✓	✓
Text, Machine-coding	GDELT, ICEWS, TERRIER	✓	✓-	✓-	✓+	✓-	✓-
Text, Hybrid	SPEED, ACLED, UCDP, MM, MMAD	✓	✓+	✓+	✓	✓	✓

Text + Processing

Approach	Example	Bias	Deduplication	Geolocation	Global Coverage	Magnitude	Protester Attributes
Text, Hand-coding	CCC, SCAD, NAVCO	✓	✓+	✓+	✓-	✓	✓
Text, Machine-coding	GDELT, ICEWS, TERRIER	✓	✓-	✓-	✓+	✓-	✓-
Text, Hybrid	SPEED, ACLED, UCDP, MM, MMAD	✓	✓+	✓+	✓	✓	✓

Three Methodologies

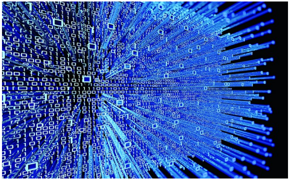
(1) Text + Manual



(2) Text + Machine

Three Methodologies

(1) Text + Manual



(2) Text + Machine

(3) Text + Hybrid

Structural Limitations with Text

Two limitations for text, regardless of dataset.

1. Difficulty measuring magnitude (size, violence) and protester attributes.

Structural Limitations with Text

Two limitations for text, regardless of dataset.

1. Difficulty measuring magnitude (size, violence) and protester attributes.
2. Bias due to space constraints*profit motive.

> *It's amazing that the amount of news that happens in the world every day always just exactly fits the newspaper. - Jerry Seinfeld*

Bias, Text

When relying on newspapers, three types of bias arise:

1. Selection (large, violent)
2. Location (urban)
3. Fatigue (novelty, coder)

See: Gobel, Christian and H. Christoph Steinhardt. 2019. "How should we measure protests in authoritarian regimes? A comparison of traditional, dissident, and social media event data from China." *Working paper.*

Duplicate Events, Text

- Early machine-coded event datasets reported the same event multiple times. New iterations of ICEWS and GDELT have improved, but duplication still exists.
- Not an issue for hand-coded or hybrid event datasets.

Geolocation

Location is hard for machine datastes, easy otherwise.

Global Coverage

Global coverage is difficult in almost all three methodologies.

- Text
  - Best: machine-coded.
  - Worst: hand-coded.

Global Coverage

Global coverage is difficult in almost all three methodologies.

- Text
  - Best: machine-coded.
  - Worst: hand-coded.
  - Tradeoff between global coverage and use of local sources.

Magnitude

Text provides coarse measurement of magnitude.

1. Size - only in manual and hybrid datasets; sourcing unclear.
2. State and protester violence - binary or ordinal variable.

Protester Attributes

No existing text-based event dataset on protester attributes.

1. Hand-coded event datasets at the movement level. MMAD at the report level.

Scope

This section, like the presentation, focuses on datasets that include some type of collective action as an event.

This choice is only about my research agenda, not about the validity.

This curating means I avoid some datasets focused on violence, like Uppsala Conflict Data or Global Terrorism Database, or the United States, like Dynamics of Collective Action.

Defunct - Conflict and Peace Database (COPDAB)

- Method: manual.
- Years: 1948-1978.
- Coverage: 135 countries, 500,000 international and domestic events.
- Sources: Newspapers, "computer-based".
- Available at ICPSR. See Azar's *JCR* article.

Defunct - World Event/Interaction Survey (WEIS)

- Method: manual.
- Years: 1966-1978.
- Coverage: Any country; about 98,000 events.
- Source: *New York Times*.
- Available from the ICPSR.

Defunct - Computational Event Data System (CEDS, KEDS)

The first automatic event data system available to academics.

- Method: automatic.
- Years: 1990-2015.
- Coverage: Primarily the Middle East; later, SE Europe and West Africa.
- Source: Reuters
- Philipp Schrodt's documentation.

Global Database of Events, Location, and Tone

The child of CEDS.

- Method: Automatic.
- Years: 1979-Present.
- Coverage: All countries, all days.
- Sources: "All global news [...] in 65 languages."
- Project website.

Because of rampant duplicate events, it is better thought of as a dataset of media coverage.

Integrated Conflict Early Warning System

The half-sibling of CEDs. Developed with Lockheed Martin for about $45 million.

- Method: Automatic.
- Years: 1991-Present, released with a 12 month delay.
- Coverage: All countries.
- Sources: Major international and local news sources, 250, translated into English.
- Available at Harvard's Dataverse.

The best automatic event dataset.

Social Conflict Analysis Database (SCAD)

- Method: Manual
- Years: 1990-2017
- Coverage: Africa, Mexico, Centra America, and the Caribbean
- Sources: Agence France Press, the Associated Press.
- Available from the Strauss Center.

Nonviolent and Violent Campaigns and Outcomes

Three distinct datasets, though versions 1 and 2 are very similar, focused on maximalist campaigns.

NAVCO 1.0

- Method: Manual
- Years: 1900-2019; no time variation.
- Coverage: Global
- Source: Heterogenous.

NAVCO 2.0

- Method: Manual.
- Years: 1945-2006*; campaign-year as unit of analysis.
- Coverage: Global.
- Source: Heterogenous.

* Will soon have campaigns through 2019.

Nonviolent and Violent Campaigns and Outcomes v3.0

Think of it like a manual ICEWS (same schema, one extra event type) with an additional focus on tactics and organizations within movements.

- Method: Manual
- Years: 1991-2012, "politically relevant events". Daily.
- Coverage: 26 countries.
- Source: Agence France Presse.

Also useful for studying event magnitude.

Crowd Counting Consortium

- Method: Hybrid
- Years: 2017-Present, released with a one month delay.
- Coverage: Primarily the United States, with some international solidarity events.
- Source: Websites for local newspapers and television stations.

The most comprehensive dataset for event size. Great location detail as well.

Mass Mobilization

Related to the CIA's Political Instability Task Force.

- Method: Manual.
- Years: 1990 - March 2017.
- Coverage: 162 countries.
- Sources: *New York Times*, *Washington Post*, *Christian Science Monitor*, *Times of London*, *Jerusalem Post*.
- Available at Harvard's Dataverse.

Mass Mobilization in Autocracies

Designed to understand the internet's effect on mobilization in autocracies.

- Method: Hybrid.
- Years: 2003-2012.
- Coverage: 68 countries.
- Sources: *Associated Press*, *Agence France Presse*, *BBC Monitoring*, *BBC Monitoring*.
- Available here.

Armed Conflict Locations and Events Dataset (ACLED)

- Method: Hybrid
- Years: 1997-Present, though start year varies by region.
- Coverage: Basically everywhere except for Europe and the U.S. plus Canada.
- Sources: No official list, but very diverse, including not newspapers.
- https://acleddata.com/#/dashboard

Image as Data

Strength of images:

1. Fewer constraints

Image as Data

Strength of images:

1. Fewer constraints → more of them[1] → less location, selection bias.

Image as Data

Strength of images:

1. Fewer constraints → more of them[1] → less location, selection bias.
2. Measurement: better for existing variables (magnitude), new variables (protester attributes).

1. Assumes from social media.

Duplicate Events, Images

- Removing duplicate images is easy (similar to plagiarism detection software).
- Especially powerful with geotagged images: assume separate cities and unique images are different events.

Duplicate Events, Images

- Removing duplicate images is easy (similar to plagiarism detection software).
- Especially powerful with geotagged images: assume separate cities and unique images are different events.

Deduplicating images is a promising entree into political communication literatures: bias in newspapers and social media, Rashomon effect, virality, &c.

Text (Emotions)

Why? *Emotions can cause or stymie mobilization, perhaps explaining the repression-dissent puzzle.*

Mobilizing: anger, joy, pride.

Demobilizing: fear, sadness, shame.

Data - Text

1. Hybrid: MMAD (Mass Mobilization in Autocracies Dataset)
2. Machine-coded: ICEWS (International Conflict Early Warning System)
3. Machine-coded: TERRIER (Temporally Extended, Regular, Reproducible Interantional Event Records)

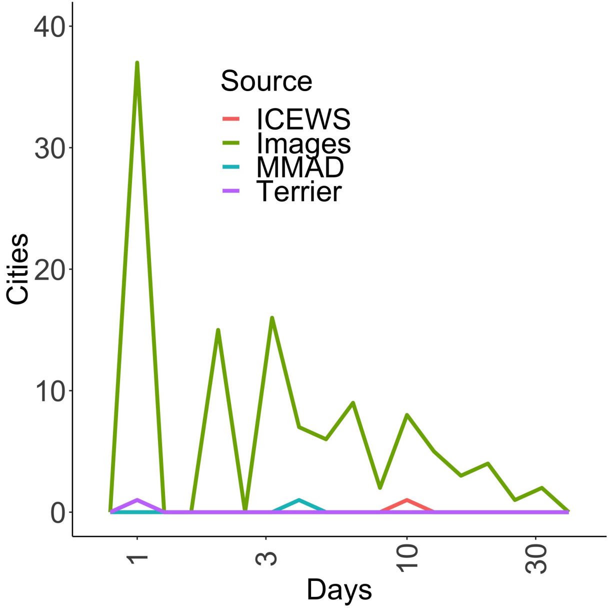
Images - Geolocated Tweets

- Have downloaded (and continue to) approximately 5 million tweets per day with latitude and longitude coordinates.
- Search specific countries, dates with known protests (**A**).
- Download all images from **A**.
- Use image classifier to identify protest images (**B**).

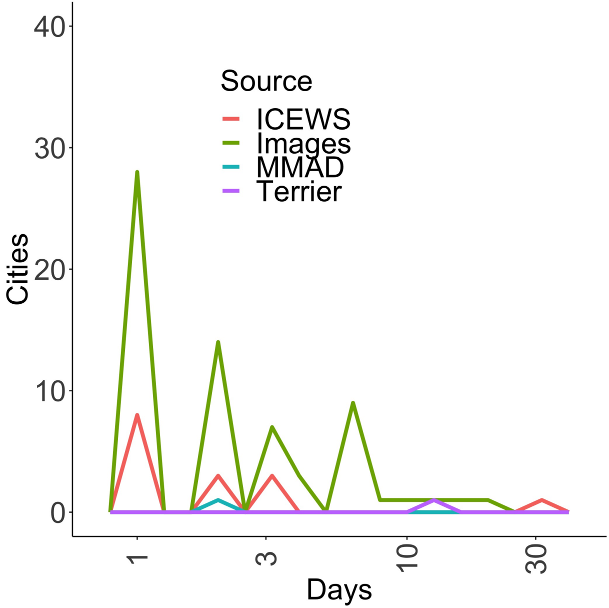
Images - Geolocated Tweets

- Have downloaded (and continue to) approximately 5 million tweets per day with latitude and longitude coordinates.
- Search specific countries, dates with known protests (**A**).
- Download all images from **A**.
- Use image classifier to identify protest images (**B**).
  - Use separate scene and face classifiers to identify magnitude and demographics in **B**.
  - Use Mohammad and Turney 2013. corpus in `syuzhet` to label tweets as containing anger or joy, fear or sadness.

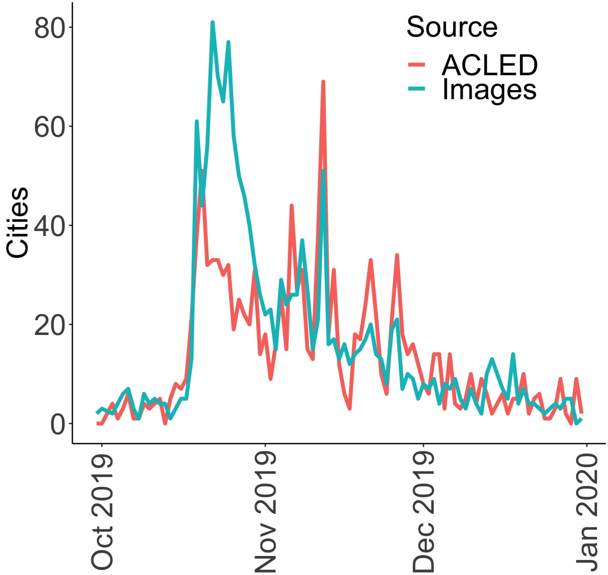
Venezuela Protest



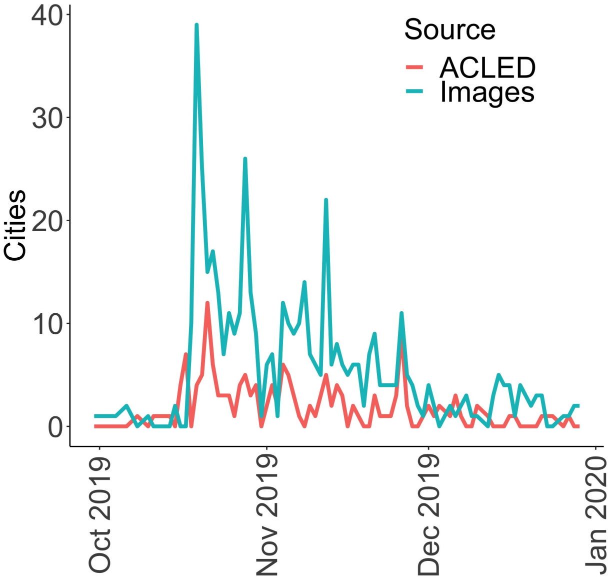
Venezuela Repression



Chile Protest



Chile Repression



Chile - Violence, Demographics, and Emotions

	Δ $\text{Log}_{10}(\text{Total Faces})_{i,t}$		
	Image	Text	Multimodal
ΔProtester Violence$_{i,t}$	$-.0499^{***}$		$-.0511^{***}$
	(.0192)		(.0191)
ΔState Violence$_{i,t}$	.3216		.3342
	(.2097)		(.2085)
ΔState Violence$^2_{i,t}$	$-.5352$		$-.5459$
	(.3616)		(.3596)
ΔFatalities$_{i,t}$	$-.0039$		$-.0036$
	(.0055)		(.0054)
Δ%Faces Male$_{i,t}$	$.5350^{***}$		$.5350^{***}$
	(.0616)		(.0617)
Δ%Faces Female $_{i,t}$	$.4897^{***}$		$.4911^{***}$
	(.0336)		(.0339)
Δ%Faces Young Adult$_{i,t}$	.0578		.0566
	(.0565)		(.0565)
ΔTweets Angry%$_{i,t}$		$-.0722$	.0039
		(.0472)	(.0328)
ΔTweets Joy%$_{i,t}$		$-.0776^{***}$	.0071
		(.0239)	(.0143)
ΔTweets Fear%$_{i,t}$		$-.0342$	$-.0169$
		(.0410)	(.0261)
ΔTweets Sadness%$_{i,t}$		$-.0032$	$-.0583^{*}$
		(.0359)	(.0342)
N	1,186	1,186	1,186
Adjusted R^2	.4530	.0008	.4518

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

City-clustered standard errors in parentheses.

Summary

- Event data have a long history in political science and the study of conflict.
- Text dominates, but images are useful.

The future of event data is bright.

The future of event data is bright.

Thank you.

APPENDIX

IMAGES AS DATA

Data - Geolocated Images

- Have downloaded (and continue to) approximately 5 million tweets per day with latitude and longitude coordinates.
- Search specific countries, dates with known protests (**A**).
- Download all images from **A**.
- Use image classifier to identify protest images **B**.
- Use separate scene and face classifiers to identify people, objects, actions in **B**.

Table: Image Data Pipeline

	Input	Output
	Protest Detection	
1. Images from Google	Keywords	100,000 images
2. Develop protest image classifier	Images from Step 1	Trained CNN
3. Protest images from Twitter corpus	Model from Step 2	115,060 images
	Scene Classification	
4. Annotate 40,746 images	Stratified sample from Step 3	Training set of images, 11 labels
5. Develop CNN	Training data from Step 4	Scene detection model
	Face Classification	
6. FairFace classifier (Xi and Joo 2019)	Yahoo YFCC100M	Gender, age, and race estimates
7. Locate faces in images	Step 3 data	Age range, gender, race

- More detail.
- On the poverty of existing facial classifiers.

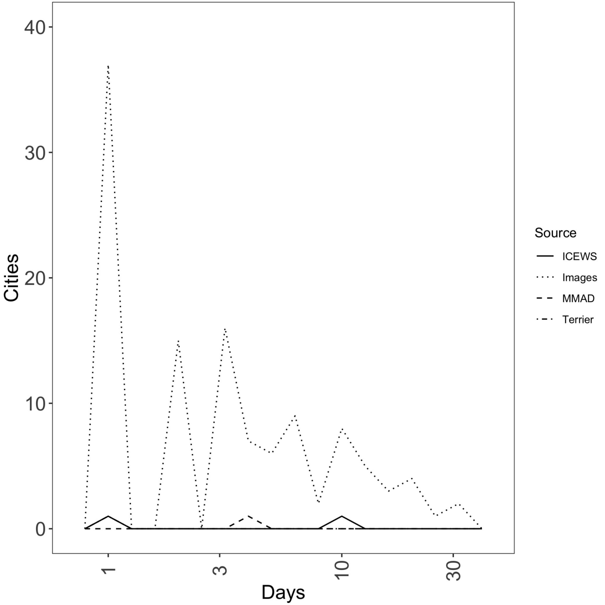
List of visual attributes.

Attribute	Description	Event Data Feature
Protester Violence	How violent protesters are.	Magnitude
State Violence	How violent the state is.	Magnitude
Gender	**Is the face male or female?**	Attributes
Race	**Is the face White, Middle Eastern, East Asian, Southeast Asian, Black, Indian, or Latino?**	Attributes
Age	**0-2, 3-9, 10-19, . . . , 70**	Attributes
Face	**Presence of a face.**	Attributes, magnitude
Group 20	There are roughly more than 20 people in the scene.	Magnitude
Group 100	There are roughly more than 100 people in the scene.	Magnitude
## Children	Children are in the scene.	Attributes
## Shout	One or more people shouting.	Magnitude
## Photo	Protesters holding signs or a photograph of a person (politicians or celebrities).	Magnitude
## Sign	Protesters holding a visual sign (on paper, panel, or wood).	Magnitude

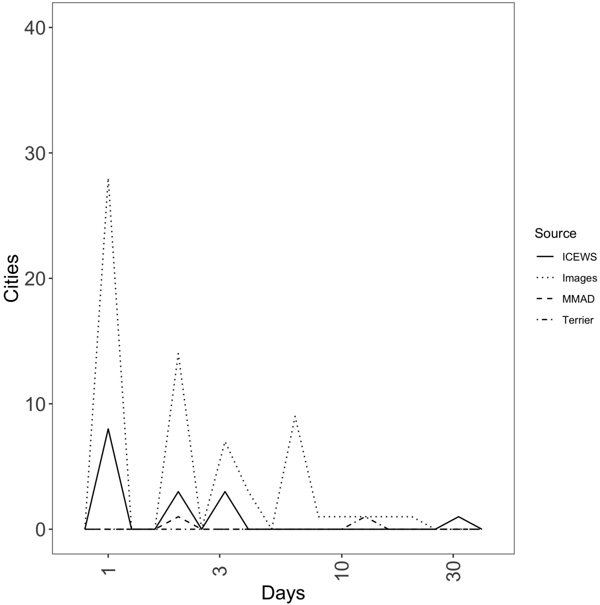
Results Summary

1. More cities and days with protest.
2. More event detail.
3. Different source of bias.
4. Regression results stronger than the text datasets.

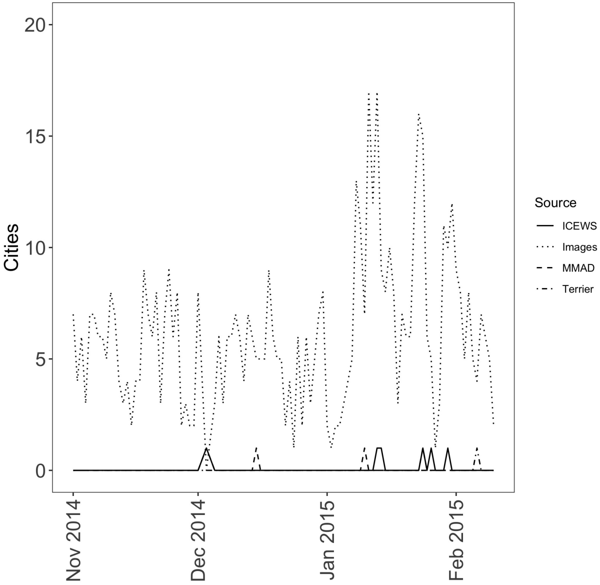
Results - Geographic Bias & Protest



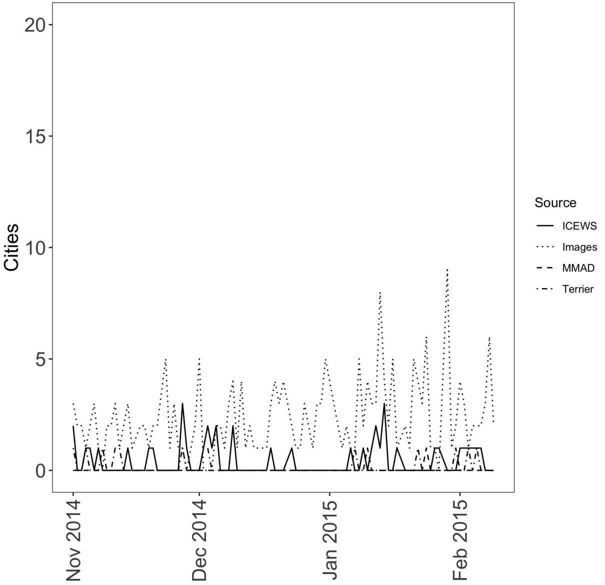
Results - Geographic Bias & Repression



Results - Novelty Bias & Protest



Results - Novelty Bias & Repression



Less Bias Means Fewer False Negatives

New event: Merida, 04.02.2015

Less Bias Means Fewer False Negatives

New event: Valencia, 06.11.2014

Images Mean More Detail on Magnitude

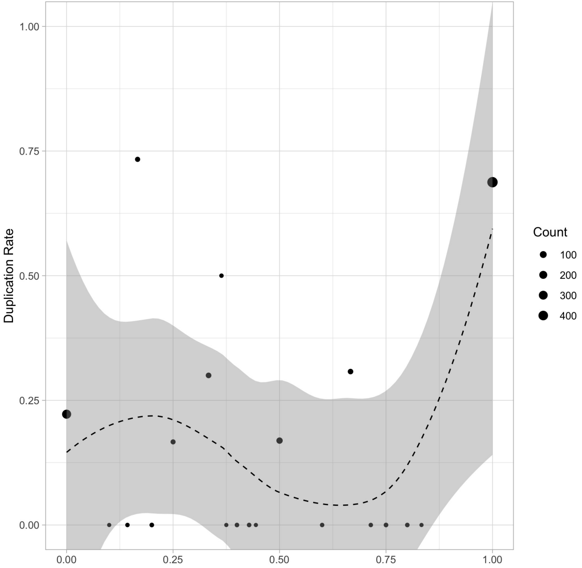
ICEWS records Caracas protest on 24.01.2015 but no information on protester behavior.

Images Mean More Detail on Magnitude

ICEWS records Caracas protest on 24.01.2015 but no information on police activity.

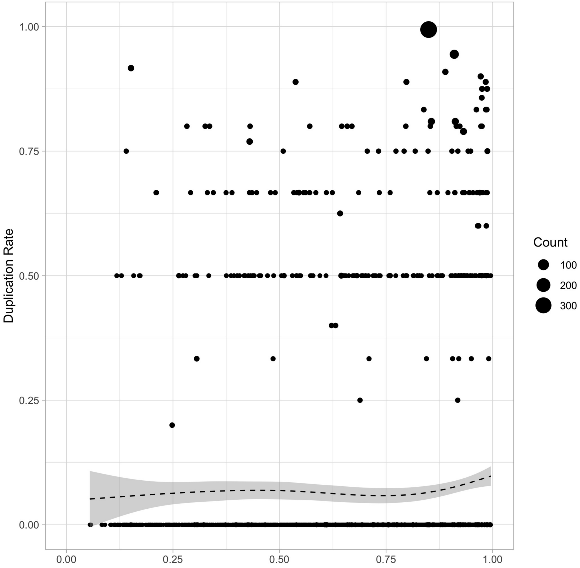
Results - Selection Bias

Duplication driven by photos with women, not violence.



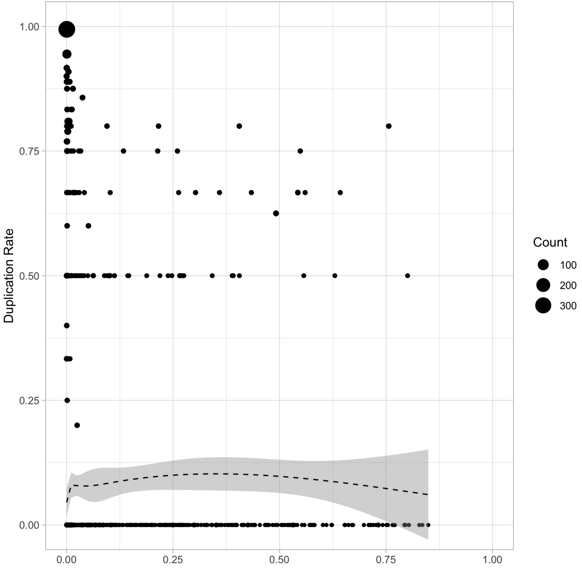
Results - Selection Bias

Duplication not driven by photos of groups.



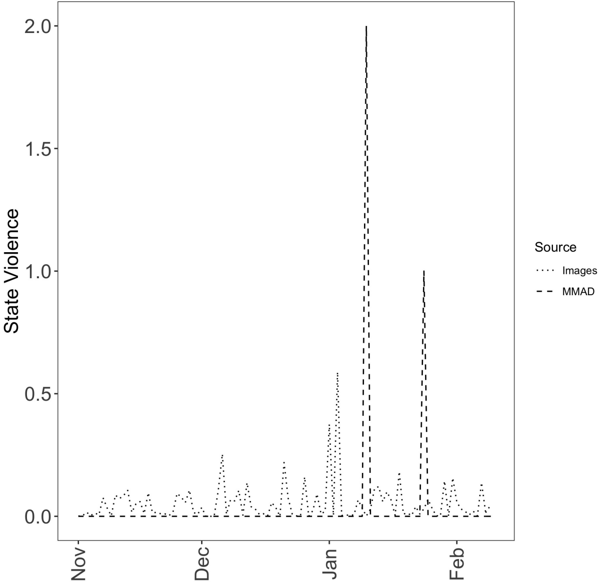
Results - Selection Bias

No preference for state or protester violence.



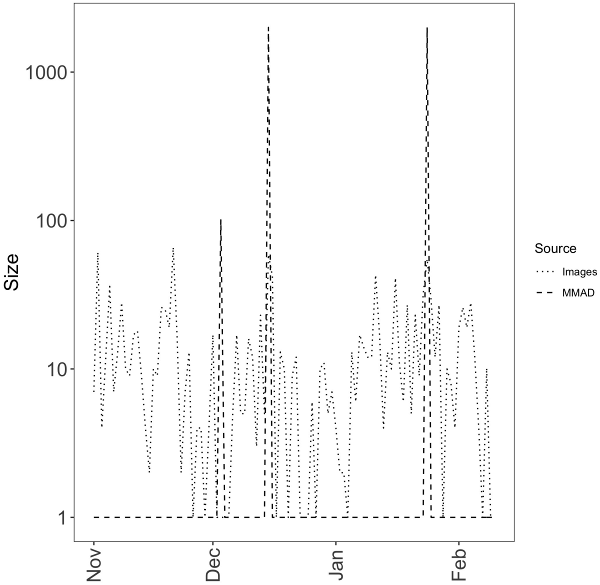
Results - State Violence

MMAD provides the closest comparison to image data. R = .08.



Results - Size

MMAD provides the closest comparison to image data. R=.98.



figures/imageAsData_regression2.png

Data from Social Media Images

Identify Protest Images

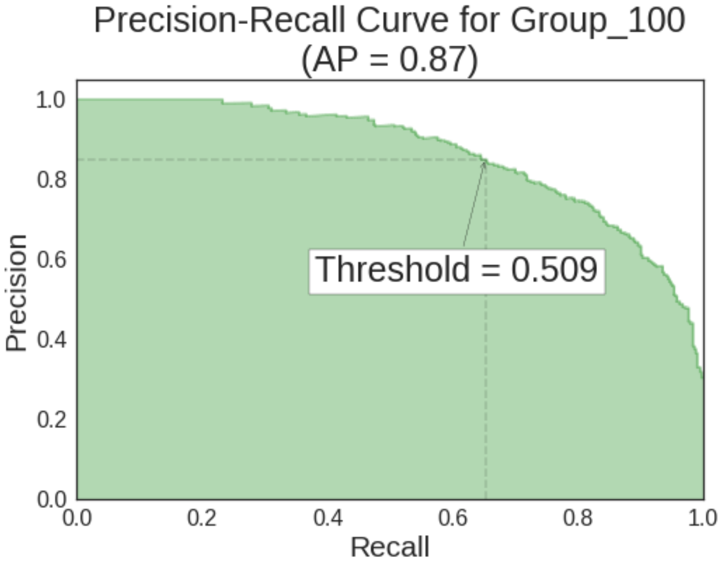
- Google Image to identify protest (10,000).
- Train convolutional neural network.
- Geolocated tweets from Twitter streaming API, find all images in tweets from protest periods.
- From these images, identify protest images ($p \geq .6$, 115,060).

Protest Data from Images

Identify Protest Image Features

- Amazon Mechanical Turk, label 40,764 of the 115,060 protest images for twelve variables.
- Build new convolutional neural network for attributes. Train ("transfer learning") face model using UTK.
- Apply CNN to each protest image.
- Make label estimates binary.
- Aggregate.

Generating Binary Attributes



Gradient-weighted Class Activation Mapping (Grad-CAM)

Figure: Police

CLASSIFIER DETAILS

Generating Binary Attributes

Table: Labels and Thresholds

Label	Threshold
Protester Violence	.021
State Violence	.01
Police	.937
Fire	.37
Gender	Highest probability gender
Race	Highest probability race
Children	.15
Group 20	.725
Group 100	.509
Face	A solved problem
Shout	.355
Photo	.815
Flag	.187
Night	.359
Sign	.744

Figure: Sample Classifier Estimates by Category: Images are ordered by their classification scores. (Blue lines mark the exact classification scores of corresponding images)

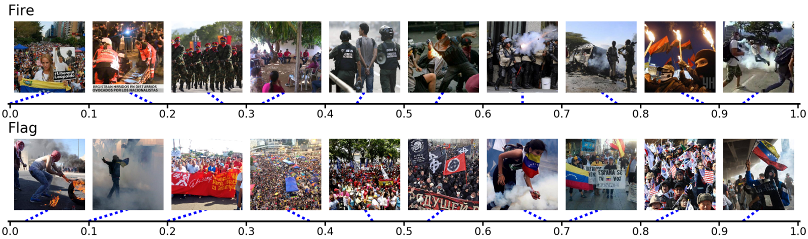


Figure: Sample Classifier Estimates by Category: Images are ordered by their classification scores. (Blue lines mark the exact classification scores of corresponding images)

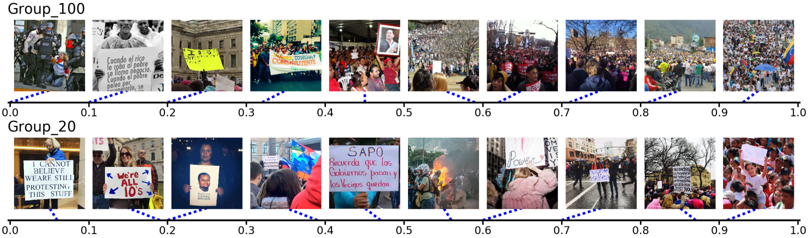


Figure: Model Performance

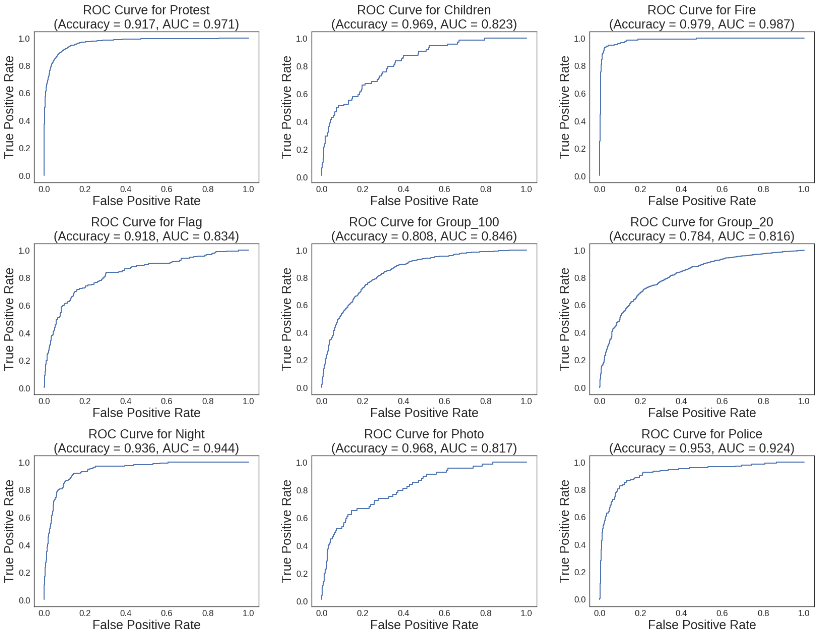


Figure: Model Performance

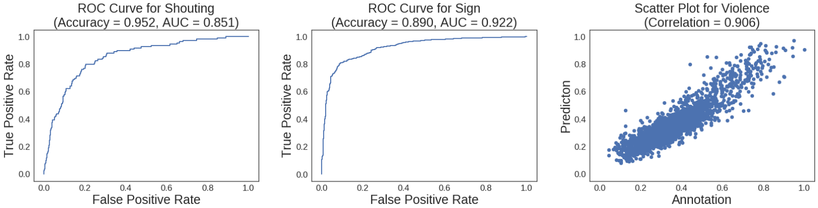


Figure: Grad-CAM

Figure: Sample Classifier Estimates by Category

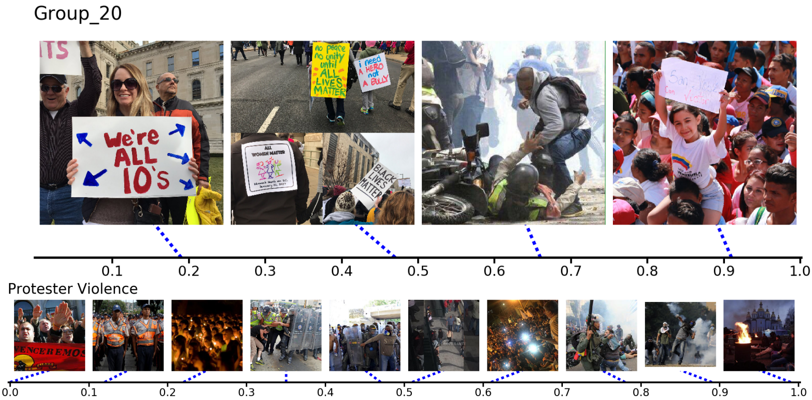
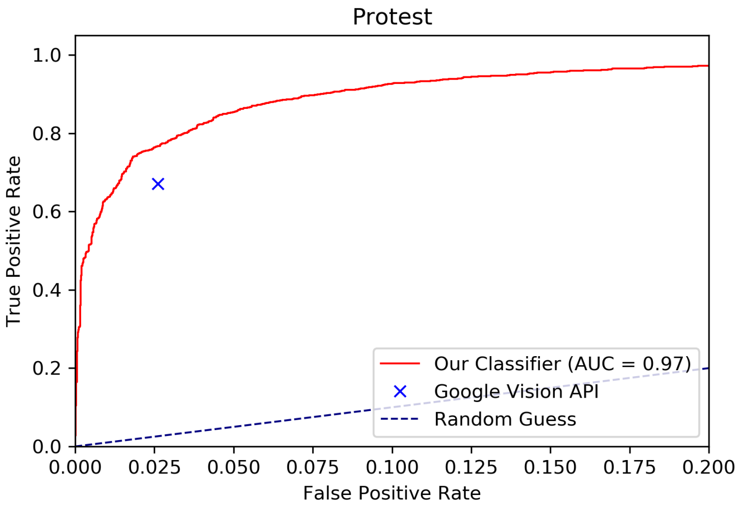


Figure: Grad-CAM

Comparing Classifier to Google Vision API

Figure: Classification performance comparison between our model and a public model offered by Google Vision API.



POVERTY OF FACIAL CLASSIFIERS

Figure: FairFace and Other Datasets

Table 1: Statistics of Face Attribute Datasets

Name	Source	# of faces	In-the-wild?	Age	Gender	Race Annotation							
						White*		Asian*		Bla-ck	Ind-ian	Lat-ino	Balan-ced?
						W	ME	E	SE				
PPB [6]	Gov. Official Profiles	1K		✓	✓	**Skin color prediction							
MORPH [44]	Public Data	55K		✓	✓	merged				✓		✓	no
PubFig [30]	Celebrity	13K	✓	Model generated predictions									no
IMDB-WIKI [45]	IMDB, WIKI	500K	✓	✓	✓								no
FotW [11]	Flickr	25K	✓	✓	✓								yes
CACD [9]	celebrity	160K	✓	✓									no
DiF [39]	Flickr	1M	✓	✓	✓	**Skin color prediction							
†CelebA [37]	CelebFace [53, 54] LFW [21]	200K	✓	✓	✓								no
LFW+ [14]	LFW [21] (Newspapers)	15K	✓	✓	✓	merged		merged					no
†LFWA+ [37]	LFW [21] (Newspapers)	13K	✓		✓	merged		merged		✓	✓		no
†UTKFace [72]	MORPH, CACD Web	20K	✓	✓	✓	merged		merged		✓	✓		yes
FairFace (Ours)	Flickr, Twitter Newspapers, Web	100K	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	yes

*FairFace (Ours) also defines East (E) Asian, Southeast (SE) Asian, Middle Eastern (ME), and Western (W) White.

Figure: Sample Faces

(a) FairFace (b) UTKFace (c) LFWA+ (d) CelebA

Figure 1: Random Samples from Face Attribute Datasets.

Figure: Racial Balance

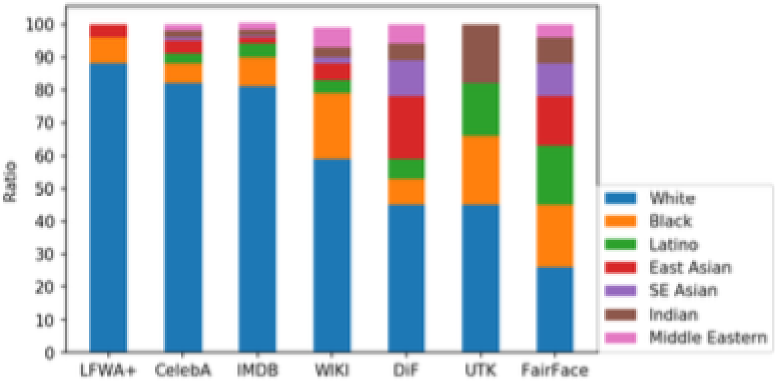


Figure: Results

Table 2: Cross-Dataset Classification Accuracy on White Race.

		Tested on								
		Race			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.981**	.959	.971	**.944**	**.948**	**.929**	.976	**.574**	.577
	UTKFace	.775	.975	.886	.851	.943	.897	.963	.285	**.649**
	LFWA+	.945	**.981**	**.997**	.758	.851	.924	.932	-	-
	CelebA	-	-	-	.866	.893	.921	**.985**	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).

Table 3: Cross-Dataset Classification Accuracy on non-White Races.

		Tested on								
		Race			Gender				Age	
		FairFace	UTKFace	LFWA+	FairFace	UTKFace	LFWA+	CelebA*	FairFace	UTKFace
Trained on	FairFace	**.743**	.549	.396	**.932**	**.929**	**.895**	.976	**.588**	.609
	UTKFace	.488	**.591**	.316	.823	.907	.877	.963	.328	**.685**
	LFWA+	.437	.353	**.439**	.734	.852	.887	.932	-	-
	CelebA	-	-	-	.794	.886	.884	**.985**	-	-

* CelebA doesn't provide race annotations. The result was obtained from the whole set (white and non-white).

MODEL DETAIL

Protest Periods (**A**)

Table: Protest Periods

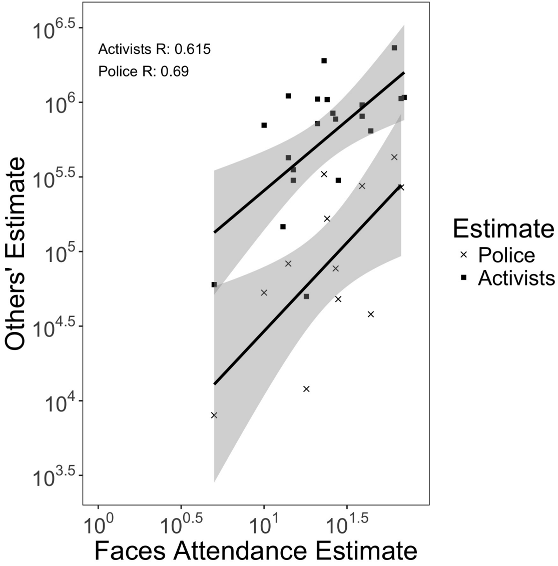
Country	Start	End	Issue	Protest Images/Day
Belarus	02.18.2017	05.02.2017	Unemployment Tax	1.64
Burundi	04.01.2015	12.01.2015	Elections	.06
Cameroon	11.01.2016	12.01.2017	Bilingualism	.06
Egypt	06.01.2017	06.31.2017	Islands to Saudi Arabia	1.68
Gabon	08.20.2016	09.27.2016	Elections	.31
Hong Kong	**2014.09.18**	**2014.12.23**	**China Reforms**	**3.41**
Pakistan	11.01.2017	11.30.2017	Religious Defamation	2.3
Russia	***03.12.2017***	***04.26.2017***	***Corruption***	***10.42***
Catalonia, Spain	***2017.09.01***	***2017.12.31***	***Secession***	***42.46***
South Korea	**2016.10.20**	**2017.03.14**	**Anti-incumbency**	**7.04**
Togo	08.01.2017	12.01.2017	Anti-incumbency	.23
Ukraine	11.21.2013	03.21.2014	European Integration	2.44
United States	2017.01.20	2017.01.22	Women's March	9,034.33
Venezuela	**2014.03.27**	**2015.02.08**	**Grievances**	**25.42**
Venezuela	**12.29.2016**	**12.17.2017**	**Anti-Maduro**	**20.01**

Summary Statistics

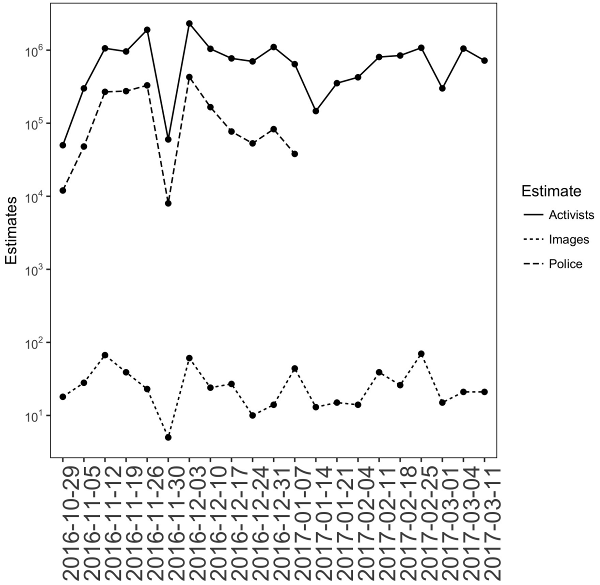
Statistic	N	Mean	St. Dev.	Min	Max
Protest Size$_{i,t}$	3,231	3.10	16.12	0	627.00
Perceived Protester Violence$_{i,t-1}$	3,203	0.03	0.13	0.00	1.00
Perceived State Violence$_{i,t-1}$	3,203	0.02	0.08	0.00	0.94
Police$_{i,t-1}$	3,203	0.01	0.10	0.00	4.00
Fire$_{i,t-1}$	3,203	0.09	0.47	0.00	7.00
Gender Diversity$_{i,t-1}$	3,203	0.07	0.18	0.00	0.69
Race Diversity$_{i,t-1}$	3,203	0.09	0.21	0.00	1.10
Any Child$_{i,t-1}$	3,203	0.03	0.22	0.00	3.00
Small Group$_{i,t-1}$	3,203	1.26	5.35	0.00	168.00
Large Group$_{i,t-1}$	3,203	0.34	1.75	0.00	50.00

	Log(Protest Size$_{i,t}$)	Perc. Prtstr. Violence$_{i,t-1}$	Perc. Stt. Violence$_{i,t-1}$	Police$_{i,t-1}$	Fire$_{i,t-1}$	Small Grp.$_{i,t-1}$	Large Grp.$_{i,t-1}$	Gend. Div.$_{i,t-1}$	Race Div.$_{i,t-1}$	Any Child$_{i,t-1}$
Any Child$_{i,t-1}$	0.21	0.05	0.03	0.03	0.19	0.25	0.18	0.09	0.08	1
Race Div.$_{i,t-1}$	0.22	0.09	0.11	0.04	0.13	0.19	0.14	0.48	1	0.08
Gend. Div.$_{i,t-1}$	0.23	0.04	0.05	0.02	0.12	0.22	0.14	1	0.48	0.09
Large Grp.$_{i,t-1}$	0.3	0.02	0.01	0.15	0.2	0.83	1	0.14	0.14	0.18
Small Grp.$_{i,t-1}$	0.4	0.04	0.05	0.09	0.28	1	0.83	0.22	0.19	0.25
Fire$_{i,t-1}$	0.3	0.52	0.13	0.03	1	0.28	0.2	0.12	0.13	0.19
Police$_{i,t-1}$	0.08	0.05	0.21	1	0.03	0.09	0.15	0.02	0.04	0.03
Perc. Stt. Violence$_{i,t-1}$	0.11	0.45	1	0.21	0.13	0.05	0.01	0.05	0.11	0.03
Perc. Prtstr. Violence$_{i,t-1}$	0.1	1	0.45	0.05	0.52	0.04	0.02	0.04	0.09	0.05
Log(Protest Size$_{i,t}$)	1	0.1	0.11	0.08	0.3	0.4	0.3	0.23	0.22	0.21

Validating Number of Faces, South Korea

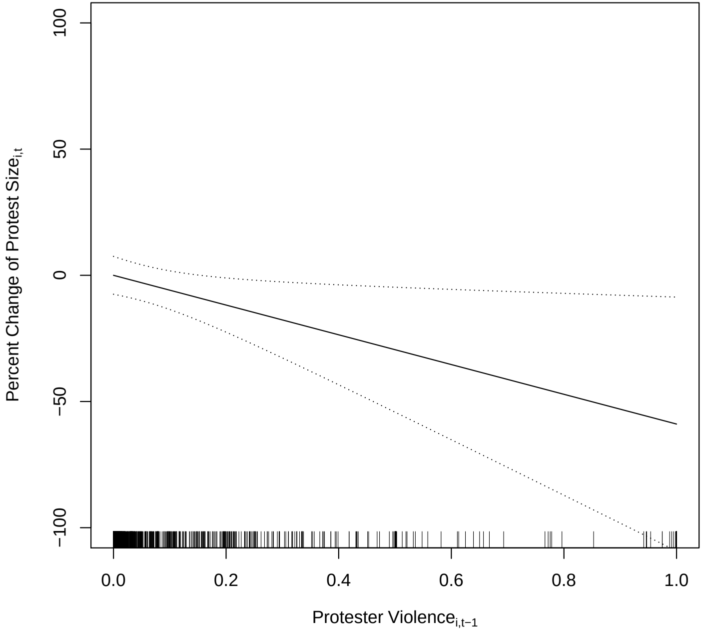


Validating Protest Size, South Korea

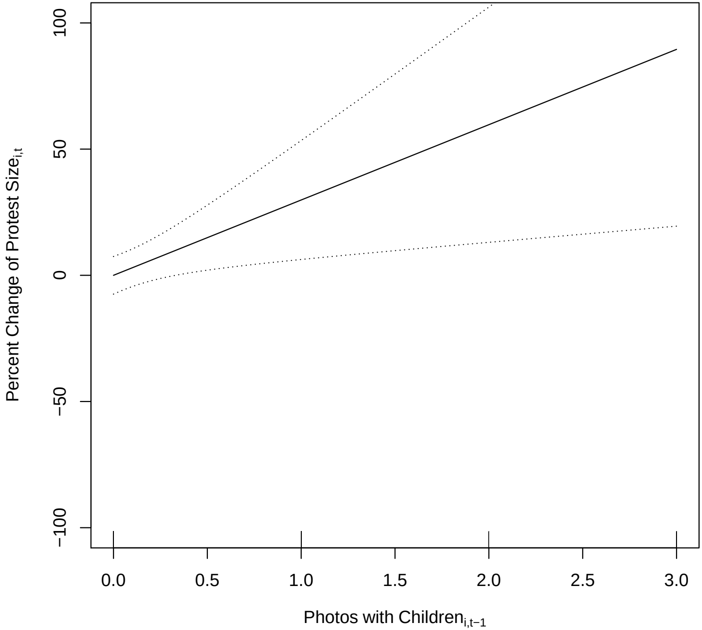


ADDITIONAL MODEL RESULTS

Marginal Effects - Protester Violence



Marginal Effects - Children



Histogram of Pairwise Image Distances

Euclidean distance between pairs of 1,000 feature vectors (ResNet50). Duplicate if $d < .2$.